Statistical inference in mixture models with random effects

by

Daniel Meddings

Thesis

Submitted to University College London

for the degree of

Doctor of Philosophy

April 2014

Department of Statistical Science

University College London

To my mother and father - the greatest people I know, and to whom I owe everything. In particular I thank them for saving me when it turned out one of the happiest phases of my life was built on false assumptions. Some people could learn a lot from them in terms of how to live life with integrity.

Starka virna, vestilie
Obadeea, obadeea
Starka, virna, vestilie
Obadeea, monye.

Stala, stoita, stonga raer
O, whit says du da bunshka baer?
O, whit says du da bunshka baer?
Litra mae vee drengie.

Saina, papa wara
Obadeea, obadeea
Saina, papa wara
Obadeea, monye.

Strong winds blow from the west
they may bring trouble and damage the boat, men.

Make sure the mast is rigged securely.
Do you think the boat will be able to carry her sail?
I'm pleased with it, boys.

Holy Father, take care of us
There may be trouble and the boat may be damaged.

Unst Boat Song: Shetland's oldest surviving song. Words from the version by Fair Isle family group Fridarey from their album "Across the Waters".

I Daniel Meddings confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm this has been indicated in the thesis.

Abstract

There is currently no existing asymptotic theory for statistical inference on the maximum likelihood estimators of the parameters in a mixture of linear mixed models (MLMMs). Despite this many researchers assume the estimators are asymptotically normally distributed with covariance matrix given by the inverse of the information matrix. Mixture models create new identifiability problems that are not inherited from the underlying linear mixed model (LMM), and this subject has not been investigated for these models. Since identifiability is a prerequisite for the existence of a consistent estimator of the model parameters, then this is an important area of research that has been neglected.

MLMMs are mixture models with random effects, and they are typically used in medical and genetics settings where random heterogeneity in repeated measures data are observed between measurement units (people, genes), but where it is assumed the units belong to one and only one of a finite number of sub-populations or components. This is expressed probabalistically by using a sub-population specific probability distribution function which are often called the component distribution functions. This thesis is motivated by the belief that the use of MLMMs in applied settings such as these is being held back by the lack of development of the statistical inference framework. Specifically this thesis has the following primary objectives;

i To investigate the quality of statistical inference provided by different information matrix based methods of confidence interval construction.

ii To investigate the impact of component distribution function separation on the quality of statistical inference, and to propose a new method to quantify this separation.

iii To determine sufficient conditions for identifiability of MLMMs.

Acknowledgements

There are not enough superlatives to describe how wonderful and important has been the love and support given to me by my parents. I could not have got this far without their support.

I would like to give a huge thanks to my supervisor Christian Hennig, who for almost four years has suffered with good grace and humour my many stupid questions. During this time I have learnt a lot about my subject, but perhaps more importantly I have learnt *how* to learn: how to ask the right questions; how to set about answering these; how to use intuition to guide this process, and often philosophy to reflect on what has been discovered. Furthermore under his tutelage I have learnt what it means *to* learn: often to accept that the results of the academic endevour in question are more complex than hoped for, and to resist the urge to accept as ”correct” one school of thought over another with respect to an often arbitrary ”truth”. The lesson instead has always been to see the merits in all the information that has been obtained, to keep an open mind so that ultimately we can do good science and have fun along the way - in this respect I cannot imagine a better teacher.

I would like to thank Roger Stafford, who for no good reason other than he likes helping people with mathematical problems, provided me some wonderfully elegant geometric analyses to help my understanding of the identifiability problems I studied in this thesis. I would also like to thank Allan Hackshaw of the University College London Cancer Trials Centre for the generous provision of the datasets whose analyses provided a very useful applied section in this thesis to complement the theoretical sections. I would also like to thank Allan for his time which he so generously gave up, and I hope he found our meetings as interesting as I did.

Finally I would like to thank the statistics department at UCL, and its various members, in particular Tom Fearn, for their tremendous support

over the last few years. During this time I moved progressively further from London, and hence the department saw ever decreasing amounts of me. Never once was this a problem, even when I asked for some extra funding, and I am exceedingly grateful for that.

Notation

$\mathbb{N}$	Set of natural numbers $\mathbb{N} := \{0, 1, 2, ...\}$
$\mathbb{N}^+$	Set of natural numbers excluding zero $\mathbb{N}^+ := \{1, 2, ...\}$
I_n	Set of integers $I_n := \{1, ..., n\}$, $n \in \mathbb{N}^+$
$\mathbb{R}^n$	Euclidean n-space for $n \in \mathbb{N}^+$
$\mathbb{R}^{m \times n}$	Set of real $m \times n$ matrices for $m, n \in \mathbb{N}^+$
$\left\{_m\, a_{ij}\right\}_{i,j=1,1}^{I,J}$	$I \times J$ matrix with elements a_{ij}, $i \in I_I$, $j \in I_J$
$\left\{_c\, a_i\right\}_{i=1}^{I}$	$I \times 1$ column vector with elements a_i, $i \in I_I$
$\left\{_r\, a_j\right\}_{j=1}^{J}$	$1 \times J$ row vector with elements a_j, $j \in I_J$
$\mathrm{tr}(\boldsymbol{A})$	For an $m \times m$ matrix $\boldsymbol{A}$ $\mathrm{tr}(\boldsymbol{A})$, the trace of $\boldsymbol{A}$, is the sum of the diagonal elements of $\boldsymbol{A}$
$\mathrm{vec}(\boldsymbol{A})$	If the $m \times n$ matrix $\boldsymbol{A}$ has $\boldsymbol{a}_i \in \mathbb{R}^m$, $i \in I_n$, as its i^{th}column, then $\mathrm{vec}(\boldsymbol{A})$ is the $mn \times 1$ vector defined by $\mathrm{vec}(\boldsymbol{A}) := \left\{_c\, \boldsymbol{a}_i\right\}_{i=1}^{n}$
$\mathrm{v}(\boldsymbol{A})$	Transformation of a $m \times m$ matrix $\boldsymbol{A}$ into a $m(m+1)/2 \times 1$ vector obtained by deleting all the elements of $\mathrm{vec}(\boldsymbol{A})$ that are above the diagonal of $\boldsymbol{A}$
$\widetilde{\boldsymbol{D}}_m$	If $\boldsymbol{A}$ is a symmetric $m \times m$ matrix then the duplication matrix $\widetilde{\boldsymbol{D}}_q$ is a $m^2 \times m(m+1)$ matrix that transforms $\mathrm{v}(\boldsymbol{A})$ into $\mathrm{vec}(\boldsymbol{A})$ via the relationship $\widetilde{\boldsymbol{D}}_m \mathrm{v}(\boldsymbol{A}) = \mathrm{vec}(\boldsymbol{A})$
$\otimes$	Kronecker product of two matrices defined so that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are $m \times n$ and $p \times q$ matrices respectively, and where $\boldsymbol{A}$ has elements a_{ij}, $i \in I_m$, $j \in I_n$, then $\boldsymbol{A} \otimes \boldsymbol{B}$ is the $mp \times nq$ matrix defined by $\boldsymbol{A} \otimes \boldsymbol{B} := \left\{_m\, a_{ij} B\right\}_{i,j=1,1}^{m,n}$

$\boldsymbol{D}_{\mathrm{vec}(\boldsymbol{X})}\left(f(\boldsymbol{X})\right)$ For $\boldsymbol{f} : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^{n\times q}$, and for $\boldsymbol{X} := \big\{_m\, x_{i,j}\big\}_{i,j=1,1}^{n,q}$, $\boldsymbol{D}_{\mathrm{vec}(\boldsymbol{X})}\left(f(\boldsymbol{X})\right)$ is the $1 \times nq$ vector of partial derivatives of f with respect to $\mathrm{vec}(\boldsymbol{X})$ defined as $\boldsymbol{D}_{\mathrm{vec}(\boldsymbol{X})}\left(f(\boldsymbol{X})\right) := [\mathrm{vec}\big\{_m\, \partial f/\partial x_{ij}\big\}_{i,j=1,1}^{n,q}]^\intercal$

$\boldsymbol{D}_{\boldsymbol{x}}\left(\boldsymbol{f}(\boldsymbol{x})\right)$ For $\boldsymbol{f} : S \longrightarrow \mathbb{R}^m$, $S \subseteq \mathbb{R}^n$, and for $\boldsymbol{x} := \big\{_c\, x_i\big\}_{i=1}^{n}$, $\boldsymbol{f} := \big\{_c\, f_j(\boldsymbol{x})\big\}_{j=1}^{m}$, $f_j : S \longrightarrow \mathbb{R}$, $j \in I_m$, $\boldsymbol{D}_{\boldsymbol{x}}\left(\boldsymbol{f}(\boldsymbol{x})\right)$ is the $m \times n$ Jacobian matrix of partial derivatives of $\boldsymbol{f}$ with respect to $\boldsymbol{x}$ defined as $\boldsymbol{D}_{\boldsymbol{x}}\left(\boldsymbol{f}(\boldsymbol{x})\right) := \big\{_m\, \partial f_j/\partial x_i\big\}_{j,i=1,1}^{m,n}$

$\boldsymbol{H}_{\mathrm{vec}(\boldsymbol{X})}\left(f(\boldsymbol{X})\right)$ For $\boldsymbol{f} : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^{n\times q}$, and for $\boldsymbol{X} := \big\{_m\, x_{i,j}\big\}_{i,j=1,1}^{n,q}$, $\boldsymbol{H}_{\mathrm{vec}(\boldsymbol{X})}\left(f(\boldsymbol{X})\right)$ is the $(nq)^2 \times (nq)^2$ Hessian matrix of second partial derivatives of $\boldsymbol{f}$ with respect to $\boldsymbol{x}$ given by $\boldsymbol{D}_{\boldsymbol{x}}\left(\boldsymbol{g}(\boldsymbol{x})\right)$ where $\boldsymbol{g}(\boldsymbol{x}) := [\boldsymbol{D}_{\mathrm{vec}(\boldsymbol{X})}\left(f(\boldsymbol{X})\right)]^\intercal$

Contents

List of Figures

List of Tables

1

Introduction

This thesis is focused on multivariate data $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$, $i \in I_N := \{1, ..., N\}$, where the distribution of the random vector $\boldsymbol{Y}_i$ is a finite mixture of G multivariate normal distributions, specifically where the distribution is induced by a Mixture of Linear Mixed Models (MLMMs), which as the name suggests consists of a finite set of G Linear Mixed Models (LMMs). The research objectives of this thesis are focused on methods to construct confidence intervals about the model parameter estimators in order to perform statistical inference on the model parameters, and on identifiability problems associated with the mixture distribution.

For the general multivariate mixture distribution (not necessarily induced by a model for the vectors $\boldsymbol{Y}_i$) we observe a random sample $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$ of N vectors $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$, $i \in I_N := \{1, ..., N\}$, where $\boldsymbol{Y}_i$ has a distribution with a density function that is a finite mixture of density functions

$$f_i(\boldsymbol{y}_i|\boldsymbol{\theta}) = \sum_{j=1}^{G} \boldsymbol{\pi}_j f_{ij}(\boldsymbol{y}_i|\boldsymbol{\theta}_j), \tag{1.1}$$

where $f_{ig}(\cdot|\boldsymbol{\theta}_g)$, $g \in I_G := \{1, ..., G\}$, is a density function with parameter $\boldsymbol{\theta}_g \in \Theta_g \subseteq \mathbb{R}^{n_\theta}$, $\boldsymbol{\theta} := \left[\boldsymbol{\theta}_1^\intercal, ..., \boldsymbol{\theta}_G^\intercal, \boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_G\right]^\intercal \in \Theta$, where

$$\Theta = \left\{ (\boldsymbol{\theta}_1^\intercal, ..., \boldsymbol{\theta}_G^\intercal, \pi_1, ..., \pi_G)^\intercal : \sum_{j=1}^{G} \pi_j = 1, \pi_j \geq 0, \boldsymbol{\theta}_j \in \Theta_j, j = 1, ..., G \right\}. \tag{1.2}$$

For MLMMs the random vectors $\boldsymbol{Y}_i$, are typically interpreted as N measurement units (entities, objects etc.) on which a "response" vector is obtained, where the within-unit responses are correlated whilst the between unit observations are independent. In chapter 2 we describe an interpretation of LMMs which we will call the hierarchical interpretation, whereby we associate with each $\boldsymbol{Y}_i$ a normally distributed vector of random effects $\boldsymbol{U}_i \in \mathbb{R}^q$ with mean zero and covariance matrix $\boldsymbol{D}_g$, where conditional on $\boldsymbol{U}_i = \boldsymbol{u}_i$, and if unit i belongs to component $g \in I_G$, then $\boldsymbol{Y}_i$ follows a Linear Mixed Model (LMM) given by

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_g + \boldsymbol{Z}_i \boldsymbol{u}_i + \boldsymbol{e}_i, \tag{1.3}$$

where $\boldsymbol{\beta}_g \in \mathbb{R}^p$ is a vector of fixed effects, $\boldsymbol{X}_i$ is a $n_i \times p$ matrix of covariate data, and $\boldsymbol{e}_i \in \mathbb{R}^{n_i}$ is a normally distributed vector of errors with mean zero, and covariance matrix $\sigma_g^2 \boldsymbol{C}_i(\boldsymbol{\phi}_g)$. By integrating the conditional distribution of $\boldsymbol{Y}_i | \boldsymbol{U}_i = \boldsymbol{u}_i$ (Verbeke and Molenberghs, 2009, p 24) with respect to $\boldsymbol{U}_i$ we obtain the marginal distribution of $\boldsymbol{Y}_i$ that has the g^{th} component density function $f_{ig}(\cdot|\boldsymbol{\theta}_g)$ in (1.1), where the covariance matrix $\boldsymbol{V}_i(\boldsymbol{\zeta}_g)$ for $\boldsymbol{Y}_i$ has the following form

$$\boldsymbol{V}_i(\boldsymbol{\zeta}_g) = \boldsymbol{Z}_i \boldsymbol{D}_g \boldsymbol{Z}_i^{\intercal} + \sigma_g^2 \boldsymbol{C}_i(\boldsymbol{\phi}_g), \tag{1.4}$$

where $\boldsymbol{Z}_i$, $\boldsymbol{D}_g$, and $\boldsymbol{C}_i(\boldsymbol{\phi}_g)$ are $n_i \times q$, $q \times q$, and $n_i \times n_i$ matrices respectively, so that $\boldsymbol{\zeta}_g = (\boldsymbol{\psi}_g^{\intercal}, \sigma_g^2, \boldsymbol{\phi}_g^{\intercal})^{\intercal}$, $\boldsymbol{\psi}_g = \mathrm{v}(\boldsymbol{D}_g)$, is a vector of covariance parameters. Here the $\mathrm{v}(\cdot)$ function stacks the supra-diagonal elements of its matrix argument one on top of each other. Thus $\boldsymbol{\psi}_g$ is a $q(q+1) \times 1$ vector of the unique elements of $\boldsymbol{D}_g$. Equation 1.4 shows that the random effects induce a covariance structure for the n_i components of $\boldsymbol{Y}_i$. When $G = 1$ we shall say $\boldsymbol{Y}_i$ follows a 1-component model, or a LMM, and we shall use these two terms interchangeably.

Historically there has been an interest in mixture distributions where the underlying model is an ordinary regression model $y_i = \boldsymbol{x}_i^{\intercal} \boldsymbol{\beta}_g + e_i$ for scalar responses y_i, and covariate vector $\boldsymbol{x}_i \in \mathbb{R}^p$, where the e_i are normally distributed errors with variance σ_g^2. By setting $n_i = 1$ for all $i \in I_N$ these component-specific models are special cases of the General Linear Model (GLM)

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_g + \boldsymbol{e}_i, \tag{1.5}$$

where $\boldsymbol{Y}_i \in \mathbb{R}_{n_i}$, $\boldsymbol{X}_i$ is a $n_i \times p$ matrix of covariate data, and $\boldsymbol{e}_i \in \mathbb{R}^{n_i}$ is a normally distributed vector of errors with mean zero, and covariance matrix $\sigma_g^2 \boldsymbol{I}_N$. In this thesis we shall refer to a mixture of the model in (1.5) as a Mixture of Linear Models (MLM).

The mixture models we have introduced here are comprised of component specific regression models, which for a 1-component model will give rise to a sample $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$ of responses that are independent by assumption, but because of the covariate data, will in general not be identically distributed. For example if the $\boldsymbol{X}_i$ contain the age of subjects in a medical study then almost certainly the ages of these subjects will be different and so too will the mean vectors $\boldsymbol{X}_i\boldsymbol{\beta}_g$. For this reason in general a sample of responses that follow a MLMM or MLM will also be independent but not identically distributed. In contrast there has been huge interest historically in *iid* samples with a mixture density given in 1.1 but where no regression model is specified for the responses. To distinguish between the two approaches we will refer to the class of finite mixture densities arising from MLMMs and MLGMs as model generated finite mixture densities, or sometimes finite mixtures densities from MLMMs or MLGMs, and as per the convention we will refer to the class of finite mixture densities not arising from regression models simply as finite mixture densities.

Normally distributed responses $\boldsymbol{Y}_i$ that consist of correlated measurements are often described as clustered data because plotted against one or more of the p covariates the within-unit correlation of the observations within the response vectors means that the $N_T := \sum_{i=1}^{N} n_i$ total responses sometimes appear grouped together in clusters. Similarly for responses $\boldsymbol{Y}_i$ from a mixture of normal distributions, plotting the responses against one or more of the p covariates can sometimes show that the N_T total responses appear to be in clusters. Indeed for this reason MGLMs with $n_i = 1$ for all $i \in I_N$ are often referred to as clusterwise regression models, and the classification of units to components or clusters that occurs when estimating the model is referred to as model-based clustering.

In this thesis we will not have much need to refer to this grouping or clustering effect (regardless of the cause), however we will often discuss the process of assignment of units to components by a mixture model. Historically this has been called clustering, but to avoid confusion we will refer to this either as the assignment of units to components, or simply as component estimation. The justification for this latter term is that correctly determining the component memberships should lead to an accurate estimation of the mixture distribution means and variances which completely characterises normal distributions, and hence the probabalistic behaviour of the components.

LMMs are used primarily for data where the n_i repeated observations in $\boldsymbol{Y}_i$ represent observations on unit i taken under different experimental conditions or at different times (longitudinal data). The primary aim of using LMMs for these data is to employ the covariance matrices $\boldsymbol{V}_i(\boldsymbol{\zeta})$ to model the within-unit correlation in the response vectors in order to obtain unbiased estimates of $\boldsymbol{\beta}$, and/or estimates of $\boldsymbol{\beta}$ with greater precision than are obtained by simply ignoring the correlation. Analogously for repeated measurements data distributed as a mixture of normals, the main motivation for the use of MLMMs is primarily to obtain unbiased and/or more precise estimates of the fixed effects compared to those obtained by simply ignoring the mixture by using a LMM. In contrast when component estimation rather than the parameters themselves are of primary concern then by definition (it is assumed sub-populations exist) mixture models or some other classification tool are used rather than LMMs or GLMs. Historically this has been the main motivation behind the use of clusterwise regression models which have been used extensively.

Two areas where MLMMs have been used is in medical and genetics settings. For example to analyse microarray data that consist of measurements on a large number of genes, where the genes were the units (Celeux et al. (2005)), and to analyse repeated measures data from patients in clinical trials, where the people are the units, for example Grün and Hornik (2011) and Xu and Hedeker (2002). For the genetic settings the main motivation for the use of mixture models is as a classification tool, whereas in the medical settings the main motivation is to obtain unbiased and/or more precise parameter estimates. One non-medical example of MLMMs is that of Coke and Tsao (2010) who apply MLMMs to electricity load series data, which are long time series of household electricity load values (households are the units) taken at hourly intervals over the time period of one year. The primary purpose of fitting a MLMM here was the component estimation rather than the parameter estimates, specifically the electricity company was interested in dividing their customers into groups that were homogeneous with respect to their electricity usage patterns.

We have described that plotting the responses from a LMM and a mixture model against one or more covariates can often reveal the clustering in the responses. For mixture models however the clusters in general will not, and indeed should not, be completely determined by any single covariate in the model, because this will cause numerical instabilities during parameter estimation due to the model being non-identifiable, or close to non-identifiable.

Of course this is an extreme example, but we would still expect similar problems if the range of covariate values within each component was very narrow rather than two single points such that the covariate "almost" identifies the two components. In this situation we might be close to having a non-identifiable mixture model. The main point is that whilst the values of the covariates should in combination serve to classify the units to the components, no single covariate should be able to perfectly, or almost perfectly achieve this classification. In more formal language we have that all of the covariate data in this example concentrate on two $(p-2)$-dimensional hyperplanes, and for clusterwise regression models Hennig (2000) has shown that this relationship between the number of components and hyperplanes can be used instrumentally in a sufficient condition for identifiability.

Using a counter example to identifiability Hennig (2000) shows for a two component model an example where the concentration of covariate data on two hyperplanes leads to a non-identifiable mixture model but at the same time identifiable one component models. Thus mixture models bring with them their own identifiability problems not inherited from the underlying model. In chapter 4 we investigate this identifiability problem in MLMMs, and in this respect prove two theorems establishing sufficient conditions for identifiability, and derive a corollary from one of these establishing sufficient conditions for identifiability for a MLMM with no autocorrelation structure in the within-unit covariance matrices. As far as we can determine these are new results for MLMMs which show that some rank restrictions on the design matrices for both the fixed and random effects can lead to the information from just a single unit identifying both the 1-component and the mixture model. Interestingly this result only holds trivially for clusterwise regression in the sense that the result holds only if a single variable is included in the regression model. The difference is caused by the greater information contained within the units for MLMMs compared to MGLMs, which can be thought of a consisting of N units each with only a single response.

In this thesis we will specify $\boldsymbol{C}_i(\boldsymbol{\phi}_g)$ as an autoregressive correlation matrix of order r, which is equivalent to assuming the within-unit errors contained in $\boldsymbol{e}_i$ follow an autoregressive process of order r. To our knowledge this use of an autoregressive correlation matrix for the correlation structure of the within-unit errors in a MLMM has not before been used, although similar assumptions have been used before but in slightly different ways. For example in a Bayesian setting Fruehwirth-Schnatter and Kaufmann (2008) use a MLMM where each y_{ij}, $j = 1, ..., n_i$, follow an AR(1) process with random coefficients, and in a frequentist setting Bartolucci et al. (2011) uses a MLMM where $\boldsymbol{U}_i$ follows a AR(1) process. The only model we can find that permits serial dependence in the within-unit errors is by Coke and Tsao (2010) who assume the errors follow an antedependence model, which is a model for non-stationary correlation. Furthermore no regression components were used, and σ_g^2 was assumed constant across components.

Although all of these methods imply the responses follow an AR process, they do so in ways that are not equivalent. For example the method we use does not imply the random effects $\boldsymbol{U}_i$ are autocorrelated, and vice-versa. We also describe in subsection A.1 that the AR process must be stationary, which refers to a state of "statistical equilibrium", in order for the resulting covariance matrix $\boldsymbol{C}_i(\boldsymbol{\phi}_g)$ of $\boldsymbol{e}_i$ to be stationary. In turn this implies certain conditions must be met by the autoregressive parameters contained in $\boldsymbol{\phi}_g$. Since the covariance matrix of the responses, the random effects and the within-unit errors have different forms, then it is likely these conditions will be different depending which of these quantities are assumed to follow the autoregressive process. Thus we feel the use of an AR correlation matrix as specified in this thesis is different enough to add value to the existing literature.

In section 3.1, for finite mixture densities we summarise the relevant literature regarding statistical inference for the model parameters using maximum likelihood estimators (MLEs). We describe there how an important asymptotic result from general maximum likelihood theory (i.e. not necessarily for mixture models) for iid samples applies to mixture models. This result is in fact an amalgamation of results from multiple authors but is widely attributed solely to Redner and Walker (1984), and states that the MLE $\hat{\theta}$ is asymptotically distributed with mean θ, and variance given by the inverse of the information matrix. To our knowledge no such analogous result exists for non-iid samples, and hence for most MLMMs, however many researchers nonetheless still use it as if it has been proven to hold.

One area that needs explaining when trying to apply such iid results is that of identifiability of the model when regression parameters are used. In particular the conditions ensuring identifiability must be specified, and so too must the method of how these conditions are maintained as N tends to infinity. In this respect Hennig (2000) has provided some sufficient conditions for identifiability of clusterwise regression models, but currently no consistency proof for these models use these conditions, nor provides an adequate alternative. Thus the question of consistency is not fully closed even for clusterwise regression, and so it is not surprising the same is true for MLMMs.

For the reasons just described it is our opinion that such generalisations from the iid to the non-iid setting are made too readily, often without even mentioning this problem. Thus one of the primary objectives of this thesis was to investigate, through simulations, whether the iid theory works well for MLMMs. Accordingly in section 3.2 we describe three closely related methods of constructing confidence intervals around the MLMM parameter estimates in order to perform inference about the model parameters. These methods are concerned with approximating the mixture model information matrix, since direct calculation is not possible. These three methods extend the work of Boldea and Magnus (2009) who use these methods for finite mixture densities, and thus represent new methods of inference for the parameteres of MLMMs.

We describe a fourth method of confidence interval construction in subsection 3.4.3 which uses the LMM information matrix within each component, or "componentwise" to provide confidence intervals for the component distribution parameters $\boldsymbol{\theta}_j$, $j = 1, ..., G$, but not the mixing proportions. The idea for componentwise inference comes from the R package "Flexmix", and also from Grün (2008), and to our knowledge this is the first time such a method has been evaluated for its ability to perform inference in the MLMM. This componentwise method of inference relies upon the model component distribution functions being "well separated" - a concept we describe in detail in subsection 3.4.1, but that briefly relates essentially to component distribution functions being easily distinguished from each other. Traditionally this metric has been based on distances between the means of the densities, but in subsection 3.4.1 we propose a new method of measuring separation based upon how easily the model parameters can be distinguished from one another.

In this thesis we describe the componentwise inference method of Grün (2008), and the way in which we apply the mixture model confidence interval methods of Boldea and Magnus (2009) to the MLMM, as "naive" methods of inference. This is because in the former case we ignore the mixture model likelihood function, and in so doing ignore the opportunity to do something more sophisticated by using a separate probability density function for each component in the population. In the latter case, and as we have described, we are ignoring the fact that the iid theory we are applying has not been proven to hold for MLMMs.

We believe the main contributions of this thesis are;

- An in depth discussion of statistical inference subjects is provided which provides a much needed clarification on issues often superficially dealt with by researchers.
- Extensive simulation results in chapter 5 provide insight into how the proposed naive methods of inference might in general perform in a realistic applied setting by using sufficiently complex rather than overly simplistic models.
- The equations derived for the mixture model information matrix approximations in chapter C will prove useful reference equations for researchers to use in order to implement these new, albeit naive, methods of inference.
- The proposed method of quantifying component separation provides a useful alternative to the many other methods that have already been proposed since the method focuses on separation of parameters rather than the component density means.

Finally in chapter 6 we present two examples of the application of MLMMs to a dataset from an oncology clinical trial investigating treatment for lung cancer where the variables analysed are quality of life scores derived from a questionnaire. The results highlight both the usefulness and the difficulties associated with fitting mixture models.

2

MLMMs: Model description and estimation

2.1 Model description

Here we formally describe the hierarchical interpretation of MLMMs to which we referred in Chapter 1. In all that follows, and indeed throughout the entire thesis, unless otherwise stated g as a subscript will denote $g \in I_G$ - that is the object with the subscript belongs to, or is associated with, component g. Let $\boldsymbol{Y} = \{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$, $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$, $i \in I_N$, be a random sample of N units from a population that consists of G subpopulations or components. Conditional on unit i belonging to component $g \in I_G$, we assume $\boldsymbol{Y}_i$ follows the LMM given by

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_g + \boldsymbol{Z}_i \boldsymbol{u}_i + \boldsymbol{e}_i, \tag{2.1}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $n_i \times p$ and $n_i \times q$ fixed matrices respectively, $\boldsymbol{\beta}_g \in \mathbb{R}^p$ is a fixed vector, $\boldsymbol{u}_i$ is the realized value of a random vector $\boldsymbol{U}_i \in \mathbb{R}^q$, and $\boldsymbol{e}_i \in \mathbb{R}^{n_i}$ is a random vector. As per convention we shall call $\boldsymbol{e}_i$ the vector of errors or within-unit errors. We shall call $\boldsymbol{X}_i$ the matrix of covariate data, $\boldsymbol{\beta}_g$ the fixed effects or vector of fixed effects, $\boldsymbol{u}_i$ the random effects vector and $\boldsymbol{Z}_i$ the matrix of random effects covariate data. Note that the model in (2.1) is conditional on $\boldsymbol{U}_i = \boldsymbol{u}_i$.

Since the N units are a random sample from the G components, the probability of component membership for each component is the same for all units. Thus we define the vector of component probabilities or mixing proportions as $\boldsymbol{\pi} := [\pi_1, ..., \pi_G]^\intercal$,

$(\boldsymbol{\pi})_j \in [0,1]$, for all $j = 1, ..., G$, where $\sum_{j=1}^{G} (\boldsymbol{\pi})_j = 1$. Describing now the component membership in terms of random variables, let the vectors $\{\boldsymbol{\Lambda}_1, ..., \boldsymbol{\Lambda}_N\}$, $\boldsymbol{\Lambda}_i \in \mathbb{R}^G$ for all $i \in I_N$, be distributed as

$$\boldsymbol{\Lambda}_i \sim \text{mult}_G(1, \boldsymbol{\pi}), \tag{2.2}$$

so that $\{\boldsymbol{\Lambda}_1, ..., \boldsymbol{\Lambda}_N\}$ models an *iid* sample from a multinomial distribution. The range of values that $\boldsymbol{\Lambda}_i$ can take will be denoted by the set $\mathcal{A} = \left\{\boldsymbol{\lambda}^{(1)}, ..., \boldsymbol{\lambda}^{(G)}\right\}$, where $\boldsymbol{\lambda}^{(g)}$, $g \in I_G$, denotes a $G \times 1$ vector with a 1 in the g^{th} element and zeroes elsewhere. The notation $\boldsymbol{\lambda}_j^{(g)}$ means the j^{th} element of $\boldsymbol{\lambda}^{(g)}$, $j = 1, ..., G$.

In terms of our sample we observe a realization $\{\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_N\}$ of the random variables $\{\boldsymbol{\Lambda}_1, ..., \boldsymbol{\Lambda}_N\}$, where the notation $\boldsymbol{\Lambda}_{ij}$ and $\boldsymbol{\lambda_{ij}}$ means the j^{th} elements of $\boldsymbol{\Lambda}_i$ and $\boldsymbol{\lambda_i}$ respectively for $j \in I_J$, $i \in I_N$. For each $\boldsymbol{\lambda}^{(g)} \in \mathcal{A}$, the probability that $\boldsymbol{\Lambda}_i$ takes on this value, i.e., the probability that the i^{th} unit belongs to the g^{th} component, is $\boldsymbol{P}\left(\boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(g)} \mid \boldsymbol{\pi}\right) = \boldsymbol{\pi}_g = h\left(\boldsymbol{\lambda}^{(g)} \mid \boldsymbol{\pi}\right)$, where h is the probability mass function of $\boldsymbol{\Lambda}_i$. For brevity in conditional density functions we shall write $\boldsymbol{\lambda}_i^{(g)}$ to mean $\boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(g)}$ for some $g \in I_G$, and $\boldsymbol{\lambda}_{ij}^{(g)}$ to mean the j^{th} element of $\boldsymbol{\lambda}_i^{(g)}$, $j \in I_G$. Accordingly $\boldsymbol{P}(\boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(g)} \mid \boldsymbol{\pi}) = \boldsymbol{\pi}_g = h(\boldsymbol{\lambda}_i^{(g)} | \boldsymbol{\pi})$.

For the $q \times 1$ vector of random effects $\boldsymbol{U}_i$ we assume

$$\boldsymbol{U}_i | \boldsymbol{\lambda}_i^{(g)} \sim N_q\left(\boldsymbol{0}, \boldsymbol{D}_g\right), \tag{2.3}$$

where $\boldsymbol{D}_g$ is a $q \times q$ unstructured covariance matrix. We shall write the density function for $\boldsymbol{U}_i | \boldsymbol{\lambda}_i^{(g)}$ as $v_{ig}(\boldsymbol{u}_i | \boldsymbol{\lambda}_i^{(g)}, \boldsymbol{D}_g)$.

For the purposes of taking derivatives of the log likelihood function with respect to elements of $\boldsymbol{D}_g$, it is convenient to parameterise $\boldsymbol{D}_g$ by exploiting the fact that the off-diagonal elements are duplicated. In this respect let $\boldsymbol{\psi}_g = \text{v}(\boldsymbol{D}_g) \in \mathbb{R}^{(q(q+1))/2}$, where $\text{v}(\cdot)$ is the half-vec operator that stacks the columns of the lower triangular matrix of $\boldsymbol{D}_g$ one of top of the other. Thus $\boldsymbol{\psi}_g$ contains the supra-diagonal elements of $\boldsymbol{D}_g$, and $\boldsymbol{D}_g$ can be obtained by "unvectorising" the vector defined by $\text{vec}(\boldsymbol{D}_g) = D_q \text{v}(\boldsymbol{D}_g)$, where $\widetilde{\boldsymbol{D}}_q$ is the $q^2 \times (q(q+1)/2)$ duplication matrix which allows $\text{vec}(\boldsymbol{D}_g)$ to be expressed as a function of $\text{v}(\boldsymbol{D}_g)$. This parameterization is important for chapter C where we will be taking derivatives of the LMM and MLMM log-likelihood functions. This is because the $q^2 \times 1$ vector $\text{vec}(\boldsymbol{D}_g)$ contains duplicate elements on account of $\boldsymbol{D}_g$ being symmetric, and so Hessian matrices of functions of $\boldsymbol{D}_g$ where the derivatives have been taken with respect to vec$(\boldsymbol{D}_g)$ will be singular and hence not invertible. In contrast the elements of the vector $\boldsymbol{\psi}_g$ are unique and this problem does not arise. Although this is not a problem for first derivatives of scalar functions (since this produces a vector not a matrix which is to be inverted) and hence for estimation, for consistency we shall estimate $\boldsymbol{\psi}_g$ rather than $\boldsymbol{D}_g$. We will also restrict $\boldsymbol{\psi}_g$ to the subset of $\mathbb{R}^{(q(q+1))/2}$ which we will denote by $\Sigma_{\mathrm{v}(\boldsymbol{D})}$, where for all $\boldsymbol{\psi}_g \in \Sigma_{\mathrm{v}(\boldsymbol{D})}$, $\boldsymbol{\psi}_g$ gives rise to a positive-definite $\boldsymbol{D}_g$.

For the within-unit errors we assume

$$\boldsymbol{e}_i|\boldsymbol{\lambda}_i^{(g)} \sim N_{n_i}\left(\boldsymbol{0}, \sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right), \tag{2.4}$$

where $\sigma_g^2 \in \mathbb{R}^+$, $\boldsymbol{\phi}_g \in \Sigma_\phi \subseteq \mathbb{R}^r$, $\boldsymbol{C}_i(\boldsymbol{\phi}_g)$ is a $n_i \times n_i$ AR(r) correlation matrix that depends on i only through n_i, and Σ_ϕ is a subset of $\mathbb{R}^r$ such that for all $\boldsymbol{\phi}_g \in \Sigma_\phi$, $\boldsymbol{\phi}$ gives rise to a stationary AR process and thus a positive-definite correlation matrix $\boldsymbol{C}_i(\boldsymbol{\phi_g})$ - see subsection A.1 for a more detailed discussion of the importance of the AR process being stationary. In this appendix section we also introduce vectors $\boldsymbol{\tau}_j$, $j = 1, ..., G$, of quantities known as partial autocorrelations that are purported to give rise to a more stable estimation process compared to using the AR parameters themselves. It turns out this is helpful to us because for each $g \in I_G$ there exists a one to one transformation between $\boldsymbol{\phi}_g$ and $\boldsymbol{\tau}_g \in \Sigma_\tau \subseteq \mathbb{R}^r$, where $\Sigma_\tau := ([-1,1]^\intercal)^r = [-1,1]^\intercal \times \cdots \times [-1,1]^\intercal \subseteq \mathbb{R}^r \times \cdots \times \mathbb{R}^r$. The subset Σ_τ is equivalent to Σ_ϕ in the sense that for all $\boldsymbol{\tau}_g \in \Sigma_\tau$, $\boldsymbol{\tau}_g$ gives rise to a stationary AR(r) process and hence a positive-definite AR correlation matrix $\boldsymbol{C}_i(\boldsymbol{\phi}_g)$. Thus we shall use this $\boldsymbol{\tau}_g$ parameterization for estimation, but continue to use the AR parameters for all other purposes, switching between the two parameterizations using the aforementioned transformations.

Now equations (2.1) and (2.4) imply $\boldsymbol{Y}_i$ has the distribution

$$\boldsymbol{Y}_i|\boldsymbol{u}_i, \boldsymbol{\lambda}_i^{(g)} \sim N_{n_i}\left(\boldsymbol{X}_i\boldsymbol{\beta}_g + \boldsymbol{Z}_i\boldsymbol{u}_i, \sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right), \tag{2.5}$$

and so integrating over $\boldsymbol{U}_i$ in (2.5) we get that the distribution for $\boldsymbol{Y}_i$ conditional on $\boldsymbol{\lambda}_i^{(g)}$ is

$$\boldsymbol{Y}_i|\boldsymbol{\lambda}_i^{(g)} \sim N_{n_i}\left(\boldsymbol{X}_i\boldsymbol{\beta}_g, \boldsymbol{V}_i(\boldsymbol{\zeta}_g)\right), \tag{2.6}$$

where

$$\boldsymbol{V}_i(\boldsymbol{\zeta}_g) = \boldsymbol{Z}_i \boldsymbol{D}_g \boldsymbol{Z}_i^{\intercal} + \sigma_g^2 \boldsymbol{C}_i(\boldsymbol{\phi}_g), \tag{2.7}$$

and $\boldsymbol{\zeta}_g := (\boldsymbol{\psi}_g^{\intercal}, \sigma_g^2, \boldsymbol{\phi}_g^{\intercal})^{\intercal} \in \Sigma_\zeta \subseteq \mathbb{R}^{n_\zeta}$, $\Sigma_\zeta := \Sigma_{\mathrm{v}(\boldsymbol{D})} \times \mathbb{R}^+ \times \Sigma_\phi$, and $n_\zeta := q(q+1)/2+1+r$. We note that because of the definitions of $\Sigma_{\mathrm{v}(\boldsymbol{D})}$ and Σ_ϕ, $\boldsymbol{V}_i(\boldsymbol{\zeta}_g)$ will be positive-definite for all $\boldsymbol{\zeta}_g \in \Sigma_\zeta$. Letting $\boldsymbol{\theta}_g := \left[\boldsymbol{\beta}_g^{\intercal}, \boldsymbol{\psi}_g, \sigma_g^2, \boldsymbol{\phi}_g^{\intercal}\right]$, we shall write the density function for $\boldsymbol{Y}_i$ conditional on $\boldsymbol{\lambda}_i^{(g)}$ as $f_{ig}(\boldsymbol{y}_i | \, \boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g)$, where $\boldsymbol{\theta}_g \in \Psi := \mathbb{R}^p \times \Sigma_\zeta \subseteq \mathbb{R}^{n_\theta}$, and $n_\theta := p + n_\zeta$. Similarly we shall write the distribution function for $\boldsymbol{Y}_i$ conditional on $\boldsymbol{\lambda}_i^{(g)}$ as $F_{ig}(\boldsymbol{y}_i | \boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g)$.

The marginal density of $\boldsymbol{Y}_i$ is given by

$$\begin{aligned} f_i(\boldsymbol{y}_i | \boldsymbol{\theta}) &= \sum_{j=1}^{G} \boldsymbol{h}(\boldsymbol{\lambda}^{(j)} | \boldsymbol{\pi}) f_{ij}(\boldsymbol{y} | \boldsymbol{\lambda}^{(j)}, \boldsymbol{\theta}_j) \\ &= \sum_{j=1}^{G} \boldsymbol{\pi}_j f_{ij}(\boldsymbol{y}_i | \, \boldsymbol{\lambda}_i^{(j)}, \boldsymbol{\theta}_j), \end{aligned} \tag{2.8}$$

where $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_1^{\intercal}, ..., \boldsymbol{\theta}_G^{\intercal}, \boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_G\right]^{\intercal} \in \Theta$, and

$$\Theta = \left\{ (\boldsymbol{\theta}_1^{\intercal}, ..., \boldsymbol{\theta}_G^{\intercal}, \pi_1, ..., \pi_G)^{\intercal} : \sum_{j=1}^{G} \pi_j = 1, \pi_j \geq 0, \boldsymbol{\theta}_j \in \Psi, j = 1, ..., G \right\}. \tag{2.9}$$

So we see that the marginal density of $\boldsymbol{Y}_i$ is a sum or a mixture of G multivariate normal densities weighted by the elements of $\boldsymbol{\pi}$. We will write the marginal distribution function for $\boldsymbol{Y}_i$ as $F(\boldsymbol{Y}_i|\boldsymbol{\theta})$. Now assuming $\boldsymbol{U}_i \perp\!\!\!\perp \boldsymbol{e}_j$, $i \neq j$, $\boldsymbol{e}_i \perp\!\!\!\perp \boldsymbol{e}_j$, $i \neq j$, and $\boldsymbol{U}_i \perp\!\!\!\perp \boldsymbol{e}_j$, $\forall (i,j)$ implies conditional on the $\boldsymbol{u}_i$ and $\boldsymbol{\lambda}_i^{(g)}$ that the responses $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$ are all independent. Thus the joint distribution function $F(\boldsymbol{Y}|\boldsymbol{\theta})$ and density function $f(\boldsymbol{Y}|\boldsymbol{\theta})$ are both an independent product of the N distribution or density functions $F(\boldsymbol{Y}_i|\boldsymbol{\theta})$ or $f(\boldsymbol{Y}_i|\boldsymbol{\theta})$ respectively. Consequently if we let $\boldsymbol{y} = (\boldsymbol{y}_1^{\intercal}, ..., \boldsymbol{y}_N^{\intercal})^{\intercal}$ be the vector of realized values then the log-likelihood for the sample is

$$
\begin{aligned}
L(\boldsymbol{\theta}|\boldsymbol{y}) &= \log\left(\prod_{i=1}^{N} f_i(\boldsymbol{y}_i|\boldsymbol{\theta})\right) \\
&= \sum_{i=1}^{N} \log f_i(\boldsymbol{y}_i|\boldsymbol{\theta}) \\
&= \sum_{i=1}^{N} \log\left(\sum_{j=1}^{G} \boldsymbol{\pi}_j f_{ij}(\boldsymbol{y}_i|\, \boldsymbol{\lambda}_i^{(j)}, \boldsymbol{\theta}_j)\right). \qquad (2.10)
\end{aligned}
$$

Throughout this thesis we will also need to discuss a different type of mixture model than a MLMM, which for special cases is sometimes known as clusterwise regression. This model has N units, on which multivariate responses $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$ are obtained. Conditional on the i^{th} observation being in component $g \in I_G$, the i^{th} response is assumed to follow the Linear Model (LM)

$$
\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_g + \boldsymbol{\epsilon}_i, \qquad (2.11)
$$

where $\boldsymbol{X}_i$ is a $n_i \times p$ matrix of fixed covariate data, $\boldsymbol{\beta}_g \in \mathbb{R}^p$ is a vector of fixed effects, and $\boldsymbol{\epsilon}_i \sim \mathrm{N}(0, \sigma_g^2 \boldsymbol{I}_N)$. All the $\boldsymbol{\epsilon}_i$ are assumed independent, and so the $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$ are independent where $\boldsymbol{Y}_i \sim \mathrm{N}_N(\boldsymbol{X}_i\boldsymbol{\beta}_g, \sigma_g^2 \boldsymbol{I}_N)$ for all $i \in I_N$. Thus each $\boldsymbol{Y}_i$ has mixture density given by (2.8) but where $f_{ig}(\boldsymbol{y}|\boldsymbol{\lambda}^{(g)}, \boldsymbol{\theta}_g)$ is a normal density function with mean vector $\boldsymbol{X}_i\boldsymbol{\beta}_g$ and covariance matrix $\sigma_g^2 \boldsymbol{I}_N$. We will call the mixture of LMs described above as a Mixture of Linear Models (MLMs). When $n_i = 1$ for all $i \in I_N$ then a MLM is known as a clusterwise regression model.

2.2 Estimating MLMMs using the EM algorithm

In terms of obtaining maximum likelihood estimates for MLMMs, the complex dependence of the mixture model likelihood function on the parameters means that in general the likelihood equations cannot be solved analytically, and so numerical methods must be used to derive approximate solutions to the likelihood equations. One such procedure which has proven to be very popular in this respect is the EM algorithm developed by Dempster et al. (1977). The reason for the popularity of this algorithm may be due to the fact that often the derivatives of the so called "complete" data likelihood used by the algorithm are much easier to derive than the derivatives of the original or "ordinary" likelihood. For example for MLMMs the ordinary log likelihood function $L_N(\boldsymbol{\theta})$

is a logarithm of a sum, and so the score vector will contain ratios. This makes the derivation of the Hessian matrix very tedious indeed, and so this is often reason enough for researchers to avoid estimation methods such as the Newton-Raphson method. In contrast, and as we will show, the complete data log likelihood function is the sum of logarithms, which lends itself more readily to being manipulated algebraically. The EM algorithm also enjoys certain desirable properties, namely that the ordinary log likelihood function increases on every iteration, and that convergence to some local optimum is guaranteed. One of the major drawbacks of the algorithm is that it can be very slow to converge, and that it does not guarantee a global optima is found.

For any given statistical model the EM algorithm proceeds by choosing some of the "data" (which often includes some of the parameters) as missing or unobservable, whilst the rest of the data is observed. The missing data is observed indirectly through the observable data. The combination of observable and missing data is called the complete data, whilst the observable data is called the incomplete data. This choice of what is missing or not is to some extent arbitrary, and can therefore lead to multiple versions or variants of the EM algorithm for the same model.

For MLLMs, and for the i^{th} unit, the two sensible choices are to either consider both the random vectors $\boldsymbol{U}_i$ and $\boldsymbol{\Lambda}_i$ as being missing, or to only consider $\boldsymbol{\Lambda}_i$ as missing. We shall call these two variants the first and second variants respectively. In Subsections 2.2.1 and 2.2.2 we outline the estimating equations of these EM algorithms that we will use to obtain maximum likelihood estimates for the parameters in MLMMs. We describe the method used to obtain starting values for these EM algorithm variants in Sub-section 5.1.2.

2.2.1 EM algorithm: first variant

Here we very briefly outline the necessary steps to take in order to implement an ECM algorithm, which is a particular version of the first variant of the EM algorithm. Let the vector $\boldsymbol{C}_i = (\boldsymbol{Y}_i^\intercal, \boldsymbol{U}_i^\intercal, \boldsymbol{\Lambda}_i^\intercal)^\intercal$ be the "complete" data vector, where we now think of the random vectors $\boldsymbol{U}_i$ and $\boldsymbol{\Lambda}_i$ as being observable. Accordingly we shall write $\boldsymbol{c}_i = (\boldsymbol{y}_i^\intercal, \boldsymbol{u}_i^\intercal, \boldsymbol{\lambda}_i^{(I_i)\intercal})^\intercal$ as the "observed" complete data vector, and so the actual observed vector $\boldsymbol{y}_i$ can be thought of as the "incomplete" data vector. We also assume $\{\boldsymbol{C}_1, ..., \boldsymbol{C}_N\}$ are independent random variables. There may be some confusion of $\boldsymbol{C}_i$ with $\boldsymbol{C}_i(\boldsymbol{\phi}_g)$, however we continue with this notation since $\boldsymbol{C}_i$ in this context (i.e. as the complete data vector) will only appear in this section of the thesis. Letting w_{ig} be the density for $(\boldsymbol{Y}_i, \boldsymbol{U}_i)|\boldsymbol{\Lambda}_i = \boldsymbol{\lambda}_i^{(g)}$, then the complete data log-likelihood from (A.18) is given by

$$\begin{aligned}\text{Ł}^c(\boldsymbol{\theta}|\boldsymbol{c}) &= \sum_{i=1}^{N}\sum_{j=1}^{G} \boldsymbol{\lambda}_{ij}^{(I_i)} \log\Big(w_{ij}(\,\boldsymbol{y}_i, \boldsymbol{u}_i|\,\boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(j)}, \boldsymbol{\theta}_j)\Big) + \sum_{i=1}^{N}\sum_{j'=1}^{G} \boldsymbol{\lambda}_{ij'}^{(I_i)} \log\left(\boldsymbol{\pi}_{j'}\right) \\ &= \sum_{i=1}^{N} \left(\boldsymbol{\lambda}_i^{(I_i)}\right)^{\intercal} T_i(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{u}_i) + \sum_{i=1}^{N} \boldsymbol{\lambda}_i^{(I_i)\intercal} \boldsymbol{U}(\boldsymbol{\pi}), \end{aligned} \tag{2.12}$$

where $T_i(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{u}_i) = \left\{_c \ \log\left\{w_{ij}(\,\boldsymbol{y}_i, \boldsymbol{u}_i|\,\boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(j)}, \boldsymbol{\theta}_j)\right\}\right\}_{j=1}^{G}$, and $\boldsymbol{U}(\boldsymbol{\pi}) = \left\{_c \ \log\left(\boldsymbol{\pi}_j\right)\right\}_{j=1}^{G}$.

The EM algorithm maximises the ordinary log-likelihood $L(\boldsymbol{\theta}|\boldsymbol{y})$ by working with $\boldsymbol{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{'})$, which is the expected value of $L^c(\boldsymbol{\theta}|\boldsymbol{C})$ conditional on $\boldsymbol{y}$ and $\boldsymbol{\theta}^{'}$. If we let s denote the current iteration of the EM algorithm, and $\hat{\boldsymbol{\theta}}^{(s)}$ the estimate obtained, then the E-step consists of calculating $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$ which from appendix A.3 is given by

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) = Q_1(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) + Q_2(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}), \tag{2.13}$$

where

$$Q_1(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) = \sum_{i=1}^{N}\sum_{k=1}^{G} \hat{p}_i(\boldsymbol{\lambda}_i^{(k)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}) Q_{1ik}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}), \tag{2.14}$$

where

$$\begin{aligned}Q_{1ig}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) = &-\left(\frac{n_i}{2}\right)\log(2\pi) - \left(\frac{n_i}{2}\right)\log(\sigma_g^2) - \frac{1}{2}\log\left(|\boldsymbol{D}_g|\right) \\ &- \frac{1}{2}\log\left(|\boldsymbol{C}_i(\boldsymbol{\phi}_g)|\right) - \frac{1}{2\sigma_g^2}\text{tr}\left(\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{E}}_i^{(s)}\right) - \frac{1}{2}\text{tr}\left(\boldsymbol{D}_g^{-1}\hat{\boldsymbol{J}}_i^{(s)}\right),\end{aligned} \tag{2.15}$$

and

$$Q_2(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) = \sum_{i=1}^{N}\sum_{k=1}^{G} \hat{p}_i(\boldsymbol{\lambda}_i^{(k)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}) \log(\boldsymbol{\pi}_k), \tag{2.16}$$

where $\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)})$ is the posterior probability of the i^{th} unit belonging to the g^{th} component, $g \in I_G$, conditional on the observed response vector for that unit, and the current estimate of $\boldsymbol{\theta}$, and is given by

$$\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}) = \frac{f_{ig}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(g)}, \hat{\boldsymbol{\theta}}_g^{(s)})\hat{\boldsymbol{\pi}_g}^{(s)}}{\sum_{l=1}^{G} f_{il}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(l)}, \hat{\boldsymbol{\theta}}_l^{(s)})\hat{\boldsymbol{\pi}_l}^{(s)}}. \tag{2.17}$$

For brevity we will often denote these posterior probabilities as $\hat{p}_{ig}$. Furthermore the equations below are needed to implement equation (2.15);

$$\hat{\boldsymbol{E}}_i^{(s)} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}_i}^{(s)} + \hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)}\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)\top}, \tag{2.18}$$

$$\hat{\boldsymbol{J}}_i^{(s)} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)} + \hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)}\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)\top}, \tag{2.19}$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)} = \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)})\boldsymbol{Z}_i^{\top}\boldsymbol{V}_i(\hat{\boldsymbol{\zeta}}_g^{(s)})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_g^{(s)}), \tag{2.20}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)} = \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)}) - \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)})\boldsymbol{Z}_i^{\top}\boldsymbol{V}_i(\hat{\boldsymbol{\zeta}}_g^{(s)})^{-1}\boldsymbol{Z}_i\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)}), \tag{2.21}$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)} = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g - \boldsymbol{Z}_i\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)}, \tag{2.22}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}_i}^{(s)} = \boldsymbol{Z}_i\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)}\boldsymbol{Z}_i^{\top}. \tag{2.23}$$

We notice in (2.13) that $\boldsymbol{\pi}$ is contained only in $Q_2(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$ and that all the remaining parameters are contained in $Q_1(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$. Accordingly we can perform two separate maximisations, one for the component probabilities and one for the parameters of the component densities. Furthermore since the parameters for the component densities do not depend on each other, we can maximise separately for each component. Note however from (2.17) that all component density parameters contribute to estimating the posterior probabilities, which in turn influence each componentwise maximisation. Thus the component density parameters are not independent of each other.

The M-step of the algorithm requires $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$ to be maximised with respect to $\boldsymbol{\theta}$ to obtain $\hat{\boldsymbol{\theta}}^{(s+1)}$. Dempster et al. (1977) point out that the EM algorithm has been criticised for being called an algorithm since it does not describe actually how to implement the E and M steps. For this reason the complexity and feasibility of these steps can vary widely depending on the application.

If the M-step is sufficiently complex it is sometimes desirable to break this step down by performing separate maximisations with respect to each component of $\boldsymbol{\theta}$ whilst fixing the other components at their current values, that is each maximisation proceeds conditional on the values of the other components being available. This is known as an expected conditional maximisation algorithm (ECM) and was developed by Meng and Rubin (1993) who show an ECM algorithm is a special class of Generalized EM algorithm (GEM) (algorithms that increase $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$ rather than maximise it on each M step) that have the same convergence properties of an EM algorithm.

We now describe the steps of the ECM algorithm for the MLMMs described here. The derivations of the derivatives for these equations can be found in chapter C. Let s denote the current iteration of the EM algorithm, and suppose that that $\hat{\boldsymbol{\theta}}^{(s)}$ is available. Then the ECM algorithm proceeds as follows;

1. For each $i \in I_N$, and $g \in I_G$, calculate the posterior probabilities using

$$\hat{p}_{ig}^{(s+1)} := \hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}) = \frac{f_{ig}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(g)}, \hat{\boldsymbol{\theta}}_g^{(s)})\hat{\boldsymbol{\pi}_g}^{(s)}}{\sum_{k=1}^{G} f_{ik}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(k)}, \hat{\boldsymbol{\theta}}_k^{(s)})\hat{\boldsymbol{\pi}_k}^{(s)}}. \tag{2.24}$$

2. For each $g \in I_G$, update $\hat{\boldsymbol{\pi}_g}^{(s)}$

$$\hat{\boldsymbol{\pi}_g}^{(s+1)} := \frac{1}{N}\sum_{i=1}^{N} \hat{p}_{ig}^{(s+1)}, \tag{2.25}$$

3. For each $g \in I_G$, update $\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)}) = \hat{\boldsymbol{D}}_g^{(s)}$

$$\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) := \hat{\boldsymbol{D}}_g^{(s+1)} = \frac{1}{\sum_{i=1}^{N} \hat{p}_{ig}^{(s+1)}} \sum_{i=1}^{N} \hat{p}_{ig}^{(s+1)} \hat{\boldsymbol{J}}_i^{(s)}, \tag{2.26}$$

where $\hat{\boldsymbol{J}}_i^{(s)}$, $\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)}$ are given in (2.18), (2.20) and (2.21) respectively.

4. For each $g \in I_G$, update $\hat{\boldsymbol{\sigma}}_g^{2(s)}$

$$\hat{\boldsymbol{\sigma}}_g^{2(s+1)} := \frac{1}{\displaystyle\sum_{i=1}^{N} n_i \hat{p}_{ig}^{(s+1)}} \sum_{i=1}^{N} \hat{p}_{ig}^{(s+1)} \mathrm{tr}\left[\boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)})^{-1} \hat{\boldsymbol{E}}_i^{(s)} \left(\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \right) \right], \tag{2.27}$$

where $\hat{\boldsymbol{E}}_i^{(s)} \left(\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \right)$ is equal to $\hat{\boldsymbol{E}}_i^{(s)}$ given in (2.18) and where

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)} = \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \boldsymbol{Z}_i^{\intercal} \left[\boldsymbol{Z}_i \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \boldsymbol{Z}_i^{\intercal} + \hat{\boldsymbol{\sigma}}_g^{2(s)} \boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)}) \right]^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_g^{(s)}), \tag{2.28}$$

and

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)} = \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) - \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \boldsymbol{Z}_i^{\intercal} \left[\boldsymbol{Z}_i \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \boldsymbol{Z}_i^{\intercal} + \hat{\boldsymbol{\sigma}}_g^{2(s)} \boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)}) \right]^{-1} \boldsymbol{Z}_i \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}). \tag{2.29}$$

5. For each $g \in I_G$, update $\hat{\boldsymbol{\beta}}_g^{(s)}$

$$\hat{\boldsymbol{\beta}}_g^{(s+1)} := \left(\sum_{i=1}^{N} \hat{p}_{ig}^{(s+1)} \boldsymbol{X}_i^{\intercal} \boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)})^{-1} \boldsymbol{X}_i \right)^{-1} \left[\sum_{i=1}^{N} \hat{p}_{ig}^{(s+1)} \boldsymbol{X}_i^{\intercal} \boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)})^{-1} \left(\boldsymbol{y}_i - \boldsymbol{Z}_i \hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)} \right) \right], \tag{2.30}$$

where

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)} = \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \boldsymbol{Z}_i^{\intercal} \left[\boldsymbol{Z}_i \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \boldsymbol{Z}_i^{\intercal} + \hat{\boldsymbol{\sigma}}_g^{2(s+1)} \boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)}) \right]^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_g^{(s)}). \tag{2.31}$$

6. For each $g \in I_G$, update $\hat{\boldsymbol{\tau}}_g^{(s)}$

$$\hat{\boldsymbol{\tau}}_g^{(s+1)} := \underset{\boldsymbol{\tau}_g \in [-1,1]^r}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \hat{p}_{ig}^{(s+1)} \left[-\log |\boldsymbol{C}_i(\boldsymbol{\tau}_g)| - \frac{1}{\hat{\sigma}_g^{2(s+1)}} \left(\boldsymbol{C}_i(\boldsymbol{\tau}_g)^{-1} \hat{\boldsymbol{E}}_i^{(s)} \left(\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \right) \right) \right] \right\}, \tag{2.32}$$

where $\hat{\boldsymbol{E}}_i^{(s)} \left(\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \right)$ is equal to $\hat{\boldsymbol{E}}_i^{(s)}$ given in (2.18) but where

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)} = \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \boldsymbol{Z}_i^{\intercal} \left[\boldsymbol{Z}_i \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \boldsymbol{Z}_i^{\intercal} + \hat{\boldsymbol{\sigma}}_g^{2(s+1)} \boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)}) \right]^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_g^{(s+1)}), \tag{2.33}$$

and

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)} = \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) - \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \boldsymbol{Z}_i^{\intercal} \left[\boldsymbol{Z}_i \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}) \boldsymbol{Z}_i^{\intercal} + \hat{\boldsymbol{\sigma}}_g^{2(s+1)} \boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)}) \right]^{-1} \boldsymbol{Z}_i \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s+1)}). \tag{2.34}$$

The maximisation in (2.32) is performed numerically, and $\hat{\boldsymbol{\tau}}_g^{(s)}$ is converted to $\hat{\boldsymbol{\phi}}_g^{(s)}$ using (A.10) in order to evaluate the equations 2.33 and 2.34.

2.2.2 EM algorithm: second variant

This variant of the EM algorithm is where we assume only the component memberships for each unit are unknown (i.e. the multinomial random vectors $\mathbf{\Lambda}_i$) rather than the component memberships and the random effects $\boldsymbol{U}_i$. Specifically if we let $\boldsymbol{C}_i = \left[\boldsymbol{Y}_i^{\intercal}, \mathbf{\Lambda}_i^{\intercal}\right]^{\intercal}$ be the complete data vector, and s denote the s^{th} EM algorithm iteration, then using the same methods shown in appendix A.3 we have

$$\begin{aligned}
Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) &= \boldsymbol{E}\left[L^c(\boldsymbol{\theta}|\boldsymbol{C})|\,\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right] \\
&= \sum_{i=1}^{N} \boldsymbol{E}\left[\mathbf{\Lambda}_{i,I_i}\log(f_{i,I_i}(\boldsymbol{y}_i, |\mathbf{\Lambda}_i, \boldsymbol{\theta}_{I_i}))|\,\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right] + \sum_{i=1}^{N} \boldsymbol{E}\left[\mathbf{\Lambda}_{i,I_i}|\,\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right]\log(\boldsymbol{\pi}_{I_i}) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{G} \hat{p}_i(\boldsymbol{\lambda}_i^{(k)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)})\log(f_{ik}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(k)}, \boldsymbol{\theta}_k)) + \sum_{i=1}^{N}\sum_{k=1}^{G} \hat{p}_i(\boldsymbol{\lambda}_i^{(k)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)})\log(\boldsymbol{\pi}_k).
\end{aligned} \tag{2.35}$$

Since the $\boldsymbol{\theta}_g$ for $g = 1, ..., G$, do not depend on each other, given that $\hat{\boldsymbol{\theta}}^{(s)}$ and $\hat{\boldsymbol{\pi}}^{(s)}$ are available, then the EM algorithm proceeds as follows;

1. For each $i \in I_N$, and $g \in I_G$, calculate the posterior probabilities using
$$\hat{p}_{ig}^{(s+1)} := \frac{f_{ig}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(g)}, \hat{\boldsymbol{\theta}}_g^{(s)})\hat{\boldsymbol{\pi}_g}^{(s)}}{\sum_{k=1}^{G} f_{ik}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(k)}, \hat{\boldsymbol{\theta}}_k^{(s)})\hat{\boldsymbol{\pi}_k}^{(s)}}. \tag{2.36}$$

2. For $g \in I_G$, compute the component density parameter vector estimates using
$$\hat{\boldsymbol{\theta}_g}^{(s+1)} := \max_{\boldsymbol{\theta}_g \in \Psi}\left\{\sum_{i=1}^{N} \hat{p}_{ig}^{(s+1)}\log(f_{ig}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g))\right\}. \tag{2.37}$$

3. For $g \in I_G$, compute the component probabilities
$$\hat{\boldsymbol{\pi}_g}^{(s+1)} := \frac{1}{N}\sum_{i=1}^{N} \hat{p}_{ig}^{(s+1)}. \tag{2.38}$$

We shall call the G separate maximisations in step 2 for the parameter vector of the component densities as componentwise maximisations, or estimating the parameters of the component densities componentwise. Note that although the componentwise maximisations in step 2 are performed separately, the component density parameters obtained depend on all the parameters of the mixture model because each maximization depends on the posterior probabilities.

3

Inference on the model parameters

In this chapter we summarise the statistical theory that is relevant to performing statistical inference about the parameters in a MLMM using maximum likelihood estimation. However on account of a lack of research into inference in non-*idd* samples that have mixture distributions, the "relevant theory" we describe is the theory for *iid* Gaussian mixtures - that is samples that are associated with mixture distributions generally (but not necessarily) not induced by a regression model. The description of this *iid* theory can be found in Section 3.1. Section 3.2 describes the problems encountered when trying to use the information matrix that the *iid* theory described in Section 3.1 tells us is the asymptotic covariance matrix of the mixture model parameters. Furthermore Section 3.2 also describes a method for quantifying the performance of confidence intervals in simulations that uses the concept of true parameter standard errors. Since the underlying model in a MLMM is a LMM then it is of interest to consider the asymptotic theory of the LMM, for example to determine if the methods used for the LMM can be extended to the MLMM. For this reason in Section 3.3 we present a brief summary of the asymptotic theory for the LMM.

Section refsec:infomatMLMM uses the previous sections to justify two methods for performing statistical inference about the parameters in a MLMM, and results of extensive simulations are reported in Chapter 5 that investigate the performance of these methods.

3.1 Inference for finite mixture densities for an *iid* sample

This section is concerned with statistical inference about the maximum likelihood estimator (MLE) of the true parameter $\boldsymbol{\theta}_0$ of the mixture density $f(\cdot|\boldsymbol{\theta}_0)$ from which we have a random sample of N identically distributed response vectors. Thus we assume an *iid* sample $\boldsymbol{Y}_i \in \mathbb{R}^n$, $i = 1, ..., N$, where each $\boldsymbol{Y}_i$ has the finite mixture density given in (1.1) but with $\boldsymbol{\theta}_0$ replacing $\boldsymbol{\theta}$. For the most part we will concentrate on finite mixture densities, but we will also touch upon mixture densities associated with clusterwise regression models. Although the samples associated with these densities are generally not identically distributed, some authors have for convenience imposed conditions on the covariate vectors in the component specific regression models to nonetheless ensure that the sample come from the same distribution.

As in Chapter 2 we shall write $\boldsymbol{Y} = (\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N)^\intercal$ for the vector of all responses, and $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_N)^\intercal$ for the vector of all realized values. Since we will be discussing asymptotic results, to show the dependence on the number of units N we shall write $L_N(\boldsymbol{Y}|\boldsymbol{\theta})$ for the log-likelihood function of the sample which is given in (2.10), and often for brevity we will shorten this to $L_N(\boldsymbol{\theta})$. Similarly we shall write $\hat{\theta}_N(\boldsymbol{Y})$ or $\hat{\theta}_N$ for the MLE of $\boldsymbol{\theta}$. For reasons we will discuss shortly, the MLE $\hat{\theta}_N$ will be defined as any $\boldsymbol{\theta} \in \Theta^\circ$ that satisfies the following equations

$$\sum_{i=1}^{N} \frac{\partial \log f(\boldsymbol{y}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r} = 0 \qquad\qquad r = 1, ..., k, \tag{3.1}$$

which are called the likelihood equations.

There are two problems associated with $L_N(\boldsymbol{\theta})$ which make the estimation of $\boldsymbol{\theta}$ different from a standard maximum likelihood problem. One problem is that $L_N(\boldsymbol{\theta})$ may have the potential to take on infinite values in the parameter space Θ, and so $L_N(\boldsymbol{\theta})$ may have no global maximum. In this respect when the component densities are all normal $L_N(\boldsymbol{\theta})$ always has this potential. For example let $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_G$, and $\boldsymbol{V}_1, ..., \boldsymbol{V}_G$, be the G mean vectors and covariance matrices of these component densities. A well known pathological example is where we assume $\boldsymbol{y}_{i'} = \boldsymbol{\mu}_1$ for one of our observed sample values $\boldsymbol{y}_{i'}$. By then letting the determinant of the covariance matrix $\boldsymbol{V}_1$ tend to zero, the log-likelihood function $L_N(\boldsymbol{\theta})$ tends to infinity. Thus each sample point in the sample has the potential to cause the log-likelihood function to be unbounded (Day

(1969)). A special case of this example occurs for univariate normal densities where the covariance matrices are scalar variances $\sigma_1^2, ..., \sigma_G^2$, and where we let σ_1^2 tend to zero.

In practice these pathological examples could occur if $\hat{\boldsymbol{\sigma}}_1^2$ or $\hat{\boldsymbol{V}}_1$ are estimated to be near zero. In the multivariate setting this will often produce a covariance matrix that is singular to working precision. Such a situation could arise if a component becomes degraded in the sense of fewer and fewer units being assigned to that component. For example if only one unit is assigned to component 1 then $\hat{\pi}_1$ will be estimated to be close to zero. The second problem associated with $L_N(\boldsymbol{\theta})$ is that due to label switching for any given local maximum there will be $G! - 1$ other local maxima that are exactly the same. Thus in combination these two problems mean $L_N(\boldsymbol{\theta})$ does not have a unique global maximum, indeed at best $L_N(\boldsymbol{\theta})$ will have a largest local maximum which is replicated at $G!$ different values of $\boldsymbol{\theta}$.

One way to prevent $L_N(\boldsymbol{\theta})$ from being unbounded is to work with a constrained parameter space, limiting in some way the magnitude of the variances or the magnitude of the determinants of the covariance matrices. For univariate component densities this approach has been taken by Hathaway (1985) who imposes restrictions on the ratios of the variances, and Tanaka and Takemura (2006) who use restrictions on the variances themselves. An alternative approach is to work with the unconstrained parameter space so that $L_N(\boldsymbol{\theta})$ is still unbounded, but to search for a local rather than a global maximser as an estimator. This latter method appears to be much more popular in the literature, and for this reason in this section we will summarise the main results that use this approach. Accordingly in this section, and unless otherwise stated, by a maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ we will mean an estimator that locally maximises $L_N(\boldsymbol{\theta})$.

It seems that the most widely cited results that concern the consistency and asymptotic distribution of $\hat{\boldsymbol{\theta}}$ are those contained in Theorems 3.1 and 3.2 in Redner and Walker (1984) which are concerned with finite mixture densities for multivariate responses $\boldsymbol{Y}_i \in \mathbb{R}^n$. The consistency proof upon which Redner and Walker rely is that of Peters and Walker (1978), which is concerned with a non-mixture maximum likelihood problem where local maxima of the log-likelihood function are admitted as solutions to the likelihood equations. Thus the estimation problem is applicable to finite mixture densities.

Another consistency proof can be found in Kiefer (1978) which is concerned with model generated finite mixture densities for univariate responses $Y_i \in \mathbb{R}^1$. Specifically Kiefer (1978) studies a bivariate normal mixture density arising from a switching regression model with two regimes, which is a specific type of MGLM. In general such model-generated densities are non-*iid*, however by assuming the covariate data are *iid* random variables, Kiefer (1978) ensures the sample are also *iid*. The proof uses Chanda (1954) who, like Peters and Walker (1978), also considers a non-mixture maximum likelihood problem where local maxima are admitted as solutions to the likelihood equations. The proof of Kiefer (1978) consists of verifying that the hypotheses of Chanda hold, and furthermore it is stated the proof will work for any number of regimes.

The models considered by Kiefer (1978) for the regimes include regression parameters, and covariate data that vary with the units. The model assumptions made by Kiefer lead to the densities for each y_i being a mixture of two univariate normal densities, but where these densities are different for each unit. Since the proof of Chanda requires an iid sample, Kiefer imposes the condition on the covariate vectors that they are bounded iid random variables. This obviously means the sample $\{y_1, ..., y_N\}$, is an iid sample with some distribution, but it is not clear that the Chanda proof would hold with the additional densities required for the covariate vectors. Unfortunately no mention of this issue is made, and Kiefer claims all that is required for the Chanda proof to hold for this model is to verify that the sufficient conditions of the Chanda proof hold. Another problem with the proof of Kiefer is that identifiability problems arising from the regression parameters are not discussed. Identifiability means for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ that $f(\boldsymbol{y}_i|\boldsymbol{\theta}) = f(\boldsymbol{y}_i|\boldsymbol{\theta}')$ for all $i \in I_N$ implies and is implied by $\boldsymbol{\theta} = \boldsymbol{\theta}'$. In this respect Hennig (2000) has shown that in general identifiability does not hold unless conditions are imposed on the covariate data. For these reasons the proof by Kiefer is not a valid one, although we will discuss the application of Chanda's proof to mixture densities without regression parameters.

The third consistency proof we will consider is that of Sundberg (1974) who considers *iid* samples with non-model generated finite mixture densities for multivariate responses $\boldsymbol{Y}_i \in \mathbb{R}^n$ (i.e it is assumed $\boldsymbol{Y}_i$ has a mixture density rather than this density being induced by a regression model). The focus of this paper is not finite mixture densities but rather responses whose distributions are from the exponential family, but where the distributions are generated by loss of information - i.e. distributions obtained by integrating out missing data from another distribution. If the component densities are from the exponential family, then Sundberg shows that the distribution function with which the mixture density is associated is in this class of distribution, where the unknown component memberships represent the loss of information. Much of the consistency proof in Sundberg (1974) uses Aitchison and Silvey (1958) whose work is concerned with non-mixture maximum likelihood problem where local maxima are admitted as solutions to the likelihood equations, and where constraints can be imposed on $\boldsymbol{\theta}$.

For any statistical model lack of identifiability precludes the existence of a consistent estimator. Thus for finite mixture densities any consistency proof must deal with the fact that identifiability does not hold due to label switching. However we can assume for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, and regardless of how close $\boldsymbol{\theta}$ is to $\boldsymbol{\theta}'$, as long as $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ that we can always find a neighbourhood around $\boldsymbol{\theta}$ in which identifiability does hold. Sundberg refers to this as local identifiability, and it is tacitly understood that his consistency proof is concerned with proving a consistent estimator exists in a neighborhood, however small, of $\boldsymbol{\theta}_0$ such that local identifiability holds (the same is true in the proof of Peters and Walker (1978)). Taking an alternate approach to overcoming the label switching problem, Kiefer (1978) states that a rule must be imposed on the parameter space Θ. One such rule which is typically used is to impose an arbitrary ordering on the mixture probabilities, for example to assume for all $\boldsymbol{\theta} \in \Theta$ that $\pi_1 > \pi_2 > ... > \pi_G$. It is worth noting that these approaches to overcome the label switching problem are only relevant theoretically in order to permit the existence of a consistent estimator of $\boldsymbol{\theta}_0$. That is to say the MLE $\hat{\boldsymbol{\theta}}$ still represents one of $G!$ possible ways in which we could re-label an estimate of $\boldsymbol{\theta}_0$. Of course this is no problem because in practice the components are unknown (in terms of what they represent) and so any labeling must necessarily be arbitrary.

For non-mixture model problems, and for a scalar parameter θ, an important consistency proof for estimators that locally maximise $L_N(\theta)$ is that of Cramér (1946, pp500). This is not only an important proof in the theory of maximum likelihood estimation in general, but also for maximum likelihood estimation in mixture models. This is because the consistency proofs in Chanda (1954), Aitchison and Silvey (1958), and Peters and Walker (1978) in which we are interested, were all aimed at generalizing Cramér's proof to the multi-parameter setting. Before we state and describe the assumptions that these three proofs use, we shall need to introduce some notation for this general (i.e. not necessarily non-mixture) maximum likelihood estimation problem. There is a re-use of some of the notation we have used before for mixture densities, however in what follows the methods we discuss are applicable to general density functions of which mixture densities are a subset. We will make it clear when we are talking about mixture densities when modifications to the methods need to be made, or where different interpretations of the results need to be used.

Let $\mathscr{P} = \{\mu_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, be a family of distributions parameterized by points in a parameter space $\Theta \subseteq \mathbb{R}^k$. For each $\boldsymbol{\theta} \in \Theta$, $\mu_{\boldsymbol{\theta}}$ is a probability measure on the measurable space $(\mathbb{R}^n, \mathscr{R}^n)$, where $\mathscr{R}^n$ are the Borel subsets of $\mathbb{R}^n$. We assume for each $\boldsymbol{\theta} \in \Theta$ that $\mu_{\boldsymbol{\theta}}$ has a density function $f_{\boldsymbol{\theta}}$ with respect to n-dimensional Lebesgue measure λ_n, where $f_{\boldsymbol{\theta}}$ is shortened notation for $f(\boldsymbol{y}|\boldsymbol{\theta})$, $\boldsymbol{y} \in \mathbb{R}^n$. We will denote by $\mathscr{P}_f := \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ the family of density functions associated with $\mathscr{P}$. Letting Θ° denote the interior of Θ, and for $\boldsymbol{\theta}_0 \in \Theta^\circ$, let $\boldsymbol{Y}_i \in \mathbb{R}^n$, $i \in I_N$, be N independent observations on a random variable $\tilde{\boldsymbol{Y}}$ with distribution $\mu_{\boldsymbol{\theta}_0} \in \mathscr{P}$. We will use $(\mathscr{P}_0, \Omega)$ to denote the underlying probability space on which the random variables $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$ are defined. We will also denote by $\boldsymbol{Y}{=}(\boldsymbol{Y}_1^\intercal, ..., \boldsymbol{Y}_N^\intercal)^\intercal \in R^{Nn}$ the joint vector of all the random variables, and similarly $\boldsymbol{y} = (\boldsymbol{y}_1^\intercal, ..., \boldsymbol{y}_N^\intercal)^\intercal \in R^{Nn}$ the joint vector of realized values.

We assume the parameter space Θ is identifiable, so that $\mu_{\boldsymbol{\theta}} = \mu_{\boldsymbol{\theta}'}$ implies and is implied by $\boldsymbol{\theta} = \boldsymbol{\theta}'$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$. Bearing in mind that we want to apply these non-mixture asymptotic results to mixture densities, we note that everything in the discussion that follows holds for mixture densities if in assumptions (i)-(iii) below we further assume $N_{\delta_1}(\boldsymbol{\theta}_0) \subseteq \Theta$ is a small enough neighborhood of $\boldsymbol{\theta}_0$ such that local identifiability holds within the neighbourhood. In the conditions that follow, and for brevity, for a general point $\boldsymbol{y} \in \mathbb{R}^n$ we will write f instead of $f(\boldsymbol{y}|\boldsymbol{\theta})$, and for any realized value $\boldsymbol{y}_i$ we will write f_i instead of $f(\boldsymbol{y}_i|\boldsymbol{\theta})$. Finally the phrase "for almost all" shall be meant with respect to n-dimensional Lebesgue measure. We can now state the assumptions used in the consistency proofs:

(i) There exists a neighborhood $N_{\delta_1}(\boldsymbol{\theta}_0) \subseteq \Theta$ such that for all $\boldsymbol{\theta} \in N_{\delta_1}(\boldsymbol{\theta}_0)$, for almost all $\boldsymbol{y} \in \mathbb{R}^{^n}$, the partial derivatives

$$\frac{\partial \text{log} f}{\partial \boldsymbol{\theta}_r}, \quad \frac{\partial^2 \text{log} f}{\partial \boldsymbol{\theta}_r \partial \boldsymbol{\theta}_s}, \quad \frac{\partial^3 \text{log} f}{\partial \boldsymbol{\theta}_r \partial \boldsymbol{\theta}_s \partial \boldsymbol{\theta}_t}, \tag{3.2}$$

exist for all $r, s, t = 1, 2, ..., k$.
(ii) For all $\boldsymbol{\theta} \in N_{\delta_1}(\boldsymbol{\theta}_0)$, for almost all $\boldsymbol{y} \in \mathbb{R}^n$

$$\left|\frac{\partial f}{\partial \boldsymbol{\theta}_r}\right| < H_r(\boldsymbol{y}), \quad \left|\frac{\partial^2 f}{\partial \boldsymbol{\theta}_r \partial \boldsymbol{\theta}_s}\right| < H_{rs}(\boldsymbol{y}), \quad \left|\frac{\partial^3 \text{log} f}{\partial \boldsymbol{\theta}_r \partial \boldsymbol{\theta}_s \partial \boldsymbol{\theta}_t}\right| < H_{rst}(\boldsymbol{y}), \tag{3.3}$$

for all $r, s, t = 1, 2, ..., k$, where $H_r(\boldsymbol{y})$ and $H_{rs}(\boldsymbol{y})$ are integrable with respect to λ_n over $\mathbb{R}^n$, and

$$\int_{\mathbb{R}^n} H_{rst}(\boldsymbol{y}) f(\boldsymbol{y}|\boldsymbol{\theta}_0) d\boldsymbol{y} < M, \tag{3.4}$$

for all $r, s, t = 1, 2, ..., k$, where $0 < M < \infty$.
(iii) For all $\boldsymbol{\theta} \in N_{\delta_1}(\boldsymbol{\theta}_0)$ the $k \times k$ information matrix $I(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[D_{\boldsymbol{\theta}}(\log f)^{\intercal} D_{\boldsymbol{\theta}}(\log f)]$ whose elements are given by

$$(I(\boldsymbol{\theta}))_{rs} = \int_{\mathbb{R}^n} \frac{\partial \text{log} f}{\partial \boldsymbol{\theta}_r} \frac{\partial \text{log} f}{\partial \boldsymbol{\theta}_s} f(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{y}, \tag{3.5}$$

for $r, s = 1, ..., k$, has a finite determinant $|I(\boldsymbol{\theta})|$, and $I(\boldsymbol{\theta}_0)$ is positive-definite.

Condition (i) ensures that $\partial \text{log} f / \partial \boldsymbol{\theta}_r$, for any $r = 1, ..., k$, and for almost all $\boldsymbol{y}$, has a Taylor series expansion as a function of $\boldsymbol{\theta}_r$ (Serfling, 1980, pp145). Note also that the existence of $\partial \text{log} f / \partial \boldsymbol{\theta}_r$ for all $r = 1, ..., k$, and $\partial^2 \text{log} f / \partial \boldsymbol{\theta}_r \partial \boldsymbol{\theta}_s$ for all $r, s = 1, ..., k$, in the neighborhood $N_{\delta_1}(\boldsymbol{\theta}_0)$ given by (i) implies that log$f$ is differentiable (in the vector sense) on Θ° (Magnus and Neudecker, 1999, Theorem 7, pp 101). In turn this implies log$f$, and hence the log-likelihood function $L(\boldsymbol{Y}|\boldsymbol{\theta})$, is continuous on Θ° (Magnus and Neudecker, 1999, Theorem 1, pp96). This is important when the parameter space is restricted to a compact subset of Θ. This is because real valued continuous functions defined on non-empty compact subsets of metric spaces will achieve a maximum and minimum value on this subset (Binmore, 1981, Theorem 19.14, pp74). The parameter space Θ is a metric space with the usual Euclidean distance as the distance function, so that $L(\boldsymbol{Y}|\boldsymbol{\theta})$ will achieve a maximum and minimum on this compact subset.

Condition (ii) means that $\int_{\mathbb{R}^n} f d\boldsymbol{y}$ and $\int_{\mathbb{R}^n} (\partial \text{log} f / \partial \boldsymbol{\theta}_r) d\boldsymbol{y}$ for any $r = 1, ..., k$, can be differentiated with respect to $\boldsymbol{\theta}_r$ under the integral sign (Serfling, 1980, pp145). This leads to $E[\partial \text{log} f / \partial \boldsymbol{\theta}_r] = 0$ for all $r = 1, ..., k$, at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. This latter result is sometimes used as an assumption instead of (ii), see for example Sundberg (1974).

We now very briefly outline the method of proof that Chanda uses to prove the existence of a consistent estimator for $\boldsymbol{\theta}_0$. We have already stated that Kiefer (1978), when using the Chanda result, applies a rule to the parameter space in order to deal with the label switching problem. However because the Chanda result is local in nature, we can instead make the assumption of local identifiability in a sufficiently small neighborhood $N_{\delta_1}(\boldsymbol{\theta}_0)$, and the proof mechanism goes through without change.

This concept of a sufficiently small neighborhood is also used to turn the mixture model estimation problem which is a constrained one, into an unconstrained one. This is important because the consistency proofs in Chanda (1954), and Peters and Walker (1978) are all concerned with unconstrained estimation problems. This approach is used by Redner and Walker (1984, p 211) and presumably works by using a continuity assumption on the parameter space Θ that guarantees $N_{\delta_1}(\boldsymbol{\theta}_0)$ can be chosen, however small, such that for all $\boldsymbol{\theta} \in N_{\delta_1}(\boldsymbol{\theta}_0)$, $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ that $\boldsymbol{\pi} \in \boldsymbol{\theta}$ satisfies the constraints $\sum_{j=1}^{G} (\boldsymbol{\pi})_j = 1, (\boldsymbol{\pi})_j \geq 0, j = 1, ..., G$. Then dropping one of the redundant mixing proportions to form a new parameter space Θ', $\Theta' \subseteq N_{\delta_1}(\boldsymbol{\theta}_0)$, means for all $\boldsymbol{\theta}' \in \Theta'$ that there exists a $\boldsymbol{\theta} \in \Theta \cap N_{\delta_1}(\boldsymbol{\theta}_0)$ such that $\boldsymbol{\theta}' \subset \boldsymbol{\theta}$ and $\boldsymbol{\pi'} \in \boldsymbol{\theta}'$, $\boldsymbol{\pi} \in \boldsymbol{\theta}$ such that $\boldsymbol{\pi'} \subset \boldsymbol{\pi}$ and $\sum_{j=1}^{G} (\boldsymbol{\pi})_j = 1, (\boldsymbol{\pi})_j \geq 0, j = 1, ..., G$. In this way the new mixture model estimation problem with $G - 1$ parameters is unconstrained but equivalent to the constrained estimation problem.

The proof by Aitchison and Silvey (1958) is concerned with constrained maximum likelihood estimation problems, but Sundberg (1974) who uses this theory is applying it to unconstrained problems, presumably by using the method described above of converting from a constrained to an unconstrained estimation problem. Borrowing where possible the notation of Chanda, the conditions (i)-(iii) are designed to imply that

$$L_r(\boldsymbol{\theta}_0) := N^{-1}\sum_{i=1}^{N}\frac{\partial \mathrm{log} f(\boldsymbol{y}_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_r} \xrightarrow{\text{as}} 0 \quad (r = 1, ..., k),$$

$$L_{rs}(\boldsymbol{\theta}_0) := -N^{-1}\sum_{i=1}^{N}\frac{\partial^2 \mathrm{log} f(\boldsymbol{y}_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_r \partial \boldsymbol{\theta}_s} \xrightarrow{\text{as}} (I(\boldsymbol{\theta}_0))_{rs} \quad (r, s = 1, ..., k),$$

$$L_{rst}(\boldsymbol{\theta}') := N^{-1}\sum_{i=1}^{N}\frac{\partial^3 \mathrm{log} f(\boldsymbol{y}_i|\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}_r \partial \boldsymbol{\theta}_s \partial \boldsymbol{\theta}_t} \xrightarrow{\text{as}} C_{rst} \quad (r, s, t = 1, ..., k), \tag{3.6}$$

where $\boldsymbol{\theta}' \in N_{\delta_1}(\boldsymbol{\theta}_0)$, and $C_{rst} < \infty$. We note that by using the weak law of large numbers Chanda gives the above results in terms of convergence in probability only. However the strong law of large numbers can be used too which gives the almost sure convergence results.

Now for any $\boldsymbol{\theta} \in N_{\delta_1}(\boldsymbol{\theta}_0)$, let $d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||$ so that $N_{d(\boldsymbol{\theta},\boldsymbol{\theta}_0)}(\boldsymbol{\theta}_0) \subseteq N_{\delta_1}(\boldsymbol{\theta}_0)$ is the neighborhood around $\boldsymbol{\theta}_0$ of radius $||\boldsymbol{\theta} - \boldsymbol{\theta}_0||$. For any $\boldsymbol{\theta} \in N_{\delta_1}(\boldsymbol{\theta}_0)$, there exists a $\boldsymbol{\theta}' \in N_{d(\boldsymbol{\theta},\boldsymbol{\theta}_0)}(\boldsymbol{\theta}_0)$ whereby $\frac{\partial \mathrm{log} f}{\partial \boldsymbol{\theta}_r}$ can be expanded about $\boldsymbol{\theta}_0$ in a Taylor series expansion such that the likelihood equations in 3.1 scaled by N^{-1} can be written

$$L_r(\boldsymbol{\theta}) = L_r(\boldsymbol{\theta}_0) - \sum_{s=1}^{k}\zeta_s L_{rs}(\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{s=1}^{k}\sum_{t=1}^{k}\zeta_s\zeta_t L_{rst}(\boldsymbol{\theta}') \qquad r = 1, ..., k, \tag{3.7}$$

where $\zeta_s := (\boldsymbol{\theta}_s - (\boldsymbol{\theta}_0)_s)$. Let $L(\boldsymbol{\theta}_0)$ be the matrix with elements $(L(\boldsymbol{\theta}_0))_{rs} := L_{rs}(\boldsymbol{\theta}_0)$. Since $I(\boldsymbol{\theta}_0)$ is positive-definite then $I(\boldsymbol{\theta}_0)^{-1}$ exists, which in turn implies $L(\boldsymbol{\theta}_0)^{-1}$ exists. We will denote by $L^{-1}_{rs}(\boldsymbol{\theta}_0)$ the elements $(L(\boldsymbol{\theta}_0)^{-1})_{rs}$, $r, s = 1, ..., k$. Putting the k equations in (3.7) into matrix form, and multiplying through by $L(\boldsymbol{\theta}_0)^{-1}$ gives

$$\zeta_r = \beta_r + \sum_{s=1}^{k}\sum_{t=1}^{k}\zeta_s\zeta_t a_{rst} \qquad r = 1, ..., k, \tag{3.8}$$

where $\beta_r = \sum_{p=1}^{k} L_p(\boldsymbol{\theta}_0)L^{-1}_{pr}(\boldsymbol{\theta}_0)$, and $a_{rst} = \sum_{p=1}^{k} L_{pst}(\boldsymbol{\theta}')L^{-1}_{pr}(\boldsymbol{\theta}_0)$.

For $r = 1, ..., k$, setting the right hand side of (3.8) to zero and solving for ζ_r is equivalent to setting the right hand side of (3.7) to zero and solving for θ_r, and the k solutions so obtained then represent the solution to the likelihood equations. If we let $\hat{\zeta}_r = \hat{\theta}_r - (\boldsymbol{\theta}_0)_r$, $r = 1, ..., k$, denote these solutions, then $\hat{\boldsymbol{\zeta}}_N(\boldsymbol{Y}) := \hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0 = (\hat{\zeta}_1, ..., \hat{\zeta}_k)^{\intercal}$ is the vector of solutions of the likelihood equations in the neighborhood $N_{\delta_1}(\boldsymbol{\theta}_0)$, regardless of how small δ_1 is.

Using the results in (3.6), Chanda shows for any $\epsilon, \delta_2 > 0$, such that $N_{\delta_2}(\boldsymbol{\theta}_0) \subseteq N_{\delta_1}(\boldsymbol{\theta}_0)$, that for all sufficiently large values of N we have

$$P[|\hat{\boldsymbol{\zeta}}_N(\boldsymbol{Y})| < \delta_2] > 1 - \epsilon. \tag{3.9}$$

This shows that no matter how small we make the neighborhood $N_{\delta_2}(\boldsymbol{\theta}_0)$, and by letting $\epsilon \to 0$, a solution of the likelihood equations exists in this neighborhood with probability tending to 1. The probability statement in (3.9) is also a probability statement about a set $S_N \in \mathscr{R}^{Nn}$ with respect to the distribution $P_{\boldsymbol{\theta}_0}$ of the joint vector $\boldsymbol{Y} \in \mathbb{R}^{Nn}$: if for $N = 1, 2, ...,$ we let $S_N \subseteq \mathbb{R}^{Nn}$ denote the set of all points $\boldsymbol{y} \in \mathbb{R}^{Nn}$ such that $|\hat{\boldsymbol{\zeta}}_N(\boldsymbol{y})| < \delta_2$ holds, then for any $\boldsymbol{y} \in \mathbb{R}^{Nn}$ we have $P_{\boldsymbol{\theta}_0}\{\boldsymbol{y} \in S_N\} > 1 - \epsilon$ for all sufficiently large N. Since $P_{\boldsymbol{\theta}_0}\{\boldsymbol{y} \in S_N\} \equiv \mathscr{P}_0\{S_N\}$ for all $\boldsymbol{y} \in S_N$ (see Cramér (1946, pp502)), then we have a sequence of random variables $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$, and sets $\{S_N\}$, such that for all sufficiently large N, the solution $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$ is in the neighborhood $N_{\delta_2}(\boldsymbol{\theta}_0)$, and $\mathscr{P}_0\{S_N\} > 1 - \epsilon$.

From (3.9) we see that for any $0 < \delta_2 < \delta_1$, and by letting $\epsilon \to 0$, we have that $\lim_{N\to\infty} P\{|\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0| < \delta_2\} = 1$, and so the sequence $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$ converges in probability to $\boldsymbol{\theta}_0$. Since this holds for all $\boldsymbol{\theta}_0 \in \Theta^0$ the sequence $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$ is consistent for θ_0. Thus a consistent solution to the scaled likelihood equations in (3.7), and thus to the likelihood equations (3.1) themselves exist. A similar result was obtained by Aitchison and Silvey (1958) in their Theorem 1.

We now discuss whether the consistent solution to the likelihood equations produces a local maximum of $L_N(\boldsymbol{\theta})$, and whether two distinct consistent sequences of solutions can exist. We will need the fact that $\log f(\boldsymbol{y}|\boldsymbol{\theta})$, and so in turn $L_N(\boldsymbol{\theta})$, are C^2 functions on $N_{\delta_1}(\boldsymbol{\theta}_0)$, that is both the first and second derivatives exist and are continuous on this neighbourhood. We can derive this from assumption (i): the existence of the third derivatives of $\log f(\boldsymbol{y}|\boldsymbol{\theta})$ in $N_{\delta_1}(\boldsymbol{\theta}_0)$ implies the second derivatives of $\log f(\boldsymbol{y}|\boldsymbol{\theta})$ are continuous functions in $N_{\delta_1}(\boldsymbol{\theta}_0)$. Similarly the existence of the second derivatives of $\log f(\boldsymbol{y}|\boldsymbol{\theta})$ implies the first derivatives of $\log f(\boldsymbol{y}|\boldsymbol{\theta})$ are continuous functions in $N_{\delta_1}(\boldsymbol{\theta}_0)$.

We first consider whether the consistent sequence of solutions produce a maximum of $L_N(\boldsymbol{\theta})$, but firstly we need the result that for sufficiently small $\delta_3 > 0$, a neighborhood $N_{\delta_3}(\boldsymbol{\theta}_0)$ exists such that $I(\boldsymbol{\theta})$ is a positive-definite matrix for all $\boldsymbol{\theta} \in N_{\delta_3}(\theta_0)$.

To see this let $A_r(\boldsymbol{\theta})$, $r = 1, ..., k$, be the $r \times r$ principle minors of $I(\boldsymbol{\theta})$, that is $A_1(\boldsymbol{\theta}), A_2(\boldsymbol{\theta}), ..., A_k(\boldsymbol{\theta})$ are the sub-matrices in the upper left 1×1, 2×2,..., and $k \times k$ corners respectively of $I(\boldsymbol{\theta})$. Using the fact that $I(\boldsymbol{\theta}_0)$ is a Hermitian matrix, we have by Sylvester's criterion that for all $r \in \{1, ..., k\}$, $\det(A_r(\boldsymbol{\theta}_0)) > 0$. Now (ii) implies that for all $r, s \in \{1, ..., k\}$, $E_{\boldsymbol{\theta}}\left[\partial^2 \text{log} f(\boldsymbol{Y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}_r\partial\boldsymbol{\theta}_s\right]$ are continuous on $N_{\delta_1}(\boldsymbol{\theta}_0)$, and so $(I(\boldsymbol{\theta}))_{rs}$ are also continuous on $N_{\delta_1}(\boldsymbol{\theta}_0)$ (for brevity we omit this derivation). Thus the maps $g_r : \boldsymbol{\theta} \mapsto \det(A_r(\boldsymbol{\theta}))$ are continuous on $N_{\delta_1}(\boldsymbol{\theta}_0)$. Putting $\eta' = \min\{\eta_1, ..., \eta_k\}$, where $0 < \eta_r < g_r(\boldsymbol{\theta}_0)$ for $r \in \{1, ...k\}$, we see that there exists an $\delta_3 > 0$ such that for all $r \in \{1, ..., k\}$, $0 < g_r(\boldsymbol{\theta}_0) - \eta' < g_r(\boldsymbol{\theta}')$ for all $\boldsymbol{\theta}' \in N_{\delta_3}(\boldsymbol{\theta}_0)$. We therefore see that all the principle minor determinants of $I(\boldsymbol{\theta}')$ are positive in the neighborhood $N_{\delta_3}(\boldsymbol{\theta}_0)$. By Sylvester's criterion we then have that $I(\boldsymbol{\theta}')$ is positive-definite for all $\boldsymbol{\theta}' \in N_{\delta_3}(\boldsymbol{\theta}_0)$.

Now for $r, s \in \{1, ..., k\}$, and from a first order Taylor expansion of $(H_{\boldsymbol{\theta}}(\boldsymbol{\theta}))_{rs} = -NL_{rs}(\boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$ we get

$$-L_{rs}(\hat{\boldsymbol{\theta}}_N) = -L_{rs}(\boldsymbol{\theta}_0) + N^{-1}L_{rst}(\boldsymbol{\theta}')((\hat{\boldsymbol{\theta}}_N)_s - (\boldsymbol{\theta}_0)_s), \tag{3.10}$$

where $\hat{\boldsymbol{\theta}}_N$ is the consistent solution for $\boldsymbol{\theta}_0$, and $\boldsymbol{\theta}' \in N_{d(\hat{\boldsymbol{\theta}}_N, \boldsymbol{\theta}_0)}(\boldsymbol{\theta}_0)$. Now if $\hat{\boldsymbol{\theta}}_N$ is close enough to $\boldsymbol{\theta}_0$ such that $N_{d(\hat{\boldsymbol{\theta}}_N, \boldsymbol{\theta}_0)}(\boldsymbol{\theta}_0) \subseteq N_{\delta_1}(\boldsymbol{\theta}_0)$, then by (ii) $N^{-1}L_{rst}(\boldsymbol{\theta}') = O_P(1)$, which implies $N^{-1}L_{rst}(\boldsymbol{\theta}')((\hat{\boldsymbol{\theta}}_N)_s - (\boldsymbol{\theta}_0)_s) = o_P(1)$ since $((\hat{\boldsymbol{\theta}}_N)_s - (\boldsymbol{\theta}_0)_s)$ tends in probability to zero. Since from (3.6) we see that $-L_{rs}(\boldsymbol{\theta}_0)$ tends in probability to $-(I(\boldsymbol{\theta}_0))_{rs}$, we have from (3.10) that

$$-L_{rs}(\hat{\boldsymbol{\theta}}_N) \xrightarrow{\text{P}} -(I(\boldsymbol{\theta}_0))_{rs} \qquad\qquad (r, s \in \{1, ..., k\}) \tag{3.11}$$

for any consistent solution $\hat{\boldsymbol{\theta}}_N$. Since $-I(\boldsymbol{\theta})$ is a negative-definite matrix for all $\boldsymbol{\theta} \in N_{\delta_3}(\theta_0)$, for any $\epsilon > 0$, we can take N large enough such that $\hat{\boldsymbol{\theta}}_N \in N_{\delta_1}(\theta_0) \cap N_{\delta_3}(\theta_0)$, and $N^{-1}H_{\boldsymbol{\theta}}(L_N(\hat{\boldsymbol{\theta}}_N))$ is negative-definite with probability greater than $1 - \epsilon$. Thus by the multivariate characterization of maxima of C^2 functions, and with probability tending to 1 as $N \to \infty$, the consistent sequence of solutions $\{\hat{\boldsymbol{\theta}}_N\}$ must produce a local maximum of the scaled likelihood equations in (3.7), and thus to the likelihood equations (3.1) themselves. Chanda proves this result in Theorem 1.

The uniqueness of the local maximum produced by the consistent solution $\hat{\boldsymbol{\theta}}_N$ can be established if we further suppose $N_{\delta_1}(\boldsymbol{\theta}_0)$ is a convex set. Given this new assumption we see from the above paragraph that $N^{-1}L_N(\boldsymbol{\theta})$ and hence $L_N(\boldsymbol{\theta})$ is strictly concave on $N_{\delta_1}(\theta_0) \cap N_{\delta_3}(\theta_0)$, which means any local maximum of $L_N(\boldsymbol{\theta})$ is also a global maximum. Thus with probability tending to 1 as $N \to \infty$ the consistent sequence of solutions $\{\hat{\boldsymbol{\theta}}_N\}$ must produce a global maximum of the likelihood equations. In terms of sequences this means that if $\{S_N\}$ and $\{S'_N\}$ are two sequences of sets that give rise to two consistent sequences of solutions $\{\hat{\boldsymbol{\theta}}_N\}$ and $\{\hat{\boldsymbol{\theta}}'_N\}$, then with probability tending to 1 as $N \to \infty$ the elements $\hat{\boldsymbol{\theta}}_N$ and $\hat{\boldsymbol{\theta}}'_N$ both tend to the same point - the global maximum in $N_{\delta_1}(\theta_0) \cap N_{\delta_3}(\theta_0)$. This means for any $\epsilon > 0$ that there exists a N' such that $|\hat{\boldsymbol{\theta}}_N - \hat{\boldsymbol{\theta}}'_N| < \epsilon$ for all $N > N'$. Thus a consistent sequence of solutions is uniquely determined up to a finite number of points, and all other sequences of solutions must be inconsistent. Chanda proved this in his Theorem 2, however it turns out his proof was incorrect. A corrected version was given by Tarone and Gruenhage (1975) who made the additional assumption to (i)-(iii) that Θ is a convex subset of $\mathbb{R}^k$. A similar convexity assumption was used by Sundberg to establish this local uniqueness property.

Given the existence of a consistent estimator $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$, Chanda also shows that

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0\right) \overset{D}{\to} N_k\left[0, I(\boldsymbol{\theta}_0)^{-1}\right]. \tag{3.12}$$

Thus to summarise: In a sufficiently small neighborhood of $\boldsymbol{\theta}_0$ such that assumptions (i)-(iii), and local identifiability holds, then with probability tending to 1 as $N \to \infty$ there exists a unique consistent sequence of solutions $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$ for $\boldsymbol{\theta}_0$ that produce local maxma of $L_N(\boldsymbol{\theta})$. Furthermore $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$ is asymptotically normally distributed with zero mean and covariance matrix $I(\boldsymbol{\theta}_0)^{-1}$. An analogue of this result using almost sure convergence is given by Redner and Walker (1984) in Theorem 3.1 where Peters and Walker (1978) is used to give the strong consistency result.

The above result may appear better than it actually is, since in practice $L_N(\boldsymbol{\theta})$ may have multiple solutions for any particular N, but the theory does not tell us which solution of the likelihood equations we should pick - only pre-knowledge of $\boldsymbol{\theta}_0$ would tell us this. As Sundberg states, the above result holds if $\boldsymbol{\theta}_0$ is replaced by any $\boldsymbol{\theta} \in \Theta$ that is equivalent to $\boldsymbol{\theta}_0$, where equivalent means $f(\boldsymbol{Y}|\boldsymbol{\theta}_0) = f(\boldsymbol{Y}|\boldsymbol{\theta})$ almost everywhere. We shall denote this equivalence as $\boldsymbol{\theta}_0 \sim \boldsymbol{\theta}$, and the set of all points equivalent to $\boldsymbol{\theta}_0$ as $\Theta(\boldsymbol{\theta}_0)$. So if $\boldsymbol{\theta}' \in \Theta(\boldsymbol{\theta}_0)$ then a consistent sequence of solutions $\{\hat{\boldsymbol{\theta}}'_N(\boldsymbol{Y})\}$ exists for $\boldsymbol{\theta}'$, and $\hat{\boldsymbol{\theta}}'_N(\boldsymbol{Y})$ is asymptotically normally distributed with zero mean and covariance matrix $I(\boldsymbol{\theta}')^{-1}$. If p_π represents a permutation of the component labels such that $p_\pi(\boldsymbol{\theta}_0) = \boldsymbol{\theta}'$ then $\{p_\pi(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))\} = \{\hat{\boldsymbol{\theta}}'_N(\boldsymbol{Y})\}$. Thus the consistent estimators of the $G!$ different versions of $\boldsymbol{\theta}_0$ contained in $\Theta(\boldsymbol{\theta}_0)$ can be obtained from $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$ by label switching.

Sundberg extends the previously described result to equivalence classes by proving a theorem concerned with restricting the parameter space to a compact subset

$K \subseteq \Theta$, where K contains an element of the equivalence class $\Theta_0 := \Theta(\theta_0)$. Using properties of distributions from the exponential family, in Theorem 7.1 Sundberg shows that if $\boldsymbol{\theta}_N^{\max} \in K$ is the maximum point of $L_N(\boldsymbol{\theta})$, then $\operatorname{dist}(\boldsymbol{\theta}_N^{\max}, \Theta_0) \overset{\text{as}}{\to} 0$, where $\operatorname{dist}(\boldsymbol{\theta}_N^{\max}, \Theta_0) = \inf_{\boldsymbol{\theta}_0 \in \Theta_0} ||\boldsymbol{\theta}_N^{\max} - \theta_0||$. The maximum point $\boldsymbol{\theta}_N^{\max}$ always exists since by assumption (i), $L_N(\boldsymbol{\theta})$ always attains a maximum and minimum value on K.

Using the above result, if we let $\boldsymbol{\theta}_N^{\max}$ denote the global maximum of $L_N(\boldsymbol{\theta})$ on the compact subset $K \subseteq \Theta$, then with probability 1 as $N \to \infty$ some member of Θ_0, $\boldsymbol{\theta}_0^{\max}$ say, is eventually equal to $\boldsymbol{\theta}_N^{\max}$. Since $\boldsymbol{\theta}_0 \in \Theta_0$ then $L_N(\boldsymbol{\theta}_0) = L_N(\boldsymbol{\theta}_0^{\max}) = L_N(\boldsymbol{\theta}_N^{\max})$ almost everywhere (using the definition of Θ_0), so we see with probability 1 as $N \to \infty$ that $\boldsymbol{\theta}_0$, and indeed any $\boldsymbol{\theta}' \in \Theta_0$, produce a log-likelihood function value that is the same as the global maximum on K. Accordingly if $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$ are a consistent sequence of roots that are local maxima of $L_N(\boldsymbol{\theta})$ proven to exist by Sundberg, and say they are consistent for $\boldsymbol{\theta}' \in \Theta_0$ in some neighborhood $N_\delta(\boldsymbol{\theta}')$, then with probability tending to 1 as $N \to \infty$ we have that $L_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ is almost everywhere equal to the global maximum of $L_N(\boldsymbol{\theta})$ on K. This result obviously holds for all points in Θ_0, and so whichever point of Θ_0 the sequence $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$ is consistent for, with probability tending to 1 as $N \to \infty$, $L_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ is almost everywhere equal to the global maximum of $L_N(\boldsymbol{\theta})$ on K. We note that some of the points of Θ_0 may be outside of K.

We defer interpretation of the Theorem 7.1 of Sundberg for mixture densities until after we have described a similar result in Theorem 3.2 of Redner and Walker (1984). The result states that if K is any compact subset of Θ such that $\boldsymbol{\theta}_0 \in K^\circ$, C is the set of points in K where $f(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})$ is almost-everywhere equal to $f(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}_0)$, and D is any closed subset of K with no points in common with C, then with probability 1

$$\lim_{N\to\infty} \sup_{\boldsymbol{\theta}\in D} \frac{\prod_{i=1}^{N} f(\boldsymbol{y}_i|\boldsymbol{\theta})}{\prod_{i=1}^{N} f(\boldsymbol{y}_i|\boldsymbol{\theta}_0)} = 0. \tag{3.13}$$

Assume that $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$ is a strongly consistent sequence of solutions for $\boldsymbol{\theta}' \in \Theta_0$ in the neighborhood $N_\delta(\boldsymbol{\theta}')$. Then with probability 1 as $N \to \infty$ we have $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) \to \boldsymbol{\theta}_0$, and also that $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$, and hence $\boldsymbol{\theta}'$, produces a local maximum of $L_N(\boldsymbol{\theta})$ on $N_\delta(\boldsymbol{\theta}')$. Since $L_N(\boldsymbol{\theta}') = L_N(\boldsymbol{\theta}'')$ almost everywhere for all $\boldsymbol{\theta}'' \in \Theta_0$, then with probability 1 as $N \to \infty$ we have $L_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})) = L_N(\boldsymbol{\theta}_0)$ almost everywhere. Let $N(C) \subseteq K$ be any open neighbourhood containing C, then $K - N(C)$ is a closed subset of K that has no points in common with C. By (3.13) we have that with probability 1 as $N \to \infty$, any $\boldsymbol{\theta} \in K$ such that $L_N(\boldsymbol{\theta}) \geq L_N(\boldsymbol{\theta}_0)$ cannot be contained in $K - N(C)$. So we must have $\boldsymbol{\theta}' \in K^c$ or $\boldsymbol{\theta}' \in N(C)$ which implies $\boldsymbol{\theta}' \in K$. We conclude with probability 1 as $N \to \infty$ that $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$ either produces a global maximum of $L_N(\boldsymbol{\theta})$ on K, or a value of $L_N(\boldsymbol{\theta})$ outside K that is almost everywhere equal to $L_N(\boldsymbol{\theta}_0)$.

Now for mixture densities since we are assuming identifiability holds up to label switching, then the set C in Theorem 3.2 of Redner and Walker is equal to the equivalence class Θ_0. This means that all the global maxima of $L_N(\boldsymbol{\theta})$ on K can be found through label-switching of $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$. However points in Θ_0 not in K cannot be guaranteed to produce global maxima of $L_N(\boldsymbol{\theta})$, but they will produce values of $L_N(\boldsymbol{\theta})$ almost everywhere equal to $L_N(\boldsymbol{\theta}_0)$. The same interpretation holds using Theorem 7.1 of Sundberg.

To conclude, it has been shown by Sundberg (1974) for mixtures of densities from the exponential family, and by Redner and Walker (1984) for mixtures of general densities, that the multi parameter analogue of the result given by Cramér (1946, p500) holds. That is with probability tending to 1 (or with probability 1 for the Redner and Walker result) there exists a unique consistent sequence of solutions $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$ to the likelihood equations in some neighborhood $N_\delta(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ that yield a local maximum of $L_N(\boldsymbol{\theta}_0)$, and that $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$ is asymptotically normally distributed with zero mean and covariance matrix $I(\boldsymbol{\theta}_0)^{-1}$. In both cases identifiability up to label switching on Θ is assumed. Furthermore the concept of local identifiability was also used which assumes the neighborhood $N_\delta(\boldsymbol{\theta}_0)$ is small enough such that no label switched versions of $\boldsymbol{\theta}_0$ exist in $N_\delta(\boldsymbol{\theta}_0)$.

For mixture densities with no regression parameters that induce an *iid* sample of random variables, the results of Chanda (1954) can also be used to derive the same result (again identifiability up to label switching and local identifiability on a neighborhood need to be used). In this respect the application of this result by Kiefer (1978) to mixture densities from a non-*iid* sample generated from a switching regression model is questionable, indeed Kiefer has a footnote effectively admitting this. Finally both Sundberg and Redner and Walker show that if a compact subset $K \subseteq \Theta$ can be found with $\boldsymbol{\theta}_0 \in K^\circ$, then at least one of the consistent sequences of solutions $\{\hat{\boldsymbol{\theta}}'_N(\boldsymbol{Y})\}$ for a point in Θ_0 approaches a global maximum on K, and all the other global maxima on K can be found through label switching of $\hat{\boldsymbol{\theta}}'_N(\boldsymbol{Y})$.

3.2 Approximating the information matrix for *iid* Gaussian mixtures

In this section we describe a simulation study by Boldea and Magnus (2009) the aim of which was to quantify for small sample sizes the performance of the estimated standard errors for $\hat{\boldsymbol{\theta}}$, where the standard errors were obtained from the inverse of the sample information matrix $I_N(\hat{\boldsymbol{\theta}})$. Boldea and Magnus study multivariate responses from *iid* samples that are distributed according to a finite Gaussian mixture, and so only mean vectors and covariance matrices of normal distributions are required to be estimated rather than regression or covariance parameters. The use of the information matrix to obtain standard errors implies Boldea and Magnus are using the theory of Redner and Walker (1984) that was described in Section 3.1. We include a section here on the work of Boldea and Magnus because in Section 3.4 we propose to adapt their approach to statistical inference and apply it to MLMMs. Boldea and Magnus describe computational difficulties regarding taking expectations of the mixture likelihood function and so they propose a number of approximations to the information matrix $I(\boldsymbol{\theta}_0)$ which are readily calculable. For this reason we include the next Subsection (3.2.1) that provides an outline of the relevant theory justifying these approximations.

3.2.1 Consistent estimators of the information matrix

For this section we introduce the random variable $\tilde{\boldsymbol{Y}} \in \mathbb{R}^n$ with distribution $\mu_{\boldsymbol{\theta}_0}$, where $\boldsymbol{\theta}_0 \in \Theta$, and $\Theta \subseteq \mathbb{R}^k$. The distribution $\mu_{\boldsymbol{\theta}_0}$ is a probability measure on the measurable space $(\mathbb{R}^n, \mathscr{R}^n)$, and $\mathscr{D} := \{\mu_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is the family of such distributions parametrized by points in Θ. We further suppose each $\mu_{\boldsymbol{\theta}}$ has a general (i.e. not necessarily a mixture) density function $f(\cdot|\boldsymbol{\theta})$ with respect to n-dimensional Lebesgue measure. We then assume we have an *iid* sample $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$ of random variables, each of which has the same distribution as $\tilde{\boldsymbol{Y}}$. These assumptions imply for any $\boldsymbol{\theta}_0 \in \Theta$ that $f(\cdot|\boldsymbol{\theta}_0)$ is the true density function for $\tilde{\boldsymbol{Y}}$, rather than another unknown density function $g(\cdot|\boldsymbol{\theta}_0)$ say. Unless otherwise stated all of the results in this section assume that $f(\cdot|\boldsymbol{\theta}_0)$ is the true density function for $\tilde{\boldsymbol{Y}}$.

The Fisher expected information matrix $I(\boldsymbol{\theta})$, or just information matrix for short, evaluated at $\boldsymbol{\theta} \in \Theta^\circ$ is defined as

$$I(\boldsymbol{\theta}) = \boldsymbol{E_\theta}\left[\boldsymbol{D_\theta}\left(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}))\right)^\intercal \boldsymbol{D_\theta}\left(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}))\right)\right], \tag{3.14}$$

where $\boldsymbol{D_\theta}\left(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}))\right)^\intercal$ is the $k \times 1$ gradient vector (or score vector) of $\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})$. Recalling that λ_n denotes n-dimensional Lebesgue measure, consider now some regularity conditions known as the Fisher information regularity conditions (Schervish, 1995, p 111)

(FI i) There exists a $B \in \mathscr{R}^n$ with $\lambda_n(B^c) = 0$, such that for all $\boldsymbol{\theta} \in \Theta^\circ$, $\partial \log f/\partial\theta_r$ exists for all $\boldsymbol{y} \in B$ and for each $r = 1, ..., k$,

(FI ii) $\int_{\mathbb{R}^n} f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y}$ can be differentiated under the integral sign with respect to $\boldsymbol{\theta}_r$ for all $r = 1, ..., k$,

(FI iii) The set $C = \{\boldsymbol{y} \in \mathbb{R}^n : f(\boldsymbol{y}|\boldsymbol{\theta}) > 0\}$ is the same for all $\boldsymbol{\theta} \in \Theta$.

Given these regularity conditions $I(\boldsymbol{\theta})$ can be written

$$I(\boldsymbol{\theta}) = \text{var}_{\boldsymbol{\theta}}\left[\boldsymbol{D_\theta}\left(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}))\right)^\intercal\right]. \tag{3.15}$$

This can easily be seen by noting that

$$\begin{aligned} E_{\boldsymbol{\theta}}\left[\boldsymbol{D_\theta}\left(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})\right)\right] &= \int_{R^n} \frac{\boldsymbol{D_\theta}[f(\boldsymbol{y}|\boldsymbol{\theta})]}{f(\boldsymbol{y}|\boldsymbol{\theta})} f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y} \\ &= \boldsymbol{D_\theta}\left[\int_{R^n} f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y}\right] \\ &= \boldsymbol{0}, \end{aligned} \tag{3.16}$$

where the last line of (3.16) follows because $f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})$ integrates to one. Note that (3.16) depends on $f(\cdot|\boldsymbol{\theta})$ being the true density function of $\tilde{\boldsymbol{Y}}$ for any $\boldsymbol{\theta} \in \Theta^\circ$. Using the fact that for any random vector X with expectation α we have $\text{var}(X) = \boldsymbol{E}[XX^\intercal] + \boldsymbol{E}[X]\boldsymbol{E}[X]^\intercal$ we get the result in (3.15).

There is another useful way of writing $I_N(\boldsymbol{\theta})$, which relies on being able to differentiate $\int_{\mathbb{R}^n} \partial f(\boldsymbol{y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}_r d\boldsymbol{y}$ under the integral sign with respect to $\boldsymbol{\theta}_r$, for all $r \in \{1, ..., k\}$, which is equivalent to modifying (FI ii) to assume $\int_{\mathbb{R}^n} f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y}$ can be differentiated twice under the integral sign. Given this assumption we have

$$I(\boldsymbol{\theta}) = -\boldsymbol{E}_{\boldsymbol{\theta}}\left[\boldsymbol{H}_{\boldsymbol{\theta}}(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})\right]. \tag{3.17}$$

This can also be easily seen by noting that

$$\boldsymbol{H}_{\boldsymbol{\theta}}[\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})] = \frac{\boldsymbol{H}_{\boldsymbol{\theta}}[f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})]}{f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})} - \frac{\boldsymbol{D}_{\boldsymbol{\theta}}[f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})]^{\intercal}\boldsymbol{D}_{\boldsymbol{\theta}}[f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})]}{f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})^2}, \tag{3.18}$$

and so

$$\begin{aligned} -\boldsymbol{E}_{\boldsymbol{\theta}}\left[\boldsymbol{H}_{\boldsymbol{\theta}}(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})\right] &= -\boldsymbol{D}_{\boldsymbol{\theta}}\left[\boldsymbol{D}_{\boldsymbol{\theta}}\left[\int_{R^n} f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y}\right]\right] \\ &\quad + \boldsymbol{E}_{\boldsymbol{\theta}}\left[\boldsymbol{D}_{\boldsymbol{\theta}}\left(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}))\right)^{\intercal}\boldsymbol{D}_{\boldsymbol{\theta}}\left(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}))\right)\right]. \end{aligned} \tag{3.19}$$

Using the same reasoning which led to the result (3.16), we see that the first term on the right hand side of (3.19) is zero, which using the definition of $I(\boldsymbol{\theta})$ in (3.14) gives the result in (3.17). Note again by virtue of relying on (3.16) that (3.19) depends on $f(\cdot|\boldsymbol{\theta})$ being the true density function of $\tilde{\boldsymbol{Y}}$ for any $\boldsymbol{\theta} \in \Theta^{\circ}$.

Defining $I_{\boldsymbol{Y}_i}(\boldsymbol{\theta})$ and $I_N(\boldsymbol{\theta})$ to be the information matrices as functions of $\boldsymbol{Y}_i$, $i \in I_N$, and $\boldsymbol{Y}$ respectively where $\boldsymbol{Y} = (\boldsymbol{Y}_1^{\intercal}, ..., \boldsymbol{Y}_N^{\intercal})^{\intercal}$, then since the sample is independent from Lehmann and Casella (1998, Theorem 5.8, pp119) we have

$$I_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} I_{\boldsymbol{Y}_i}(\boldsymbol{\theta}), \tag{3.20}$$

where

$$\begin{aligned} I_{\boldsymbol{Y}_i}(\boldsymbol{\theta}) &= \boldsymbol{E}_{\boldsymbol{\theta}}\left[\boldsymbol{D}_{\boldsymbol{\theta}}\left(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta}))\right)^{\intercal}\boldsymbol{D}_{\boldsymbol{\theta}}\left(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta}))\right)\right], \\ &= \operatorname{var}_{\boldsymbol{\theta}}\left[\boldsymbol{D}_{\boldsymbol{\theta}}\left(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta}))\right)^{\intercal}\right], \\ &= -\boldsymbol{E}_{\boldsymbol{\theta}}\left[\boldsymbol{H}_{\boldsymbol{\theta}}(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta})\right], \end{aligned} \tag{3.21}$$

using the identities in (3.14), (3.15) and (3.17). Now assume $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$ is a strongly consistent sequence of solutions for $\boldsymbol{\theta}_0$, and that $I(\boldsymbol{\theta})$ is continuous on $\mathbb{R}^k$, so that by the Continuous Mapping Theorem (Van Der Vaart, 1998, p 7) we have $I(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})) \xrightarrow{a.s.} I(\boldsymbol{\theta}_0)$. Since the sample is *iid* then $I_N(\boldsymbol{\theta}) = NI_{\tilde{\boldsymbol{Y}}}(\boldsymbol{\theta}) = NI(\boldsymbol{\theta})$, which implies $N^{-1}I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})) \xrightarrow{a.s.} I(\boldsymbol{\theta}_0)$. This holds for all $\theta_0 \in \Theta^{\circ}$ and so $N^{-1}I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ is a strongly consistent estimator of $I(\boldsymbol{\theta}_0)$.

We now introduce and discuss two more random matrices which again under certain regularity conditions converge in probability to $I(\boldsymbol{\theta}_0)$. Define the observed Fisher information matrix as

$$J(\boldsymbol{\theta}) = -\boldsymbol{H}_{\boldsymbol{\theta}}(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})), \tag{3.22}$$

and the outer product of the score vector as

$$S(\boldsymbol{\theta}) = \boldsymbol{D}_{\boldsymbol{\theta}}(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}))^{\intercal} \boldsymbol{D}_{\boldsymbol{\theta}}(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta})). \tag{3.23}$$

Letting $J_N(\boldsymbol{\theta})$ and $J_{\boldsymbol{Y}_i}(\boldsymbol{\theta})$ be the observed information matrices for $\boldsymbol{Y}$ and all the $\boldsymbol{Y}_i$ respectively, then again by independence we have that $J_N(\boldsymbol{\theta}) = \sum_{i=1}^N J_{\boldsymbol{Y}_i}(\boldsymbol{\theta})$. Similarly letting $S_N(\boldsymbol{\theta})$ and $S_{\boldsymbol{Y}_i}(\boldsymbol{\theta})$ be the outer product of the score vectors for $\boldsymbol{Y}$ and all the $\boldsymbol{Y}_i$ respectively, then $S_N(\boldsymbol{\theta}) = \sum_{i=1}^N S_{\boldsymbol{Y}_i}(\boldsymbol{\theta})$.

Now $(J_N(\boldsymbol{\theta}))_{rs} = -\sum_{i=1}^N (\boldsymbol{H}_{\boldsymbol{\theta}}(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta}))_{rs}$ is element (r, s) of $J_N(\boldsymbol{\theta})$, for $r, s \in \{1, ..., k\}$. If the sample is iid, and if we assume $\boldsymbol{E}_{\boldsymbol{\theta}}[(\boldsymbol{H}_{\boldsymbol{\theta}}(\log f(\boldsymbol{Y}_1|\boldsymbol{\theta}))_{rs}] < \infty$, then by the strong law of large numbers (SLLN) the average of $(J_N(\boldsymbol{\theta}))_{rs}$ converges almost surely to $-\boldsymbol{E}_{\boldsymbol{\theta}}[(\boldsymbol{H}_{\boldsymbol{\theta}}(\log f(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}))_{rs}] = (I_{\tilde{\boldsymbol{Y}}}(\boldsymbol{\theta}))_{rs} = (I(\boldsymbol{\theta}))_{rs}$ using (3.17). Thus $N^{-1}(J_N(\boldsymbol{\theta}))_{rs} \xrightarrow{a.s.} (I(\boldsymbol{\theta}))_{rs}$ for all $r, s \in \{1, ..., k\}$, and so $N^{-1}J_N(\boldsymbol{\theta}) \xrightarrow{a.s.} I(\boldsymbol{\theta})$. Unfortunately we cannot simply conclude $N^{-1}J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})) \xrightarrow{a.s.} I(\boldsymbol{\theta}_0)$ by arguing as before for the estimator $N^{-1}I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ of $I(\boldsymbol{\theta}_0)$. An alternative approach uses a uniform law of large numbers (ULLN). We now describe this approach.

Let us temporarily adopt the notation $J(\tilde{\boldsymbol{Y}}, \boldsymbol{\theta}) := J(\boldsymbol{\theta})$ in order to be explicit about the dependence of $J(\boldsymbol{\theta})$ on the random vector $\tilde{\boldsymbol{Y}}$. We shall also work elementwise with $(J(\tilde{\boldsymbol{Y}}, \boldsymbol{\theta}))_{rs}$ and $(J_N(\boldsymbol{\theta}))_{rs} = \sum_{i=1}^N (J(\boldsymbol{Y}_i, \boldsymbol{\theta}))_{rs}$, $r, s \in \{1, ..., k\}$, for this discussion. The function $(J(\cdot, \boldsymbol{\theta}))_{rs}$ is real-valued on the set $\mathbb{R}^n \times \Theta^\circ$, and we will suppose that it is Lebesgue measurable for every $\boldsymbol{\theta} \in \Theta^\circ$. A uniform (strong) law of large numbers defines a set of conditions under which

$$\begin{aligned} &\sup_{\boldsymbol{\theta}\in\Theta} \left|N^{-1}(J_N(\boldsymbol{\theta}))_{rs} - \boldsymbol{E}_{\boldsymbol{\theta}}\left[(J(\boldsymbol{Y}_1, \boldsymbol{\theta}))_{rs}\right]\right| = \\ \sup_{\boldsymbol{\theta}\in\Theta} &\left|N^{-1} \sum\nolimits_{i=1}^N (J(\boldsymbol{Y}_i, \boldsymbol{\theta}))_{rs} - (I(\boldsymbol{\theta}))_{rs}\right| \xrightarrow{a.s} 0. \end{aligned} \tag{3.24}$$

The conditions that must be satisfied in order that (3.24) holds are (a) Θ° is a compact set; (b) $(J(\tilde{\boldsymbol{Y}},\boldsymbol{\theta}))_{rs}$ is a continuous function on Θ° with probability 1; (c) for each $\boldsymbol{\theta} \in \Theta^{\circ}$ $(J(\tilde{\boldsymbol{Y}},\boldsymbol{\theta}))_{rs}$ is dominated by a function $h(\tilde{\boldsymbol{Y}})$, i.e. $|(J(\tilde{\boldsymbol{Y}},\boldsymbol{\theta}))_{rs}| < h(\tilde{\boldsymbol{Y}})$; and (d) for each $\boldsymbol{\theta} \in \Theta^{\circ}$ $E_{\boldsymbol{\theta}}[h(\tilde{\boldsymbol{Y}})] < \infty$;. These conditions come from Jennrich (1969, Theorem 2).

Now for any $\boldsymbol{y}_i \in \mathbb{R}^n$, $i \in I_N$ and $\boldsymbol{\theta}' \in S \subseteq \Theta^{\circ}$, the following inequality obviously holds

$$\left| N^{-1} \sum\nolimits_{i=1}^{N} (J(\boldsymbol{y}_i,\boldsymbol{\theta}'))_{rs} - (I(\boldsymbol{\theta}'))_{rs} \right| \leq \sup_{\boldsymbol{\theta}\in S} \left| N^{-1} \sum\nolimits_{i=1}^{N} (J(\boldsymbol{y}_i,\boldsymbol{\theta}))_{rs} - (I(\boldsymbol{\theta}))_{rs} \right|. \tag{3.25}$$

Suppose that $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$ is a strongly consistent sequence of estimators for $\boldsymbol{\theta}_0$, and let $\Theta_{N_1} = B_{\delta_{N_1}}(\boldsymbol{\theta}_0) \subseteq K \subseteq \Theta^{\circ}$ be an open ball in $\mathbb{R}^k$ with radius $\delta_{N_1} \to 0$ as $N_1 \to \infty$, and suppose K is compact. Then since $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) \in \Theta_{N_1}$ for fixed N_1, and for N sufficiently large enough, then we have that $P[\lim_{N}\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) \in \Theta_{N_1}\}] = 1$ for sufficiently large N. Together with (3.25) this implies

$$P\left[\lim_{N\to\infty} \left\{ \left| N^{-1} \sum\nolimits_{i=1}^{N} (J(\boldsymbol{Y}_i,\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} - (I(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} \right| \leq \right.\right.$$
$$\left.\left. \sup_{\boldsymbol{\theta}\in\Theta_{N_1}} \left| N^{-1} \sum\nolimits_{i=1}^{N} (J(\boldsymbol{Y}_i,\boldsymbol{\theta}))_{rs} - (I(\boldsymbol{\theta}))_{rs} \right| \right\} \right] = 1. \tag{3.26}$$

Now $\Theta_{N_1} \subseteq \Theta^{\circ}$ implies conditions (a)-(d) of Jennrich (1969, Theorem 2) apply to Θ_{N_1}. Thus (3.24) and (3.26) imply

$$P\left[\lim_{N\to\infty} \left\{ \left| N^{-1} \sum\nolimits_{i=1}^{N} (J(\boldsymbol{Y}_i,\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} - (I(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} \right| = 0 \right\} \right] = 1. \tag{3.27}$$

Since $(I(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} \xrightarrow{a.s.} I(\boldsymbol{\theta}_0)$ then (3.27) implies that $N^{-1}(J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} \xrightarrow{a.s.} (I(\boldsymbol{\theta}_0))_{rs}$. Note that (3.26) holds however small Θ_{N_1} is, and so the result in (3.27) is independent of the choice of N_1 other than N_1 must be chosen such that $\Theta_{N_1} \subseteq \Theta^{\circ}$. This result holds for all $r, s = 1, ..., k$, and so in terms of matrices we have $N^{-1}J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})) \xrightarrow{a.s.} I(\boldsymbol{\theta}_0)$.

A similar analysis applies to $S_N(\boldsymbol{\theta})$ where we use the fact that the average of the summands in $(S_N(\boldsymbol{\theta}))_{rs} = \sum_{i=1}^{N}(\boldsymbol{D}_{\boldsymbol{\theta}}(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta}))^{\intercal}\boldsymbol{D}_{\boldsymbol{\theta}}(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta})))_{rs}$ converge almost surely to $(I(\boldsymbol{\theta}))_{rs}$ using the SLLN and (3.15). Thus $N^{-1}(S_N(\boldsymbol{\theta}))_{rs} \xrightarrow{a.s.} (I(\boldsymbol{\theta}))_{rs}$ for all $r, s = 1, ..., k$, and so $N^{-1}S_N(\boldsymbol{\theta}) \xrightarrow{a.s.} I(\boldsymbol{\theta})$. Again we cannot use the Continuous

Mapping Theorem (Van Der Vaart, 1998, p 7) to conclude $N^{-1}S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})) \xrightarrow{a.s.} I(\boldsymbol{\theta}_0)$ but rather we apply conditions (a)-(d) of Jennrich (1969, Theorem 2) to $S(\tilde{\boldsymbol{Y}}, \boldsymbol{\theta}) := S(\boldsymbol{\theta})$ to obtain a strong ULLN as before.

An alternative approach to the ULLN is as follows, and concerns weak rather than strong convergence. We know from the WLLN for any $\epsilon > 0$ that

$$\lim_{N\to\infty} P\left[\left|N^{-1}(S_N(\boldsymbol{\theta}_0))_{rs} - (I(\boldsymbol{\theta}_0))_{rs}\right| < \epsilon\right] = 1. \tag{3.28}$$

If we can then show that

$$\lim_{N\to\infty} P\left[\left|N^{-1}(S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} - N^{-1}(S_N(\boldsymbol{\theta}_0))_{rs}\right| < \epsilon\right] = 1, \tag{3.29}$$

i.e. if $N^{-1}(S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} \xrightarrow{P} \lim_{N\to\infty}(N^{-1}(S_N(\boldsymbol{\theta}_0))_{rs})$, then

$$\begin{aligned}
&\left|(N^{-1}(S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} - I(\boldsymbol{\theta}_0)\right| = \\
&\left|N^{-1}(S_N(\boldsymbol{\theta}_0))_{rs} - (I(\boldsymbol{\theta}_0))_{rs} + N^{-1}(S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} - N^{-1}(S_N(\boldsymbol{\theta}_0))_{rs}\right| \leq \\
&\left|N^{-1}(S_N(\boldsymbol{\theta}_0))_{rs} - (I(\boldsymbol{\theta}_0))_{rs}\right| + \left|N^{-1}(S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} - N^{-1}(S_N(\boldsymbol{\theta}_0))_{rs}\right|,
\end{aligned} \tag{3.30}$$

implies

$$\lim_{N\to\infty} P\left[\left|(N^{-1}(S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} - I(\boldsymbol{\theta}_0)\right| < \epsilon\right] = 1, \tag{3.31}$$

which gives the desired result that $N^{-1}(S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))_{rs} \xrightarrow{P} (I(\boldsymbol{\theta}_0))_{rs})$. The key result in (3.29) is obtained if the collection of random variables $\mathscr{G} := \{N^{-1}(S_N(\boldsymbol{\theta}))_{rs} : \boldsymbol{\theta} \in \Theta^\circ\}_{n\in\mathbb{N}}$ is stochastically equicontinuous at $\boldsymbol{\theta}_0 \in \Theta^\circ$.

To describe this derivation, in what follows we use and expand the discussion in Jordan (2007) about stochastic equicontinuity. A collection of stochastic processes $\mathscr{Z} := \{Z_n(t) : t \in \mathscr{T}\}_{N\in\mathbb{N}}$ is defined to be stochastic equicontinuous at $t_0 \in \mathscr{T}$ if for all $\eta > 0$, and for all $\epsilon > 0$, there exists a neighborhood $U(\epsilon, \eta)$ of t_0 such that

$$\limsup_n P\left[\sup_{t\in U(\epsilon,\eta)} |Z_n(t) - Z_n(t_0)| > \eta\right] < \epsilon. \tag{3.32}$$

The first thing to note is that the limsup in (3.32) relates to real numbers, in particular numbers in the range $[0, 1]$, and so a simple interpretation is possible. For a sequence of real numbers $\{x_n\}$ we have an alternative definition of the limsup as $\limsup_n x_n = \lim_{n\to\infty} \sup_{m\geq n} x_m$. If $\limsup_n x_n = b$ say, then for every $\epsilon > 0$ there exists an $N > 0$ such that $x_n < b + \epsilon$ for all $n > N$. Thus any number greater than the limsup of the sequence $\{x_n\}$ is an eventual upper bound for the sequence - i.e. only a finite number of points of the sequence are greater than $b + \epsilon$. In particular $b = 0$ implies $x_n \to 0$ as $n \to \infty$ for a sequence of positive reals $\{|x_n|\}$. Thus if $\{A_n\}$ are sets in some σ-field $\mathscr{F}$, and $(\Omega, \mathscr{F}, P)$ is a probability space, then $\limsup_n P(A_n) = 0$ implies $P(A_n) \to 0$ as $n \to \infty$.

Now if we assume $\mathscr{Z}$ is stochastic equicontinuous at t_0 then from (3.32) by letting $\eta \to 0$, and $\epsilon \to 0$ we see that $\limsup_n P[\sup_{t\in U_0}|Z_n(t) - Z_n(t_0)| > 0] = 0$, where U_0 is the neighborhood of t_0 that exists in the limit as $\eta \to 0$ and $\epsilon \to 0$. This implies $P[\sup_{t\in U_0}|Z_n(t) - Z_n(t_0)| > 0] \to 0$ as $n \to \infty$. Suppose now that $t_n \xrightarrow{P} t_0$. Then for any $\eta > 0$, $\epsilon > 0$, we have $\limsup_n P(t_n \notin U(\epsilon, \eta)) < \epsilon$, and so $\limsup_n P(t_n \notin U_0) = 0$ by again letting $\eta \to 0$, and $\epsilon \to 0$. Thus $P(t_n \notin U_0) \to 0$ as $n \to \infty$. Now from our assumptions we have that for any $\epsilon > 0$, $\eta > 0$

$$|Z_n(t_n) - Z_n(t_0)| > \eta \implies (t_n \notin U(\epsilon, \eta)) \text{ OR } \left(\sup_{t\in U(\epsilon,\eta))} |Z_n(t) - Z_n(t_0)| > \eta\right), \tag{3.33}$$

and so by letting $\eta \to 0$, $\epsilon \to 0$, for any $\epsilon'/2$ there exists an N such that for all $n > N$

$$P[|Z_n(t_n) - Z_n(t_0)| > 0] \leq P(t_n \notin U_0) + P\left[\sup_{t\in U_0}|Z_n(t) - Z_n(t_0)| > 0\right] < \frac{\epsilon'}{2} + \frac{\epsilon'}{2} = \epsilon'. \tag{3.34}$$

We conclude that $Z_n(\tau_n) \xrightarrow{P} Z_n(\tau_0)$ as $n \to \infty$. Applying this result to the collection of random variables $\mathscr{G}$ which we assume to be stochastically equicontinuous at $\boldsymbol{\theta}_0 \in \Theta^\circ$, and with $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$ in place of t_n, then gives the result in (3.29). We can similarly derive the analogue of (3.29) for $J_N(\boldsymbol{\theta})$ as well.

An interesting question of course is what conditions do we need to impose on $\mathscr{G}$ in order for it to be stochastically equicontinuous at $\boldsymbol{\theta}_0$. Reverting back for one moment to the collection of random variables $\mathscr{Z}$, Theorem 3 of (Li) gives a condition that is sufficient for stochastic uniform equicontinuity, which will imply stochastic equicontinuity at any point in the set $\mathscr{T}$ over which $\mathscr{Z}$ is stochastically uniformly equicontinuous. We write $Z_n(\omega, t)$ for $\omega \in \Omega$ to denote the dependence of Z_n on some underlying probability space $(\Omega, \mathscr{F}, P)$, and let $E \in \mathscr{F}$ be a zero-probability event, that is $P(E) = 0$. Then $\mathscr{Z}$ is stochastically uniformly equicontinuous if there exists an $N \in \mathbb{N}$ such that

$$\big\{\omega \in \Omega : \big|Z_n(\omega, t) - Z_n(\omega, t')\big| \leq B_n h(d(t, t'))\big\}^c \subseteq E, \tag{3.35}$$

holds for all $t, t' \in \mathscr{T}$, and for all $n > N$, where h is a non-random function, $h(x) \downarrow 0$ as $x \downarrow 0$, $B_n = O_p(1)$, and $d(\cdot, \cdot)$ is just the normal Euclidean distance. If we let $M > 0$ denote the finite real number that bounds in probability the random sequence $\{B_n\}$, then obviously (3.35) holds trivially as M gets very large. Conversely we can think of the Z_n as (almost surely) being more continuous on $\mathscr{T}$ the smaller M becomes. In this sense for any $n > N$ (3.35) can be thought of as a stochastic version of a Lipschitz continuity condition for deterministic functions, which is a stronger form of continuity even than uniform continuity. This is most clear when $h(x) = x$ since the condition in (3.35) becomes $|Z_n(\omega, t) - Z_n(\omega, t')| \leq B_n d(t, t')$ which for any $n > N$ is precisely a Lipschitz continuity condition with a stochastically bounded coefficient B_n.

Since for any single function Lipschitz continuity is quite a strong form of continuity, it is likely that (3.35) will be too strong a condition to be satisfied by $\mathscr{G}$. Indeed (Li) states that although the advantage of (3.35) for forms of $Z_n(t)$ such as $Z_n(t) = n^{-1} \sum_{i=1}^{n} z(X_i, t)$ (which we are interested in) is that the random variables $\{X_i\}$ do not need to be independent and/or identically distributed, the condition on $Z_n(t)$ is often transferred to $z(X_i, t)$ which is often too strong. Of course weaker conditions may obtained if we search for sufficient conditions for $\mathscr{G}$ to be stochastically equicontinuous at $\boldsymbol{\theta}_0$ rather than the uniformly so.

A third type of matrix which we shall call the sandwich estimator, is

$$SW_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})) = J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))\left(S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))\right)^{-1} J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})). \tag{3.36}$$

Noting

$$N^{-1}SW_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})) = \left(N^{-1}J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))\right)\left(N(S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{-1}\right)\left(N^{-1}J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))\right), \tag{3.37}$$

and that $N(S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{-1} \xrightarrow{as} I(\boldsymbol{\theta}_0)^{-1}$ by the Continuous Mapping Theorem (Van Der Vaart, 1998, p 7), we see that $N^{-1}SW_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})) \xrightarrow{as} I(\boldsymbol{\theta}_0)I(\boldsymbol{\theta}_0)^{-1}I(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0)$ using Slutsky's theorem (Van Der Vaart, 1998, p 11).

In summary the three random matrices $N^{-1}J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, $N^{-1}S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, and $N^{-1}SW_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ are all consistent estimators of $I(\boldsymbol{\theta}_0)$. Because these three estimators should be close to $N^{-1}I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ for large enough N, then we will also refer to these estimators as approximations to the sample information matrix. The derivations of these consistency results all rely on the two identities (3.15) and (3.17), both of which in turn rely on the assumption for $\boldsymbol{\theta}_0 \in \Theta$ that $f(\cdot|\boldsymbol{\theta}_0)$ is the true density function for $\tilde{\boldsymbol{Y}}$. However if this density function has been misspecified and the true density function is $g(\cdot|\boldsymbol{\theta}_0)$ say, then $N^{-1}(SW_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{-1}$ is still a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$. For this reason $(SW_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{-1}$ is sometimes called the "Robust" covariance matrix. This result is given in White (1982, Theorem 3.2) and rests upon very similar assumptions that are outlined in Section 3.1. Because in this work we are not focusing on misspecified models we do not describe the details of this result here.

It was described in Section 2.2 that the EM algorithm is easily the most popular method of estimating the parameters in a mixture model, partly to avoid taking derivatives of the ordinary log-likelihood function (in particular second derivatives). Thus for the most part many researchers will be eager to avoid using any of the three consistent estimators of $I(\boldsymbol{\theta}_0)$ that we have described here. For *iid* samples however there is a way to compute $S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ using the $\boldsymbol{D}_{\boldsymbol{\theta}}(\log f(\boldsymbol{C}_i|\boldsymbol{\theta}))^\intercal$ (the score vectors of the complete data log-likelihood function $L^c(\boldsymbol{\theta}|\boldsymbol{C}_i)$ for the i^{th} unit), which are are typically easier to derive than using the ordinary log-likelihood function. The result we require comes from Louis (1982) and states that for an independent but not necessarily identically distributed sample we have

$$\begin{aligned}\boldsymbol{D_\theta}(\log f(\boldsymbol{y}|\boldsymbol{\theta}))^\intercal &= \sum_{i=1}^{N} \boldsymbol{D_\theta}(\log f(\boldsymbol{y}_i|\boldsymbol{\theta}))^\intercal \\ &= \sum_{i=1}^{N} \boldsymbol{E_\theta}\left[\boldsymbol{D_\theta}(\log f(\boldsymbol{C}_i|\boldsymbol{\theta}))^\intercal|\boldsymbol{y}_i\right]. \end{aligned} \tag{3.38}$$

If the sample is *iid*, and considering the incomplete data vector $\boldsymbol{y}_i$ which we condition on as being random, then $\boldsymbol{D_\theta}(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta}))^\intercal = \boldsymbol{E_\theta}[\boldsymbol{D_\theta}(\log f(\boldsymbol{C}_1|\boldsymbol{\theta}))^\intercal|\boldsymbol{Y}_1]$ for all $i \in I_N$, and so (3.38) implies

$$\begin{aligned} S_N(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \boldsymbol{D_\theta}(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta}))^\intercal \boldsymbol{D_\theta}(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta})) \\ &= N\boldsymbol{E_\theta}\left[\boldsymbol{D_\theta}(\log f(\boldsymbol{C}_1|\boldsymbol{\theta}))^\intercal|\boldsymbol{Y}_1\right]\boldsymbol{E_\theta}\left[\boldsymbol{D_\theta}(\log f(\boldsymbol{C}_1|\boldsymbol{\theta}))|\boldsymbol{Y}_1\right]. \end{aligned} \tag{3.39}$$

Thus the consistent estimator $N^{-1}S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ of $I(\boldsymbol{\theta}_0)$ can be obtained using just the conditional expected values of the score vectors of the complete data log-likelihood function from one unit.

Typically a consistent estimator of the information matrix $I(\boldsymbol{\theta}_0)$ is used with an asymptotic result such as

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0) \xrightarrow{D} N_k(\boldsymbol{0}, (I(\boldsymbol{\theta}_0))^{-1}), \tag{3.40}$$

for the purpose of performing asymptotic inference on the estimators of the model parameters in $\boldsymbol{\theta}_0$. For example using a consistent estimator of $I(\boldsymbol{\theta}_0)$, say $N^{-1}I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, we have that

$$\begin{aligned} &\left(N^{-1}I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))\right)^{1/2}\sqrt{N}(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0) = \\ &\qquad (I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{1/2}(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0) \xrightarrow{D} (I(\boldsymbol{\theta}_0))^{1/2}\boldsymbol{Z}, \end{aligned} \tag{3.41}$$

where $\boldsymbol{Z} \sim N_k(\boldsymbol{0}, (I(\boldsymbol{\theta}_0))^{-1})$.

Now since $I(\boldsymbol{\theta}_0)$ is a symmetric matrix then it is possible to construct a set $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\}$ of n orthonormal eigenvectors corresponding to the set of k eigenvalues $\{\lambda_1, ..., \lambda_k\}$ of $I(\boldsymbol{\theta}_0)$. We can then construct an orthogonal matrix $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_k)$ to obtain the spectral decomposition $I(\boldsymbol{\theta}_0) = \boldsymbol{X}\Lambda\boldsymbol{X}^\intercal$, where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_k)$. Since $I(\boldsymbol{\theta}_0)$ is positive definite then $\lambda_j > 0$ for all $j = 1, ..., k$, and so the square root matrix $\Lambda^{1/2}$ is real and positive-definite too. This means we can find a nonnegative definite $k \times k$ matrix $(I(\boldsymbol{\theta}_0))^{1/2} = \boldsymbol{X}\Lambda^{1/2}\boldsymbol{X}^\intercal$ such that $(I(\boldsymbol{\theta}_0))^{1/2}(I(\boldsymbol{\theta}_0))^{1/2} = I(\boldsymbol{\theta}_0)$, that is $(I(\boldsymbol{\theta}_0))^{1/2}$ is a square root matrix of $I(\boldsymbol{\theta}_0)$. Thus the covariance matrix of the random vector in the left-hand side of (3.41) in the limit is given by $(I(\boldsymbol{\theta}_0))^{1/2}[(I(\boldsymbol{\theta}_0))^{1/2}(I(\boldsymbol{\theta}_0))^{1/2}]^{-1}(I(\boldsymbol{\theta}_0))^{1/2} = \boldsymbol{I}_k$, and so we have using Slutsky's Theorem (Van Der Vaart, 1998, p 11)

$$(I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{1/2}(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I}_k). \tag{3.42}$$

Now this result says in the limit as $N \to \infty$ that the random sequence of vectors $\{(I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{1/2}(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0)\}$ converges to a random variable with distribution $N_k(\boldsymbol{0}, \boldsymbol{I}_k)$. It does not say that $(I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{1/2}(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0)$ has this distribution for any finite N however large. But perhaps for N large enough the distribution of $(I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{1/2}(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0)$ might be approximately $N_k(\boldsymbol{0}, \boldsymbol{I}_k)$. So if we assume this holds, and if we ignore the fact that $(I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{1/2}$ is random we get

$$\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) \approx N_k(\boldsymbol{\theta}_0, (I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{-1}), \tag{3.43}$$

where $\approx$ means "distributed approximately". Of course (3.43) can be derived in the same way with $J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, $S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, and $SW_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ replacing $I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ since they are all consistent estimators of $I(\boldsymbol{\theta}_0)$.

3.2.2 Quantifying confidence interval performance - true standard errors

We now describe the simulation study by Boldea and Magnus (2009) to which we referred at the beginning of Section 3.2. In this study Boldea and Magnus used $J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, $S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, and $SW_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ which are the approximations to sample information matrix $I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ which we described in subsection (3.2.1), and three versions of a bootstrap procedure, in order to obtain standard errors of the parameter estimates in a 2-component Gaussian mixture using an iid sample. The parameters to be estimated were the elements of the mean vectors and covariance matrices of the component distribution functions in the mixture.

The approach taken by Boldea and Magnus to evaluate the performance of these six different methods of obtaining parameter estimate standard errors was to compare these standard errors to estimates of the true parameter estimate standard errors. For any given model parameter this method worked as follows: an estimate of the true parameter estimate standard error was calculated by obtaining a very large number (50,000) of parameter estimates and defining the true standard error to be equal to the standard deviation of this sample of estimates. A further 10,000 parameter estimates and their estimated standard errors were then obtained (different from the previous 50,000) using the above approximations to the sample information matrix, and the three bootstrap procedures. These samples of estimated standard errors were then used to quantify the performance of these different methods of standard error calculation. Specifically the MSE of these samples were calculated using the "true" standard error calculated from the 50,000 replications.

The three bootstrap methods used by Boldea and Magnus were the parametric, non-parametric, and a "weighted" bootstrap method. Sample sizes of $N = 100$ and $N = 500$ were used, and the component distribution means were specified to be well separated enough such the parameter estimates were unbiased, since the intention was to focus solely on the performance of the estimates of the standard errors. The root mean square error (RMSE) was used as a measure of closeness of the estimated standard errors to the true ones. Boldea and Magnus investigated four scenarios including one where the model was correctly specified, that is the model fitted to the simulated data sets was the same as the data generating process. We only describe these results since this is the scenario we are interested in in this thesis.

When the model is correctly specified, and using the summary of the results for $N = 500$ presented in table 4, the Hessian estimator produced a lower RMSE than the score estimator, which in turn produced a slightly lower RMSE than the sandwich estimator. For the bootstrap procedures the parametric method was best, producing lower RSME values than the other two methods which produced the same RMSE values as each other. In terms of comparing the information matrix based estimators with the bootstrap ones, the Hessian estimator was better than the parametric bootstrap estimator, which in turn was slightly better than the score estimator, and clearly better than the sandwich estimator. The score and sandwich estimators were better than the non-parametric and weighted bootstrap estimators. For $N = 100$ almost the same results are obtained with the exception that the RMSE for the parametric bootstrap is very slightly lower than that produced by the Hessian estimator.

Thus for these simple (two components only) well separated Gaussian mixtures, for $N = 500$ the Hessian based estimator proved the superior method for estimating the true parameter standard errors for these small samples sizes, although not by much. For $N = 100$ the parametric bootstrap was nominally better but the difference compared to the Hessian estimator was so small that we can say these two methods were the same. For both sample sizes the Score and sandwich estimators were better than the non-parametric and the weighted bootstrap methods. Finally all of the RMSE's for all methods were close to zero, which Boldea and Magnus point out is in contrast to the claim made by some authors that very large sample sizes are required for accurate results using the information matrix. Furthermore, and as expected, the RMSE are much smaller for $N = 500$ compared to $N = 100$.

It is worth noting that in general, and for both information matrix and bootstrap methods, Boldea and Magnus state that the contribution of the bias to the RMSE is small compared to the contribution from the variance, for example for the Hessian estimator for $N = 500$, and averaged over all the model parameters, the ratio of the absolute bias to the RMSE is approximately 9%. The corresponding figure is not given for $N = 100$, but presumably it is larger. Furthermore Boldea and Magnus note that the bias tends to be negative for all methods of standard error estimation. This suggests that in small sample sizes confidence intervals constructed using these standard error estimates may be slightly shorter than they should be, leading to a false impression of the precision with which we can estimate the model parameters. Since the bias is small for $N = 500$ then this may not be too much of a problem, however for smaller sample sizes this bias may well be much larger. The fact that the majority of the RMSE for $N = 500$ comes from the variance of the estimators is perhaps not surprising for low samples sizes such as this.

3.3 Inference for the LMM

The following is a very brief summary of the relevant asymptotic theory for parameter inference in a LMM where the within-unit errors are assumed to follow a stationary AR(r) process. There are other similar results for LMMs that use a simple within-unit covariance structure, however the result described here will apply to those models by setting $\boldsymbol{\phi} = \mathbf{0}$. We will describe the result given by Wang and Fan (2009) who actually focus on LMMs for multiple response variables, that is where $\boldsymbol{Y}_i$ is a $n_i \times h$ matrix - Wang and Fan call this multivariate longitudinal data. The results are applicable to a LMM by setting $h = 1$. For brevity we will use the same notation that we introduced in chapter 2 by setting $G = 1$.

Now since the sample $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$ for a LMM are independent but generally not identically distributed, it is not necessarily the case that we have $\boldsymbol{E}_{\boldsymbol{\theta}}[\boldsymbol{D}_{\boldsymbol{\theta}}\left(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta})\right)^\intercal \boldsymbol{D}_{\boldsymbol{\theta}}\left(\log f(\boldsymbol{Y}_i|\boldsymbol{\theta})\right)] = \boldsymbol{A}$ for all $i \in I_N$, where $\boldsymbol{A}$ is a $n_\theta \times n_\theta$ finite matrix. Thus for these non-*iid* samples it is not as natural to assume that $N^{-1}I_N(\boldsymbol{\theta})$ converges to an asymptotic information matrix $I(\boldsymbol{\theta})$, nor to impose conditions such that this occurs, although of course such assumptions can nonetheless still be used. Two approaches can be used for these non-*iid* samples. One approach is called the deterministic scheme or fixed design, whilst another approach is called the stochastic scheme or random design (Demidenko, 2004). The fixed scheme assumes the $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, and n_i are fixed, and attempts to impose conditions on these fixed data that ensures $N^{-1}I_N(\boldsymbol{\theta})$ converges to a matrix $I(\boldsymbol{\theta})$. The easiest assumption in this respect is to simply assume this holds, and this is the approach taken by Wang and Fan. In the stochastic scheme it is assumed that the $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, and n_i are random variables with distributions that may depend on unknown parameters (but not $\boldsymbol{\theta}$), and that the triples $\{\boldsymbol{Y}_i, \boldsymbol{X}_i, n_i\}$ are *iid*, which in turn ensures the $\{\boldsymbol{Y}_i\}$ are *iid* - see Demidenko (2004) for an outline of this approach.

Wang and Fan present present two asymptotic results, one for the fixed effects parameter $\boldsymbol{\beta}$, and one for the vector of covariance parameters $\boldsymbol{\zeta} = (\sigma^2, \boldsymbol{\psi}^\intercal, \boldsymbol{\phi}^\intercal)^\intercal$. For this partitioning of $\boldsymbol{\theta}$ we will denote the sample information matrices as $I_N(\boldsymbol{\beta})$ and $I_N(\boldsymbol{\zeta})$ for $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ respectively, where the definition of the sample information matrix can be found in subsection (3.2.1). For this partitioning of $\boldsymbol{\theta}$, and as we described above, Wang and Fan assume that $N^{-1}I_N(\boldsymbol{\beta}) \to I(\boldsymbol{\beta})$, and $N^{-1}I_N(\boldsymbol{\zeta}) \to I(\boldsymbol{\zeta})$ as $N \to \infty$

where $I(\boldsymbol{\beta})$, and $I(\boldsymbol{\zeta})$ are finite positive-definite matrices. The other assumptions they use are: The n_i are bounded above; the parameter spaces for $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ are compact; $\boldsymbol{\theta}_0 \in \Theta^\circ$; and $\max_i\{\boldsymbol{a}^\intercal \boldsymbol{X}_i^\intercal \boldsymbol{X}_i \boldsymbol{a}/\boldsymbol{a}^\intercal \sum_{i=1}^N \boldsymbol{X}_i^\intercal \boldsymbol{X}_i \boldsymbol{a}\} \to 0$ for all non-zero $\boldsymbol{a} \in \mathbb{R}^p$, which ensures that $\sum_{i=1}^N \boldsymbol{X}_i \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{X}_i$ is of order $O(N)$.

Given these assumptions and some unstated regularity conditions, Wang and Fan state that $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) \xrightarrow{P} \boldsymbol{\theta}_0$, $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, (I(\boldsymbol{\beta}_0))^{-1})$, and $\sqrt{N}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \xrightarrow{D} N(\mathbf{0}, (I(\boldsymbol{\zeta}_0))^{-1})$. Now for the LMM, and from (C.28), and (C.2), we see that $I_N(\boldsymbol{\theta})$ has a block structure with $I_N(\boldsymbol{\beta})$ in the top-left block, $I_N(\boldsymbol{\zeta})$ in the bottom-right block, and zeros elsewhere. Thus from Schott (2005, Theorem 7.1, pp256) we see that $I_N(\hat{\boldsymbol{\theta}})^{-1}$ has $(I_N(\hat{\boldsymbol{\beta}}))^{-1}$ in the top-left block, $(I_N(\hat{\boldsymbol{\zeta}}))^{-1}$ in the bottom-right block, and zeros elsewhere. This means that $(I_N(\hat{\boldsymbol{\theta}}))^{-1}$ can be used to simultaneously calculate both inverse information matrices $(I(\boldsymbol{\beta}_0))^{-1})$, and $(I(\boldsymbol{\zeta}_0))^{-1})$.

3.4 Mixture model naive inference

In this section we propose two methods of statistical inference which we shall call ”naive”, since they are based on unproven theory, but nonetheless possess a certain level of credibility that makes them worthy of further study. Both methods are concerned with constructing approximate asymptotic confidence intervals around the parameter estimates in MLMMs. The methods described in subsection (3.4.2) focus on approximating the information matrix of a MLMM, whilst the methods in subsection (3.4.3) focus on using G separate LMM information matrices to approximate some of the information in the full mixture model information matrix. Subsection (3.4.1) is concerned with quantifying how well separated the parameters in a MLMM are, which essentially quantifies how difficult it is to tell the model parameters apart between the components. The motivation for this subsection is that this separation will determine how well both methods of inference proposed here will work.

3.4.1 Quantifying component separation

This subsection is concerned with proposing an index to quantify how well separated the components of a mixture distribution function are, specifically where the distribution function is induced by an MLMM. The word ”separation” implies, as it is intended to, that the notion of the separatedness of the components of the mixture distribution function relates to some sort of distance-based measure on the component distribution functions. This is because it is clear that if the component distributions are too close to each other then it is likely that classifying units to components will prove very difficult, and thus it is likely this classification will be carried out with high error rates.

If the primary purpose of fitting a mixture model is the classification itself, then the propensity of a mixture model to produce a poor classification is obviously not a desirable property. However if the primary interest is in the parameter estimates then high classification error rates are also not good. This is because it is almost certain that the quality of the parameter estimates will be heavily influenced by the quality of the classification of units to components (and vice-versa). This is because the accuracy of the estimated posterior probabilities will be dependent on the accuracy of the classification, and in turn the parameter estimates all depend on the estimated posterior probabilities. Because the focus of this work is not on classification we make this statement without further explanation or justification, nor do we attempt to directly investigate this in any of the simulations we perform in Section 5.1, but instead we do this indirectly by investigating the relationship between component separation and classification error rates.

We note that since the intention of the separation index is essentially prognostic in terms of quantifying how well the mixture model works, then any separation index should be based on the true mean vectors and covariance matrices, or the true regression and covariance parameters. In this way the index will be independent of the estimation procedure itself. In contrast since estimation and thus inference on the estimators is a function of the data, it is probably desirable not to have an index that is based on just the true parameters alone, since this would not be very informative. Instead it would be preferable to have the index be a function of the fixed covariate data $\{\boldsymbol{X}_i\}_{i \in I_N}$ and $\{\boldsymbol{Z}_i\}_{i \in I_N}$, since in combination the covariate data as well as the model parameters will determine to a large extent the separatedness of the components.

Regardless of whether the mean vectors and covariance matrices of a mixture model are parameterized by regression and covariance parameters, we feel an obvious choice as a basis for a separation index is some function of both the distance between the true mean vectors, and the magnitude of the entries in the true covariance matrices of the component distribution functions. The intuition behind this choice is that for fixed values of the component covariance matrices as the distance between the mean vectors reduces then the distributions become closer together and hence less well separated. Clearly the task of classifying units to components should get more difficult if this convergence of mean vectors results in significant overlapping of the tails of the distributions. Of course this overlapping is also a function of the component covariance matrices. For example there may be no overlapping at all for even small amounts of separation between the mean vectors if the covariance matrices all have small elements.

Another approach to quantifying component separation is to recognise that quantifying the separatedness of the true parameters (again based upon distance) should also give a good idea of how easy the classification of units to components is likely to be. This may be a much better strategy when the mean vectors and covariance matrices are parameterized by regression parameters and covariance parameters, as they are for MLMMs and MLMs. The reason for this is that on account of the complex relationship between the covariate data $\{\boldsymbol{X}_i\}_{i \in I_N}$, $\{\boldsymbol{Z}\}_{i \in I_N}$, and the regression and covariance parameters, it is conceivable we could have say two mean vectors that are separated in terms of distance, but that some of the regression parameters are not. Furthermore this method has extra intuitive appeal in the sense that when there are only mean vectors and covariance matrices to estimate, then this concept of separatedness of parameters in terms of distance will in essence reduce to the previous interpretation of separatedness of components in terms of distance.

The index we now propose will measure the separation based upon distance between two scalar parameters in a pair of components, where the parameters "correspond" in the two components. Formally, and recalling that n_θ is the total number of parameters in each $\boldsymbol{\theta}_g$, $g \in I_G$, then for any two components $g, g' \in I_G$, and any $s \in \{1, ..., n_\theta\}$, let $(\boldsymbol{\theta}_g)_s$ and $(\boldsymbol{\theta}_{g'})_s$ be called corresponding parameters in the component pair (g, g'). Then let $SI^s(g, g')$ denote the separation index between the s^{th} corresponding pair of parameters in the component pair (g, g'). Later we will also need to aggregate in some way across the n_θ different separation indexes in order to derive an overall measure of separatedness of the component pair (g, g').

If for a given MLMM the true component memberships are known, then each $\boldsymbol{\theta}_g$, $g \in I_G$, can be estimated componentwise as we described in subsection (3.4.3). Now since there is nothing stopping us from constructing the componentwise "confidence intervals" for the true parameters, then by doing so we will immediately obtain a mechanism to determine the separation of the two true parameters that is grounded in statistical theory. Thus we will base $SI^s(g,g')$ on the componentwise "confidence intervals" for $\boldsymbol{\theta}_g$, and $\boldsymbol{\theta}_{g'}$. We will call these intervals the true parameter intervals (TPIs). Specifically

$$TPI_{(1-\alpha)}((\boldsymbol{\theta}_g)_s) = (\boldsymbol{\theta}_g)_s \pm z_{\alpha/2}\sqrt{((I_N(\boldsymbol{\theta}_g))^{-1})_{ss}}, \tag{3.44}$$

where $z_{\alpha/2}$ is the value of a standard normal random variable Z such that $P[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] = 1-\alpha$, for nominal confidence level $\alpha \in [0,1]$. This is simply the $(1-\alpha)*$ 100% approximate confidence interval for $(\hat{\boldsymbol{\theta}}_g(\boldsymbol{Y}))_s$ derived from (3.48) but with $(\boldsymbol{\theta}_g)_s$ replacing $(\hat{\boldsymbol{\theta}}_g(\boldsymbol{Y}))_s$.

Since the the TPIs in (3.44) are derived from (3.48) then the TPIs are a function of the asymptotic distribution of $\hat{\boldsymbol{\theta}}_g(\boldsymbol{Y})$. As a result the use of the additive term $z_{\alpha/2}\sqrt{((I_N(\boldsymbol{\theta}_g))^{-1})_{ss}}$ to create an interval is based upon sound statistical theory. However an obvious question is how can we interpret the non-randomness of $(\boldsymbol{\theta}_g)_s$ around which we construct the TPI? To answer this question we note that $I_N(\boldsymbol{\theta}_g)$ depends on the sample through the non-random quantities $\{\boldsymbol{X}_i\}_{i\in I_N}$, $\{\boldsymbol{Z}\}_{i\in I_N}$, and thus also the sample sizes N and n_i. Thus since we can think of $\boldsymbol{\theta}_g$ as the most accurate estimate of $\boldsymbol{\theta}_g$ that we can obtain, then one interpretation of the TPI in (3.44) is that the interval represents the most accurate error bound for $(\hat{\boldsymbol{\theta}}_g(\boldsymbol{Y}))_s$ we can obtain given the fixed covariate data $\{\boldsymbol{X}_i\}_{i\in I_N}$, $\{\boldsymbol{Z}\}_{i\in I_N}$, and sample sizes $\{n_i\}_{i\in I_N}$, and N.

We now describe how we will calculate $SI^s(g,g')$. Let $I_1 = (a_1, b_1)$ and $I_2 = (a_2, b_2)$ be the TPI for $(\boldsymbol{\theta}_g)_s$ and $(\boldsymbol{\theta}_{g'})_s$ respectively. We will base our definition of $SI^s(g,g')$ on the following scenarios relating to how I_1 and I_2 can overlap: nested (NEST) where $a_1 \leq a_2 < b_2 \leq b_1$; overlap (OVLP) where $a_1 < a_2 \leq b_1 < b_2$; separate (SEP) where $b_1 < a_2$. For OVLP and SEP these scenarios define the situation when I_2 is to the right (either totally or in part) of I_1, whilst for NEST they define $I_2 \subseteq I_1$. For I_1 to the right of I_2, or $I_1 \subseteq I_2$ simply reverse the roles of the two intervals. Then we define the separation index $SI^s(g,g')$ as

$$SI^s(g,g') = \begin{cases} \frac{a_2-b_2}{b_1-a_1}, & \text{if } a_1 \leq a_2 < b_2 \leq b_1 \text{ (NEST)} \\ \\ \frac{a_2-b_1}{b_2-a_1}, & \text{if } a_1 < a_2 \leq b_1 < b_2 \text{ (OVLP)} \\ \\ \frac{a_2-b_1}{(b_2-a_2)+(b_1-a_1)}, & \text{if } b_1 < a_2 \text{ (SEP)}, \end{cases} \tag{3.45}$$

where it can be verified that the ranges $SI^s(g,g')$ take on are

$$SI^s(g,g') \in \begin{cases} [-1,0), & \text{for } a_1 \leq a_2 < b_2 \leq b_1 \text{ (NEST)} \\ (-1,0], & \text{for } a_1 < a_2 \leq b_1 < b_2 \text{ (OVLP)} \\ (0,\infty), & \text{for } b_1 < a_2 \text{ (SEP)}. \end{cases} \tag{3.46}$$

We will use $\alpha = 0.975$ so that the TPIs are 95% confidence intervals, and so match the width of the confidence intervals we will use when we fit MLMMs in our simulations. It may be however that some other value for α for the TPIs gives better SIs in some respect, but we did not investiagte this subject, nor did we investigate how sensitive the SIs are with respect to changes in α. Some of the desirable properties satisfied by $SI^s(g,g')$ are:

1. For fixed interval lengths $||I_1||$ and $||I_2||$, and for OVLP and SEP, $SI^s(g,g')$ is an increasing function of $a_2 - b_1$.

2. $SI^s(g,g')$ attains a minimum value of -1 if and only if $I_1 = I_2$

3. $SI^s(g,g') = 0$ if and only if $a_2 = b_1$.

4. $SI^s(g,g') < 0$ for $a_2 < b_1$.

5. $SI^s(g,g') > 0$ for $b_1 < a_2$.

Property 1 means for OVLP and SEP that $SI^s(g,g')$ is determined by the distance between the left hand end point a_2 of I_2, and the right hand end point b_1 of I_1. Thus as I_2 moves further away from I_1 (to the right) then $SI^s(g,g')$ increases, which obviously has intuitive appeal. Property 2 implies we are assuming maximum unseparatedness of two intervals occurs only when the two intervals are the same (this occurs in the NEST scenario). In turn this implies we are assuming that if I_1 and I_2 are such that $I_2 \subseteq I_1$ holds strictly, and regardless of how small the difference in the lengths of the interval are, then the two intervals are more separated than when $I_1 = I_2$. Furthermore property 4 means that the range of values $SI^s(g,g')$ takes on in the NEST and OVLP scenarios are almost the same. This implies we are assuming that two intervals that overlap by a certain amount should not be viewed as being more or less separated than two intervals that overlap by the same amount but in the OVLP scenario. Property 5 implies we view the separation of the two intervals as tending to infinity as the distance I_2 is to the right of I_1 tends to infinity.

Finally we need some way of aggregating the set of separation indices of all of the model parameters for the component pair (g,g'). In this respect we will define

$$SI(g,g') = \max\{SI^1(g,g'),...,SI^{n_\theta}(g,g')\}, \tag{3.47}$$

to represent a measure of the overall separatedness of components g and g'. The reason we use the maximum of the separation indexes (rather than say the average) is that it is possible important performance metrics such as estimator bias and variance of a MLMM are influenced by the separatedness of even one parameter, if that separatedness is large enough.

3.4.2 Approximating the information matrix for MLMMs

It was described in Section 3.1 that for *iid* mixture densities there exists a unique strongly consistent sequence of solutions $\{\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})\}$ to the likelihood equations, that with probability 1 as $N \to \infty$ these solutions are a local maximum of $L_N(\boldsymbol{\theta})$, and that $\sqrt{N}(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) - \boldsymbol{\theta}_0)$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix $I(\boldsymbol{\theta}_0)^{-1}$. Given this result, and if $I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ is a consistent estimator of $I(\boldsymbol{\theta}_0)$, we then get

$$\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}) \sim N_{n_\theta}(\boldsymbol{\theta}_0, (I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})))^{-1}), \tag{3.48}$$

where we recall that $n_{\boldsymbol{\theta}}$ is the number of parameters in the 1-component MLMM and $n_{\Theta} = (G * n_{\boldsymbol{\theta}}) + G$ is the number of parameters in the MLMM. We are currently not aware of such a result as 3.48 for non-*iid* samples, however notwithstanding this some authors apply this result to MLMMs with non-*idd* samples as if such a result has been proven. Furthermore they make no mention of this, and in doing so give a false impression that this problem has been solved.

For example Xu and Hedeker (2002), and Grün and Hornik (2011) both study a MLMM for a non-iid sample with a simple within-unit error covariance structure, and an unstructured random effects covariance structure. Xu and Hedeker use a Fisher Scoring procedure to estimate the parameters of the model, but where the complete data information matrix $I_N^c(\boldsymbol{\theta})$ is used instead of $I_N(\boldsymbol{\theta})$. Standard errors for the parameter estimates are then obtained by inverting $I_N^c(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ however no justification for doing this is given. Similarly they further make the claim that $I_N^c(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ is almost equal to $I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ without any justification. In contrast Grün and Hornik obtain standard errors for the parameter estimates by inverting the matrix $S_N(\boldsymbol{\theta})$, which for *iid* samples we know can be a consistent estimator of $I(\boldsymbol{\theta}_0)$ subject to certain assumptions (see subsection 3.2.1). They calculate $S_N(\boldsymbol{\theta})$ with the matrix obtained from the outer product of the conditional expected value of the score vector given in equation 3.39 which is a result which again relies on an *iid* sample. No discussion is given justifying their methods.

By using estimators of $I_N(\boldsymbol{\theta})$, or of $S_N(\boldsymbol{\theta})$, both of which for *iid* samples have been proven to be consistent estimators of $I(\boldsymbol{\theta}_0)$, it is clear that both Xu and Hedeker (2002), and Grün and Hornik (2011) anticipate a result such as (3.48) for the non-*iid* samples associated with their MLMMs. In this section we call this approach to statistical inference as "naive" in the sense that unproven results are being superficially applied. Now despite the ambiguity of Xu and Hedeker (2002), and Grün and Hornik (2011), their naive approach to inference may be justified in the sense that there probably is good reason to hope that such a result may well hold, not least because from Section 3.3 we see that a similar result holds for the LMM, which in general induces a non-*iid* sample. However the methods employed there may not easily generalise to mixtures, and furthermore identifiability problems caused by the regression and covariance parameters will also need to be addressed. In this respect there does not even exist a valid consistency proof for clusterwise regression models (MLMs with $n_i = 1$ for all $i \in I_N$) since the widely cited result given in Kiefer (1978) has some flaws (see the discussion in Section 3.1.

Using this naive approach to inference we propose to construct approximate asymptotic confidence intervals about the estimates for the parameters in the MLMMs we introduced in chapter 2. Specifically we will use the Redner and Walker (1984) result "naively" for MLMMs, in the sense that such a result has not been proven for these models. From this result the approximate asymptotic distribution of $\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y})$ can be derived and is given by equation 3.48. Thus naive inference involves using the sample information matrix $I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ to construct approximate asymptotic normal confidence intervals about the mixture model parameter estimates, and the intention is to investigate using simulations the performance of these confidence intervals. In order to implement this we need to derive $I_N(\boldsymbol{\theta})$ in explicit form for the models we introduced in Chapter 2, but unfortunately there are computational problems with this. However these have been addressed by Boldea and Magnus (2009) by using alternative matrices which are also consistent estimators of $I(\boldsymbol{\theta}_0)$, and so for large enough N will approximate $I_N(\boldsymbol{\theta})$.

The work of Boldea and Magnus is concerned with using the result in equation 3.48 to perform statistical inference on the parameters from finite mixture densities with an *iid* sample, and thus they are justified in using this result. The problem encountered when calculating $I_N(\boldsymbol{\theta})$ is that the N summands in the log-likelihood $L(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{N} \log(\sum_{j=1}^{G} \boldsymbol{\pi}_j f_{ij}(\boldsymbol{y}_i| \boldsymbol{\lambda}_i^{(j)}, \boldsymbol{\theta}_j))$ are logarithms of sums which cannot be simplified, and so taking derivatives results in fractions. As a result, according to Boldea and Magnus, computing the expectations is typically unfeasible. In this respect we have already mentioned Xu and Hedeker (2002) have derived $I_N^c(\boldsymbol{\theta})$, where one big advantage of using the complete data log-likelihood given in (2.12) is that its mathematical form, by being the sum of logarithms rather than the logarithm of sums, and due to the exponential term in the Normal density function, means we end up with a sum of terms whose expectations can be computed.

The three consistent estimators of $I(\boldsymbol{\theta}_0)$ that Boldea and Magnus use to approximate $I_N(\boldsymbol{\theta})$ are $N^{-1}J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, $N^{-1}S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, and $N^{-1}SW_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, and we described these in subsection 3.2.1. The nice property of these estimators is that no expectations are involved, and Boldea and Magnus have shown it is possible to derive them at least for MLMMs with no regression components or covariance parameters. Thus we will use and adapt the methods of Boldea and Magnus to derive these approximations for the MLMMs we are concerned with in this thesis, and these derivations are detailed in subsection (C.2), and furthermore we are not currently not aware of any such derivations for these classes of MLMMs.

Thus in practice naive inference consists of using equation 3.48 to construct asymptotic confidence intervals about the MLMM parameter estimates, but where we replace $I_N(\hat{\boldsymbol{\theta}})$ with $J_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, $S_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$, or $SW_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$. We note that since our naive approach to inference in this subsection is to anticipate the result in 3.48, then implicit in this is also the assumption that these three information matrix estimators are consistent estimators of $I(\boldsymbol{\theta}_0)$ (and thus approximate $I_N(\hat{\boldsymbol{\theta}}_N(\boldsymbol{Y}))$ for large N). However because the MLMMs we work with in this thesis induce non-*iid* samples, then the assumptions which lead to these three estimators being consistent for $I(\boldsymbol{\theta}_0)$ will not necessarily work for these models, and so even if equation 3.48 does hold for non-*iid* samples, the extent of the success of naive inference will also depend to a great extent on the accuracy of these approximations.

We described in subsection 3.2.2 how Boldea and Magnus compared the parameter estimate standard errors obtained from the three approximations to $I_N(\hat{\boldsymbol{\theta}})$ to the "true" standard errors. We criticised this method primarily because there is no clear relationship between how well one of these methods estimates the true standard errors, and the coverage of the confidence intervals obtained from these estimates. Obviously coverage probabilities do not suffer from the latter problem, but of course they do not themselves tell us anything about the standard errors used in the construction of the intervals. Furthermore they do not tell us about the length of the confidence intervals with which we have obtained this coverage, which will be often be important. For example it is clear that good coverage can be attained even if parameter estimates are very biased if the confidence intervals are long enough. Despite these disadvantages we prefer the more direct quantification of the quality of inference that can be derived from a particular confidence interval method that coverage probabilities permits us to make. But we can overcome these disadvantages by also calculating the means of both the confidence interval lengths and the parameter estimate standard errors. Thus to investigate the quality of inference provided by the three approximations to the sample information matrix that we have described here, we will calculate coverage probabilities, average confidence interval lengths, and averages of the standard errors. These investigations are described in Section 5.1.

3.4.3 Componentwise Inference for MLMMs

The concept of componentwise inference, or more specifically the conditions under which we expect it to produce valid statistical inference about the component density parameters in a MLMM, is linked to the quality of the classification of units to components we can obtain using the estimated posterior probabilities. For $g \in I_G$ the estimated posterior probabilities are given by

$$\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{Y}_i, \hat{\boldsymbol{\theta}}) = \frac{f_g(\boldsymbol{Y}_i|\boldsymbol{\lambda}_i^{(g)}, \hat{\boldsymbol{\theta}}_g)\hat{\boldsymbol{\pi}}_g}{\sum_{k=1}^{G} f_k(\boldsymbol{Y}_i|\boldsymbol{\lambda}_i^{(k)}, \hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\pi}}_k}, \tag{3.49}$$

or $\hat{p}_{ig}$ for short, which we introduced in Section 2.2 within the context of the EM algorithm, although here we are working with the random vector $\boldsymbol{Y}_i$ rather than its realized value $\boldsymbol{y}_i$. The analogous quantity using the true parameter $\boldsymbol{\theta}_0$ rather than the estimator $\hat{\boldsymbol{\theta}}$ we will denote by p_{ig}.

Now if the components induced by the MLMM are not well separated, and if unit i belongs to component $g \in I_G$, then $f_g(\boldsymbol{Y}_i|\boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g) > f_j(\boldsymbol{Y}_i|\boldsymbol{\lambda}_i^{(j)}, \boldsymbol{\theta}_j)$ for all $j \in I_G$, $j \neq g$, but we will probably also have that $f_j(\boldsymbol{Y}_i|\boldsymbol{\lambda}_i^{(j)}, \hat{\boldsymbol{\theta}}_j)$ is substantially greater than zero for all $j \in I_G$. Accordingly the posterior probabilities p_{ij} for all $j = 1, ..., G$ will take values in the whole range $[0, 1]$, and so too will the estimated posterior probabilities $\hat{p}_{ij}$ regardless of how close the estimator $\hat{\boldsymbol{\theta}}$ is to $\boldsymbol{\theta}_0$. In contrast when the components induced by the MLMM are well-separated then $f_g(\boldsymbol{Y}_i|\boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g) \approx 1$ and $f_j(\boldsymbol{Y}_i|\boldsymbol{\lambda}_i^{(j)}, \boldsymbol{\theta}_j) \approx 0$ for all $j \in I_G$, $j \neq g$, and so we will have $p_{ig} \approx 1$ whilst $p_{ij} \approx 0$ for all $j \in I_G$, $j \neq g$. We will call such a classification "crisp" since the posterior probabilities clearly indicate to which component each unit belongs. In this situation when the estimator $\hat{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}_0$ then we will also have $\hat{p}_{ig} \approx 1$ whilst $\hat{p}_{ij} \approx 0$ for all $j \in I_G$, $j \neq g$, however when the estimator $\hat{\boldsymbol{\theta}}$ is not close to $\boldsymbol{\theta}_0$ then as for non well separated components it is possible the estimated posterior probabilities will take on values in the whole range $[0, 1]$.

Given the above discussion we see that when a MLMM induces well separated components in the population, and when the MLMM (and its associated fixed covariate data) admits a consistent estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$, that the classification obtained using the estimated posterior probabilities will be crisp as long as N is large enough, and furthermore that the classification will be almost error free. The ability to obtain a crisp classification of units to components with negligible amounts of error is a crucial concept for componentwise inference, and thus in turn we will have reason to frequently refer to the MLMMs which permit such a classification - accordingly we will call such MLMMs "well behaved". In contrast the classification we obtain using the estimated posterior probabilities when the components are not well separated will not necessarily be crisp (and therefore not necessarily accurate), even with consistency, and regardless of how large N is.

We are now in a position to introduce componentwise inference, the intuition behind which is simple: if we knew the component memberships of the N units then we could simply estimate the G 1-component models separately, and so we would not need to estimate a mixture model at all. If however a MLMM is well behaved then we might be close to this ideal situation, that is we might be able to classify the units to components with negligible error, and so the LMM information matrix may be all we need to perform valid inference on the component density parameters $\boldsymbol{\theta}_j$, $j = 1, ..., G$. This very roughly speaking is what we mean by componentwise inference. Thus to perform componentwise inference is to ignore that the estimators $\hat{\boldsymbol{\theta}}_j$, $j = 1, ..., G$ come from a mixture model, and to instead assume they have been estimated from G separate LMMs. Another way of stating this is that componentwise inference naively ignores the uncertainty in estimating the posterior probabilities, and hence the mixing proportions.

The idea for componentwise inference comes primarily from the R package called Flexmix (although the phrase componentwise inference is not used), and also from some ideas in Grün (2008), where we note that Bettina Grün is one of the authors of the Flexmix package. With the exception of a warning in the Flexmix manual that componentwise inference ignores the uncertainty in estimating the posterior probabilities there is no discussion of the rationale behind this method. In this respect, and for Flexmix in particular, one compelling reason to use componentwise inference is convenience. This is related to the fact that through user-defined code Flexmix permits users to fit 1-component models whose form is unknown to Flexmix, and so inference is not possible to implement by Flexmix itself. Flexmix "solves" this problem by allowing the user to perform inference on each of the component density parameters in turn using the theory behind the inference of the 1-component model. This is possible through the use of the second variant of the EM algorithm described in Subsection 2.2.2. For example for MLMMs the user of Flexmix gets Flexmix to call code to perform the weighted maximisation in equation (2.37) and thus permits confidence intervals to be constructed about $\hat{\boldsymbol{\theta}}_g$.

Other than convenience, one justification for using componentwise inference can be found by considering the second variant of the EM algorithm that we described in Subsection 2.2.2. We can see that the componentwise maximisations in step 2 require a weighted log-likelihood function of a LMM to be maximised, where the weights are the posterior probabilities for the component in question. Clearly if the estimated posterior probabilities are all either one or zero, and if this classification is correct, then these G estimators $\hat{\boldsymbol{\theta}}_j$, $j = 1, ..., G$, will be precisely the estimators obtained by splitting the data into the G components and estimating G separate LMMs.

Another way of seeing the link to the LMM is to note that for an MLMM with simple within-unit errors Grün (2008) has shown that for each component this weighted maximisation is equivalent to maximising an unweighted log-likelihood function of a LMM transformed in a particular way with the estimated posterior probabilities for that component. In this case each of the estimators of the component density parameters $\hat{\boldsymbol{\theta}}_j$, $j = 1, ..., G$, produced by the second variant of the EM algorithm will have an asymptotic normal distribution the same as that of the estimator from a LMM for a transformed response. When the transformation is such the estimated posterior probabilities are all either one or zero, and when the classification is correct, again we have that the estimators of the component density parameters will be the same as those obtained from estimating G separate LMMs. In the next sub-subsection we describe in greater detail, and justify more formally, the use of componentwise inference.

3.4.3.1 A justification for componentwise inference

Before we justify componentwise inference more formally we firstly introduce some new notation, and also adjust some of our previous notation we used in Chapter 2 so that we can simultaneously discuss estimators of the component density parameters from a MLMM, and from 1-component models. In this respect for $g \in I_G$, let $\boldsymbol{Y}^{(g)} =$

$((\boldsymbol{Y}_1^{(g)})^\intercal, ..., (\boldsymbol{Y}_{N_g}^{(g)})^\intercal)^\intercal$ be the subset of $N_g \leq N$ response vectors of $\boldsymbol{Y} = (\boldsymbol{Y}_1^\intercal, ..., \boldsymbol{Y}_N^\intercal)^\intercal$ that follow the g^{th} 1-component model defined by a MLMM, where $\boldsymbol{Y}_k^{(g)}$ for $k \in I_{N_g} := \{1, ..., N_g\}$ denotes that the response vector $\boldsymbol{Y}_k^{(g)} = \boldsymbol{Y}_i$ for some $i \in I_N$, and that unit i belongs to component g. Note that we also have $\sum_{k=1}^G N_k = N$. Similarly let $\hat{\boldsymbol{Y}}^{(g)} = ((\hat{\boldsymbol{Y}}_1^{(g)})^\intercal, ..., (\hat{\boldsymbol{Y}}_{N_g}^{(g)})^\intercal)^\intercal$ for $g \in I_G$ denote an "estimate" of the true assignment of units to component g contained in $\boldsymbol{Y}^{(g)}$, where we assume this particular assignment has been made because $\hat{p}_{ig} = \max\{\hat{p}_{i1}, ..., \hat{p}_{iG}\}$.

Letting $f_i^1(\cdot|\boldsymbol{\theta}_g)$ denote the i^{th} density function for the g^{th} 1-component model, then we will use $L^1(\boldsymbol{Y}^{(g)}|\boldsymbol{\theta}_g) = \sum_{i=1}^{N_g} \log f_i^1(\boldsymbol{Y}_i^{(g)}|\boldsymbol{\theta}_g)$ to denote the log-likelihood function for the g^{th} 1-component model which for brevity we may also shorten to just $L^1(\boldsymbol{\theta}_g)$. Letting $\boldsymbol{\theta}_j^0 \in \Psi$, $j = 1, ..., G$, denote the true component density parameters of the MLMM, then we use $\hat{\boldsymbol{\theta}}_g^1(\boldsymbol{Y}^{(g)})$ to denote the MLE of $\boldsymbol{\theta}_g^0$, that is $\hat{\boldsymbol{\theta}}_g^1(\boldsymbol{Y}^{(g)})$ is the estimator of $\boldsymbol{\theta}_g^0$ obtained from maximising $L^1(\boldsymbol{Y}^{(g)}|\boldsymbol{\theta}_g)$. For brevity we may also shorten $\hat{\boldsymbol{\theta}}_g^1(\boldsymbol{Y}^{(g)})$ to $\hat{\boldsymbol{\theta}}_g^1$. We now slightly adjust our notation we used previously in Chapter 2 for the MLMM estimator, that is $\hat{\boldsymbol{\alpha}}(\boldsymbol{Y}) = (\hat{\boldsymbol{\alpha}}_1(\boldsymbol{Y})^\intercal, ..., \hat{\boldsymbol{\alpha}}_G(\boldsymbol{Y})^\intercal, \hat{\boldsymbol{\pi}}(\boldsymbol{Y})^\intercal)^\intercal$ or $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_1^\intercal, ..., \hat{\boldsymbol{\alpha}}_G^\intercal, \hat{\boldsymbol{\pi}}^\intercal)$ for short, will denote the MLMM estimator of $\boldsymbol{\theta}_0 \in \Theta$. Thus using this notation, for any $g \in I_G$, $\hat{\boldsymbol{\theta}}_g^1$ is the 1-component model estimator of $\boldsymbol{\theta}_g^0$ using $\boldsymbol{Y}^{(g)}$, whilst $\hat{\boldsymbol{\alpha}}_g$ is the MLMM estimator of $\boldsymbol{\theta}_g^0$ using $\boldsymbol{Y}$.

We now rephrase the inference results for the LMM from Section 3.3 using this notation. For any $g \in I_G$, we have that if $\hat{\boldsymbol{\theta}}_g^1(\boldsymbol{Y}^{(g)})$ is a consistent estimator of $\boldsymbol{\theta}_g^0$ then $N^{-1} I_{N_g}^1(\hat{\boldsymbol{\theta}}_g^1(\boldsymbol{Y}^{(g)}))$ is a consistent estimator of a matrix $I^1(\boldsymbol{\theta}_g^0)$, where for any $\boldsymbol{\theta}_g \in \Psi$, $I_{N_g}^1(\boldsymbol{\theta}_g)$ is the sample information matrix for the g^{th} 1-component model and is defined as

$$I_{N_g}^1(\boldsymbol{\theta}_g) = -\sum_{i=1}^{N_g} \boldsymbol{E}_{\boldsymbol{\theta}_g} \left[\boldsymbol{H}_{\boldsymbol{\theta}}(\log f_i^1(\boldsymbol{Y}_i^{(g)}|\boldsymbol{\theta}_g) \right]. \tag{3.50}$$

This combined with the fact

$$\sqrt{N_g}(\hat{\boldsymbol{\theta}}_g^1(\boldsymbol{Y}^{(g)}) - \boldsymbol{\theta}_g^0) \xrightarrow{D} N_{n_\theta}(\mathbf{0}, (I^1(\boldsymbol{\theta}_g^0))^{-1}), \tag{3.51}$$

leads to the result that

$$\hat{\boldsymbol{\theta}}_g^1(\boldsymbol{Y}^{(g)}) \approx N_{n_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_g^0, [I_{N_g}^1\{\hat{\boldsymbol{\theta}}_g^1(\boldsymbol{Y}^{(g)})\}]^{-1}). \tag{3.52}$$

The assumption that componentwise inference might work for a well-behaved MLMM rests upon the assumption that such a model produces "almost" independent estimators of the component density parameters. The reasoning for this is as follows. Firstly because the MLMM is able to partition with negligible amounts of error the random vector $\boldsymbol{Y}$ into G separate response vectors $\hat{\boldsymbol{Y}}^{(j)}$ (from the definition of a well-behaved MLMM described in Subsection 3.4.3), $j = 1, ..., G$, and because $\boldsymbol{Y}$ has been sampled randomly from the "mixture" population of response vectors, then for any $g \in I_G$ we have that $\hat{\boldsymbol{Y}}^{(g)}$ is approximately equal to the true subset of responses $\boldsymbol{Y}^{(g)}$ for component g. Thus for any $g \in I_G$ the vector $\hat{\boldsymbol{Y}}^{(g)}$ can be thought of as having been "almost" sampled randomly from the g^{th} sub-population. Another consequence of being able to classify units to components relatively error free is that the variance of the estimator $\hat{\boldsymbol{\pi}}(\boldsymbol{Y})$ of $\boldsymbol{\pi}_0$ should be approximately zero, and furthermore so too should the covariances between $\hat{\boldsymbol{\pi}}(\boldsymbol{Y})$ and the $\hat{\boldsymbol{\alpha}}_j(\boldsymbol{Y})$, $j = 1, ..., G$.

Secondly since the components of the population are well separated, then for any $g \in I_G$, and if $\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})$ is close to $\boldsymbol{\theta}_g^0$, then $f_i^1(\boldsymbol{Y}_i|\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})) \approx 0$ if unit i does not belong to component g, and so in general only those units that belong to component g should contribute significantly to the estimator $\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})$ - that is we should have $\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}) \approx \hat{\boldsymbol{\alpha}}_g(\hat{\boldsymbol{Y}}^{(g)})$. Since from Subsection 3.4.2 we are assuming naively that the MLMM estimator $\hat{\boldsymbol{\alpha}}(\boldsymbol{Y})$ is consistent for $\boldsymbol{\theta}^0$, then this in turn implies $\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})$ will be close to $\boldsymbol{\theta}_g^0$. In this way the well separated components induced by the MLMM, and consistency of the MLMM estimator mean we should have that the estimator $\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})$ is in fact predominantly a function of $\hat{\boldsymbol{Y}}^{(g)}$. Furthermore $\hat{\boldsymbol{Y}}^{(g)} \approx \boldsymbol{Y}^{(g)}$ because the MLMM permits us to determine component memberships with negligible error. Accordingly we have $\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}) \approx \hat{\boldsymbol{\alpha}}_g(\hat{\boldsymbol{Y}}^{(g)}) \approx \hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}^{(g)})$, which means the estimators $\hat{\boldsymbol{\alpha}}_j(\boldsymbol{Y})$, and $\hat{\boldsymbol{\alpha}}_k(\boldsymbol{Y})$, for $j, k \in I_G$, $j \neq k$ should be approximately independent.

This is the justification of why a well behaved MLMM should give rise to almost independent estimators of the component density parameters, which henceforth we shall assume to hold true. Given this assumption we now justify more formally the concept of componentwise inference. We start by building on the assumptions we made in subsection 3.4.2 about $\hat{\boldsymbol{\alpha}}(\boldsymbol{Y})$ where we assumed naively that

$$\sqrt{N}(\hat{\boldsymbol{\alpha}}(\boldsymbol{Y}) - \boldsymbol{\theta}_0) \xrightarrow{D} N_{n_\Theta}(\mathbf{0}, (I(\boldsymbol{\theta}_0))^{-1}). \tag{3.53}$$

Given the approximate independence of the component density estimators $\hat{\boldsymbol{\alpha}}_j(\boldsymbol{Y})$, $j = 1, ..., G$, we have that $(I(\boldsymbol{\theta}_0))^{-1} \approx \text{diag}\{\boldsymbol{A}_1, ..., \boldsymbol{A}_G, \boldsymbol{A}_{G+1}\}$ where $\boldsymbol{A}_g$, $g \in I_G$, and $\boldsymbol{A}_{G+1} = \boldsymbol{H}_{\tilde{\boldsymbol{\pi}}}\ (L(\tilde{\boldsymbol{\pi}}))$, are the asymptotic covariance matrices respectively of $\sqrt{N}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}) - \boldsymbol{\theta}_g^0)$, and $\sqrt{N}(\hat{\boldsymbol{\pi}}(\boldsymbol{Y}) - \boldsymbol{\pi}^0)$, which are the g^{th} and $(G+1)^{th}$ sub-vectors of $\sqrt{N}(\hat{\boldsymbol{\alpha}}(\boldsymbol{Y}) - \boldsymbol{\theta}_0)$ respectively, and where $\tilde{\boldsymbol{\pi}}$ is the vector $\boldsymbol{\pi}$ with the G^{th} parameter removed (see Chapter C.2.2 for an explanation of why this is needed). Now because $\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}) \approx \hat{\boldsymbol{\alpha}}_g(\hat{\boldsymbol{Y}}^{(g)}) \approx \hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}^{(g)})$ we have

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}) - \boldsymbol{\theta}_g^0) &\approx \sqrt{N}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}^{(g)}) - \boldsymbol{\theta}_g^0) \\ &= \sqrt{\frac{N}{N_g}}\sqrt{N_g}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}^{(g)}) - \boldsymbol{\theta}_g^0) \\ &= \sqrt{\boldsymbol{\pi}_g^{-1}}\sqrt{N_g}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}^{(g)}) - \boldsymbol{\theta}_g^0). \qquad (3.54)\end{aligned}$$

We make the reasonable assumption that $(I^1(\boldsymbol{\theta}_g^0))^{-1}$, the covariance matrix of $\sqrt{N_g}(\hat{\boldsymbol{\theta}}_g^1(\boldsymbol{Y}^{(g)}) - \boldsymbol{\theta}_g^0)$, and $\boldsymbol{A}_g$, the covariance matrix of $\sqrt{N_g}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}^{(g)}) - \boldsymbol{\theta}_g^0)$, are approximately the same. This is a reasonable assumption since the diagonal form of $I(\boldsymbol{\theta}_0))^{-1}$ implies the independence of $\sqrt{N_j}(\hat{\boldsymbol{\alpha}}_j(\boldsymbol{Y}^{(j)}) - \boldsymbol{\theta}_j^0)$ and $\sqrt{N_k}(\hat{\boldsymbol{\alpha}}_k(\boldsymbol{Y}^{(k)}) - \boldsymbol{\theta}_k^0)$ for $j \neq k$, and so the variation of $\sqrt{N_j}(\hat{\boldsymbol{\alpha}}_j(\boldsymbol{Y}^{(j)}) - \boldsymbol{\theta}_j^0)$ should be determined only by the variation of $\boldsymbol{Y}^{(j)}$ and not by the variation of $\boldsymbol{Y}^{(k)}$ for $j \neq k$. Furthermore since $N_g/N = \boldsymbol{\pi}_g$ for all N as N tends to infinity, then using 3.54 and 3.53 we see that this assumption is equivalent to the assumption

$$\begin{aligned}\text{var}\left[\sqrt{N}(\hat{\boldsymbol{\alpha}}_g - \boldsymbol{\theta}_g^0)\right] &\longrightarrow \boldsymbol{\pi}_g^{-1}(\boldsymbol{A}_g) \\ &\approx \boldsymbol{\pi}_g^{-1}(I^1(\boldsymbol{\theta}_g^0))^{-1}, \qquad (3.55)\end{aligned}$$

as $N \to \infty$, and so from (3.53) and (3.55), and for any $g \in I_G$, we then have that

$$\sqrt{N}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}) - \boldsymbol{\theta}_g^0) \approx \sqrt{N}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}^{(g)}) - \boldsymbol{\theta}_g^0) \xrightarrow{D} N_{n_\theta}(\boldsymbol{0}, \boldsymbol{\pi}_g^{-1}(I^1(\boldsymbol{\theta}_g^0))^{-1}). \qquad (3.56)$$

Finally if the MLMM estimator $\hat{\boldsymbol{\alpha}}(\boldsymbol{Y})$ is consistent for $\boldsymbol{\theta}_0$ then $\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})$ is consistent for $\boldsymbol{\theta}_g^0$, and so $\boldsymbol{\pi}_g N_g^{-1} I_{N_g}^1(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}))$ is a consistent estimator of $\boldsymbol{\pi}_g I^1(\boldsymbol{\theta}_g^0) \approx \boldsymbol{\pi}_g \boldsymbol{A}_g^{-1}$, and so using 3.56 it is reasonable to assert that

$$\begin{aligned}&\left(\boldsymbol{\pi}_g N_g^{-1} I^1_{N_g}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}))\right)^{\frac{1}{2}} \sqrt{N}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}) - \boldsymbol{\theta}^0_g)\\ &\qquad = \left(I^1_{N_g}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}))\right)^{\frac{1}{2}} \sqrt{N}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}) - \boldsymbol{\theta}^0_g) \xrightarrow{D} N_{n_\theta}(\mathbf{0}, \boldsymbol{I}_{n_\theta}),\end{aligned} \tag{3.57}$$

and so

$$\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}^{(g)}) \approx N_{n_{\boldsymbol{\theta}}}(\boldsymbol{\theta}^0_g, (I^1_{N_g}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})))^{-1}). \tag{3.58}$$

If we knew the true component memberships, then from 3.58 for each $\hat{\boldsymbol{\alpha}}_j(\boldsymbol{Y}^{(g)})$, $j = 1, ..., G$, we could derive approximate Normal confidence intervals. We will call these confidence intervals componentwise confidence intervals, and similarly $I^1_{N_g}(\cdot)$ the componentwise sample information matrix. In words equation 3.58 means we use the sample of response vectors $\boldsymbol{Y}^{(g)}$ known to be in component g in order to calculate the componentwise sample information matrix $I^1_{N_g}(\cdot)$ in (3.50), but we evaluate this information matrix at the estimator $\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})$ of $\boldsymbol{\theta}^0_g$ obtained from the MLMM using the whole sample $\boldsymbol{Y}$.

In practice of course we do not know the true component memberships, and so even under the favorable assumption of a well behaved MLMM, componentwise inference as we have described it here will not work. In this respect the obvious modification to equation 3.58 is to work with the vector of all responses $\boldsymbol{Y}$ in $I^1_{N_g}(\cdot)$, but to rely on the estimated posterior probabilities $\hat{p}_{ig}$ to ensure only those units that belong to component g actually contribute significantly to the 1-component model information matrix. That is instead of (3.50) we use should instead use the following equation

$$CW_N(\hat{\boldsymbol{\alpha}}_g) = -\sum_{i=1}^{N} \boldsymbol{E}_{\hat{\boldsymbol{\alpha}}_g}\left[\boldsymbol{H}_{\boldsymbol{\theta}}(\hat{p}_{ig}\text{log}f^1_i(\boldsymbol{Y}^{(g)}_i|\hat{\boldsymbol{\alpha}}_g)\right], \tag{3.59}$$

where the CW stands for componentwise. When the estimated posterior probabilities achieve a crisp classification of units to components, and when this classification is correct, then $CW_N(\boldsymbol{\theta}_g) \approx I^1_{N_g}(\boldsymbol{\theta}_g)$ for all $\boldsymbol{\theta}_g \in \Psi$. We note again that the estimated posterior probabilities need not achieve a crisp classification even if the MLMM parameters have all been estimated very accurately, rather a crisp classification requires both accurate parameter estimates combined with an MLMM that produces well separated components.

In this respect we are assuming a well behaved MLMM that produces well separated components, and consistency of the MLMM parameter estimator (that we are naively assuming) ensures the required accuracy of the estimates for a large enough number of units. Thus in turn these assumptions imply that the estimated posterior probabilities get closer and closer to the true classification of units to components as the number of units tends to infinity, and that this classification becomes crisp. We then have that $CW_N(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})) \to I^1_{N_g}(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}))$ as $N \to \infty$ for any $g \in I_G$. But since $\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})$ is consistent for $\boldsymbol{\theta}^0_g$, and since $N^{-1}I^1_{N_g}(\boldsymbol{\theta}_g)$ converges in probability to $I^1(\boldsymbol{\theta}_g)$ for all $\boldsymbol{\theta}_g \in \Psi$, we have that $N^{-1}CW_N(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})) \xrightarrow{P} I^1(\boldsymbol{\theta}^0_g)$ as $N \to \infty$.

Thus if the estimated posterior probabilities converge to values that give the true classification of units to well separated components, then $N^{-1}CW_N(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}))$ is a consistent estimator of $I^1(\boldsymbol{\theta}^0_g)$, which justifies us replacing equation 3.58 with

$$\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y}) \approx N_{n_{\boldsymbol{\theta}}}(\boldsymbol{\theta}^0_g, (CW_N(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{Y})))^{-1}). \tag{3.60}$$

We propose to use equations 3.60 and 3.59 to derive approximate Normal confidence intervals for the component density parameters in a MLMM, and this is what we mean by performing componentwise inference.

In Section C.3 we present an alternative and more formal justification for componentwise inference based on the derivatives of the log-likelihood function we derive in Appendix C.

4

Identifiability

This chapter is concerned with establishing sufficient conditions under which an MLMM is identifiable, that is to say we want to establish the conditions under which, probabilistically, the parametrization of the model is unique. Theoretically such a concept is crucial because non-uniqueness in the parametrization means there do not exist either asymptotically unbiased nor consistent estimators of the model parameters, regardless of the methods used to obtain them (San Martin and Quintana, 2002). We will also discuss identifiability for MLMs since this discussion will motivate our discussion for MLMMs.

In section 4.1 we give an alternative formulation of an MLMM to the hierarchical one we described in chapter 2, which we will call the mixing distribution formulation. The idea behind this formulation is that the mixture distribution function of the sample $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$ is generated from an underlying 1-component model distribution function by the mixing distribution. Section (4.2) defines MLMMs and MLMs using the mixing distribution formulation introduced in section 4.1, and introduces the definitions of identifiability we shall need later in the chapter. Section 4.3 contains the main material of this chapter, wherein we will use the mixing distribution introduced in section (4.1) to parametrize families of distribution functions for MLMMs, and it is with respect to this parametrization that we shall consider the problem of identifiability. The section contains two theorems that give different sufficient conditions for identifiability of a MLMM with an unspecified covariance structure, and a corollary to one of the theorems for a MLMM with a simple within-unit error covariance structure.

4.1 Mixing distribution

In this section we introduce the concept of a latent or mixing distribution that generates the mixture density given in 2.8, and we follow the description given in Lindsay (1995, section 1.2, pp6) to do this. There are two reasons for introducing this formulation. Firstly this formulation makes clear that, by being the building block of the mixture distribution function, that it is the 1-component model distribution function that should be the focus of attention concerning identifiability issues. In this respect since the 1-component model distribution function is a multivariate normal distribution which is completely specified by its mean vector and covariance matrix, we should focus our attention on how identifiability problems with the parametrization of these mean vectors and covariance matrices might cause identifiability problems with the mixture. Secondly for theoretical work focusing on identifiability the notation used with the latent distribution formulation makes the proofs more compact and clearer to read. here

As in chapter 2 we assume a sample of response vectors $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$, $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$, $i = 1, ..., N$, from N units out of a population of units that consists of an unknown number of subpopulations or components. We shall use s for the unknown numbers of components, and to be consistent with the notation of chapter 2, $s = G$ when the number of components is considered known. It is assumed each unit belongs to only one of the s components, and that the units are sampled randomly. We will write π_g for the proportion of the population in component g, for $g \in \{1, 2, ..., s\}$, so that $\sum_{j=1}^{s} \pi_j = 1$. As in chapter 2, and unless otherwise stated, we will always use the index g to denote a particular component chosen from the s possible components, whilst we will use another index, usually j, when we want to reference all the $j = 1, ..., s$ components.

Let the random variable I_i denote the component membership of each $\boldsymbol{Y}_i$, so that $I_i = g$ and $P[I_i = g] = \pi_g$ for some $g \in \{1, 2, ..., s\}$. It is further assumed that, conditional on $I_i = g$, $\boldsymbol{Y}_i$ has density function $f_{ig}(\boldsymbol{y}_i | I_i = g, \boldsymbol{\theta_g})$, where $\boldsymbol{\theta_g} \in T \subseteq \mathbb{R}^{n_\theta}$, and n_θ is the total number of parameters in $\boldsymbol{\theta}$. These densities will be called the component densities, and we note that for $j = 1, ..., s$, the component densities $f_{ij}(\boldsymbol{y}_i | I_i = j, \boldsymbol{\theta_j})$ using the notation of this section are equal to the density functions $f_{ij}(\boldsymbol{y}_i | \boldsymbol{\lambda}_i^{(j)}, \boldsymbol{\theta}_j)$ in chapter 2.

For this section we will leave the f_{ig} unspecified, but in section 4.3 these will be density functions of multivariate normal distributions induced by a LMM. The density function for the joint random variable $(\boldsymbol{Y}_i, I_i)$ is given by $f_{ig}(\boldsymbol{y}_i|I_i = g, \boldsymbol{\theta}_g)\pi_g$, and the marginal density function for $\boldsymbol{Y}_i$ is $f_i(\boldsymbol{y}_i|\boldsymbol{\theta}) = \sum_{j=1}^{s} f_{ij}(\boldsymbol{y}_i|I_i = j, \boldsymbol{\theta}_j)\pi_j$, where $\boldsymbol{\theta} \in \Theta$, and

$$\Theta = \left\{ (\boldsymbol{\theta}_1^{\intercal}, ..., \boldsymbol{\theta}_s^{\intercal}, \pi_1, ..., \pi_s)^{\intercal} : \sum_{j=1}^{s} \pi_j = 1, \pi_j \geq 0, \boldsymbol{\theta}_j \in T, j = 1, ..., s \right\}. \tag{4.1}$$

This is called a G-component finite mixture model if it is known there are $s = G$ components.

We now introduce latent random variables $\{\boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_N\}$ which we define by $I_i = g \Leftrightarrow \boldsymbol{\Phi}_i = \boldsymbol{\theta}_g$ for all $i = 1, ..., N$, and $g \in \{1, ..., s\}$, so that for each unit the realized value of $\boldsymbol{\Phi}_i$ determines the component to which the unit belongs, and vice-versa. The realized value of each $\boldsymbol{\Phi}_i$ will be written $\boldsymbol{\phi}_i$, so $\boldsymbol{\phi_i} \in T$ for all $i = 1, ..., N$. We note that in this chapter we refer to the $\boldsymbol{\tau}$ parametrization for the autocorrelation matrices of the within-unit errors, thus the latent random variables $\boldsymbol{\phi}_i$, $i \in I_N$ we use here need not be confused with the autoregressive parameters $\boldsymbol{\phi}_g$, $g \in I_G$ we refer to in other chapters.

Letting ϕ denote a general point in T, then for each $\phi \in T$ let $\tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i = \phi)$ be the density function for the random variable $\boldsymbol{Y}_i$ conditional on the value of $\boldsymbol{\Phi}_i$, so that $\tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i = \phi) = f_{ig}(\boldsymbol{y}_i|I_i = g, \boldsymbol{\theta}_g)$ when $\boldsymbol{\Phi}_i = \boldsymbol{\phi} = \boldsymbol{\theta}_g$. Using this notation, for each $i = 1, ..., N$, we have that $\tilde{f}_i$ is in the family of density functions $\mathscr{F}_i = \{\tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i = \phi) : \boldsymbol{\phi} \in T\}$, and that the s density functions $f_{ij}(\boldsymbol{y}_i|I_i = j, \boldsymbol{\theta}_j)$, $j = 1, ..., s$, are all contained $\mathscr{F}_i$.

It is assumed the $\boldsymbol{\Phi}_i$ are a *iid* sample from a distribution J which is a discrete probability measure that assigns "masses" $\pi_1, ..., \pi_s$ at the points $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_s$. The measurable space on which J is defined is $(T, \mathscr{T})$, where $\mathscr{T}$ is a σ-field of subsets of T. The support set $S(J) = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_s\}$ of J is a minimal support set in the sense $S(J)$ does not contain any points to which J assigns zero probability. Using these definitions we have $J(A) = \sum_k \pi_k$ for any $A \in \mathscr{T}$, where the sum extends over all the $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_s$ in A.

The distribution J is called the latent or mixing distribution, and the assumptions just introduced mean we have the relation

$$P[\boldsymbol{\Phi}_i = \boldsymbol{\theta}_g] = J(\{\boldsymbol{\theta}_g\}) = \pi_g. \tag{4.2}$$

Since probability measures are uniquely defined on the σ-field of the measurable space, then we can equate the unknown parameters $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_s\}$, and $\{\pi_1, ..., \pi_s\}$ uniquely with the distribution J on the parameter space T. Thus estimating the parameters of the mixture model $f_i(\boldsymbol{y}_i|\boldsymbol{\theta}) = \sum_{j=1}^{s} f_{ij}(\boldsymbol{y}_i|I_i = j, \boldsymbol{\theta}_j)\pi_j$ is the same thing as estimating the unknown mixing distribution J with minimal support set $S(J)$ on T (Lindsay, 1995, pp7). We will denote the set of mixing distributions with finite support on T as $\mathfrak{J}(T)$, and for any $J \in \mathfrak{J}(T)$, $s = |S(J)|$ will be the number of points in the support set of J.

It is worth noting that the label-switching problem associated with mixture models simply does not occur with the mixing distribution formulation described here. This is because J assigns masses to points in $S(J)$ regardless of what we call or label the points as. Of course we introduce the label switching problem as soon as we label the points, as we must do for practical and convenience purposes. However for theoretical purposes the only type of identifiability problems are of the non-trivial type when using the mixing distribution formulation.

Lindsay (1995, section 1.2, pp7) states that, since the component density $\tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i = \boldsymbol{\phi})$ depends on the component only through the parameter $\boldsymbol{\phi}$, the mixture density $f_i(\boldsymbol{y}_i|\boldsymbol{\theta})$ can be written as an expectation of the component density $\tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i)$ with respect to the mixing distribution J. To see this let $\boldsymbol{\phi}$ denote a general point in T, and choose disjoint $\mathscr{T}$-sets $A_1 = \{\boldsymbol{\theta_1}\}, ..., A_s = \{\boldsymbol{\theta_s}\}, A_{s+1} = T \bigcap \left(\bigcup_{j=1}^{s} A_j\right)^c$, so that $T = \bigcup_{k=1}^{s+1} A_k$, and $J(\boldsymbol{\phi}) = 0$ for all $\boldsymbol{\phi} \in A_{s+1}$. For any set $A \in \mathscr{T}$, let $I_A(\boldsymbol{\phi})$ be the indicator function that is one when $\boldsymbol{\phi} \in A$, and zero if not. For brevity we will write $\tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i = \phi)$ as $\tilde{f}_i(\boldsymbol{y}_i|\phi)$. Then since $\tilde{f}_i$ is non-negative, from Billingsley (1995, Theorem 16.9, pp212) we have

$$\begin{aligned}E_J[\tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i)] &= \int_T \tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\phi})dJ(\boldsymbol{\phi}) \\ &= \sum_{j=1}^{s+1} \int_{A_j} \tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\phi})dJ(\boldsymbol{\phi}) \\ &= \sum_{j=1}^{s} \int_T I_{A_j}(\boldsymbol{\phi})\tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\phi})dJ(\boldsymbol{\phi}) + \int_T I_{A_{s+1}}(\boldsymbol{\phi})\tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\phi})dJ(\boldsymbol{\phi}) \\ &= \sum_{j=1}^{s} \int_T I_{\{\boldsymbol{\theta}_j\}}(\boldsymbol{\phi})\tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\theta}_j)dJ(\boldsymbol{\phi}) + 0 \\ &= \sum_{j=1}^{s} \tilde{f}_i(\boldsymbol{y}_i|\boldsymbol{\theta}_j)J(\{\boldsymbol{\theta}_j\}) \\ &= \sum_{j=1}^{s} f_{ij}(\boldsymbol{y}_i|I_i = j, \boldsymbol{\theta_j})\pi_j \\ &= f_i(\boldsymbol{y}_i|\boldsymbol{\theta}). \end{aligned} \tag{4.3}$$

In terms of distribution functions, let $F_{ig}(\boldsymbol{y}_i|I_i = g, \boldsymbol{\theta_g})$ be the distribution function of $\boldsymbol{Y}_i$ conditional on unit i being in component g, and we note that this distribution function is equivalent to $F_{ig}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g)$ introduced in Chapter 2. Similar to the density function $\tilde{f}_i$, for any $\phi \in T$ let $\tilde{F}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i = \phi)$ be the distribution function of $\boldsymbol{Y}_i$ conditional on $\boldsymbol{\Phi}_i = \phi$, so that $\tilde{F}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i = \phi) = F_{ig}(\boldsymbol{y}_i|I_i = g, \boldsymbol{\theta_g})$ when $\boldsymbol{\Phi}_i = \boldsymbol{\phi} = \boldsymbol{\theta}_g$. Using this notation, for each $i = 1, ..., N$, we have that $\tilde{F}_i$ is in the family of distribution functions $\mathscr{D}_i = \{\tilde{F}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i = \phi) : \boldsymbol{\phi} \in T\}$, and that the s distribution functions $F_{ij}(\boldsymbol{y}_i|I_i = j, \boldsymbol{\theta}_j)$, $j = 1, ..., s$, are all contained $\mathscr{D}_i$.

If we now let $F_i(\boldsymbol{y}_i|\boldsymbol{\theta})$ be the mixture distribution function for the i^{th} unit, so that $F_i(\boldsymbol{y}_i|\boldsymbol{\theta}) = \sum_{j=1}^{s} F_{ij}(\boldsymbol{y}_i|I_i = j, \boldsymbol{\theta}_j)\pi_j$, then similar to 4.3 we have

$$\begin{aligned}E_J[\tilde{F}_i(\boldsymbol{y}_i|\boldsymbol{\Phi}_i)] &= \int_T \tilde{F}_i(\boldsymbol{y}_i|\phi)dJ(\phi) \\ &= \sum_{j=1}^{s} F_{ij}(\boldsymbol{y}_i|I_i = j, \boldsymbol{\theta}_j)\pi_j \\ &= F_i(\boldsymbol{y}_i|\boldsymbol{\theta}). \end{aligned} \tag{4.4}$$

If we denote $\int_T \tilde{F}_i(\boldsymbol{y}_i|\phi)dJ(\phi)$ by $F_i(\boldsymbol{y}_i|J)$, then (4.4) shows that $F_i(\boldsymbol{y}_i|J)$ is the mixture distribution function for $\boldsymbol{Y}_i$, where the mixture is generated by the mixing distribution

J. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1^\intercal, ..., \boldsymbol{Y}_N^\intercal)^\intercal$ denote the joint vector of the N unit response vectors in our sample, and recall that $F(\boldsymbol{y}|\boldsymbol{\theta})$ is the mixture distribution function of the sample which was introduced in chapter 2. Analogously if we define $F(\boldsymbol{y}|J) = \prod_{i=1}^N F_i(\boldsymbol{y}_i|J)$ to be the distribution function for $\boldsymbol{Y}$ parametrized by J, then from (4.4) we have the following relationship between $F(\boldsymbol{y}|\boldsymbol{\theta})$ and $F(\boldsymbol{y}|J)$

$$\begin{aligned} F(\boldsymbol{y}|\boldsymbol{\theta}) &= \prod_{i=1}^N F_i(\boldsymbol{y}_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^N \int_T \tilde{F}_i(\boldsymbol{y}_i|\boldsymbol{\phi}) dJ(\boldsymbol{\phi}) \\ &= \prod_{i=1}^N F_i(\boldsymbol{y}_i|J) \\ &= F(\boldsymbol{y}|J). \end{aligned} \tag{4.5}$$

Now for any $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\intercal, ..., \boldsymbol{\theta}_s^\intercal, \pi_1, ..., \pi_s)^\intercal \in \Theta$ there exists a unique $J \in \mathfrak{J}(T)$ such that J has support set $S(J) = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_s\}$, and masses $\{\pi_1, ..., \pi_s\}$. Conversely for any $J \in \mathfrak{J}(T)$ with support set $S(J) = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_s\}$, and masses $\{\pi_1, ..., \pi_s\}$, there exists a unique $\boldsymbol{\theta} \in \Theta$ that satisfies $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\intercal, ..., \boldsymbol{\theta}_s^\intercal, \pi_1, ..., \pi_s)^\intercal$. Thus letting $\mathcal{D}$ denote the family of distribution functions for the sample response vector $\boldsymbol{Y}$ we have

$$\mathcal{D} := \{F(\boldsymbol{y}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\} = \{F(\boldsymbol{y}|J) : J \in \mathfrak{J}(T)\}, \tag{4.6}$$

which shows that $\mathcal{D}$ can be parametrized by either points in Θ or by the mixing distributions in $\mathfrak{J}(T)$. In the next section we shall use the mixing distribution parametrization of $\mathcal{D}$ to investigate identifiability problems in MLMMs.

4.2 Definitions for identifiability

In the first part of this section we introduce model notation for the distribution functions of two samples of random variables, one of which is assumed to follow a MLMM, the other a MLM. Both models induce multivariate normal distributions in these samples. The second part of this section introduces the definitions of identifiability we shall use in section 4.3, which concern the parametrization of families of distribution functions induced by the mixture models.

We will consider two samples of random variables $(\boldsymbol{Y}_i)_{i\in I}$, $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$, indexed by a general index set I, and we will use $\mathscr{L}(\{\boldsymbol{Y}_i\}_{i\in I})$ to denote the probability laws of these samples. Starting with a MLMM, if each $\boldsymbol{Y}_i$ follows the model in 2.1, in combination with all the distributional assumptions that follow it, then the marginal distribution function of the sample can be written

Model 1: MLMM marginal distribution function

$$F(\boldsymbol{Y}|J) := \mathscr{L}(\{\boldsymbol{Y}_i\}_{i\in I}) = \bigotimes_{i\in I} F_i(\boldsymbol{y}_i|J) \text{ where}$$

$$F_i(\boldsymbol{y}_i|J) = \int_{\Psi_1} \boldsymbol{\Phi}_{\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{V}_i(\boldsymbol{\zeta})}(\boldsymbol{y}_i)dJ(\boldsymbol{\beta},\boldsymbol{\zeta}), \quad \Psi_1 := \mathbb{R}^p \times \Psi_{\boldsymbol{\zeta}},$$

$$\boldsymbol{\beta} \in \mathbb{R}^p, \quad \boldsymbol{\zeta} = (\mathrm{v}(\boldsymbol{D})^\intercal, \sigma^2, \boldsymbol{\tau}^\intercal)^\intercal \in \Sigma_{\boldsymbol{\zeta}}, \quad \Sigma_{\boldsymbol{\zeta}} := \Sigma_{\mathrm{v}(\boldsymbol{D})} \times \mathbb{R}^+ \times \Sigma_\tau,$$

$$\Sigma_\tau := ([-1,1]^\intercal)^r = [-1,1]^\intercal \times \cdots \times [-1,1]^\intercal \subseteq \mathbb{R}^r \times \cdots \times \mathbb{R}^r,$$

$$\Sigma_{\mathrm{v}(\boldsymbol{D})} \subseteq \mathbb{R}^{q(q+1)/2}, \quad \boldsymbol{V}_i(\boldsymbol{\zeta}) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\intercal + \sigma^2 C_i(\boldsymbol{\tau}),$$

$$J \in \Omega_1 := \mathfrak{J}(\Psi_1).$$

where $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$ for all $i \in I$, $\boldsymbol{Y} = (\boldsymbol{Y}_1^\intercal, ..., \boldsymbol{Y}_N^\intercal)^\intercal$, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $n_i \times p$ and $n_i \times q$ fixed matrices respectively. The parameter space for the covariance parameters, $\Sigma_{\mathrm{v}(\boldsymbol{D})}$, is the subset of $\mathbb{R}^{q(q+1)/2}$ that gives rise to a positive definite symmetric $q \times q$ matrix $\boldsymbol{D}$ obtained by "unvectorising" the vector defined by $\mathrm{vec}(\boldsymbol{D}) = D_q\mathrm{v}(\boldsymbol{D})$, where D_q is the $q^2 \times (q(q+1)/2)$ duplication matrix. Similarly Σ_τ is such that $C_i(\boldsymbol{\tau})$ is a positive-definite AR correlation matrix for all $\boldsymbol{\tau} \in \Sigma_\tau$. The symbol "$\bigotimes$" means an independent product of distributions, and $\boldsymbol{\Phi}_{\mu_i,\Sigma_i}(\cdot)$ denotes the cumulative distribution function of the n_i-dimensional Normal density function with mean vector μ_i, and positive definite covariance matrix Σ_i.

Model 1 assumes the n_i observations for each unit to be equally spaced with no missing values, however when $\boldsymbol{\tau}$ is restricted to be the zero vector we will permit missing values to occur. This avoids defining a second model on account of this difference which is not important in what follows. Note that $\boldsymbol{V}_i(\boldsymbol{\zeta}) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\intercal + \sigma^2 C_i(\boldsymbol{\tau})$ is positive definite since $\boldsymbol{D}$, and $C_i(\boldsymbol{\tau})$ are positive definite.

Remembering that $s = |S(J)|$, then for each $i \in I$, J generates the following mixture distribution

$$F(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i \in I} \sum_{j=1}^{s} F_{ij}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(j)}, \boldsymbol{\theta}_j)\boldsymbol{\pi}_j, \tag{4.7}$$

where $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j^\intercal, \boldsymbol{\zeta}_j^\intercal)^\intercal \in \Psi_1$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\intercal, ..., \boldsymbol{\theta}_G^\intercal, \pi_1, ..., \pi_G)^\intercal \in \Theta_1$, and

$$\Theta_1 = \left\{ (\boldsymbol{\theta}_1^\intercal, ..., \boldsymbol{\theta}_G^\intercal, \pi_1, ..., \pi_G)^\intercal : \sum_{j=1}^{G} \pi_j = 1, \pi_j \geq 0, \boldsymbol{\theta}_j \in \Psi_1, j = 1, ..., G \right\}. \tag{4.8}$$

Now if we have a sample $(\boldsymbol{Y}_i)_{i \in I}$ where each $\boldsymbol{Y}_i$ follows a 1-component version of Model 1, or a LMM, then the marginal distribution function (after integrating out the random effects) of the sample is given by

LMM marginal distribution function

$$F^1(\boldsymbol{Y}|\boldsymbol{\theta}) := \mathscr{L}(\{\boldsymbol{Y}_i\}_{i \in I}) = \bigotimes_{i \in I} F_i^1(\boldsymbol{y}_i|\boldsymbol{\theta}) \text{ where}$$

$$F_i^1(\boldsymbol{y}_i|\boldsymbol{\theta}) = \Phi_{\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i(\boldsymbol{\zeta})}(\boldsymbol{y}_i), \quad \boldsymbol{\theta} := (\boldsymbol{\beta}^\intercal, \boldsymbol{\zeta}^\intercal)^\intercal \in \Psi_1,$$

$$\boldsymbol{\beta} \in \mathbb{R}^p, \quad \boldsymbol{\zeta} := (\mathrm{v}(\boldsymbol{D})^\intercal, \sigma^2, \boldsymbol{\tau}^\intercal)^\intercal \in \Sigma_{\boldsymbol{\zeta}}, \quad \Psi_{\boldsymbol{\zeta}} := \Sigma_{\mathrm{v}(\boldsymbol{D})} \times \mathbb{R}^+ \times \Sigma_{\boldsymbol{\tau}},$$

$$\Sigma_\tau := ([-1,1]^\intercal)^r = [-1,1]^\intercal \times \cdots \times [-1,1]^\intercal \subseteq \mathbb{R}^r \times \cdots \times \mathbb{R}^r,$$

$$\Sigma_{\mathrm{v}(\boldsymbol{D})} \subseteq \mathbb{R}^{(q(q+1)/2)}, \quad \Psi_1 := \mathbb{R}^p \times \Psi_{\boldsymbol{\zeta}},$$

$$\boldsymbol{V}_i(\boldsymbol{\zeta}) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\intercal + \sigma^2 C_i(\boldsymbol{\tau}),$$

where $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$ for all $i \in I$, and $\boldsymbol{Y} = (\boldsymbol{Y}_1^\intercal, ..., \boldsymbol{Y}_N^\intercal)^\intercal$, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $n_i \times p$ and $n_i \times q$ fixed matrices respectively. The superscripts 1 denote 1-component as opposed to mixture distribution functions. Note that we are not using the mixing distribution in the above definition.

We will also need an analogous definition of (4.2) for MLMs as follows

Model 2: MLM distribution function

$$F(\boldsymbol{Y}|J) := \mathscr{L}(\{\boldsymbol{y}_i\}_{i\in I}) = \bigotimes_{i\in I} F_i(\boldsymbol{y}_i|J) \text{ where}$$

$$F_i(\boldsymbol{y}_i|J) = \int_{\Psi_2} \Phi_{\boldsymbol{X}_i\boldsymbol{\beta},\sigma^2\boldsymbol{I}_{n_i}}(\boldsymbol{y}_i) dJ(\boldsymbol{\beta},\sigma^2), \quad \Psi_2 := \mathbb{R}^p \times \mathbb{R}^+,$$

$$\boldsymbol{\beta} \in \mathbb{R}^p, \quad \sigma^2 \in \mathbb{R}^+, \quad J \in \Omega_2 := \mathfrak{J}(\Psi_2),$$

where $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$ for all $i \in I$, and $\boldsymbol{Y} = (\boldsymbol{Y}_1^\intercal, ..., \boldsymbol{Y}_N^\intercal)^\intercal$. For clusterwise regression models ($n_i = 1$ for all $i \in I$) Model 2 can be written

Model 2: Clusterwise regression distribution function

$$F(\boldsymbol{Y}|J) := \mathscr{L}(\{y_i\}_{i\in I}) = \bigotimes_{i\in I} F_i(y_i|J) \text{ where}$$

$$F_i(y_i|J) = \int_{\Psi_2} \Phi_{\boldsymbol{x}_i^\intercal\boldsymbol{\beta},\sigma^2}(y_i) dJ(\boldsymbol{\beta},\sigma^2), \quad \Psi_2 := \mathbb{R}^p \times \mathbb{R}^+,$$

$$\boldsymbol{\beta} \in \mathbb{R}^p, \quad \sigma^2 \in \mathbb{R}^+, \quad J \in \Omega_2 := \mathfrak{J}(\Psi_2),$$

where $y_i \in \mathbb{R}$ for all $i \in I$, $\boldsymbol{Y} = (y, ..., y_N)^\intercal$, and $x_i \in \mathbb{R}^p$. The mixture distribution function for Model 2 parameterized without the mixing distribution J will be the same as is given in (4.7) but where the covariance matrices of the $F_{ij}(\boldsymbol{y}_i|\boldsymbol{\lambda}_i^{(j)}, \boldsymbol{\theta}_j)$ will be equal to $\sigma^2\boldsymbol{I}_{n_i}$ for MLMs or σ^2 for clusterwise regression.

We now introduce some definitions of identifiability of the models we have introduced above. Define a family of mixture distribution functions for a mixture model as

$$\mathscr{D}_l := \left\{ F(\cdot|J) : F(\cdot|J) = \bigotimes_{i\in I} F_i(\cdot|J), J \in \Omega_l \right\}, \tag{4.9}$$

where $l = 1, 2$, denotes the members of $\mathscr{D}_l$ are distribution functions for Model l. Then we have the following definition of identifiability for $\mathscr{D}_l$

Definition 4.2.1 *$\mathscr{D}_l$ is identifiable with respect to Ω_l if*

$$\forall J, \hat{J} \in \Omega_l : \qquad F(\cdot|J) = F(\cdot|\hat{J}) \Leftrightarrow J = \hat{J}, \tag{4.10}$$

or in terms of the distribution functions

Definition 4.2.2 *$\mathscr{D}_l$ is identifiable with respect to Ω_l if*

$$\forall J, \hat{J} \in \Omega_l : \qquad F_i(\cdot|J) = F_i(\cdot|\hat{J}) \quad \forall i \in I \Leftrightarrow J = \hat{J}. \tag{4.11}$$

An equivalent definition of identifiability to 4.2.2, based on the one given by Yakowitz and Spragins (1968), is that the mapping $F_i(\cdot|J) = \int_T \Phi_{\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{V}_i(\boldsymbol{\zeta})}(\cdot)dJ(\boldsymbol{\beta},\boldsymbol{\zeta})$ from Ω_l to $\mathscr{D}_l$ is one-one for all $i \in I$.

We see from these definitions that identifiability relates to the whole sample: if $J = \hat{J}$, identifiability can fail to hold even if the distribution functions of just a single $i \in I$ are not equal under both mixing distributions. Similarly if identifiability holds, and $J \neq \hat{J}$, then $F_i(\cdot|J) \neq F_i(\cdot|\hat{J})$ for at least one $i \in I$, but maybe only one i - in this case a single unit alone identifies the parameters. This raises an interesting question regarding the identifiability of $\mathscr{D}_l$ as N tends to infinity, if no further units are added to the sample that do identify the parameters. Theoretically $\mathscr{D}_l$ will still be identifiable no matter how large N becomes, however the information in the units that do not identify the parameters may "swamp" the information in units that does identify the parameters. Perhaps in this case $\mathscr{D}_l$ may become close to non-identifiable in some sense, rather like a matrix can be near to being collinear.

The potential problem described above does not occur when the sample is *iid*, and may be relevant for any consistency proof of parameter estimators in the MLMM. This issue has been discussed by Hennig (2000) for clusterwise regression, wherein the term "observational model" is used to describe the situation above where, as $N \to \infty$, units with different covariate data, and hence different distribution functions, are introduced.

Hennig also describes an alternative interpretation of the index set I, called a "repeatable design", which for MLMMs means that the $M \leq N$ distinct distribution functions are repeated. This implies the fixed covariate data $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are repeated, which could happen in a designed experiment, where we imagine the whole experiment is repeated. If the covariate data arise from observation rather than design, then the repeatable design interpretation is probably unrealistic. The logic behind the repeatable design interpretation is that conditions are imposed to ensure the first N units identify $\mathscr{D}_l$, and then the data $(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)_{i \in I}$ are repeated *iid*. This approach would not only overcome the "swamping" of information problem, but would also perhaps enable easier consistency proofs by being able to use *iid* theory.

In section (4.3) we will need to discuss the identifiability of the LMM, and so we now define identifiability for this model too. First define the family of LMM distribution functions $\mathscr{D}^1$ as

$$\mathscr{D}^1 := \left\{ F^1(\cdot|\boldsymbol{\theta}) : F^1(\cdot|\boldsymbol{\theta}) = \bigotimes_{i \in I} F_i^1(\cdot|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Psi_1 \right\}. \tag{4.12}$$

Since $\mathscr{D}^1$ is a special case of $\mathscr{D}$, the identifiability definitions (4.2.1) and (4.2.2) apply. However a more specific definition can be obtained for the LMM since the distribution functions $F_i^1(\cdot|\boldsymbol{\theta})$ are normal distributions. This definition is given in proposition 10 of Demidenko (2004, pp 118) and concerns the distribution function of the whole sample $F^1(\cdot|\boldsymbol{\theta})$. Restating this definition in terms of the I units we get

Definition 4.2.3 *$\mathscr{D}^1$ is identifiable with respect to Ψ_1 if*

$$\forall \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \in \Psi_1 : \quad \boldsymbol{X}_i \boldsymbol{\beta} = \boldsymbol{X}_i \hat{\boldsymbol{\beta}} \text{ and } \boldsymbol{V}_i(\boldsymbol{\zeta}) = \boldsymbol{V}_i(\hat{\boldsymbol{\zeta}}) \quad \forall i \in I \Leftrightarrow \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}. \tag{4.13}$$

Partly based upon this result, for LMMs with a simple within-unit error covariance structure Demidenko (2004) gives the following sufficient conditions for identifiability

Theorem 4.2.4 *(Theorem 11, Demidenko, 2004, p 118) For the MLMM with $\boldsymbol{V}_i(\boldsymbol{\zeta}) = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^\intercal + \sigma^2 \boldsymbol{I}_{n_i}$ for all $i = 1, ..., N$, then $\mathscr{D}^1$ is identifiable with respect to Ψ_1 if at least one $\boldsymbol{Z}_i$ is full rank, $\tilde{\boldsymbol{X}}$ is full rank, and $\sum_{i=1}^N (n_i - q) > 0$.*

Hennig (2000) gives sufficient conditions for the identifiability of clusterwise regression models in terms of the number of hyperplanes the fixed effects covariate data concentrate on. In the next section we will discuss this result, and how it relates to mixtures of multivariate normal distributions with regression components. For this reason we need the following definition of an $(m-1)$-dimensional hyperplane $H_{m-1}(\boldsymbol{\alpha}, c)$

$$H_{m-1}(\boldsymbol{\alpha}, c) = \{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{\alpha}^\intercal \boldsymbol{x} = c, \boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\alpha} \neq \boldsymbol{0}, c \in \mathbb{R}\}. \tag{4.14}$$

We will also need the theorem below, and a corollary of it, which relate the rank of a matrix to the above hyperplane definition - a proof of the Theorem can be found in Appendix A.2. The theorem and corollary use the following notation: for a $n \times p$

matrix $\boldsymbol{X}$, $\boldsymbol{X}^-$ means the $n \times (p-1)$ matrix obtained by removing the first column from $\boldsymbol{X}$; $(\boldsymbol{X})_{j\cdot}$, $j = 1, ..., n$, denotes the j^{th} row of $\boldsymbol{X}$ (similarly for $\boldsymbol{X}^-$); and $S_{\boldsymbol{X}} = \text{span}\{(\boldsymbol{X})_{1\cdot}, ..., (\boldsymbol{X})_{n\cdot}\}$, and $S_{\boldsymbol{X}^-} = \text{span}\{(\boldsymbol{X}^-)_{1\cdot}, ..., (\boldsymbol{X}^-)_{n\cdot}\}$ are the row spaces of $\boldsymbol{X}$ and $\boldsymbol{X}^-$.

Theorem 4.2.5 *For any $n \times p$ matrix $\boldsymbol{X}$*

$$\begin{aligned} rank(\boldsymbol{X}) = p-1 &\iff dim(S_{\boldsymbol{X}}) = p-1 \\ &\iff S_{\boldsymbol{X}} = H_{p-1}(\boldsymbol{\alpha}, 0) \\ &\iff (\boldsymbol{X})_{j\cdot} \in H_{p-1}(\boldsymbol{\alpha}, 0) \textit{ for all } j = 1, ..., n, \end{aligned} \tag{4.15}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^p$.

Corollary 4.2.6 *For any $n \times p$ matrix $\boldsymbol{X}$ where the first column of $\boldsymbol{X}$ is a column of 1's we have*

$$\begin{aligned} rank(\boldsymbol{X}) = p-1 &\Longrightarrow dim(S_{\boldsymbol{X}}) = p-1 \\ &\Longrightarrow S_{\boldsymbol{X}^-} = H_{p-2}(\boldsymbol{\alpha}, 0) \\ &\Longrightarrow (\boldsymbol{X}^-)_{j\cdot} \in H_{p-2}(\boldsymbol{\alpha}, 0) \textit{ for all } j = 1, ..., n, \end{aligned} \tag{4.16}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^{p-1}$.

4.3 Identifiability of MLMMs

This section presents two Theorems, Theorem 4.3.2 and Theorem 4.3.4, giving two different sufficient conditions for identifiability of Model 1, and a Corollary (Corollary 4.3.3) applying the result of Theorem 4.3.4 to MLMMs with a simple within-unit error covariance structure. To motivate the choice of the sufficient conditions, preceding the proofs we give some specific counter examples to identifiability in MLMMs, and discuss the considerations that led to them. As a starting point we find it instructive to consider the identifiability problem presented by Model 2 which has been addressed by Hennig (2000), not least because the logic used there will be used instrumentally in the proof of Theorem 4.3.4, but also because the logic gives valuable insight which is employed to good use in Theorem 4.3.2.

Compared to Model 2, for Model 1 we have the additional problem of simultaneously identifying the covariance matrices parametrized by $\boldsymbol{\zeta}$. However Model 1 is a mixture of LMM model distribution functions $F_i^1(\cdot|\boldsymbol{\theta})$ which are n_i-dimensional normal distributions with mean vectors $\boldsymbol{X}_i\boldsymbol{\beta}$, and covariance matrices $\boldsymbol{V}_i(\boldsymbol{\zeta})$ which depend on $\boldsymbol{Z}_i$ and not on $\boldsymbol{X}_i$. Consequently although the additional challenge of identifying $\boldsymbol{\zeta}$ should lead to stronger sufficient conditions for identifiability of Model 1 than for Model 2, it should not make the identification of the fixed effects any more complicated. For this reason we might expect the problem of identification of the mean vectors parametrized by $\boldsymbol{\beta}$, and how this relates to the identifiability of the mixture, to be fundamentally the same as the corresponding identifiability problem for Model 2. With this in mind we relate the theory developed by (Hennig, 2000) to the MLMM defined in Model 1, and use this insight to derive counter examples for Model 1 where the regression parameter $\boldsymbol{\beta}$ is not identified. The result of (Hennig, 2000) states (in the notation we are using in this thesis)

Theorem 4.3.1 *(Hennig 2000) For Model 2 with an intercept, let h be the minimum number of $(p-2)$-dimensional hyperplanes that cover the $\boldsymbol{x}_i^-$, $i \in I$, then Model 2 is identifiable with respect to $\mathscr{D}_2$ if $|S(J)| < h$.*

In particular, for $p = 2$ then each $\boldsymbol{x}_i^-$ is a scalar which lies on the 0-dimensional hyperplane defined by itself, and so we can cover the N scalars with a minimum number of $(p-2)$-dimensional hyperplanes $h \leq N$ (some $\boldsymbol{x}_i$ may be the same). Thus if we consider Model 2 with $|S(J)| = N$ then $h \leq |S(J)|$, and so the resulting MLM will not be identifiable. This type of counter example was used by Hennig for Model 2, and we now obtain similar counter examples for MLMMs.

In order to relate Theorem 4.3.1 to Model 1, we introduce some notation. For the purposes of the following discussion we will alternate between the general form of Model 1 which does not necessarily contain an intercept in the $\boldsymbol{X}_i$ matrices, and assuming Model 1 has an intercept, in which case we will assume the first column of every $\boldsymbol{X}_i$ contains the required column of ones. We will use the notation $\boldsymbol{X}_i^-$ to denote $\boldsymbol{X}_i$ with the first column removed, and $(\boldsymbol{X}_i)_{j\cdot}$, $j = 1, ..., n_i$, to denote the j^{th} row of $\boldsymbol{X}_i$ (similarly for $\boldsymbol{X}_i^-$). We will also need Corollary (4.2.6), which relates hyperplanes with the rank of a matrix.

Assuming Model 1 has an intercept, and for all $i \in I$ that $\mathrm{rank}(\boldsymbol{X}_i) = p - 1$, then from Corollary (4.2.6) we know that all the rows of each $\boldsymbol{X}_i^-$ will lie on their own common $(p-2)$-dimensional hyperplane. That is for all $i \in I$ we have

$$\mathrm{rank}(\boldsymbol{X}_i) = p - 1 \Longrightarrow (\boldsymbol{X}_i^-)_{j\cdot} \in H_{p-2}(\boldsymbol{\alpha}_i, 0) \text{ for all } j = 1, ..., n_i, \tag{4.17}$$

for some $\boldsymbol{\alpha}_i \in \mathbb{R}^{p-1}$. Setting $N = |S(J)|$, and using (4.17) together with Theorem 4.3.1 implies that all the rows of $\tilde{X} := (\boldsymbol{X}_1^\intercal, ..., \boldsymbol{X}_N^\intercal)^\intercal$ lie on no more than N common $(p-2)$-dimensional hyperplanes. Thus $h \leq N = |S(J)|$, and so $\boldsymbol{\beta} \in \mathbb{R}^p$ will not be identified by the rows of $\tilde{X}$. It is likely Theorem 4.3.1 can be used in the same way for the general version of Model 1, except we need the analogue result of (4.17) using Theorem (4.2.5)

$$\mathrm{rank}(\boldsymbol{X}_i) = p - 1 \Longleftrightarrow (\boldsymbol{X}_i)_{j\cdot} \in H_{p-1}(\boldsymbol{\alpha}_i, 0) \text{ for all } j = 1, ..., n_i. \tag{4.18}$$

We now present three counter examples motivated by the theory just described where lack of identifiability will be caused by a non-uniqueness of the parametrization of the mean vectors and covariance matrices. That is the problems will be caused by being able to find, for each $i \in I$, fixed effects $\boldsymbol{\beta} \neq \hat{\boldsymbol{\beta}}$ such that $\boldsymbol{X}_i\boldsymbol{\beta} = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}$, or covariance parameters $\boldsymbol{\zeta} \neq \hat{\boldsymbol{\zeta}}$ such that $\boldsymbol{V}_i(\boldsymbol{\zeta}) = \boldsymbol{V}_i(\hat{\boldsymbol{\zeta}})$ resulting in, for each $i \in I$, the equality of the distribution functions $F_i(\cdot|J)$ and $F_i(\cdot|\hat{J})$ for $J \neq \hat{J}$. We do not present the opposite type of counter examples where the mean vectors and covariance matrices are identified, but where these distribution function equalities still hold due the sums of the component distribution functions not being uniquely defined.

Counter example 1

For Model 1 with $N = G = 2$, $n_i = 6$ for $i = 1, 2$, and $p = 3$, consider the following choices of covariate data and regression parameters;

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 4 & 8 \\ 1 & 3 & 6 \\ 1 & 2.3 & 4.6 \\ 1 & 3.4 & 6.8 \\ 1 & 8 & 16 \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} 1 & -6 & -33 \\ 1 & 3 & 16.5 \\ 1 & -6.4 & -35.2 \\ 1 & 9 & 49.5 \\ 1 & 5 & 27.5 \\ 1 & 12 & 66 \end{bmatrix},$$

$$\boldsymbol{\beta}_1 = \begin{bmatrix} 5 \\ 12 \\ -6 \end{bmatrix}, \quad \boldsymbol{\beta}_2 = \begin{bmatrix} 5 \\ -13 \\ 3 \end{bmatrix}, \quad \hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} 5 \\ 1 \\ -4 \end{bmatrix}, \quad \hat{\boldsymbol{\beta}}_2 = \begin{bmatrix} 5 \\ -2 \\ 1 \end{bmatrix}.$$

We have $\text{rank}(\boldsymbol{X}_1) = \text{rank}(\boldsymbol{X}_2) = 2 = p - 1$, and apart from small rounding errors

these covariate data and regression parameters produce the following mean vectors

$$\boldsymbol{X}_1\boldsymbol{\beta}_1 = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_2 = \begin{bmatrix} 5 \\ 5 \\ 5 \\ 5 \\ 5 \\ 5 \end{bmatrix}, \quad \boldsymbol{X}_1\boldsymbol{\beta}_2 = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} -9 \\ -23 \\ -16 \\ -11.1 \\ -18.8 \\ -51 \end{bmatrix},$$

$$\boldsymbol{X}_2\boldsymbol{\beta}_1 = \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} 131 \\ -58 \\ 139.4 \\ -184 \\ -100 \\ -247 \end{bmatrix}, \quad \boldsymbol{X}_2\boldsymbol{\beta}_2 = \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \begin{bmatrix} -16 \\ 15.5 \\ -17.4 \\ 36.5 \\ 22.5 \\ 47 \end{bmatrix},$$

for $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 \neq \hat{\boldsymbol{\beta}}_1 \neq \hat{\boldsymbol{\beta}}_2$. For $j = 1, 2$, let $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \hat{\boldsymbol{\zeta}}_1, \hat{\boldsymbol{\zeta}}_2, \boldsymbol{\zeta} \in \Psi_{\boldsymbol{\zeta}}$ such that $\boldsymbol{\zeta}_j = \hat{\boldsymbol{\zeta}}_j = \boldsymbol{\zeta}$ for all j, so that $\boldsymbol{V}_i(\boldsymbol{\zeta}_j) = \boldsymbol{V}_i(\hat{\boldsymbol{\zeta}}_j)$ for all i and all j. Let $S(J) = \{(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1), (\boldsymbol{\beta}_2, \boldsymbol{\zeta}_2)\}$, $S(\hat{J}) = \{(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1), (\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\zeta}}_2)\}$, and for all j set $J\{(\boldsymbol{\beta}_j, \boldsymbol{\zeta}_j)\} = \hat{J}\{(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\zeta}}_j)\} = 1/2$. Then for each i we have

$$\begin{aligned} F_i(\cdot|J) &= \left(\frac{1}{2}\right) \Phi_{\boldsymbol{X}_i\boldsymbol{\beta}_1, \boldsymbol{V}_i(\boldsymbol{\zeta}_1)}(\cdot) + \left(\frac{1}{2}\right) \Phi_{\boldsymbol{X}_i\boldsymbol{\beta}_2, \boldsymbol{V}_i(\boldsymbol{\zeta}_2)}(\cdot) \\ =F_i(\cdot|\hat{J}) &= \left(\frac{1}{2}\right) \Phi_{\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_1, \boldsymbol{V}_i(\hat{\boldsymbol{\zeta}}_1)}(\cdot) + \left(\frac{1}{2}\right) \Phi_{\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_2, \boldsymbol{V}_i(\hat{\boldsymbol{\zeta}}_2)}(\cdot), \end{aligned} \tag{4.19}$$

and so $F(\cdot|\hat{J}) = F(\cdot|\hat{J})$ for $J \neq \hat{J}$. Thus $\mathscr{D}_1$ is not identifiable with respect to Ω_1. Note however that ignoring either component 1 or component 2 of this model gives two examples of 1-component models that are not counter examples to identifiability. So we see that as for MLMs, 1-component model identifiability is not a sufficient condition for the identifiability of the mixture.

Counter example 2

For this counter example we will not include an intercept in the $\boldsymbol{X}_i$ matrices. Again assume the same set up as for counter example 1, but consider the following choices of covariate data and regression parameters;

$$X_1 = \begin{bmatrix} -9.1 & 2.3 & 4.6 \\ 4.9 & 1.7 & 3.4 \\ 5.4 & -3.7 & -7.4 \\ 6.1 & -23.4 & -46.8 \\ -7.1 & -16.8 & -33.6 \\ 6.1 & 10.4 & 20.8 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 6.7 & 9.1 & 50.05 \\ -4.8 & 2.9 & 15.95 \\ -3.9 & -0.47 & -2.585 \\ 7.8 & 0.63 & 3.465 \\ -18.9 & -1.36 & -7.48 \\ 56.1 & 21.9 & 120.45 \end{bmatrix},$$

$$\boldsymbol{\beta}_1 = \begin{bmatrix} 9.1 \\ 29.9 \\ -19.9 \end{bmatrix}, \quad \boldsymbol{\beta}_2 = \begin{bmatrix} 9.1 \\ -68.3 \\ 6.1 \end{bmatrix}, \quad \hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} 9.1 \\ -42.7 \\ -6.7 \end{bmatrix}, \quad \hat{\boldsymbol{\beta}}_2 = \begin{bmatrix} 9.1 \\ 4.3 \\ -7.1 \end{bmatrix}.$$

We have $\text{rank}(\boldsymbol{X}_1) = \text{rank}(\boldsymbol{X}_2) = 2 = p - 1$, and apart from small rounding errors these data produce the following mean vectors

$$\boldsymbol{X}_1\boldsymbol{\beta}_1 = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_2 = \begin{bmatrix} -105.58 \\ 27.76 \\ 85.77 \\ 287.17 \\ 101.71 \\ -47.45 \end{bmatrix}, \quad \boldsymbol{X}_1\boldsymbol{\beta}_2 = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} -211.84 \\ -50.78 \\ 256.71 \\ 1368.2 \\ 877.87 \\ -527.93 \end{bmatrix},$$

$$\boldsymbol{X}_2\boldsymbol{\beta}_1 = \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} -662.935 \\ -274.375 \\ 1.8985 \\ 20.8635 \\ -63.8020 \\ -1231.6 \end{bmatrix}, \quad \boldsymbol{X}_2\boldsymbol{\beta}_2 = \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \begin{bmatrix} -255.255 \\ -144.455 \\ -19.1575 \\ 49.0875 \\ -124.73 \\ -250.5150 \end{bmatrix},$$

for $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 \neq \hat{\boldsymbol{\beta}}_1 \neq \hat{\boldsymbol{\beta}}_2$. Again (4.19) holds, and so $\mathscr{D}_1$ is not identifiable with respect to Ω_1. Once again ignoring either component 1 or component 2 of this model gives two examples of 1-component models that are not counter examples to identifiability.

The two counter examples presented above relied on both the $\boldsymbol{X}_i$ matrices having rank equal to $p - 1$ in order to use Theorems 4.3.1, 4.2.5 and Corollary (4.2.6) as a guide to how to avoid models that we know are identifiable. It is perhaps not surprising that the rank of the design matrices plays such a key role since the rank of matrices is fundamental to the identification of vectors in systems of homogeneous equations. Specifically we have that if $\boldsymbol{A}$ is an $m \times n$ matrix, and $\boldsymbol{x} \in \mathbb{R}^n$, then $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ has non-trivial solutions if and only if $\text{rank}(\boldsymbol{A}) < n$. Consequently the existence of an $i \in I$ such that $\text{rank}(\boldsymbol{X}_i) = p$ implies $\boldsymbol{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ if and only if $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ for all $\boldsymbol{\beta}, \hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, which means $\boldsymbol{\beta}$ is identified. Now for any $i \in I$, $\text{rank}(\boldsymbol{X}_i) = \min(n_i, p)$, and so $\text{rank}(\boldsymbol{X}_i) = p$

implies $\boldsymbol{X}_i$ is full rank, and so for Model 1, and regardless of whether the $\boldsymbol{X}_i$ matrices have an intercept or not, at least one $\boldsymbol{X}_i$ being of full rank will be sufficient to prevent the counter examples above from working. It should also be clear that $n_i < p$ for all $i \in I$ implies $n_i = \operatorname{rank}(\boldsymbol{X}_i) < p$, which means $\boldsymbol{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ has non-trivial solutions for all units. Thus $\boldsymbol{\beta}$ is never identified if for all units there are less rows in $\boldsymbol{X}_i$ than columns.

In light of the above discussion it is likely that a sufficient condition for identifiability for Model 1 can be formulated that demands the existence of a single unit $i \in I$ that, by having a full rank design matrix $\boldsymbol{X}_i$, can alone identify the fixed effects $\boldsymbol{\beta}$ through the parametrization of the mean vector $\boldsymbol{X}_i\boldsymbol{\beta}$. This rank condition is a fairly restrictive, for example the design matrix of the identifying unit cannot contain any column that is constant for all rows when an intercept is included in the model. We will discuss this in more detail later.

We now use the ideas we have just discussed, in particular the concept of the rank of a matrix, to search for rank conditions we can impose on one of the matrices $\boldsymbol{Z}_i$ such that a single unit identifies $\boldsymbol{\zeta}$ through the parametrization of the covariance matrix $\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\intercal} + \sigma^2\boldsymbol{C}_i(\boldsymbol{\phi})$. For a MLMM with a simple error covariance structure we can deduce the conditions we need in order to ensure the existence of this identifying unit for $\boldsymbol{\zeta}$, and we do this by considering one unit from our sample, $\boldsymbol{Y}_1$ say. Denote the family of distribution functions for this sample of one to be $\mathscr{D}^1(1)$. From Theorem 4.2.4 we have that if $\boldsymbol{X}_1$ and $\boldsymbol{Z}_1$ are both full rank, and if $n_1 > q$ (this implies $\operatorname{rank}(\boldsymbol{Z}_i) = q$), then $\mathscr{D}^1(1)$ is identifiable with respect to Ψ_1. Since the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ are independent of each other in the normal distribution, this result gives the conditions we seek: let $\boldsymbol{Z}$ be any $n \times q$ matrix, and define $\boldsymbol{V}(\boldsymbol{\zeta}) = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^{\intercal} + \sigma^2\boldsymbol{I}_n$, then

$$\forall \boldsymbol{\zeta}, \hat{\boldsymbol{\zeta}} \in \Psi_{\boldsymbol{\zeta}} : \quad \boldsymbol{V}(\boldsymbol{\zeta}) = \boldsymbol{V}(\hat{\boldsymbol{\zeta}}) \Leftrightarrow \boldsymbol{\zeta} = \hat{\boldsymbol{\zeta}} \text{ if and only if } \operatorname{rank}(\boldsymbol{Z}) = q \text{ and } n > q. \quad (4.20)$$

We will use the above result in conjunction with Theorem 4.3.2 later on in this section. We now present a counter example to identifiability where the conditions of (4.20) are not satisfied.

<u>Counter example 3</u>

We use Model 1 with a simple within unit error covariance structure. Choose $N = G =$

2, $n_i = 3$ for $i = 1, 2$, $q = 2$, leave p to be arbitrary, and consider the following choices of covariate data and covariance parameters

$$\boldsymbol{D}_1 = \begin{bmatrix} 5.0455 & -1.8503 \\ -1.8503 & 4.1440 \end{bmatrix}, \quad \boldsymbol{D}_2 = \begin{bmatrix} 4 & -1.5678 \\ -1.5678 & 6.3021 \end{bmatrix}, \quad \sigma_1^2 = \sigma_2^2 = 9,$$

$$\hat{\boldsymbol{D}}_1 = \begin{bmatrix} 4.2049 & -1.6232 \\ -1.6232 & 5.8791 \end{bmatrix}, \quad \hat{\boldsymbol{D}}_2 = \begin{bmatrix} 5.5997 & -2 \\ -2 & 3 \end{bmatrix}, \quad \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = 9,$$

$$\boldsymbol{Z}_1 = \begin{bmatrix} 0.33215 & -0.27871 \\ -0.26247 & 0.22023 \\ -2.74121 & 2.30015 \end{bmatrix}, \quad \boldsymbol{Z}_2 = \begin{bmatrix} 2.39840 & 1.38472 \\ -1.16904 & -0.67494 \\ 2.62832 & 1.51746 \end{bmatrix},$$

where all the random effects covariance matrices are positive definite as required, but where $\operatorname{rank}(\boldsymbol{Z}_1) = \operatorname{rank}(\boldsymbol{Z}_2) = 1 < q$, which means both matrices are rank deficient. For $j = 1, 2$, let $\zeta_j = (\mathrm{v}(\boldsymbol{D}_j)^\intercal, \sigma_j^2)^\intercal$, $\hat{\zeta}_j = (\mathrm{v}(\hat{\boldsymbol{D}}_j)^\intercal, \hat{\sigma}_j^2)^\intercal$, then the above data give the following positive definite covariance matrices for the response vectors

$$\boldsymbol{V}_1(\boldsymbol{\zeta}_1) = \boldsymbol{V}_1(\hat{\boldsymbol{\zeta}}_2) = \boldsymbol{V}_1(\boldsymbol{\zeta}_2) = \boldsymbol{V}_1(\hat{\boldsymbol{\zeta}}_1) = \begin{bmatrix} 10.8777 & 6.4044 & -7.8885 \\ 6.4044 & 30.8442 & -26.9061 \\ -7.8885 & -26.9061 & 42.1410 \end{bmatrix},$$

$$\boldsymbol{V}_2(\boldsymbol{\zeta}_1) = \boldsymbol{V}_2(\hat{\boldsymbol{\zeta}}_1) = \boldsymbol{V}_2(\boldsymbol{\zeta}_2) = \boldsymbol{V}_2(\hat{\boldsymbol{\zeta}}_2) = \begin{bmatrix} 11.3919 & 0.8843 & -3.6278 \\ 0.8843 & 9.3270 & -1.3413 \\ -3.6278 & -1.3413 & 14.5025 \end{bmatrix}.$$

Let $S(J) = \{(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1), (\boldsymbol{\beta}_2, \boldsymbol{\zeta}_2)\}$, $S(\hat{J}) = \{(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1), (\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\zeta}}_2)\}$. For all j let $J\{(\boldsymbol{\beta}_j, \boldsymbol{\zeta}_j)\} = \hat{J}\{(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\zeta}}_j)\} = 1/2$, and $\boldsymbol{\beta}_j = \hat{\boldsymbol{\beta}}_j = \boldsymbol{\beta}$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$, so that $\boldsymbol{X}_i\boldsymbol{\beta}_j = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_j$ for all i.

Then we again get the equality in (4.19), and so the MLMM is not identifiable.

By using (4.20) it is clear we can prevent counter example 3 from working if we demand rank($\boldsymbol{Z}_i$) $= q$, and $n_i > q$, for at least one $i \in I$, and so it is likely these two conditions can be used to form part of a sufficient condition for identifiability for Model 1 with a simple error covariance structure. For Model 1 in general (i.e. with an autoregressive within unit error covariance structure) it remains to determine under what conditions this "identifying" unit can be guaranteed to exist. Unlike counter examples 1 and 2, for counter example 3 we note that ignoring either of the two components give counter examples to identifiability for a 1-component model, which is not surprising since we see that the $\boldsymbol{Z}_i$ matrices do not meet the assumptions of theorem (4.2.4).

In counter example 3 by choosing the fixed effects to be equal, the MLMM used there was a particular example (2 units, 2 within-unit observations, 2 random effects) of a MLMM where there are no fixed effects parameters for the component distributions but rather mean vectors $\mu_{ig} := \boldsymbol{X}_i \boldsymbol{\beta}_g$, for all $i \in I$, and $g \in I_G$, which are always identified. Specifically for each $i = 1, 2$ we set $\mu_{ig} = \hat{\mu}_{ig} = c_i$ for all $g = 1, 2$. For this particular MLMM a detailed analysis shows that any counter example necessarily involves both $\boldsymbol{Z}_1$, and $\boldsymbol{Z}_2$ being rank deficient, and hence any counter example to identifiability will also yield counter examples to identifiability for the 1-component model, or equivalently an identifiable 1-component model implies an identifiable MLMM. Since non-identifiability of any 1-component model implies non-identifiability of the MLMM, or equivalently that identifiability of any MLMM implies the identifiability of the 1-component model, then for this particular MLMM with no regression components we have that the 1-component model is identifiable if and only if the MLMM is identifiable.

From the above discussion we see for that particular MLMM with only the covariance matrices parametrized that no further identifiability problems are introduced by the mixture than are already encountered in the 1-component model. The reason this is not the case for the examples of MLMMs with regression parameters we have considered in counter examples 1 and 2 is that we can have both design matrices being rank deficient, yet the matrix $\tilde{\boldsymbol{X}}$ formed by stacking both one on top of the other is still of full rank, thus ensuring the identifiability of the 1-component model.

No analogy of this combining of information from the individual units to avoid identifiability problems can be employed with the covariance matrices. For example let $\tilde{\boldsymbol{V}}(\boldsymbol{\zeta}) := \text{diag}\{\boldsymbol{V}_1(\boldsymbol{\zeta}), \boldsymbol{V}_2(\boldsymbol{\zeta})\}$ where rank($\boldsymbol{V}_1(\boldsymbol{\zeta})$) $< (n_1 + n_2)$ and rank($\boldsymbol{V}_2(\boldsymbol{\zeta})$) $<$

$(n_1 + n_2)$. Then from Theorem 2.12 (Schott, 2005) we have that the rank of the $2(n_1+n_2) \times 2(n_1+n_2)$ matrix $\tilde{\boldsymbol{V}}(\boldsymbol{\zeta})$ is the sum of the ranks of the two diagonal matrices, which must be less than $2(n_1 + n_2)$. Thus a single unit having a rank deficient matrix prevents the whole matrix from having full rank. It should be clear that retaining units 1 and 2, but adding more covariance matrices will not overcome this problem. Now suppose $\tilde{\boldsymbol{V}}(\hat{\boldsymbol{\zeta}}) := \text{diag}\{\boldsymbol{V}_1(\hat{\boldsymbol{\zeta}}), \boldsymbol{V}_2(\hat{\boldsymbol{\zeta}})\}$ for $\boldsymbol{\zeta} \neq \hat{\boldsymbol{\zeta}}$. It is clear the covariance parameter is identified even if a single unit identifies the parameter (we show later this occurs if the $\boldsymbol{Z}_i$ matrix is full rank). This stands in contrast to the fixed effects where the whole sample can identify the parameter whilst every unit might not.

In summary, for simple 2-component MLMMs we have discussed counter examples to identifiability of a certain type - those that involve lack of identification of $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ through the parametrization of the mean vector and the covariance matrix respectively. We have seen that the counter examples presented cannot occur if we demand that at least one unit identifies $\boldsymbol{\beta}$, and at least one unit identifies $\boldsymbol{\zeta}$. For MLMMs with a simple error covariance structure we have also discussed conditions that can be imposed on the fixed sample data $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, and n_i, for $i \in I$, to ensure the existence of such units. With these things in mind we present Theorem 4.3.2 which gives sufficient conditions for the identifiability for Model 1, and Corollary (4.3.3) relating the theorem to a MLMM with a simple error covariance structure.

The key result we will use repeatedly in the proof of Theorem 4.3.2 is that of Yakowitz and Spragins (1968) which states that mixtures of multivariate normal distributions (i.e. normal mixtures without regression and covariance parameters) are identifiable. To see how we use this result, consider unit i, and the mixing distributions J and $\hat{J}$ with support sets $S(J)$ and $S(\hat{J})$ respectively. For say $(\boldsymbol{\beta}', \boldsymbol{\zeta})' \in \Psi_1$, let $A_i(\boldsymbol{\beta}', \boldsymbol{\zeta}')$ be the set of all parameters in Ψ_1 that give rise to the mean vector $\boldsymbol{X}_i\boldsymbol{\beta}'$ and covariance matrix $\boldsymbol{V}_i(\boldsymbol{\zeta}')$, i.e.

$$A_i(\boldsymbol{\beta}', \boldsymbol{\zeta}') = \{(\boldsymbol{\beta}, \boldsymbol{\zeta}) : \boldsymbol{X}_i\boldsymbol{\beta} = \boldsymbol{X}_i\boldsymbol{\beta}', \boldsymbol{V}_i(\boldsymbol{\zeta}) = \boldsymbol{V}_i(\boldsymbol{\zeta}'), (\boldsymbol{\beta}, \boldsymbol{\zeta}) \in \Psi_1\}. \tag{4.21}$$

Since the result of Yakowitz and Spragins (1968) means that mean vectors and covariance matrices are identified, then $J = \hat{J}$ implies $J(A_i(\boldsymbol{\beta}', \boldsymbol{\zeta}')) = \hat{J}(A_i(\boldsymbol{\beta}', \boldsymbol{\zeta}'))$. In the proof of Theorem 4.3.2 we will simply say $J(A_i(\boldsymbol{\beta}', \boldsymbol{\zeta}')) = \hat{J}(A_i(\boldsymbol{\beta}', \boldsymbol{\zeta}'))$ follows by identifiability of multivariate normal mixtures for unit i.

Finally in the proof of Theorem 4.3.2 we shall refer to the following equivalence relation on Ψ_1

$$(\boldsymbol{\beta}, \boldsymbol{\zeta}) \overset{\zeta}{\sim} (\boldsymbol{\beta}', \boldsymbol{\zeta}') : \quad \boldsymbol{\zeta} = \boldsymbol{\zeta}' \quad \forall (\boldsymbol{\beta}, \boldsymbol{\zeta}), (\boldsymbol{\beta}', \boldsymbol{\zeta}') \in \Psi_1, \tag{4.22}$$

where we will denote by $[(\boldsymbol{\beta}', \boldsymbol{\zeta}')]^{\zeta}$ the equivalence class of $(\boldsymbol{\beta}', \boldsymbol{\zeta}')$ in Ψ_1 under $\overset{\zeta}{\sim}$. The collection of all equivalence classes within a set $A \subseteq \Psi_1$ under $\overset{\zeta}{\sim}$ will be denoted by $A/\overset{\zeta}{\sim}$. This is known as the quotient set of A under $\overset{\zeta}{\sim}$, and forms a partition of A. Finally $|A/\overset{\zeta}{\sim}|$ shall mean the number of equivalence classes under $\overset{\zeta}{\sim}$ that partition the set A. We are now in a position to present Theorem 4.3.2.

Theorem 4.3.2 *$\mathscr{D}_1$ is identifiable with respect to Ω_1 according to definition (4.2.1) if for all $\boldsymbol{\beta}, \hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ we have ($\boldsymbol{X}_i\boldsymbol{\beta} = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ if and only if $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$) for at least one $i \in I$, and for all $\boldsymbol{\zeta}, \hat{\boldsymbol{\zeta}} \in \Psi_{\boldsymbol{\zeta}}$ we have ($\boldsymbol{V}_i(\boldsymbol{\zeta}) = \boldsymbol{V}_i(\hat{\boldsymbol{\zeta}})$ if and only if $\boldsymbol{\zeta} = \hat{\boldsymbol{\zeta}}$) for at least one $i \in I$.*

Proof. Let unit $j \in I$ be the unit that satisfies $\boldsymbol{X}_j\boldsymbol{\beta} = \boldsymbol{X}_j\hat{\boldsymbol{\beta}}$ if and only if $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, and unit $k \in I$ be the unit that satisfies $\boldsymbol{V}_k(\boldsymbol{\zeta}) = \boldsymbol{V}_k(\hat{\boldsymbol{\zeta}})$ if and only if $\boldsymbol{\zeta} = \hat{\boldsymbol{\zeta}}$. Let $J, \hat{J} \in \Omega_1$, and assume $J = \hat{J}$. Since the normal distribution is completely determined by its mean vector and covariance matrix, this implies $F_i(\cdot|J) = F_i(\cdot|\hat{J})$ for all $i \in I$. Thus $J = \hat{J} \Rightarrow F(\cdot|J) = F(\cdot|\hat{J})$. Identifiability will follow if we can show $F(\cdot|J) = F(\cdot|\hat{J}) \Rightarrow J = \hat{J}$, or equivalently $F_i(\cdot|J) = F_i(\cdot|\hat{J})\ \forall i \in I \Rightarrow J = \hat{J}$.

Assume first that $j = k$, $F_i(\cdot|J) = F_i(\cdot|\hat{J})\ \forall i \in I$. Without loss of generality take $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1) \in S(J)$, and assume $|S(J)| \leq |S(\hat{J})|$. By the identifiability of multivariate normal mixtures for unit j we have

$$\begin{aligned} &J\{(\boldsymbol{\beta}, \boldsymbol{\zeta}) : \boldsymbol{X}_j\boldsymbol{\beta} = \boldsymbol{X}_j\boldsymbol{\beta}_1, \boldsymbol{V}_j(\boldsymbol{\zeta}) = \boldsymbol{V}_j(\boldsymbol{\zeta}_1), (\boldsymbol{\beta}, \boldsymbol{\zeta}) \in \Psi_1\} \\ =&\hat{J}\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) : \boldsymbol{X}_j\hat{\boldsymbol{\beta}} = \boldsymbol{X}_j\boldsymbol{\beta}_1, \boldsymbol{V}_j(\hat{\boldsymbol{\zeta}}) = \boldsymbol{V}_j(\boldsymbol{\zeta}_1), (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in \Psi_1)\} \\ \Longleftrightarrow &J\{(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)\} = \hat{J}\{(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)\}. \end{aligned} \tag{4.23}$$

Now $\hat{J}\{(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)\} > 0$ since $J\{(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)\} > 0$, which implies $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1) \in S(\hat{J})$. This result, and the last equality in (4.23), must apply to all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J)$ since $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)$ was picked arbitrarily from $S(J)$. Thus we have $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(\hat{J})$, and $J\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\} = \hat{J}\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\}$ for all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J)$. Our assumption $|S(J)| \leq |S(\hat{J})|$ then implies $S(J) \subseteq S(\hat{J})$. Now repeat all of the above arguments starting from the paragraph above (4.23) but reverse the roles of J and $\hat{J}$. We then get $S(\hat{J}) \subseteq S(J)$, where $J\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})\} = \hat{J}\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})\}$ for all $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in S(\hat{J})$. This then implies $S(J) = S(\hat{J})$, where $J\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\} = \hat{J}\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\}$ for all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J)$, and so $J = \hat{J}$.

For $j \neq k$, assume $F_i(\cdot|J) = F_i(\cdot|\hat{J})\ \forall i \in I$, and without loss of generality take $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1) \in S(J)$. By the identifiability of multivariate normal mixtures for unit k we have

$$\begin{aligned} & J\{(\boldsymbol{\beta}, \boldsymbol{\zeta}) : \boldsymbol{X}_k\boldsymbol{\beta} = \boldsymbol{X}_k\boldsymbol{\beta}_1, \boldsymbol{V}_k(\boldsymbol{\zeta}) = \boldsymbol{V}_k(\boldsymbol{\zeta}_1), (\boldsymbol{\beta}, \boldsymbol{\zeta}) \in \Psi_1\} \\ =& \hat{J}\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) : \boldsymbol{X}_k\hat{\boldsymbol{\beta}} = \boldsymbol{X}_k\boldsymbol{\beta}_1, \boldsymbol{V}_k(\hat{\boldsymbol{\zeta}}) = \boldsymbol{V}_k(\boldsymbol{\zeta}_1), (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in \Psi_1\} \\ \Longleftrightarrow & J\{[(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\zeta}\} = \hat{J}\{[(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\zeta}\}. \end{aligned} \tag{4.24}$$

Now because $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1) \in [(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\zeta}$ we must have $J\{[(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\zeta}\} > 0$, which implies $\hat{J}\{[(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\zeta}\} > 0$, and so $[(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\zeta} \cap S(\hat{J}) \neq \emptyset$. Furthermore this must hold for all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J)$ since $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)$ was picked arbitrarily from $S(J)$. Thus we have

$$\forall (\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J),\ \exists (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in S(\hat{J}) \text{ such that } (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \overset{\zeta}{\sim} (\boldsymbol{\beta}, \boldsymbol{\zeta}). \tag{4.25}$$

The possibility remains that the converse of (4.25) does not hold, that is to say there may be some points in $S(\hat{J})$ not equivalent under $\overset{\zeta}{\sim}$ to any points in $S(J)$. We now show however that this cannot be true. Without loss of generality, assume for $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1) \in S(\hat{J})$ that there does not exist a $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J)$ such that $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \overset{\zeta}{\sim} (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)$. Once again by the identifiability of multivariate normal mixtures for unit k we have

$$\begin{aligned} & J\{(\boldsymbol{\beta}, \boldsymbol{\zeta}) : \boldsymbol{X}_k\boldsymbol{\beta} = \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_1, \boldsymbol{V}_k(\boldsymbol{\zeta}) = \boldsymbol{V}_k(\hat{\boldsymbol{\zeta}}_1), (\boldsymbol{\beta}, \boldsymbol{\zeta}) \in \Psi_1\} \\ =& \hat{J}\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) : \boldsymbol{X}_k\hat{\boldsymbol{\beta}} = \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_1, \boldsymbol{V}_k(\hat{\boldsymbol{\zeta}}) = \boldsymbol{V}_k(\hat{\boldsymbol{\zeta}}_1), (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in \Psi_1\} \\ \Longleftrightarrow & J\{[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\zeta}\} = \hat{J}\{[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\zeta}\}. \end{aligned} \tag{4.26}$$

Now because $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1) \in [(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\zeta}$ we must have $\hat{J}\{[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\zeta}\} > 0$, which implies $J\{[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\zeta}\} > 0$, and so $[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\zeta} \cap S(J) \neq \emptyset$. This contradicts our assumption that no point in $S(J)$ is equivalent to $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)$ under $\overset{\zeta}{\sim}$, and so the converse of (4.25) holds

$$\forall(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in S(\hat{J}),\ \exists(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J) \text{ such that } (\boldsymbol{\beta}, \boldsymbol{\zeta}) \overset{\zeta}{\sim} (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}), \tag{4.27}$$

and so (4.25) and (4.27) together imply $S(J)$ and $S(\hat{J})$ are partitioned into the same number of equivalence classes under $\overset{\zeta}{\sim}$, i.e. we have $s(\boldsymbol{\zeta}) := |S(J)/\overset{\zeta}{\sim}| = |S(\hat{J})/\overset{\zeta}{\sim}|$.

The equations (4.25), and (4.27) establish a relationship between $S(J)$, and $S(\hat{J})$ in terms of equivalence classes with respect to $\overset{\zeta}{\sim}$. We now turn our attention to unit j that identifies $\boldsymbol{\beta}$. Let $\{A_1, ..., A_{s(\boldsymbol{\zeta})}\}$ and $\{\hat{A}_1, ..., \hat{A}_{s(\boldsymbol{\zeta})}\}$ denote the $s(\boldsymbol{\zeta})$ equivalence classes that partition $S(J)$, and $S(\hat{J})$ respectively, and let $l = 1, ..., s(\boldsymbol{\zeta})$. We note that the definition of the support sets $S(J)$ and $S(\hat{J})$ does not explicitly forbid the presence of duplicate vectors, however standard set theory does, and hence A_l and $\hat{A}_l$ for all l cannot contain duplicate vectors. Without loss of generality assume $|S(J)| \leq |S(\hat{J})|$, and pick a $(\boldsymbol{\beta}', \boldsymbol{\zeta}') \in A_1$, which implies all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in A_1$ satisfy $\boldsymbol{\zeta} = \boldsymbol{\zeta}'$. By the identifiability of multivariate normal mixtures for unit j we have

$$\begin{aligned} &J\{(\boldsymbol{\beta}, \boldsymbol{\zeta}) : \boldsymbol{X}_j\boldsymbol{\beta} = \boldsymbol{X}_j\boldsymbol{\beta}', \boldsymbol{V}_j(\boldsymbol{\zeta}) = \boldsymbol{V}_j(\boldsymbol{\zeta}'), (\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J) \cap A_1\} \\ =&\hat{J}\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) : \boldsymbol{X}_j\hat{\boldsymbol{\beta}} = \boldsymbol{X}_j\boldsymbol{\beta}', \boldsymbol{V}_j(\hat{\boldsymbol{\zeta}}) = \boldsymbol{V}_j(\boldsymbol{\zeta}'), (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in S(\hat{J}) \cap A_1\} \\ \Longleftrightarrow &J\{(\boldsymbol{\beta}', \boldsymbol{\zeta}')\} = \hat{J}\{(\boldsymbol{\beta}', \boldsymbol{\zeta}')\}. \end{aligned} \tag{4.28}$$

Now $\hat{J}\{(\boldsymbol{\beta}', \boldsymbol{\zeta}')\} > 0$ since $J\{(\boldsymbol{\beta}', \boldsymbol{\zeta}')\} > 0$, which implies $(\boldsymbol{\beta}', \boldsymbol{\zeta}') \in S(\hat{J})$. This result, and (4.28), must hold for all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in A_1$ since $(\boldsymbol{\beta}', \boldsymbol{\zeta}')$ was picked arbitrarily from there. Thus $A_1 \subseteq S(\hat{J})$, and $J\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\} = \hat{J}\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\}$ for all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in A_1$. Repeating this argument for all A_l we get $A_l \subseteq S(\hat{J})$, and $J\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\} = \hat{J}\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\}$ for all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in A_l$. From our assumption $|S(J)| \leq |S(\hat{J})|$ we then get $S(J) = \cup_{l=1}^{s(\boldsymbol{\zeta})} A_l \subseteq S(\hat{J})$, and $J\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\} = \hat{J}\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\}$ for all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J)$.

Now assume $|S(J)| \geq |S(\hat{J})|$, and pick a $(\boldsymbol{\beta}', \boldsymbol{\zeta}') \in \hat{A}_1$. Repeating the above arguments starting from (4.28) gives $S(\hat{J}) = \cup_{l=1}^{s(\boldsymbol{\zeta})} \hat{A}_l \subseteq S(J)$, and $J\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})\} = \hat{J}\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})\}$ for all $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in S(\hat{J})$. Thus we must have $S(J) = S(\hat{J})$, and $J\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\} = \hat{J}\{(\boldsymbol{\beta}, \boldsymbol{\zeta})\}$ for all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J)$ which implies $J = \hat{J}$. ■

We note that the sufficient condition for identifiability in the above theorem requires the 1-component model to be identifiable by just a single unit, which is a much stronger condition than just requiring the 1-component model to be identifiable. Furthermore counter examples 1 and 2 are not counter examples to the above theorem, since in those examples the 1-component models were only identified for those particular choices of the fixed effects. The fact that the design matrices were rank deficient mean that this does not hold for the entire parameter space - i.e. we will be able to find choices of the fixed effects yielding counter examples to the 1-component model.

We also note that the analogy of Theorem 4.3.2 holds also for Model 2 when we make the obvious changes in notation. Thus $\mathscr{D}_2$ is identifiable with respect to Ω_2 if and only if $\boldsymbol{x}_i^\intercal \boldsymbol{\beta} = \boldsymbol{x}_i^\intercal \hat{\boldsymbol{\beta}}$ for at least one $i \in I$, that is if and only if the 1-component model is identifiable from the data from at least a single unit (but perhaps only a single unit). Viewing $\boldsymbol{x}_i^\intercal$ as a $1 \times p$ matrix, a sufficient condition for this to hold is $\boldsymbol{x}_i^\intercal$ being full rank, which implies $p = 1$. Thus we get the trivial result that a clusterwise regression model is identifiable if it contains only one variable, and if at least one $\boldsymbol{x}_i \in \mathbb{R}$ is non-zero which shows that the usefulness of Theorem 4.3.2 is largely dependent on the specific conditions which guarantee the existence of the identifying units. For Model 1 we have already discussed these conditions: for all versions of Model 1 we need $\text{rank}(\boldsymbol{X}_i) = p$ for at least one unit i (which implies $p \leq n_i$). Unlike for clusterwise regression this condition however does not in general impose a very restricted model, although as we will discuss shortly it does restrict the model much more than we would like. Furthermore, and again as we will discuss shortly, the rank condition in 4.3.2 on the covariance matrices also does not imply a very restricted model, which shows that the within-unit sample sizes for MLMMs being greater than 1 lead to a sufficient condition for identifiability which is more than just a trivial result. This illustrates the beneficial effect of having greater within-unit information, and in statistics more information is always better - identifiability problems are no exception.

The above point also illustrates, slightly counter-intuitively, that identifiability problems in MLMMs can be easier to characterise, and also to understand, than for the simpler clusterwise regression - i.e. if we include a unit in our dataset that has particularly "informative" data about the model parameters, then this unit alone can identify both the 1-component and the mixture model, and that the rest of our sample may not play a large part, or indeed any part, in determining whether or not the mixture model is identifiable.

In contrast for clusterwise regression a much more abstract sufficient condition for identifiability as given by (Hennig, 2000) seems to be the only sufficient condition for identifiability currently known, and this does not explicitly guarantee that the 1-component model is also identifiable. In terms of this abstract sufficient condition, later in this chapter we will present an alternate theorem giving sufficient conditions for identifiability that involve the same concepts as given in (Hennig, 2000). We will also show with a few examples that these two theorems are not equivalent, although they will nonetheless overlap to a great extent in terms of the situations for which they guarantee the identifiability of the MLMM.

For Model 1 with a simple within unit error covariance structure, and for the aforementioned rank condition on the random effects covariance matrices to hold, we need rank$(\boldsymbol{Z}_i) = q$, and $n_i > q$ for at least one unit i. Using this fact in combination with the discussion following Theorem 4.3.2, we thus obtain a Corollary of Theorem 4.3.2.

Corollary 4.3.3 *For Model 1 with a simple within unit error covariance structure, $\mathscr{D}_1$ is identifiable with respect to Ω_1 according to definition (4.2.1) if rank$(\boldsymbol{X}_i) = p$ for at least one unit i, and rank$(\boldsymbol{Z}_i) = q$, and $n_i > q$ for at least one unit i.*

The rank condition on the $\boldsymbol{Z}_i$ matrix should be satisfied in most samples since for LMMs the $\boldsymbol{Z}_i$ matrices are supposed to contain "observational" level data that should vary by row within a unit for each column. So whilst the model definitions for both the LMM and MLMM do not preclude say having two columns of some of the $\boldsymbol{Z}_i$ matrices being constants (and thus linearly related), in general we should not. Again although it is not precluded by definition, most $\boldsymbol{Z}_i$ matrices will not contain non-constant columns that are linearly related to the other non-constant columns, and so rank deficiency of all the $\boldsymbol{Z}_i$ matrices should not occur often for this reason.

Unfortunately the rank condition on the $\boldsymbol{X}_i$ matrix is more restrictive. Similar to the $\boldsymbol{Z}_i$ matrices, although it is not precluded by definition, most of the $\boldsymbol{X}_i$ matrices will not contain non-constant columns that are linearly related to the other non-constant columns, and so rank deficiency of all the $\boldsymbol{X}_i$ matrices should not occur often for this reason. The problem occurs because for LMMs the $\boldsymbol{X}_i$ matrices contain both "observational" and "global" level data. The global variables tend to vary within the sample, but frequently not within any unit. For example in medical studies age and hospital of treatment will often be in the $\boldsymbol{X}_i$ matrices, but these variables will often be fixed for a unit due to the duration of the study being a number of weeks or days only. If we assume for example that we have an intercept and age in the model, then all the $\boldsymbol{X}_i$ will be rank deficient, since for all i age will be linearly dependent with the intercept.

We see from the above discussion that the conditions of Theorem 4.3.2 preclude MLMMs that have all the $\boldsymbol{X}_i$ matrices having at least one constant column and an intercept. However Theorem 4.3.2 is a positive one, and so it does not tell us whether models not satisfying its hypotheses are not identifiable. This applies in particular to these MLMMs with constant columns. Since the utility of the MLMM is greatly reduced by restricting the $\boldsymbol{X}_i$ to contain only non-constant columns in the presence of an intercept, it is of interest to find sufficient conditions that do permit such models.

In this respect for MLMMs we can think of the N_T total scalar responses in the N response vectors as N_T scalar responses from N_T subjects. Thus by ignoring the units we can view a MLMM as a clusterwise regression model, and so the sufficient conditions for identifiability of clusterwise regression models given in Theorem 4.3.1 should also be sufficient to identify the fixed effects in MLMMs, and this theorem specifically includes an intercept in the model.

In light of the above discussion we now present a second theorem giving sufficient conditions for identifiability of MLMMs, but where the sufficient conditions for identification of the fixed effects is now given in terms of the minimum number of hyperplanes on which the rows of $\tilde{\boldsymbol{X}}$ lie. The first half of the proof is exactly the same as Theorem 4.3.4, whilst the second half is based on the method demonstrated in Theorem 4.3.1. We note that we still work with the original definition of Model 1, which does not assume an intercept is in the model. In the proof the index sets I_{n_i}, $i \in I$, index the n_i observations in $\boldsymbol{Y}_i$.

Theorem 4.3.4 *Let h denote the minimum number of $(p-1)$-dimensional hyperplanes on which the rows of $\tilde{\boldsymbol{X}}$ lie. Then $\mathscr{D}_1$ is identifiable with respect to Ω_1 according to definition (4.2.1) if $h > |S(J)|$, and if for all $\boldsymbol{\zeta}, \hat{\boldsymbol{\zeta}} \in \Psi_{\boldsymbol{\zeta}}$ we have $\boldsymbol{V}_i(\boldsymbol{\zeta}) = \boldsymbol{V}_i(\hat{\boldsymbol{\zeta}})$ if and only if $\boldsymbol{\zeta} = \hat{\boldsymbol{\zeta}}$ for at least one $i \in I$.*

Proof. Let $J, \hat{J} \in \Omega_1$, and assume $J = \hat{J}$. Since the normal distribution is completely determined by its mean vector and covariance matrix, this implies $F_i(\cdot|J) = F_i(\cdot|\hat{J})$ for all $i \in I$. Thus $J = \hat{J} \Rightarrow F(\cdot|J) = F(\cdot|\hat{J})$. Identifiability will follow if we can show $F(\cdot|J) = F(\cdot|\hat{J}) \Rightarrow J = \hat{J}$, or equivalently $F_i(\cdot|J) = F_i(\cdot|\hat{J})$ $\forall i \in I \Rightarrow J = \hat{J}$. To this end let $F_i(\cdot|J) = F_i(\cdot|\hat{J})\ \forall i \in I$, and suppose $J \neq \hat{J}$: we now seek a contradiction which will lead us to conclude $J = J$.

Without loss of generality assume for $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1) \in S(J)$ that $J\{(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)\} \neq \hat{J}\{(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)\}$, and $|S(J)| \geq |S(\hat{J})|$. Letting unit $k \in I$ be the unit that satisfies $\boldsymbol{V}_k(\boldsymbol{\zeta}) = \boldsymbol{V}_k(\hat{\boldsymbol{\zeta}})$ if and only if $\boldsymbol{\zeta} = \hat{\boldsymbol{\zeta}}$, by the identifiability of multivariate normal mixtures for unit k we have

$$\begin{aligned}
&J\{(\boldsymbol{\beta}, \boldsymbol{\zeta}) : \boldsymbol{X}_k\boldsymbol{\beta} = \boldsymbol{X}_k\boldsymbol{\beta}_1, \boldsymbol{V}_k(\boldsymbol{\zeta}) = \boldsymbol{V}_k(\boldsymbol{\zeta}_1), (\boldsymbol{\beta}, \boldsymbol{\zeta}) \in \Psi_1\}\\
=&\hat{J}\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) : \boldsymbol{X}_k\hat{\boldsymbol{\beta}} = \boldsymbol{X}_k\boldsymbol{\beta}_1, \boldsymbol{V}_k(\hat{\boldsymbol{\zeta}}) = \boldsymbol{V}_k(\boldsymbol{\zeta}_1), (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in \Psi_1\}\\
\Longleftrightarrow &J\{[(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\boldsymbol{\zeta}}\} = \hat{J}\{[(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\boldsymbol{\zeta}}\}.
\end{aligned} \tag{4.29}$$

Now because $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1) \in [(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\boldsymbol{\zeta}}$ we must have $J\{[(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\boldsymbol{\zeta}}\} > 0$, which implies $\hat{J}\{[(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\boldsymbol{\zeta}}\} > 0$, and so $[(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)]^{\boldsymbol{\zeta}} \cap S(\hat{J}) \neq \emptyset$. Furthermore this must hold for all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J)$ since $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)$ was picked arbitrarily from $S(J)$. Thus we have

$$\forall(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J),\ \ \exists(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in S(\hat{J}) \text{ such that } (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \overset{\boldsymbol{\zeta}}{\sim} (\boldsymbol{\beta}, \boldsymbol{\zeta}). \tag{4.30}$$

The possibility remains that the converse of (4.30) does not hold, that is to say there may be some points in $S(\hat{J})$ not equivalent under $\overset{\boldsymbol{\zeta}}{\sim}$ to any points in $S(J)$. We now show however that this cannot be true. Assume for $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1) \in S(\hat{J})$ that there does not exist a $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J)$ such that $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \overset{\boldsymbol{\zeta}}{\sim} (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)$. Once again by the identifiability of multivariate normal mixtures for unit k we have

$$\begin{aligned}
&J\{(\boldsymbol{\beta}, \boldsymbol{\zeta}) : \boldsymbol{X}_k\boldsymbol{\beta} = \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_1, \boldsymbol{V}_k(\boldsymbol{\zeta}) = \boldsymbol{V}_k(\hat{\boldsymbol{\zeta}}_1), (\boldsymbol{\beta}, \boldsymbol{\zeta}) \in \Psi_1\}\\
=&\hat{J}\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) : \boldsymbol{X}_k\hat{\boldsymbol{\beta}} = \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_1, \boldsymbol{V}_k(\hat{\boldsymbol{\zeta}}) = \boldsymbol{V}_k(\hat{\boldsymbol{\zeta}}_1), (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in \Psi_1\}\\
\Longleftrightarrow &J\{[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\boldsymbol{\zeta}}\} = \hat{J}\{[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\boldsymbol{\zeta}}\}.
\end{aligned} \tag{4.31}$$

Now because $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1) \in [(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\boldsymbol{\zeta}}$ we must have $\hat{J}\{[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\boldsymbol{\zeta}}\} > 0$, which implies $J\{[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\boldsymbol{\zeta}}\} > 0$, and so $[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)]^{\boldsymbol{\zeta}} \cap S(J) \neq \emptyset$. This contradicts our assumption that no point in $S(J)$ is equivalent to $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\zeta}}_1)$ under $\overset{\boldsymbol{\zeta}}{\sim}$, and so the converse of (4.30) holds:

$$\forall(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in S(\hat{J}),\ \ \exists(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J) \text{ such that } (\boldsymbol{\beta}, \boldsymbol{\zeta}) \overset{\boldsymbol{\zeta}}{\sim} (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}). \tag{4.32}$$

Together the equations (4.30) and (4.32) imply $S(J)$ and $S(\hat{J})$ are partitioned into the same number of equivalence classes under $\overset{\boldsymbol{\zeta}}{\sim}$, i.e. we have $s(\boldsymbol{\zeta}) := |S(J)/ \overset{\boldsymbol{\zeta}}{\sim} | = |S(\hat{J})/ \overset{\boldsymbol{\zeta}}{\sim} |$. We will let $\{A_1, ..., A_{s(\boldsymbol{\zeta})}\}$ and $\{\hat{A}_1, ..., \hat{A}_{s(\boldsymbol{\zeta})}\}$ denote these $s(\boldsymbol{\zeta})$ equivalence classes that partition $S(J)$ and $S(\hat{J})$ respectively, and we will use $l := 1, ..., s(\boldsymbol{\zeta})$ to index these classes.

We now turn our attention to applying the hyperplane condition of the theorem within these equivalence classes. Firstly assume the following statement holds

$$\begin{aligned}&\exists(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J), \forall i \in I, \forall j \in I_{n_i}, \exists(\hat{\boldsymbol{\beta}}(i,j), \hat{\boldsymbol{\zeta}}(i,j)) \in S(\hat{J}) :\\ &\qquad (\boldsymbol{\beta}, \boldsymbol{\zeta}) \overset{\boldsymbol{\beta}}{\nsim} (\hat{\boldsymbol{\beta}}(i,j), \hat{\boldsymbol{\zeta}}(i,j)) \Rightarrow (\boldsymbol{X}_i)_{j\cdot}\boldsymbol{\beta} = (\boldsymbol{X}_i)_{j\cdot}\hat{\boldsymbol{\beta}}(i,j).\end{aligned} \tag{4.33}$$

If say $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1)$ is the point in $S(J)$ guaranteed by (4.33) to exist, then for all $i \in I$, and $j \in I_{n_i}$ we have $((\boldsymbol{X}_i)_{j\cdot})^{\intercal} \in \{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{x}^{\intercal}(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}(i,j)) = \boldsymbol{0}\} = H_{p-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}(\hat{i}, j), \boldsymbol{0})$ for some $(\hat{\boldsymbol{\beta}}(i,j), \hat{\boldsymbol{\zeta}}(i,j)) \in S(\hat{J})$ where $\hat{\boldsymbol{\beta}}(i,j) \neq \boldsymbol{\beta}$. This means h, the minimum number of $(p-1)$-dimensional hyperplanes that cover the rows of $\tilde{\boldsymbol{X}}$ satisfies $h \leq |S(\hat{J})| \leq |S(J)|$, which contradicts our hypothesis that $h > |S(J)|$. Thus the negation of (4.33) must be true:

$$\begin{aligned}&\forall(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in S(J), \exists i \in I, \exists j \in I_{n_i}, \forall(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in S(\hat{J}) :\\ &\qquad\qquad (\boldsymbol{X}_i)_{j\cdot}\boldsymbol{\beta} = (\boldsymbol{X}_i)_{j\cdot}\hat{\boldsymbol{\beta}} \Rightarrow (\boldsymbol{\beta}, \boldsymbol{\zeta}) \overset{\boldsymbol{\beta}}{\sim} (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}).\end{aligned} \tag{4.34}$$

Now suppose $(\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1) \in A_1$, and that row m of unit j satisfies (4.34) for this point. Since all $(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in A_1$ satisfy $\boldsymbol{\zeta} = \boldsymbol{\zeta}_1$, and noting that the support set $S(\hat{J})$ does not contain duplicate vectors, we have

$$\forall(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) \in S(\hat{J}) \cap A_1 : (\boldsymbol{X}_j)_{m\cdot}\boldsymbol{\beta}_1 = (\boldsymbol{X}_j)_{m\cdot}\hat{\boldsymbol{\beta}} \Rightarrow (\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1) = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}), \tag{4.35}$$

and so by the identifiability of univariate normal mixtures

$$
\begin{aligned}
&J\{(\boldsymbol{\beta},\boldsymbol{\zeta}) : (\boldsymbol{X}_j)_{m\cdot}\boldsymbol{\beta} = (\boldsymbol{X}_j)_{m\cdot}\boldsymbol{\beta}_1, (\boldsymbol{V}_j(\boldsymbol{\zeta}))_{mm} = (\boldsymbol{V}_j(\boldsymbol{\zeta}_1))_{mm}, (\boldsymbol{\beta},\boldsymbol{\zeta}) \in S(J) \cap A_1\} \\
=&\hat{J}\{(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\zeta}}) : (\boldsymbol{X}_j)_{m\cdot}\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_j)_{m\cdot}\boldsymbol{\beta}_1, (\boldsymbol{V}_j(\hat{\boldsymbol{\zeta}}))_{mm} = (\boldsymbol{V}_j(\boldsymbol{\zeta}_1))_{mm}, (\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\zeta}}) \in S(\hat{J}) \cap A_1\} \\
\Longleftrightarrow &J\{(\boldsymbol{\beta},\boldsymbol{\zeta}) : (\boldsymbol{X}_j)_{m\cdot}\boldsymbol{\beta} = (\boldsymbol{X}_j)_{m\cdot}\boldsymbol{\beta}_1, (\boldsymbol{V}_j(\boldsymbol{\zeta}))_{mm} = (\boldsymbol{V}_j(\boldsymbol{\zeta}_1))_{mm}, (\boldsymbol{\beta},\boldsymbol{\zeta}) \in S(J) \cap A_1\} \\
=&\hat{J}\{(\boldsymbol{\beta}_1,\boldsymbol{\zeta}_1)\}. \qquad (4.36)
\end{aligned}
$$

Since $(\boldsymbol{\beta}_1,\boldsymbol{\zeta}_1) \in S(J) \cap A_1$ and $J\{(\boldsymbol{\beta}_1,\boldsymbol{\zeta}_1)\} > 0$, we must have $\hat{J}\{(\boldsymbol{\beta}_1,\boldsymbol{\zeta}_1)\} > 0$, which implies $(\boldsymbol{\beta}_1,\boldsymbol{\zeta}_1) \in S(\hat{J}) \cap A_1$. Now (4.36), and our assumption $J\{(\boldsymbol{\beta}_1,\boldsymbol{\zeta}_1)\} \neq \hat{J}\{(\boldsymbol{\beta}_1,\boldsymbol{\zeta}_1)\}$ imply

$$
\begin{aligned}
&\exists(\boldsymbol{\beta}_2,\boldsymbol{\zeta}_2) \in S(J) \cap A_1, (\boldsymbol{\beta}_2,\boldsymbol{\zeta}_2) \neq (\boldsymbol{\beta}_1,\boldsymbol{\zeta}_1) : \\
&\qquad (\boldsymbol{X}_j)_{m\cdot}\boldsymbol{\beta}_2 = (\boldsymbol{X}_j)_{m\cdot}\boldsymbol{\beta}_1 \text{ and } (\boldsymbol{V}_j(\boldsymbol{\zeta}_2))_{mm} = (\boldsymbol{V}_j(\boldsymbol{\zeta}_1))_{mm}. \qquad (4.37)
\end{aligned}
$$

Suppose row m' of unit j' satisfies (4.34) for this point $(\boldsymbol{\beta}_2,\boldsymbol{\zeta}_2)$ in $S(J) \cap A_1$. Then repeating the arguments starting from the paragraph above (4.35) but with $(\boldsymbol{\beta}_2,\boldsymbol{\zeta}_2)$ instead of $(\boldsymbol{\beta}_1,\boldsymbol{\zeta}_1)$ leads us to conclude $(\boldsymbol{\beta}_2,\boldsymbol{\zeta}_2) \in S(\hat{J}) \cap A_1$. But from (4.35) $(\boldsymbol{\beta}_2,\boldsymbol{\zeta}_2) \neq (\boldsymbol{\beta}_1,\boldsymbol{\zeta}_1)$ implies $(\boldsymbol{X}_j)_{m\cdot}\boldsymbol{\beta}_2 \neq (\boldsymbol{X}_j)_{m\cdot}\boldsymbol{\beta}_1$, which contradicts 4.37. ■

The question naturally arises as to whether either of Theorems 4.3.2 and 4.3.4 strictly imply the other. Using two examples we now show that they do not, and so we conclude that both theorems can provide sufficient conditions for identifiability in situations where the other cannot.

Consider a MLMM with $N = 3$ and $n_i = n$ for all $i = 1, 2, 3$, and assume one of the units identifies the covariance parameters. Suppose each $\boldsymbol{X}_i$ contains an intercept and constant column $\mathbf{1}_n a_i$, so that $((\boldsymbol{X}_i)_{j\cdot})^{\intercal} = (1, a_i)^{\intercal} \in \mathbb{R}^2$ for all i, and for all $j = 1, ..., n$. Now for $\boldsymbol{\alpha} \in \mathbb{R}^2$ the 1-dimensional hyperplane $H_1(\boldsymbol{\alpha}, c) = \{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{\alpha}^{\intercal}\boldsymbol{x} = c\}$ is a line in $\mathbb{R}^2$, and for each i we can draw a line in $\mathbb{R}^2$ through the point $(1, a_i)$ which intersects the vertical axis. If we assume $a_1 \neq a_2 \neq a_3$ (all non-zero) then these lines must all be different, and so the rows of $\tilde{\boldsymbol{X}}$ lie on three distinct 1-dimensional hyperplanes. If the MLMM has two components then by Theorem 4.3.4 this MLMM is identifiable since we have $h = 3 > |S(J)|$. However even though $\tilde{\boldsymbol{X}}$ is full rank, each $\boldsymbol{X}_i$ has less than full column rank, which means each $\boldsymbol{X}_i$ cannot identify the fixed effects. Thus the hypothesis of Theorem 4.3.2 is not satisfied, and so we cannot use Theorem 4.3.2 to tell us if this MLMM is identifiable. So in this example, confirming the identifiability of the model depends on the number of components in the model.

For the reverse of the above example, assume $N = 2$, $n_i = n$ for all $i = 1, 2$, and as before assume that one of the units identifies the covariance parameters. This time let the $\boldsymbol{X}_i$ contain an intercept and one other column that is not necessarily constant. Let the n rows of $\boldsymbol{X}_1$ consist of at least one row $(1, a)$ and one row $(1, b)$, where $a \neq b$, and let $\boldsymbol{X}_2$ consist of an intercept and a constant column $\mathbf{1}_n c$, where $a \neq b \neq c$. Using this setup we have that $\boldsymbol{X}_2$ is rank deficient but that $\boldsymbol{X}_1$ has full column rank, and hence identifies the fixed effects, which means that the model is identifiable by Theorem 4.3.2. Unlike the previous example however we can confirm that identifiability holds regardless of how many components the model has. For example if h is the minimum number of hyperplanes that cover the rows of the covariate data, then choosing $|S(J)| = h$ would not allow us to use 4.3.4 to confirm the identifiability of this model.

5

Simulations

This chapter is concerned with evaluating through simulations the "naive" methods of statistical inference we proposed in section 3.4. In this respect the models we use in these simulations are intended to be realistically complex both in terms of the numbers of parameters and units used, and in terms of how similar the components these parameters define are, so that the dual challenge of parameter estimation and classification of units to components is of an order of difficulty approximately equal to that encountered in many real world scenarios. In this respect there are two major factors that determine this level of difficulty - sample sizes and component separation, which are of course linked to some extent. The reason for restricting sample sizes is that it is very easy to specify a model such that on a decent but nonetheless ordinary personal computer the resulting model would take a prohibitively long period of time to converge. For example for the types of models we are concerned with in this thesis even $N = 5000$ with $n_i = 5$ for all $i \in I_N$, or $N = 100$ with $n_i = 250$ for all $i \in I_N$ would take at least a week to converge. Thankfully for many real world scenarios, particularly in many medical studies, $N \leq 1000$, and $n_i \leq 10$ for all $i \in I_N$ are the typical sample size ranges encountered.

The reason for restricting the level of component separation (i.e. avoiding components too close together) is that there is an element of triviality about this, that is to say we can always specify a mixture model that is impossible to estimate well by setting the parameters of all components to be the same - in such a scenario there can be no method of inference on the model parameters that would perform well.

Since our primary aim is to take a "first look" at the inference methods proposed in section 3.4 we were careful to ensure these methods had every chance of success and so we avoided these trivially difficult models and instead limited the maximum difficulty of the models to be moderately difficult where the closeness of the components was less extreme. This is not to say however that we present a biased view of how well the proposed inference methods work but rather that we knew from our testing periods the limitations of these inference methods. In this respect it is evident from the results that these inference methods perform very poorly even for some of these moderately difficult mixture models and so to include models with extremely close components would be to labor the point since the reader could easily guess the outcome.

Section 5.1 is concerned with describing the quantities we will use to quantify the performance of a particular model from a simulation, and also to give details of the simulation procedures so they can if needed be replicated. In section 5.2 we describe some simulations investigating the performance of the first and second variants of the EM algorithm in terms of parameter inference. In section 5.3 we describe simulations investigating which factors associated with a model (number of units, within-unit samples sizes, magnitude of the covariance parameters etc.) influence parameter inference.

5.1 Simulation methods

Firstly by a model we mean a data generating process. For example a MLMM with G components, p fixed effects, q random effects, an unstructured random effects covariance structure, and a simple within-units error covariance structure would be a model, but one with $G+1$ components would be a different model. By a model version we mean a particular choice of number of parameters, and fixed covariate data for the model. For example one version of a model might be to choose all of the fixed effects as being continuous variables, whilst another version might be to have some factor and continuous variables. One version of a model might be to have a certain choice of fixed data in the random effects design matrix, whilst another would be to change that data. One version of a model might be to choose one set of parameters, whilst another version would be to choose another set of parameters whereby at least one of the parameters is different in value from the other set. One version of a model might be to choose a certain number of units whilst another would be to choose a different number of units. Furthermore combinations of these choices combine in the obvious way to define further model versions. In this way a model is general and can have an infinite number of versions associated with it, whilst a model version is unique apart from the random data - i.e. the within-unit errors and the component memberships.

5.1.1 Data generation

For any model, and any given model version, we can generate the random data and then run the EM algorithm to estimate the model parameters for that version. We will call this a replication of the model version, or just a replication for short. Each simulation we will describe will be associated with one model and a certain number of versions of that model. We will perform a given number of individual replications in the simulation, which we will denote by $Nsim$, whereby we will collect information from the results of the individual replications - parameter estimates primarily.
Let $s \in I_{Nsim} := \{1, 2, ..., Nsim\}$ index the simulation number. We assume $n_c + n_f = p$ fixed variables in each of the N fixed effects design matrices $\boldsymbol{X}_i$, where n_c and n_f are the number of continuous and factor variables respectively. Although it is not an assumption of MLMMs, for these simulations we choose $q \leq p$, and we choose the q variables in the N random effects design matrices $\boldsymbol{Z}_i$ to be a subset of the p variables in the fixed effects design matrices $\boldsymbol{X}_i$. These choices however do reflect good practice in that to include a variable in the $\boldsymbol{Z}_i$ matrices that is not in the $\boldsymbol{X}_i$ matrices is to imply we are assuming that the effect with which the variable is associated with works only at a unit level, and not at a “global” level. Such an individual but not global effect will often be difficult to justify.

For any given model the following steps are performed once in order to generate the fixed data in a MLMM;

Within-unit sample sizes: Let $max\text{-}n_i$ denote the maximum number of observations we will allow the N units to have, and let $t = t_k$ for $k = 1, 2, ..., max\text{-}n_i$, denote the time variable that the data we will generate will correspond to. This means when time is entered into a model as a continuous variable that the complete set of time points are equally spaced. This is convenient because when the within unit errors follow an autoregressive process we must have equally spaced time points. Of course we could have used unequally spaced time points for models without an autoregressive process for the within unit errors.

Now for the i^{th} unit, $i \in I_N$, and for $k = 1, ..., max\text{-}n_i$, we draw a value x_k from the Bernoulli distribution with probability p. If $x_k = 1$ then in subsequent steps (see below) we will generate data for this unit at time point $t = k$, but if $x_k = 0$ then this unit will not have data at time point $t = k$. In this way if X_k, $k = 1, ..., max\text{-}n_i$, are the Bernoulli random variables, then we are assuming $X := \sum_{k=1}^{max\text{-}n_i} X_k \sim Bin(max\text{-}n_i, p)$. Since $\boldsymbol{E}[X] = (max\text{-}n_i)p$, then for all units $n_{pc} := (\boldsymbol{E}[X]/max\text{-}n_i) * 100 = p * 100$ is the average number observations, expressed as a percentage of $max\text{-}n_i$, that we might expect a unit to have. For the models used in sections 5.2 and 5.3 we will choose p to give a desired amount of unbalancedness in the within-unit sample sizes by calculating $p = n_{pc}/100$, that is by choosing p in this way, for all units $100 - n_{pc}$ is the average number of missing values we might expect, expressed as a percentage of $max\text{-}n_i$. Setting $p = 1$ gives balanced within-unit sample sizes where $n_i = max\text{-}n_i$ for all $i \in I_N$.

Time variable: The time variable $t = t_k$ for $k = 1, 2, ..., max\text{-}n_i$, that we introduced above can be entered into a model as either a continuous or a factor variable. When we wish to think of t as continuous then we choose to use a centered variable $t_k^c = t_k - \bar{t}$, for all k, where $\bar{t}$ is the mean of the $max\text{-}n_i$ time points. Although we do not include t^2 as a variable in any of the models we consider, this centering can have beneficial effects on the estimation of the parameter associated with the polynomial time effect (Cnaan and Slasor (1997)).

Factor variables: For each $i \in I_N$, $k = 1, 2, ..., n_f$, let $x_{i,k}^f \in \mathbb{R}^{n_i}$ denote the k^{th} factor variable that has l_k levels. If we allow this factor variable to vary within units then we randomly draw n_i values $m_j \in \{1, .., l_k\}$, $j = 1, 2, ..., n_i$, from a l_k-dimensional Multinomial distribution with parameters 1 and p_k, where p_k is a $l_k \times 1$ vector with entries l_k^{-1}. That is we assume that m_j is the realized value of a Multinomial random variable M_j with equal group probabilities, and is distributed as $M_j \sim \text{mult}_{l_k}(1, p_k)$. If we do not allow this factor variable to vary within units, then we draw just a single value from this distribution, and then copy it n_i times into $x_{i,k}^f$. The fixed variable $x_{i,k}^f$ is then split into l_k separate $n_i \times 1$ dummy variables that take on the values 0 or 1 that indicate to which level of this factor variable the j^{th} response for the i^{th} unit belongs, and it is these l_k dummy variables that are included in the fixed-effects design matrices $\boldsymbol{X}_i$, and the random effects design matrices $\boldsymbol{Z}_i$. Since for the i^{th} unit we may have $n_i < max\text{-}n_i$, then the n_i observations in each of these factor variables will not necessarily correspond to sequential observations taken at $t = 1, 2, ..., max\text{-}n_i$.

Continuous variables: For each $i \in I_N$, $s = 1, 2, ..., n_c$, let $\boldsymbol{X}_i^c$ be a $n_i \times n_c$ matrix whose s^{th} column contains the s^{th} continuous variable. If the continuous variables are allowed to vary within units, then we make n_i random draws of a n_c-dimensional vector $\boldsymbol{x}_j$, $j = 1, 2, ..., n_i$, from a normal distribution with mean $\boldsymbol{\mu}_s := (\mu_1, ..., \mu_{n_c})^\intercal \in \mathbb{R}^{n_c}$, and $n_c \times n_c$ covariance matrix $\Sigma_s := \text{diag}\{\sigma_1^2, ..., \sigma_{n_c}^2\}$. We then set the n_i rows of $\boldsymbol{X}_i^c$ to be equal to these vectors, that is $\boldsymbol{x}_j^\intercal$ is the j^{th} row of $\boldsymbol{X}_i^c$, where we assume $\boldsymbol{x}_j$ is the realized value of a random vector $\boldsymbol{X}_j$ that is distributed as $\boldsymbol{X}_j \sim N_{n_c}(\boldsymbol{\mu}_s, \Sigma_s)$. If one or more of the n_c variables are not permitted to vary within units, then we copy the relevant entries of the first row of $\boldsymbol{X}_i^c$ into the relevant entries of all the subsequent rows. Since for the i^{th} unit we may have $n_i < max\text{-}n_i$, then the n_i observations in each of these continuous variables will not necessarily correspond to sequential observations taken at $t = 1, 2, ..., max\text{-}n_i$.

For each $s \in I_{Nsim}$ we generate the random data of a model in the following way;

Component memberships: Using the notation from Chapter 2, for each unit $i \in I_N$, we draw the $G \times 1$ random vectors $\boldsymbol{\Lambda}_i^{(s)}$ (which denote component membership) from a G-dimensional Multinomial distribution with parameters 1 and $\boldsymbol{\pi}$. That is $\boldsymbol{\Lambda}_i^{(s)} \sim \text{mult}_G(1, \boldsymbol{\pi})$ for all i.

Responses: Using the notation from Chapter 2 we let $\boldsymbol{\lambda}_i^{(s,g)}$ denote $\boldsymbol{\Lambda}_i^{(s)} = \boldsymbol{\lambda}_i^{(g)}$ for $g \in I_G$. Thus $\boldsymbol{\lambda}_i^{(s,g)}$ is a variable that indicates for the s^{th} simulation that the i^{th} unit belongs to the g^{th} component. For all $i \in I_N$ we then draw the random effects vector $\boldsymbol{u}_i^{(s)}$ from the distribution of $\boldsymbol{U}_i|\boldsymbol{\lambda}_i^{(s,g)}$ which is given by $\boldsymbol{U}_i|\boldsymbol{\lambda}_i^{(s,g)} \sim N_q(\boldsymbol{0}, \boldsymbol{D}_g)$ and the within unit error vector $\epsilon_i^{(s)}$ from the distribution of $\boldsymbol{e}_i^{(s)}|\boldsymbol{\lambda}_i^{(s,g)}$ which is given by $\boldsymbol{e}_i^{(s)}|\boldsymbol{\lambda}_i^{(s,g)} \sim N_{n_i}\left(\boldsymbol{0}, \sigma_g^2 \boldsymbol{C}_i(\boldsymbol{\phi}_g)\right)$. Conditional on $\boldsymbol{\lambda}_i^{(s,g)}$ and $\boldsymbol{u}_i^{(s)}$ we then generate the i^{th} response vector $\boldsymbol{y}_i^{(s)}$ as $\boldsymbol{y}_i^{(s)} = \boldsymbol{X}_i\boldsymbol{\beta}_g + \boldsymbol{Z}_i\boldsymbol{u}_i^{(s)} + \epsilon_i^{(s)}$.

5.1.2 Parameter Estimation

All parameters were estimated using either the first or second variant of the EM algorithm which were described in section 2.2. For parameter estimation the following procedure was followed for each of the G components: for component $g \in I_G$ randomly choose $i' = 1, ..., N'$ units where N' is quite small (typically 30-50 units). Compute an initial estimate of the fixed effects $\hat{\boldsymbol{\beta}_g}^{(0)}$ using the ordinary least squares (OLS) estimator

$$\hat{\boldsymbol{\beta}_g}^{(0)} = \left(\sum_{i'=1}^{N'} \boldsymbol{X}_{i'}^{\intercal} \boldsymbol{X}_{i'} \right)^{-1} \sum_{i'=1}^{N'} \boldsymbol{X}_{i'}^{\intercal} \boldsymbol{y}_{i'}. \tag{5.1}$$

For the within-unit covariance parameters we ignore any autoregressive parameters and only compute estimates of the within-unit variances - this is equivalent to assuming $\hat{\boldsymbol{\phi}}_g^{(0)} = \mathbf{0}$. The initial within-unit variances were estimated as the average of the residual sum of squares from the above OLS regression, that is

$$\hat{\sigma_g}^{2(0)} = \frac{\sum_{i'=1}^{N'} \left(\boldsymbol{y}_{i'} - \boldsymbol{X}_{i'}^{\intercal} \hat{\boldsymbol{\beta}_g}^{(0)} \right)^{\intercal} \left(\boldsymbol{y}_{i'} - \boldsymbol{X}_{i'}^{\intercal} \hat{\boldsymbol{\beta}_g}^{(0)} \right)}{N'}. \tag{5.2}$$

For the random effects covariance parameters we simply set $\hat{\boldsymbol{D}}_g^{(0)} = \boldsymbol{I}_q$ which implies we are assuming the random effects are uncorrelated and have unit variances. For the mixing proportions we set the initial estimates to be $1/G$ for all the G components, that is $\hat{\boldsymbol{\pi}_g}^{(0)} = G^{-1}\mathbf{1}_G$.

Following this parameter estimate initialisation we estimated the mixture model starting with these initial estimates, but where we only ran the EM algorithm for a very small number of iterations (typically five), and recorded the log-likelihood after the final EM iteration. We repeated this parameter initialisation and mixture model estimation procedure five times and determined the repetition with the highest log-likelihood. We then took the final mixture model parameter estimates from this repetition and used these as the starting values for the parameter estimates in a full run of the EM algorithm. In this full run we ran the EM algorithm until either convergence, or until a maximum number of iterations had been achieved without convergence - this was set at 5000 and 50 for the first and second variants of the EM algorithm respectively. The large disparity in these maximum values is because the first variant of the EM algorithm is very slow to converge, whereas the second variant is comparatively very fast to converge - see section 5.2 for a discussion of this difference. We defined the EM algorithm to have converged (for both variants) when $\frac{LL_s - LL_{s-1}}{|LL_s| + 0.1} < 1e^{-8}$, where LL_s and LL_{s-1} are the log-likelihood values on the s^{th} and $(s-1)^{th}$ iterations of the EM algorithm respectively.

5.1.3 Summary measures within replications

In this subsection we will continue to use the index $s \in I_{Nsim}$ that we introduced at the end of subsection 5.1.1, and which indexes the $Nsim$ replications of a particular model version. Recalling that $n_{\boldsymbol{\theta}}$, and $n_{\Theta} = (G * n_{\boldsymbol{\theta}}) + G$ denote the number of parameters in the 1-component model and MLMM respectively, then $(\boldsymbol{\theta})_t$ for $t \in I_{n_\Theta} := \{1, ..., n_\Theta\}$, is a scalar parameter. In addition to the parameter estimates $\hat{\boldsymbol{\theta}}^{(s)}$ we shall also calculate the following quantities for each replication of a simulation.

Asymptotic standard errors: For the parameter estimate $\hat{\boldsymbol{\theta}}^{(s)}$ from the s^{th} replication, let $\hat{I}_M\left(\hat{\boldsymbol{\theta}}^{(s)}\right)$, $M \in \{1, 2, 3, 4\}$ denote one of four methods we will use to approximate the sample information matrix $I_N\left(\hat{\boldsymbol{\theta}}^{(s)}\right)$ that we described in (3.4.2) and (3.4.3). In this respect we let $\hat{I}_1\left(\hat{\boldsymbol{\theta}}^{(s)}\right) = S_N\left(\hat{\boldsymbol{\theta}}^{(s)}\right)$, $\hat{I}_2\left(\hat{\boldsymbol{\theta}}^{(s)}\right) = J_N\left(\hat{\boldsymbol{\theta}}^{(s)}\right)$, $\hat{I}_3\left(\hat{\boldsymbol{\theta}}_g^{(s)}\right) = CW_N\left(\hat{\boldsymbol{\theta}}_g^{(s)}\right)$, $g \in I_G$, and $\hat{I}_4\left(\hat{\boldsymbol{\theta}}^{(s)}\right) = SW_N\left(\hat{\boldsymbol{\theta}}^{(s)}\right)$. Then from the assumed asymptotic normal distribution of the estimator $\hat{\boldsymbol{\theta}}^{(s)}$ of $\boldsymbol{\theta}$ given in (3.48), we see for $M = 1, 2, 4$ that

$$SE_M\left(\left(\hat{\boldsymbol{\theta}}_g^{(s)}\right)_t\right) = \sqrt{\left(\left(\hat{I}_M\left(\hat{\boldsymbol{\theta}}^{(s)}\right)\right)^{-1}\right)_{t,t}}, \tag{5.3}$$

will give asymptotic standard errors for $\left(\hat{\boldsymbol{\theta}}^{(s)}\right)_t$, and for $g \in I_G$

$$SE_3\left(\left(\hat{\boldsymbol{\theta}}_g^{(s)}\right)_t\right) = \sqrt{\left(\left(\hat{I}_3\left(\hat{\boldsymbol{\theta}}_g^{(s)}\right)\right)^{-1}\right)_{t,t}}, \tag{5.4}$$

will give asymptotic standard errors for $\left(\hat{\boldsymbol{\theta}}_g^{(s)}\right)_t$.

Asymptotic confidence intervals:

From the assumed asymptotic normal distribution of the estimator $\hat{\boldsymbol{\theta}}^{(s)}$ of $\boldsymbol{\theta}$ given in (3.48), for $M = 1, 2, 4$, and for $t \in I_{n_\Theta}$, we will calculate $CI_M^{(s)}((\boldsymbol{\theta})_t)$, the $(1-\alpha)*100\%$ approximate normal confidence interval for $(\boldsymbol{\theta})_t$, as

$$\begin{aligned} CI_M^{(s)}((\boldsymbol{\theta})_t) = \Big[\left(\hat{\boldsymbol{\theta}}^{(s)}\right)_t - Z_{1-\alpha/2} SE_M \left(\left(\hat{\boldsymbol{\theta}}_g^{(s)} \right)_t \right) \; , \\ \left(\hat{\boldsymbol{\theta}}^{(s)}\right)_t + Z_{1-\alpha/2} SE_M \left(\left(\hat{\boldsymbol{\theta}}_g^{(s)} \right)_t \right) \Big] , \end{aligned} \tag{5.5}$$

and similarly for $M = 3$, and for any $g \in I_G$, we will calculate $CI_M^{(s)}((\boldsymbol{\theta}_g)_t) = CI_3^{(s)}((\boldsymbol{\theta}_g)_t)$ as

$$\begin{aligned} CI_3^{(s)}((\boldsymbol{\theta}_g)_t) = \Big[\left(\hat{\boldsymbol{\theta}}_g^{(s)}\right)_t - Z_{1-\alpha/2} SE_3 \left(\left(\hat{\boldsymbol{\theta}}_g^{(s)} \right)_t \right) \; , \\ \left(\hat{\boldsymbol{\theta}}_g^{(s)}\right)_t + Z_{1-\alpha/2} SE_3 \left(\left(\hat{\boldsymbol{\theta}}_g^{(s)} \right)_t \right) \Big] . \end{aligned} \tag{5.6}$$

We shall also calculate the standardised confidence intervals by multiplying the confidence intervals in (5.5) and (5.6) by the reciprocal of the modulus of the true parameter value, that is we calculate

$$StCI_M^{(s)}((\boldsymbol{\theta})_t) = \left(\frac{1}{(|\boldsymbol{\theta}|)_t} \right) CI_M^{st}((\boldsymbol{\theta})_t), \tag{5.7}$$

for $M = 1, 2, 4$, and

$$StCI_M^{(s)}((\boldsymbol{\theta}_g)_t) = \left(\frac{1}{(|\boldsymbol{\theta}_g|)_t} \right) CI_M^{st}((\boldsymbol{\theta}_g)_t), \tag{5.8}$$

for $M = 3$, and $g \in I_G$.

Confidence interval lengths:

We will calculate the lengths of the symmetric confidence intervals given in (5.5), (5.7), (5.6) and (5.8), which we will denote respectively by $CIL_M^{(s)}((\boldsymbol{\theta})_t)$ and $stCIL_M^{(s)}((\boldsymbol{\theta})_t)$ for $M = 1, 2, 4$, and $CIL_3^{(s)}((\boldsymbol{\theta}_g)_t)$ and $stCIL_3^{(s)}((\boldsymbol{\theta}_g)_t)$ for $M = 3$ and $g \in I_G$. The lengths of the standardised confidence intervals should behave in a way such that a confidence interval for a parameter that is large in value should not necessarily be long, nor should a confidence interval for a parameter that is small in value necessarily be short.

Classification errors:

For each $i \in I_N$, let $g_i^{(s)} \in I_G$ denote the component to which unit i is assigned using the following rule: $\hat{p}_{ig}^{(s)} = \max\{\hat{p}_{i1}^{(s)}, ..., \hat{p}_{iG}^{(s)}\}, g \in I_G \to g_i^{(s)} = g$. Letting $N_g^{(s)}$ denote the number of the N_g units whose $g_i^{(s)}$ values do not equal g, then $CE_g^{(s)} := (N_g^{(s)}/N_g) * 100$ is the percentage of units that belong to component g that have been incorrectly classified to one of the other $G-1$ components. Using this definition then $CE_T^{(s)} := (N_T^{(s)}/N) * 100$ for $N_T^{(s)} := \sum_{j=1}^{G} N_j^{(s)}$ is the percentage of the total N units that have been incorrectly classified.

5.1.4 Summary measures over all replications

We shall calculate averages and standard deviations over the $Nsim$ replications for all the quantities we described in subsection 5.1.3. In addition we also calculate the mean square error (MSE) of the sequence of parameter estimates and their standard errors. For a general sequence of scalar quantities $\{x_s\}_{s=1}^{Nsim}$, where each x_s is supposed to estimate μ, letting $\bar{x}$ and S_x be the sample mean and variance respectively of this sequence, then using

$$BIAS(\bar{x}) = \bar{x} - \mu, \tag{5.9}$$

and

$$SE(\bar{x}) = \sqrt{\frac{1}{Nsim} \sum_{j=1}^{Nsim} (x_j - \mu)^2}, \tag{5.10}$$

then the MSE of $\bar{x}$ is given by

$$MSE(\bar{x}) = SE(\bar{x})^2 + BIAS(\bar{x})^2. \tag{5.11}$$

In addition to the mean and standard deviation of the confidence interval lengths we will also use coverage probabilities to quantify the performance of the confidence intervals over the $Nsim$ replications. For any given model, any given confidence interval method M, and any $t \in I_{n_\Theta}$, let $c_M^{(t)}$ and $c_3^{(g,t)}$ be the number of times the $Nsim$ confidence intervals respectively contain the true parameter $(\boldsymbol{\theta})_t$ (for M=1,2,4) or $(\boldsymbol{\theta}_g)_t$ for

M=3. Then we will calculate the estimated coverage probabilities as $CP^t_M = c^{(t)}_M/Nsim$ or $CP^{(g,t)}_3 = c^{(g,t)}_3/Nsim$. Since $CP^{(t)}_M$ and $CP^{(g,t)}_3$ are proportions then we shall also construct approximate 95% Binomial confidence intervals about these proportions using the following formulae

$$BCI^{(t)}_M \pm z_{0.975}\sqrt{\frac{CP^{(t)}_M}{(1-CP^{(t)}_M)Nsim}}, \tag{5.12}$$

and

$$BCI^{(g,t)}_3 \pm z_{0.975}\sqrt{\frac{CP^{(g,t)}_3}{(1-CP^{(g,t)}_3)Nsim}}, \tag{5.13}$$

where $z_{0.975}$ is the value of a standard normal random variable Z such that $P[|Z| > z_{0.975}] \leq 0.05$. We will use the term "range of coverage" to mean the values between and including the end-points of these Binomial confidence intervals. Unless otherwise stated we will consider two coverage probabilities to be similar if the ranges of those coverage probabilities intersect, and different if they do not. Similarly unless otherwise stated if the nominal level α is contained in the range of coverage of a confidence interval then we will consider that confidence interval to have attained the nominal level.

Despite the merits of presenting both coverage probabilities and means of confidence interval lengths, for simulations that investigate many different models, and where the models have many parameters, it is tedious to look at two sets of information in order to asses the quality of the confidence intervals. In this respect we now propose the construction of an index that combines both pieces of information.

For any $t \in I_{n_\Theta}$, and $g \in I_G$, let $\overline{StCIL^{(t)}_M}$ for $M = 1, 2, 4$, and $\overline{StCIL^{(g,t)}_3}$ for $M = 3$ denote the means of the lengths of the standardised confidence intervals $StCI^{(s)}_M((\boldsymbol{\theta})_t)$ and $StCI^{(s)}_3((\boldsymbol{\theta}_g)_t)$ respectively taken over all the $Nsim$ replications. Furthermore denote the lower and upper endpoints of the Binomial confidence intervals $BCI^{(t)}_M$ and $BCI^{(g,t)}_M$ by a and b respectively. Then we define the "coverage probability and length" index, which we will denote by $CPLI^{(t)}_M$, and $CPLI^{(g,t)}_3$ as follows

$$CPLI^{(t)}_M = \frac{CP^{(t)}_M}{\overline{StCIL^{(t)}_M} + \left[((\overline{StCIL^{(t)}_M})d^{(t)}_M) + 1\right]^2}, \tag{5.14}$$

and

$$CPLI_3^{(g,t)} = \frac{CP_3^{(g,t)}}{\overline{StCIL_3^{(g,t)}} + \left[((\overline{StCIL_M^{(g,t)}})d_3^{(g,t)}) + 1\right]^2}, \tag{5.15}$$

where

$$d_M^{(t)} = \begin{cases} \max\{d(a,\alpha), d(b,\alpha)\}, & \text{if } \alpha \notin BCI_M^{(t)}, \\ \\ 0, & \text{otherwise}, \end{cases} \tag{5.16}$$

and

$$d_3^{(g,t)} = \begin{cases} \max\{d(a,\alpha), d(b,\alpha)\}, & \text{if } \alpha \notin BCI_M^{(g,t)}, \\ \\ 0, & \text{otherwise}, \end{cases} \tag{5.17}$$

and where $d(x,y) = |x-y|$ for any two real numbers x and y. We note that since the nominal level α is typically set to be high, for example 0.95, then $d_M^{(t)}$ and $d_3^{(g,t)}$ will almost all of the time be the distance of the lower end point a from α. The exception is when the Binomial confidence intervals are particularly short and centered over α.

We see for fixed $CP_M^{(t)}$, and $d_M^{(t)}$ that $CPLI_M^{(t)} \to 0$ as $\overline{StCIL_M^{(t)}} \to \infty$, and for fixed $\overline{StCIL_M^{(t)}}$, and $d_M^{(t)}$ that $CPLI_M^{(t)} \to 0$ as $CP_M^{(t)} \to 0$. The purpose of the squared term in the denominator of 5.14 is to penalise coverage probabilities either because they have a Binomial confidence interval one of whose endpoints is far away from the nominal level, or because that coverage probability is associated with confidence intervals that tend on average to be long. The purpose of the "+1" term is simply to prevent $CPLI_M^{(t)}$ from tending to infinity as $\overline{StCIL_M^{(t)}} \to 0$ (for fixed $d_M^{(t)}$). Indeed in this situation, which can be interpreted as the $Nsim$ confidence intervals all becoming infinitely precise, then we simply take $CPLI_M^{(t)}$ to be $CP_M^{(t)}$. In contrast the effect of large $\overline{StCIL_M^{(t)}}$ and/or $d_M^{(t)}$ is to down weight $CPLI_M^{(t)}$. Finally by setting $d_M^{(t)}$ to be zero when $BCI_M^{(t)}$ contains α we are implying that $CPLI_M^{(t)}$ should, all other things being equal, be higher than $CPL_M^{(t)}$ when the Binomial confidence interval does not contain α. This is the "reward" for attaining the nominal level. The behavior we have just described for $CPLI_M^{(t)}$ obviously applies to $CPLI_M^{(g,t)}$ as well.

In addition to the above summary measures we shall also use various plots to summarise the simulation results. For the coverage probabilities we will use errorbar plots where the errorbars are the binomial confidence intervals. For the parameter estimates and the confidence interval lengths we will use box plots, and for the coverage probability and length indices we will use simple bar charts. For some of the quantities plotted, namely the parameter estimates and confidence interval lengths, some very large outliers were observed. Some of these outliers were so large that they dominated the boxplots at the expense of showing other relevant information such as the median value, and the inter-quartile ranges of the quantities of interest. Furthermore by displaying these outliers, it is impossible to meaningfully compare the results of different model versions (for section 5.2) or for different factorial variable settings (for section 5.3).

For this reason we chose to use the "compress" value of the "extrememode" parameter in the Matlab routines we used to produce the boxplots. Using this parameter setting Matlab truncates any data point outside a user supplied range, and displays these truncated values in a "compression region"' whilst maintaining the relative position of the points. This has the effect of showing the reader the number of outlying values, and the threshold lower and upper values the outliers exceed, but does not show the exact value of the outliers in order to not stretch the y-axis of the plots. We usually specified the threshold values in terms of percentiles of the data, so that data points outside say the 5^{th} and 95^{th} percentiles were plotted in the compression regions.

For the factorial simulations we conduct in 5.3, for each of the three models we investigate we will use many different model versions (between 64 and 128). Accordingly the simulations will produce a lot of information, and whilst the plots will be valuable tools to get an overview of the results, they may not by themselves provide an easy method of determining the relationship between our quantities of interest (parameter estimate MSEs, parameter CPs, CILs, and CPLIs) and the simulation variables. In order to determine these relationships we will fit robust linear models to each of these quantities, with the simulation variables as the covariates. Specifically we will obtain M-estimates of the effects of the simulation variables using the "robustreg" procedure in the SAS statistical software system (Cary, NC: SAS Institute). We shall use the default settings of this procedure which uses the bisquare function and median as the weight functions ρ and ρ_{scale} associated with estimating the location and scale parameters respectively of the unknown data distribution. The cut-off value used by SAS for ρ is 4.68, meaning that any model residuals whose values are close to 0 are almost unweighted, residuals whose values equal or exceed ± 4.68 have values of zero, and the residuals in the range -4.68 to 4.68 are down weighted by weights that follow a symmetrical bell-shaped curve that on each side goes from 1 to 0. The denominator used by SAS in ρ_{scale} is 0.675 meaning that the scale estimator is consistent for the true scale parameter when the data distribution is Normal.

5.2 EM first and second variants

We described in subsection 3.4.3 how the second variant of the EM algorithm can motivate the use of componentwise inference, where we used some ideas presented in Grün (2008). However the primary focus of Grün (2008) was not componentwise inference at all, but rather it was to draw attention to the second variant of the EM algorithm as both a conceptually and practically easier method to implement than the first variant we described in subsection 2.2.1. The second variant is easier to implement from a practical viewpoint since if existing software or code libraries can maximise the weighted log-likelihood required in step 2, then no new code need be written to perform this step other than calling the required functions or methods with the correct weights. Even if a given software package or code library cannot perform this weighted estimation then a transformed model can be estimated instead (Grün, 2008) which gives the same parameter estimates as using weighted estimation.

In terms of the differences between the two EM algorithm variants, an obvious question is whether both variants of the EM algorithm give the same parameter estimates. It is appealing if they do, for if they do not then the parameter estimates obtained will depend on which data are thought of as missing, and this decision can sometimes be arbitrary when there really is no "missing" data. If the two variants are equivalent then this should be proven, however it is not obvious how to do this since the two variants are very different.

Other than this fundamental question there are of course questions regarding performance of the two variants. For this reason Grün (2008) poses some questions regarding this choice of EM method and postulates that the first variant, since it has more missing data, should need more iterations but that each iteration will be faster than an iteration of the second variant algorithm because the M step will be in closed form. However the maximisations in step 2 of the second variant involve fitting a LMM which utilise a Newton-Raphson or Fisher Scoring algorithm which are typically fast to converge, thus it is not necessarily clear which variant will be the quickest.

In light of these questions, and in addition to the main simulations in section (5.3), we will include here a comparison of the EM first and second variants. Specifically we want to see if the same parameter estimates are obtained from both variants, and to compare computational performance. Since the quality of the parameter estimates in terms of bias and variability will also affect the quality of the confidence intervals in terms of coverage probabilities and confidence interval lengths, then we will also compare the confidence intervals obtained using both variants of the EM algorithm.

From our experience developing the code to estimate MLMMs, our general impression is that there are often no large differences between the two EM algorithm variants in terms of the quality of the parameter estimates or confidence intervals, and that the first variant is vastly slower to converge than the second. Furthermore we also noticed when the within-unit sample sizes are low that the first variant struggles to estimate models where some of fixed effect variables in the $\boldsymbol{X}_i$ matrices are constant within a unit. This is an important issue to investigate because in medical statistics in particular many variables that are often included in LMMs are constant within a unit - i.e. age and sex of subjects. For this reason the simulation study we now describe will also investigate this effect of variables being constant within a unit or not, since we will focus on other issues in the main factorial experiments in section (5.3). Thus for this section, and for any given model (we shall introduce two shortly), we will compare the quality of parameter estimates in terms of their levels of bias and variability, and the performance of the parameter confidence intervals as measured by coverage probabilities, and confidence interval lengths. For brevity we will refer to the first and second variants of the EM algorithm as EM1 and EM2 respectively.

The two models we will use for these simulations we will call Model 1 and Model 2, and both have $N = 1000$ units, and $G = 3$ components. Model 1 will use a simple random effects covariance structure, and an $AR(3)$ process for the within-unit errors covariance structure. Model 2 will use an unstructured covariance structure for the random effects, and a simple within-unit errors covariance structure. Both models will contain the following variables and fixed effects for $j = 1, ..., G$: an intercept (β_j^0); a factor variable f_1 with two levels and parameters $\beta_j^{f_{11}}$ and $\beta_j^{f_{12}}$; a factor variable f_2 with three levels and parameters $\beta_j^{f_{21}}$, $\beta_j^{f_{22}}$ and $\beta_j^{f_{23}}$; two continuous variables c_1 and c_2 with parameters $\beta_j^{c_1}$ and $\beta_j^{c_2}$ respectively; and time as a continuous variable with parameter β_j^{tc}. We will let the variables c_2, f_1 and f_2 be either constant or not constant within units, whereas c_1 will always vary within a unit.

In terms of the covariance parameters, for Model 2 we have a 2×2 random effects covariance matrix $\boldsymbol{D}_j$ with diagonal elements d_j^{11} and d_j^{22} corresponding to the variances for the random intercept and random effect of time respectively, and off-diagonal element d_j^{21} corresponding to the covariance between these two random effects. Model 2 also has one within-unit variance parameter σ_j^2. For Model 1 there is only $q = 1$ one random intercept with parameter d_j^{11}. Model 1 also has one within-unit variance parameter σ_j^2, and three autoregressive parameters ϕ_j^v, $v = 1, ..., 3$.

For Model 1 we will use only two model versions: CON and NCON, both using $max\text{-}n_i = 20$, which means $n_i = 20$ for all i, since there are autoregressive parameters in the model and so we do not permit missing observations. For Model 2 we will use three model versions which we shall call CON6, NCON6, and CON15 where the numbers denote respectively that $max\text{-}n_i = 6$ and $max\text{-}n_i = 15$. We will also make Model 2 slightly unbalanced so that for each level of $max\text{-}n_i$ there are approximately 5% missing values within each unit. The "CON" and "NCON" means that the variables c_2, f_1 and f_2 were either generated to be either constant or non-constant respectively within units.

5.2.1 Model 1

In this subsection we describe the simulation results of Model 1. We described in Section A.1 that during parameter estimation if we can ensure the estimates of $\boldsymbol{\phi}_g$ always give rise to a stationary AR process then the estimates of $\boldsymbol{C}_i(\boldsymbol{\phi}_g)$ will always be positive-definite. This is true theoretically, however in practice during estimation the elements of $\boldsymbol{\tau}$ had to be kept less equal to a number, m say, such that $|m| < 1$ and $1 - |m| < \epsilon$ for $\epsilon > 0$ being small. When running Model 1 we chose $|m| = 0.999$ which turned out to be too close to 1 which resulted in many replications of the simulations producing a $\boldsymbol{C}_i(\boldsymbol{\phi}_g)$ which was not positive definite numerically. This resulted in many replications being aborted since we automatically tested for this condition. This problem only affected the EM1 method since $\boldsymbol{\tau}$ was obtained using a general optimisation algorithm.

Consequently for EM1 although we aimed for $Nsim = 1500$ we achieved only $Nsim = 1016$ replications for the CON model version and $Nsim = 438$ replications for the NCON model version. To save computation time for EM2 we used $Nsim = 300$ replications for both the CON and NCON model versions. Using a value of $|m| = 0.99$ avoided these issues.

The supplementary materials contain all of the plots of these simulation results, which comprise boxplots of the individual $Nsim$ parameter estimates and confidence interval lengths, and errorbar plots of the coverage proportions computed from the $Nsim$ runs (the errorbars are Binomial confidence intervals). Figures 5.1 through to 5.3 show examples of these plots for ϕ_2^3 which is the third order autoregressive parameter for the second component. Tables A.1 through to (A.4) show the parameter estimate and confidence interval length results averaged over the $Nsim$ runs, and also the coverage probability results.

Figure 5.1: Boxplots of simulated parameter estimates for ϕ_2^3 from model 1 using the EM1 (top two charts) and EM2 (bottom two charts) algorithms. The x axis displays two different versions of model 1: constant (CON) and non-constant (NON) fixed variables, for all estimates (left-hand charts) and excluding estimates outside the 10^{th} and 90^{th} percentiles of both model versions combined (right-hand charts). For EM1 $Nsim = 1016$ for CON, and $Nsim = 438$ for NON. For EM2 $Nsim = 300$ for CON and NON.

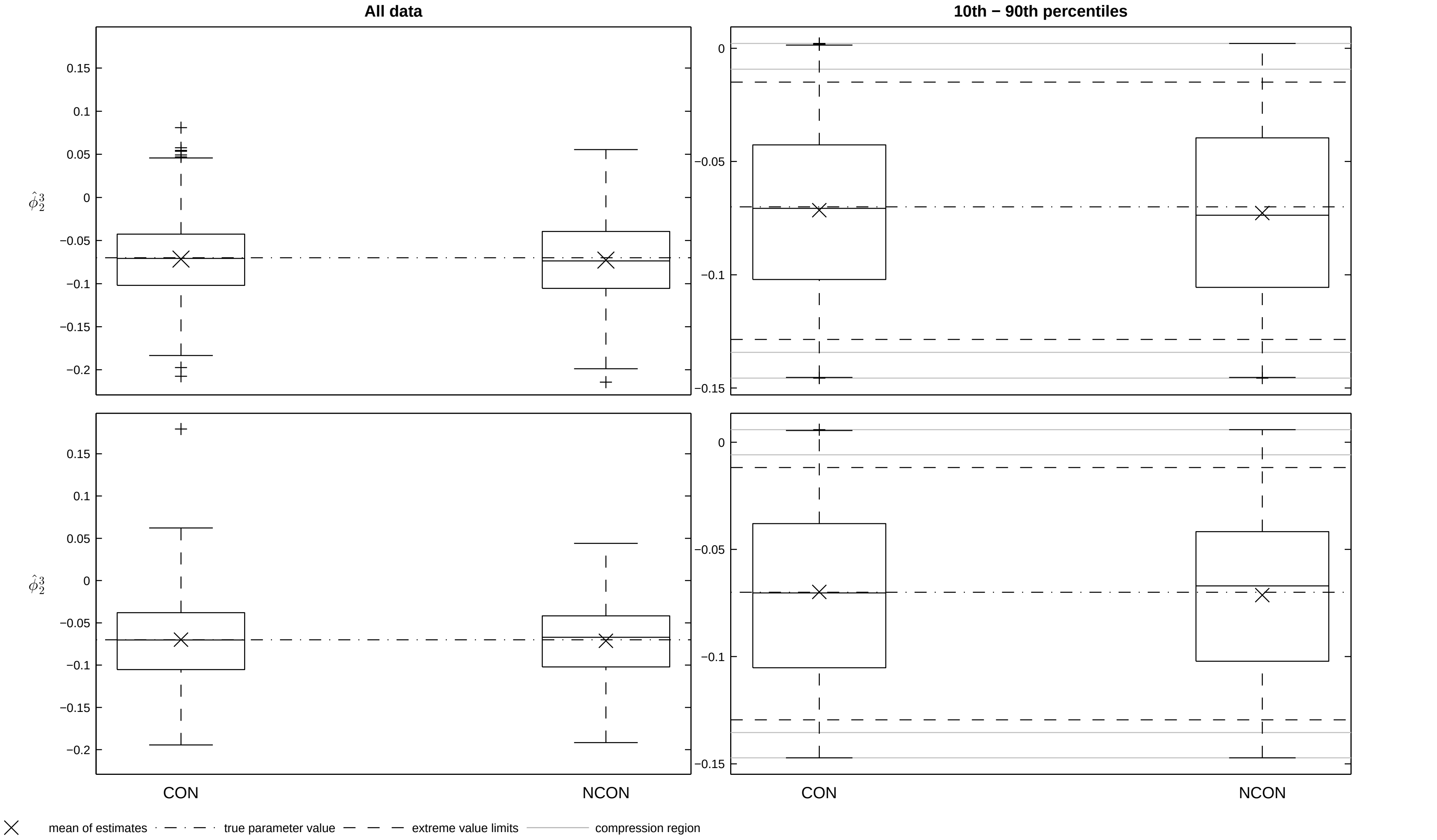


Figure 5.2: Coverage probabilities of simulated parameter estimate confidence intervals for ϕ_2^3 from model 1 with 95% approximate confidence intervals on the proportions. Each chart displays a different method of confidence interval construction on which the coverage probabilities are based. The x axis displays two different versions of model 1: constant (CON) and non-constant (NON) fixed variables. For EM1 $Nsim = 1016$ for CON, and $Nsim = 438$ for NON. For EM2 $Nsim = 300$ for CON and NON.

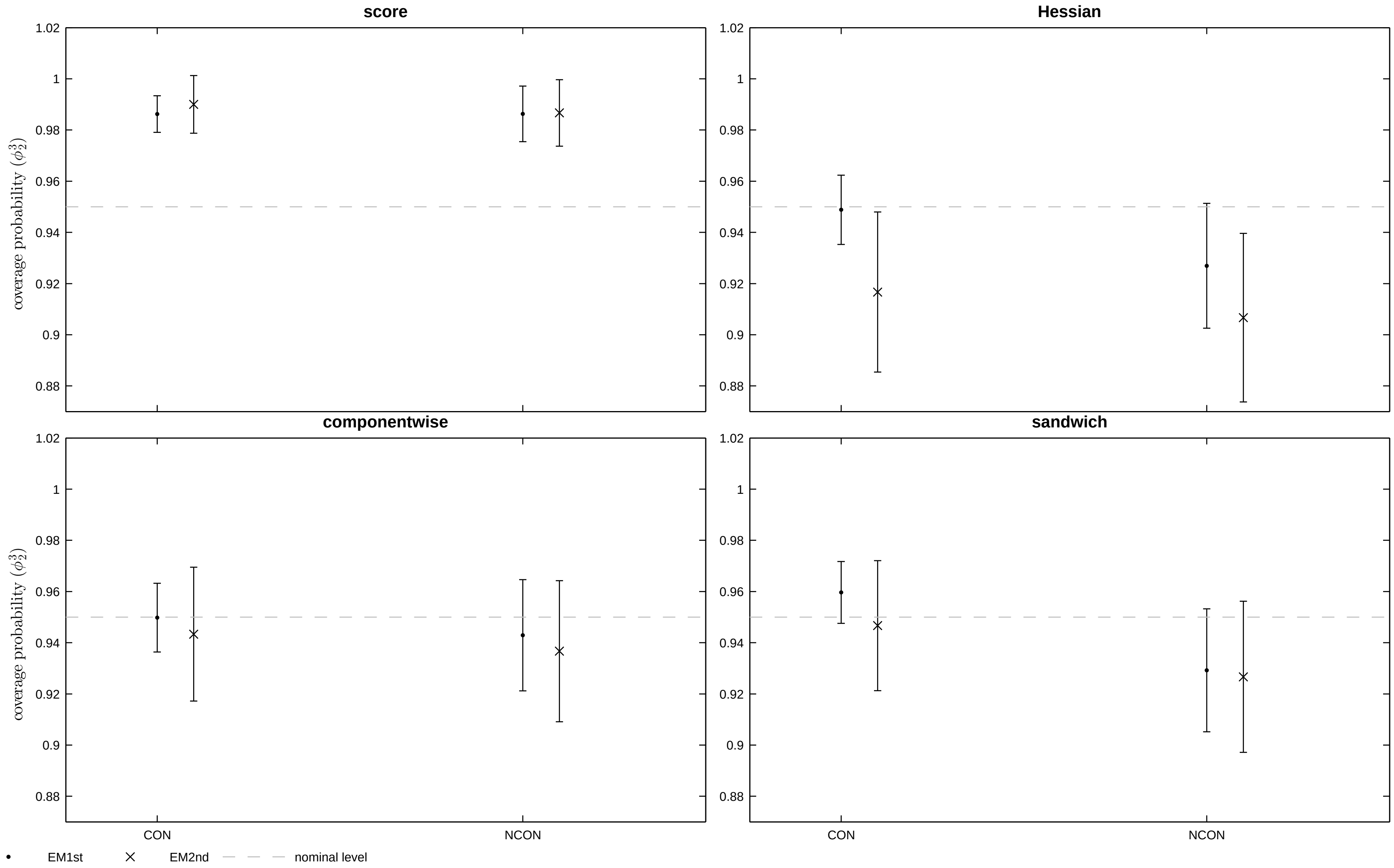
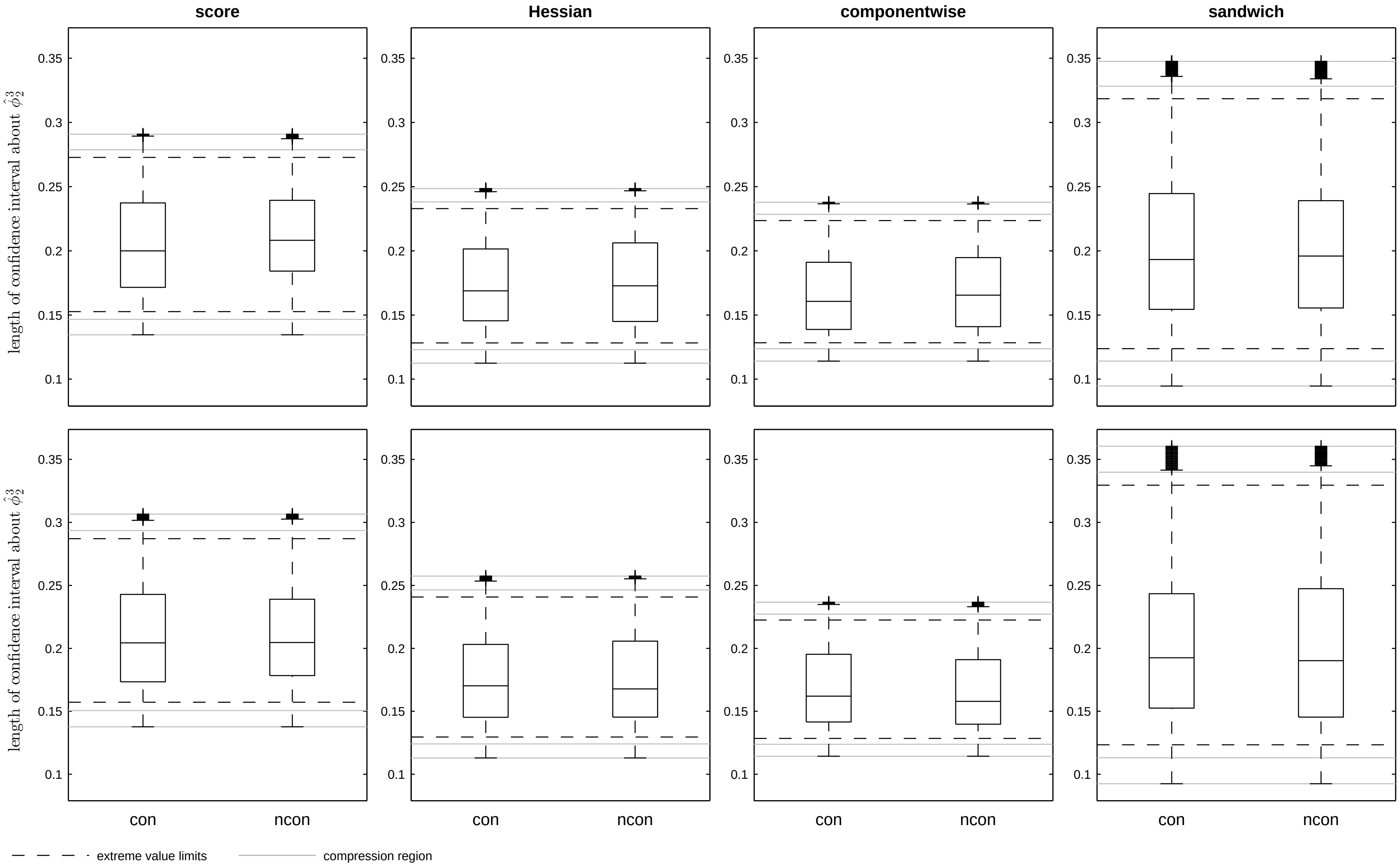


Figure 5.3: Boxplots of confidence interval lengths for ϕ_2^3 from model 1 using the EM1 (top four charts) and EM2 (bottom four charts) algorithms. The x axis displays two different versions of model 1: constant (CON) and non-constant (NON) fixed variables. Due to large variation in the data, for both EM1 and EM2, data outside the 5^{th} and 95^{th} percentiles (calculated using the data from both model versions combined) have been excluded. For EM1 $Nsim = 1016$ for CON, and $Nsim = 438$ for NON. For EM2 $Nsim = 300$ for CON and NON.



We firstly discuss the estimation results. For both model versions and for most model parameters, the parameter estimates were reasonably unbiased, as manifested by low MSEs, for both EM1 and EM2. The exception to this was the estimates of the within-unit variances which were clearly much more biased than the other parameters, particularly for EM2. For each component the ACF for the autoregressive process the three AR parameters collectively define is characterised by an exponential reduction so that by lag 10 the ACF is approximately zero. However despite this rapid reduction in the ACF, certainly the first 3 lags have reasonably high levels of autocorrelation which, if ignored might be attributed by the model instead to the within-unit variance, causing it to become inflated. This might explain why these within-unit variances have not been estimated as well as the other parameters, nor as well as the estimates for the within-unit variances for Model 2 where there were no AR parameters. It may well be that more than $N = 100$ units are required in order to estimate the within-unit variances well in the presence of high or even moderately high levels of autocorrelation. Furthermore if high levels of autocorrelation were being attributed to the within-unit variances then this did not adversely affect the estimation of the autoregressive parameters themselves, which in general were well estimated.

For both EM1 and EM2, and for most of the model parameters, the estimates for the CON model version had higher levels of variation, but similar levels of bias (looking at the mean rather than the median of the estimates) compared to the NCON version. These differences in variation of the estimates for CON compared to NCON are more pronounced for EM2 compared to EM1, primarily because of many outlying values for

CON for EM2. Thus just as for Model 2 it appears that having most of the fixed effects covariates being constant within a unit has a detrimental effect on the quality of the parameter estimation in terms of increasing variability of the estimates. However this effect was not as strong as for Model 2, nor were the estimates more biased for CON compared to NCON as they were for Model 2 (CON6 and NCON6 model versions). This may be because $max\text{-}n_i = 20$ for all i are sufficiently high numbers of within-unit observations to offset to some extent the loss in information that occurs when most of the fixed effects covariates are constant within a unit. Furthermore for Model 2 it was EM1 rather than EM2 that performed the worst when most of the fixed effects covariates were constant within a unit.

For most of the model parameters the coverage probabilities for EM2 appear to be slightly lower compared to EM1, however the ranges of coverage substantially overlap, and so in this sense there are no real differences between EM1 and EM2 in terms of coverage. For both EM1 and EM2 it is also clear that CI1 produces the highest ranges of coverage which tends to be around or sometimes higher than the nominal level. There appears to be no real difference between the ranges of coverage for the other three methods which are often reasonably close to the nominal level (80%-95%), and this result is similar to the one obtained for Model 2. In general the ranges of coverage for NCON were slightly higher than for CON, and this was a much weaker effect than was observed for Model 2 (CON6 versus NCON6).

For the fixed effects there appeared to be no large differences in the confidence interval lengths between the confidence interval methods. This holds too for the covariance parameters with the exception that the confidence interval lengths for $CI4$ were much more variable than the other three methods. For all model parameters there appeared to be no large differences between EM1 and EM2. These results are in contrast to those of Model 2 (CON6 and NCON6 model versions) where large differences in the variability of the confidence intervals between the confidence interval methods, and between EM1 and EM2 were observed. For all confidence interval methods, and for both EM1 and EM2, for the fixed effects the variability in the confidence interval lengths was much larger, and the median lengths reasonably higher for CON compared to NCON - there were no such large differences between CON and NCON for the covariance parameters. These differences between CON and NCON are a weaker version of the results observed for Model 2 (CON6 and NCON6 model versions).

In conclusion parameter estimates were reasonably unbiased for most model parameters for both EM1 and EM2. The exception to this was the within-unit variances which were estimated in some cases rather poorly, which may be because of the influence of high levels of autocorrelation in the within-unit errors. EM2 produced the highest levels of bias in these parameter estimates for the CON model version. Excluding the within-unit variances, parameter estimates displayed more variation but similar levels of bias for CON compared to NCON. No large differences could be observed between EM1 and EM2 for either the coverage or confidence interval length results. Just as for Model 2, $CI1$ produced the best coverage results, often attaining the nominal level. The other three methods were not far behind in producing only slightly lower ranges of coverage. Again as with Model 2 it is notable in this respect that $CI3$ performs just as well as $CI2$ and $CI3$, and not too worse than $CI1$.

5.2.2 Model 2

In this subsection we describe the simulation results of Model 2. For EM1 we aimed for $NSim = 1000$ replications for all three model versions, but obtained $Nsim = 999$ for the CON6 and CON15 model versions, and $Nsim = 964$ for the NCON6 model version. The reason for these lost replications is due to the covariance matrix of the responses $\boldsymbol{V}_i(\boldsymbol{\zeta}_g)$ occasionally not being positive definite since we tested for this condition during estimation and discarded these replications. For EM2 we aimed for and achieved $Nsim = 1000$ replications for all model versions. The supplementary materials contain all of the plots of these simulation results, which comprise boxplots of the individual $Nsim$ parameter estimates and confidence interval lengths, and errorbar plots of the coverage proportions computed from the $Nsim$ runs (the errorbars are Binomial confidence intervals). Figures (5.4) through to (5.6) show examples of these plots for $\boldsymbol{\beta}_2^0$ which is the model intercept for the second component. Tables (A.5) through to (A.10) show the parameter estimate and confidence interval length results averaged over the $Nsim$ runs, and also the coverage probability results.

Figure 5.4: Boxplots of simulated parameter estimates (Nsim=1000) for β_2^0 from model 2 using the EM1 (top two charts) and EM2 (bottom two charts) algorithms. The x axis displays three different versions of model 2: constant fixed variables/max-n_i=6 (CON6), constant fixed variables/max-n_i=15 (CON15), and non-constant fixed variables/max-n_i=6 (NCON6), for all estimates (left-hand charts) and excluding estimates outside the 10^{th} and 90^{th} percentiles of all model versions combined (right-hand charts).

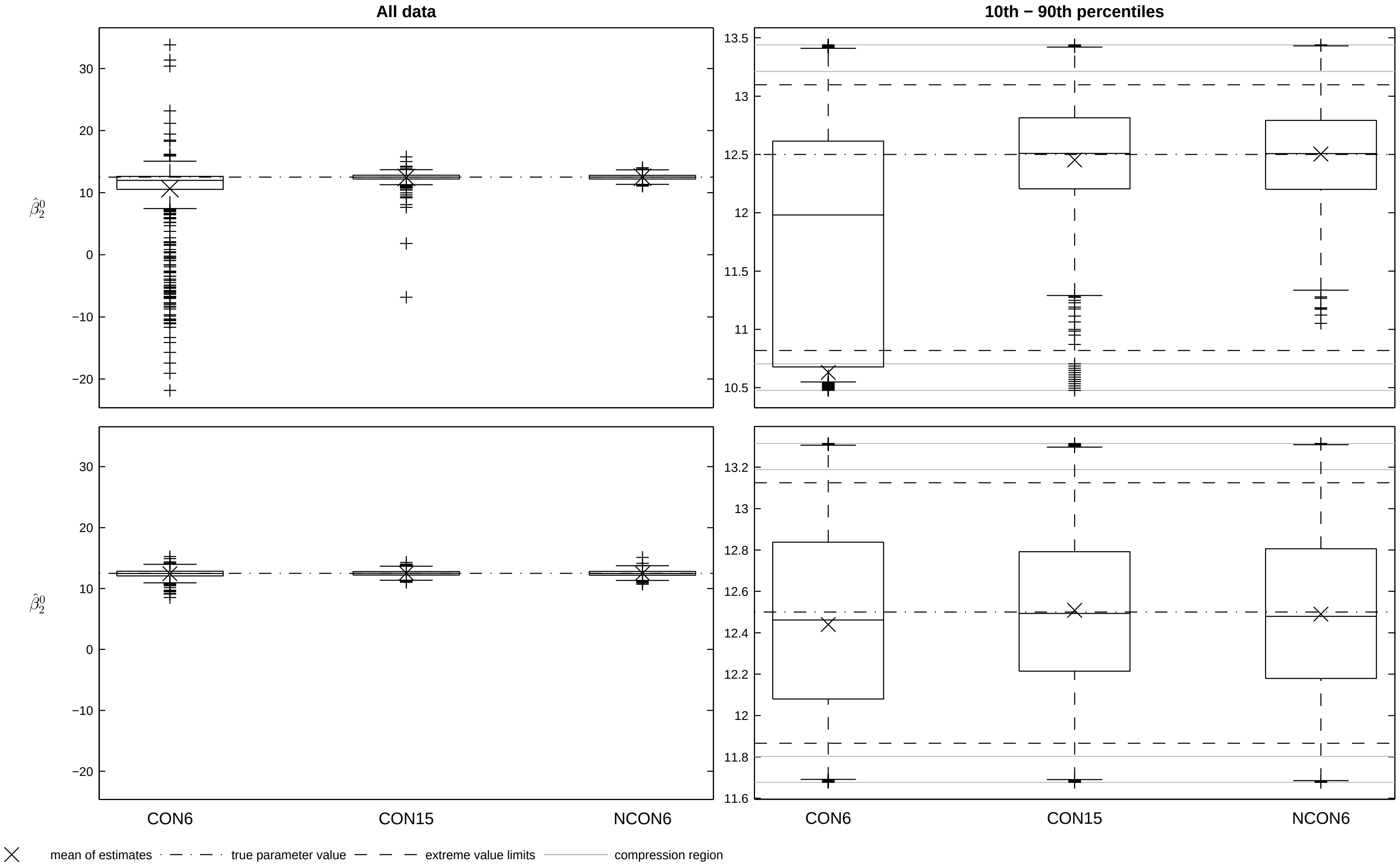


Figure 5.5: Coverage probabilities of simulated parameter estimate confidence intervals (Nsim=1000) for β_2^0 from model 2 with 95% approximate confidence intervals on the proportions. Each chart displays a different method of confidence interval construction on which the coverage probabilities are based. The x axis displays three different versions of model 2: constant fixed variables/max-n_i=6 (CON6), constant fixed variables/max-n_i=15 (CON15), and non-constant fixed variables/max-$n_i = 6$ (NCON6). For EM1, and due to low coverage, the constant fixed variables/max-$n_i = 6$ data point has been omitted.

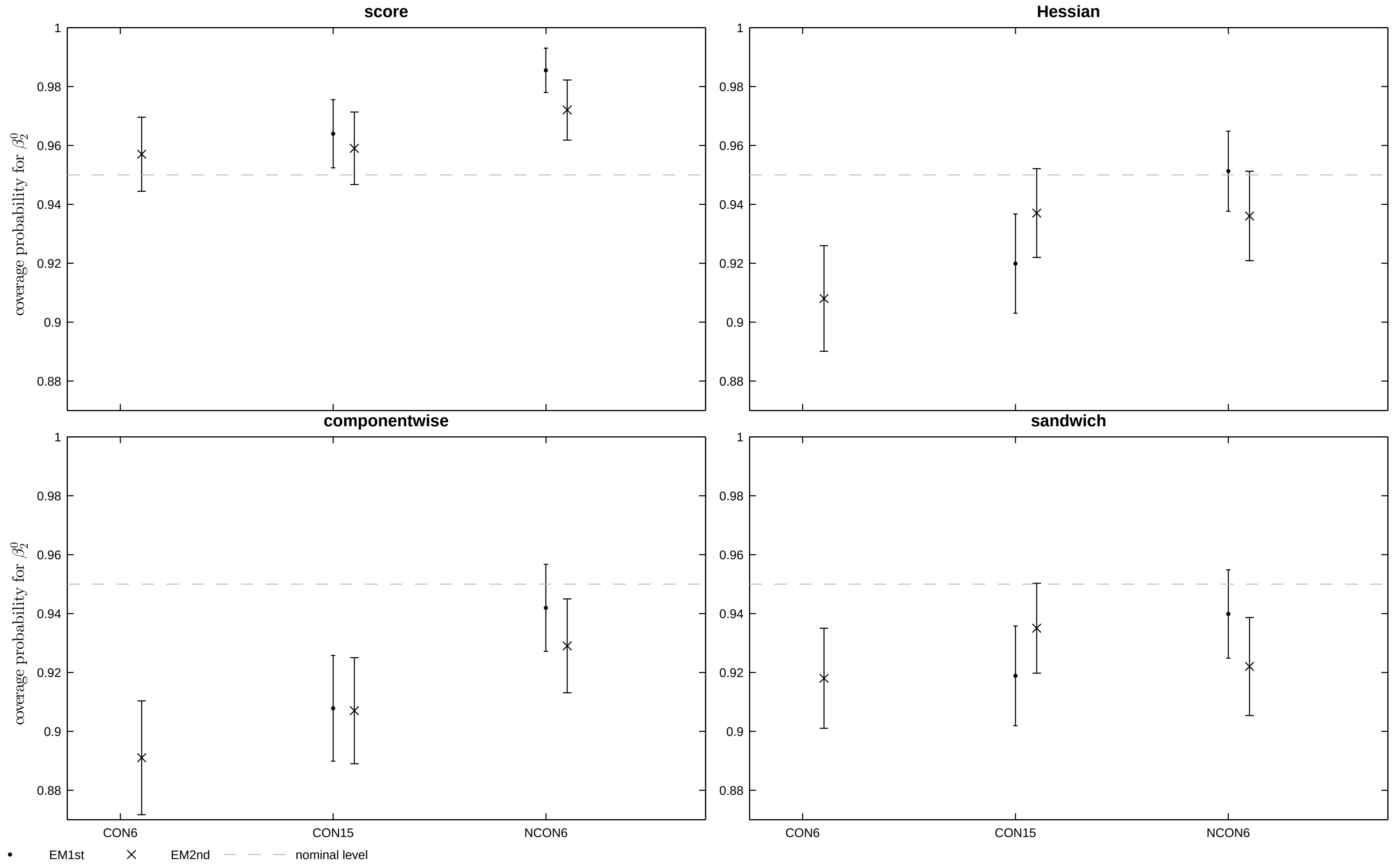
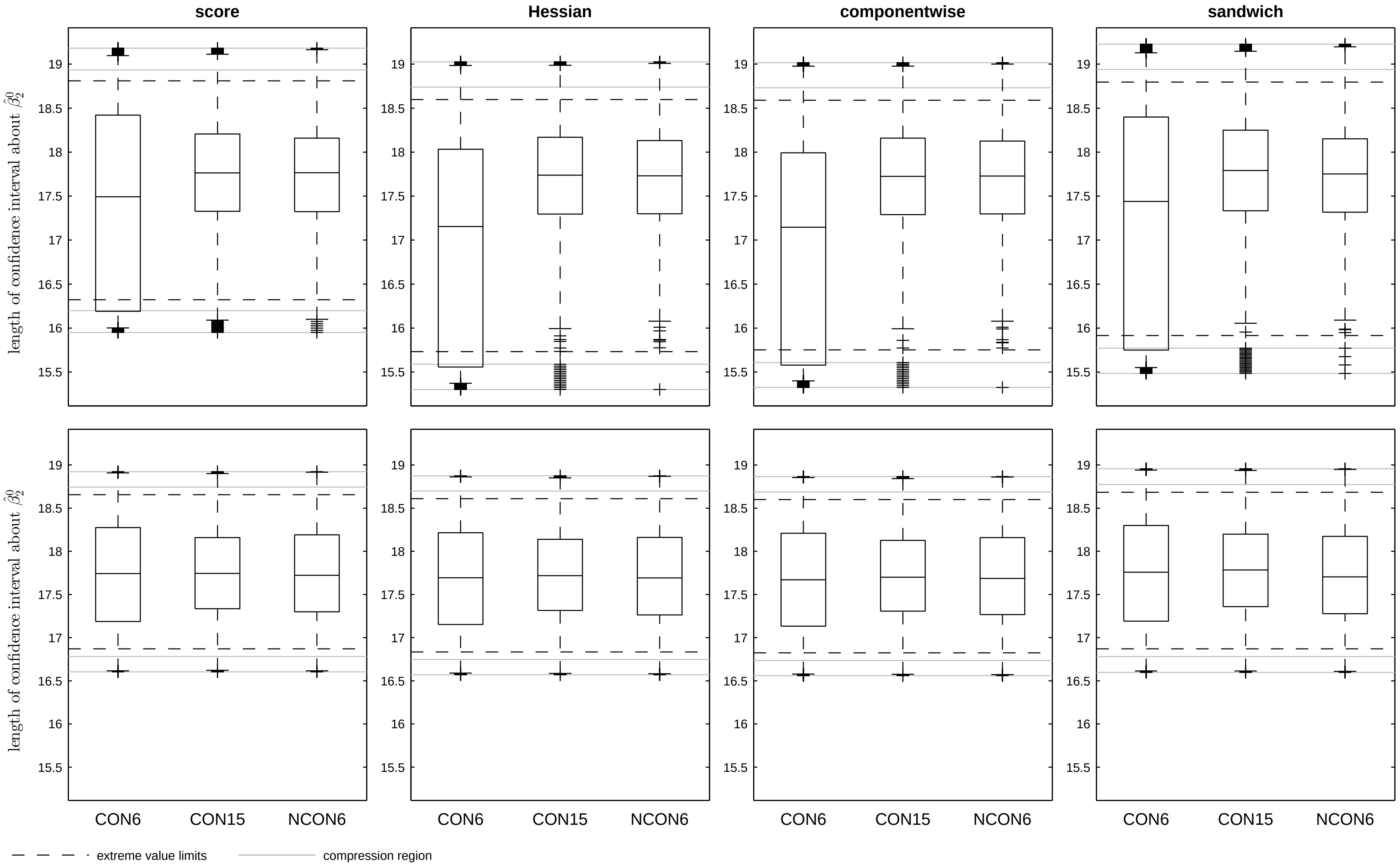


Figure 5.6: Boxplots of confidence interval lengths (Nsim=1000) for β_2^0 from model 2 using the EM1 (top four charts) and EM2 (bottom four charts) algorithms. The x axis displays three different versions of model 2: constant fixed variables/max-n_i=6 (CON6), constant fixed variables/max-n_i=15 (CON15), and non-constant fixed variables/max-n_i=6 (NCON6). Due to large variation in the EM1 data, for both EM1 and EM2, data outside the 5th and 95th percentiles (calculated using the data from all model versions combined) have been excluded.



We will firstly discuss the parameter estimation results, and in this respect we will start with the comparison of the two model versions CON6 and CON15. It is clear that the most striking feature of these results is that for EM1 and all the parameters that the variation in the parameter estimates for CON6 was much larger than for CON15. In addition many of the estimates for CON6 were reasonably biased, whereas for CON15 they were reasonably unbiased. Similar differences between these two model versions can be observed too for EM2, however the size of these differences is much smaller than for EM1. For EM1 but not EM2 it is clear that the increases in bias with which the parameter estimates for CON6 displayed compared to those for CON15 were particularly high for the parameters associated with the factor variables. Thus it seems for EM1 that it is much harder to estimate factor variables than continuous ones at low values of $max\text{-}n_i$ when the variables are constant within a unit. Furthermore it is clear that for both EM1 and EM2 that increasing $max\text{-}n_i$ improves how well the parameters for all the model parameters can be estimated.

The results we have just described for the CON6 versus the CON15 model version were also observed for the CON6 versus the NCON6 model version comparison. This shows for EM1 that the poor estimation of the parameters at low values of $max\text{-}n_i$ when the variables with which they are associated are constant within a unit, disappear when the variables are permitted to vary within a unit. This effect is particularly strong for the factor variable fixed effects parameters. Of course this observed effect makes sense because allowing the variable with which a parameter is associated to vary within units gives more information on the variable-response relationship compared to restricting it to be constant within a unit. It is not clear however why this effect for the fixed effect parameters associated with the continuous variables was less pronounced than for the factor variable fixed effect parameters. Furthermore we are not aware of these problems with the LMM (which is not estimated with the EM algorithm), and so we suspect this problem is specific to the first variant of the EM algorithm rather than a characteristic of mixture models in general.

These differences between EM1 and EM2 in terms of how poorly EM1 estimates parameters that are associated with variables that do not vary within a unit, might be explained by looking at the "poor" runs that occurred out of the $Nsim$ simulations within the CON6 model version. We define "poor" runs here as those runs producing classification errors greater than 10% for any component. Firstly we see that for EM2 only approximately 6% of the runs were poor compared to 82% of the runs for EM1. For EM2 all of these poor runs converged, whilst for EM1 only a few failed to converge. Thus in both cases poor runs were generally characterised by the EM algorithm simply converging to poor final estimates, usually following poor initial values in the sense of being far from the true parameters, however for EM1 the frequency of this occuring was much greater.

A check of some of the poor runs for the EM1 simulations shows that often one of the mixing proportions for the three components has been estimated to be approximately zero. The majority of the units that belong to this "zero" component have been incorrectly assigned to only one of the remaining two components, whilst a few of the units from the other two components have been incorrectly assigned to the zero component. Let $g^0 \in I_G$ denote the zero component, and $g' \in I_G$, $g' \neq g^0$ denote the component to which most or all of the units from component g^0 have been assigned. Then another observation of these poor runs is that one or more of the elements of $\boldsymbol{D}_{g'}$ have been initialised and estimated to be very large, and where the final estimates are reasonably similar to the initial ones. Presumably due to the inclusion of many of the units from component g^0, a number of the elements of $\boldsymbol{\beta}_{g'}$ have also been estimated poorly.

A possible explanation of how the majority of the units from component g^0 might be assigned incorrectly to component g' is that the first variant of the EM algorithm is in fact an ECM rather than a standard EM algorithm. As we described in section 2.2, for any given parameter, and at each ECM iteration of a EM iteration, the ECM algorithm updates a sub-vector of $\boldsymbol{\theta}$ conditional on estimates of the other sub-vectors of $\boldsymbol{\theta}$ from the previous ECM iterations. Thus on the $(s+1)^{th}$ iteration of the EM algorithm, both variants update the parameter estimates conditional upon $\boldsymbol{\theta}^{(s)}$, but the second variant does this indirectly through the posterior probabilities, whilst the first variant does this more directly by conditioning on the sub-vectors of $\boldsymbol{\theta}^{(s)}$. We now describe why this might explain these results.

If s denotes the s^{th} iteration of the EM algorithm, suppose that $\hat{\boldsymbol{\psi}}_g^{(s)}$ is such that $\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)})$ has large estimated variances and covariances, and further suppose that $\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_g^{(s)}$ is large. Regardless of whether $\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)})$ is close to the true random effects covariance matrix $\boldsymbol{D}(\boldsymbol{\psi}_g^{(s)})$ or not, if we believe $\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)})$ is the true covariance matrix, and if we believe unit i belongs to component g, we would expect $\boldsymbol{U}_i$ to vary greatly about its mean value, often taking large values. Under these assumptions we would put the probability to be high that $\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_g^{(s)}$ will be large even if $\hat{\boldsymbol{\beta}}_g^{(s)}$ is close to the true parameter $\boldsymbol{\beta}_g$. Thus conditional on knowing $\hat{\boldsymbol{\psi}}_g^{(s)}$, and without knowing the value of $\boldsymbol{U}_i = \boldsymbol{u}_i$, it may be difficult to decide if $\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_g^{(s)}$ being large is due to the high variation in the random effects, or because unit i has been incorrectly assigned to component g.

Using this reasoning, because on each ECM step we have that the parameters are updated conditional on those parameters that have already been updated, it might be that $\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)})$ having large estimated variances and covariances causes the EM algorithm to have no reason to assign unit i to another component, and so "tolerates" this unit being assigned to it. This situation may manifest itself mathematically by small gradients of $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$ with respect to all of the parameters on the ECM steps, and so this could result in little parameter updating. If the classification of units to components is poor then the EM algorithm may then nonetheless still converge to incorrect estimates.

We now turn our attention to the coverage results. Because of the poor estimation results for EM1 for the CON model version, for all of the model parameters the coverage probabilities for all of the confidence interval methods are low (approximately 70%). For this reason this model version for EM1 has been omitted from the plots. Accordingly when we discuss the CON6 model version in terms of coverage results, we refer only to EM2. These coverage results show that there is no difference in the ranges of coverage produced when using EM1 or EM2. For all the model versions the $CI1$ method produces the superior confidence intervals in terms of producing ranges of coverage that often intersect the nominal level, and there are no significant differences between the other three methods which produce slightly lower ranges of coverage. The fact however that $CI3$ is just as good as $CI2$, and $CI4$, and not too much worse than $CI1$, is an important result. This is because the confidence intervals used by $CI3$ are based upon the LMM information matrix, whereas the intervals produced by the other three methods are based upon approximations to the MLMM information matrix.

For most of the model parameters and all the confidence interval methods, and for EM2 only, the coverage ranges for CON6 are clearly lower than for NCON6. Similarly for most of the model parameters and all the confidence interval methods, and for both EM1 and EM2, the coverage ranges for CON15 are often lower than for NCON6. These results show that having variables in the $\boldsymbol{X}_i$ matrices that are constant within units leads to a reasonable degradation in the performance of the confidence intervals produced by all the confidence interval methods. For EM2 only it is also clear that for most parameters and all the confidence interval methods, the ranges of coverage for CON15 were reasonably higher than for CON6. This shows that more within-unit information can to some extent offset the loss in information associated with having some fixed effects covariates being constant within units.

The confidence interval length results show that the variation in lengths for EM1 is much higher than for EM2, and that this variation is extremely high for the lengths produced by $CI1$ in the $CON6$ model version. This extreme variation occurs as a result of a few very large parameter estimate standard errors when the estimates have converged to poor values. In contrast, for EM1 and for $CON6$, a reasonable number of the confidence intervals for $CI2$ were complex valued as a result of negative standard errors. This can happen because $CI2$ is based on the Hessian matrix, which is only guaranteed to be a covariance matrix in the limit as N tends to infinity. So for finite sample sizes we might have a non-positive definite Hessian which could give rise to negative diagonal entries of the inverse Hessian. Thus it seems as though the confidence intervals for $CI1$ and $CI2$ behave quite differently when the model estimates have converged to poor values. The confidence intervals for $CI3$ and $CI4$ did not suffer from either of these major drawbacks. Despite the larger variation in confidence interval lengths for $CI1$ in the $CON6$ model version, there were no differences in the median confidence interval lengths between either the three model versions, or the four confidence interval methods. However the main result here is that the propensity of some of the confidence intervals of $CI1$ and $CI2$ to be either very long or complex valued respectively when parameter estimates are poor, means we can argue that $CI3$ and $CI4$ produce better confidence intervals in the sense of being more invariant to estimation quality.

In conclusion it is clear when estimation was difficult (as in the $CON6$ model version) that EM1 produced much more biased parameter estimates, and that the estimates displayed considerably more variation than the other two model versions. This was also observed for EM2 but to a much lesser extent. Thus the influence of having all fixed effects covariates varying within units, and increasing the within-unit sample sizes was to improve the quality of the estimates. When parameter estimation was easier (CON15 and NCON6) EM1 and EM2 produced reasonably similar estimates. The reasons why EM1 produces poor quality estimates when estimation is difficult may be due to the fact that the first variant of the EM algorithm is an ECM algorithm that conditions on the sub-vectors of $\boldsymbol{\theta}^{(s)}$ in order to derive updated parameter estimates on the $(s+1)^{th}$ iteration. The best coverage was obtained by $CI1$, where the ranges of coverage often attained the nominal level, whilst the other three methods produced reasonably similar ranges of coverage that were slightly lower than the nominal level. The effect of having fixed effects covariates that are constant within a unit was to reduce the levels of the ranges of coverage. There were no differences in the median confidence interval lengths between the four confidence interval methods or between EM1 and EM2, however the variation in the confidence interval lengths produced by EM1 was much larger than for EM2. The confidence intervals produced by $CI1$ and $CI2$ have the propensity to become either very long or complex valued respectively when estimation is difficult.

Finally it is also important to point out that not only did EM1 not cope very well for CON6 when variables in the $\boldsymbol{X}_i$ matrices are constant within units, but it was also very slow to converge compared to EM2. In this respect all the simulations were run on a fairly decent workstation: a Microsoft Windows machine running a 64-bit operating system with a reasonably modern (2012) quad core processor running at 2.4GHz, and with 16 GB of RAM. For CON6 on average (over all the 1000 runs) it took EM1 144 seconds to fit an individual run, whilst for EM2 it took only 7.2 seconds. Similarly for CON15 on average it again took EM1 144 seconds to fit an individual run, whilst for

EM2 it took 5.3 seconds. Thus there is a huge increase in computation time for EM1 compared to EM2.

5.3 Factorial simulations

In these simulations, which we call “factorial simulations”, we will use three models which we will denote by Model 1-Model 3. We call these simulations “factorial” simulations because for each model we will use $k = 6$ variables (simulation variables), most of which have two levels, where we will perform $Nsim$ replications at different combination of the simulation variables, where for any given model each combination of the factorial variables (to be introduced shortly) will be called a model version. Since we will use all the combinations of these simulation variables then our simulations represent a completely crossed designed experiment, and so represents a factorial experiment. In total Model 1 and Model 2 will have 128 different combinations of the simulation variables - i.e. 128 different model versions, whilst Model 3 will have 64 model versions.

Model 1 will have $G = 3$ components, an unstructured random effects covariance matrix with $q = 2$ random effects, and a simple within-unit covariance structure. Model 2 will have $G = 2$ components, an unstructured random effects covariance matrix with $q = 2$ random effects, and an $AR(2)$ within-unit covariance structure. Model 3 will have $G = 4$ components, and a simple covariance structure for both the random effects and the within-unit variances.

Model 1 will have the following fixed effects and associated variables for $j = 1, ..., G$:

an intercept β_j^0; a factor variable f_1 with two levels and parameters (including redundant parameters) $\beta_j^{f_{11}}$ and $\beta_j^{f_{12}}$; a factor variable f_2 with three levels and parameters (including redundant parameters) $\beta_j^{f_{21}}$, $\beta_j^{f_{22}}$ and $\beta_j^{f_{23}}$; two continuous variables c_1 and c_2 with parameters $\beta_j^{c_1}$ and $\beta_j^{c_2}$ respectively; and time tc as a continuous variable with parameter β_j^{tc}. The random effects for Model 1 will be the intercept and the time variable tc. Thus a 2×2 random effects covariance matrix $\boldsymbol{D}_j$ is obtained with diagonal elements d_j^{11}, and d_j^{22} corresponding to the variances for the random intercept and random effect of time respectively, and off-diagonal element d_j^{21} corresponding to the covariance between these two random effects. Model 1 also has one within-unit variance parameter σ_j^2.

We will use the same fixed effects and variables for Model 2 as we did for Model 1 (although the actual covariate data will be generated to be different). Similarly the same random effects will be used for Model 2 as they were for Model 1. Model 2 also has one within-unit variance parameter σ_j^2, and two autoregressive parameters ϕ_j^v, $v = 1, ..., 2$. For Model 3 we again use the same fixed effects and variables as for Model 1, but we also include an interaction variable between f_1 and f_2 which has parameters $\beta_j^{f_{11}*f_{21}}$ and $\beta_j^{f_{11}*f_{22}}$. Model 3 has only one random effect covariance parameter d_j^{11}, and one within-unit variance parameter σ_j^2.

In terms of the simulation variables, for the many that take on only two values these values were chosen to represent low and high values, where for most of these variables the "low" and "high" settings are self-explanatory and are denoted by L and H respectively. In contrast for some variables the meaning of "low" and "high" is less clear, and this terminology is retained only for consistency with the other variables. For all the simulation variables we will write the values the variable can take in an order such that the first listed value represents intuitively the setting that should make model estimation easiest. We will call this first setting the reference setting, or the reference value. Accordingly setting all the simulation variables to their reference settings should yield a model version that is the easiest to estimate out of all the other combinations. This model version we will call the reference model version.

The variables common to all models are: unit sample sizes ($N \in \{H, L\}$); the maximum within-unit sample size across all the N units ($max\text{-}n_i \in \{H, L\}$); the within-unit variances ($\sigma^2 \in \{L, H\}$)); the unbalancedness of the mixing proportions (π-*unbalance* $\in \{L, H\}$), where $L := BAL$ means balanced (i.e a low level of unbalancedness) and $H := UNBAL$ means unbalanced (i.e a higher level of unbalancedness than BAL). For Model 1 and Model 2 we will have a simulation variable that represents the random effects covariance matrix, or more specifically a combination of the diagonal elements being either low or high, and the off-diagonal element being either positive or negative. Denoting this variable by D, the four values this variable can take are $D \in \{LPOS, LNEG, HPOS, HNEG\}$ where $LPOS$ and $LNEG$ mean low and positive, and low and negative respectively, and where $HPOS$ and $HNEG$ mean high and positive, and high and negative respectively. For Model 3 there is no off-diagonal element of D, and so this variable will be set at just a low or high setting - i.e. $D \in \{L, H\}$.

For Model 2 we will have a simulation variable that represents the amount of autocorrelation present in the within-unit errors, which we will control with the ACF of the autoregressive parameters, which in turn will be controlled with the autoregressive parameters. We will denote this variable by $ACF \in \{L, H\}$. The high setting H will correspond to slowly declining ACF whereby the ACF reaches near zero approximately at lag 10. The low setting L will correspond to a comparatively rapidly declining ACF, usually such that the ACF function is near zero by approximately lag 4. We note that although the $ACF = L$ should make estimation easier regardless of the $max\text{-}n_i$ setting, 4 lags will represent a smaller proportion of the $max\text{-}n_i$ lags when $max\text{-}n_i = H$, compared to when $max\text{-}n_i = L$, and so the greatest benefit of $ACF = L$ may well be observed for the larger within-unit sample sizes. For similar reasons the anticipated detrimental effect of $ACF = H$ may well be most strongly observed when $max\text{-}n_i = L$. The specific values chosen for the simulation variables can be found in table 5.1, and the specific values chosen for the fixed effects parameters can be found in table 5.2.

In subsection 3.4.1 we constructed an index of separation to quantify how well separated two corresponding parameters in different components are in terms of how easily an estimation procedure should be able to tell them apart. In general for two components indexed by $g, g' \in I_G$, making the fixed effects parameters of the two components to be further and further apart from each other will increase the separation indices between this pair of fixed effects, which in turn will increase $SI(g, g')$. Furthermore this increase in separation will occur no matter how large the fixed effects become in magnitude. In contrast making covariance parameters large in order to make them different from corresponding parameters in other components increases the noise present in the data within those components, which in turn reduces the separation between these two components.

For these reasons we see that our manipulation of the covariance parameters as part of the factorial simulations will also result in the separation indices of the components being affected. This is preferable to specifying *a priori* a given level of separation for each of the components, and to then choose values for the model parameters to achieve these pre-specified levels of separation. This is because it would not only be very tedious to do this, but there will in general be a uniqueness problem in that many different choices of parameters will lead to the same separation indices. Furthermore during code development casual observation of the separation indices for different models revealed that often quite different amounts of component separation were required to obtain the same level of difficulty of parameter estimation. Thus the base level of component separation and therefore estimation difficulty appears to be very model-specific, although it was also clear that for all models increasing component separation made parameter estimation easier.

Even though our approach to manipulating the separation indices was indirect, we nonetheless attempted to obtain a specific range of separation indices across the different model versions - from not well-separated to very well separated. Firstly we chose the reference values for the simulation variables and the values of the fixed effects parameters in order to obtained high enough levels of separation between the G components such that the resulting models were easy to estimate, as reflected by low mean square errors of parameter estimates, and zero classification errors. Although no attempts were made to ensure these default model versions achieved the easiest estimation problem possible, it suits our purposes here to consider these as the "gold standard" models that should achieve the best performing parameter estimators and classification of units to components.

Given this process of choosing the default model versions, setting the non-reference levels of the simulation variables to any other values that are "worse" for estimation than the reference levels then served to reduce the separation indices, and in turn increase the difficulty of estimation compared to the "gold standard". In this respect we will make clear in the results that one of the main findings, and perhaps not surprisingly, is that it is very easy to pick values of the simulation variables so as to produce components that are extremely close together, in terms of having small or even negative separation indices. For reasons we described at the start of this chapter, we chose instead to set the non-reference levels of the simulation variables in such a way as to produce moderately difficult rather than very difficult models to estimate. In this way we have "calibrated" the models in such a way as to permit us to examine the effects of the simulation variables on statistical inference, or equivalently the effects of component separation on statistical inference, without specifying models that would give very poor results.

Table 5.1: Simulation parameter settings for all models (L =low, H =high, BAL =balanced, $UNBA$ =unbalanced, $LPOS$ =low and positive, $LNEG$ =low and negative, $HPOS$ =high and positive, $HNEG$ =high and negative).

		Model 1		
Simulation parameter	$\boldsymbol{L}$	$\boldsymbol{H}$		
N	100	1000		
$max\text{-}n_i$	5	10		
$\sigma^2 : (\sigma_1^2, ..., \sigma_G^2)$	(1.9,1.8,1.75)	(9.5,9,8.75)		
$n_i\text{-}unbalance : E(X) = \%$ of $max\text{-}n_i$	95	65		
$ACF : (\phi_1), (\phi_2)$	-	-		
	$\boldsymbol{BAL}$	$\boldsymbol{UNBA}$		
$\pi\text{-}unbalance : (\pi_1, ..., \pi_G)$	(0.333,0.333,0.333)	(0.2,0.4,0.4)		
	$\boldsymbol{LPOS}$	$\boldsymbol{LNEG}$	$\boldsymbol{HPOS}$	$\boldsymbol{HNEG}$
$D : (\mathrm{v}(D_1)), ..., (\mathrm{v}(D_G))$	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)
		Model 2		
Simulation parameter	$\boldsymbol{L}$	$\boldsymbol{H}$		
N	100	500		
$max\text{-}n_i$	10	15		
$\sigma^2 : (\sigma_1^2, ..., \sigma_G^2)$	(1.3,1.2)	(8.3,8.2)		
$n_i\text{-}unbalance : E(X) = \%$ of $max\text{-}n_i$	-	-		
$ACF : (\phi_1), (\phi_2)$	(0.4,-0.1),(0.38,-0.12)	(0.4,0.1),(0.38,0.12)		
	$\boldsymbol{BAL}$	$\boldsymbol{UNBA}$		
$\pi\text{-}unbalance : (\pi_1, ..., \pi_G)$	(0.5,0.5)	(0.1,0.9)		
	$\boldsymbol{LPOS}$	$\boldsymbol{LNEG}$	$\boldsymbol{HPOS}$	$\boldsymbol{HNEG}$
$D : (\mathrm{v}(D_1)), ..., (\mathrm{v}(D_G))$	(1,0.5,2.1),(2,0.9,1.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(5,4.5,10.5),(10,5.5,7.5)	(5,-4.5,10.5),(10,-5.5,7.5)
		Model 3		
Simulation parameter	$\boldsymbol{L}$	$\boldsymbol{H}$		
N	100	1000		
$max\text{-}n_i$	6	10		
$\sigma^2 : (\sigma_1^2, ..., \sigma_G^2)$	(1,1.2,1.1,0.9)	(6,6.2,6.1,6.9)		
$n_i\text{-}unbalance : E(X) = \%$ of $max\text{-}n_i$	95	65		
$ACF : (\phi_1), (\phi_2)$	-	-		
	$\boldsymbol{BAL}$	$\boldsymbol{UNBA}$		
$\pi\text{-}unbalance : (\pi_1, ..., \pi_G)$	(0.25,0.25,0.25,0.25)	(0.15,0.2833,0.2833,0.2833)		
	$\boldsymbol{L}$	$\boldsymbol{H}$		
$D : (D_1), ..., (D_G)$	(0.8,0.3,0.5,0.7)	(5.8,5.3,5.5,5.7)		

Table 5.2: Simulation settings for the fixed effects for all models.

	Model 1			**Model 2**		**Model 3**			
fixed effect	comp 1	comp 2	comp 3	comp 1	comp 2	comp 1	comp 2	comp 3	comp 4
β^0	5	15	30	8	-1	-20	18	25	-10
β^{c1}	-2	-3	-0.5	1.2	1.5	-2	-3	-0.5	-2.5
β^{c2}	3	1.2	5.5	2	1.8	3	1.5	3.8	4.8
β^{f11}	1	3	6	-1	-1.5	0.5	-0.5	3	2
β^{f12}	-1	-3	-6	1	1.5	-0.5	0.5	-3	-4
β^{f21}	4	1	9	4.5	7	1	-1	1	-0.5
β^{f22}	7	4	-6	4	7	1.5	-0.9	1.2	2.5
β^{f23}	3.4	6	1.4	3.4	6	2	-4	1.4	4
$\beta^{f11*f21}$	-	-	-	-	-	2	-2	-2	-2.7
$\beta^{f11*f22}$	-	-	-	-	-	4	-4	-3	-4.3
$\beta^{f11*f23}$	-	-	-	-	-	6	-6	-4	5.4
$\beta^{f12*f21}$	-	-	-	-	-	1.5	-1.5	-3	-1.2
$\beta^{f12*f22}$	-	-	-	-	-	2.5	-0.5	-1	5.5
$\beta^{f12*f23}$	-	-	-	-	-	1.5	-1.5	3	-2
β^{tc}	4.5	-2.5	0.5	0.5	1.5	-	-	-	-

5.3.1 Model 1

We first describe the simulation results for the MSE of the mixture model parameter estimators. Looking at the estimate plots in the supplementary materials, as expected it is clear that for most mixture model parameters estimation was better (estimates had less bias and variability) for $\sigma^2 = L$ compared to $\sigma^2 = H$, and that this relationship was stronger for $N = H$ compared to $N = L$. Furthermore whilst all of the parameters seemed to be estimated worst when $N = H$, and $\sigma^2 = H$ together, the random effects covariance parameters appear to be estimated particularly poorly at these simulation variable settings. It also appears that the balancedness of the mixing proportions does not influence these relationships. In general the estimation quality for the mixture model parameters was good, particularly when $N = H$ and $\sigma^2 = L$.

We now look at the robust model M-estimates shown in table 5.4 when MSE is the response variable. Since the scale of the MSE will to some extent be determined by which parameter in the mixture model we are looking at, it is not surprising to see the *param* main effect featuring amongst this strongest set of parameters. In this respect the *param* effects for π, β^{c_1}, β^{c_2}, and β^{tc} show that the MSE for the estimators of these mixture model parameters are significantly reduced compared to the MSE for the estimator of the mixture model intercept. It is interesting to note that these particularly well estimated mixture model parameters do not include any of the factor variable fixed effects or the covariance parameters.

The $comp * \pi\text{-}balance$ effect for $comp = 1$ and $\pi\text{-}balance = UNBAL$ was associated with a significant increase in the MSE of the estimators of all the parameters in the mixture model compared to those of component 3 when the mixing proportions are balanced. This is not surprising since we chose component 1 to be the component with the smallest proportion of 0.2. Perhaps for the same reason the $comp * N$ effect for $N = L$ and $comp = 1$ shows that the MSEs for all of the mixture model parameter estimators was significantly higher than when $N = H$ and $comp = 3$. Thus the effect of unit sample size on MSE was stronger when the mixing proportions were unbalanced.

In terms of the covariance parameters, the $param * N$ effects when $param = d_{11}$ and $N = L$, and when $param = d_{21}$ and $N = L$ both show that the estimators of these random effects covariance parameters were associated with significantly higher MSEs when the number of units were low compared to the intercept when $N = H$. This is quite logical since the number of units can be expected to influence the estimation of the random effects parameters more than the other parameters. The effect of D at the $D = HPOS$ and $D = HNEG$ levels show that the MSEs for the estimators of all the mixture model parameters were significantly higher when the random effects covariance matrices were large compared to when they were small and positive. Furthermore whether the random effects are positively or negatively correlated seemed to not make much of a difference. Similarly the effect σ^2 shows that the MSE of all the model parameters increased when $\sigma^2 = H$ compared to $\sigma^2 = L$.

In terms of sample sizes, the N effect shows that the MSEs for the estimators of all the mixture model parameters were significantly increased when $N = L$ compared to when $N = H$. The effect $N * \sigma^2$ when $N = L$ and $\sigma^2 = H$ shows that the estimators of all the model parameters were associated with higher MSEs when the number of units were low compared to when they were high. The effect $N * max\text{-}n_i$ when N and $max\text{-}n_i$ were both low, and the effect $max\text{-}n_i * \sigma^2$ when $max\text{-}n_i = L$ and $\sigma^2 = H$ both show that these less optimal settings of the simulation variables significantly increased the MSEs of the estimators of the mixture model parameters compared to the optimal settings.

Thus the main result for the MSEs is that the MSEs increased when either the unit sample size, and/or the within-unit sample sizes reduced, and that this effect was particularly strong for the random effects covariance parameters. Furthermore large variances and covariances of and between the random effects also increased the MSEs, as did high within-unit error covariances.

We will now discuss the results from the charts in the supplementary materials. Firstly, and for the CPLIs (coverage probability length indices), CPs (coverage probabilities), and the CILs (confidence interval lengths), there was no major interaction between the simulation variables. Thus we will generally concentrate only on the differences between the confidence interval methods. Starting with the CPLI we see that $CI1$ produced slightly higher CPLIs than the other methods, and that $CI3$ was very slightly higher than $CI2$ which in turn was very slightly higher than $CI4$. For the CPs we have that $CI1$ consistently produced slightly higher ranges of coverage probabilities which often intersected the nominal level compared to the other three methods. Indeed on average it appears as if the coverage probabilities for $CI1$ were about 95%, whilst those for the other three methods were about 90%. It also appears as though $CI2$ and $CI4$ were very similar, and that both produced very slightly higher coverage probabilities than $CI3$. This difference was quite small however. It is noteworthy that the confidence intervals produced by $CI3$ performed as well as those of $CI2$ and $CI4$, and that they were not too much worse than those produced by $CI1$. For the CILs $CI4$ and $CI1$ produced similar confidence intervals which were slightly longer than those produced by $CI2$ and $CI3$. The methods $CI2$ and $CI3$ produced confidence intervals of a similar length. Thus the very good coverage attained by $CI1$ was not achieved trivially in the sense of producing confidence intervals that were very long.

We now discuss the robust model M-estimates in table 5.5 for the CPLI as the response. We firstly describe effects that relate to all confidence interval methods, and all the mixture model parameters (i.e. effects that do not contain CI or *param*). The following effects $N * max\text{-}n_i$, $N * \sigma^2$, $N * n_i\text{-}unbalance$, and $max\text{-}n_i * \sigma^2$ all show that as expected the CPLI reduced when the two simulation variables involved were set to the non-optimal settings compared to the optimal ones. For example the $N * max\text{-}n_i$ effect shows that the CPLI reduced when $N = L$ and $max\text{-}n_i = L$ compared to when $N = H$ and $max\text{-}n_i = H$. The $N * D$ effect shows that CPLI reduced when $N = L$ and $D = HNEG$ compared to when $N = H$ and $D = LPOS$. A similar but weaker effect slightly lower down the table was observed when $N = L$ and $D = HPOS$. This suggests the variances and covariances of the random effects being high and negative respectively resulted in a larger and more significant reduction in the CPLI (compared to when the variances and covariances were low and positive respectively) than when the variances of the random effects were high and positively correlated.

The differences between the confidence interval methods is demonstrated by a few strong interaction effects involving the factor CI. It is important to note that we do not want to over interpret effects such as these involving CI since they are all relative to $CI1$ and the other reference categories. This is because we have no good reason to make $CI1$ the reference category. For example we have no reason to believe $CI1$ is the best method for generating confidence intervals, nor is it an established "gold standard" method. Thus we do not want to put too much emphasis on the effect size itself because if we were to change the reference category for CI we would get a different effect size. For this reason all we wish to conclude is that the effect on the CPLI of the mixing proportions being unbalanced was strongly dependent on the confidence interval method. Similarly the two $CI * N$ effects involving $CI2$ and $CI3$ show that the effect of N on the CPLI was also strongly dependent on the confidence interval method. Apart from these two interactions there are no other effects containing CI in the top twenty effects in the table. This suggests that there were not many differences between the four confidence interval methods in terms of the factors that very strongly influence the CPLI. In contrast there are many effects that contain CI lower down the table, suggesting that there were many differences between the four confidence interval methods in terms of factors that moderately influence the CPLI.

In terms of individual mixture model parameters, the three $param * N$ effects involving all of the random effects parameters show that the CPLI for these mixture model parameters were reduced the most compared to the CPLI for the mixture model intercept when $N = L$ compared to when $N = H$. This is to be expected since the number of units is really the effective sample size for the random effects covariance parameters. The $param * comp$ effects involving π_1, $\beta_2^{c_1}$, β_1^{tc}, ,β_2^{tc}, and d_{21} for component 2, show that these parameters had higher CPLIs than the mixture model intercept for component 3. The reason for the superior CPLIs for some of these parameters may well be the lower MSEs associated with their estimators that we alluded to earlier. However these lower MSEs were not associated with the estimators of d_{21}, and so it is likely that the CPLI is not completely determined by estimator quality. Again these effects are difficult to interpret, but what is noteworthy is that the effects do not concern either the factor variable fixed effects, nor many of the covariance parameters. Thus perhaps confidence interval quality is superior in the continuous fixed effects, and for the mixing proportions.

To conclude from the M-estimates, the simulation variables N, *max*-n_i, n_i-*unbalance*, σ^2, and D all influenced the CPLI in the expected way when they were set at their non-optimal settings. For the simulation variable D it also appears that the detrimental impact on confidence interval quality when D was high was larger when the random effects were negatively as opposed to positively correlated. In terms of differences between mixture model parameters it appears that the factor variable fixed effects, and the covariance parameters had worse quality confidence intervals than the continuous variable fixed effects, and the mixing proportions. Finally it was only the influence of π-*balance* and N on the CPLI that differed strongly between the confidence interval methods.

Table 5.3: Simulation variable settings for the 128 runs of Model 1.

simnumber	$\boldsymbol{N}$	***max-n_i***	***π-unbalance*** $\boldsymbol{: (\pi_1, \pi_2, \pi_3)}$	$\boldsymbol{D : (\mathrm{v}(D_1)), (\mathrm{v}(D_2)), (\mathrm{v}(D_3))}$	$\boldsymbol{\sigma^2 : (\sigma_1^2, \sigma_2^2, \sigma_3^2)}$	***n_i-unbalance*** $\boldsymbol{: E(X) = np}$ **of** $\boldsymbol{n}$
1	100	5	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	4.75 of 5
2	100	5	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	3.25 of 5
3	100	5	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	4.75 of 5
4	100	5	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	3.25 of 5
5	100	5	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	4.75 of 5
6	100	5	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	3.25 of 5
7	100	5	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	4.75 of 5
8	100	5	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	3.25 of 5
9	100	5	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	4.75 of 5
10	100	5	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	3.25 of 5
11	100	5	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	4.75 of 5
12	100	5	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	3.25 of 5
13	100	5	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	4.75 of 5
14	100	5	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	3.25 of 5
15	100	5	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	4.75 of 5
16	100	5	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	3.25 of 5
17	100	5	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	4.75 of 5
18	100	5	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	3.25 of 5
19	100	5	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	4.75 of 5
20	100	5	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	3.25 of 5
21	100	5	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	4.75 of 5
22	100	5	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	3.25 of 5
23	100	5	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	4.75 of 5
24	100	5	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	3.25 of 5
25	100	5	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	4.75 of 5
26	100	5	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	3.25 of 5
27	100	5	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	4.75 of 5
28	100	5	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	3.25 of 5
29	100	5	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	4.75 of 5
30	100	5	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	3.25 of 5
31	100	5	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	4.75 of 5
32	100	5	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	3.25 of 5
33	100	10	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	9.5 of 10
34	100	10	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	6.5 of 10
35	100	10	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	9.5 of 10
36	100	10	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	6.5 of 10
37	100	10	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	9.5 of 10
38	100	10	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	6.5 of 10
39	100	10	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	9.5 of 10
40	100	10	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	6.5 of 10

Table 5.3 continued

simnumber	***N***	***max-n_i***	***π-unbalance* : (π_1, π_2, π_3)**	**D : (v(D_1)), (v(D_2)), (v(D_3))**	**σ^2 : $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$**	***n_i-unbalance* : $E(X) = np$ of n**
41	100	10	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	9.5 of 10
42	100	10	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	6.5 of 10
43	100	10	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	9.5 of 10
44	100	10	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	6.5 of 10
45	100	10	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	9.5 of 10
46	100	10	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	6.5 of 10
47	100	10	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	9.5 of 10
48	100	10	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	6.5 of 10
49	100	10	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	9.5 of 10
50	100	10	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	6.5 of 10
51	100	10	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	9.5 of 10
52	100	10	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	6.5 of 10
53	100	10	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	9.5 of 10
54	100	10	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	6.5 of 10
55	100	10	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	9.5 of 10
56	100	10	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	6.5 of 10
57	100	10	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	9.5 of 10
58	100	10	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	6.5 of 10
59	100	10	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	9.5 of 10
60	100	10	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	6.5 of 10
61	100	10	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	9.5 of 10
62	100	10	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	6.5 of 10
63	100	10	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	9.5 of 10
64	100	10	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	6.5 of 10
65	1000	5	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	4.75 of 5
66	1000	5	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	3.25 of 5
67	1000	5	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	4.75 of 5
68	1000	5	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	3.25 of 5
69	1000	5	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	4.75 of 5
70	1000	5	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	3.25 of 5
71	1000	5	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	4.75 of 5
72	1000	5	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	3.25 of 5
73	1000	5	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	4.75 of 5
74	1000	5	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	3.25 of 5
75	1000	5	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	4.75 of 5
76	1000	5	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	3.25 of 5
77	1000	5	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	4.75 of 5
78	1000	5	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	3.25 of 5
79	1000	5	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	4.75 of 5
80	1000	5	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	3.25 of 5

Table 5.3 continued

simnumber	$\boldsymbol{N}$	$\boldsymbol{max\text{-}n_i}$	$\boldsymbol{\pi\text{-}unbalance : (\pi_1, \pi_2, \pi_3)}$	$\boldsymbol{D : (\mathrm{v}(D_1)), (\mathrm{v}(D_2)), (\mathrm{v}(D_3))}$	$\boldsymbol{\sigma^2 : (\sigma_1^2, \sigma_2^2, \sigma_3^2)}$	$\boldsymbol{n_i\text{-}unbalance : E(X) = np}$ **of** $\boldsymbol{n}$
81	1000	5	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	4.75 of 5
82	1000	5	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	3.25 of 5
83	1000	5	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	4.75 of 5
84	1000	5	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	3.25 of 5
85	1000	5	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	4.75 of 5
86	1000	5	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	3.25 of 5
87	1000	5	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	4.75 of 5
88	1000	5	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	3.25 of 5
89	1000	5	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	4.75 of 5
90	1000	5	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	3.25 of 5
91	1000	5	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	4.75 of 5
92	1000	5	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	3.25 of 5
93	1000	5	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	4.75 of 5
94	1000	5	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	3.25 of 5
95	1000	5	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	4.75 of 5
96	1000	5	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	3.25 of 5
97	1000	10	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	9.5 of 10
98	1000	10	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	6.5 of 10
99	1000	10	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	9.5 of 10
100	1000	10	(0.333,0.333,0.333)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	6.5 of 10
101	1000	10	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	9.5 of 10
102	1000	10	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	6.5 of 10
103	1000	10	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	9.5 of 10
104	1000	10	(0.333,0.333,0.333)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	6.5 of 10
105	1000	10	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	9.5 of 10
106	1000	10	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	6.5 of 10
107	1000	10	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	9.5 of 10
108	1000	10	(0.333,0.333,0.333)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	6.5 of 10
109	1000	10	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	9.5 of 10
110	1000	10	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	6.5 of 10
111	1000	10	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	9.5 of 10
112	1000	10	(0.333,0.333,0.333)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	6.5 of 10
113	1000	10	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	9.5 of 10
114	1000	10	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(1.9,1.8,1.75)	6.5 of 10
115	1000	10	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	9.5 of 10
116	1000	10	(0.2,0.4,0.4)	(1,-0.5,2.1),(2,-0.9,1.5),(1.5,-0.3,1.3)	(9.5,9,8.75)	6.5 of 10
117	1000	10	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	9.5 of 10
118	1000	10	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(1.9,1.8,1.75)	6.5 of 10
119	1000	10	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	9.5 of 10
120	1000	10	(0.2,0.4,0.4)	(1,0.5,2.1),(2,0.9,1.5),(1.5,0.3,1.3)	(9.5,9,8.75)	6.5 of 10

Table 5.3 continued

simnumber	N	*max*-n_i	π-*unbalance* : (π_1, π_2, π_3)	$D : (\mathrm{v}(D_1)), (\mathrm{v}(D_2)), (\mathrm{v}(D_3))$	$\sigma^2 : (\sigma_1^2, \sigma_2^2, \sigma_3^2)$	n_i-*unbalance* : $E(X) = np$ of n
121	1000	10	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	9.5 of 10
122	1000	10	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(1.9,1.8,1.75)	6.5 of 10
123	1000	10	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	9.5 of 10
124	1000	10	(0.2,0.4,0.4)	(5,-4.5,10.5),(10,-5.5,7.5),(7.5,-5,6.5)	(9.5,9,8.75)	6.5 of 10
125	1000	10	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	9.5 of 10
126	1000	10	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(1.9,1.8,1.75)	6.5 of 10
127	1000	10	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	9.5 of 10
128	1000	10	(0.2,0.4,0.4)	(5,4.5,10.5),(10,5.5,7.5),(7.5,5,6.5)	(9.5,9,8.75)	6.5 of 10

Figure 5.7: Boxplots of parameter estimates for β_2^{c2} with estimates outside the 10^{th} and 90^{th} percentiles displayed in a compression region - for an explanation see Subsection 5.1.4. Each of the four x-axis labels on each subplot denote whether the simulation variables *max-*n_i and n_i*-unbalanced* are high or low respectively, thus "H/L" denotes *max-*$n_i = H$ and n_i*-unbalanced* $= L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the $HNEG$ setting for the simulation variable D.

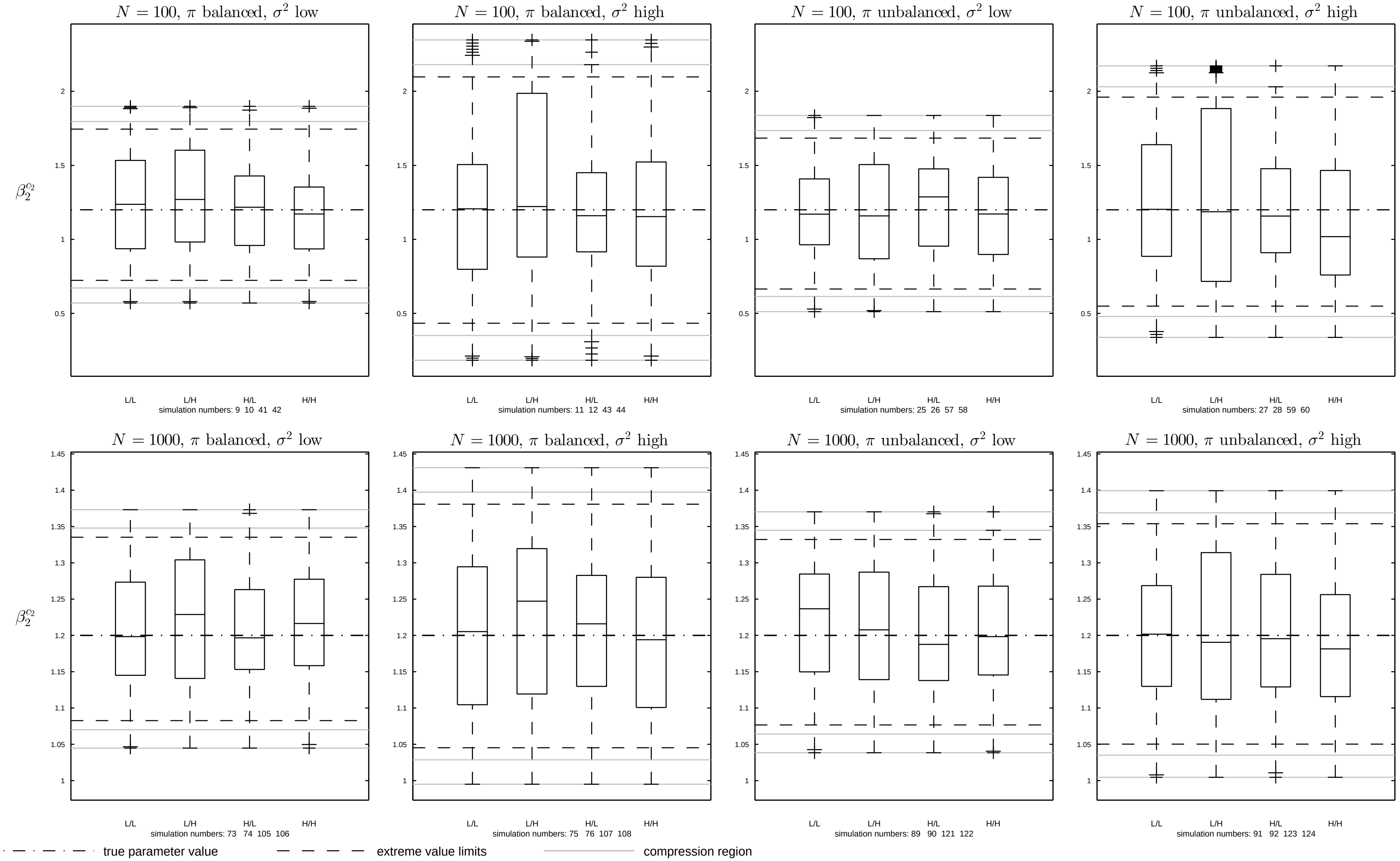


Figure 5.8: Coverage probabilities for β_2^{c2} with 95% approximate Binomial confidence intervals for each type of confidence interval construction method. Each of the four x-axis labels on each subplot denote whether the simulation variables $max\text{-}n_i$ and $n_i\text{-}unbalanced$ are high or low respectively, thus "H/L" denotes $max\text{-}n_i = H$ and $n_i\text{-}unbalanced = L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the $HNEG$ setting for the simulation variable D.

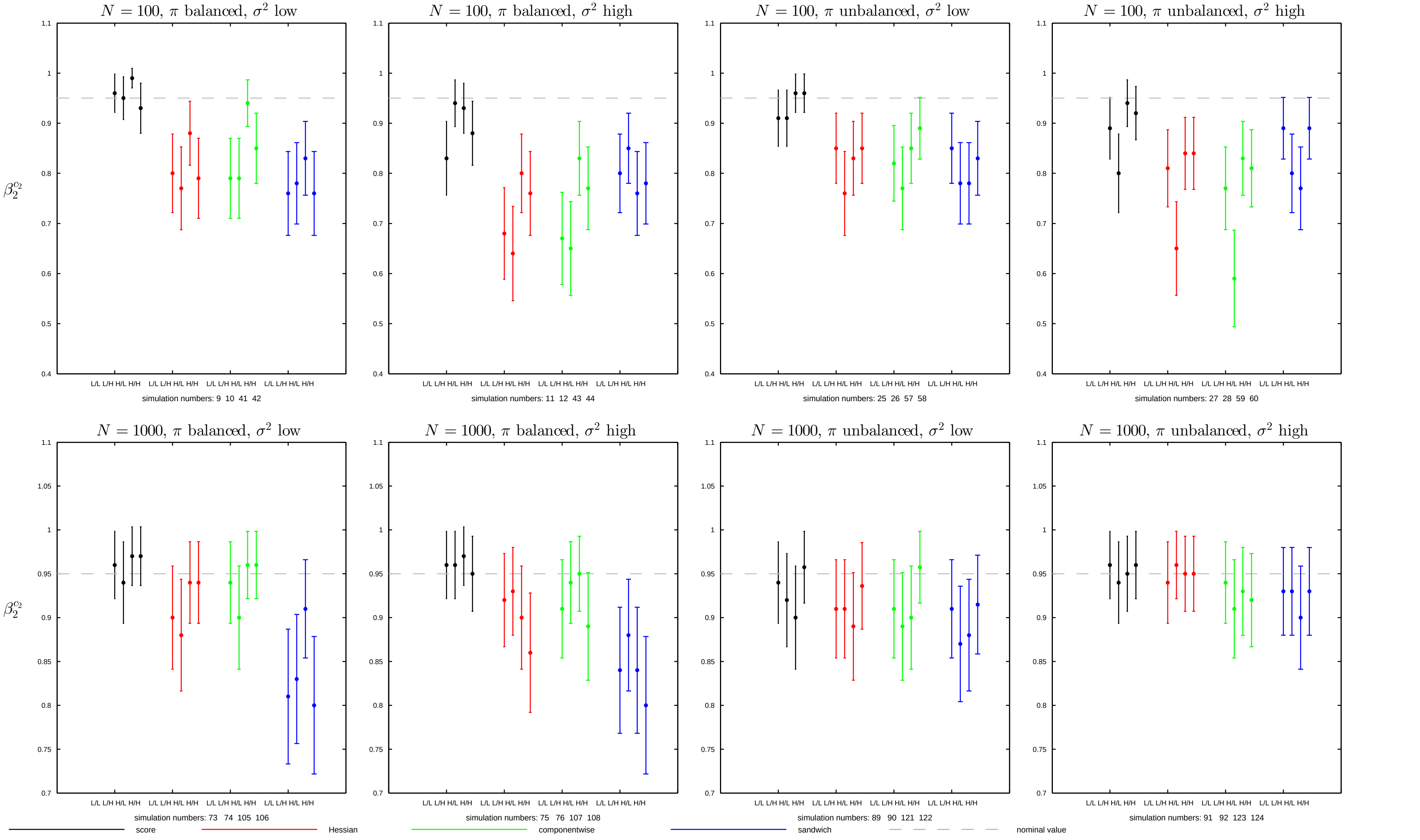


Figure 5.9: Boxplots of confidence interval lengths for β_2^{c2} for each type of confidence interval construction method. Confidence interval lengths outside the 10^{th} and 90^{th} percentiles are displayed in a compression region - for an explanation see Subsection 5.1.4. Each of the four x-axis labels on each subplot denote whether the simulation variables *max*-n_i and n_i-*unbalanced* are high or low respectively, thus ”H/L” denotes *max*-n_i $= H$ and n_i-*unbalanced* $= L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the $HNEG$ setting for the simulation variable D.

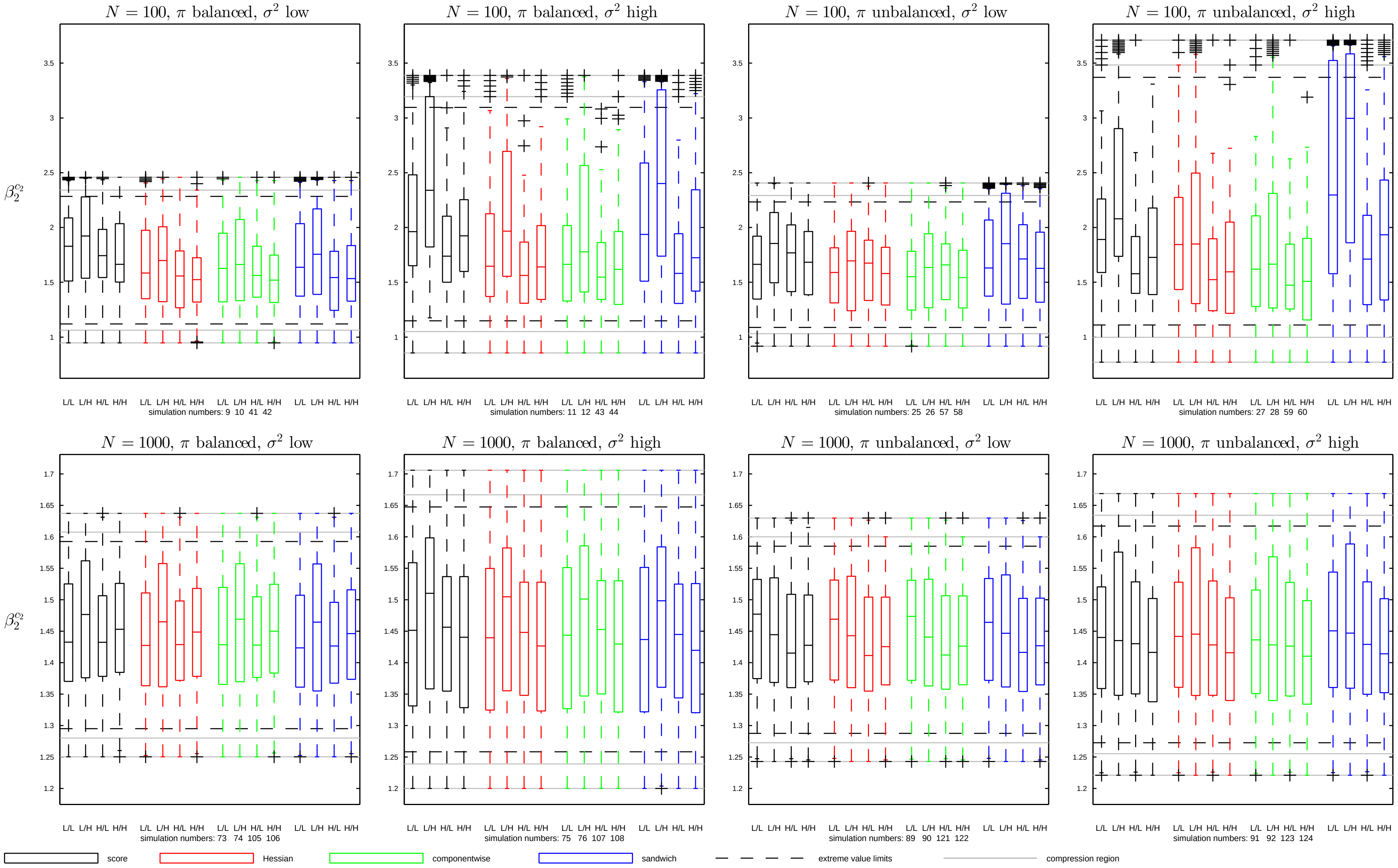


Figure 5.10: CPL indices for β_2^{c2} for each type of confidence interval construction method. Each of the four x-axis labels on each subplot denote whether the simulation variables *max-n_i* and *n_i-unbalanced* are high or low respectively, thus ”H/L” denotes *max-n_i* $= H$ and *n_i-unbalanced* $= L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the $HNEG$ setting for the simulation variable D.

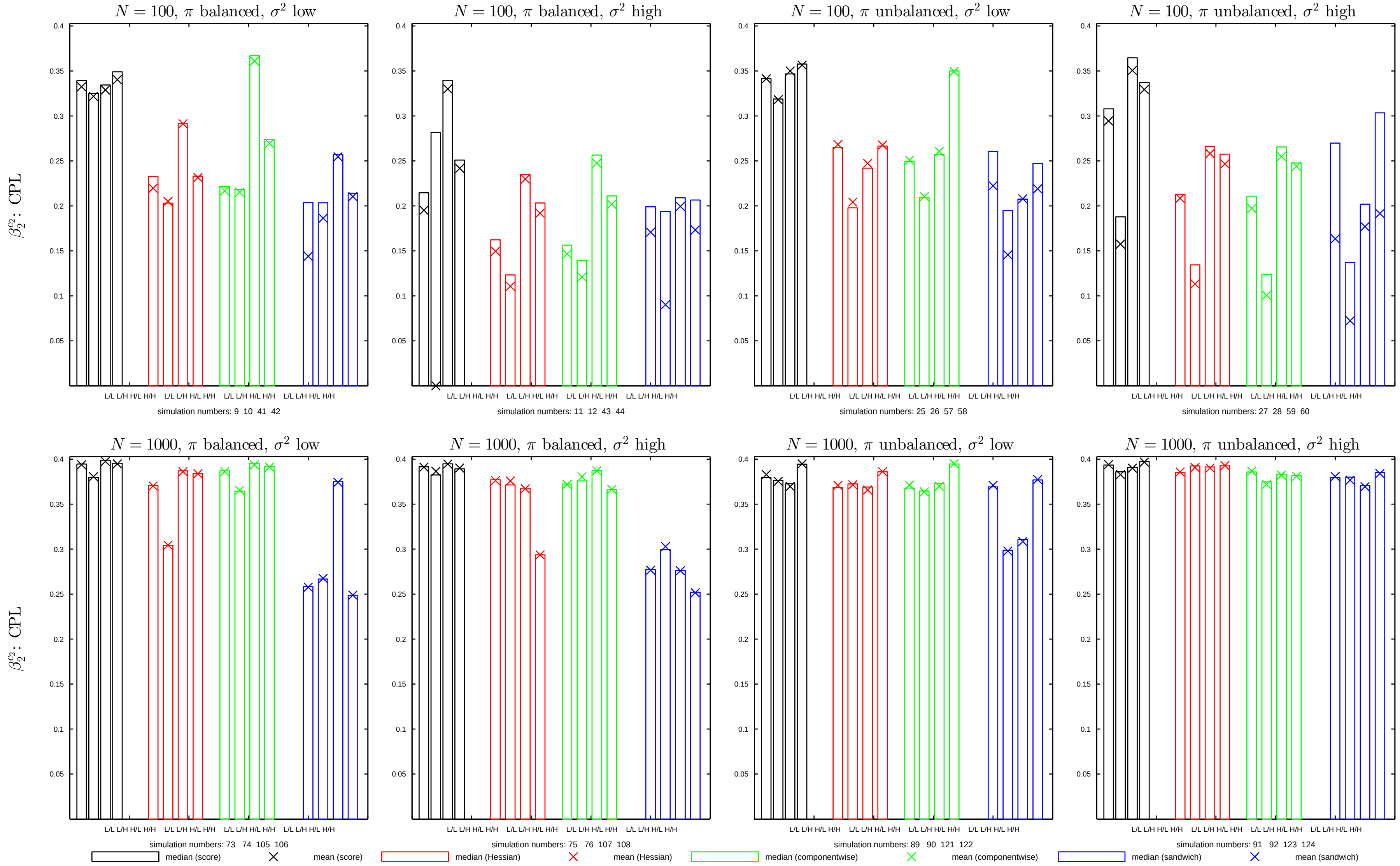


Table 5.4: Simulation parameter M-estimates with p-values less than 0.001 for MSE as the response.

Parameter	Level1	Level2	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$param$	π		-5.5994	0.1210	-5.8366	-5.3622	2140.54	0.00000
$param$	β^{c_1}		-4.3253	0.1210	-4.5625	-4.0881	1277.23	0.00000
$Intercept$			-3.2046	0.1077	-3.4157	-2.9935	885.62	0.00000
$param$	β^{c_2}		-2.9433	0.1210	-3.1805	-2.7061	591.43	0.00000
N	L		1.9345	0.0924	1.7533	2.1157	437.99	0.00000
D	$HPOS$		1.4968	0.0755	1.3489	1.6447	393.31	0.00000
D	$HNEG$		1.3712	0.0755	1.2233	1.5191	330.07	0.00000
$param$	β^{tc}		-2.1343	0.1210	-2.3715	-1.8971	310.99	0.00000
$comp * \pi\text{-}balance$	1	$UNBA$	0.7775	0.0494	0.6807	0.8744	247.65	0.00000
$comp * N$	1	L	0.7742	0.0494	0.6774	0.8710	245.53	0.00000
$N * \sigma^2$	L	H	0.5675	0.0403	0.4884	0.6466	197.88	0.00000
$N * max\text{-}n_i$	L	L	0.5398	0.0403	0.4607	0.6189	179.04	0.00000
$param * N$	d_{11}	L	1.1230	0.0988	0.9294	1.3167	129.16	0.00000
$param * N$	d_{21}	L	1.0905	0.0988	0.8968	1.2841	121.78	0.00000
$max\text{-}n_i * \sigma^2$	L	H	0.4355	0.0403	0.3564	0.5145	116.52	0.00000
$param$	d_{21}		-1.1631	0.1210	-1.4003	-0.9259	92.36	0.00000
$param * N$	σ^2	L	-0.9231	0.0988	-1.1167	-0.7294	87.25	0.00000
$comp * N$	2	L	0.4291	0.0494	0.3322	0.5259	75.42	0.00000
$param * comp$	d_{22}	1	0.9471	0.1210	0.7099	1.1843	61.24	0.00000
$param * comp$	d_{21}	1	0.9086	0.1210	0.6714	1.1458	56.36	0.00000
σ^2	H		0.4591	0.0638	0.3340	0.5841	51.79	0.00000
$param$	$\beta^{f_{11}}$		-0.8559	0.1210	-1.0931	-0.6187	50.01	0.00000
$param * comp$	d_{21}	2	0.8349	0.1210	0.5977	1.0721	47.59	0.00000
$D * \sigma^2$	$HNEG$	H	-0.3822	0.0571	-0.4941	-0.2704	44.89	0.00000
$comp * \sigma^2$	1	H	0.2995	0.0494	0.2027	0.3964	36.75	0.00000
$param * N$	π	L	-0.5679	0.0988	-0.7616	-0.3742	33.03	0.00000
$param * comp$	d_{11}	2	0.6842	0.1210	0.4470	0.9214	31.96	0.00000
$param * comp$	π	1	-0.6829	0.1210	-0.9201	-0.4457	31.84	0.00000
$N * n_i\text{-}unbalance$	L	H	0.2239	0.0403	0.1449	0.3030	30.81	0.00000
$comp$	1		-0.5978	0.1105	-0.8143	-0.3812	29.27	0.00000
$\sigma^2 * n_i\text{-}unbalance$	H	H	0.2091	0.0403	0.1300	0.2882	26.86	0.00000
$param * max\text{-}n_i$	d_{11}	L	0.5031	0.0988	0.3095	0.6968	25.92	0.00000
$param * N$	β^{c1}	L	-0.4533	0.0988	-0.6470	-0.2596	21.04	0.00000
$param$	$\beta^{f_{21}}$		-0.5540	0.1210	-0.7912	-0.3168	20.96	0.00000

Table 5.4 continued.

Parameter	Level1	Level2	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$D * \sigma^2$	$HPOS$	H	-0.2599	0.0571	-0.3717	-0.1481	20.75	0.00001
$param$	σ^2		-0.5494	0.1210	-0.7866	-0.3122	20.60	0.00001
$param * max\text{-}n_i$	β^{c1}	L	0.4232	0.0988	0.2295	0.6168	18.34	0.00002
$param * max\text{-}n_i$	π	L	-0.4126	0.0988	-0.6063	-0.2190	17.44	0.00003
$param * comp$	β^{tc}	1	0.4926	0.1210	0.2554	0.7298	16.57	0.00005
$param * comp$	d_{22}	2	0.4653	0.1210	0.2281	0.7025	14.78	0.00012
$comp * \sigma^2$	2	H	0.1882	0.0494	0.0913	0.2850	14.51	0.00014
$param$	β^{f22}		-0.4598	0.1210	-0.6970	-0.2225	14.43	0.00015
$comp * max\text{-}n_i$	1	L	0.1868	0.0494	0.0899	0.2836	14.29	0.00016
$param$	d_{22}		-0.4415	0.1210	-0.6788	-0.2043	13.31	0.00026
$\pi\text{-}balance * \sigma^2$	$UNBA$	H	-0.1388	0.0403	-0.2179	-0.0598	11.84	0.00058
$param * max\text{-}n_i$	d_{21}	L	0.3255	0.0988	0.1318	0.5192	10.85	0.00099

Table 5.5: Simulation parameter M-estimates with p-values less than 0.001 for median-based CPL as the response.

Parameter	Level1	Level2	Level3	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$Intercept$				-0.3144	0.0004	-0.3153	-0.3135	490090.0130	0.00000
$param * comp$	β^{f21}	1		-0.0083	0.0003	-0.0089	-0.0077	771.1530	0.00000
$N * max\text{-}n_i$	L	L		-0.0028	0.0001	-0.0030	-0.0026	763.0287	0.00000
$param * comp$	β^{tc}	2		0.0082	0.0003	0.0076	0.0088	757.7667	0.00000
$N * \sigma^2$	L	H		-0.0023	0.0001	-0.0025	-0.0021	527.0298	0.00000
$CI * comp * \pi\text{-}balance$	$CI4$	1	UNBA	-0.0078	0.0003	-0.0085	-0.0072	518.5061	0.00000
$param * comp$	β^{tc}	1		0.0056	0.0003	0.0050	0.0061	347.7905	0.00000
$param * N$	d_{21}	L		-0.0085	0.0005	-0.0095	-0.0076	307.7763	0.00000
$param * N$	d_{11}	L		-0.0084	0.0005	-0.0093	-0.0074	297.3850	0.00000
$CI * comp * \pi\text{-}balance$	$CI2$	1	UNBA	-0.0053	0.0003	-0.0060	-0.0046	235.5190	0.00000
$N * n_i\text{-}unbalance$	L	H		-0.0015	0.0001	-0.0017	-0.0013	215.7445	0.00000
$N * D$	L	$HNEG$		-0.0020	0.0001	-0.0023	-0.0018	206.8946	0.00000
$comp * N$	1	L		-0.0032	0.0002	-0.0036	-0.0027	170.0627	0.00000
$param * comp$	β^{c1}	2		0.0037	0.0003	0.0031	0.0043	154.1157	0.00000
CI*N	$CI3$	L		-0.0065	0.0005	-0.0075	-0.0054	151.0138	0.00000
$param * comp$	d_{21}	2		0.0036	0.0003	0.0030	0.0042	147.8138	0.00000
$CI * comp * N$	$CI4$	1	L	0.0039	0.0003	0.0032	0.0046	127.3020	0.00000
$param$	β^{tc}			-0.0057	0.0005	-0.0068	-0.0047	124.0435	0.00000
CI*N	$CI2$	L		-0.0058	0.0005	-0.0069	-0.0048	123.3287	0.00000
$param * N$	d_{22}	L		-0.0050	0.0005	-0.0059	-0.0040	104.3460	0.00000
$param * comp$	π	1		0.0032	0.0003	0.0026	0.0039	100.4884	0.00000
$max\text{-}n_i * \sigma^2$	L	H		-0.0010	0.0001	-0.0012	-0.0008	94.8897	0.00000
$N * D$	L	$HPOS$		-0.0013	0.0001	-0.0016	-0.0010	86.4703	0.00000
$param * N$	β^{f22}	L		-0.0045	0.0005	-0.0054	-0.0035	84.6471	0.00000
$param * comp$	β^{f22}	2		-0.0027	0.0003	-0.0033	-0.0021	83.4511	0.00000
$param * comp$	d_{21}	1		0.0027	0.0003	0.0021	0.0032	79.2233	0.00000
$CI * param * N$	$CI3$	d_{21}	L	0.0060	0.0007	0.0046	0.0073	76.0287	0.00000
$param$	d_{21}			-0.0043	0.0005	-0.0054	-0.0033	70.9371	0.00000
$CI * param * N$	$CI2$	d_{21}	L	0.0053	0.0007	0.0040	0.0066	59.3997	0.00000
$CI * param * N$	$CI3$	σ^2	L	0.0050	0.0007	0.0036	0.0063	52.6641	0.00000
N	L			0.0028	0.0004	0.0021	0.0036	51.7932	0.00000
$comp * \pi\text{-}balance$	1	$UNBA$		-0.0017	0.0002	-0.0022	-0.0012	49.7426	0.00000
$CI * param * N$	$CI3$	β^{c1}	L	0.0047	0.0007	0.0034	0.0061	47.5419	0.00000
$param * comp$	β^{c1}	1		0.0021	0.0003	0.0015	0.0026	47.4881	0.00000

Table 5.5 continued.

Parameter	Level1	Level2	Level3	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$CI * param * N$	$CI2$	π	L	0.0047	0.0007	0.0034	0.0061	47.2080	0.00000
$N * \pi\text{-}balance$	L	$UNBA$		0.0007	0.0001	0.0005	0.0009	46.0562	0.00000
$CI * param * N$	$CI2$	d_{22}	L	0.0046	0.0007	0.0032	0.0059	44.1595	0.00000
$CI * param * N$	$CI2$	σ^2	L	0.0045	0.0007	0.0031	0.0058	42.2096	0.00000
$CI * param * N$	$CI3$	β^{tc}	L	0.0044	0.0007	0.0030	0.0057	40.6500	0.00000
$param * N$	β^{tc}	L		-0.0030	0.0005	-0.0040	-0.0021	39.2002	0.00000
$CI * param * N$	$CI2$	β^{c1}	L	0.0043	0.0007	0.0029	0.0056	38.9142	0.00000
$param * comp$	β^{f11}	1		-0.0019	0.0003	-0.0024	-0.0013	38.8380	0.00000
CI*N	$CI4$	L		-0.0033	0.0005	-0.0043	-0.0022	38.8183	0.00000
$param * N$	β^{f21}	L		-0.0030	0.0005	-0.0040	-0.0021	38.6600	0.00000
D	$HNEG$			0.0017	0.0003	0.0012	0.0023	38.0340	0.00000
$CI * param * N$	$CI2$	β^{tc}	L	0.0042	0.0007	0.0029	0.0056	37.8163	0.00000
$CI * param$	$CI4$	d_{11}		-0.0040	0.0007	-0.0054	-0.0027	34.4033	0.00000
$CI * param * N$	$CI4$	d_{11}	L	0.0040	0.0007	0.0027	0.0054	34.2506	0.00000
$CI * comp * N$	$CI2$	1	L	0.0020	0.0003	0.0013	0.0027	33.6613	0.00000
$param * comp$	d_{11}	1		0.0017	0.0003	0.0011	0.0023	33.6433	0.00000
$param * comp$	d_{22}	1		-0.0017	0.0003	-0.0023	-0.0011	33.2687	0.00000
$CI * comp$	$CI4$	1		-0.0029	0.0005	-0.0039	-0.0019	31.6745	0.00000
$CI * param * N$	$CI3$	d_{22}	L	0.0038	0.0007	0.0025	0.0052	31.1341	0.00000
σ^2	H			0.0012	0.0002	0.0008	0.0016	30.9030	0.00000
$CI * comp * \pi\text{-}balance$	$CI4$	2	UNBA	0.0019	0.0003	0.0012	0.0026	30.8096	0.00000
$CI * param * N$	$CI4$	d_{21}	L	0.0037	0.0007	0.0024	0.0050	28.9500	0.00000
$param * comp$	σ^2	2		0.0016	0.0003	0.0010	0.0022	28.8070	0.00000
$max\text{-}n_i * \pi\text{-}balance$	L	$UNBA$		0.0005	0.0001	0.0003	0.0007	27.7022	0.00000
$CI * comp$	$CI4$	2		-0.0027	0.0005	-0.0037	-0.0017	27.6370	0.00000
$CI * comp * D$	$CI4$	2	HNEG	-0.0025	0.0005	-0.0035	-0.0016	27.2247	0.00000
$CI * \sigma^2$	$CI3$	H		-0.0013	0.0002	-0.0018	-0.0008	26.3076	0.00000
$max\text{-}n_i * D$	L	$HNEG$		-0.0007	0.0001	-0.0010	-0.0004	25.3016	0.00000
$D * \sigma^2$	$HPOS$	H		-0.0007	0.0001	-0.0010	-0.0004	24.7976	0.00000
$CI * param * N$	$CI2$	d_{11}	L	0.0034	0.0007	0.0021	0.0048	24.6474	0.00000
$\sigma^2 * n_i\text{-}unbalance$	H	H		-0.0005	0.0001	-0.0007	-0.0003	24.5895	0.00000
$max\text{-}n_i * n_i\text{-}unbalance$	L	H		-0.0005	0.0001	-0.0007	-0.0003	24.3961	0.00000
$param * comp$	π	2		0.0016	0.0003	0.0009	0.0022	23.6101	0.00000
$comp * N$	2	L		-0.0012	0.0002	-0.0016	-0.0007	22.3683	0.00000

Table 5.5 continued.

Parameter	Level1	Level2	Level3	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$CI * param$	$CI4$	π		0.0032	0.0007	0.0018	0.0045	21.3535	0.00000
$comp * D$	2	$HNEG$		-0.0016	0.0003	-0.0023	-0.0009	21.1128	0.00000
$CI * param * N$	$CI3$	d_{11}	L	0.0032	0.0007	0.0018	0.0045	21.0592	0.00000
CI*D	$CI3$	$HNEG$		-0.0015	0.0004	-0.0022	-0.0008	17.9521	0.00002
$CI * param * N$	$CI3$	β^{f21}	L	0.0029	0.0007	0.0016	0.0043	17.8436	0.00002
$D * \sigma^2$	$HNEG$	H		-0.0006	0.0001	-0.0009	-0.0003	17.8405	0.00002
$CI * param$	$CI3$	d_{11}		-0.0029	0.0007	-0.0042	-0.0015	17.3275	0.00003
$param * N$	β^{f11}	L		-0.0020	0.0005	-0.0030	-0.0011	17.3092	0.00003
$param$	β^{c1}			-0.0021	0.0005	-0.0031	-0.0011	15.9776	0.00006
$CI * param$	$CI2$	d_{22}		-0.0027	0.0007	-0.0041	-0.0014	15.9434	0.00007
D	$HPOS$			0.0011	0.0003	0.0006	0.0017	15.6855	0.00007
$CI * param * N$	$CI2$	β^{f21}	L	0.0027	0.0007	0.0014	0.0041	15.4697	0.00008
$param$	π			-0.0020	0.0005	-0.0031	-0.0010	15.1831	0.00010
$CI * param$	$CI4$	d_{22}		-0.0026	0.0007	-0.0040	-0.0013	14.7316	0.00012
$CI * param * N$	$CI2$	β^{f22}	L	0.0026	0.0007	0.0013	0.0040	14.5140	0.00014
$CI * comp * D$	$CI4$	2	HPOS	-0.0018	0.0005	-0.0028	-0.0009	14.0085	0.00018
$param * comp$	d_{22}	2		0.0011	0.0003	0.0005	0.0016	12.7199	0.00036
$CI * comp$	$CI2$	1		-0.0018	0.0005	-0.0028	-0.0008	12.2022	0.00048
$CI * comp * \sigma^2$	$CI4$	2	H	0.0012	0.0003	0.0005	0.0019	11.7800	0.00060
$CI * param * N$	$CI3$	β^{f22}	L	0.0023	0.0007	0.0010	0.0037	11.3742	0.00074
$max\text{-}n_i * D$	L	$HPOS$		-0.0005	0.0001	-0.0007	-0.0002	10.9798	0.00092

5.3.2 Model 2

Looking at the estimate plots in the supplementary materials, and just as for Model 1, it is clear that for most mixture model parameters estimation was better (estimates were less biased and had less variability) for $\sigma^2 = L$ compared to $\sigma^2 = H$, and that this relationship was stronger for $N = H$ compared to $N = L$. However in contrast to Model 1 this interaction between N and σ^2 was modified by whether or not the mixing proportions were balanced or not. In particular it seems the increased levels of bias and variation observed in the mixture model parameter estimators as σ^2 changes from low to high was more marked when the mixing proportions were unbalanced. This was observed for both $N = L$ and $N = H$. These observations suggest a three way interaction between N, σ^2, and the balancedness of the mixing proportions. Furthermore this three way interaction seemed slightly weaker for the estimators of the within-unit variances and the autoregressive parameters. In general the estimation quality for the mixture model parameters was good, particularly when $N = H$, $\sigma^2 = L$, and the mixing proportions were balanced.

We now look at the robust model M-estimates shown in table 5.7 when MSE is the response variable. As expected we see that the strongest effect is $comp * \pi$-*balance*, which shows that the MSE increased when the mixing proportions were unbalanced (again for *comp* $= 1$ since component 1 was arbitrarily selected to have the smallest mixing proportion) compared to when the mixing proportions were balanced. The obvious reason why this effect is so strong compared to Model 1 is that the unbalancedness of $\boldsymbol{\pi} = (0.1, 0.9)$ is greater than $\boldsymbol{\pi} = (0.2, 0.4, 0.4)$ of Model 1. Another striking difference compared to Model 1 is the number of $param$ effects at the top of the table. This shows there was much greater variation in the MSEs of the parameter estimators for this model. In particular we see compared to the mixture model intercept that the fixed effect parameters of the continuous and the factor variables, the autoregressive parameters, and the mixing proportions were all estimated with lower MSEs, whilst the random effects covariance parameters were all estimated with higher MSEs.

Just as for Model 1 the effects N, $N * \sigma^2$, and σ^2 show that the MSE of the mixture model parameters are increased when the levels of the variables involved are set at the non-optimal levels compared to the optimal ones. Similarly the effect $D = HNEG$ shows that the MSEs increase when the random effects have both high variances and are negatively correlated. In contrast to Model 1 we see that the $D = HPOS$ effect is not a strong one. In terms of the ACF simulation variable, the $param * ACF$ effect when $param = \sigma^2$ and $ACF = H$ shows that the MSEs for the within-unit variances was increased when the ACF was high. When looking at the individual runs it is clear that this effect was most severe when the other simulation variables were all not set to their default values. In contrast when these variables were set at their defaults the estimation of the within-unit variances was very good regardless of whether the ACF variable was high or low. However these three way interactions between ACF, $param = \sigma^2$ and the other simulation variables were not the very strong ones for this model and so they do not appear in Table 5.7.

In summary the MSEs were influenced in a similar way as in Model 1 by the unit sample sizes, and by the covariance parameters. One exception is that only negatively correlated random effects with high variances seemed to adversely influence parameter estimation. Because of the more extreme unbalancedness of the mixing proportions compared to Model 1, the $comp * \pi$-*balance* was the strongest effect. High levels of serial dependence in the within-unit responses adversely affected estimation of the within-unit variances, although this effect was most severe when the other simulation variables were not set at their default levels.

We now look at the CPLIs of the mixture model parameter estimates, focusing first on the charts in the supplementary materials. The same strong three way interaction between N, σ^2, and the balancedness of the mixing proportions is observable, where the CPLIs reduced when these simulation variables were set at their non-optimal levels. Similar to Model 1 there was not a great difference between the three confidence interval methods, however $CI1$ consistently produced slightly higher CPLIs than the other three methods. Furthermore $CI3$ produced CPLIs which were as high as those produced by $CI2$ and $CI4$, and only slightly lower than those produced by $CI1$. Thus $CI1$ appears to produce superior confidence intervals on the estimators of the mixture model parameters compared to the other three methods which are very similar.

The plots of the CPs reveal a strong interaction between σ^2, the balancedness of the mixing proportions, and the confidence interval method. They show that $CI1$ was the superior method, regardless of whether σ^2 was high or low, and regardless of whether $\boldsymbol{\pi}$ was balanced or not. These results also suggest that $CI4$, which is based on the robust estimator of the mixture model information matrix, performs better than $CI2$ and $CI3$ when the within-unit variation is high. This might be because the properties of the robust estimator that make it robust to model misspecification may be beneficial in other circumstances where the model appears to have been misspecified - i.e. low N and/or high levels of noise in the data. It is difficult to explain however why $CI4$ performed worse than $CI3$ when $\sigma^2 = L$ when the mixing proportions were unbalanced.

In contrast to Model 1, all of the confidence interval methods produced different levels of coverage: $CI1$ produced the highest, $CI4$ the next highest, and then $CI3$ and $CI2$. In contrast to Model 1, on average the coverage for $CI1$ is approximately 5% off the nominal level, however this is probably on account of the unbalancedness of the mixing proportions being more extreme. The coverage of $CI4$ is reasonable, however the coverage of $CI3$ and $CI2$ are slightly low. Furthermore the fact that $CI3$ produces similar coverage probabilities to $CI2$ is notable.

Looking now at the plots for the CILs we see again the strong interaction between N, σ^2, and the balancedness of the mixing proportions. For balanced mixing proportions, in general we see that the CILs were longer when $\sigma^2 = H$ compared to when $\sigma^2 = L$, and that this difference was greater when $N = L$. Furthermore when $\sigma^2 = H$ it appears $CI4$ produced the longest confidence intervals, and so this may explain the better coverage probabilities observed for $CI4$ we alluded to previously. This relationship between σ^2 and N was less clear when the mixing proportions were unbalanced. Instead we observed for component 1 that the $CI1$ CILs were the longest, but for component 2 the CILs for $CI4$ were the longest. This result is in contrast to Model 1 where both $CI1$ and $CI4$ produced the longest lengths. On average $CI1$ produced longer CILs than the other three methods, $CI2$ produced longer CILs than $CI3$, and the CILs produced by

$CI4$ were no different from $CI2$ and $CI3$.

In summary from the plots we conclude that $CI1$ was the best method overall. The fact that on average that $CI1$ produced both the highest CPLIs and CILs suggests the superior coverage produced by $CI1$ was not of the trivial kind - i.e. due to length alone. A similar result was obtained for Model 1. However caution must be exercised when drawing conclusions on averages because a strong three way interaction between N, σ^2, and the balancedness of the mixing proportions was present. In this respect good coverage probabilities (approximately $> 80\%$) were obtained by all methods when the mixing proportions were balanced, and generally by only $CI1$ (approximately 90%) when they were unbalanced. These coverage probabilities were slightly lower than for Model 1. Finally, and again similar to Model 1, the $CI3$ method performed well compared to $CI2$ and $CI3$, and even to $CI1$ when the mixing proportions were balanced.

We now look at the M-estimates in Table 5.8. Considering the above discussion from the results of the plots, it is not surprising to see that one of the strongest effects by far is $comp * \pi\text{-}balance$ which shows the CPLIs reduce when $comp = 1$ and the mixing proportions were unbalanced. Even considering this very strong interaction, the negative main effect for π-*balance* is still reasonably strong. These results confirm that the unbalancedness of the mixing proportions in this two component model had a greater influence than the unbalancedness did in the three component Model 1.

The next very strong effect is $CI * \sigma^2$ which shows that the CPLIs for all parameters reduced when $CI = 3$ and $\sigma^2 = H$ compared to when $CI = 1$ and $\sigma^2 = L$. Similarly the next strongest effect involves $CI2$, but this is much weaker than for $CI3$. Again, and as we described in subsection 5.3.1, it is important to note that we do not want to over interpret effects such as these involving CI since they are all relative to $CI1$, and in this case to $\sigma^2 = L$. Thus all we wish to conclude is that strong differences exist between the confidence interval methods with respect to the simulation variable σ^2.

The strong $N * \pi$-*balance* effect is unfortunately counter-intuitive to what we expect in that it shows the CPLIs increase when $N = L$, and when the mixing proportions were unbalanced. However this might be an example of where the expected effects of the constituent variables have been captured by other effects. In this respect the large negative effect of the mixing proportions being unbalanced has certainly been captured by the $comp * \pi$-*balance* effect, and the negative effect of $N = L$ has been captured in component 1 at least by the $comp * N$ effect. Similarly the strong positive π-*balance* $* \sigma^2$ effect is also counter-intuitive. However again the $comp * \pi$-*balance* effect has captured a lot of the negative effect of the mixing proportions being unbalanced, and the σ^2 effect has captured a lot of the negative effect when $\sigma^2 = H$ compared to when $\sigma^2 = L$. In terms of the ACF simulation variable, and in contrast to the MSEs, there is a weak $param * ACF$ effect (low down table 5.8) for $param = \sigma^2$ and $ACF = H$ which shows that the CPLIs for the within-unit variances were slightly reduced when the ACF was high. Thus poorer estimation of the within-unit variances when the ACF variable was high was not associated with a strong reduction in the performance of the confidence intervals for these estimates.

Finally, and in contrast to Model 1, there are a reasonable number of interaction effects at the top of the table involving CI and the simulation variables $param$, $comp$,

σ^2, and π-*balance*. Again because we do not wish to over-interpret these effects we will merely conclude that the effect of these simulation variables is strongly influenced by the confidence interval method. However similar to Model 1 there are also some $param * comp$, and $param * N$ effects at the top of the table which show there were strong differences between the CPLIs of the mixture model parameters, and that these differences varied by component and the level of N.

To summarise, from the M-estimates we see that many more differences between the confidence interval methods exist compared to Model 1, and similarly the effect of the mixing proportions being unbalanced was much stronger compared to Model 1. Another difference compared to Model 1 is that the random effects covariance matrix $\boldsymbol{D}$ does not feature amongst the strongest effects on the CPLIs. In terms of similarities to Model 1 the effects of N and σ^2 on the CPLIs were strong.

Table 5.6: Simulation variable settings for the 128 runs of Model 2.

$\boldsymbol{simnumber}$	$\boldsymbol{N}$	$\boldsymbol{max\text{-}n_i}$	$\boldsymbol{\pi\text{-}unbalance : (\pi_1, \pi_2)}$	$\boldsymbol{D : (\mathrm{v}(D_1)), (\mathrm{v}(D_2))}$	$\boldsymbol{\sigma^2 : (\sigma_1^2, \sigma_2^2)}$	$\boldsymbol{ACF : (\phi_1), (\phi_2)}$
1	100	10	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
2	100	10	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
3	100	10	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
4	100	10	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
5	100	10	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
6	100	10	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
7	100	10	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
8	100	10	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
9	100	10	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
10	100	10	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
11	100	10	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
12	100	10	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
13	100	10	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
14	100	10	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
15	100	10	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
16	100	10	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
17	100	10	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
18	100	10	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
19	100	10	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
20	100	10	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
21	100	10	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
22	100	10	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
23	100	10	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
24	100	10	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
25	100	10	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
26	100	10	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
27	100	10	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
28	100	10	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
29	100	10	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
30	100	10	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
31	100	10	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
32	100	10	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
33	100	15	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
34	100	15	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
35	100	15	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
36	100	15	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
37	100	15	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
38	100	15	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
39	100	15	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
40	100	15	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)

Table 5.6 continued

$\boldsymbol{simnumber}$	$\boldsymbol{N}$	$\boldsymbol{max\text{-}n_i}$	$\boldsymbol{\pi\text{-}unbalance : (\pi_1, \pi_2)}$	$\boldsymbol{D : (\mathrm{v}(D_1)), (\mathrm{v}(D_2))}$	$\boldsymbol{\sigma^2 : (\sigma_1^2, \sigma_2^2)}$	$\boldsymbol{ACF : (\phi_1), (\phi_2)}$
41	100	15	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
42	100	15	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
43	100	15	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
44	100	15	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
45	100	15	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
46	100	15	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
47	100	15	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
48	100	15	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
49	100	15	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
50	100	15	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
51	100	15	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
52	100	15	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
53	100	15	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
54	100	15	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
55	100	15	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
56	100	15	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
57	100	15	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
58	100	15	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
59	100	15	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
60	100	15	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
61	100	15	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
62	100	15	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
63	100	15	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
64	100	15	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
65	500	10	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
66	500	10	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
67	500	10	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
68	500	10	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
69	500	10	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
70	500	10	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
71	500	10	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
72	500	10	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
73	500	10	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
74	500	10	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
75	500	10	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
76	500	10	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
77	500	10	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
78	500	10	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
79	500	10	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
80	500	10	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)

Table 5.6 continued

$\boldsymbol{simnumber}$	$\boldsymbol{N}$	$\boldsymbol{max\text{-}n_i}$	$\boldsymbol{\pi\text{-}unbalance : (\pi_1, \pi_2)}$	$\boldsymbol{D : (\mathrm{v}(D_1)), (\mathrm{v}(D_2))}$	$\boldsymbol{\sigma^2 : (\sigma_1^2, \sigma_2^2)}$	$\boldsymbol{ACF : (\phi_1), (\phi_2)}$
81	500	10	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
82	500	10	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
83	500	10	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
84	500	10	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
85	500	10	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
86	500	10	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
87	500	10	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
88	500	10	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
89	500	10	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
90	500	10	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
91	500	10	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
92	500	10	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
93	500	10	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
94	500	10	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
95	500	10	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
96	500	10	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
97	500	15	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
98	500	15	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
99	500	15	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
100	500	15	(0.5,0.5)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
101	500	15	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
102	500	15	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
103	500	15	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
104	500	15	(0.5,0.5)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
105	500	15	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
106	500	15	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
107	500	15	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
108	500	15	(0.5,0.5)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
109	500	15	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
110	500	15	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
111	500	15	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
112	500	15	(0.5,0.5)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
113	500	15	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
114	500	15	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
115	500	15	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
116	500	15	(0.1,0.9)	(1,-0.5,2.1),(2,-0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
117	500	15	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
118	500	15	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
119	500	15	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
120	500	15	(0.1,0.9)	(1,0.5,2.1),(2,0.9,1.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)

Table 5.6 continued

simnumber	$\boldsymbol{N}$	$\boldsymbol{max\text{-}n_i}$	$\boldsymbol{\pi\text{-}unbalance : (\pi_1, \pi_2)}$	$\boldsymbol{D : (\mathrm{v}(D_1)), (\mathrm{v}(D_2))}$	$\boldsymbol{\sigma^2 : (\sigma_1^2, \sigma_2^2)}$	$\boldsymbol{ACF : (\phi_1), (\phi_2)}$
121	500	15	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
122	500	15	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
123	500	15	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
124	500	15	(0.1,0.9)	(5,-4.5,10.5),(10,-5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)
125	500	15	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,0.1),(0.38,0.12)
126	500	15	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(1.3,1.2)	(0.4,-0.1),(0.38,-0.12)
127	500	15	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,0.1),(0.38,0.12)
128	500	15	(0.1,0.9)	(5,4.5,10.5),(10,5.5,7.5)	(8.3,8.2)	(0.4,-0.1),(0.38,-0.12)

Figure 5.11: Boxplots of parameter estimates for ϕ_1^2 with estimates outside the 10^{th} and 90^{th} percentiles displayed in a compression region - for an explanation see Subsection 5.1.4. Each of the four x-axis labels on each subplot denote whether the simulation variables $max\text{-}n_i$ and ACF are high or low respectively, thus "H/L" denotes $max\text{-}n_i = H$ and $ACF = L$ respectively. Note that as the simulation variable ACF changes then so too do the true values for ϕ_1^2, and so there are two broken lines on the plots indicating these true values. In total all of the plots show results for 32 combinations of the simulation variables, all at the $LNEG$ setting for the simulation variable D.

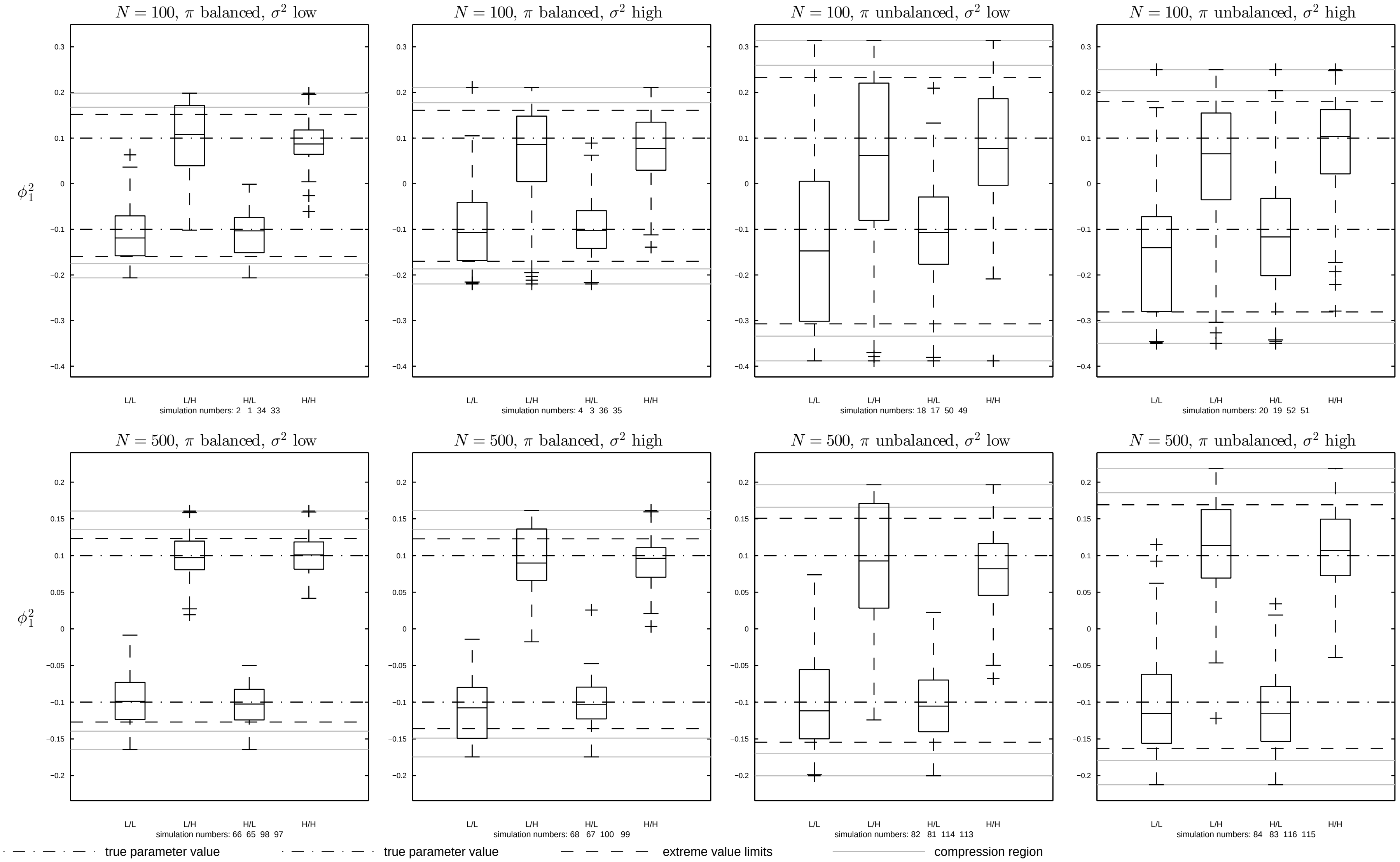


Figure 5.12: Coverage probabilities for ϕ_1^2 with 95% approximate Binomial confidence intervals for each type of confidence interval construction method. Each of the four x-axis labels on each subplot denote whether the simulation variables $max\text{-}n_i$ and ACF are high or low respectively, thus "H/L" denotes $max\text{-}n_i = H$ and $ACF = L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the $LNEG$ setting for the simulation variable D.

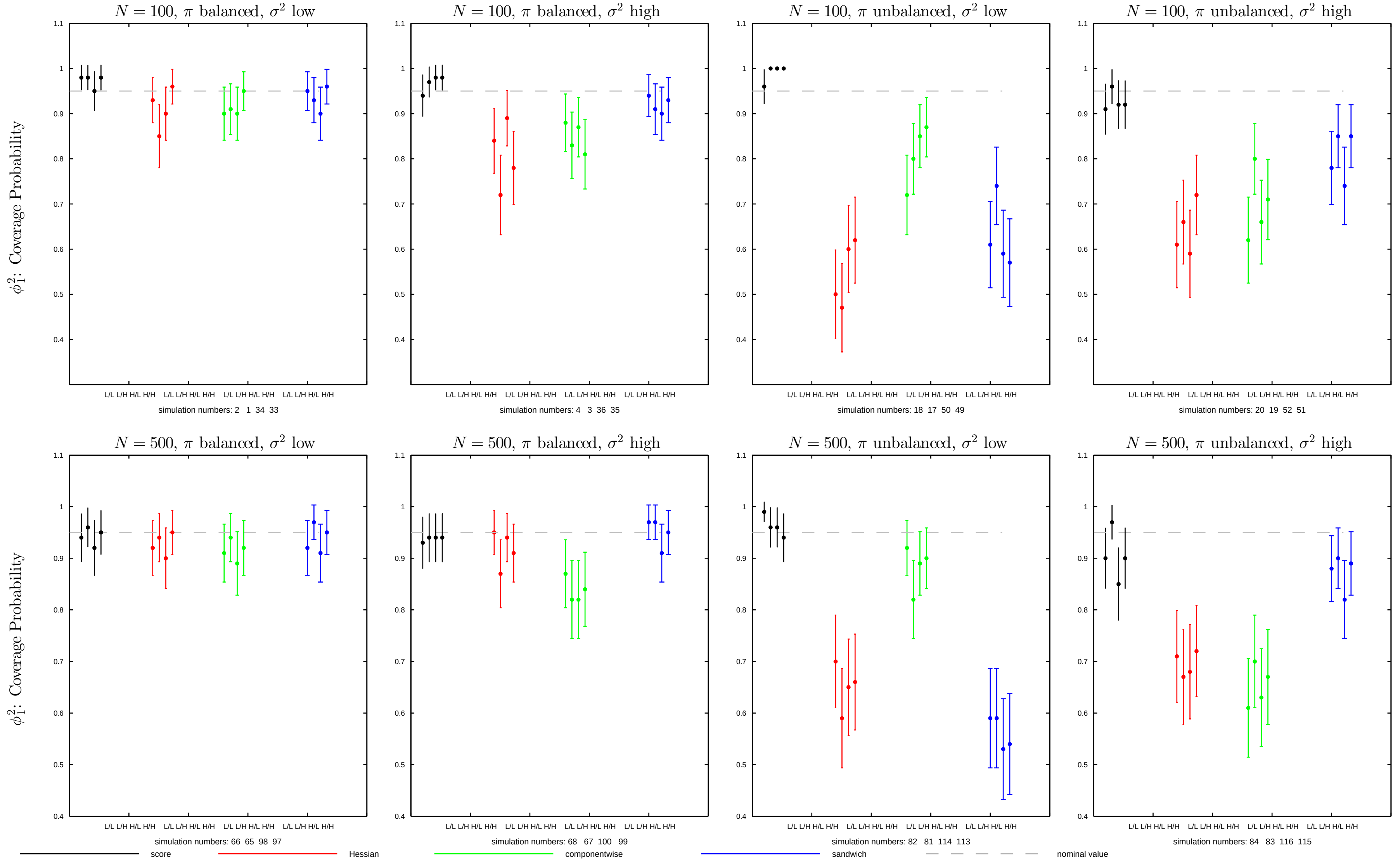


Figure 5.13: Boxplots of confidence interval lengths for ϕ_1^2 for each type of confidence interval construction method. Confidence interval lengths outside the 10^{th} and 90^{th} percentiles are displayed in a compression region - for an explanation see Subsection 5.1.4. Each of the four x-axis labels on each subplot denote whether the simulation variables $max\text{-}n_i$ and ACF are high or low respectively, thus "H/L" denotes $max\text{-}n_i = H$ and $ACF = L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the $LNEG$ setting for the simulation variable D.

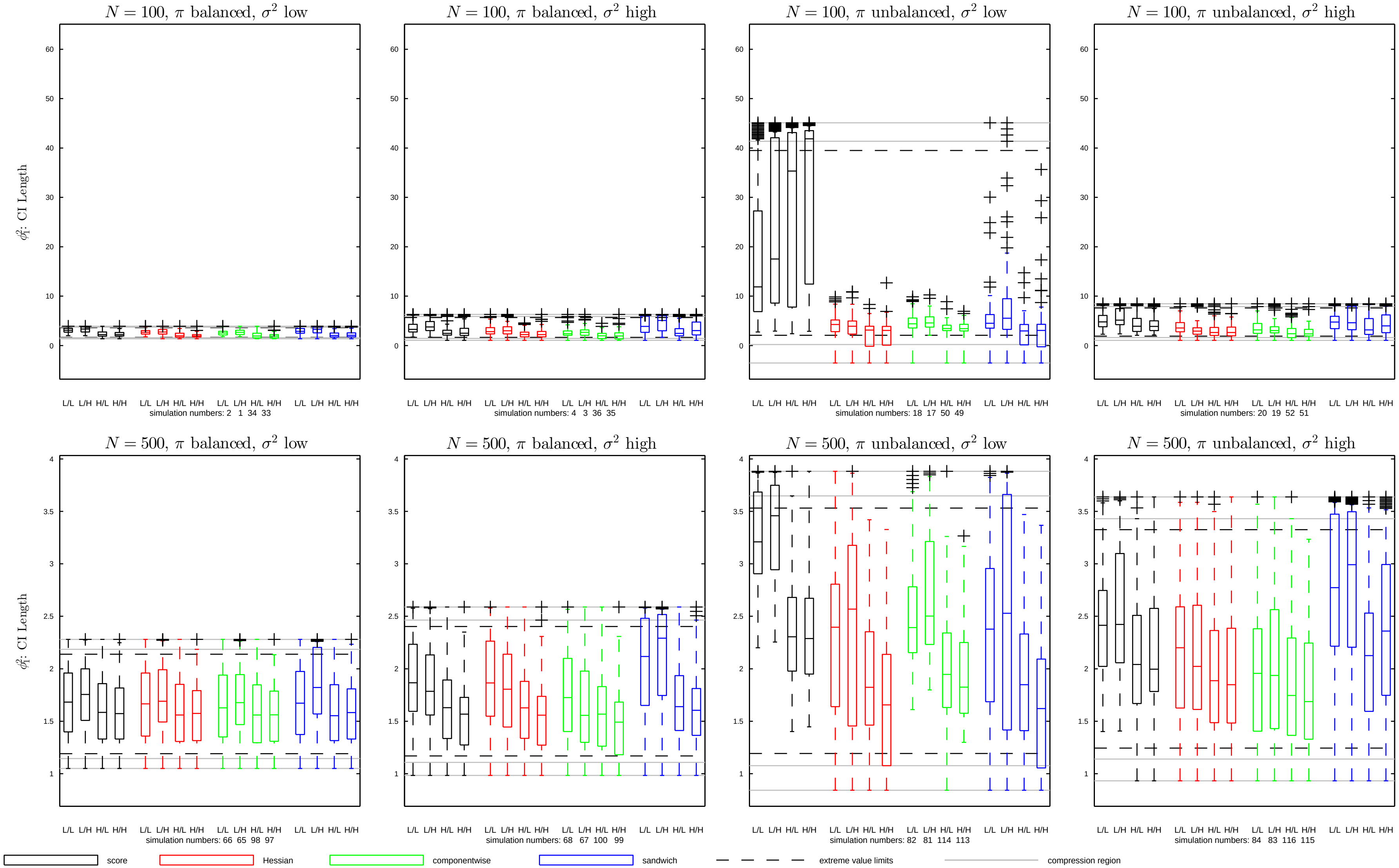


Figure 5.14: CPL indices for ϕ_1^2 for each type of confidence interval construction method. Each of the four x-axis labels on each subplot denote whether the simulation variables $max\text{-}n_i$ and ACF are high or low respectively, thus "H/L" denotes $max\text{-}n_i = H$ and $ACF = L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the $LNEG$ setting for the simulation variable D.

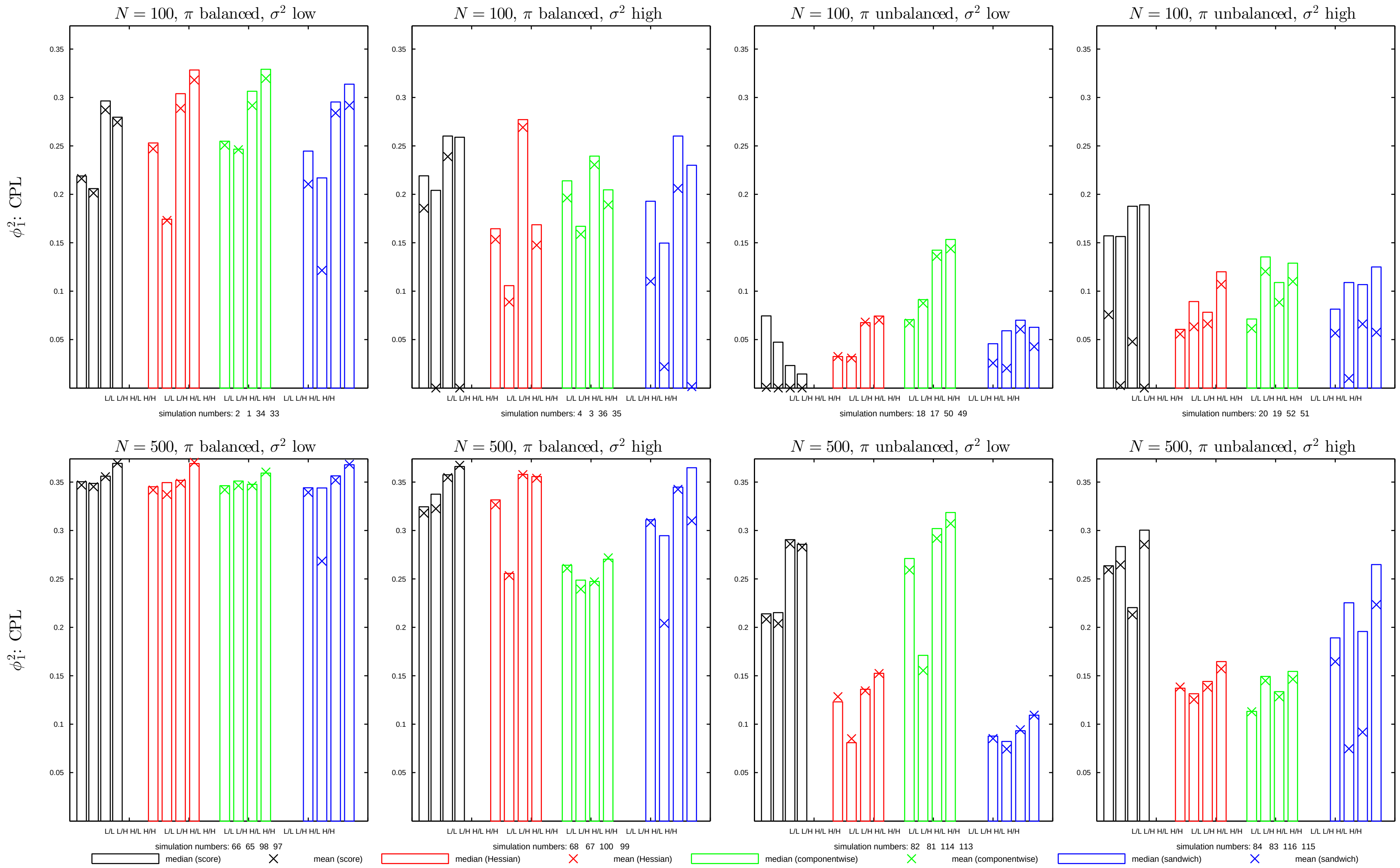


Table 5.7: Simulation parameter M-estimates with p-values less than 0.001 for MSE as the response.

Parameter	Level1	Level2	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$comp * \pi\text{-}balance$	1	$UNBA$	2.0015	0.0670	1.8701	2.1328	892.44	0.00000
$Intercept$			-4.7421	0.1708	-5.0769	-4.4074	770.77	0.00000
$param$	β^{c2}		-5.3212	0.1982	-5.7096	-4.9328	720.93	0.00000
$param$	β^{c1}		-5.1468	0.1982	-5.5352	-4.7583	674.44	0.00000
σ^2	H		2.4505	0.1005	2.2535	2.6475	594.58	0.00000
$param$	ϕ^2		-4.2044	0.1982	-4.5928	-3.8159	450.07	0.00000
$param$	ϕ^1		-4.1063	0.1982	-4.4947	-3.7178	429.31	0.00000
$param$	π		-2.7293	0.1982	-3.1178	-2.3409	189.67	0.00000
N	L		2.0183	0.1571	1.7103	2.3262	165.00	0.00000
D	$HNEG$		1.2549	0.1005	1.0579	1.4519	155.93	0.00000
$param$	β^{f11}		-2.3304	0.1982	-2.7188	-1.9420	138.27	0.00000
$param$	β^{f21}		-2.1686	0.1982	-2.5570	-1.7802	119.74	0.00000
$param$	β^{f22}		-2.0930	0.1982	-2.4814	-1.7045	111.53	0.00000
$param$	d_{11}		1.8568	0.1982	1.4683	2.2452	87.78	0.00000
$param * ACF$	σ^2	H	1.6063	0.1773	1.2589	1.9537	82.12	0.00000
$N * \sigma^2$	L	H	-0.4830	0.0670	-0.6143	-0.3517	51.97	0.00000
$param$	σ^2		-1.4281	0.1982	-1.8165	-1.0396	51.92	0.00000
$param$	d_{22}		1.2465	0.1982	0.8581	1.6349	39.56	0.00000
$param * max\text{-}n_i$	σ^2	L	1.0796	0.1773	0.7321	1.4270	37.09	0.00000
$comp * \sigma^2$	1	H	-0.3824	0.0670	-0.5137	-0.2510	32.57	0.00000
$param$	d_{21}		1.0216	0.1982	0.6331	1.4100	26.57	0.00000
$N * \pi\text{-}balance$	L	$UNBA$	-0.3080	0.0670	-0.4393	-0.1767	21.14	0.00000
$param$	β^{tc}		-0.7096	0.1982	-1.0981	-0.3212	12.82	0.00034
$param * comp$	β^{c1}	1	0.6311	0.1773	0.2836	0.9785	12.67	0.00037
$param * comp$	d_{22}	1	0.6308	0.1773	0.2834	0.9782	12.66	0.00037
$param * comp$	β^{tc}	1	0.6153	0.1773	0.2678	0.9627	12.05	0.00052
$param * N$	σ^2	L	0.6133	0.1773	0.2659	0.9608	11.97	0.00054
$param * comp$	d_{11}	1	-0.6012	0.1773	-0.9486	-0.2537	11.50	0.00070

Table 5.8: Simulation parameter M-estimates with p-values less than 0.001 for median-based CPL as the response.

Parameter	Level1	Level2	Level3	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$Intercept$				-0.4497	0.0024	-0.4545	-0.4449	33709.8778	0.00000
$comp * \pi\text{-}balance$	1	$UNBA$		-0.0384	0.0011	-0.0405	-0.0362	1199.2398	0.00000
$CI * \sigma^2$	$CI3$	H		-0.0384	0.0011	-0.0406	-0.0361	1145.9379	0.00000
$CI * \sigma^2$	$CI2$	H		-0.0228	0.0011	-0.0250	-0.0206	420.3780	0.00000
$N * \pi\text{-}balance$	L	$UNBA$		0.0107	0.0006	0.0096	0.0118	365.3803	0.00000
σ^2	H			-0.0179	0.0011	-0.0200	-0.0158	274.5516	0.00000
$CI * param$	$CI4$	π		-0.0645	0.0041	-0.0726	-0.0564	243.6831	0.00000
$param * comp$	β^{tc}	1		-0.0223	0.0015	-0.0252	-0.0195	233.3143	0.00000
$comp * N$	1	L		-0.0153	0.0011	-0.0175	-0.0131	190.4535	0.00000
$CI * comp * \sigma^2$	$CI2$	1	H	0.0207	0.0016	0.0176	0.0238	174.8206	0.00000
$CI * param * N$	$CI4$	π	L	0.0535	0.0041	0.0454	0.0616	167.5364	0.00000
$param * comp$	π	1		0.0201	0.0016	0.0170	0.0232	160.2531	0.00000
$\pi\text{-}balance$	$UNBA$			-0.0123	0.0010	-0.0143	-0.0102	138.5212	0.00000
$\pi\text{-}balance * \sigma^2$	$UNBA$	H		0.0065	0.0006	0.0054	0.0076	133.4656	0.00000
$CI * comp * \pi\text{-}balance$	$CI4$	1	$UNBA$	-0.0170	0.0016	-0.0200	-0.0139	117.1459	0.00000
$CI * comp * N$	$CI4$	1	L	0.0167	0.0016	0.0136	0.0198	112.9919	0.00000
$CI * comp * N$	$CI2$	1	L	0.0157	0.0016	0.0126	0.0188	99.9572	0.00000
$param * N$	ϕ_2	L		-0.0279	0.0029	-0.0336	-0.0222	91.2789	0.00000
$param * N$	π	L		-0.0279	0.0029	-0.0336	-0.0222	91.1711	0.00000
$CI * comp * \sigma^2$	$CI3$	1	H	0.0145	0.0016	0.0114	0.0177	82.7882	0.00000
$CI * comp$	$CI2$	1		-0.0194	0.0023	-0.0240	-0.0148	68.2752	0.00000
$CI * comp$	$CI4$	1		-0.0177	0.0023	-0.0223	-0.0131	56.9802	0.00000
$param * N$	d_{21}	L		-0.0218	0.0029	-0.0275	-0.0160	55.4689	0.00000
$CI * comp * \sigma^2$	$CI4$	1	H	0.0114	0.0016	0.0083	0.0145	52.9364	0.00000
$CI * comp * \pi\text{-}balance$	$CI3$	1	$UNBA$	0.0115	0.0016	0.0083	0.0146	51.5084	0.00000
$CI * comp * \pi\text{-}balance$	$CI2$	1	$UNBA$	-0.0111	0.0016	-0.0142	-0.0081	50.4329	0.00000
$CI * param$	$CI2$	π		-0.0287	0.0041	-0.0368	-0.0206	48.2830	0.00000
$param * N$	d_{11}	L		-0.0200	0.0029	-0.0258	-0.0143	47.0201	0.00000
$CI * param * N$	$CI2$	π	L	0.0277	0.0041	0.0196	0.0358	45.0293	0.00000
$param * comp$	β^{f22}	1		-0.0090	0.0015	-0.0118	-0.0061	37.5882	0.00000
$param * comp$	β^{f11}	1		-0.0086	0.0015	-0.0115	-0.0057	34.6398	0.00000
$param * N$	β^{tc}	L		-0.0164	0.0029	-0.0221	-0.0106	31.3617	0.00000
$param * comp$	ϕ_1	1		0.0084	0.0015	0.0054	0.0113	30.4536	0.00000
$CI * N$	$CI2$	L		-0.0156	0.0030	-0.0216	-0.0097	26.6623	0.00000

Table 5.8 continued.

Parameter	Level1	Level2	Level3	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$CI * N$	$CI4$	L		-0.0156	0.0030	-0.0215	-0.0096	26.4630	0.00000
$param * comp$	β^{c2}	1		-0.0064	0.0015	-0.0093	-0.0035	19.1410	0.00001
$N * D$	L	$HNEG$		-0.0030	0.0007	-0.0043	-0.0016	19.0725	0.00001
$param$	σ^2			0.0131	0.0030	0.0072	0.0190	19.0183	0.00001
$N * max\text{-}n_i$	L	L		-0.0024	0.0006	-0.0035	-0.0013	18.4809	0.00002
$CI * \pi\text{-}balance$	$CI3$	$UNBA$		-0.0049	0.0011	-0.0071	-0.0026	18.4308	0.00002
$param * comp$	β^{c1}	1		-0.0061	0.0015	-0.0089	-0.0032	17.3435	0.00003
$param * N$	β^{f21}	L		-0.0119	0.0029	-0.0176	-0.0061	16.4748	0.00005
$param * max\text{-}n_i$	π	L		-0.0117	0.0029	-0.0175	-0.0060	16.1214	0.00006
$CI * param * N$	$CI3$	ϕ_2	L	0.0163	0.0041	0.0082	0.0244	15.5807	0.00008
$CI * param * N$	$CI4$	β^{c1}	L	0.0162	0.0041	0.0081	0.0243	15.4518	0.00008
$param * max\text{-}n_i$	ϕ_2	L		-0.0114	0.0029	-0.0171	-0.0057	15.1842	0.00010
$param * N$	d_{22}	L		-0.0113	0.0029	-0.0170	-0.0055	14.8362	0.00012
$\pi\text{-}balance * D$	$UNBA$	$HNEG$		0.0026	0.0007	0.0013	0.0039	14.4063	0.00015
$CI * param * N$	$CI3$	d_{21}	L	0.0155	0.0041	0.0074	0.0236	14.0926	0.00017
$param * N$	β^{f22}	L		-0.0109	0.0029	-0.0167	-0.0052	14.0158	0.00018
$CI * param * N$	$CI3$	d_{11}	L	0.0155	0.0041	0.0074	0.0236	14.0143	0.00018
$param * ACF$	σ^2	H		-0.0105	0.0029	-0.0162	-0.0047	12.8552	0.00034
$param * comp$	σ^2	1		0.0052	0.0015	0.0023	0.0081	12.6273	0.00038
$CI * comp * N$	$CI3$	1	L	0.0055	0.0016	0.0023	0.0086	11.6786	0.00063
$param * N$	σ^2	L		-0.0100	0.0029	-0.0157	-0.0043	11.6705	0.00063
$param * comp$	β^{f21}	1		0.0050	0.0015	0.0021	0.0078	11.4905	0.00070
$comp * \sigma^2$	1	H		-0.0037	0.0011	-0.0059	-0.0016	11.3896	0.00074
$param * N$	β^{f11}	L		-0.0097	0.0029	-0.0154	-0.0040	11.0547	0.00088

5.3.3 Model 3

We first discuss the plots of the mixture model parameter estimates in the supplementary materials. The main result is that there was a clear relationship between estimation quality and both N and σ^2 when the mixing proportions were balanced. That is to say the estimates were more biased and had greater variability for $\sigma^2 = H$ compared to $\sigma^2 = L$, and that this relationship was stronger for $N = L$ compared to $N = H$. This relationship was still observable when the mixing proportions were unbalanced but it was less clear, and in this sense these results are more similar to Model 2 than to Model 1. In general the estimation quality for the mixture model parameters was good, particularly when $N = H$, $\sigma^2 = L$, and the mixing proportions were unbalanced.

Looking now at the M-estimates in Table 5.11 we see that by far the strongest four effects involve individual mixture model parameters. Compared to the mixture model intercept we see that the mixing proportions, β^{c1}, and β^{c2} are estimated with lower MSEs. In contrast d_{21} is estimated with a higher MSE than the mixture model intercept when $N = L$. The main effect for $param = d_{21}$ however shows that d_{21} is estimated well when interactions with other variables are accounted for. These results involving the mixture model parameters are broadly similar to those of Model 2 except that the factor variable fixed effects (which show these parameters were well estimated) are a little weaker than those of Model 1, and so appear lower down the table. The next two strongest effects are N and σ^2 which show the MSEs of all the mixture model parameters increased when $N = L$ and $\sigma^2 = H$. These effects were also observed in both Models1 and 2. However the three $comp * \sigma^2$ effects (one for components 1-3) show that there was a lot of between component variation in the effect of σ^2 on the MSEs.

Compared to the mixture model intercept the negative $param = \sigma^2$ effect suggests that σ^2 is estimated well, however since this is a main effect, and since σ^2 is involved in interactions, then this must be interpreted carefully. In this respect the two positive $param * comp$ effects for σ^2 for components 2 and 3 show there were strong differences between the components in terms of how well σ^2 was estimated. The $comp * \pi$-*balance* effect shows the MSE of all the mixture model parameter estimates increased when the mixing proportions were unbalanced (again in component 1 by design). The strength of this effect compared to the other effects is moderate, and so in this respect Model 3 is comparable to Model 1, but not to Model 2 where this was by far the strongest effect. Finally the *max*-n_i effect shows the MSEs of all the mixture model parameter estimates increased when the within-sample sizes were low, which is a similar to the result obtained for Model 1.

To summarise from the M-estimates for the MSEs, there was considerable variation between the mixture model parameters in terms of estimation quality, where the fixed effects parameters of the continuous variables were estimated the best. The effects of N, *max*-n_i, and the mixing proportions were to increase the MSEs at the non-optimal settings. Finally the quality of the parameter estimates was in general good, especially at the optimal settings of the simulation variables. One difference compared to Model 1 and Model 2 is that the levels of the random effects covariance parameters did not strongly influence the MSEs, nor were the estimates of the parameters themselves affected strongly by the other simulation variables.

We look now at the plots of the CPLIs in the supplementary materials. In contrast to Model 1 and 2 there does not appear to be an interaction between σ^2, N and π-*balance*. In general $CI4$ appears to produce the lowest CPLIs, whilst there appears to be no difference between the other three methods. The highest CPLIs were produced by $CI1$, whilst slightly lower CPLIs were produced by $CI3$ and then by $CI2$, although these latter two methods did not differ by much. This superiority of $CI1$ was observed in both Model 1 and Model 2, whilst $CI4$ being the worst method was also observed in Model 1, although the method was not as poor compared to the other methods as it is here. Furthermore the CPLIs for $CI3$ for both this model and Model 1 were the second highest, and in Model 2 they were similar to all the methods apart from $CI1$. Thus the important result here is that $CI3$ which is based on the LMM information matrix consistently produces CPLIs that compare favorably with the other methods that are based on approximations to the MLMM information matrix.

Looking at the plots of the CPs, we see for balanced mixing proportions that $CI1$ produced the highest coverage probabilities, and that there was not much difference between the other three methods. When the mixing proportions were unbalanced we see that $CI1$ produced coverage probabilities that were more similar to $CI2$ and $CI3$. Furthermore $CI4$ produced the lowest coverage probabilities. For all methods, and particularly when the mixing proportions were balanced, the ranges of coverage often intersected the nominal level or were close to it. In contrast to Model 2 there is no strong effect of σ^2. Again we have the important result here that $CI3$, which is based on the LMM information matrix, consistently produces CPs that compare favorably with the other methods that are based on approximations to the MLMM information matrix.

Looking at the CIL plots we see a $N{*}\sigma^2{*}\pi$-*balance* interaction. That is for balanced mixing proportions we see higher median lengths and greater variability for $\sigma^2 = H$ compared to $\sigma^2 = L$, and that this effect was larger for $N = L$ compared to $N = H$. When $\sigma^2 = H$ it is clear that $CI1$ produced slightly higher median lengths than the other three methods which themselves produce similar lengths. For unbalanced mixing proportions this relationship between N and σ^2 is slightly more pronounced. Overall $CI1$ produces the longest confidence interval lengths, whilst the lengths produced by the other three methods are similar. This result is similar to Model 1.

To summarise from the plots, we see that $CI1$ produces the best confidence intervals in terms of both the CPLI index and the coverage probabilities, and that $CI3$ is the next best method. Because the CPLI index accounts for confidence interval lengths then we can conclude the superiority of $CI1$ and $CI2$ was not a result of excessively long confidence intervals. The good performance here of $CI3$ is stronger than for Model 1 and Model 2. Finally the effects of σ^2, N, and the balancedness of the mixing proportions on the CPs and CILs were reasonably strong, however their effects on the CPLIs were weak.

We now look at the M-estimates for CPLI as the response variable in Table 5.12. Many of the strongest effects are comprised of interactions that involve $CI4$, many of which are negative. We have mentioned before that since we have no reason to pick $CI1$ as the reference category that we should not over-interpret these effects involving CI. Thus we should conclude only that the relationship between the CPLIs and *comp*, σ^2, and π-*balance* were very different for $CI4$ and $CI1$. Similarly there are a few weaker effects involving interactions between $CI2$, $CI3$ and *comp*, σ^2, π-*balance*, and N. Thus just as for Model 2 there was considerable variation between the confidence interval methods and a number of the simulation variables. There are also some $param * comp$, and $param * N$ effects at the top of the table which show there were strong differences between the CPLIs of the mixture model parameters, and that these differences also varied by component and the level of N. These results are similar to the results for Model 1 and Model 2.

The effects $N * \sigma^2$, and $N * max\text{-}n_i$ show that the CPLIs increase for all of the mixture model parameters when these simulation variables were set to the non-optimal levels compared to the optimal ones. These results are similar to Model 2 but not Model 1. The effect of the mixing proportions being unbalanced only comes through strongly in the three way interaction effects $CI * comp * \pi\text{-}balance$ involving $CI4$ and $CI2$ (and again in component 1), and this result is similar to Model 1 but not for Model 2 where the two way interaction $comp * \pi\text{-}balance$ was much stronger than these three way interactions.

In summary for the CPLIs, there was considerable variation between the confidence interval methods with respect to the effect on the CPLIs of the simulation variables *comp*, σ^2, and π-*balance*. This variation with confidence interval method is similar to the results of Model 2 but not Model 1. There was considerable variation between the mixture model parameters with respect to the effects on the CPLIs of the simulation variables *comp* and N, and this is similar to the results obtained for both Model 1 and Model 2. The effects on the CPLIs of the mixing proportions being unbalanced, and of the simulation variables N and σ^2 were also strong, and these results are similar to both Model 1 and Model 2. The effect of *max*-n_i on the CPLIs was strong and this is a similar result to Model 1 but not Model 2. Finally, and similar to Model 2, there were no strong effects involving the random effects covariance matrix $\boldsymbol{D}$ (which for this model is a scalar parameter).

Table 5.9: Simulation variable settings for the 64 runs of Model 3.

simnumber	N	*max*-n_i	π-*unbalance* : $(\pi_1,\pi_2,\pi_3,\pi_4)$	$D : (D_1,D_2,D_3,D_4)$	$\sigma^2 : (\sigma_1^2,\sigma_2^2,\sigma_3^2,\sigma_4^2)$	n_i-*unbalance* : $E(X) = np$ of n
1	100	6	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	5.7 of 6
2	100	6	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	3.9 of 6
3	100	6	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	5.7 of 6
4	100	6	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	3.9 of 6
5	100	6	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	5.7 of 6
6	100	6	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	3.9 of 6
7	100	6	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	5.7 of 6
8	100	6	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	3.9 of 6
9	100	6	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	5.7 of 6
10	100	6	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	3.9 of 6
11	100	6	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	5.7 of 6
12	100	6	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	3.9 of 6
13	100	6	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	5.7 of 6
14	100	6	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	3.9 of 6
15	100	6	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	5.7 of 6
16	100	6	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	3.9 of 6
17	100	10	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	9.5 of 10
18	100	10	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	6.5 of 10
19	100	10	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	9.5 of 10
20	100	10	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	6.5 of 10
21	100	10	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	9.5 of 10
22	100	10	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	6.5 of 10
23	100	10	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	9.5 of 10
24	100	10	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	6.5 of 10
25	100	10	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	9.5 of 10
26	100	10	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	6.5 of 10
27	100	10	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	9.5 of 10
28	100	10	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	6.5 of 10
29	100	10	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	9.5 of 10
30	100	10	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	6.5 of 10
31	100	10	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	9.5 of 10
32	100	10	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	6.5 of 10
33	1000	6	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	5.7 of 6
34	1000	6	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	3.9 of 6
35	1000	6	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	5.7 of 6
36	1000	6	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	3.9 of 6
37	1000	6	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	5.7 of 6
38	1000	6	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	3.9 of 6
39	1000	6	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	5.7 of 6
40	1000	6	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	3.9 of 6

Table 5.10: Simulation variable settings for the 64 runs of Model 3.

simnumber	$\boldsymbol{N}$	***max-n_i***	$\boldsymbol{\pi}$***-unbalance*** : $(\pi_1,\pi_2,\pi_3,\pi_4)$	$D:(D_1,D_2,D_3,D_4)$	$\sigma^2:(\sigma_1^2,\sigma_2^2,\sigma_3^2,\sigma_4^2)$	n_i***-unbalance*** : $E(X)=np$ **of** n
41	1000	6	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	5.7 of 6
42	1000	6	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	3.9 of 6
43	1000	6	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	5.7 of 6
44	1000	6	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	3.9 of 6
45	1000	6	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	5.7 of 6
46	1000	6	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	3.9 of 6
47	1000	6	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	5.7 of 6
48	1000	6	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	3.9 of 6
49	1000	10	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	9.5 of 10
50	1000	10	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	6.5 of 10
51	1000	10	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	9.5 of 10
52	1000	10	(0.25,0.25,0.25,0.25)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	6.5 of 10
53	1000	10	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	9.5 of 10
54	1000	10	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	6.5 of 10
55	1000	10	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	9.5 of 10
56	1000	10	(0.25,0.25,0.25,0.25)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	6.5 of 10
57	1000	10	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	9.5 of 10
58	1000	10	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(1,1.2,1.1,1.9)	6.5 of 10
59	1000	10	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	9.5 of 10
60	1000	10	(0.15,0.2833,0.2833,0.2833)	(0.8,0.3,0.5,0.7)	(6,6.2,6.1,6.9)	6.5 of 10
61	1000	10	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	9.5 of 10
62	1000	10	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(1,1.2,1.1,1.9)	6.5 of 10
63	1000	10	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	9.5 of 10
64	1000	10	(0.15,0.2833,0.2833,0.2833)	(5.8,5.3,5.5,5.7)	(6,6.2,6.1,6.9)	6.5 of 10

Figure 5.15: Boxplots of parameter estimates for σ_4^2 with estimates outside the 10^{th} and 90^{th} percentiles displayed in a compression region - for an explanation see Subsection 5.1.4. Each of the four x-axis labels on each subplot denote whether the simulation variables *max-*n_i and n_i*-unbalance* are high or low respectively, thus "H/L" denotes *max-*$n_i = H$ and n_i*-unbalance* $= L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the L setting for the simulation variable D.

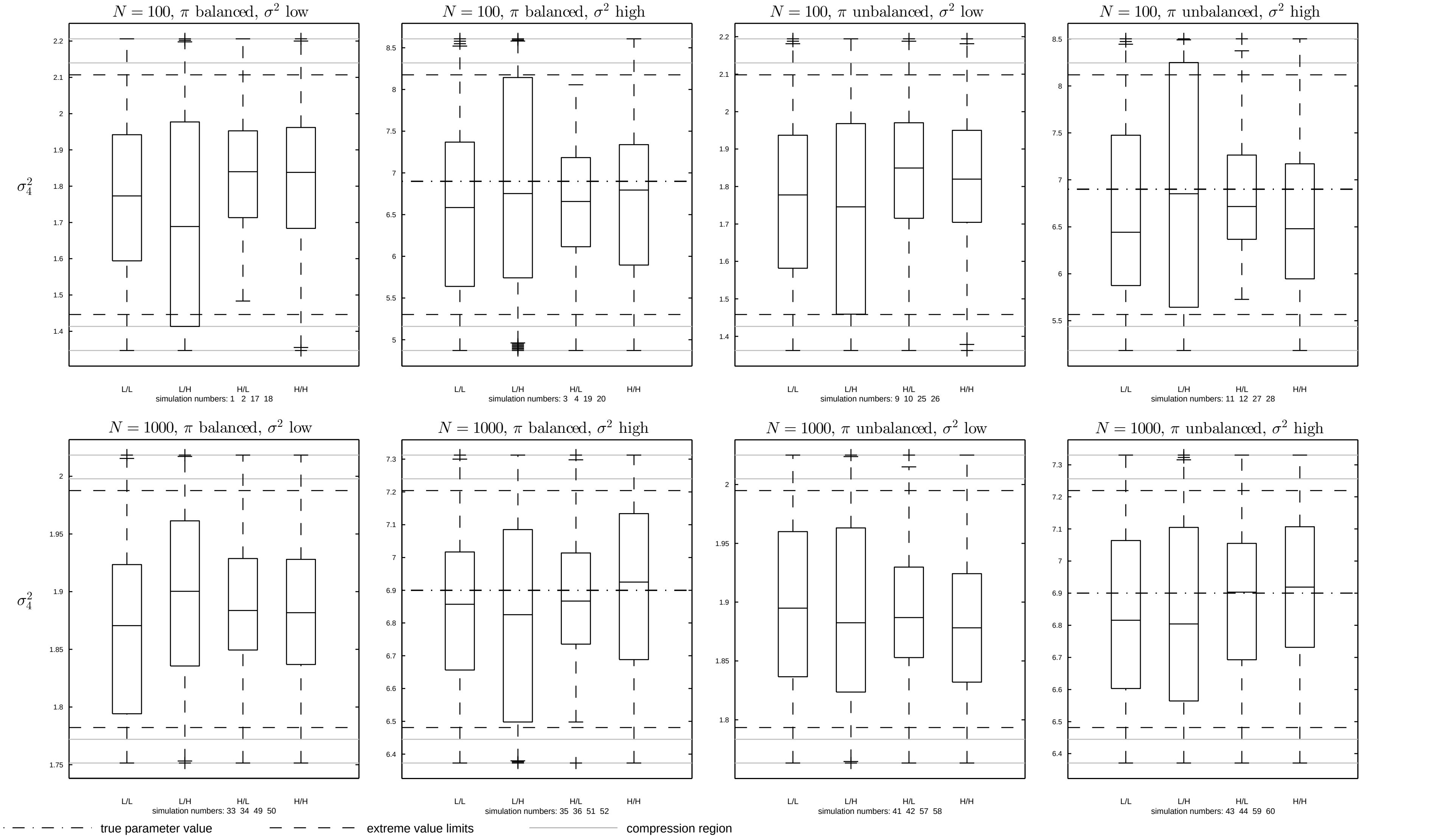


Figure 5.16: Coverage probabilities for σ_4^2 with 95% approximate Binomial confidence intervals for each type of confidence interval construction method. Each of the four x-axis labels on each subplot denote whether the simulation variables *max-*n_i and n_i*-unbalance* are high or low respectively, thus "H/L" denotes *max-*$n_i = H$ and n_i*-unbalance* $= L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the L setting for the simulation variable D.

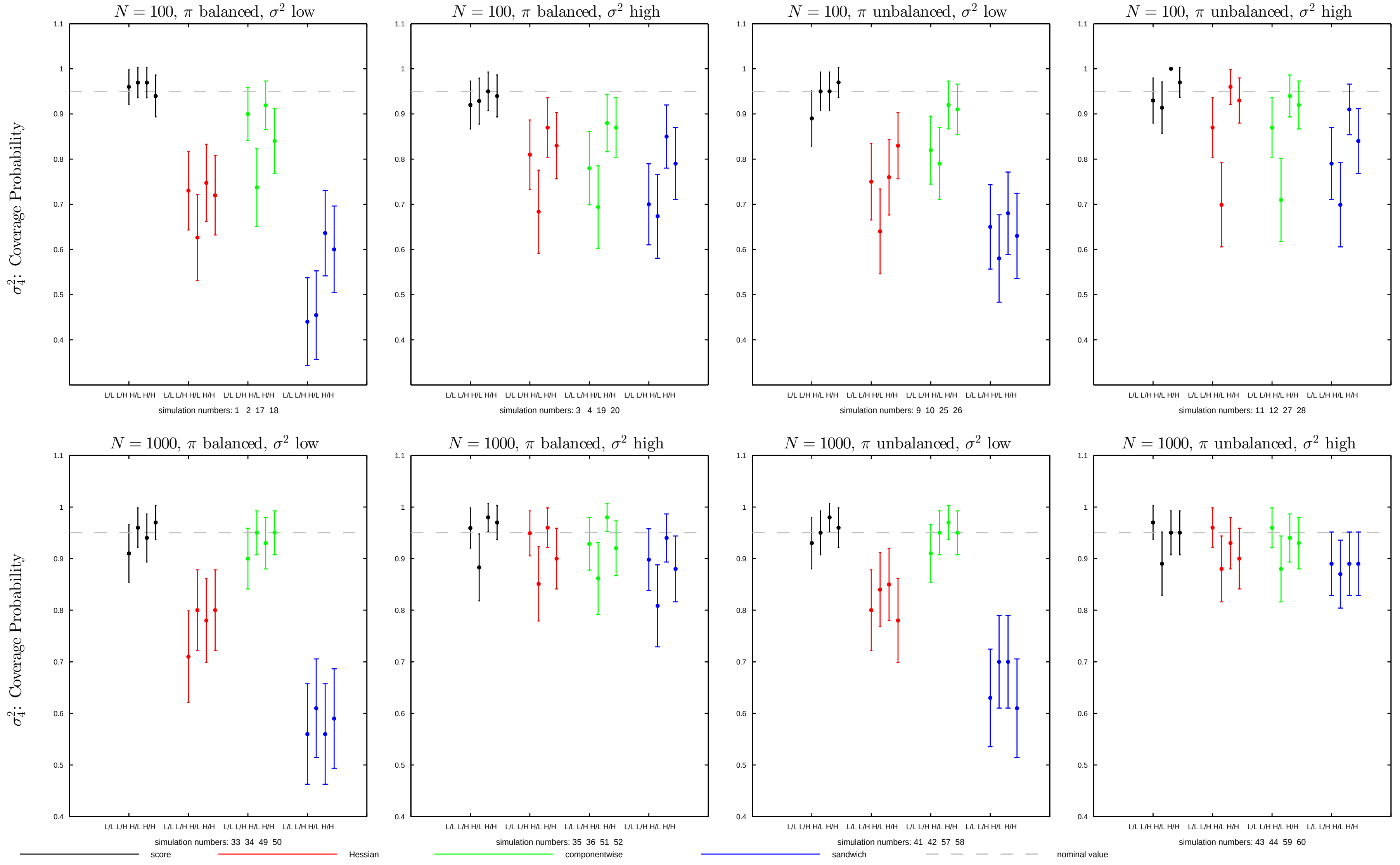


Figure 5.17: Boxplots of confidence interval lengths for σ_4^2 for each type of confidence interval construction method. Confidence interval lengths outside the 10^{th} and 90^{th} percentiles are displayed in a compression region - for an explanation see Subsection 5.1.4. Each of the four x-axis labels on each subplot denote whether the simulation variables $max\text{-}n_i$ and $n_i\text{-}unbalance$ are high or low respectively, thus "H/L" denotes $max\text{-}n_i = H$ and $n_i\text{-}unbalance = L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the L setting for the simulation variable D.

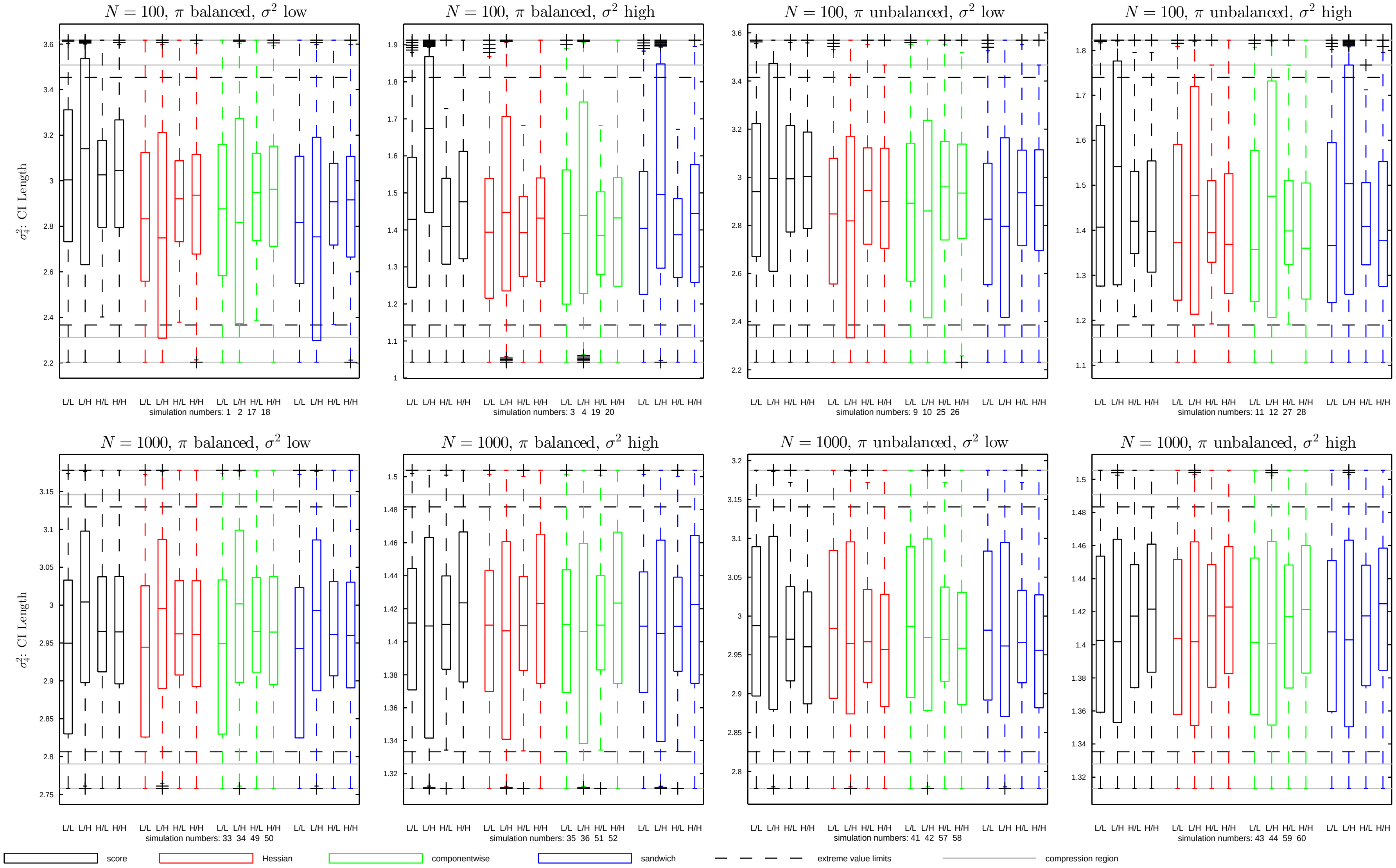


Figure 5.18: CPL indices for σ_4^2 for each type of confidence interval construction method. Each of the four x-axis labels on each subplot denote whether the simulation variables *max-*n_i and n_i*-unbalance* are high or low respectively, thus "H/L" denotes *max-*$n_i = H$ and n_i*-unbalance* $= L$ respectively. In total all of the plots show results for 32 combinations of the simulation variables, all at the L setting for the simulation variable D.

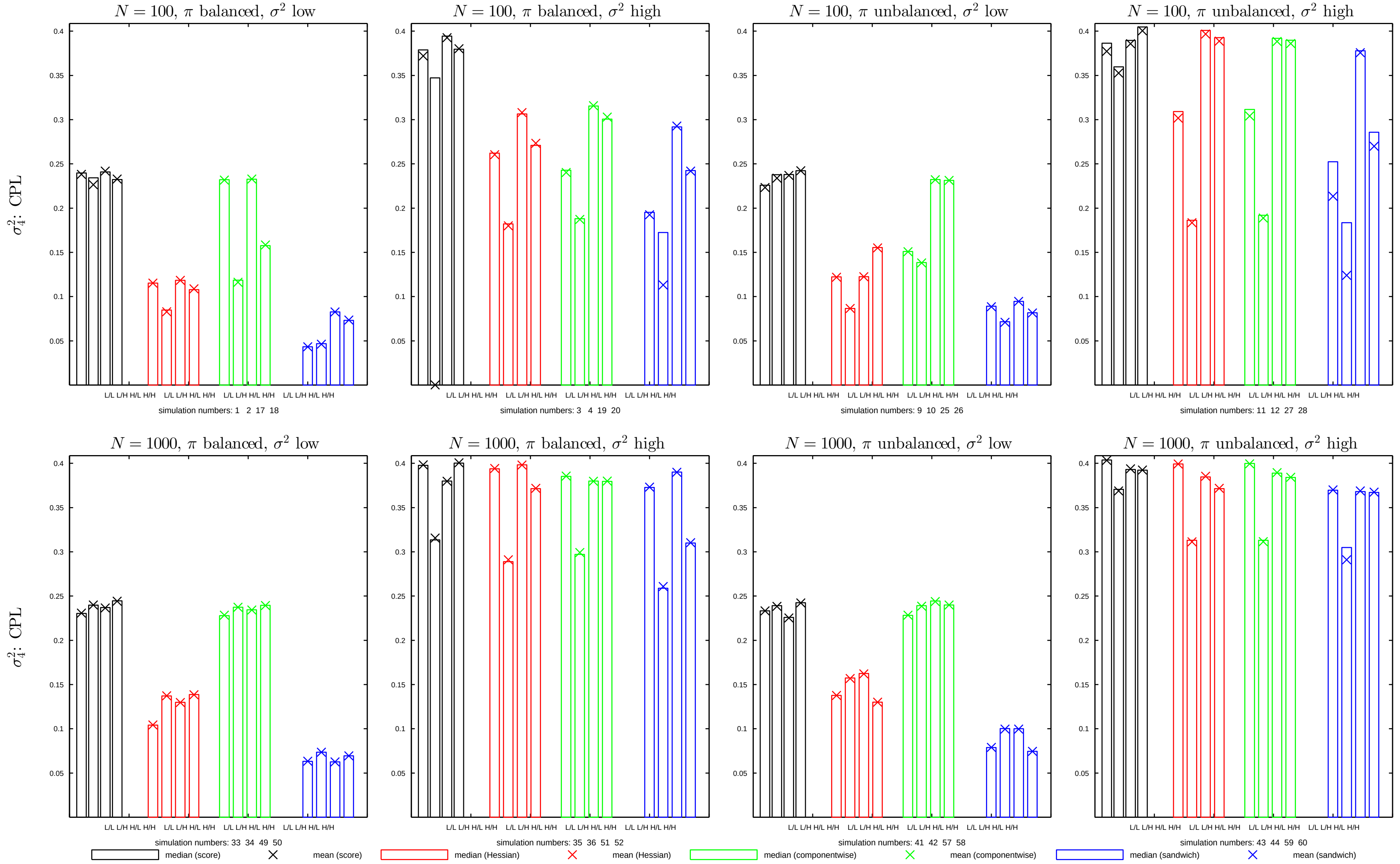


Table 5.11: Simulation parameter M-estimates with p-values less than 0.001 for MSE as the response.

Parameter	Level1	Level2	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$Intercept$			-4.1671	0.0414	-4.2483	-4.0859	10114.64	0.00000
$param$	π		-4.3633	0.0488	-4.4590	-4.2677	7993.33	0.00000
$param * N$	d_{21}	L	2.9133	0.0369	2.8410	2.9856	6235.83	0.00000
$param$	β^{c2}		-3.5694	0.0488	-3.6650	-3.4737	5348.92	0.00000
$param$	β^{c1}		-3.5465	0.0488	-3.6422	-3.4509	5280.67	0.00000
N	L		2.2324	0.0343	2.1652	2.2996	4239.56	0.00000
σ^2	H		1.2679	0.0236	1.2216	1.3141	2887.04	0.00000
$param$	σ^2		-1.5276	0.0488	-1.6232	-1.4319	979.69	0.00000
$comp * \pi\text{-}balance$	1	$UNBA$	0.6831	0.0222	0.6395	0.7267	942.88	0.00000
$param * comp$	σ^2	2	1.2929	0.0522	1.1906	1.3951	614.07	0.00000
$param * comp$	σ^2	3	1.2842	0.0522	1.1819	1.3864	605.82	0.00000
$comp * \sigma^2$	1	H	0.5358	0.0222	0.4922	0.5794	580.08	0.00000
$param$	d_{21}		-0.9890	0.0488	-1.0846	-0.8933	410.62	0.00000
$max\text{-}n_i$	L		0.5861	0.0343	0.5189	0.6533	292.27	0.00000
$param * max\text{-}n_i$	π	L	-0.5835	0.0369	-0.6558	-0.5112	250.17	0.00000
$comp$	1		-0.6522	0.0459	-0.7421	-0.5623	202.24	0.00000
$param * comp$	σ^2	1	-0.7096	0.0522	-0.8118	-0.6073	184.97	0.00000
$comp * \sigma^2$	3	H	0.2752	0.0222	0.2316	0.3189	153.08	0.00000
$param * comp$	d_{21}	2	-0.6044	0.0522	-0.7066	-0.5021	134.19	0.00000
$param * comp$	π	1	-0.5770	0.0522	-0.6793	-0.4748	122.31	0.00000
$comp * \sigma^2$	2	H	0.2351	0.0222	0.1915	0.2787	111.69	0.00000
$param * comp$	π	3	0.5466	0.0522	0.4444	0.6489	109.77	0.00000
$param * comp$	π	2	0.5395	0.0522	0.4372	0.6417	106.91	0.00000
$comp$	2		-0.4741	0.0459	-0.5640	-0.3842	106.86	0.00000
$n_i\text{-}unbalance$	H		0.3123	0.0324	0.2488	0.3759	92.75	0.00000
$comp$	3		-0.4187	0.0459	-0.5086	-0.3288	83.35	0.00000
$param$	β^{f22}		-0.4064	0.0488	-0.5020	-0.3107	69.33	0.00000
$param$	β^{f21}		-0.4036	0.0488	-0.4992	-0.3079	68.37	0.00000
$param * n_i\text{-}unbalance$	π	H	-0.2874	0.0369	-0.3597	-0.2151	60.70	0.00000
$param$	β^{f11}		-0.3734	0.0488	-0.4690	-0.2777	58.53	0.00000
$N * max\text{-}n_i$	L	L	0.1185	0.0157	0.0877	0.1494	56.79	0.00000
$comp * max\text{-}n_i$	3	L	-0.1504	0.0222	-0.1940	-0.1068	45.68	0.00000
$param$	$\beta^{f11*f22}$		0.3148	0.0488	0.2191	0.4104	41.60	0.00000
$N * n_i\text{-}unbalance$	L	H	0.0993	0.0157	0.0685	0.1301	39.83	0.00000

Table 5.11 continued.

Parameter	Level1	Level2	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$param * comp$	d_{21}	3	-0.3243	0.0522	-0.4265	-0.2220	38.63	0.00000
$comp * max\text{-}n_i$	2	L	-0.1365	0.0222	-0.1802	-0.0929	37.67	0.00000
$param$	$\beta^{f11*f21}$		0.2724	0.0488	0.1768	0.3681	31.16	0.00000
$\pi\text{-}balance$	$UNBA$		-0.1161	0.0222	-0.1597	-0.0725	27.24	0.00000
$max\text{-}n_i * \sigma^2$	L	H	0.0792	0.0157	0.0484	0.1101	25.37	0.00000
D	H		0.0840	0.0222	0.0404	0.1276	14.26	0.00016
$param * n_i\text{-}unbalance$	σ^2	H	0.1373	0.0369	0.0650	0.2096	13.85	0.00020
$N * \sigma^2$	L	H	0.0539	0.0157	0.0231	0.0848	11.76	0.00061
$N * D$	L	H	0.0523	0.0157	0.0215	0.0831	11.05	0.00089

Table 5.12: Simulation parameter M-estimates with p-values less than 0.001 for median-based CPL as the response.

Parameter	Level1	Level2	Level3	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$Intercept$				-0.3245	0.0005	-0.3255	-0.3235	411899.5641	0.00000
$CI * \sigma^2$	$CI4$	H		0.0101	0.0003	0.0095	0.0107	984.8718	0.00000
$CI * comp * \pi\text{-}balance$	$CI4$	1	$UNBA$	-0.0138	0.0005	-0.0147	-0.0129	920.9299	0.00000
$CI * comp * \sigma^2$	$CI4$	1	H	-0.0115	0.0005	-0.0124	-0.0106	640.7803	0.00000
$param * comp$	β^{f22}	1		-0.0094	0.0004	-0.0101	-0.0087	618.2331	0.00000
$CI * comp * \sigma^2$	$CI4$	3	H	-0.0104	0.0005	-0.0112	-0.0095	516.7816	0.00000
$param * comp$	$\beta^{f11*f22}$	2		-0.0085	0.0004	-0.0093	-0.0078	510.7403	0.00000
$CI * comp$	$CI4$	3		0.0135	0.0006	0.0124	0.0147	505.1034	0.00000
$CI * comp$	$CI4$	2		0.0130	0.0006	0.0119	0.0142	467.9571	0.00000
$CI * comp$	$CI4$	1		0.0128	0.0006	0.0117	0.0140	454.5160	0.00000
$CI * comp * \sigma^2$	$CI4$	2	H	-0.0095	0.0005	-0.0104	-0.0087	439.5729	0.00000
CI	$CI4$			-0.0132	0.0007	-0.0145	-0.0119	397.1133	0.00000
$N * \sigma^2$	L	H		-0.0021	0.0001	-0.0023	-0.0019	337.6337	0.00000
$N * max\text{-}n_i$	L	L		-0.0020	0.0001	-0.0023	-0.0018	312.8770	0.00000
$param * N$	d_{21}	L		-0.0094	0.0005	-0.0105	-0.0084	311.4418	0.00000
$CI * comp * \pi\text{-}balance$	$CI2$	1	$UNBA$	-0.0077	0.0005	-0.0086	-0.0068	282.9383	0.00000
$param * comp$	β^{f11}	3		-0.0053	0.0004	-0.0061	-0.0046	200.3536	0.00000
$comp * N$	1	L		-0.0043	0.0003	-0.0050	-0.0037	181.9072	0.00000
$CI * param * N$	$CI3$	σ^2	L	-0.0095	0.0008	-0.0110	-0.0080	158.3917	0.00000
$param * comp$	σ^2	3		0.0047	0.0004	0.0039	0.0054	153.4811	0.00000
$CI * comp * \sigma^2$	$CI2$	1	H	-0.0056	0.0005	-0.0065	-0.0047	153.4179	0.00000
$param * comp$	σ^2	1		0.0047	0.0004	0.0039	0.0054	152.6736	0.00000
$param * comp$	β^{f21}	1		-0.0043	0.0004	-0.0051	-0.0036	132.3080	0.00000
$param * comp$	σ^2	2		0.0043	0.0004	0.0036	0.0051	132.0061	0.00000
$param * N$	$\beta^{f11*f22}$	L		-0.0061	0.0005	-0.0072	-0.0051	131.8725	0.00000
$N * n_i\text{-}unbalance$	L	H		-0.0012	0.0001	-0.0014	-0.0009	100.6428	0.00000
$CI * comp * N$	$CI4$	1	L	0.0045	0.0005	0.0036	0.0053	95.6291	0.00000
$CI * comp * N$	$CI2$	1	L	0.0044	0.0005	0.0035	0.0052	91.4812	0.00000
$CI * comp * N$	$CI2$	2	L	0.0042	0.0005	0.0033	0.0051	86.7428	0.00000
$max\text{-}n_i * \sigma^2$	L	H		-0.0011	0.0001	-0.0013	-0.0008	83.2277	0.00000
$CI * \pi\text{-}balance$	$CI4$	$UNBA$		0.0029	0.0003	0.0022	0.0035	79.8971	0.00000
$max\text{-}n_i * n_i\text{-}unbalance$	L	H		-0.0010	0.0001	-0.0012	-0.0008	75.1577	0.00000
$param * N$	β^{f21}	L		-0.0045	0.0005	-0.0056	-0.0035	71.7741	0.00000
$CI * comp * N$	$CI2$	3	L	0.0037	0.0005	0.0028	0.0046	66.1908	0.00000

Table 5.12 continued.

Parameter	Level1	Level2	Level3	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$CI * comp * max\text{-}n_i$	$CI2$	3	L	0.0035	0.0005	0.0026	0.0044	58.3387	0.00000
$CI * comp * max\text{-}n_i$	$CI2$	2	L	0.0035	0.0005	0.0026	0.0044	57.6209	0.00000
$CI * \sigma^2$	$CI2$	H		0.0024	0.0003	0.0018	0.0030	56.0438	0.00000
$\sigma^2 * n_i\text{-}unbalance$	H	H		-0.0009	0.0001	-0.0011	-0.0006	55.7746	0.00000
$CI * \pi\text{-}balance$	$CI2$	$UNBA$		0.0024	0.0003	0.0017	0.0030	54.0271	0.00000
$CI * N$	$CI2$	L		-0.0044	0.0006	-0.0056	-0.0032	52.6275	0.00000
$param * N$	β^{f22}	L		-0.0039	0.0005	-0.0049	-0.0028	52.4136	0.00000
N	L			0.0031	0.0004	0.0022	0.0040	48.9769	0.00000
$CI * comp$	$CI2$	1		0.0041	0.0006	0.0030	0.0053	47.3245	0.00000
$comp * \sigma^2$	1	H		-0.0022	0.0003	-0.0028	-0.0015	45.4237	0.00000
$CI * comp * \pi\text{-}balance$	$CI4$	3	$UNBA$	-0.0030	0.0005	-0.0039	-0.0021	42.9438	0.00000
$CI * max\text{-}n_i$	$CI2$	L		-0.0038	0.0006	-0.0050	-0.0027	40.4347	0.00000
$param$	σ^2			-0.0034	0.0006	-0.0045	-0.0023	34.1581	0.00000
$param * N$	$\beta^{f11*f21}$	L		-0.0031	0.0005	-0.0041	-0.0021	33.7166	0.00000
$CI * comp * \sigma^2$	$CI2$	3	H	-0.0026	0.0005	-0.0035	-0.0017	31.5268	0.00000
$CI * comp * \pi\text{-}balance$	$CI2$	3	$UNBA$	-0.0026	0.0005	-0.0034	-0.0017	31.4936	0.00000
$CI * comp * \pi\text{-}balance$	$CI4$	2	$UNBA$	-0.0026	0.0005	-0.0034	-0.0017	31.4466	0.00000
$param * N$	β^{f11}	L		-0.0029	0.0005	-0.0040	-0.0019	30.3786	0.00000
$comp * \pi\text{-}balance$	1	$UNBA$		-0.0017	0.0003	-0.0024	-0.0011	29.3203	0.00000
σ^2	H			0.0014	0.0003	0.0009	0.0019	28.6525	0.00000
$comp * N$	2	L		-0.0017	0.0003	-0.0023	-0.0011	27.8753	0.00000
$comp$	1			0.0026	0.0005	0.0016	0.0036	27.4127	0.00000
$param$	$\beta^{f11*f22}$			0.0030	0.0006	0.0018	0.0041	26.3972	0.00000
$param * comp$	β^{c1}	3		-0.0019	0.0004	-0.0027	-0.0012	26.1862	0.00000
$param * comp$	d_{21}	2		-0.0019	0.0004	-0.0026	-0.0012	25.5182	0.00000
$CI * comp * \pi\text{-}balance$	$CI2$	2	$UNBA$	-0.0023	0.0005	-0.0032	-0.0014	24.8066	0.00000
$max\text{-}n_i$	L			0.0021	0.0004	0.0013	0.0030	23.2620	0.00000
$CI * comp * max\text{-}n_i$	$CI2$	1	L	0.0022	0.0005	0.0013	0.0031	22.9386	0.00000
$CI * param * N$	$CI4$	σ^2	L	-0.0036	0.0008	-0.0051	-0.0021	22.9129	0.00000
$CI * param$	$CI4$	d_{21}		-0.0036	0.0008	-0.0050	-0.0021	22.3321	0.00000
$param * comp$	π	3		-0.0019	0.0004	-0.0027	-0.0011	21.9448	0.00000
$CI * n_i\text{-}unbalance$	$CI2$	H		-0.0028	0.0006	-0.0039	-0.0016	20.9490	0.00000
$CI * comp * \sigma^2$	$CI2$	2	H	-0.0021	0.0005	-0.0030	-0.0012	20.8180	0.00001
$CI * param * N$	$CI2$	σ^2	L	-0.0033	0.0008	-0.0048	-0.0018	18.8586	0.00001

Table 5.12 continued.

Parameter	Level1	Level2	Level3	Estimate	StdErr	LowerCL	UpperCL	ChiSq	ProbChiSq
$param * comp$	π	2		-0.0018	0.0004	-0.0026	-0.0010	18.6906	0.00002
$CI * comp * n_i\text{-}unbalance$	$CI2$	2	H	0.0019	0.0005	0.0011	0.0028	18.2801	0.00002
$N * \pi\text{-}balance$	L	$UNBA$		0.0005	0.0001	0.0003	0.0007	18.2185	0.00002
$param * comp$	$\beta^{f11*f22}$	1		-0.0015	0.0004	-0.0023	-0.0008	16.3110	0.00005
$CI * comp * n_i\text{-}unbalance$	$CI2$	3	H	0.0018	0.0005	0.0009	0.0027	15.9208	0.00007
$CI * N$	$CI4$	L		-0.0023	0.0006	-0.0035	-0.0011	14.3926	0.00015
$CI * param * N$	$CI2$	$\beta^{f11*f22}$	L	0.0027	0.0008	0.0012	0.0042	12.5552	0.00040
$param * comp$	$\beta^{f11*f21}$	2		-0.0013	0.0004	-0.0021	-0.0006	12.4682	0.00041
$CI * param * N$	$CI3$	$\beta^{f11*f22}$	L	0.0026	0.0008	0.0011	0.0041	11.9864	0.00054
$CI * max\text{-}n_i$	$CI4$	L		-0.0021	0.0006	-0.0033	-0.0009	11.8884	0.00056
$CI * comp * max\text{-}n_i$	$CI4$	2	L	0.0015	0.0005	0.0007	0.0024	11.4965	0.00070
$param * max\text{-}n_i$	d_{21}	L		-0.0018	0.0005	-0.0028	-0.0007	10.9141	0.00095

5.3.4 Separation index

In this subsection we briefly discuss the effect on the separatedness of the components of the factorial simulation variables, and for simplicity we focus generally on each simulation variable in isolation with the aid of some simple plots on the next few pages. However in the first instance if we ignore the simulation variables then as expected is clear that there is an inverse relationship between the separation index (SI) and the classification error (CE). The plots show this relationship is non-linear in nature, whereby there is a rapid decline in the CEs as a model changes from not well separated to moderately well separated (SI from 0 to 5), but thereafter for increasing levels of separation there is a far more gradual decline in the CEs. In this respect the plots also show that our attempts at "calibrating" the models were not entirely successful. For example for Model 2 it can be observed that the simulation variables did not elicit highly separated models, whereas for Model 3 the opposite has occurred. In contrast Model 1 has attained both high and low levels of separation across all of the simulation variables. In combination however the data from all three models gives us a good picture of the CE/SI relationship - with the exception of the negative part of the SI index which has not been attained for any of the models. These observations show that a more thorough calibration process must be carried out in order to be more confident in the ranges of the SI that the simulations are likely to produce.

In terms of the simulation variables it is clear that the number of units is one of the variables that most strongly influences the CE/SI relationship. Figure 5.22 shows that larger numbers of units shifts the CE/SI curve to the right. This shows that if two model versions attain the same level of classification errors then the components produced by the version with the larger numbers of units will be more separated than the components produced by the version with the smaller numbers of units. This of course makes sense because of the fundamental role that confidence intervals play in the calculation of the separation index (see subsection 3.4.1), and in turn the strong influence on confidence interval lengths that the numbers of units will have. More importantly it appears as though the CE/SI relationship is the same for both low and high numbers of units, although there is a hint for poorly separated models that the reduction in the CEs with increasing levels of separation is less pronounced when the numbers of units are large compared to when they are small.

In terms of the within-unit sample sizes figure 5.23 shows for both large and small numbers of units that higher within-unit sample sizes leads to higher levels of component separation and thus lower classification errors compared to the lower within-unit sample sizes. This relationship is less clear for Model 2, however from figure 5.19 we see this relationship is indeed clear for low values of the within-unit error covariance parameters. In contrast to these clear relationships for the number of units and the within-unit sample sizes, figure 5.21 shows the unbalancedness of the within-unit sample sizes did not influence the CE/SI relationship.

Figure 5.19 shows for both large and small numbers of units that large values of the within-unit covariance parameters lead to less well separated models and thus higher classification errors than for smaller values of the within-unit covariance parameters. This result is as expected, and this relationship is very strong for Model 2 which also had serially correlated within-unit errors. In this respect figure 5.24 shows that the AR parameters did not themselves influence the CE/SI relationship which suggests this strong effect of the within-unit variances for Model 2 is not to do with the serial dependence in the within-unit errors. In contrast to the within-unit error covariance parameters figure 5.25 shows that the random effects covariance parameters did not influence very much the CE/SI relationship. It is likely higher values of these parameters are required in order to observe any effect.

Finally figure 5.20 shows that the balancedness of the mixing proportions only had an effect in Model 2. Specifically for both large and small numbers of units, and for both low and high values of the within-unit error covariance parameters, unbalanced mixing proportions produced less well separated models and thus higher classification errors. The lack of an effect in the other two models is probably because the unbalanced setting of the factorial variable were not extreme enough.

Figure 5.19: Plots of the average classification error by the average separation index as a function of the within-unit error covariance. In general the red data points should theoretically correspond to lower separation index values and hence higher classification error values than the blue data points. For each plot, and for each data point (each model version) the average classification error is the average taken over all the simulation replications and over all the components, whilst the average Separation Index is the average taken over all the simulation replications and over all the pairwise combinations of components.

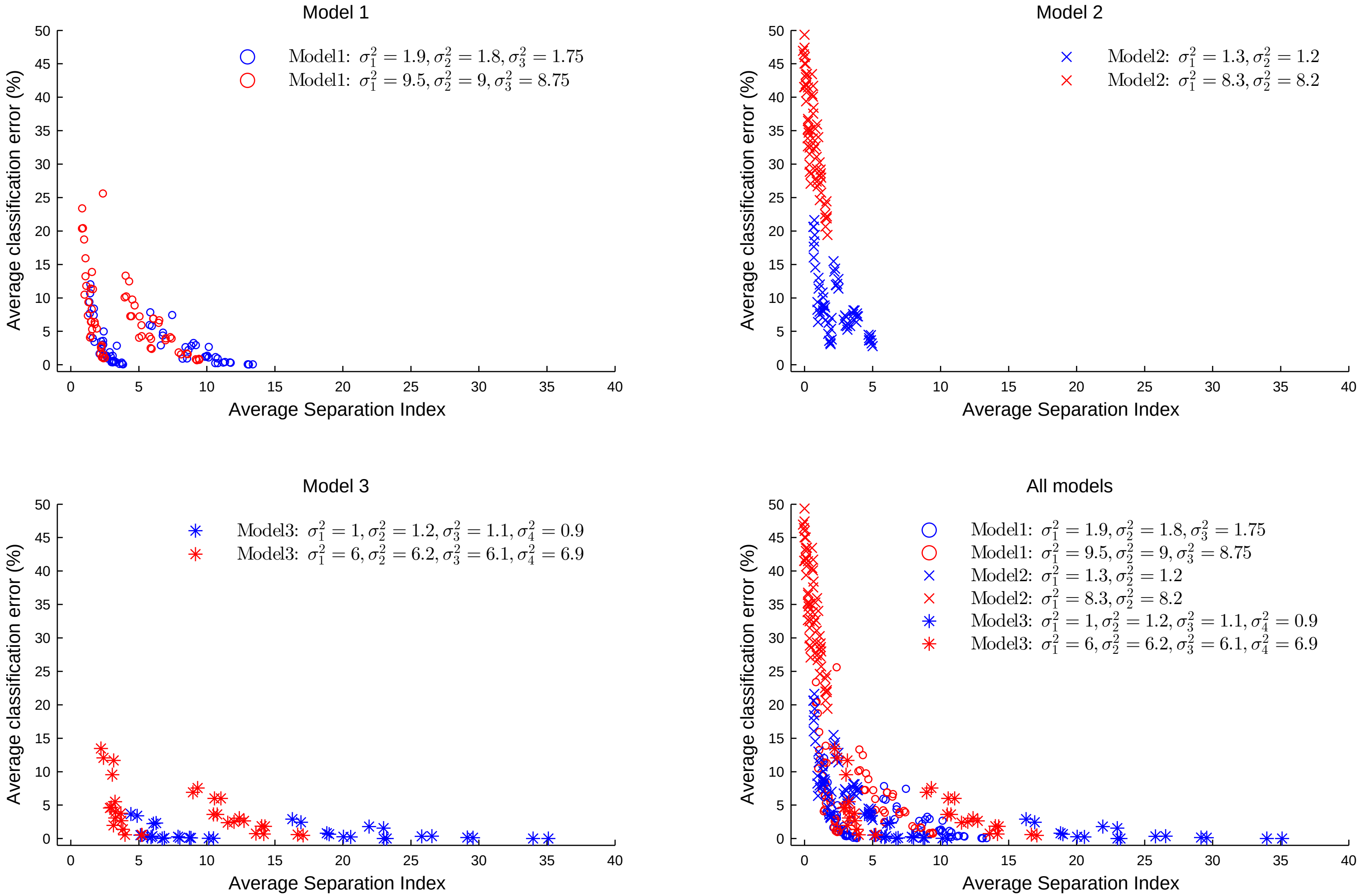


Figure 5.20: Plots of the average classification error by the average separation index as a function of the balancedness of the mixing proportions. In general the red data points should theoretically correspond to lower separation index values and hence higher classification error values than the blue data points. For each plot, and for each data point (each model version) the average classification error is the average taken over all the simulation replications and over all the components, whilst the average Separation Index is the average taken over all the simulation replications and over all the pairwise combinations of components.

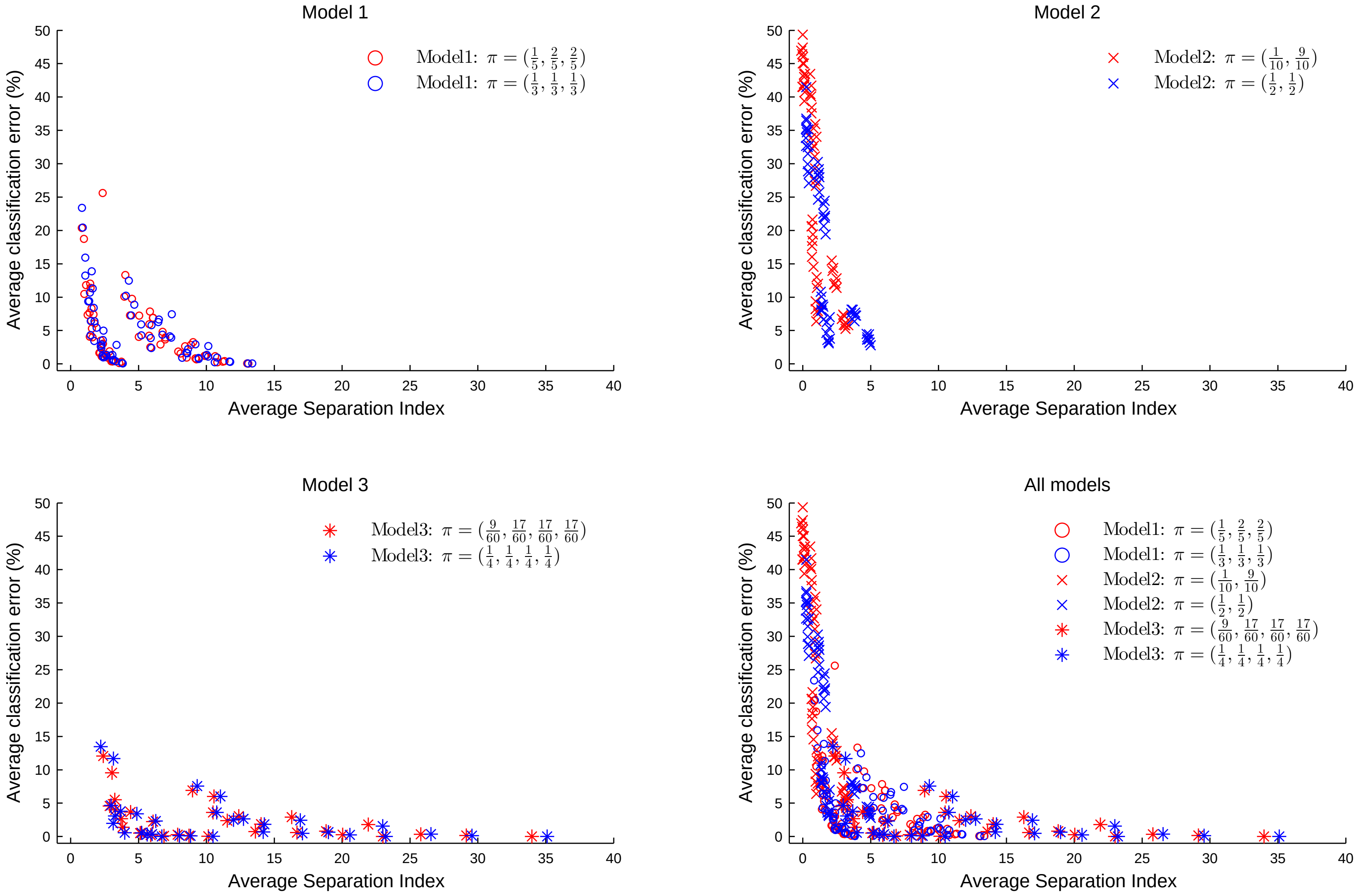


Figure 5.21: Plots of the average classification error by the average separation index as a function of the unbalancedness of the within-unit sample sizes. In general the red data points should theoretically correspond to lower separation index values and hence higher classification error values than the blue data points. For each plot, and for each data point (each model version) the average classification error is the average taken over all the simulation replications and over all the components, whilst the average Separation Index is the average taken over all the simulation replications and over all the pairwise combinations of components.

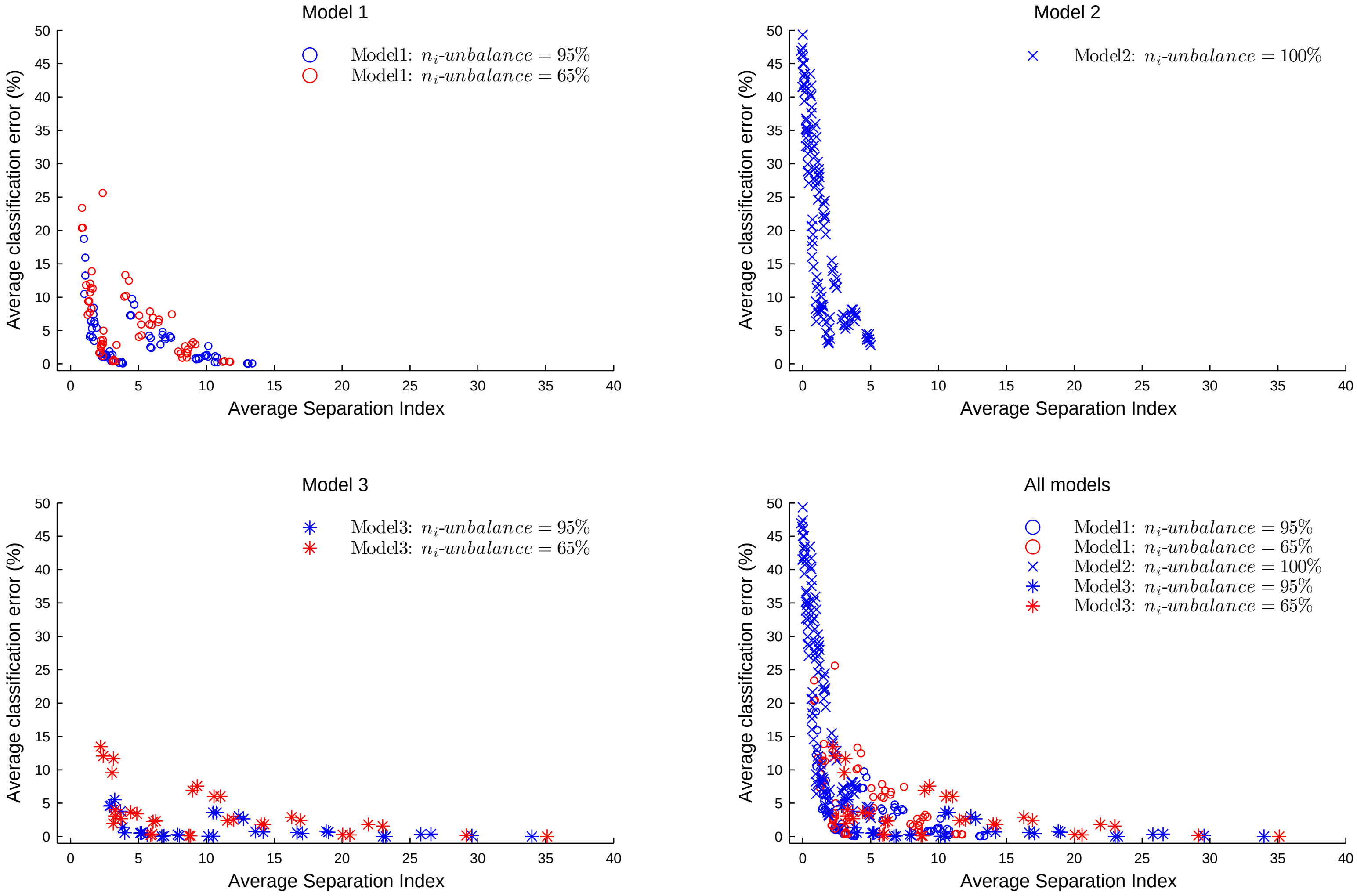


Figure 5.22: Plots of the average classification error by the average separation index as a function of the number of units. In general the red data points should theoretically correspond to lower separation index values and hence higher classification error values than the blue data points. For each plot, and for each data point (each model version) the average classification error is the average taken over all the simulation replications and over all the components, whilst the average Separation Index is the average taken over all the simulation replications and over all the pairwise combinations of components.

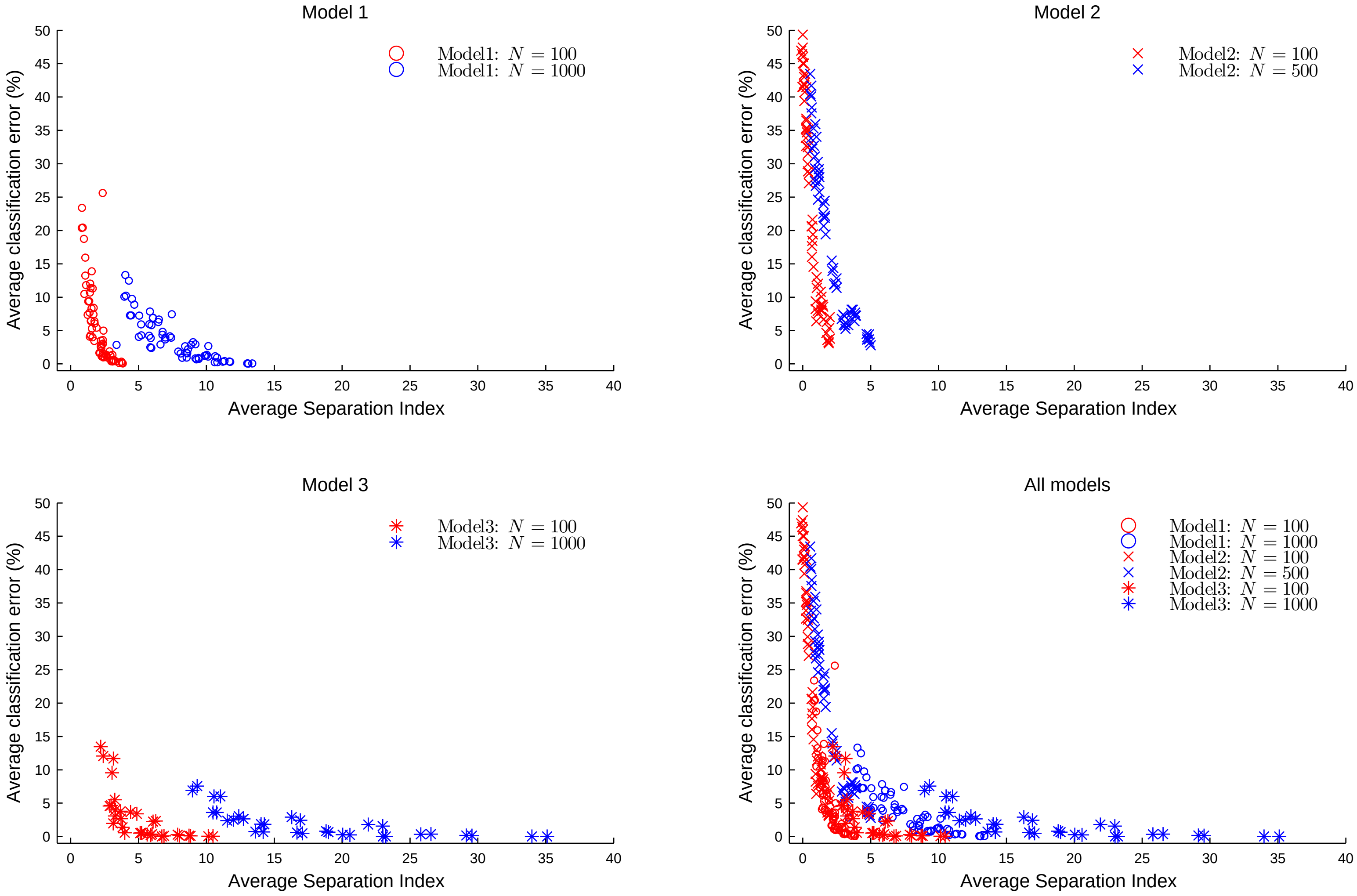


Figure 5.23: Plots of the average classification error by the average separation index as a function of the maximum sample size of the units. In general the red data points should theoretically correspond to lower separation index values and hence higher classification error values than the blue data points. For each plot, and for each data point (each model version) the average classification error is the average taken over all the simulation replications and over all the components, whilst the average Separation Index is the average taken over all the simulation replications and over all the pairwise combinations of components.

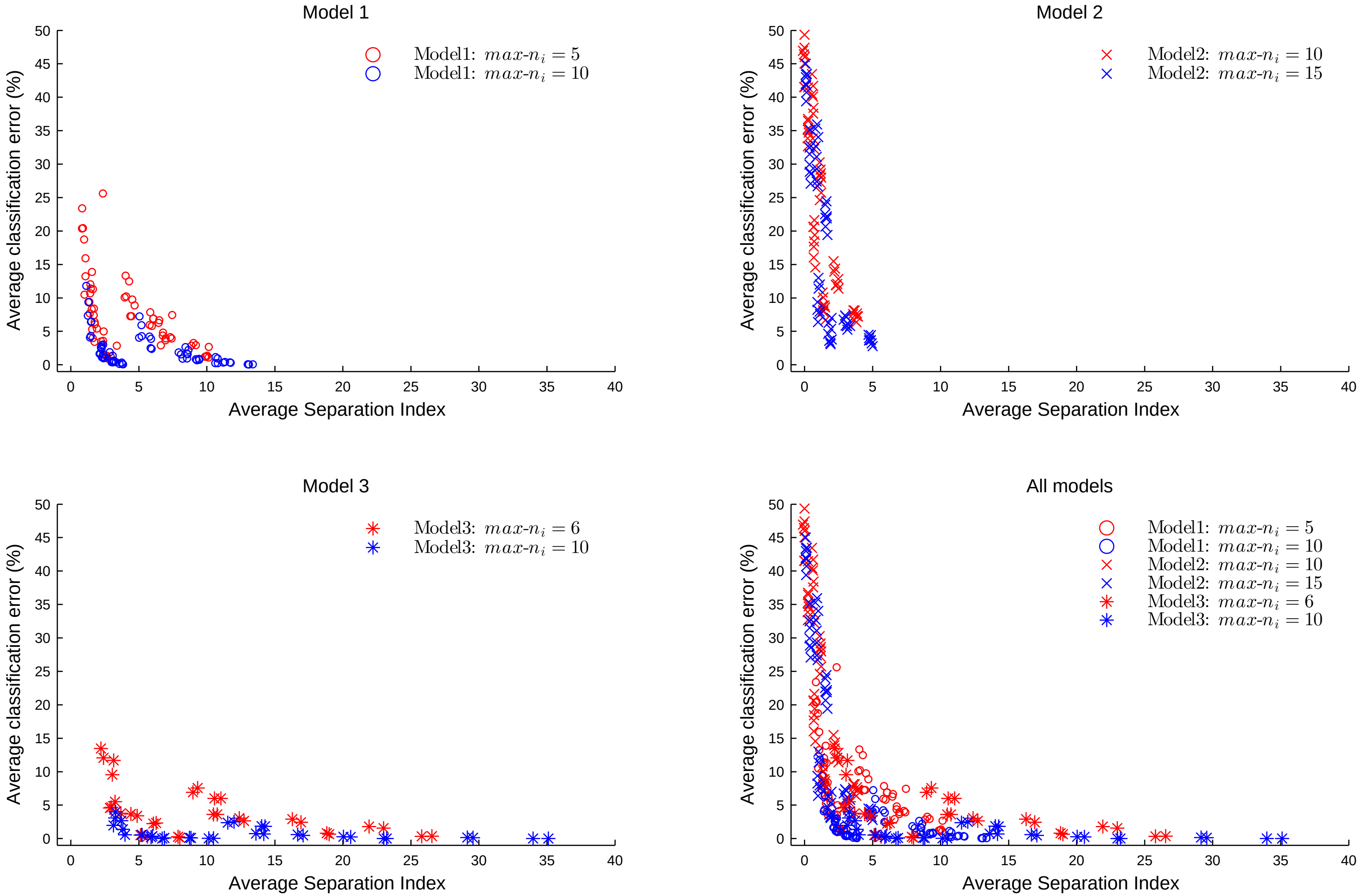


Figure 5.24: Plots of the average classification error by the average separation index as a function of the autocorrelation function of the within-unit errors. In general the red data points should theoretically correspond to lower separation index values and hence higher classification error values than the blue data points. For each plot, and for each data point (each model version) the average classification error is the average taken over all the simulation replications and over all the components, whilst the average Separation Index is the average taken over all the simulation replications and over all the pairwise combinations of components.

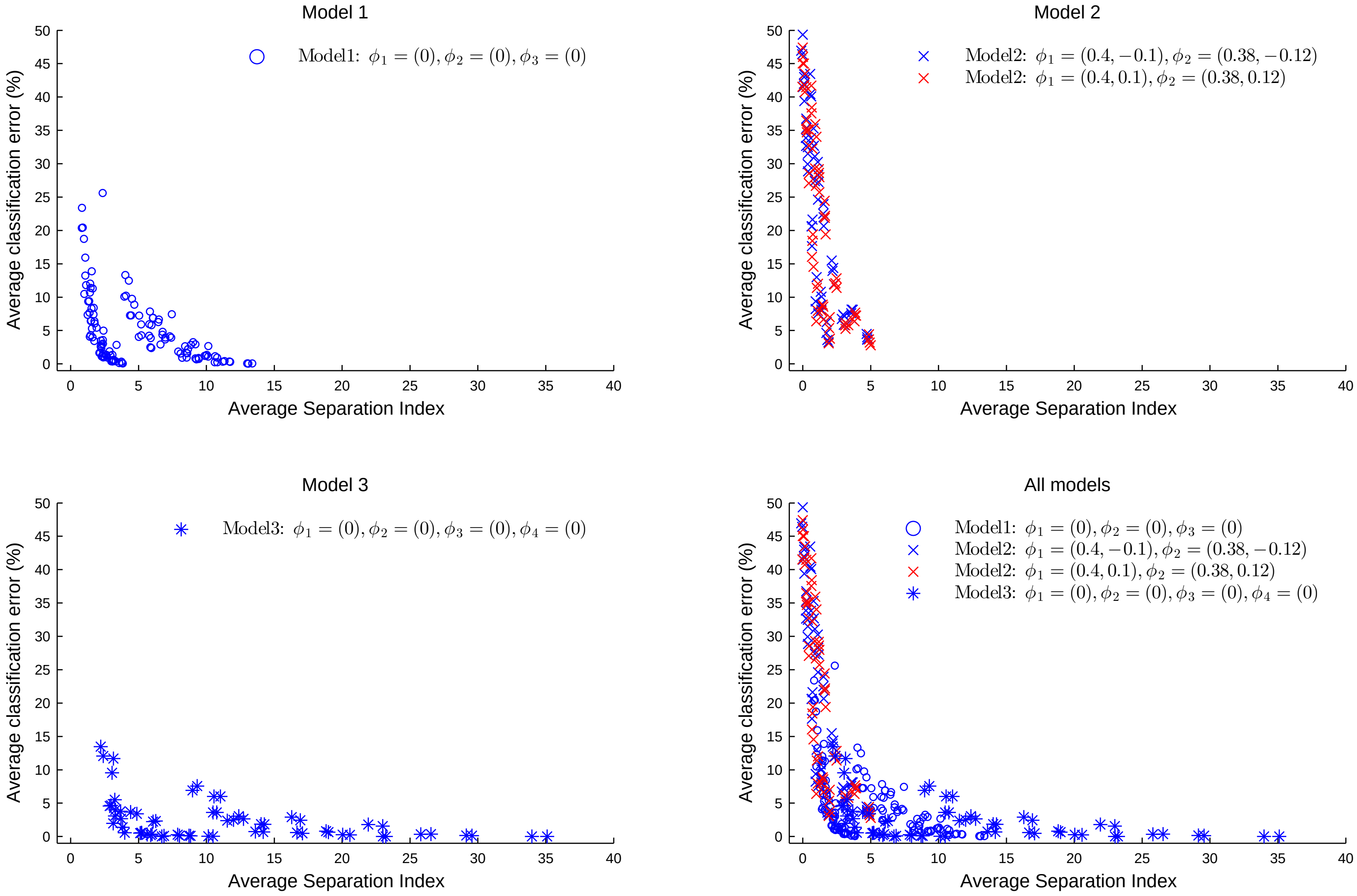
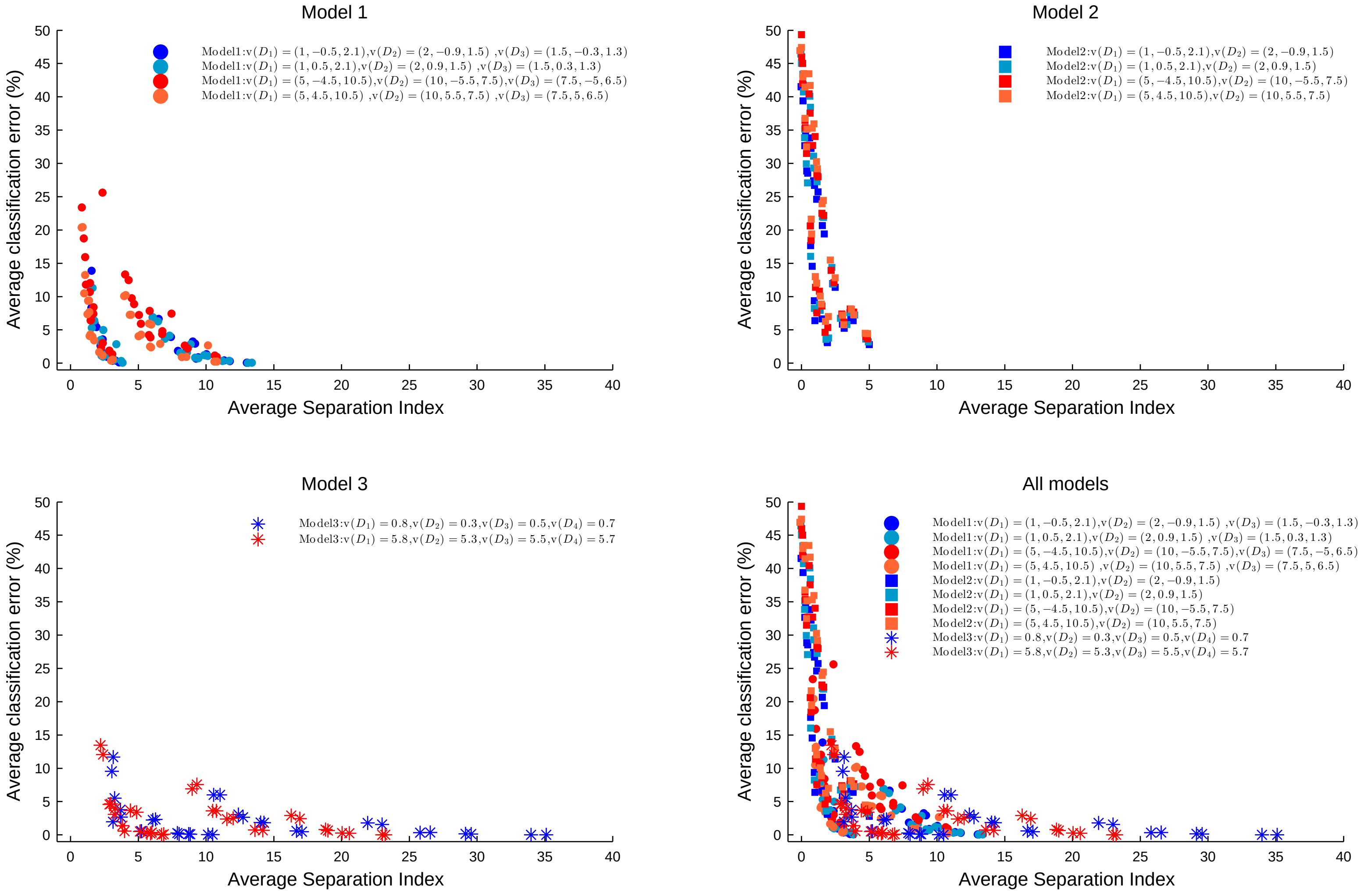


Figure 5.25: Plots of the average classification error by the average separation index as a function of the random effects covariance matrix. In general the red and orange data points should theoretically correspond to lower separation index values and hence higher classification error values than the blue and light blue data points. For each plot, and for each data point (each model version) the average classification error is the average taken over all the simulation replications and over all the components, whilst the average Separation Index is the average taken over all the simulation replications and over all the pairwise combinations of components.



5.4 Summary

For the comparison of the first with the second variants of the EM algorithm one of the main results from section (5.2) was that the first variant of the EM algorithm often converged to poor final parameter estimates when both the within-unit sample sizes were low ($max\text{-}n_i = 6$), and most of the fixed effects covariates were constant within units. This situation is a particularly difficult parameter estimation problem, however it is a problem the second variant of the EM algorithm coped well with. Accordingly for EM1 compared to EM2 the parameter estimates had more variability and were more biased, the confidence interval lengths were occasionally far longer, and the ranges of coverage much lower. Thus this is a major drawback of the first variant compared to the second variant of the EM algorithm. In contrast when the parameter estimation problem was easier there were no large differences between the two EM algorithm variants in terms of the quality of the estimates, coverage probabilities or the confidence interval lengths.

Another main result from section (5.2) was that the score based confidence intervals of $CI1$ produced the best confidence intervals in the sense that ranges of coverage probabilities often intersected the nominal level. In comparison the other three methods produced confidence intervals with similar ranges of coverage that were slightly lower than those of $CI1$, and so intersected the nominal level less often. The fact that the componentwise confidence intervals of $CI3$ performed no worse than those of $CI2$ and $CI4$, and were not much worse than those of $CI1$ is also an important result. The models in section (5.2) were well separated, and so this result represents some evidence in favor of our speculation in subsection (3.4.3) that a well separated model should lead to a comparable level of performance of the componentwise compared to the mixture model confidence intervals.

Finally for the results from section (5.2) we also observed that some of the within-unit variances were poorly estimated by both EM1 and EM2, but in general the estimation was worst for EM2. This poor estimation may have been because the high levels of autocorrelation that are induced by the AR parameters in the within-unit errors in the early time periods might have been captured by the within-unit error variances. If this was occurring then this did not adversely affect the estimation of the autoregressive parameters themselves, which in general were well estimated. In this respect we have already noted this effect was also observed for Model 2 from the factorial simulations. However we also noted this was most severe when the other factorial simulation variables were not set at their default levels. Accordingly we might infer for Model 1 from section (5.2) that the within-unit sample size of $N = 100$ was sufficiently low that high levels of autocorrelation adversely affected estimation of the within-unit variances.

For the three factorial simulations one of the main results was that estimation of all the mixture model parameters was generally good (low MSEs), particularly when the simulation variables σ^2, N, and π-*balance* were set at their optimal levels. Other simulation variables that affected the MSEs were $max\text{-}n_i$, and D, and again these variables increased the MSEs when they were set to their non-default levels. Despite this general result, and as we have just described, there was an adverse affect on the

MSEs but not the CPLIs for the within-unit variances when there were high levels of autocorrelation in the within-unit errors.

In terms of confidence interval performance $CI1$ produced both the highest CPLIs, and ranges of coverage that were closest to the nominal level (often intersecting it), which shows that the superior coverage of $CI1$ was achieved without excessive confidence interval lengths. This confirms the results from section (5.2) that $CI1$ produced the superior confidence intervals. In addition $CI3$ produced similar, and sometimes better CPLIs and CPs compared to $CI2$ and $CI4$. In terms of coverage we also noted this effect for Model 1 and Model 2 from the comparisons of the first and second variants of the EM algorithm. Furthermore $CI3$ produced comparable confidence intervals to these mixture model confidence interval methods even when the simulation variables were set at the non-default levels - i.e. simulation variable levels producing models that were not very well separated. Thus it might be that the different combinations of simulation variable settings all produced models that were sufficiently well separated to prevent the $CI3$ confidence intervals from performing poorly. In this respect we noted in subsection 5.3.4 that the three models were not "calibrated" well enough to produce negative separation indices. Thus it might be that models that produce these negative values have components close enough together to reduce the performance of the $CI3$ confidence intervals.

A very important point to note is that whilst the coverage probabilities were generally good (approximately 80-95%) when the simulation variables were set at the optimal levels, these coverage probabilities often became quite poor (sometimes as low as 40-50%) at the non-optimal levels. Whilst $CI1$ was definitely less prone to producing these low coverage values, we can nonetheless conclude that it is very easy indeed to produce poor confidence intervals from all methods. Since in general the separatedness of the components reduced as the factors change from their optimal to their non-optimal settings, as expected we infer that the quality of inference reduces as component separation reduces. Thus it is very easy to produce poor confidence intervals form all methods when the components are not well separated. The fact that the non-default settings of the simulation variables produced low coverage probabilities explains the many significant effects observed in the robust linear models with the CPLI as the response, and the strongest of these effects were *comp*, σ^2, π-*balance*, N, *max*-n_i, and D (although not all models were affected by all these variables). It is also worth noting that the ACF variable for Model 2 did not feature as a strong effect for the CPLIs, even though it did for the MSEs.

The main conclusion for the practitioner is that the score based confidence interval method should be used, and that this provides good coverage (approximately $80-95\%$) when the estimation problem is easy (simulation variables set at their optimal levels in these simulations). However when the estimation problem is dfficult this method will not give adequate coverage. In this instance a bootstrap procedure might give better results.

6

Data analysis: Quality of life in a lung cancer clinical trial

In this chapter we analyse subject quality of life (QoL) data collected during a clinical trial whose primary aim was to determine the effect on survival for patients with lung cancer who took a particular treatment in conjunction with chemotherapy. The clinical trial was organised and run by Cancer Research UK and University College London Cancer Trials Centre, and the full description and results can be found in Siow et al. (2009). In that paper the trial is referred to as “Study 14”, and so at times we will also use this description.

The statistical analysis of the QoL data from Study 14 employed LMMs, and in this section we will compare the results of LMMs (although not precisely the same models as used in the published study) with two component MLMMs. The aim is to highlight the potential for a MLMM to be a more valid method of analysis than a LMM , although in this respect the use of this particular clinical trial data to achieve this aim is merely speculative, that is to say we have no good reason to believe the LMMs employed in this trial are in any sense invalid. Indeed the use of LMMs for normally distributed repeated measures data such as QoL is common place in medical studies, and so for this type of data in general the validity of the method is beyond question. However there may be specific datasets that appear particularly non-normally distributed and so might be better suited to being analysed with mixture models.

As with any newer and comparatively less established method, the burden of proof that must be carried and successfully discharged in order to prove the claim of superiority (in some sense) must be high, and in this respect we have no such lofty ambitions here. Rather our aim is to highlight some of the difficulties that can be faced when using real datasets in assessing the evidence for and against mixture models when comparing them to the homogeneous model (the one component model). In this respect it is sometimes the case with statistical models that the outcome of such a decision depends on our personal confidence in models that are statistically well justified but that lack the high levels of real world interpretability that we would like. In many situations it is right that our confidence in such models is low, for often it is not enough to have a statistically significant model without it making sense in all respects. By chance the two examples we present in this section illustrate these considerations well rather than showing clear evidence either for or against the use of two component models in favour of LMMs.

6.0.1 Description of the trial

Study 14 took place between June 2003 and September 2005 in 66 centers in the National Cancer Research Institute network. Patients (henceforth subjects) had all been diagnosed with non-small-cell lung cancer (NSCLC) which accounts for around 80% of all the lung cancer deaths wordwide each year. Most patients with NSCLC present with the disease in an advanced state so that surgery or radiotherapy are unsuitable forms of treatment, and so for this reason treatment for these subjects consists of chemotherapy. The proliferatoin of new blood vessels within a tumor, referred to as angiogenesis, is neccessary for tumors to grow, and hence for the cancer to become more severe. In order to combat this process, thalidomide is an oral antiangiogenic agent which has a synergistic activity when combined with cytoxic agents that can be used in chemotherapy, and it is this agent that constituted the treatment in Study 14. Specifically, all subjects (who all had advanced stage NSCLC) were randomised to either a treatment group which consisted of chemotherapy plus thalidomide, or to a placebo group which consisted of chemotherapy plus placebo capsules.

Subjects were randomised to treatment within strata formed by the levels of the factor variables disease stage (*stage*), Eastern Cooperative Oncology Group performance status (*ECOG*), and center (*center*). Subjects underwent repeated cycles of chemotherapy which lasted approximately three weeks each up to a maximum of four cycles. Thus if no delays occurred between cycles, a subject could have up to a maximum of 12 weeks of chemotherapy. Various physical examinations were conducted on the subjects and their QoL data collected via a questionnaire at multiple time points:

before chemotherapy started (baseline measurements); at the start of each chemotherapy cycle; every two months for two years following the end of chemotherapy; and then every three months up to a maximum of 2 years. Thus the maximum time of follow up for a subject who survives for the entire study duration was approximately 4 years.

Because of the severity of NSCLC, and the advanced stage of disease with which subjects present themselves, the mortality rate for NSCLC is high. For Study 14 the median survival time was 8.5 and 8.9 months for the treatment and placebo groups respectively. For overall survival (using just the date of death), which is denoted by (OS), the hazard ratio (HR) of the two treatment groups obtained from a proportional hazards model that adjusts for the variables used in the randomisation was 1.14 with a 95% confidence interval of $0.97 - 1.34$. The HR for progression free survival (PFS), which was calculated using the date of first recurrence of cancer or death, was 1.10 with a 95% confidence interval of $0.95 - 1.28$. The HR of the two treatment groups is the probability of an event for a subject in the treatment group at any time point conditional on that subject having had no event up until that time divided by the same conditional probability for a subject in the placebo group, where the event is death for OS and death or disease recurrence for PFS. These results show that the conditional probability of an event are higher for subjects in the treatment group than in the placebo group, but that this difference is not significant.

The QoL data were obtained using the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire. We will call the overall QoL score for a subject the global QoL index, and this is comprised of multiple sub-indices that measure QoL with respect to function (i.e. physical, emotional, cognitive function), symptoms (i.e. fatigue, pain, insomnia) of which some are lung cancer specific, for example hoarseness, coughing and peripheral neuropathy. For the purposes of this section we will analyse the global and the peripheral neuropathy QoL indices, both of which range from 0 to 100. For the global index 0 and 100 indicate poor and good health respectively, whilst for the peripheral neuropathy index 0 and 100 indicate no symptoms and a high level of symptoms respectively.

In the published paper the QoL data were analysed with a LMM with treatment ($treat$), baseline QoL score ($base$), time of QoL measurement ($tQoL$), and a time of QoL measurement by treatment interaction ($tQoL * treat$) as the fixed effects variables. The $tQoL$ variable was treated as a factor variable, and in doing so no assumption of a linear relationship with time was assumed. A simple random effects and within-unit errors covariance structure was used. Using this model, but without the interaction term, the difference in the mean global QoL scores (as predicted by the model) for treatment minus placebo was -2.1 with a p-value of 0.11. We note that without the interaction this difference in mean scores is just the parameter estimate for treatment, which shows treatment reduced quality of life, although this estimate was not significant. Using this model with the interaction, the differences in mean peripheral neuropathy QoL scores at 12 and 24 weeks into the study were 3.7 and 6.9 respectively, both with p-values less than 0.001. This shows treatment reduced the quality of life with respect to the peripheral neuropathy QoL index (i.e. increased the symptoms) at 12 weeks, and that this adverse affect on QoL almost doubled at 24 weeks. Furthermore these mean score differences were highly significant.

As part of the Study 14 analysis a post hoc analysis was conducted which compared survival in the treatment and placebo groups for subjects with two different types of tumor histology types - squamous and nonsquamous. The two sets of survival curves suggested that treatment might be beneficial for subjects with squamous histology type tumors after approximately 18 months following randomisation. Under the suspicion that this modification of the effect of treatment on survival by tumor histology type might be due to the inclusion of subjects who are unlikely to benefit from chemotherapy (as characterised by subjects whose tumors continued to grow and/or disease continues to progress even after two cycles of chemotherapy), a more thorough post hoc analysis was conducted to investigate this, the results of which have been published in Siow and Hackshaw (2013). We describe this analysis since we believed based on the strength of the results obtained that subgroups may exist within the QoL response variables that could be identified by mixture models.

In general (i.e. for any disease), and under the assumption that criteria exist that can be used to prospectively identify subgroups in the population who stand to benefit more from treatment than other subgroups, then any group of subjects belonging to this subgroup represent an “enriched” patient population with respect to their potential to respond to treatment since the group has not been ”diluted” by subjects whose potential to respond to treatment is low. For this reason we will refer to the post hoc analysis in Siow and Hackshaw (2013) as an ”enriched” version of Study 14, even though it is not a prospective study. Such subgroup targeting as proposed in Siow and

Hackshaw (2013) is becoming more common in very early phase exploratory studies, but where the target subgroups are typically identified through biomarker data, for example such as gene expression levels for genes that are involved in some way in the physiological processes thought or known to be linked to the action of the proposed treatment. This approach is sometimes called translational medicine, and refers to the translation of findings from the laboratory "bench", driven by biomarker data, through to the "bedside" of subjects in clinical trials. The post hoc approach taken by Siow and Hackshaw (2013) is suggested to be suitable for diseases where there are no biomarkers known that predict treatment response.

Even though translational medicine is to some extent a marketing term, this "bench to bedside" approach is important in that it has the potential to lead to the widespread development of "personalised medicine", whereby two different individuals with the same disease may be prescribed different drugs based upon say some genetic difference between the two individuals. Theoretically there is a potential and desirable symbiosis between patients and drugs companies using this approach to drug development, that is if we assume multiple drugs can better serve the needs of a potential patient group than can a single drug, then translational medicine should be welcomed. The symbiosis of course is business related: based on the promise of greater statistical power in enriched study designs, drug companies stand to benefit by reducing their costs by virtue of potentially reducing the failure rate of clinical trials, and/or by being able to run smaller trials.

In the analysis described in Siow and Hackshaw (2013), a variable was defined called “tumor response” with levels “stable disease” (i.e. disease not worsening), “partial response” (partial tumor response), “complete response” (complete tumor response), “progressive disease” (disease worsening), and “not evaluable” (tumor not evaluable for any reason). For each subject their tumor response was determined from data collected after the end of cycle 2 of the chemotherapy. Subjects who had either stable disease, or partial or complete tumor responses were collectively referred to as “nonprogressors”, whilst those with progressive disease were called “progressors”. The histology of the tumors were presumably determined by some sort of biopsy procedure prior to chemotherapy. Whilst separate clinical trials for subjects with these two different histology sub-types are now common (each group appears to benefit from a different combination of cytoxic drugs), as are separate trials for responders (maintenance dose studies), this post hoc analysis combined the two approaches to determine if a statistically significant beneficial effect of treatment could be found for the subgroup squamous/nonprogressors (Study 14 used the cytoxic drugs that appear to benefit squamous subjects). In total the tumor response and histology variables define four subgroups defined by the combinations of squamous/nonsquamous histology, and by progressors/nonprogressors.

The results of this post hoc analysis did indeed show that survival was significantly improved for the squamous/nonprogressor group for PFS only (HR of 0.71 with a p-value of 0.04), but also that survival was significantly worse for the nonsquamous/nonprogressors group for both OS and PFS, and for the nonsquamous group ignoring tumor response for OS and PFS. The conclusion from this study was that patients with squamous type tumors, and whose tumors responded or who had a stable disease after two cycles of chemotherapy may benefit from thalidomide being added to their remaining two chemotherapy cycles, and/or for their maintenance chemotherapy. On account of the strength of these post hoc findings the UCL Cancer Institute has decided to investigate this hypothesis further through prospective randomised controlled trials.

6.0.2 Analysis methods

For the purposes of these examples we chose to use a different LMM to analyse the global QoL and peripheral neuropathy QoL indices than the LMMs used in the published paper, although no attempt was made to compare which model was more suitable since the aim in this section was to compare a one with a two component model of any type. The LMM we used included the same fixed effects variables as the LMM used in Study 14, that is $base$, $treat$, $tQoL$, $tQoL * treat$, but we also included the age and gender of the subjects (age) and ($gender$) respectively. Another difference compared to the Study 14 LMM was that we chose to treat $tQoL$ as a continuous variable expressed in weeks since randomisation occurred. In terms of the covariance parameters we chose an unstructured covariance matrix for the random effects, as well as a simple within-unit error covariance structure. We chose to include two random effects in the LMM - an intercept and $tQoL$, giving a 2×2 random effects covariance matrix.

The treatment variable $treat$ has two levels $treat_1$ which is thalidomide and $treat_2$ which is placebo. We chose $treat_2$ as the reference level, and so only $treat_1$ was estimated. In all that follows by treatment we will mean the $treat_1$ level of the $treat$ variable rather than the whole variable. We also included two of the factors used in the randomisation process - *stage* and *ECOG* - we omitted the third factor *center* since the inclusion of the many parameters caused convergence problems in the mixture model. For global QoL only two centers had a significant estimate in the LMM, and the *center* effect as a whole had a p-value of 0.28, and so this variable was not an important predictor anyway. For the peripheral neuropathy QoL index only one center had a significant estimate which did result in the *center* effect being significant. However *center* was by far the weakest of the variables that were significant, and so omitting this variable probably would not have made any important differences to the results.

The *ECOG* variable was a variable derived from a variable we will call *ECOGraw* which takes on the values 0, 1, or 2, and means the subject had full activity levels, restricted activity levels, or could not carry out work activities respectively. The *ECOG* variable had two levels, $ecog_1$ which meant a subject had either *ECOGraw* scores of 0 or 1, and $ecog_2$ which meant a subject had *ECOGraw* scores of 2. The *stage* variable had two levels $stage_1$ and $stage_2$ which meant the subject had limited and extensive disease respectively. For *ECOG* and *stage* we chose the reference levels to be $ecog_2$ and $stage_2$. We chose the reference level of the *gender* variable to be $gender_2$ which was females.

In order to determine statistically if the two component model is "better" than the one component model we used a likelihood ratio test (LRT). In this context a

LRT tests the null hypothesis that the true parameter for the mixture model (that we assumed generated our observed data) contains zeros for all the parameters in one of the components, and that the mixing proportion for that component is also equal to zero - i.e. the null hypothesis is that the true model is a one component model (sometimes referred to as the homogeneous model). The alternate hypothesis is that the true model is a two component model. The likelihood ratio (LR) statistic is the ratio of the log-likelihood function evaluated at the parameter value that maximises the "restricted" parameter space associated with the null hypothesis (i.e. the parameter space giving rise to one component models) divided by the log-likelihood function evaluated at the parameter value that maximises the "unrestricted" parameter space associated with the alternate hypothesis (i.e. the parameter space giving rise to two component models). Large values of the LR statistic are supposed to constitute evidence in favour of the alternate hypothesis. However for mixture models since the log-likelihood function can under some circumstances theoretically tend to infinity even for parameters far away form the true ones, it cannot be ruled out that a high LR statistic in fact provides no evidence in favour of the alternate hypothesis. However notwithstanding this limitation we use the LRT since it is one of the few tests we have to help us decide whether a two component model fits the data better than a one component model.

For mixture models the null distribution of the LR is unknown, and so a popular method to calculate the p-value of an observed LR statistic is to use a parametric bootstrap procedure. Accordingly we too adopted this method, and this procedure consisted of generating response vectors according to the estimated one component model - i.e. by using the estimated parameters and fixed covariate data but randomly generating within-unit errors and unit level random effects on top of the predicted response. To each newly generated random vector we then estimated both a LMM and a two component model again using the fixed covariate data, and recorded the LR each time. We aimed to repeat this 1000 times, but due to numerical problems estimating the mixture models we in fact only achieved around 800 replications of this procedure. These numerical problems occurred because we were trying to estimate two components when in fact there was only one component in the data. As a result one component would often "degrade" by having fewer and fewer units assigned to the component, and accordingly the parameter estimates would tend to zero. This results in covariance matrices that are approximately zero with matrix determinants that are too small to be stored on a computer - and thus estimation fails.

In terms of the enriched study design described in Siow and Hackshaw (2013), even though the results from this study concerned survival, based on the strength of the results we thought perhaps that some of the four subgroups defined there might appear as subpopulations in the QoL responses we analysed. Accordingly we inspected the estimated components from the two component models to determine if assignment of subjects to these components discriminated between either of these four sub-groups. For example we looked to see if most of the subjects that were in the squamous group were also assigned to a different component than those subjects that were in the non-squamous group. Finally we also used score-based confidence intervals since the results of Chapter 5 showed these were the best intervals of the four types presented in this thesis.

6.0.3 Anaysis results

We first discuss the results from the analysis of the global QoL index, which are presented in table 6.1. With t-statistics of 17.296, 5.192, 2.422 and 2.051, we see in order of strength from highest to lowest that the estimates of $base$, $tQoL$, $treat_1$, $treat_1 * tQoL$ were significant at the 95% confidence level. Remembering that higher values of this index indicates better health, the negative $tQoL$ effect shows that QoL decreased for both treatment groups as time increased, but the positive $treat_1 * tQoL$ effect shows this rate of decline was less pronounced for the treatment compared to the placebo group. However the negative $treat_1$ effect shows the average QoL was lower for the treatment compared to the placebo group. The positive $base$ effect and its strength are consistent with what we would expect, that is the higher a patients' baseline QoL score is, the higher their overall QoL score is over all the time points. The estimates of d_{11}, d_{22} and σ^2, with t-statistics of 9.75, 2.57 and 31.05 respectively, were all significant, highly so for the estimate of σ^2.

These results show that the beneficial effects of treatment in terms of reducing the rate at which QoL decreases with time when having chemotherapy treatment are outweighed by the fact that treatment also reduces the overall level of QoL compared to placebo. The results also show statistically significant levels of random heterogeneity in QoL both between and within the units.

The two component model has a highly significant likelihood ratio statistic from the bootstrap procedure which is strong evidence in favour of the hypothesis that there are two components in the data rather than just one. The estimates of the mixing proportions show that approximately half of the subjects are in one component whilst the other half are in the other component. With t-statistics given in the order (component 1,component 2), the t-statistics for the estimates of the *base*, $ecog_1$, and $tQoL$ parameters were $(19.67, 4.372)$, $(2.494, 2.891)$, and $(2.64, 3.379)$ respectively. All three of these estimates were significant at the 95% confidence level, and furthermore the confidence intervals for the *base* and $ecog_1$ estimates between the components do not intersect. Thus whilst the effect of time since randomisation is not really different between the components (again negative and of a similar magnitude as in the 1-component model), the effects of baseline QoL and ECOG are very different between the components. Furthermore the treatment effect for the one component model is no longer significant in the two component model.

The estimates of the *base* parameters show that increases in baseline QoL for subjects in component 1 increased average levels of QoL more than they did for subjects in component 2. Whilst these estimates make sense, the estimates of the $ecog_1$ parameters do not: for component 1 the estimate of β^{ecog_1}, by being negative, suggests subjects whose disease severely restricts their activity levels have a higher QoL than those subjects whose disease restricts their activities much less severely. One possible explanation for this that seems likely is that whilst this two component model has been found to give a higher log-likelihood than the one component model, this does not guarantee that any interpretation can be attached to the components.

For example we might imagine the sample comes from a distribution with longer tails than the normal, and thus that there is a violation of model assumptions. We might then fit the distribution better with two normal distributions - one that has a small variance and thus a sharp peak so that it fits the sample points well that are in the range that contains the modal frequency of the data, and another distribution with a large variance and thus long tails so that it fits the sample points well that are far away from the range that contains the modal frequency. Since these two distributions will overlap to a large extent then the assignment of sample points to one of the two distributions around the range containing the modal frequency will be arbitrary, and thus no clear interpretation of the components will be possible. In this way it is eminently possible that the estimated components of a mixture model which give rise to a statistically significant LR statistic may be mathematical rather than real world entities to which a sensible interpretation can be attached.

For this analysis, evidence in favour of "non-interpretable" components can be found by inspecting figure 6.3. This shows considerable overlap between the two components as manifested by half of the subjects having posterior probabilities for both components that are in the range $[0.2, 0.8]$. The distribution for global QoL did not suggest a long tailed distribution, however it was not a smooth normal distribution either on account of the data being reasonably discrete.

Further evidence in favour of "non-interpretable" components can be found by inspecting table 6.3 which shows no large differences between the components with respect to the covariates in the model, although subjects in component 1 on average have a higher global QoL than subjects in component 2. The biggest covariate difference is with respect to the tQoL variable with subjects in component 2 having their quality of life measurements taken on average three weeks later than those subjects in component 1. However this only means that subjects in component 2 had on average approximately half an extra observation than those subjects in component 1. Although large differences between components with respect to covariates in the model is not an assumption of mixture models, we might observe such differences if for example there is a non-linear relationship between a covariate and the response - i.e. two linear parameters covering different parts of the covariate range will often be better than a single linear parameter.

In terms of differences between the components with respect to variables not in the model, the percentage of subjects in component 1 that survived or were censored was twice as large as the same percentage of subjects in component 2. Furthermore the time to event (alive/censored or death) was on average ten weeks shorter for subjects in component 1 than those in component 2. Since deaths make up the majority of these events in both components, we can say that on average time to death was ten weeks shorter for subjects in component 1 than in component 2 even though more subjects in component 1 survived than in component 2. Thus it might be that the disease progressed faster or was more severe for those subjects in component 1 that died compared to those subjects in component 2 that died. Furthermore there were also no differences between the two components in terms of either tumor response, or histology type of the tumor.

Despite these small differences between the components in terms of survival, the lack of component separation strongly suggests the components cannot be interpreted. In this respect we also note that the estimate of the within-unit variance of component 1 is far lower than the estimate of the within-unit variance for component 2. The reason for this difference is not immediately clear when looking at figure 6.1 since this plot is rather busy. Notwithstanding this, for component 1 it can be determined in the time range where most of the QoL measurements fall, that is $0 - 25$ weeks, that the bulk of the QoL measurements fall in the range $35 - 85$ whereas for component 2 the bulk of the QoL measurements in the same time range fall in the range $25 - 85$. Furthermore for time values greater than 25 weeks it is clear the range of the QoL measurements for component 1 are less than that for component 2. Thus it may be that the two estimated components are characterised by low (component 1) and high (component 2) levels of within-unit variation but that no other interpretation can be applied to them in terms of either the covariates in the model or known external variables not in the model.

The obvious question that remains unanswered is whether to focus on the results of the one component model, which makes sense (i.e. the $ecog_1$ parameter estimate is positive), or to focus on the results of the two component model that are mathematically superior in terms of providing a statistically significant LR statistic, but where the results lack a satisfactory interpretation. In this particular instance if we make this choice then we also choose whether to regard the negative effect of treatment on QoL as statistically significant or not.

Turning our attention now to the results for the peripheral neuropathy QoL index, from table 6.2 we see that estimates of the *base* and $treat_1 * tQoL$ parameters, with t-statistics of 10.479 and 3.457 respectively, are statistically significant. Recalling for this index that 100 represents lots of symptoms and 0 no symptoms, then just as for the global QoL index wee see that improvements (i.e. reductions) in the baseline QoL are associated with improvements in the QoL index. Similarly the interaction estimate and the estimate for *tQoL* together show that whilst QoL reduces over time for both treatment and placebo groups, QoL reduces faster for the treatment than for the placebo group. Again as for the global QoL index we have statistically significant parameter estimates for all of the covariance parameters.

The LR statistic for the two component model is again highly significant, where it is clear that none of the covariates have any affect on QoL for those subjects belonging to component 2, indeed even the intercept has been estimated to be close to zero. The only significant parameter estimate is for the within-unit variance which is very small compared to the same estimate for component 1. A glance at figure 6.2 shows why this is so: most subjects in component 2 have zero QoL scores for all time points. Thus almost all of these subjects either had no symptoms of peripheral neuropathy throughout the entire course of their treatment, or they were answering the questionnaires without due care and attention - i.e. by putting zero down for all time points. The estimates of the mixing proportions show that these subjects accounted for a substantial number of the total number of units (approximately 38%).

In contrast the subjects in component 1 have a pattern of QoL measurements that we might expect. For this component, and with t-statistics of 9.891, 3.559, and 2.67, wee see that the parameter estimates of $base$, $ecog_1$, and $treat_1 * tQoL$ are all significant. Again increases of baseline QoL are associated with increases in QoL, and QoL reduces over time faster in the treatment than in the placebo group, and again all of the covariance parameter estimates are significant. In contrast to the two component model for the global QoL index, the $ecog_1$ parameter estimate, by being negative, shows that those subjects whose disease meant they were most severely restricted in their activities had a worse QoL score than those subjects whose activity levels were less restricted.

Thus there is at least one plausible interpretation of these two estimated components: subjects in component 2 did not suffer any symptoms of peripheral neuropathy at all and thus yield no information on covariate-response relationships, or else they provided consistent and invalid questionnaire responses, whilst subjects in component 1 were "normal" in the reverse sense, that is experiencing symptoms and providing valid questionnaire responses, and thus providing valid information on covariate-response relationships. Table 6.4 confirms that the only real difference between the two components with respect to known variables concerns the response itself. In particular we again see that there were no differences between the two components in terms of either tumor response, or histology type of the tumor.

In contrast to the global QoL model, the two components are fairly well separated - this can be seen in figure 6.3. This suggests an arbitrary assignment of units to components is not being made where two normal distributions have been fitted to one non-normal distribution, as we described previously. Thus if we accept that the one component model has been incorrectly influenced by these ”no symptom” subjects, then the two component model has revealed that ECOG does in fact significantly affect QoL. The key question of course is the validity in treating the component 2 subjects separately - i.e. by fitting a two component model or by removing them before fitting a one component model. As with all of these types of decisions it is of paramount importance to at least ensure the lack of symptoms is not to do with the treatment itself, however table 6.4 shows this is probably not the case. Thus one could justifiably suggest that these subjects simply add noise to the data since they stand no chance of contributing valid information for determining either the efficacy of the treatment on the response, or any covariate-response relationships.

Table 6.1: Comparison of the homogeneous model with a two component mixture model for the global QoL index. The p-value for the likelihood ratio statistic comes from 713 replications of a parametric bootstrap procedure using the parameter estimates from the homogeneous model. A "$*$" signifies that the parameter estimate is significant at the 95% confidence level, whilst a "$**$" signifies not only that the parameter estimate is significant, but also that the confidence intervals from both components for this parameter do not intersect. The confidence intervals for the mixture model use standard errors based on the score vector approximation to the mixture model information matrix.

	LMM					**MLMM**									
LL	-11125.989					-10952.928									
LR						1.016									
$P[LR > lr]$						< 0.01									
	Component 1					**Component 1**					**Component 2**				
Parameter	Estimate	StdErr	LowerCL	UpperCL		Estimate	StdErr	LowerCL	UpperCL		Estimate	StdErr	LowerCL	UpperCL	
β^0	38.609	4.696	29.405	47.814	*	35.272	6.887	21.774	48.770	*	35.557	6.722	22.383	48.732	*
β^{base}	0.467	0.027	0.414	0.520	*	0.728	0.037	0.654	0.801	**	0.188	0.043	0.104	0.273	**
β^{age}	-0.099	0.067	-0.230	0.032		-0.099	0.095	-0.284	0.086		0.041	0.103	-0.160	0.242	
β^{gender_1}	0.696	1.211	-1.677	3.068		-0.782	1.879	-4.464	2.900		2.375	1.801	-1.155	5.905	
β^{treat_1}	-3.028	1.258	-5.493	-0.563	*	-1.365	1.771	-4.836	2.106		-3.326	2.199	-7.637	0.985	
β^{ecog_1}	1.144	2.075	-2.922	5.211		-8.151	3.268	-14.555	-1.746	**	7.997	2.766	2.575	13.419	**
β^{stage_1}	-1.543	1.159	-3.815	0.728		-2.494	1.672	-5.771	0.784		-0.138	1.837	-3.737	3.462	
β^{tQoL}	-0.135	0.026	-0.185	-0.084	*	-0.132	0.050	-0.230	-0.035	*	-0.098	0.029	-0.156	-0.041	*
$\beta^{treat_1 * tqol}$	0.080	0.039	0.002	0.157	*	0.060	0.067	-0.070	0.191		0.062	0.070	-0.075	0.199	
d_{11}	133.274	13.670	106.482	160.067	*	104.816	15.622	74.196	135.435	*	80.257	24.146	32.932	127.583	*
d_{21}	-0.058	0.290	-0.627	0.511		-0.222	0.421	-1.047	0.603		-0.440	0.432	-1.286	0.407	
d_{22}	0.018	0.007	0.005	0.032	*	0.063	0.017	0.030	0.095	**	0.002	0.008	-0.013	0.018	
σ^2	227.971	7.341	213.582	242.359	*	69.103	4.561	60.163	78.044	**	380.643	16.670	347.971	413.315	**
π						0.492	0.033	0.427	0.557	*	0.508	0.033	0.444	0.573	*

Table 6.2: Comparison of the homogeneous model with a two component mixture model for the peripheral neuropathy QoL index. The p-value for the likelihood ratio statistic comes from 713 replications of a parametric bootstrap procedure using the parameter estimates from the homogeneous model. A "$*$" signifies that the parameter estimate is significant at the 95% confidence level, whilst a "$**$" signifies not only that the parameter estimate is significant, but also that the confidence intervals from both components for this parameter do not intersect. The confidence intervals for the mixture model use standard errors based on the score vector approximation to the mixture model information matrix.

	LMM					**MLMM**									
LL	-10914.111					-9784.701									
LR						1.115									
$P[LR > lr]$						< 0.01									
	Component 1					**Component 1**					**Component 2**				
Parameter	Estimate	StdErr	LowerCL	UpperCL		Estimate	StdErr	LowerCL	UpperCL		Estimate	StdErr	LowerCL	UpperCL	
β^0	9.275	4.684	0.094	18.455	*	15.230	6.603	2.288	28.172	*	-0.299	6.104	-12.262	11.664	
β^{base}	0.503	0.048	0.409	0.598	*	0.455	0.046	0.366	0.545	**	0.009	0.044	-0.077	0.094	
β^{age}	0.004	0.069	-0.131	0.139		0.085	0.101	-0.112	0.283		0.001	0.063	-0.122	0.125	
β^{gender_1}	0.035	1.246	-2.408	2.478		0.397	1.802	-3.136	3.929		0.113	1.804	-3.424	3.649	
β^{treat_1}	2.268	1.334	-0.345	4.882		1.822	2.007	-2.112	5.756		0.102	2.531	-4.858	5.062	
β^{ecog_1}	-2.949	2.036	-6.939	1.041		-7.814	2.189	-12.104	-3.524	*	0.136	4.827	-9.326	9.597	
β^{stage_1}	0.198	1.189	-2.133	2.529		-0.544	1.621	-3.721	2.633		-0.056	1.750	-3.485	3.373	
β^{tqol}	0.076	0.041	-0.004	0.156		0.105	0.070	-0.033	0.243		0.012	0.071	-0.128	0.152	
$\beta^{treat_1 * tqol}$	0.204	0.059	0.089	0.319	*	0.222	0.083	0.059	0.386	*	-0.010	0.133	-0.271	0.251	
d_{11}	176.627	15.468	146.310	206.944	*	186.061	20.358	146.159	225.962	**	0.190	3.909	-7.471	7.851	
d_{21}	-1.403	0.491	-2.365	-0.440	*	-2.497	0.666	-3.803	-1.191	**	-0.015	0.054	-0.122	0.091	
d_{22}	0.142	0.020	0.101	0.182	*	0.153	0.029	0.096	0.211	**	0.001	0.002	-0.004	0.006	
σ^2	161.316	5.376	150.779	171.853	*	238.101	7.321	223.752	252.450	**	10.172	3.270	3.763	16.581	*
π						0.615	0.029	0.559	0.672	**	0.385	0.029	0.328	0.441	**

Table 6.3: Characteristics of the estimated components from the mixture model fitted to the Global QoL index. From 722 subjects, 92 subjects were removed due to having only a baseline observation, and 13 subjects were removed due to all of their non-baseline observations containing one or more missing values for either the Global QoL index or the covariates in the model, thus 105 subjects were removed in total.

		Component 1 (N=335)		**Component 2 (N=282)**	
Variable	**Level**	**Count**	**Percentage**	**Count**	**Percentage**
Gender	**Male**	229	68.36	165	58.51
	Female	106	31.64	117	41.49
ECOG	**Full or restricted activity levels**	310	92.54	249	88.30
	Cannot do work activities	25	7.46	33	11.70
Treatment arm	**Treatment**	182	54.33	139	49.29
	Placebo	153	45.67	143	50.71
ECOG raw	**Full activity levels**	98	29.25	97	34.40
	Restricted activity levels	212	63.28	152	53.90
	Cannot do work activities	25	7.46	33	11.70
Stage	**IIIb (disease stage limited)**	152	45.37	124	43.97
	IV (disease stage extensive)	183	54.63	158	56.03
Tumor response	**stable disease**	151	45.07	126	44.68
	partial response	79	23.58	68	24.11
	complete response	2	0.60	1	0.35
	progressive disease	5	1.49	6	2.13
	not evaluable	98	29.25	81	28.72
Histology	**Squamous**	116	34.63	87	30.85
	Non-Squamous	219	65.37	195	69.15
Survival	**Alive/censored**	36	10.75	16	5.67
	Dead	299	89.25	266	94.33
		Mean	**Stdv**	**Mean**	**Stdv**
Global QoL		65.35	19.85	55.74	21.84
Baseline global QoL		64.38	21.11	63.62	20.89
Age (years)		60.94	8.77	61.44	8.34
Number of QoL measurements		4.10	2.17	4.58	2.74
tQoL (weeks)		16.62	19.54	20.72	26.12
Time to event (weeks)		65.02	42.78	74.71	49.91

Table 6.4: Characteristics of the estimated components from the mixture model fitted to the peripheral neuropathy QoL index. From 722 subjects, 92 subjects were removed due to having only a baseline observation, and 5 subjects were removed due to all of their non-baseline observations containing one or more missing values for either the peripheral neuropathy QoL index or the covariates in the model, thus 97 subjects were removed in total.

		Component 1 (N=378)		**Component 2 (N=247)**	
Variable	**Level**	**Count**	**Percentage**	**Count**	**Percentage**
Gender	**Male**	242	64.02	160	64.78
	Female	136	35.98	87	35.22
ECOG	**Full or restricted activity levels**	343	90.74	221	89.47
	Cannot do work activities	35	9.26	26	10.53
Treatment arm	**Treatment**	207	54.76	117	47.37
	Placebo	171	45.24	130	52.63
ECOG raw	**Full activity levels**	124	32.80	74	29.96
	Restricted activity levels	219	57.94	147	59.51
	Cannot do work activities	35	9.26	26	10.53
Stage	**IIIb (disease stage limited)**	174	46.03	110	44.53
	IV (disease stage extensive)	204	53.97	137	55.47
Tumor response	**stable disease**	180	47.62	97	39.27
	partial response	93	24.60	56	22.67
	complete response	1	0.26	2	0.81
	progressive disease	8	2.12	3	1.21
	not evaluable	96	25.40	89	36.03
Histology	**Squamous**	126	33.33	81	32.79
	Non-squamous	252	66.67	166	67.21
Survival	**Alive/censored**	31	8.20	22	8.91
	Dead	347	91.80	225	91.09
		Mean	**Stdv**	**Mean**	**Stdv**
Peripheral Neuropathy QoL		20.75	22.18	0.11	1.37
Baseline peripheral Neuropathy QoL		6.26	12.82	1.58	7.05
Age (years)		61.07	8.69	61.72	8.19
Number of QoL measurements		4.72	2.46	3.58	2.28
tQoL (weeks)		19.74	24.15	15.81	19.66
Time to event (weeks)		71.94	46.28	64.28	46.30

Figure 6.1: Global QoL for component 1 (top plot) and component 2 (bottom plot) for each subject (each line is a subject) with the average QoL ±1 standard deviation calculated for the time ranges $[0, 5)$, $[5, 10)$, $[10, 15)$, $[15, 20)$, $[20, 30)$, $[30, 40)$, $[40, 50)$, $[50, 75)$, $[75, 100)$, and $[100, 150)$ weeks, and plotted at the range mid-points 2.5, 7.5, 12.5, 17.5, 25, 35, 45, 62.5, 87.5, 125, 175 and 225 weeks.

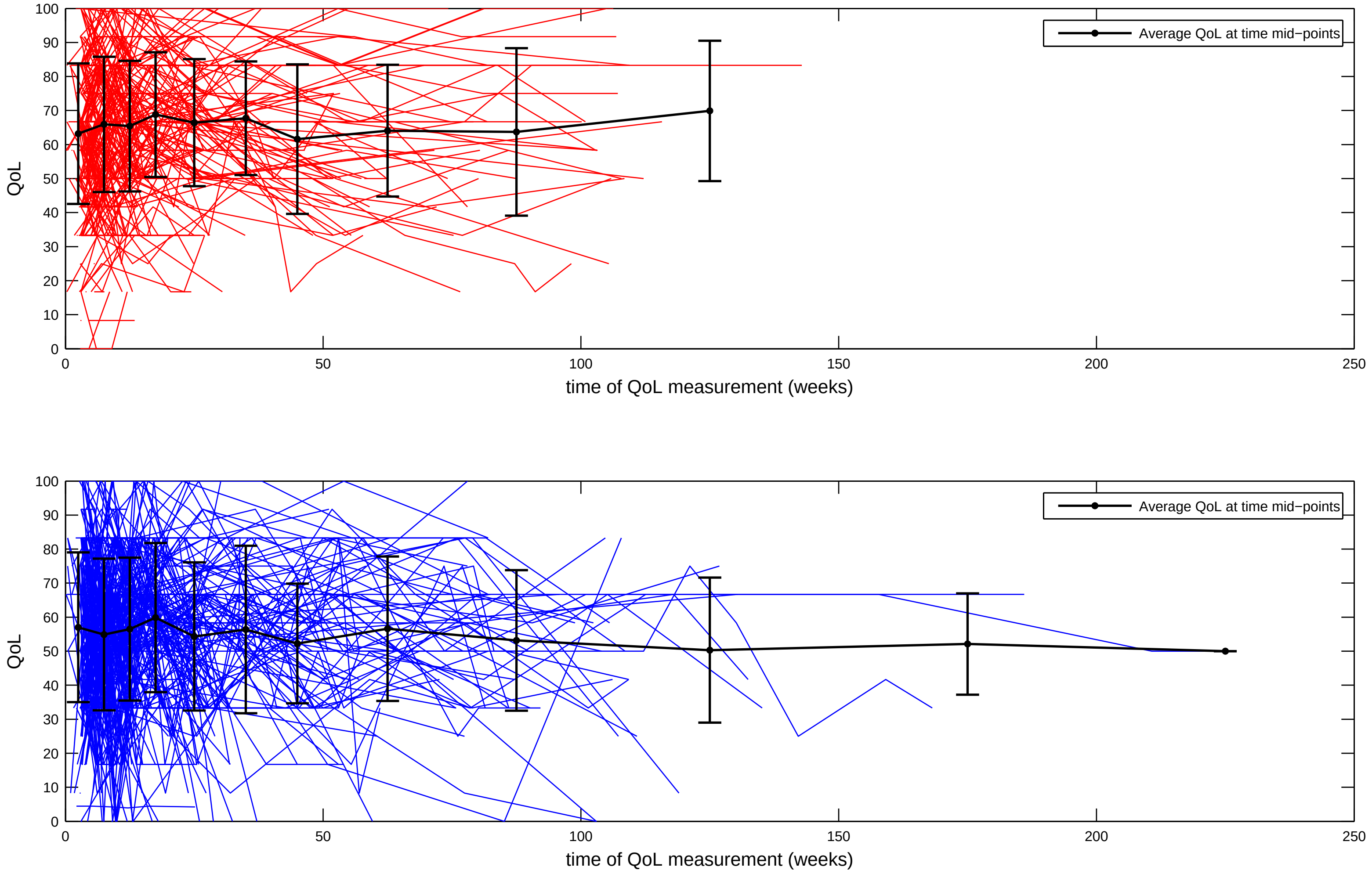


Figure 6.2: Peripheral neuropathy QoL for component 1 (top plot) and component 2 (bottom plot) for each subject (each line is a subject) with the average QoL ±1 standard deviation calculated for the time ranges $[0, 5)$, $[5, 10)$, $[10, 15)$, $[15, 20)$, $[20, 30)$, $[30, 40)$, $[40, 50)$, $[50, 75)$, $[75, 100)$, and $[100, 150)$ weeks, and plotted at the range mid-points 2.5, 7.5, 12.5, 17.5, 25, 35, 45, 62.5, 87.5, 125, 175 and 225 weeks.

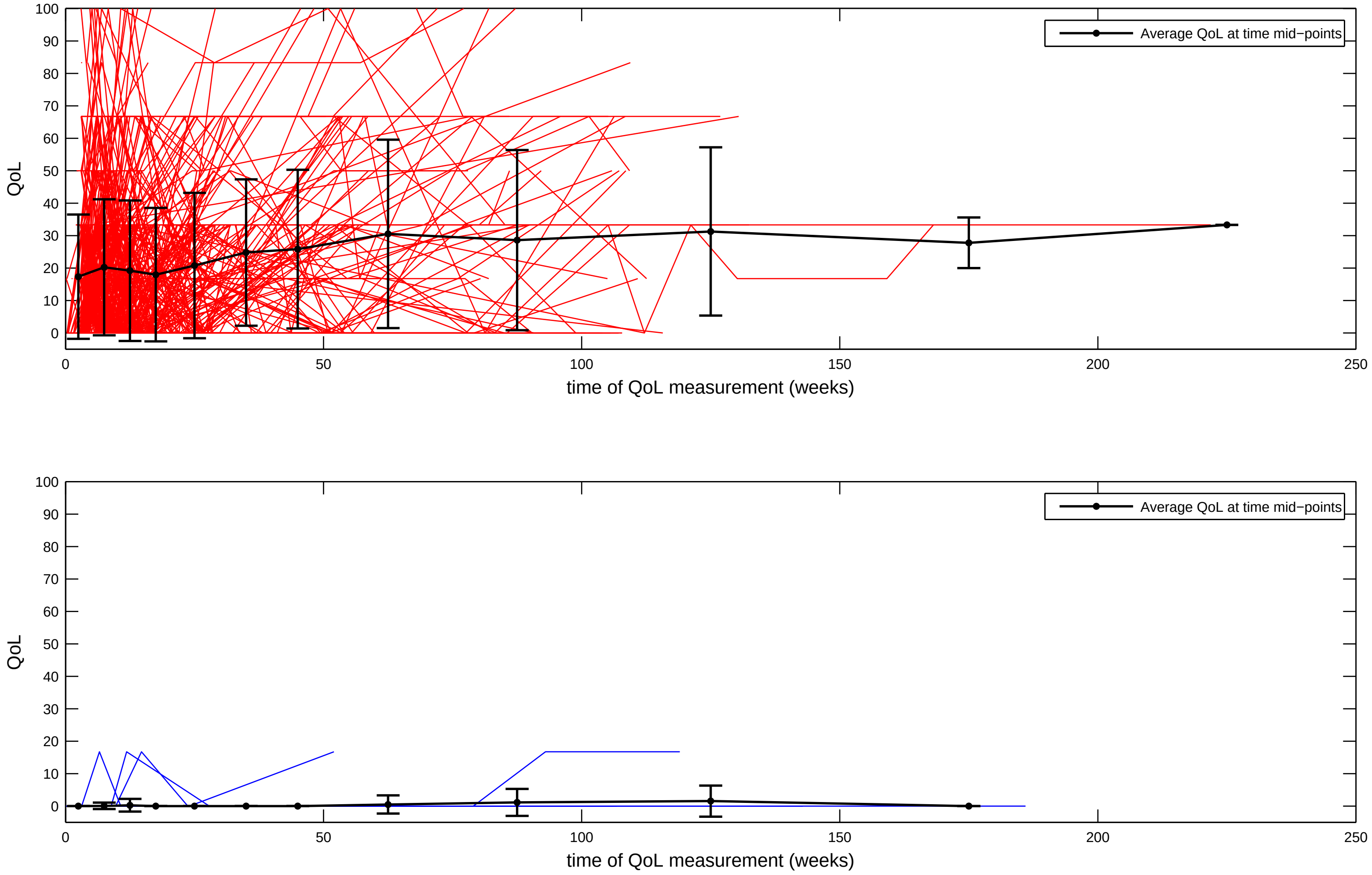
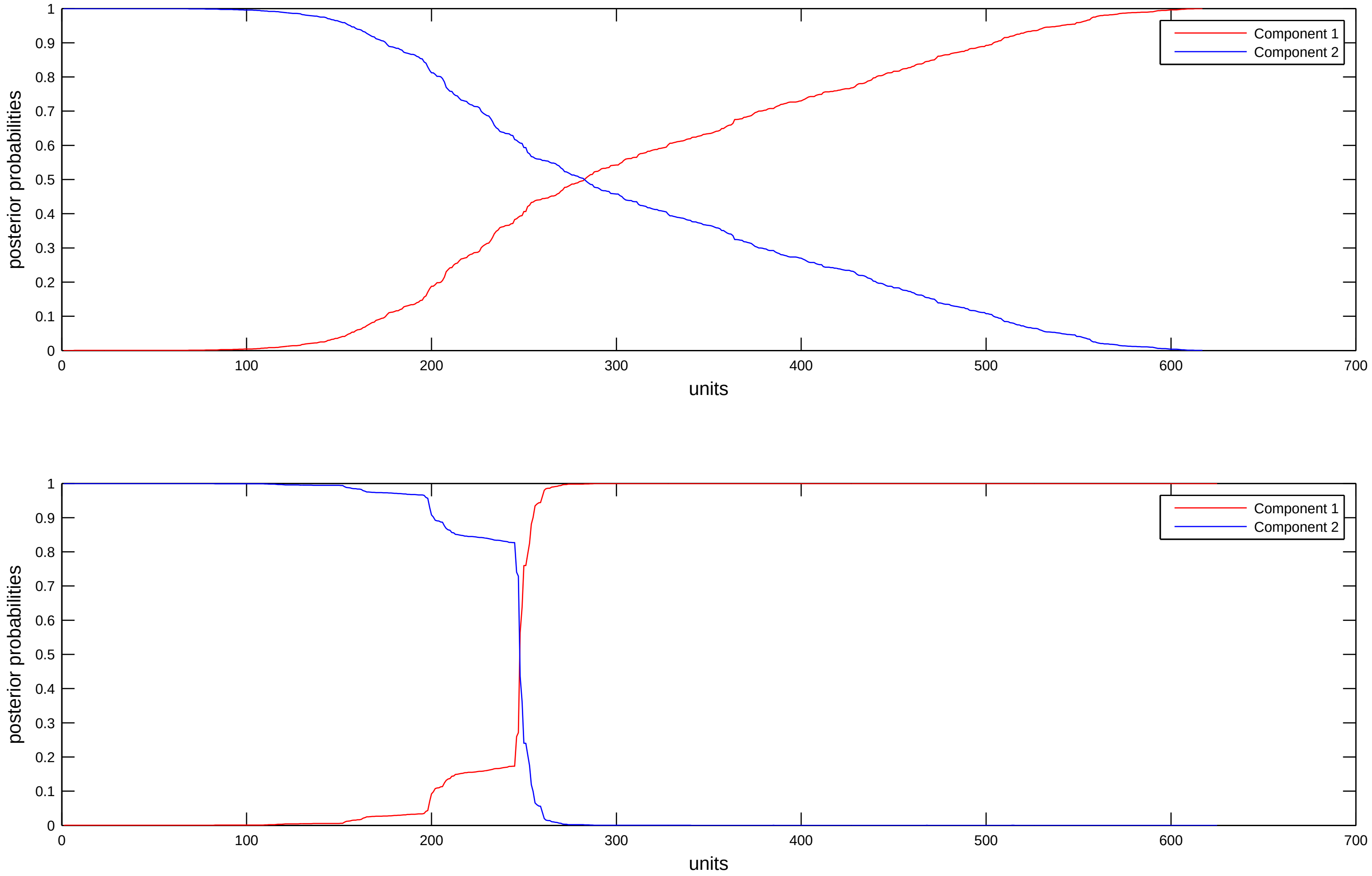


Figure 6.3: Plots of the posterior probabilities for components 1 and 2 ordered by the posterior probabilities for component 1 in ascending sequence. The top and bottom figures show the posterior probabilities from the mixture models fitted to the global and peripheral neuropathy QoL indices respectively.



.

7

Conclusions

We presented in Chapter 4 two theorems giving sufficient conditions for identifiability of the MLMM. Theorem 4.3.2 requires there to exist at least one unit that identifies the fixed effects, and at least one unit that identifies the covariance parameters (this might be the same unit). Corollary 4.3.3 applies Theorem 4.3.2 to a MLMM with a simple covariance structure, and shows the sufficient conditions of the theorem for that model translate into rank conditions on both the fixed effects and random effects design matrices respectively. The rank condition on the random effects design matrices is mild, but the rank condition on the fixed effects design matrix precludes the inclusion of covariates in the model that are constant within a unit, for example like age and sex. This is very restrictive and so alternative conditions guaranteeing identifiability were sought. These are provided by Theorem 4.3.4 which uses a hyperplane condition adapted from the one used by Hennig (2000) in clusterwise regression models. The condition in this second theorem requires that the minimum number of $(p-1)$-dimensional hyperplanes that cover all of the rows of covariate data are greater than the number of components in the mixture model.

As required Theorem 4.3.4 does permit covariates in the model that are constant within units, and so in this sense Theorem 4.3.4 is much more useful than Theorem 4.3.2. However examples of MLMMs can be found where Theorem 4.3.2 can be used to guarantee the identifiability of a model but where Theorem 4.3.4 cannot. Thus the two theorems are not equivalent to one another, and so a preference of one over the other may be determined by the particular covariate data obtained. For example it can be very difficult to verify the hyperplane condition, however this has to be balanced against the need to include covariates in the model that are constant within a unit.

In Section 5.2 we compared two different variants of the EM algorithm which we denoted by EM1 (random effects considered missing) and EM2 (random effects considered known). We found when estimation was difficult (say when sample sizes were low and covariance parameters were large) that EM1 often converged to very poor final estimates whereas for EM2 this did not happen. When parameter estimation was easier there were no large differences between the two methods in terms of the quality of estimates produced, coverage probabilities, or confidence interval lengths. Furthermore EM1 was found to be significantly slower at converging than EM2.

In Section 5.3 we conducted simulations on three models to investigate the influence of various factors on the quality of estimates produced, and on the performance of the naive methods of inference proposed in Section 3.4 in terms of coverage probabilities and confidence interval lengths. In terms of the methods of inference, we found that CI1 (score based confidence intervals) tended to produce the highest quality intervals by producing coverage probabilities either attaining or being close to the nominal level, and where this good coverage was attained without excessively long interval lengths. A close second seemed to be CI3 (componentwise confidence intervals), beating the theoretically superior CI2 (Hessian based confidence intervals) and CI4 (sandwich estimator based confidence intervals). This is a very noteworthy result since the intervals used in componentwise inference ignore the uncertainty in estimating the mixing proportions. However it is important to note that this good performance of componentwise inference may be because we did not specify components that were close enough together, in terms of separation indices, to degrade the performance of these componentwise intervals (which we theoretically expected).

In terms of the absolute level of coverage offered by these methods, coverages of approximately $80\% - 90\%$ were generally obtained when the simulation variables were set at their "optimal" levels to make estimation easy. However when estimation was made difficult, this covarage for all methods became very low $40\% - 50\%$. Thus a major point to remember is that all of these methods can often produce very poor coverage results. The factors strongly influencing coverage and confidence interval lengths were the within-unit variances, balancedness of the mixing proportions, number of units, balancedness of the within-unit sample sizes, and the random effects covariance parameters. The effect of these factors was as expected - inference quality offered by the intervals improved when the factors were set to their optimal levels. For the MSEs of the parameter estimators the same factors were also influential, but with the added factor of the ACF also being important.

Since we have shown that the naive methods of inference proposed in Section 3.4 can often produce good results, some researchers might wish to implement these methods themselves, since they offer a computationally quick way of performing inference on the model parameters compared to a bootstrap procedure. In this respect in chapter C we have derived all the derivatives required in order to do this.

In chapter 6 we analysed quality of life questionnaire (QoL) data from a lung cancer clinical trial. We showed that an MLMM can identify components within the data that sometimes cannot be easily interpreted. This occurred for an overall QoL index, where not surprisingly the classification of units to components was not very crisp - this was characterised by many units having posterior probabilities for both components around 0.5. In contrast for a more specific QoL index (peripheral neuropathy) the two estimated components could be interpreted as those patients possibly experiencing symptoms and giving "normal" questionnaire responses (and thus yielding covariate-response information), and those patients possibly not experiencing symptoms (and thus not yielding any covariate-response information). This however is just one possible interpretation that could be made of these estimated components. For this QoL index the classification of units to components was fairly crisp.

Finally we have demonstrated, and as expected, that the quality of inference provided by all methods of confidence intervals reduces quite dramatically as the separation of the components reduce. Thus in terms of future work, it would be useful to develop a method of measuring component separation that does not rely on knowing the true model parameter values. Such a measure could then be used in an applied setting to help predict if the naive methods of inference proposed here would give valid confidence intervals. In terms of future work regarding inference, there is a need to provide a proof showing the existence of a consistent estimator of the MLMM parameters. In this respect we imagine a “repeatable” design might be the best approach here. This was suggested by Hennig (2000) for clusterwsie regression models, and consists of ”repeating” the covariate data of a set of units that identify the mixture distribution function such that as N tends to infinity the identifiability of the model is maintained. This may have practical implications since by not repeating this covariate data, the inclusion of more and more covariate data from units that do not identify the model may well “swamp” the data from the identifying units, thus producing a model that is close to being non-identifiable.

Appendix A

Miscellaneous results

A.1 Autoregressive process for the within-unit errors

The following section contains a very brief summary of the theory for linear stationary time-series models, of which a purely autoregressive process is a subset. We are interested in AR processes because the AR(r) correlation matrices $\boldsymbol{C}(\boldsymbol{\phi}_g)$, $g \in I_G$, that we will use for some MLMMs is equivalent to assuming the within-unit errors for the i^{th} unit follow an AR(r) process. The material on which this summary is based can be found in Box and Reinsel (1994, Chapter 3).

A discrete time infinite AR(r) process $\{..., e_{-2}, e_{-1}, e_0, e_1, e_2, ...\}$ is defined as

$$e_t = \boldsymbol{\phi}_1 e_{t-1} + ... + \boldsymbol{\phi}_r e_{t-r} + a_t, \tag{A.1}$$

for $t \in \mathbb{Z}$, where the random variables a_t follow a white noise process, that is they are uncorrelated with zero mean and constant variance: $E[a_t] = 0$, and $\text{Var}[a_t] = \gamma_0 = \sigma_a^2$ for all t, so that for $k \in \mathbb{Z}$ this implies $\text{cov}[a_t, a_{t+k}] = \gamma_k = \sigma_a^2$ for $k = 0$, and 0 otherwise. We assume each within-unit error vector $\boldsymbol{e}_i$ in the MLMM is comprised of

n_i consecutive observed values of such a AR(r) process.

A very important type or class of AR processes are stationary AR processes, which are AR processes that are in a state of statistical equilibrium. Specifically for discrete AR processes this means that for any t, the joint distribution of $e_t, e_{t+1}, ..., e_{t+n}$, is the same as the joint distribution of $e_{t+k}, e_{t+k+1}, ..., e_{t+k+n}$. Thus the joint distribution of n consecutive observations from a stationary AR process is unaltered by shifting those observations forward or backwards by k time periods. If the process given in (A.1) is stationary then the mean and variances of e_t are the same for all t, and the covariances $\mathrm{cov}[e_t, e_{t'}] = \gamma_{|t-t'|}$ depend only on the lag between t and t'.

For $s \in \mathbb{N}^+$, let $\rho_s = \gamma_s/\gamma_0$ be the correlation between e_t and e_{t+s}, and let $\rho_0 = 1$. For a stationary AR process we have that ρ_s is given by

$$\rho_s(\phi) = \boldsymbol{\phi}_1\rho_{s-1} + ... + \boldsymbol{\phi}_r\rho_{s-r}, \quad s \in \mathbb{N}^+, \tag{A.2}$$

which is called the autocorrelation function, or ACF, of the AR process. For stationary AR processes we have $\rho_{-s} = \rho_s$ since $\gamma_{-s} = \gamma_s$, and so as for the covariances only the lag $|t-t'|$ determines the correlation between e_t and $e_{t'}$ and not the actual time periods t and t'.

For an AR(r) process it is necessary first to estimate the r autocorrelations $\rho_1, ..., \rho_r$ before being able to use A.2 to sequentially calculate $\rho_{r+1}, \rho_{r+2}, \rho_{r+3}, ...$ This is done by using (A.2) to obtain a set of r linear equations for $\phi_1, ..., \phi_r$ in terms of $\rho_1(\phi), ..., \rho_r(\phi)$ - these are called the Yule-Walker equations. Solving these equations for $\phi_1, ..., \phi_r$ gives the starting values we require in order to calculate all the autocorrelations given in (A.2). For an AR(1) process this method gives

$$\rho_s = \phi_1^s \qquad s \in \mathbb{N}^+, \tag{A.3}$$

for an AR(2) process this method gives

$$\begin{aligned}\rho_1 &= \frac{\phi_1}{1-\phi_2},\\ \rho_2 &= \phi_2 + \frac{\phi_1^2}{1-\phi_2},\end{aligned} \tag{A.4}$$

whilst for an AR(3) process this method gives

$$\begin{aligned}\rho_1 &= \frac{\phi_1 - \phi_1\phi_2 - \phi_3\phi_2^2 + \phi_3\phi_2}{1 - 2\phi_2 - \phi_1\phi_3 - \phi_3^2 + \phi_2^2 + \phi_1\phi_2\phi_3 + \phi_2\phi_3^2},\\ \rho_2 &= \frac{\phi_1^2 - \phi_2^2 + \phi_2 + \phi_3\phi_1}{1 - \phi_2 - \phi_1\phi_3 - \phi_3^2},\\ \rho_3 &= \frac{\phi_1^3 - \phi_1\phi_2^2 + \phi_1\phi_2 + \phi_1^2\phi_3}{1 - \phi_2 - \phi_1\phi_3 - \phi_3^2} + \frac{\phi_1\phi_2 - \phi_1\phi_2^2 - \phi_3\phi_2^2 + \phi_2^2\phi_3}{1 - 2\phi_2 - \phi_1\phi_3 - \phi_3^2 + \phi_2^2 + \phi_1\phi_2\phi_3 + \phi_2\phi_3^2}\\ &\quad + \phi_3.\end{aligned} \tag{A.5}$$

In terms of applying this theory to the vectors of within-unit errors for a MLMM, for any unit $i \in I_N$, and conditional on unit i belonging to component $g \in I_G$, then assuming the within-unit errors $e_{i,1}, ..., e_{i,n_i}$ contained in $\boldsymbol{e}_i$ are n_i consecutive realized values from a stationary AR(r) process is equivalent to assuming that the variance of $\boldsymbol{e}_i$ conditional on $\boldsymbol{\lambda}_i^{(g)}$ has the following form

$$\mathrm{Var}[\boldsymbol{e}_i|\boldsymbol{\lambda}_i^{(g)}] = \sigma_g^2 \boldsymbol{C}_i(\boldsymbol{\phi}_g) = \sigma_g^2 \{_m \;\; \rho_{|t-t'|(\boldsymbol{\phi}_g)}\}_{t=1,\, t'=1}^{ni \quad\; ni}, \tag{A.6}$$

where

$$\sigma_g^2 = \frac{\sigma_a^2}{1 - \rho_1\phi_1 - \rho_2\phi_2 - ... - \rho_r\phi_r}, \tag{A.7}$$

and

$$\rho_s(\boldsymbol{\phi}_g) = \boldsymbol{\phi}_{g,1}\rho_{s-1} + ... + \boldsymbol{\phi}_{g,r}\rho_{s-r}, \quad \rho_0 = 1, \quad s = 1, ..., n_i - 1, \tag{A.8}$$

and furthermore the matrix in A.6 is always positive-definite. If we further suppose the a_t in the underlying infinite AR process are normally distributed then $\sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g)$ is the covariance matrix of the normal distribution of $\boldsymbol{Y}_i$ conditional on $\boldsymbol{U}_i = \boldsymbol{u}_i$ (the density function of which is given in (2.5)). Thus we have the important result that if $\boldsymbol{e}_i$ follows a stationary AR(r) process then the covariance matrix $\sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g)$ is positive-definite. Accordingly during parameter estimation, if we can ensure the estimates of $\boldsymbol{\phi}_g$ always give rise to a stationary AR process then the estimates of $\boldsymbol{C}_i(\boldsymbol{\phi}_g)$ will always be positive-definite.

In order to understand the conditions we need to impose on the r individual AR parameters in $\boldsymbol{\phi}$ so that the AR process is stationary, we need to introduce the characteristic equation $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_r B^r = 0$. In this function B is the backward shift operator, and is considered to be a variable that can take on complex values. The backward shift operator B operates on the time index of a variable that it is multiplied with: $Be_t = e_{t-1}$, and for $k \in \mathbb{Z}$, $B^k e_t = e_{t-k}$. Thus the model in (A.1) can be written in terms of the backward shift operator as $\phi(B)e_t = a_t$. Now $\phi(B)$ is a function of $\boldsymbol{\phi}$, and so it turns out that the conditions on $\boldsymbol{\phi}$ we need for stationarity can be described in terms of the roots of $\phi(B)$: the characteristic equation can be factored as $\phi(B) = (1 - G_1 B)(1 - G_2 B)...(1 - G_r B)$, where $G_1^{-1}, ..., G_r^{-1}$, are the roots of $\phi(B) = 0$. For an AR(r) process to be stationary we must have $\left|G_{\mathrm{v}}^{-1}\right| > 1$ for all $\mathrm{v} \in \{1, ..., r\}$, so that roots of the characteristic equation must all lie outside the unit circle.

For low values of r, the conditions on $\boldsymbol{\phi}$ that ensure the roots of $\phi(B)$ lie outside of the unit circle are reasonably simple to calculate and succinct. For $r = 1$ we need $-1 < \phi_1 < 1$, whilst for $r = 2$ we need the three equations: $\phi_2 + \phi_1 < 1$, $\phi_2 - \phi_1 < 1$, and $-1 < \phi_2 < 1$. For higher orders of r the calculation of these conditions is more difficult, and they comprise many equations. A much simpler method of specifying conditions on $\boldsymbol{\phi}$ that ensure a stationary AR process is to specify conditions instead on the vector of partial autocorrelations, which we will denote by $\boldsymbol{\tau} = (\tau_1, ..., \tau_r)^{\intercal}$, and so τ_{v}, for $\mathrm{v} = 1, .., r$, is the v^{th} partial autocorrelation. This is the approach taken by Wang and Fan (2009), who also state that estimating the partial autocorrelations is a more stable procedure than estimating the AR parameters themselves. Importantly there is a one to one mapping of $\boldsymbol{\phi}$ to $\boldsymbol{\tau}$ which can be easily calculated.

For these reasons we too adopt the approach of Wang and Fan (2009) and estimate the partial autocorrelation vectors $\boldsymbol{\tau}_j = ((\boldsymbol{\tau}_j)_1, \cdots, (\boldsymbol{\tau}_j)_r)^{\intercal}$ instead of $\boldsymbol{\phi}_j$ for $j = 1, ..., G$. We now briefly describe how the partial autocorrelations of an AR process can be calculated, and finish with the equation which gives the one to one mapping of $\boldsymbol{\phi}$ to $\boldsymbol{\tau}$. Let the process $\{e_t^{\mathrm{v}}\}$ be an AR(v) process, for $\mathrm{v} \in \mathbb{N}^+$, and let $\boldsymbol{\phi}_s^{(\mathrm{v})}$ be the s^{th}

parameter, $s = 1, ..., \mathrm{v}$. Then from $(A.2)$ the ACF of the process $\{e_t^{\mathrm{v}}\}$ is given by

$$\rho_s = \boldsymbol{\phi}_1^{(\mathrm{v})} \rho_{s-1} + ... + \boldsymbol{\phi}_{\mathrm{v}-1}^{(\mathrm{v})} \rho_{s-\mathrm{v}+1} + \boldsymbol{\phi}_{\mathrm{v}}^{(\mathrm{v})} \rho_{s-\mathrm{v}} \quad s = 1, ..., \mathrm{v}. \tag{A.9}$$

The set of v equations in (A.9) are called the Yule-Walker equations, and which for $\mathrm{v} = 1, 2, ...,$ can be solved in turn for $\boldsymbol{\phi}_{\mathrm{v}} := (\phi_1^{(\mathrm{v})}, \cdots, \phi_{\mathrm{v}}^{(\mathrm{v})})^{\intercal}$. From these solutions it is the quantity $\phi_{\mathrm{v}}^{(\mathrm{v})}$, when viewed as a function of the lag v, which is defined to be the partial autocorrelation function. The partial autocorrelation function $\phi_{\mathrm{v}}^{(\mathrm{v})}$ is defined for any stationary AR(r) process $\{e_t\}$, and is so called because it can be shown to be equal to the correlation between e_t and $e_{t-\mathrm{v}}$, but where this partial autocorrelation is not accounted for by the intermediate values $e_{t-1}, ..., e_{t-\mathrm{v}-1}$. For details of how to calculate $\phi_{\mathrm{v}}^{(\mathrm{v})}$ to help illustrate this concept of partial autocorrelation see Box and Reinsel (1994, Chapter 3, pp66-67). A useful property of the partial autocorrelations is that $\phi_{\mathrm{v}}^{(\mathrm{v})} = 0$ for all $\mathrm{v} > r$, and so they are a useful tool to help identify the order of the AR process.

If we define $\tau_{\mathrm{v}} = \boldsymbol{\phi}_{\mathrm{v}}^{(\mathrm{v})}$ for $\mathrm{v} = 1, ..., r$, then $\boldsymbol{\tau}$ contains the individual values of the partial autocorrelation function as v varies up to and including r. From Wang and Fan (2009) we have the following relationship between the AR parameters and the r partial autocorrelations

$$\begin{aligned} \boldsymbol{\phi}_r^{(r)} &= \tau_r \\ \boldsymbol{\phi}_{\mathrm{v}}^{(r)} &= \boldsymbol{\phi}_{\mathrm{v}}^{(r-1)} - \tau_r \boldsymbol{\phi}_{r-\mathrm{v}}^{(r-1)} \\ &= \tau_{\mathrm{v}} - \tau_{\mathrm{v}+1} \boldsymbol{\phi}_1^{\mathrm{v}} - \tau_{\mathrm{v}+2} \boldsymbol{\phi}_2^{\mathrm{v}+1} - ... - \tau_r \boldsymbol{\phi}_{r-\mathrm{v}}^{r-1}, \end{aligned} \tag{A.10}$$

where $\mathrm{v} \in \{1, ..., r-1\}$. The relationship (A.10) defines a one-to-one mapping from $\boldsymbol{\phi}$

to $\boldsymbol{\tau}$, and furthermore the conditions on $\boldsymbol{\tau}$ for the AR(r) process to be stationary are that $\tau_{\mathrm{v}} \in [-1, 1]$ for all $\mathrm{v} = 1, ..., r$, or equivalently that $\boldsymbol{\tau} \in [-1, 1]^r$. Compared to the conditions described previously regarding the roots of $\phi(B)$, this condition has the big advantage in that it is simple to interpret and implement for all values of r.

A.2 Rank of matrices and hyperplanes

Theorem A.2.1 *For any $n \times p$ matrix $\boldsymbol{X}$*

$$\begin{aligned} \mathit{rank}(\boldsymbol{X}) = p - 1 &\Longleftrightarrow \mathit{dim}(S_{\boldsymbol{X}}) = p - 1 \\ &\Longleftrightarrow S_{\boldsymbol{X}} = H_{p-1}(\boldsymbol{\alpha}, 0) \\ &\Longleftrightarrow (\boldsymbol{X})_{j\cdot} \in H_{p-1}(\boldsymbol{\alpha}, 0) \textit{ for all } j = 1, ..., n, \end{aligned} \tag{A.11}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^p$. If the first column of $\boldsymbol{X}$ is a column of 1's we have

$$\begin{aligned} \mathit{rank}(\boldsymbol{X}) = p - 1 &\Longrightarrow \mathit{dim}(S_{\boldsymbol{X}}) = p - 1 \\ &\Longrightarrow S_{\boldsymbol{X}^-} = H_{p-2}(\boldsymbol{\alpha}, 0) \\ &\Longrightarrow (\boldsymbol{X}^-)_{j\cdot} \in H_{p-2}(\boldsymbol{\alpha}, 0) \textit{ for all } j = 1, ..., n, \end{aligned} \tag{A.12}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^{p-1}$.

Proof. We will use the notation $\boldsymbol{X}^-$ to mean the $n \times (p-1)$ matrix obtained by removing the first column from $\boldsymbol{X}$. Let $S_{\boldsymbol{X}} = \mathrm{span}\{(\boldsymbol{X})_{1\cdot}, ..., (\boldsymbol{X})_{n\cdot}\}$, and $S_{\boldsymbol{X}^-} = \mathrm{span}\{(\boldsymbol{X}^-)_{1\cdot}, ..., (\boldsymbol{X}^-)_{n\cdot}\}$ be the row spaces of $\boldsymbol{X}$ and $\boldsymbol{X}^-$ respectively.

We firstly relate $\dim(S_{\boldsymbol{X}})$ with $\mathrm{rank}(\boldsymbol{X})$. Assuming $\dim(S_{\boldsymbol{X}}) = r$, $1 \leq r \leq p-1$, means that any basis set for $S_{\boldsymbol{X}_i}$ will contain r vectors. Now the rows of $\boldsymbol{X}$ are obviously a spanning set for $S_{\boldsymbol{X}}$, and this set can be reduced to a basis set by the removal of appropriate rows that are linearly related. This implies r of the rows of $\boldsymbol{X}$ are linearly independent. Since by definition the rank of $\boldsymbol{X}$ must equal the number of linearly independent rows of $\boldsymbol{X}$, or the number of linearly independent columns of $\boldsymbol{X}$ (these numbers are the same), then we must have $\text{rank}(\boldsymbol{X}) = r$. Conversely the assumption $\text{rank}(\boldsymbol{X}) = r$ means that $\boldsymbol{X}$ has r linearly independent rows, and so $\dim(S_{\boldsymbol{X}}) = r$. This gives the following result

$$\text{rank}(\boldsymbol{X}) = r \Longleftrightarrow \dim(S_{\boldsymbol{X}}) = r, \tag{A.13}$$

where $1 \leq r \leq p - 1$.

We now relate the dimension of $S_{\boldsymbol{X}}$ with the hyperplane definition in 4.14. Assuming $\dim(S_{\boldsymbol{X}}) = p - 1$, means that $\boldsymbol{X}$ has $p - 1$ linearly independent rows that form a basis set for $S_{\boldsymbol{X}}$. Let $B = \{\boldsymbol{e}_1, ..., \boldsymbol{e}_{p-1}\}$ be one of these basis sets, where for each $l = 1, ..., p - 1$, $\boldsymbol{e}_l = (\boldsymbol{X})_{m\cdot}$, for some $m = 1, ..., p$. In terms of elements we shall write $\boldsymbol{e}_l = (e_{l1}, ..., e_{lp})^\intercal \in \mathbb{R}^p$, for all l. Any $\boldsymbol{x} = (x_1, ..., x_p)^\intercal \in \mathbb{R}^p$ such that $\boldsymbol{x} \in S_{\boldsymbol{X}}$, has the parametric form $\boldsymbol{x} = \sum_{l=1}^{p-1} \alpha_l \boldsymbol{e}_l$, where α_l for all l are independent scalars. We can find a non-zero vector $\boldsymbol{n} = (n_1, ..., n_p)^\intercal \in \mathbb{R}^p$ that is orthogonal to the $p - 1$ basis vectors in B, and so $\boldsymbol{n}^\intercal \boldsymbol{x} = 0$. Thus $\boldsymbol{x}$ satisfies $n_1 x_1 + ... n_p x_p = 0$, which shows $\boldsymbol{x}$ lies on a $(p - 1)$-dimensional hyperplane $H_{p-1}(\boldsymbol{n}, 0)$, which implies $S_{\boldsymbol{X}} = H_{p-1}(\boldsymbol{n}, 0)$. In particular since all the rows of $\boldsymbol{X}$ are in $S_{\boldsymbol{X}}$ we must have $(\boldsymbol{X})_{j\cdot} \in H_{p-1}(\boldsymbol{n}, 0)$, for $j = 1, ..., n$. If the first column of $\boldsymbol{X}$ is a column of 1's then any $\boldsymbol{x} \in S_{\boldsymbol{X}}$ satisfies $n_1(1) + n_2 x_2 + ... n_p x_p = 0$, or equivalently $n_2 x_2 + ... n_p x_p = c$, for $c = -n_1$, which implies $\boldsymbol{x}^- \in H_{p-2}(\boldsymbol{n}', c)$, where $\boldsymbol{n}' = (n_2, ..., n_p)^\intercal \in \mathbb{R}^{p-1}$, and so $S_{\boldsymbol{X}^-} = H_{p-2}(\boldsymbol{n}', c)$. This implies $(\boldsymbol{X}^-)_{j\cdot} \in H_{p-2}(\boldsymbol{n}', c)$, for $j = 1, ..., n$.

Conversely assume all the rows of $\boldsymbol{X}$ lie on a $(p-1)$-dimensional hyperplane $H_{p-1}(\boldsymbol{\alpha}, 0)$, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p) \in \mathbb{R}^p$. Since $H_{p-1}(\boldsymbol{\alpha}, 0)$ goes through the origin it is a is a vector space, and so linear combinations of vectors in $H_{p-1}(\boldsymbol{\alpha}, 0)$ will also be in $H_{p-1}(\boldsymbol{\alpha}, 0)$. Thus $S_{\boldsymbol{X}} = H_{p-1}(\boldsymbol{\alpha}, 0)$, and so any $\boldsymbol{x} = (x_1, ..., x_p)^{\intercal} \in S_{\boldsymbol{X}}$ satisfies $\alpha_1 x_1 + ... \alpha_p x_p = 0$. Assume without loss of generality that $\alpha_1 \neq 0$, then $\boldsymbol{\alpha}^{\intercal}\boldsymbol{x} = 0$ can be written $x_1 = -\sum_{l \neq 1}^{p} \alpha_1^{-1}\alpha_l x_l$, so that $\boldsymbol{x} = \sum_{l \neq 1}^{p} \boldsymbol{z_l} x_l$, where $\boldsymbol{z}_1 = (\alpha_1^{-1}\alpha_2, 1, 0, ..., 0)^{\intercal}$,..., $\boldsymbol{z}_{p-1} = (\alpha_1^{-1}\alpha_p, 0, ..., 0, 1)^{\intercal}$, and where $\{\boldsymbol{z}_1, ..., \boldsymbol{z}_{p-1}\}$ are linearly independent. So for all $\boldsymbol{x} \in S_{\boldsymbol{X}}$, $\boldsymbol{x} \in \text{span}\{\boldsymbol{z}_1, ..., \boldsymbol{z}_{p-1}\}$, which shows $\{\boldsymbol{z}_1, ..., \boldsymbol{z}_{p-1}\}$ is a basis set for $S_{\boldsymbol{X}}$. This implies $\dim(S_{\boldsymbol{X}}) = p - 1$.

Now assume the first column of $\boldsymbol{X}$ is a column of 1's. Then the assumption all the rows of $\boldsymbol{X}^-$ lie on a $(p-2)$-dimensional hyperplane $H_{p-2}(\boldsymbol{\alpha}, 0)$, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{p-1}) \in \mathbb{R}^{p-1}$, leads to the conclusion $\dim(S_{\boldsymbol{X}^-}) = p - 2$, by repeating the arguments in the above paragraph, but by making the obvious changes to the dimensions of vectors. However we do not in general have $\dim(S_{\boldsymbol{X}^-}) = p - 2 \Rightarrow \dim(S_{\boldsymbol{X}}) = p - 1$, since one or more of the columns of $\boldsymbol{X}^-$ may be linearly dependent with the intercept. Thus we have the following results

$$\dim(S_{\boldsymbol{X}}) = p - 1 \Longleftrightarrow S_{\boldsymbol{X}} = H_{p-1}(\boldsymbol{\alpha}, 0) \quad \Longleftrightarrow (\boldsymbol{X})_{j\cdot} \in H_{p-1}(\boldsymbol{\alpha}, 0) \text{ for all } j = 1, ..., n, \tag{A.14}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^p$. If the first column of $\boldsymbol{X}$ is a column of 1's we have

$$\dim(S_{\boldsymbol{X}}) = p - 1 \Longrightarrow S_{\boldsymbol{X}^-} = H_{p-2}(\boldsymbol{\alpha}, 0) \quad \Longrightarrow (\boldsymbol{X}^-)_{j\cdot} \in H_{p-2}(\boldsymbol{\alpha}, 0) \text{ for all } j = 1, ..., n, \tag{A.15}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^{p-1}$, and

$$\dim(S_{\boldsymbol{X}^-}) = p-2 \Longleftarrow S_{\boldsymbol{X}^-} = H_{p-2}(\boldsymbol{\alpha}, 0) \quad \Longleftarrow (\boldsymbol{X}^-)_{j\cdot} \in H_{p-2}(\boldsymbol{\alpha}, 0) \text{ for all } j = 1, ..., n, \tag{A.16}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^{p-1}$. The results (A.14) and (A.15), in combination with (A.13) then give the results (A.11) and (A.12).■

A.3 Derivation of ECM algorithm estimating equations

This section describes the derivation of the ECM estimating equations given in subsection 2.2.1. We will need a function $I_i = g$ when unit i is in component $g \in I_G$ and $I_i = 0$ when not. We firstly derive the log-likelihood function for the i^{th} complete data vector $\boldsymbol{c}_i = (\boldsymbol{y}_i^{\intercal}, \boldsymbol{u}_i^{\intercal}, \boldsymbol{\lambda}_i^{(I_i)\intercal})^{\intercal}$. Letting w_{ig} be the density for $(\boldsymbol{Y}_i, \boldsymbol{U}_i)|\boldsymbol{\Lambda}_i = \boldsymbol{\lambda}_i^{(g)}$, we now write the complete data density f_i^c conditional on $\boldsymbol{\lambda}_i^{(I_i)}$ as a product involving all G components, that is

$$\begin{aligned} f_i^c(\boldsymbol{c}_i|\boldsymbol{\theta}_{I_i}) &= w_{i,I_i}(\boldsymbol{y}_i, \boldsymbol{u}_i|\,\boldsymbol{\lambda}_i^{(,I_i)}, \boldsymbol{\theta}_{,I_i})h(\boldsymbol{\lambda}_i^{(I_i)}|\boldsymbol{\pi}_{I_i}) \\ &= \prod_{j=1}^{G}\left(w_{ij}(\boldsymbol{y}_i, \boldsymbol{u}_i|\,\boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(j)}, \boldsymbol{\theta}_j)^{\boldsymbol{\lambda}_{ij}^{(I_i)}}\right)\prod_{j'=1}^{G}\left(\boldsymbol{\pi}_{j'}^{\boldsymbol{\lambda}_{ij'}^{(I_i)}}\right), \end{aligned} \tag{A.17}$$

where the last line holds since only the I_i^{th} element of $\boldsymbol{\lambda}_{i,}^{(I_i)}$ is equal to 1 whilst the others are zero. Letting $\boldsymbol{C} = (\boldsymbol{C}_1^{\intercal}, ..., \boldsymbol{C}_N^{\intercal})^{\intercal}$ and $\boldsymbol{c} = (\boldsymbol{c}_1^{\intercal}, ..., \boldsymbol{c}_N^{\intercal})^{\intercal}$, then from independence of the random variables $\{\boldsymbol{C}_1, ..., \boldsymbol{C}_N\}$, the complete data log-likelihood $\text{Ł}^c(\boldsymbol{\theta}|\boldsymbol{c})$ is

$$
\begin{aligned}
\text{Ł}^{c}(\boldsymbol{\theta}|\boldsymbol{c}) &= \sum_{i=1}^{N}\sum_{j=1}^{G} \boldsymbol{\lambda}_{ij}^{(I_i)} \log\Big(w_{ij}(\boldsymbol{y}_i, \boldsymbol{u}_i|\, \boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(j)}, \boldsymbol{\theta}_j)\Big) + \sum_{i=1}^{N}\sum_{j'=1}^{G} \boldsymbol{\lambda}_{ij'}^{(I_i)} \log\left(\boldsymbol{\pi}_{j'}\right) \\
&= \sum_{i=1}^{N} \Big(\boldsymbol{\lambda}_{i}^{(I_i)}\Big)^{\intercal} T_i(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{u}_i) + \sum_{i=1}^{N} \boldsymbol{\lambda}_{i}^{(I_i)\intercal} \boldsymbol{U}(\boldsymbol{\pi}), \qquad \text{(A.18)}
\end{aligned}
$$

where $T_i(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{u}_i) = \{_{c}\ \log\{w_{ij}(\boldsymbol{y}_i, \boldsymbol{u}_i|\, \boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(j)}, \boldsymbol{\theta}_j)\}\}_{j=1}^{G}$, and $\boldsymbol{U}(\boldsymbol{\pi}) = \{_{c}\ \log(\boldsymbol{\pi}_j)\}_{j=1}^{G}$.

The EM algorithm maximises the ordinary log-likelihood $L(\boldsymbol{\theta}|\boldsymbol{y})$ by working with $\boldsymbol{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{'})$, which is the expected value of $L^{c}(\boldsymbol{\theta}|\boldsymbol{C})$ conditional on $\boldsymbol{y}$ and $\boldsymbol{\theta}^{'}$. If we let s denote the current iteration of the EM algorithm, and $\hat{\boldsymbol{\theta}}^{(s)}$ the estimate obtained, then the E-step consists of calculating $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$ which is given by

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) &= \boldsymbol{E}\left[L^{c}(\boldsymbol{\theta}|\boldsymbol{C})|\, \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right] \\
&= \sum_{i=1}^{N} \boldsymbol{E}\left[\boldsymbol{\Lambda}_{i,I_i} \log\Big(f_{i,I_i}(\boldsymbol{y}_i, \boldsymbol{U}_i|\, \boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(I_i)}, \boldsymbol{\theta}_{I_i})\Big)\Big|\, \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right] \\
&\quad + \sum_{i=1}^{N} \boldsymbol{E}\left[\boldsymbol{\Lambda}_{i,I_i}|\, \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right] \log(\boldsymbol{\pi}_{I_i}). \qquad \text{(A.19)}
\end{aligned}
$$

Using the notation defined in chapter 2 we note that $\boldsymbol{\Lambda}_{ij} = \boldsymbol{\lambda}_{ik}^{(k)}$ implies that $\boldsymbol{\Lambda}_{ik} = 1$ since $\boldsymbol{\lambda}_{ik}^{(k)} = 1$, or equivalently that $\boldsymbol{\Lambda}_i = \boldsymbol{\lambda}_i^{(k)}$. For this reason we also have that $\boldsymbol{P}[\boldsymbol{\Lambda}_{ij} = \boldsymbol{\lambda}_{ik}^{(k)}|\hat{\boldsymbol{\pi}}^{(s)})]$ means the same thing as $\boldsymbol{P}[\boldsymbol{\Lambda}_i = \boldsymbol{\lambda}_i^{(k)}|\hat{\boldsymbol{\pi}}^{(s)})]$. Furthermore we also note that theoretically $\boldsymbol{\Lambda}_{i,I_i}$ can take on values that are either 0 or 1, where additionally I_i can take on any value in the set I_G. Thus the range of values of $\boldsymbol{\Lambda}_{i,I_i}$ can take can be enumerated by the values $\boldsymbol{\lambda}_{ik'}^{(k)}$ for $k', k = 1, ..., G$, and this enumeration is the same regardless of the component membership of the i^{th} unit. This is why the double summation appears in (A.20) as a result of the expectation operator acting on $\boldsymbol{\Lambda}_{i,I_i}$, and why this is true for any $i \in I_N$. With these things in mind we first calculate $\boldsymbol{E}\left[\boldsymbol{\Lambda}_{i,I_i} | \, \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right]$ which is given by

$$
\begin{aligned}
\boldsymbol{E}\left[\boldsymbol{\Lambda}_{i,I_i} | \, \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right] &= \sum_{k'=1}^{G} \sum_{k=1}^{G} \boldsymbol{\lambda}_{ik'}^{(k)} \boldsymbol{P}[\boldsymbol{\Lambda}_{ij} = \boldsymbol{\lambda}_{ik'}^{(k)} | \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}] \\
&= \sum_{k=1}^{G} \boldsymbol{\lambda}_{ik}^{(k)} \boldsymbol{P}[\boldsymbol{\Lambda}_{ij} = \boldsymbol{\lambda}_{ik}^{(k)} | \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}] \\
&= \frac{\sum_{k=1}^{G} \boldsymbol{P}[\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{\Lambda}_{ij} = \boldsymbol{\lambda}_{ik}^{(k)}, \hat{\boldsymbol{\theta}}^{(s)}] \boldsymbol{P}[\boldsymbol{\Lambda}_{ij} = \boldsymbol{\lambda}_{ik}^{(k)} | \hat{\boldsymbol{\pi}}^{(s)})]}{\boldsymbol{P}[\boldsymbol{Y}_i = \boldsymbol{y}_i | \hat{\boldsymbol{\theta}}^{(s)}]} \\
&= \sum_{k=1}^{G} \frac{f_{ik}(\boldsymbol{y}_i | \boldsymbol{\lambda}_i^{(k)}, \hat{\boldsymbol{\theta}}_k^{(s)}) \boldsymbol{h}(\boldsymbol{\lambda}_i^{(k)} | \hat{\boldsymbol{\pi}}^{(s)})}{f_i(\boldsymbol{y}_i | \hat{\boldsymbol{\theta}}^{(s)})} \\
&= \sum_{k=1}^{G} \frac{f_{ik}(\boldsymbol{y}_i | \boldsymbol{\lambda}_i^{(k)}, \hat{\boldsymbol{\theta}}_k^{(s)}) \hat{\boldsymbol{\pi}_k}^{(s)}}{f_i(\boldsymbol{y}_i | \hat{\boldsymbol{\theta}}^{(s)})} \\
&= \sum_{k=1}^{G} \left(\frac{f_{ik}(\boldsymbol{y}_i | \boldsymbol{\lambda}_i^{(k)}, \hat{\boldsymbol{\theta}}_k^{(s)}) \hat{\boldsymbol{\pi}_k}^{(s)}}{\sum_{l=1}^{G} f_{il}(\boldsymbol{y}_i | \boldsymbol{\lambda}_i^{(l)}, \hat{\boldsymbol{\theta}}_l^{(s)}) \hat{\boldsymbol{\pi}_l}^{(s)}} \right) \\
&= \sum_{k=1}^{G} \hat{p}_i(\boldsymbol{\lambda}_i^{(k)} | \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}), \qquad \text{(A.20)}
\end{aligned}
$$

where

$$
\hat{p}_i(\boldsymbol{\lambda}_i^{(g)} | \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}) = \frac{f_{ig}(\boldsymbol{y}_i | \boldsymbol{\lambda}_i^{(g)}, \hat{\boldsymbol{\theta}}_g^{(s)}) \hat{\boldsymbol{\pi}_g}^{(s)}}{\sum_{l=1}^{G} f_{il}(\boldsymbol{y}_i | \boldsymbol{\lambda}_i^{(l)}, \hat{\boldsymbol{\theta}}_l^{(s)}) \hat{\boldsymbol{\pi}_l}^{(s)}}, \qquad \text{(A.21)}
$$

is the posterior probability of the i^{th} unit belonging to the g^{th} component, conditional on the observed response vector for that unit, and the current estimate of $\boldsymbol{\theta}$.

Before we calculate $\boldsymbol{E}\left[\boldsymbol{\Lambda}_{i,I_i} \log\left(w_{i,I_i}(\boldsymbol{y}_i, \boldsymbol{U}_i | \, \boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(I_i)}, \boldsymbol{\theta}_{I_i})\right) \middle| \, \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right]$, we first need to define some more density functions: let w_{ig} be the density for $(\boldsymbol{Y}_i, \boldsymbol{U}_i) | \boldsymbol{\Lambda}_i = \boldsymbol{\lambda}_i^{(g)}$, z_i the density for $\boldsymbol{U}_i, \boldsymbol{\Lambda}_i | \boldsymbol{Y}_i$, and t_{ig} the density for $\boldsymbol{U}_i | \boldsymbol{Y}_i, \boldsymbol{\Lambda}_i = \boldsymbol{\lambda}_i^{(g)}$. We can now calculate $\boldsymbol{E}\left[\boldsymbol{\Lambda}_{i,I_i} \log\left(w_{i,I_i}(\boldsymbol{y}_i, \boldsymbol{U}_i | \, \boldsymbol{\Lambda}_i = \boldsymbol{\lambda}^{(I_i)}, \boldsymbol{\theta}_{I_i})\right) \middle| \, \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right]$ which is given by

$$
\begin{aligned}
&\boldsymbol{E}\left[\boldsymbol{\Lambda}_{i,I_i}\log\left(w_{i,I_i}(\boldsymbol{y}_i,\boldsymbol{U}_i|\,\boldsymbol{\Lambda}_i=\boldsymbol{\lambda}^{(I_i)},\boldsymbol{\theta}_{,I_i})\right)\Big|\,\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)}\right]\\
&=\sum_{k=1}^{G}\sum_{k'=1}^{G}\left\{\boldsymbol{\lambda}_{ik'}^{(k)}\int_{\mathbb{R}^q}\left[\log\left(w_{ik}(\boldsymbol{y}_i,\boldsymbol{u}|\,\boldsymbol{\Lambda}_i=\boldsymbol{\lambda}_i^{(k)},\boldsymbol{\theta}_k)\right)z_i(\boldsymbol{u},\boldsymbol{\lambda}_i^{(k)}\Big|\,\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\right]d\boldsymbol{u}\right\}\\
&=\sum_{k=1}^{G}\left\{\boldsymbol{\lambda}_{ik}^{(k)}\int_{\mathbb{R}^q}\left[\log\left(w_{ik}(\boldsymbol{y}_i,\boldsymbol{u}|\,\boldsymbol{\Lambda}_i=\boldsymbol{\lambda}_i^{(k)},\boldsymbol{\theta}_k)\right)z_i(\boldsymbol{u},\boldsymbol{\lambda}_i^{(k)}\Big|\,\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\right]d\boldsymbol{u}\right\}\\
&=\sum_{k=1}^{G}\left\{\int_{\mathbb{R}^q}\left[\log\left(w_{ik}(\boldsymbol{y}_i,\boldsymbol{u}|\,\boldsymbol{\lambda}_i^{(k)},\boldsymbol{\theta}_k)\right)\right.\right.\\
&\qquad\left.\left.\times\frac{t_{ik}(\boldsymbol{u}|\,\boldsymbol{y}_i,\boldsymbol{\lambda}_i^{(k)},\hat{\boldsymbol{\theta}_k}^{(s)})h(\boldsymbol{\lambda}_i^{(k)}|\hat{\boldsymbol{\pi}_k}^{(s)})f_{ik}(\boldsymbol{y}_i|\,\boldsymbol{\lambda}_i^{(k)},\hat{\boldsymbol{\theta}_k}^{(s)})}{f_i(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}^{(s)})}\right]d\boldsymbol{u}\right\}\\
&=\sum_{k=1}^{G}\frac{h(\boldsymbol{\lambda}_i^{(k)}|\hat{\boldsymbol{\pi}_k}^{(s)})f_{ik}(\boldsymbol{y}_i|\,\boldsymbol{\lambda}_i^{(k)},\hat{\boldsymbol{\theta}_k}^{(s)})}{f_i(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}^{(s)})}\\
&\quad\times\int_{\mathbb{R}^q}\left[\log\left(w_{ik}(\boldsymbol{y}_i,\boldsymbol{u}|\,\boldsymbol{\lambda}_i^{(k)},\boldsymbol{\theta}_k)\right)t_{ik}(\boldsymbol{u}|\,\boldsymbol{y}_i,\boldsymbol{\lambda}_i^{(k)},\hat{\boldsymbol{\theta}_k}^{(s)})\right]d\boldsymbol{u}\\
&=\sum_{k=1}^{G}\hat{p}_i(\boldsymbol{\lambda}_i^{(k)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\boldsymbol{E}\left[\log\left(w_{ik}(\boldsymbol{y}_i,\boldsymbol{U}_i|\,\boldsymbol{\lambda}_i^{(k)},\boldsymbol{\theta}_k^{(s)})\right)\Big|\,\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)}\right].
\end{aligned}
\tag{A.22}
$$

The integration in (A.22) is with respect to q-dimensional Lebesgue measure, and $\boldsymbol{u}$ is a vector in $\mathbb{R}^q$. Using (A.53) we have for any $g \in I_G$

$$
\begin{aligned}
\boldsymbol{E}\left[\log\left(w_{ig}(\boldsymbol{y}_i,\boldsymbol{U}_i|\,\boldsymbol{\lambda}_i^{(g)},\boldsymbol{\theta}_g^{(s)})\right)\Big|\,\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)}\right]=&-\left(\frac{n_i}{2}\right)\log(2\pi)-\left(\frac{n_i}{2}\right)\log(\sigma_g^2)-\frac{1}{2}\log\left(|\boldsymbol{D}_g|\right)\\
&-\frac{1}{2}\log\left(|\boldsymbol{C}_i(\boldsymbol{\phi}_g)|\right)-\frac{1}{2\sigma_g^2}\boldsymbol{E}\left[\boldsymbol{e}_i^\intercal\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{e}_i|\,\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)}\right]\\
&-\frac{1}{2}\boldsymbol{E}\left[\boldsymbol{U}_i^\intercal\boldsymbol{D}_g^{-1}\boldsymbol{U}_i|\,\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)}\right].
\end{aligned}
\tag{A.23}
$$

Let $\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)}=\boldsymbol{E}\left[\boldsymbol{U}_i|\,\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)}\right]$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)}=\mathrm{Var}\left[\boldsymbol{U}_i|\,\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)}\right]$, where from (A.55) and (A.56) these are given by

$$
\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)}=\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)})\boldsymbol{Z}_i^\intercal\boldsymbol{V}_i(\hat{\boldsymbol{\zeta}}_g^{(s)})^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_g^{(s)}),
\tag{A.24}
$$

and

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)} = \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)}) - \boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)})\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\hat{\boldsymbol{\zeta}}_g^{(s)})^{-1}\boldsymbol{Z}_i\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)}), \tag{A.25}$$

where $\boldsymbol{V}_i(\hat{\boldsymbol{\zeta}}_g^{(s)}) = \boldsymbol{Z}_i\boldsymbol{D}(\hat{\boldsymbol{\psi}}_g^{(s)})\boldsymbol{Z}_i^{\intercal} + \hat{\sigma}_g^{2(s)}\boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)})$. Also let $\hat{\boldsymbol{E}}_i^{(s)} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}_i}^{(s)} + \hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)}\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)\intercal}$, where $\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)} = \boldsymbol{E}\left[\boldsymbol{e}_i \middle| \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right]$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}_i}^{(s)} = \mathrm{Var}\left[\boldsymbol{e}_i \middle| \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right]$ which are given by

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)} &= \boldsymbol{E}\left[\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g - \boldsymbol{Z}_i\boldsymbol{U}_i \middle| \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right] \\ &= \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g - \boldsymbol{Z}_i\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)},\end{aligned} \tag{A.26}$$

and

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}_i}^{(s)} &= \mathrm{Cov}\left[\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g - \boldsymbol{Z}_i\boldsymbol{U}_i, \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g - \boldsymbol{Z}_i\boldsymbol{U}_i|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}\right] \\ &= \boldsymbol{Z}_i\mathrm{Cov}\left[\boldsymbol{Z}_i\boldsymbol{U}_i, \boldsymbol{Z}_i\boldsymbol{U}_i|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}\right]\boldsymbol{Z}_i^{\intercal} \\ &= \boldsymbol{Z}_i\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)}\boldsymbol{Z}_i^{\intercal}.\end{aligned} \tag{A.27}$$

Then we have

$$\begin{aligned}\boldsymbol{E}\left[\boldsymbol{e}_i^{\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{e}_i \middle| \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right] &= \mathrm{tr}\left(\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)}\right) + \hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)} \\ &= \mathrm{tr}\left(\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)}\right) + \boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)}\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)\intercal} \\ &= \mathrm{tr}\left\{\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)} + \hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)}\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)\intercal}\right)\right\} \\ &= \mathrm{tr}\left(\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{E}}_i^{(s)}\right),\end{aligned} \tag{A.28}$$

and in exactly the same fashion

$$\boldsymbol{E}\left[\boldsymbol{U}_i^{\intercal}\boldsymbol{D}_g^{-1}\boldsymbol{U}_i \middle| \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right] = \mathrm{tr}\left(\boldsymbol{D}_g^{-1}\hat{\boldsymbol{J}}_i^{(s)}\right), \tag{A.29}$$

where $\hat{\boldsymbol{J}}_i^{(s)} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}_i}^{(s)} + \hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)} \hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)\intercal}$. So if for any $g \in I_G$ we let $\boldsymbol{E}\left[\log\left(f_{ig}(\boldsymbol{y}_i, \boldsymbol{U}_i |\, \boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g^{(s)})\right)\Big|\, \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}\right] = Q_{1ig}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$ then

$$\begin{aligned} Q_{1ig}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) = &-\left(\frac{n_i}{2}\right)\log(2\pi) - \left(\frac{n_i}{2}\right)\log(\sigma_g^2) - \frac{1}{2}\log\left(|\boldsymbol{D}_g|\right) \\ &- \frac{1}{2}\log\left(|\boldsymbol{C}_i(\boldsymbol{\phi}_g)|\right) - \frac{1}{2\sigma_g^2}\mathrm{tr}\left(\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{E}}_i^{(s)}\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{D}_g^{-1}\hat{\boldsymbol{J}}_i^{(s)}\right). \end{aligned} \tag{A.30}$$

So from (A.22), (A.30) and (A.20), the conditional expectation of $L^c(\boldsymbol{\theta}|\boldsymbol{C})$ in (A.19) can be written

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) = Q_1(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) + Q_2(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}), \tag{A.31}$$

where

$$Q_1(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) = \sum_{i=1}^{N}\sum_{k=1}^{G} \hat{p}_i(\boldsymbol{\lambda}_i^{(k)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}) Q_{1ik}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}), \tag{A.32}$$

and

$$Q_2(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}) = \sum_{i=1}^{N}\sum_{k=1}^{G} \hat{p}_i(\boldsymbol{\lambda}_i^{(k)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}) \log(\boldsymbol{\pi}_k). \tag{A.33}$$

For the component density parameters we now find the derivative vectors of $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$ with respect to the components of $\boldsymbol{\theta}$ in turn, set the resultant expressions to zero, and solve for the parameter of interest. To avoid repetition we note that for any $g \in I_G$, if we compute the differential of $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$ with respect to $\boldsymbol{\theta}_g$ we have $\boldsymbol{d}(Q_1(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})) = \sum_{i=1}^{N} \hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)})\boldsymbol{d}(Q_{1ig}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)}))$. Thus we shall compute the differentials of $Q_{1ig}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$

with respect to the components of $\boldsymbol{\theta}$ and then bring in the summation and posterior probabilities at the end of the derivations.

When $\boldsymbol{D_g}(\boldsymbol{\psi}) = \boldsymbol{D}_g$ we need to calculate $\boldsymbol{D}_{\text{vec}(\boldsymbol{D}_g)}\left(\boldsymbol{Q}(\text{vec}(\boldsymbol{D}_g))\right)$. Now $\boldsymbol{Q}(\text{vec}(\boldsymbol{D}_g)) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^{q^2}$, so that by the first identification table (Table B.1) we have that if $\boldsymbol{d}\left(\boldsymbol{Q}(\text{vec}(\boldsymbol{D}_g))\right) = \boldsymbol{a}^\intercal \boldsymbol{d}\left(\text{vec}(\boldsymbol{D}_g)\right)$ for $\boldsymbol{a} \in \mathbb{R}^{q^2}$ then $\boldsymbol{D}_{\text{vec}(\boldsymbol{D}_g)}\left(\boldsymbol{Q}(\text{vec}(\boldsymbol{D}_g))\right) = \boldsymbol{a}^\intercal$. Now for $Q_{1ig}(\text{vec}(\boldsymbol{D}_g))$ we have

$$\begin{aligned}
\boldsymbol{d}\left(Q_{1ig}(\text{vec}(\boldsymbol{D}_g))\right) &= -\frac{1}{2}\boldsymbol{d}\left(\log|\boldsymbol{D}_g|\right) - \frac{1}{2}\text{tr}\left[\boldsymbol{d}\left(\boldsymbol{D}_g^{-1}\right)\hat{\boldsymbol{J}}_i^{(s)}\right] \\
&= -\frac{1}{2}\text{tr}\left[\boldsymbol{d}\left(\boldsymbol{D}_g\right)\boldsymbol{D}_g^{-1}\right] + \frac{1}{2}\text{tr}\left[\boldsymbol{d}\left(\boldsymbol{D}_g\right)^\intercal \boldsymbol{D}_g^{-1}\hat{\boldsymbol{J}}_i^{(s)\intercal}\boldsymbol{D}_g^{-1}\right] \\
&= -\frac{1}{2}\left[\text{vec}(\boldsymbol{D}_g)\right]^\intercal \text{vec}(\boldsymbol{D}_g^{-1}) + \frac{1}{2}\left[\text{vec}(\boldsymbol{D}_g)\right]^\intercal (\boldsymbol{D}_g^{-1}\otimes\boldsymbol{D}_g^{-1})\text{vec}(\hat{\boldsymbol{J}}_i^{(s)}) \\
&= -\frac{1}{2}\text{vec}(\boldsymbol{D}_g^{-1})^\intercal\left[\text{vec}(\boldsymbol{D}_g)\right] + \frac{1}{2}\left[\text{vec}(\hat{\boldsymbol{J}}_i^{(s)})\right]^\intercal (\boldsymbol{D}_g^{-1}\otimes\boldsymbol{D}_g^{-1})\text{vec}(\boldsymbol{D}_g) \\
&= \left\{-\frac{1}{2}\text{vec}(\boldsymbol{D}_g^{-1})^\intercal + \frac{1}{2}\left[\text{vec}(\hat{\boldsymbol{J}}_i^{(s)})\right]^\intercal (\boldsymbol{D}_g^{-1}\otimes\boldsymbol{D}_g^{-1})\right\}\boldsymbol{d}\left(\text{vec}(\boldsymbol{D}_g)\right),
\end{aligned} \tag{A.34}$$

so that

$$\boldsymbol{d}\left(\boldsymbol{Q}(\text{vec}(\boldsymbol{D}_g))\right) = \frac{1}{2}\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\left\{\left[\text{vec}(\hat{\boldsymbol{J}}_i^{(s)})\right]^\intercal(\boldsymbol{D}_g^{-1}\otimes\boldsymbol{D}_g^{-1}) - \text{vec}(\boldsymbol{D}_g^{-1})^\intercal\right\}\boldsymbol{d}\left(\text{vec}(\boldsymbol{D}_g)\right), \tag{A.35}$$

and thus we see that the $1\times q^2$ vector of partial derivatives is

$$\boldsymbol{D}_{\text{vec}(\boldsymbol{D}_g)}\left(\boldsymbol{Q}(\text{vec}(\boldsymbol{D}_g))\right) = \frac{1}{2}\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\left(\left[\text{vec}(\hat{\boldsymbol{J}}_i^{(s)})\right]^\intercal(\boldsymbol{D}_g^{-1}\otimes\boldsymbol{D}_g^{-1}) - \text{vec}(\boldsymbol{D}_g^{-1})^\intercal\right). \tag{A.36}$$

Setting $\boldsymbol{D}_{\text{vec}(\boldsymbol{D}_g)}\left(\boldsymbol{Q}(\text{vec}(\boldsymbol{D}_g))\right)$ in (A.36) to zero, multiplying by 2, and transposing both sides we get

$$\begin{aligned}\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\text{vec}(\boldsymbol{D}_g^{-1}) &= \sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})(\boldsymbol{D}_g^{-1}\otimes\boldsymbol{D}_g^{-1})\text{vec}(\hat{\boldsymbol{J}}_i^{(s)}) \\ &= \sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\text{vec}(\boldsymbol{D}_g^{-1}\hat{\boldsymbol{J}}_i^{(s)}\boldsymbol{D}_g^{-1}), \qquad \text{(A.37)}\end{aligned}$$

and so by un-vectorising both sides we get

$$\begin{aligned}\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\boldsymbol{D}_g^{-1} &= \sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\boldsymbol{D}_g^{-1}\hat{\boldsymbol{J}}_i^{(s)}\boldsymbol{D}_g^{-1} \\ \Longleftrightarrow \sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\boldsymbol{D}_g &= \sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\hat{\boldsymbol{J}}_i^{(s)} \\ \Longleftrightarrow \hat{\boldsymbol{D}}_g^{(s+1)} &= \frac{1}{\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})}\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\hat{\boldsymbol{J}}_i^{(s)}. \qquad \text{(A.38)}\end{aligned}$$

We now derive $\boldsymbol{D}_{\sigma_g^2}\left(\boldsymbol{Q}(\sigma_g^2)\right)$. We have that $\boldsymbol{Q}(\sigma_g^2) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}$, so that by the first identification table (Table B.1), $\boldsymbol{d}\left(\boldsymbol{Q}(\sigma_g^2)\right) = \alpha\boldsymbol{d}\sigma_g^2$ for $\alpha \in \mathbb{R}$ implies that $\boldsymbol{D}_{\sigma_g^2}\left(\boldsymbol{Q}(\sigma_g^2)\right) = \alpha$. For $Q_{1ig}(\sigma_g^2)$ we have $\boldsymbol{d}\left(Q_{1ig}(\sigma_g^2)\right) = -(n_i/2)\boldsymbol{d}\left(\log(\sigma_g^2)\right) - (1/2)\boldsymbol{d}\left(\boldsymbol{\sigma}_g^{-2}\right)\text{tr}(\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{E}}_i^{(s)}) = \{-(n_i/2\sigma_g^2) + (1/2\boldsymbol{\sigma}^4)\text{tr}(\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{E}}_i^{(s)})\}\boldsymbol{d}\sigma_g^2$. So we have $\boldsymbol{d}\left(\boldsymbol{Q}(\sigma_g^2)\right) = \sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\{-(n_i/2\sigma_g^2) + (1/2\boldsymbol{\sigma}_g^4)\text{tr}(\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{E}}_i^{(s)})\}\boldsymbol{d}\sigma_g^2$. Thus the scalar derivative $\boldsymbol{D}_{\sigma_g^2}\left(\boldsymbol{Q}(\sigma_g^2)\right)$ is

$$\boldsymbol{D}_{\sigma_g^2}\left(\boldsymbol{Q}(\sigma_g^2)\right) = \sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\left(\frac{1}{2\boldsymbol{\sigma}_g^4}\text{tr}(\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{E}}_i^{(s)}) - \frac{n_i}{2\sigma_g^2}\right). \qquad \text{(A.39)}$$

Then equating (A.39) to zero and solving for σ_g^2 we get

$$\hat{\boldsymbol{\sigma}}_g^{2(s+1)} = \frac{1}{\sum_{i=1}^{N}n_i\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})}\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\text{tr}\left[\boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)})^{-1}\hat{\boldsymbol{E}}_i^{(s)}\right]. \qquad \text{(A.40)}$$

We will now derive $\boldsymbol{D}_{\boldsymbol{\beta}_g}\left(\boldsymbol{Q}(\boldsymbol{\beta}_g)\right)$. Now $\boldsymbol{Q}(\boldsymbol{\beta}_g) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^p$, so that by the first identification table (Table B.1) we have that if $\boldsymbol{d}\left(\boldsymbol{Q}(\boldsymbol{\beta}_g)\right) = \boldsymbol{a}^{\intercal}\boldsymbol{d}\boldsymbol{\beta}_g$ for $\boldsymbol{a} \in \mathbb{R}^p$ then $\boldsymbol{D}_{\boldsymbol{\beta}_g}\left(\boldsymbol{Q}(\boldsymbol{\beta}_g)\right) = \boldsymbol{a}^{\intercal}$. Now for $Q_{1ig}(\boldsymbol{\beta}_g)$ we have

$$\begin{aligned}
\boldsymbol{d}\left(Q_{1ig}(\boldsymbol{\beta}_g)\right) &= -\frac{1}{2\sigma_g^2}\operatorname{tr}\left[\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{d}\left(\hat{\boldsymbol{E}}_i^{(s)}\right)\right] \\
&= -\frac{1}{2\sigma_g^2}\operatorname{tr}\left[-\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{X}_i\boldsymbol{d}(\boldsymbol{\beta}_g)\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)\intercal} - \boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)}\boldsymbol{d}(\boldsymbol{\beta}_g)^{\intercal}\boldsymbol{X}_i^{\intercal}\right] \\
&= \frac{1}{2\sigma_g^2}\operatorname{tr}\left[\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{X}_i\boldsymbol{d}(\boldsymbol{\beta}_g)\right] \\
&= \frac{1}{2\sigma_g^2}\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{X}_i\boldsymbol{d}(\boldsymbol{\beta}_g), \qquad \text{(A.41)}
\end{aligned}$$

so that

$$\boldsymbol{d}\left(\boldsymbol{Q}(\boldsymbol{\beta}_g)\right) = \left\{\frac{1}{2\sigma_g^2}\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{X}_i\right\}\boldsymbol{d}(\boldsymbol{\beta}_g). \qquad \text{(A.42)}$$

Thus the $1 \times p$ vector of partial derivatives $\boldsymbol{D}_{\boldsymbol{\beta}_g}\left(\boldsymbol{Q}(\boldsymbol{\beta}_g)\right)$ is

$$\boldsymbol{D}_{\boldsymbol{\beta}_g}\left(\boldsymbol{Q}(\boldsymbol{\beta}_g)\right) = \frac{1}{2\sigma_g^2}\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\hat{\boldsymbol{\mu}}_{\boldsymbol{e}_i}^{(s)\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{X}_i. \qquad \text{(A.43)}$$

So on setting (A.43) to zero, multiplying by $2\sigma_g^2$ and transposing both sides we get

$$\begin{aligned}
&\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\boldsymbol{X}_i^{\intercal}\boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g) = \sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\boldsymbol{X}_i^{\intercal}\boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)})^{-1}\boldsymbol{Z}_i\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)} \\
\Longleftrightarrow \hat{\boldsymbol{\beta}}_g^{(s+1)} &= \left(\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\boldsymbol{X}_i^{\intercal}\boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)})^{-1}\boldsymbol{X}_i\right)^{-1} \\
&\quad \times \left[\sum_{i=1}^{N}\hat{p}_i(\boldsymbol{\lambda}_i^{(g)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\boldsymbol{X}_i^{\intercal}\boldsymbol{C}_i(\hat{\boldsymbol{\phi}}_g^{(s)})^{-1}\left(\boldsymbol{y}_i - \boldsymbol{Z}_i\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i}^{(s)}\right)\right]. \qquad \text{(A.44)}
\end{aligned}$$

We will maximise $Q_2(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(s)})$ in (A.33) with respect to $\boldsymbol{\pi}_g$, $g \in I_G$, by finding the stationary values of the Lagrange function $l(\boldsymbol{\pi},\kappa)$ given by

$$l(\boldsymbol{\pi},\kappa) = \sum_{i=1}^{N}\sum_{k=1}^{G}\hat{p}_i(\boldsymbol{\lambda}_i^{(k)}|\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}^{(s)})\log(\boldsymbol{\pi}_k) - \kappa\left(\sum_{k'=1}^{G}\boldsymbol{\pi}_{k'} - 1\right). \qquad \text{(A.45)}$$

Differentiating l with respect to $\boldsymbol{\pi}_g$ we get $\frac{\partial l(\boldsymbol{\pi}_g)}{\partial \boldsymbol{\pi}_g} = \boldsymbol{\pi}_g^{-1} \sum_{i=1}^{N} \hat{p}_i(\boldsymbol{\lambda}_i^{(g)} | \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}) - \kappa$. Setting this equation equal to zero and summing both sides over $j = 1, ..., G$, implies that $\kappa = \sum_{j=1}^{G} \boldsymbol{\pi}_j = 1$. Substituting this into the original equation and solving for $\boldsymbol{\pi}_g$ gives

$$\hat{\boldsymbol{\pi}_g}^{(s+1)} = \frac{1}{N} \sum_{i=1}^{N} \hat{p}_i(\boldsymbol{\lambda}_i^{(g)} | \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(s)}), \tag{A.46}$$

A.3.1 Some distributional results

Here we show the derivation of the distribution of $\boldsymbol{Y}_i$ conditional on $\boldsymbol{\lambda}_i^{(g)}$ given in (2.6). We do this by deriving the joint distribution of $\left(\boldsymbol{Y}_i^{\intercal}, \boldsymbol{U}_i^{\intercal}\right)^{\intercal}$ conditional on $\boldsymbol{\lambda}_i^{(g)}$. Letting

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{I}_{n_i} & \boldsymbol{Z}_i \\ \boldsymbol{0} & \boldsymbol{I}_q \end{bmatrix},$$

$\boldsymbol{s} = \left[\boldsymbol{e}_i^{\intercal}, \boldsymbol{U}_i^{\intercal}\right]^{\intercal}$ and $\boldsymbol{t} = \left[\boldsymbol{\beta}_g^{\intercal} \boldsymbol{X}_i^{\intercal}, \boldsymbol{0}\right]^{\intercal}$ then the joint vector can be written

$$\begin{aligned} \begin{bmatrix} \boldsymbol{Y}_i \\ \boldsymbol{U}_i \end{bmatrix} &= \begin{bmatrix} \boldsymbol{I}_{n_i} & \boldsymbol{Z}_i \\ \boldsymbol{0} & \boldsymbol{I}_q \end{bmatrix} \begin{bmatrix} \boldsymbol{e}_i \\ \boldsymbol{U}_i \end{bmatrix} + \begin{bmatrix} \boldsymbol{X}_i \boldsymbol{\beta}_g \\ \boldsymbol{0} \end{bmatrix} \\ &= \boldsymbol{C}\boldsymbol{s} + \boldsymbol{t}, \end{aligned} \tag{A.47}$$

where $\boldsymbol{C}$ is the $(n_i + q) \times (n_i + q)$ matrix with elements $\boldsymbol{I}_{n_i}$, $\boldsymbol{Z}_i$, $\boldsymbol{0}$ and $\boldsymbol{I}_q$, $\boldsymbol{s} \in \mathbb{R}^{n_i+q}$ is the random vector $(\boldsymbol{e}_i^{\intercal}, \boldsymbol{U}_i^{\intercal})^{\intercal}$, and $\boldsymbol{t} \in \mathbb{R}^{n_i+q}$ is the fixed vector $((\boldsymbol{X}_i\boldsymbol{\beta}_g)^{\intercal}, \boldsymbol{0}^{\intercal})^{\intercal}$. Now for any $\boldsymbol{a}_1 \in \mathbb{R}^{n_i}$ and $\boldsymbol{a}_2 \in \mathbb{R}^{q}$, $\boldsymbol{a}_1^{\intercal}\boldsymbol{e}_i$ and $\boldsymbol{a}_2^{\intercal}\boldsymbol{U}_i$ are, conditional on $\boldsymbol{\lambda}_i^{(g)}$, distributed as independent univariate normal random variables. If we let $\boldsymbol{a} = \left[\boldsymbol{a}_1^{\intercal}, \boldsymbol{a}_2^{\intercal}\right]^{\intercal} \in \mathbb{R}^{n_i+q}$

then $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{s}$ is the sum of two univariate independent normal random variables and so is itself a univariate normal random variable. Then we have $\boldsymbol{s}|\boldsymbol{\lambda}_i^{(g)} \sim N_{n_i+q}\left(\boldsymbol{\mu_s}, \boldsymbol{\Sigma_s}\right)$, where $\boldsymbol{\mu_s} = \boldsymbol{E}[\boldsymbol{s}] = \boldsymbol{0}$ and $\boldsymbol{\Sigma_s} = \mathrm{Var}[\boldsymbol{s}]$ is a block-diagonal matrix with diagonal elements $(\boldsymbol{\Sigma_s})_{1,1} = \sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g)$ and $(\boldsymbol{\Sigma_s})_{2,2} = \boldsymbol{D}_g$, where $\mathrm{rank}(\boldsymbol{\Sigma_s}) = \mathrm{rank}(\sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g)) + \mathrm{rank}(\boldsymbol{D}_g) = n_i + q$, by Schott (2005, Theorem 2.12, pp48). Then from Seber and Lee (2003, Theorem 2.2, pp20) we have $\left(\boldsymbol{Y}_i^{\mathsf{T}}, \boldsymbol{U}_i^{\mathsf{T}}\right)^{\mathsf{T}}|\boldsymbol{\lambda}_i^{(g)} \sim N_{n_i+q}\left(\boldsymbol{C}\boldsymbol{\mu_s} + \boldsymbol{t}, \boldsymbol{C}\boldsymbol{\Sigma_s}\boldsymbol{C}^{\mathsf{T}}\right)$ so that

$$\left.\begin{bmatrix} \boldsymbol{Y}_i \\ \boldsymbol{U}_i \end{bmatrix}\right|_{\boldsymbol{\lambda}_i^{(g)}} \sim N_{n_i+q}\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta}_g \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{Z}_i\boldsymbol{D}_g\boldsymbol{Z}_i^{\mathsf{T}} + \sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g) & \boldsymbol{Z}_i\boldsymbol{D}_g \\ \boldsymbol{D}_g\boldsymbol{Z}_i^{\mathsf{T}} & \boldsymbol{D}_g \end{bmatrix}\right). \tag{A.48}$$

Letting $\boldsymbol{\theta_g} = \left[\boldsymbol{\beta}_g^{\mathsf{T}}, \sigma_g^2, \boldsymbol{\phi}_g^{\mathsf{T}}, \boldsymbol{\psi}_g^{\mathsf{T}}\right]^{\mathsf{T}}$ then we shall write the density for the joint distribution of $\left(\boldsymbol{Y}_i^{\mathsf{T}}, \boldsymbol{U}_i^{\mathsf{T}}\right)^{\mathsf{T}}$ conditional on $\boldsymbol{\lambda}_i^{(g)}$ given by A.48 as $w_{ig}(\boldsymbol{y}_i, \boldsymbol{u}_i|\,\boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g)$. Now let $\boldsymbol{\zeta}_g = \left[\sigma_g^2, \boldsymbol{\phi}_g^{\mathsf{T}}, \boldsymbol{\psi}_g^{\mathsf{T}}\right]^{\mathsf{T}}$, then using (A.48) and standard multivariate normal theory, we immediately see that the distribution for $\boldsymbol{Y}_i$ conditional on $\boldsymbol{\lambda}_i^{(g)}$ is as given in (2.6).

Now let $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i,\boldsymbol{U}_i}$ be the covariance matrix of $\left(\boldsymbol{Y}_i^{\mathsf{T}}, \boldsymbol{U}_i^{\mathsf{T}}\right)^{\mathsf{T}}$ in (A.48). We will now derive explicit forms for $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i,\boldsymbol{U}_i}^{-1}$ and $|\boldsymbol{\Sigma}_{\boldsymbol{Y}_i,\boldsymbol{U}_i}|$, which we will use to write down the joint density $w_{ig}(\boldsymbol{y}_i, \boldsymbol{u}_i|\,\boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g)$. This density will be needed for the complete data density used by the EM algorithm. Let

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}_i,\boldsymbol{U}_i} = \begin{bmatrix} \boldsymbol{Z}_i\boldsymbol{D}_g\boldsymbol{Z}_i^{\mathsf{T}} + \sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g) & \boldsymbol{Z}_i\boldsymbol{D}_g \\ \boldsymbol{D}_g\boldsymbol{Z}_i^{\mathsf{T}} & \boldsymbol{D}_g \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}, \tag{A.49}$$

then using Schott (2005, Theorem 7.1, pp256)

$$\boldsymbol{\Sigma}_{\mathbf{Y}_i,\boldsymbol{U}_i}^{-1} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix}, \tag{A.50}$$

where

$$\begin{aligned}
\boldsymbol{B}_{11} &= \left(\boldsymbol{Z}_i\boldsymbol{D}_g\boldsymbol{Z}_i^{\intercal} + \sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g) - \boldsymbol{Z}_i\boldsymbol{D}_g\boldsymbol{D}_g^{-1}\boldsymbol{D}_g\boldsymbol{Z}_i^{\intercal}\right)^{-1} \\
&= \boldsymbol{\sigma}_g^{-2}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}, \\
\boldsymbol{B}_{12} &= -\boldsymbol{B}_{11}\boldsymbol{Z}_i\boldsymbol{D}_g\boldsymbol{D}_g^{-1} \\
&= -\boldsymbol{\sigma}_g^{-2}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{Z}_i, \\
\boldsymbol{B}_{21} &= -\boldsymbol{D}_g^{-1}\boldsymbol{D}_g\boldsymbol{Z}_i^{\intercal}\boldsymbol{B}_{11} \\
&= -\boldsymbol{\sigma}_g^{-2}\boldsymbol{Z}_i^{\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}, \\
\boldsymbol{B}_{22} &= \boldsymbol{D}_g^{-1} + \boldsymbol{D}_g^{-1}\boldsymbol{D}_g\boldsymbol{Z}_i^{\intercal}\boldsymbol{B}_{11}\boldsymbol{Z}_i\boldsymbol{D}_g\boldsymbol{D}_g^{-1} \\
&= \boldsymbol{D}_g^{-1} + \boldsymbol{\sigma}_g^{-2}\boldsymbol{Z}_i^{\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{Z}_i.
\end{aligned}$$

Thus we have

$$\boldsymbol{\Sigma}_{\mathbf{Y}_i,\boldsymbol{U}_i}^{-1} = \begin{bmatrix} \boldsymbol{\sigma}_g^{-2}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1} & -\boldsymbol{\sigma}_g^{-2}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{Z}_i \\ -\boldsymbol{\sigma}_g^{-2}\boldsymbol{Z}_i^{\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1} & \boldsymbol{D}_g^{-1} + \boldsymbol{\sigma}_g^{-2}\boldsymbol{Z}_i^{\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{Z}_i \end{bmatrix}. \tag{A.51}$$

From Schott (2005, Theorem 7.4, pp259)

$$\begin{aligned}
\left|\boldsymbol{\Sigma}_{\mathbf{Y}_i,\boldsymbol{U}_i}\right| &= \left|\boldsymbol{D}_g\right| \left|\boldsymbol{Z}_i\boldsymbol{D}_g\boldsymbol{Z}_i^{\intercal} + \sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g) - \boldsymbol{Z}_i\boldsymbol{D}_g\boldsymbol{D}_g^{-1}\boldsymbol{D}_g\boldsymbol{Z}_i^{\intercal}\right| \\
&= \left|\boldsymbol{D}_g\right| \left|\sigma_g^2\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right| \\
&= \left(\sigma_g^2\right)^{n_i} \left|\boldsymbol{D}_g\right| \left|\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right|.
\end{aligned} \tag{A.52}$$

Using (A.51) and (A.52) we get

$$w_{ig}(\boldsymbol{y}_i, \boldsymbol{u}_i|\,\boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g) = (2\pi)^{-\frac{n_i}{2}} \left|\boldsymbol{\Sigma}_{\mathbf{Y}_i,\boldsymbol{U}_i}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i^{\intercal} - \boldsymbol{\beta}_g^{\intercal}\boldsymbol{X}_i^{\intercal}, \boldsymbol{u}_i^{\intercal}\right)\boldsymbol{\Sigma}_{\mathbf{Y}_i,\boldsymbol{U}_i}^{-1}\left[\begin{array}{c}\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g\\ \\ \boldsymbol{u}_i\end{array}\right]\right\}$$

$$= (2\pi)^{-\frac{n_i}{2}}\left(\sigma_g^2\right)^{-\frac{n_i}{2}}\left|\boldsymbol{D}_g\right|^{-\frac{1}{2}}\left|\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right|^{-\frac{1}{2}} \times$$

$$\exp\left\{-\frac{\boldsymbol{\sigma}_g^{-2}}{2}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g - \boldsymbol{Z}_i\boldsymbol{u}_i\right)^{\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g - \boldsymbol{Z}_i\boldsymbol{u}_i\right)\right.$$

$$\left.-\frac{1}{2}\boldsymbol{u}_i^{\intercal}\boldsymbol{D}_g^{-1}\boldsymbol{u}_i\right\}$$

$$= (2\pi)^{-\frac{n_i}{2}}\left(\sigma_g^2\right)^{-\frac{n_i}{2}}\left|\boldsymbol{D}_g\right|^{-\frac{1}{2}}\left|\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_g^2}\boldsymbol{e}_i{}^{\intercal}\boldsymbol{C}_i(\boldsymbol{\phi}_g)^{-1}\boldsymbol{e}_i\right.$$

$$\left.-\frac{1}{2}\boldsymbol{u}_i^{\intercal}\boldsymbol{D}_g^{-1}\boldsymbol{u}_i\right\}, \tag{A.53}$$

where $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g - \boldsymbol{Z}_i\boldsymbol{u}_i$.

We can use the above results and standard multivariate normal distribution theory to calculate the mean and covariance matrix of t_{ig}, the density function of $\boldsymbol{U}_i|\boldsymbol{Y}_i, \boldsymbol{\Lambda}_i$. That is from using (A.48) we see $t_{ig}(\boldsymbol{u}_i|\,\boldsymbol{y}_i, \boldsymbol{\lambda}_i^{(g)}, \boldsymbol{\theta}_g)$ is the density function of the random variable where

$$\boldsymbol{U}_i|\boldsymbol{y}_i, \boldsymbol{\lambda}_i^{(g)} \sim N_q\left(\boldsymbol{\mu}(\boldsymbol{\theta}_g), \boldsymbol{\Sigma}(\boldsymbol{\zeta}_g)\right), \tag{A.54}$$

where

$$\boldsymbol{\mu}(\boldsymbol{\theta}_g) = \boldsymbol{D}_g\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g), \tag{A.55}$$

and

$$\boldsymbol{\Sigma}(\boldsymbol{\zeta}_g) = \boldsymbol{D}_g - \boldsymbol{D}_g\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i\boldsymbol{D}_g. \tag{A.56}$$

A.4 Simulation results

A.4.1 Model 1

Table A.1: EM1st variant simulation results for CON

parm	est avg	est std	est mse	SE1 avg	SE1 std	CI1 CP	SE2 avg	SE2 std	CI2 CP	SE3 avg	SE3 std	CI3 CP	SE4 avg	SE4 std	CI4 CP	CI1BLCL	CI1BUCL	CI2BLCL	CI2BUCL	CI3BLCL	CI3BUCL	CI4BLCL	CI4BUCL
β_1^0	12.3761	0.8947	0.8011	0.8498	0.2837	0.9665	0.6412	0.1182	0.9242	0.6149	0.1196	0.9213	0.7617	0.3848	0.9203	0.9555	0.9776	0.9079	0.9405	0.9047	0.9378	0.9036	0.9369
β_1^{c1}	-1.0002	0.0950	0.0090	0.0175	0.0039	0.9734	0.0132	0.0015	0.9419	0.0132	0.0015	0.9429	0.0139	0.0127	0.9242	0.9635	0.9833	0.9276	0.9563	0.9287	0.9572	0.9079	0.9405
β_1^{c2}	-0.7489	0.1660	0.0276	0.1369	0.0440	0.9774	0.1025	0.0195	0.9291	0.0974	0.0196	0.9262	0.1245	0.0661	0.9252	0.9682	0.9865	0.9134	0.9449	0.9101	0.9423	0.9090	0.9414
β_1^{f11}	1.9997	0.5423	0.2941	0.6614	0.1808	0.9724	0.5065	0.0862	0.9350	0.4849	0.0879	0.9232	0.6158	0.2679	0.9282	0.9624	0.9825	0.9199	0.9502	0.9069	0.9396	0.9123	0.9440
β_1^{f21}	-2.3730	0.6636	0.4411	0.8099	0.2620	0.9705	0.6088	0.1066	0.9282	0.5852	0.1085	0.9232	0.7257	0.3335	0.9134	0.9601	0.9809	0.9123	0.9440	0.9069	0.9396	0.8961	0.9307
β_1^{f22}	-1.0842	0.6682	0.4467	0.8203	0.3092	0.9715	0.6135	0.1073	0.9272	0.5886	0.1089	0.9203	0.7382	0.3316	0.9173	0.9612	0.9817	0.9112	0.9431	0.9036	0.9369	0.9004	0.9343
β_1^{tc}	0.4998	0.0433	0.0019	0.0172	0.0051	0.9803	0.0146	0.0022	0.9606	0.0132	0.0021	0.9459	0.0169	0.0061	0.9695	0.9718	0.9889	0.9487	0.9726	0.9320	0.9598	0.9589	0.9801
d_1^{11}	1.4059	0.4616	0.2996	0.6802	0.9309	0.8917	0.4857	0.1505	0.8701	0.4598	0.7125	0.7845	0.8598	1.9820	0.8908	0.8726	0.9108	0.8494	0.8908	0.7592	0.8097	0.8716	0.9099
σ_1^2	1.4224	1.7796	3.1819	0.2699	1.0047	0.9813	0.2372	0.1404	0.9724	0.1958	0.7502	0.9488	0.6979	2.2105	0.9557	0.9730	0.9896	0.9624	0.9825	0.9353	0.9624	0.9431	0.9684
ϕ_1^1	0.6029	0.0421	0.0018	0.0573	0.0147	0.9852	0.0437	0.0096	0.9518	0.0418	0.0030	0.9528	0.0651	0.0970	0.9518	0.9778	0.9927	0.9386	0.9650	0.9397	0.9658	0.9386	0.9650
ϕ_1^2	0.1013	0.0492	0.0024	0.0659	0.0200	0.9823	0.0528	0.0232	0.9341	0.0483	0.0036	0.9508	0.0840	0.2044	0.9390	0.9742	0.9904	0.9188	0.9493	0.9375	0.9641	0.9243	0.9537
ϕ_1^3	-0.0493	0.0459	0.0021	0.0599	0.0132	0.9852	0.0488	0.0149	0.9409	0.0448	0.0032	0.9557	0.0782	0.1453	0.9537	0.9778	0.9927	0.9265	0.9554	0.9431	0.9684	0.9408	0.9667
β_2^0	-5.9947	0.7664	0.5874	0.9447	0.3059	0.9636	0.6427	0.1293	0.8858	0.6800	0.1256	0.9154	0.7276	1.0169	0.8494	0.9521	0.9751	0.8663	0.9054	0.8982	0.9325	0.8274	0.8714
β_2^{c1}	1.4996	0.0134	0.0002	0.0177	0.0041	0.9833	0.0133	0.0017	0.9488	0.0132	0.0011	0.9498	0.0138	0.0107	0.9232	0.9754	0.9912	0.9353	0.9624	0.9364	0.9632	0.9069	0.9396
β_2^{c2}	3.0047	0.1236	0.0153	0.1553	0.0472	0.9784	0.1032	0.0225	0.8750	0.1082	0.0201	0.9065	0.1156	0.1205	0.8484	0.9694	0.9873	0.8547	0.8953	0.8886	0.9244	0.8264	0.8705
β_2^{f11}	4.0129	0.5919	0.3505	0.7472	0.2130	0.9793	0.5083	0.1030	0.8858	0.5330	0.0901	0.9144	0.5805	0.6008	0.8721	0.9706	0.9881	0.8663	0.9054	0.8972	0.9316	0.8515	0.8926
β_2^{f21}	-5.0357	0.6826	0.4672	0.9008	0.3010	0.9803	0.6073	0.1206	0.9114	0.6406	0.1110	0.9341	0.7143	1.7382	0.8898	0.9718	0.9889	0.8940	0.9289	0.9188	0.9493	0.8705	0.9090
β_2^{f22}	-2.0084	0.7210	0.5199	0.8989	0.2828	0.9724	0.6147	0.1239	0.8967	0.6457	0.1122	0.9232	0.7185	1.3175	0.8750	0.9624	0.9825	0.8779	0.9154	0.9069	0.9396	0.8547	0.8953
β_2^{tc}	1.5013	0.0156	0.0002	0.0195	0.0044	0.9734	0.0148	0.0029	0.9291	0.0148	0.0018	0.9321	0.0160	0.0186	0.9114	0.9635	0.9833	0.9134	0.9449	0.9166	0.9476	0.8940	0.9289
d_2^{11}	1.6065	0.5632	0.4721	0.7806	0.2835	0.8976	0.4862	0.1619	0.7805	0.5217	0.1395	0.7746	1.1487	8.7328	0.7569	0.8790	0.9163	0.7551	0.8060	0.7489	0.8003	0.7305	0.7833
σ_2^2	1.5112	0.2236	0.0501	0.3296	0.1683	0.9705	0.2423	0.1540	0.8947	0.2300	0.0852	0.9400	1.1692	15.5466	0.9173	0.9601	0.9809	0.8758	0.9136	0.9254	0.9546	0.9004	0.9343
ϕ_2^1	0.5792	0.0408	0.0017	0.0583	0.0139	0.9892	0.0443	0.0125	0.9537	0.0420	0.0031	0.9498	0.0651	0.1700	0.9547	0.9828	0.9955	0.9408	0.9667	0.9364	0.9632	0.9419	0.9675
ϕ_2^2	0.2178	0.0499	0.0025	0.0654	0.0166	0.9734	0.0523	0.0205	0.9380	0.0477	0.0036	0.9409	0.0738	0.1658	0.9360	0.9635	0.9833	0.9232	0.9528	0.9265	0.9554	0.9210	0.9511
ϕ_2^3	-0.0715	0.0439	0.0019	0.0615	0.0148	0.9862	0.0495	0.0234	0.9488	0.0451	0.0033	0.9498	0.0818	0.3134	0.9597	0.9791	0.9934	0.9353	0.9624	0.9364	0.9632	0.9476	0.9718
β_3^0	3.3724	0.9561	0.9149	1.0610	5.1360	0.9685	0.6540	0.1267	0.9134	0.6414	0.1195	0.9272	0.7885	0.8323	0.8947	0.9578	0.9792	0.8961	0.9307	0.9112	0.9431	0.8758	0.9136
β_3^{c1}	-2.9852	0.2480	0.0617	0.0194	0.0591	0.9734	0.0133	0.0012	0.9252	0.0132	0.0015	0.9301	0.0144	0.0215	0.9124	0.9635	0.9833	0.9090	0.9414	0.9144	0.9458	0.8950	0.9298
β_3^{c2}	1.0021	0.1624	0.0264	0.1515	0.1058	0.9764	0.1034	0.0198	0.9104	0.1033	0.0196	0.9222	0.1232	0.1345	0.8898	0.9670	0.9857	0.8929	0.9280	0.9058	0.9387	0.8705	0.9090
β_3^{f11}	5.9779	0.5770	0.3334	0.8559	5.2065	0.9724	0.5080	0.0863	0.9163	0.5035	0.0867	0.9095	0.6070	0.4935	0.9114	0.9624	0.9825	0.8993	0.9334	0.8918	0.9271	0.8940	0.9289
β_3^{f21}	-0.3968	0.7106	0.5049	0.8669	0.7828	0.9813	0.6142	0.1102	0.9095	0.6028	0.1068	0.9095	0.7332	0.8237	0.9026	0.9730	0.9896	0.8918	0.9271	0.8918	0.9271	0.8843	0.9208
β_3^{f22}	-0.1712	0.6949	0.4838	0.8709	0.6630	0.9656	0.6189	0.1098	0.8967	0.6114	0.1119	0.9006	0.7197	0.3378	0.8957	0.9543	0.9768	0.8779	0.9154	0.8822	0.9190	0.8769	0.9145
β_3^{tc}	-0.4937	0.1094	0.0120	0.0216	0.0119	0.9784	0.0146	0.0022	0.9075	0.0160	0.0022	0.9360	0.0146	0.0165	0.8622	0.9694	0.9873	0.8897	0.9253	0.9210	0.9511	0.8410	0.8834
d_3^{11}	1.2378	0.5387	0.4213	0.7733	0.5150	0.9193	0.4873	0.1616	0.8199	0.5122	0.4188	0.8081	0.6252	1.3220	0.8140	0.9025	0.9360	0.7963	0.8435	0.7839	0.8323	0.7901	0.8379
σ_3^2	1.7490	1.0197	1.0421	0.4664	0.5170	0.9547	0.2337	0.1563	0.8012	0.3255	0.4455	0.9055	0.3805	1.2120	0.8799	0.9419	0.9675	0.7766	0.8257	0.8875	0.9235	0.8599	0.8999
ϕ_3^1	0.6191	0.0453	0.0021	0.0647	0.1752	0.9843	0.0438	0.0121	0.9360	0.0425	0.0038	0.9469	0.0694	0.2476	0.9577	0.9766	0.9919	0.9210	0.9511	0.9331	0.9606	0.9453	0.9701
ϕ_3^2	0.1747	0.0503	0.0026	0.0708	0.0909	0.9784	0.0512	0.0139	0.9272	0.0491	0.0043	0.9341	0.0741	0.1895	0.9272	0.9694	0.9873	0.9112	0.9431	0.9188	0.9493	0.9112	0.9431
ϕ_3^3	-0.0314	0.0490	0.0024	0.0648	0.0558	0.9793	0.0487	0.0286	0.9262	0.0455	0.0040	0.9272	0.0964	0.6302	0.9439	0.9706	0.9881	0.9101	0.9423	0.9112	0.9431	0.9298	0.9581
π_1	0.3386	0.0490	0.0024	0.0472	0.0038	0.9626	0.0470	0.0018	0.9616				0.0481	0.0325	0.9626	0.9509	0.9743	0.9498	0.9734			0.9509	0.9743
π_2	0.3313	0.0467	0.0022	0.0469	0.0028	0.9557	0.0468	0.0019	0.9537				0.0505	0.0696	0.9557	0.9431	0.9684	0.9408	0.9667			0.9431	0.9684
π_3	0.3301	0.0481	0.0023	0.0470	0.0043	0.9626	0.0467	0.0022	0.9606				0.0494	0.0616	0.9626	0.9509	0.9743	0.9487	0.9726			0.9509	0.9743

Table A.1 continued

parm	CI1L avg	CI1L std	CI1L n	CI2L avg	CI2L std	CI2L n	CI3L avg	CI3L std	CI3L n	CI4L avg	CI4L std	CI4L n
β_1^0	17.6951	0.9939	1016	17.6118	1.0140	1016	17.6052	1.0048	1016	17.6804	0.9675	1016
β_1^{c1}	1.4195	0.0800	1016	1.4191	0.0800	1016	1.4191	0.0801	1016	1.4194	0.0831	1016
β_1^{c2}	1.1417	0.1789	1016	1.1084	0.1785	1016	1.1052	0.1764	1016	1.1356	0.1996	1016
β_1^{f11}	3.4166	0.7246	1016	3.1843	0.6887	1016	3.1553	0.7039	1016	3.3750	0.8126	1016
β_1^{f21}	4.1090	0.9090	1016	3.7924	0.8339	1016	3.7628	0.8398	1016	4.0078	0.9902	1016
β_1^{f22}	2.8790	0.9274	1016	2.4121	0.6372	1016	2.3597	0.6548	1016	2.7098	0.9657	1016
β_1^{tc}	0.7090	0.0554	1016	0.7081	0.0599	1016	0.7080	0.0585	1016	0.7087	0.0589	1016
d_1^{11}	2.8059	2.5921	1016	2.4415	0.6395	1015	2.4175	2.0149	1016	3.3710	5.3733	1014
σ_1^2	2.1786	3.7349	1016	2.1475	2.5216	1015	2.1016	3.2530	1016	3.1841	6.4467	1010
ϕ_1^1	0.8682	0.0601	1016	0.8616	0.0591	1012	0.8605	0.0591	1016	0.8887	0.2134	1015
ϕ_1^2	0.2407	0.0622	1016	0.2137	0.0731	1002	0.2030	0.0467	1016	0.2956	0.5595	1016
ϕ_1^3	0.1906	0.0406	1016	0.1636	0.0486	1002	0.1537	0.0303	1016	0.2411	0.4001	1015
β_2^0	8.9165	1.0586	1016	8.6735	1.0588	1016	8.6955	1.0537	1016	8.8423	2.6219	1016
β_2^{c1}	2.1214	0.0189	1016	2.1211	0.0189	1016	2.1211	0.0189	1016	2.1213	0.0197	1015
β_2^{c2}	4.2731	0.1739	1016	4.2594	0.1743	1016	4.2603	0.1743	1016	4.2712	0.2415	1016
β_2^{f11}	6.0735	0.8117	1016	5.8579	0.8122	1016	5.8738	0.8074	1016	5.9865	1.5598	1016
β_2^{f21}	7.5897	0.9880	1016	7.3285	0.9423	1016	7.3497	0.9373	1016	7.5569	4.6557	1016
β_2^{f22}	3.8799	0.9358	1016	3.3729	0.8667	1016	3.4221	0.8343	1016	3.6322	3.6353	1016
β_2^{tc}	2.1239	0.0220	1016	2.1236	0.0220	1016	2.1236	0.0220	1016	2.1242	0.0261	1016
d_2^{11}	3.1724	1.0156	1016	2.6909	0.7659	1013	2.7039	0.8519	1016	4.5264	24.1116	1014
σ_2^2	2.3432	0.4785	1016	2.2667	0.4020	1012	2.2343	0.3708	1016	4.7093	43.0109	1014
ϕ_2^1	0.8358	0.0573	1016	0.8286	0.0595	1011	0.8275	0.0570	1016	0.8593	0.4363	1013
ϕ_2^2	0.3612	0.0659	1016	0.3443	0.0754	1009	0.3368	0.0633	1016	0.3937	0.4449	1014
ϕ_2^3	0.2064	0.0473	1016	0.1794	0.0704	1010	0.1686	0.0369	1016	0.2649	0.8659	1014
β_3^0	5.9432	14.1621	1016	5.1878	1.0348	1016	5.1752	1.0307	1016	5.4339	2.2684	1016
β_3^{c1}	4.2383	0.1132	1016	4.2346	0.1235	1016	4.2346	0.1234	1016	4.2353	0.1145	1016
β_3^{c2}	1.4866	0.3368	1016	1.4475	0.2259	1016	1.4475	0.2249	1016	1.4727	0.3843	1016
β_3^{f11}	9.1369	14.2319	1016	8.5749	0.8040	1016	8.5733	0.7994	1016	8.6811	1.2147	1016
β_3^{f21}	2.6428	2.1957	1016	2.0089	0.5294	1016	1.9782	0.5393	1016	2.3237	2.2951	1016
β_3^{f22}	2.6121	1.8454	1016	1.9682	0.4318	1016	1.9465	0.4512	1016	2.2381	0.9334	1016
β_3^{tc}	0.7134	0.0835	1016	0.7117	0.0808	1016	0.7119	0.0803	1016	0.7123	0.0879	1016
d_3^{11}	2.8407	1.4855	1016	2.2782	0.6987	1013	2.3098	1.2933	1016	2.6770	3.5933	1015
σ_3^2	2.8288	1.9798	1016	2.5758	1.4695	1014	2.6467	1.8791	1016	2.8593	3.5230	1014
ϕ_3^1	0.9070	0.4648	1016	0.8844	0.0647	1008	0.8836	0.0625	1016	0.9261	0.6488	1016
ϕ_3^2	0.3243	0.2508	1016	0.2887	0.0647	1007	0.2850	0.0598	1016	0.3461	0.5140	1015
ϕ_3^3	0.1971	0.1554	1016	0.1564	0.0820	1006	0.1480	0.0291	1016	0.2870	1.7455	1016
π_1	0.4967	0.0684	999	0.4965	0.0681	1016				0.4992	0.1040	1016
π_2	0.4865	0.0644	999	0.4863	0.0648	1016				0.4957	0.1868	1016
π_3	0.4852	0.0651	999	0.4847	0.0670	1016				0.4919	0.1689	1016

Table A.2: EM1st variant simulation results for NCON

parm	est avg	est std	est mse	SE1 avg	SE1 std	CI1 CP	SE2 avg	SE2 std	CI2 CP	SE3 avg	SE3 std	CI3 CP	SE4 avg	SE4 std	CI4 CP	CI1BLCL	CI1BUCL	CI2BLCL	CI2BUCL	CI3BLCL	CI3BUCL	CI4BLCL	CI4BUCL
β_1^0	12.2601	1.6168	2.6334	0.3862	0.8689	0.9680	0.2707	0.0468	0.9338	0.2545	0.0476	0.9292	0.3051	0.1310	0.9224	0.9516	0.9845	0.9105	0.9571	0.9052	0.9532	0.8973	0.9474
β_1^{c1}	-0.9895	0.2199	0.0485	0.0247	0.1215	0.9612	0.0131	0.0035	0.9178	0.0129	0.0029	0.9155	0.0141	0.0131	0.9041	0.9431	0.9793	0.8921	0.9435	0.8895	0.9416	0.8765	0.9317
β_1^{c2}	-0.7195	0.3172	0.1016	0.0216	0.0662	0.9726	0.0137	0.0034	0.9224	0.0134	0.0031	0.9338	0.0152	0.0159	0.9132	0.9573	0.9879	0.8973	0.9474	0.9105	0.9571	0.8869	0.9396
β_1^{f11}	2.0282	0.2419	0.0593	0.1130	0.4701	0.9772	0.0613	0.0156	0.9475	0.0602	0.0135	0.9543	0.0676	0.0620	0.9315	0.9632	0.9912	0.9266	0.9684	0.9348	0.9739	0.9079	0.9552
β_1^{f21}	-2.4128	0.2462	0.0608	0.1139	0.2646	0.9726	0.0732	0.0186	0.9361	0.0718	0.0166	0.9315	0.0792	0.0672	0.9155	0.9573	0.9879	0.9132	0.9590	0.9079	0.9552	0.8895	0.9416
β_1^{f22}	-1.1044	0.1254	0.0158	0.1316	0.6099	0.9680	0.0727	0.0180	0.9155	0.0716	0.0165	0.9087	0.0812	0.0896	0.9087	0.9516	0.9845	0.8895	0.9416	0.8817	0.9357	0.8817	0.9357
β_1^{tc}	0.5034	0.0920	0.0085	0.0281	0.1890	0.9635	0.0151	0.0040	0.9475	0.0134	0.0033	0.9338	0.0184	0.0167	0.9498	0.9459	0.9810	0.9266	0.9684	0.9105	0.9571	0.9293	0.9702
d_1^{11}	1.5664	0.4792	0.2475	1.3100	11.8065	0.9521	0.5634	0.3209	0.9361	0.4912	0.3446	0.8836	1.4684	8.4107	0.9155	0.9321	0.9721	0.9132	0.9590	0.8535	0.9136	0.8895	0.9416
σ_1^2	1.5163	2.2858	5.2715	0.4852	4.1066	0.9795	0.2416	0.1903	0.9589	0.1899	0.3931	0.9589	0.8151	2.8785	0.9498	0.9662	0.9927	0.9403	0.9775	0.9403	0.9775	0.9293	0.9702
ϕ_1^1	0.6032	0.0522	0.0027	0.0769	0.2812	0.9726	0.0444	0.0080	0.9338	0.0419	0.0047	0.9292	0.0679	0.0995	0.9406	0.9573	0.9879	0.9105	0.9571	0.9052	0.9532	0.9185	0.9628
ϕ_1^2	0.0997	0.0522	0.0027	0.0786	0.2059	0.9726	0.0537	0.0188	0.9224	0.0484	0.0060	0.9361	0.0954	0.3083	0.9384	0.9573	0.9879	0.8973	0.9474	0.9132	0.9590	0.9158	0.9609
ϕ_1^3	-0.0443	0.0520	0.0027	0.0913	0.4714	0.9817	0.0491	0.0127	0.9064	0.0448	0.0051	0.9338	0.0889	0.2889	0.9269	0.9692	0.9943	0.8791	0.9337	0.9105	0.9571	0.9026	0.9513
β_2^0	-6.0723	0.2764	0.0816	0.3757	0.0907	0.9886	0.2732	0.1312	0.9132	0.2844	0.0409	0.9384	0.4061	2.8120	0.8813	0.9786	0.9985	0.8869	0.9396	0.9158	0.9609	0.8510	0.9116
β_2^{c1}	1.5002	0.0135	0.0002	0.0174	0.0043	0.9817	0.0129	0.0044	0.9247	0.0127	0.0011	0.9361	0.0176	0.0906	0.9018	0.9692	0.9943	0.8999	0.9494	0.9132	0.9590	0.8740	0.9297
β_2^{c2}	3.0013	0.0132	0.0002	0.0182	0.0050	0.9977	0.0133	0.0021	0.9452	0.0133	0.0012	0.9566	0.0148	0.0241	0.9224	0.9933	1.0022	0.9239	0.9665	0.9375	0.9757	0.8973	0.9474
β_2^{f11}	3.9972	0.0630	0.0040	0.0820	0.0215	0.9817	0.0609	0.0383	0.9247	0.0596	0.0049	0.9406	0.1026	0.8398	0.9087	0.9692	0.9943	0.8999	0.9494	0.9185	0.9628	0.8817	0.9357
β_2^{f21}	-5.0067	0.0732	0.0054	0.0960	0.0250	0.9795	0.0713	0.0186	0.9406	0.0709	0.0059	0.9498	0.0927	0.3788	0.9132	0.9662	0.9927	0.9185	0.9628	0.9293	0.9702	0.8869	0.9396
β_2^{f22}	-2.0041	0.0771	0.0060	0.0965	0.0232	0.9772	0.0713	0.0196	0.9110	0.0713	0.0060	0.9201	0.0926	0.4023	0.9018	0.9632	0.9912	0.8843	0.9376	0.8947	0.9455	0.8740	0.9297
β_2^{tc}	1.4998	0.0152	0.0002	0.0200	0.0050	0.9886	0.0149	0.0034	0.9338	0.0149	0.0017	0.9315	0.0177	0.0566	0.9018	0.9786	0.9985	0.9105	0.9571	0.9079	0.9552	0.8740	0.9297
d_2^{11}	1.9252	0.6235	0.3943	0.8773	0.3319	0.9521	0.5444	0.1671	0.8813	0.6043	0.1621	0.8950	0.7188	2.5984	0.8196	0.9321	0.9721	0.8510	0.9116	0.8663	0.9237	0.7836	0.8556
σ_2^2	1.4926	0.2127	0.0453	0.3315	0.1382	0.9703	0.2394	0.1425	0.9361	0.2287	0.0803	0.9269	0.5379	2.9240	0.9452	0.9544	0.9862	0.9132	0.9590	0.9026	0.9513	0.9239	0.9665
ϕ_2^1	0.5825	0.0401	0.0016	0.0601	0.0154	0.9909	0.0438	0.0062	0.9452	0.0426	0.0032	0.9589	0.0658	0.1828	0.9635	0.9820	0.9998	0.9239	0.9665	0.9403	0.9775	0.9459	0.9810
ϕ_2^2	0.2152	0.0526	0.0028	0.0677	0.0151	0.9772	0.0524	0.0135	0.9315	0.0485	0.0038	0.9338	0.0851	0.3183	0.9269	0.9632	0.9912	0.9079	0.9552	0.9105	0.9571	0.9026	0.9513
ϕ_2^3	-0.0727	0.0474	0.0023	0.0642	0.0159	0.9863	0.0490	0.0107	0.9269	0.0457	0.0034	0.9429	0.0790	0.2267	0.9292	0.9754	0.9972	0.9026	0.9513	0.9212	0.9647	0.9052	0.9532
β_3^0	3.3892	0.8976	0.8059	12.8561	249.2545	0.9726	0.2720	0.0661	0.9155	0.2681	0.0420	0.9247	0.4472	2.6132	0.9064	0.9573	0.9879	0.8895	0.9416	0.8999	0.9494	0.8791	0.9337
β_3^{c1}	-2.9642	0.3734	0.1407	0.5647	10.1734	0.9772	0.0131	0.0053	0.9155	0.0128	0.0033	0.9178	0.0364	0.3994	0.8995	0.9632	0.9912	0.8895	0.9416	0.8921	0.9435	0.8714	0.9277
β_3^{c2}	1.0086	0.1806	0.0327	0.7116	13.8406	0.9772	0.0134	0.0026	0.9429	0.0135	0.0035	0.9429	0.0206	0.1375	0.9201	0.9632	0.9912	0.9212	0.9647	0.9212	0.9647	0.8947	0.9455
β_3^{f11}	5.9747	0.2444	0.0604	5.6631	114.0431	0.9703	0.0600	0.0139	0.9406	0.0600	0.0154	0.9384	0.1360	1.2131	0.9338	0.9544	0.9862	0.9185	0.9628	0.9158	0.9609	0.9105	0.9571
β_3^{f21}	-0.4300	0.3771	0.1431	3.0162	55.9722	0.9772	0.0725	0.0312	0.9201	0.0717	0.0182	0.9269	0.1535	1.1695	0.9247	0.9632	0.9912	0.8947	0.9455	0.9026	0.9513	0.8999	0.9494
β_3^{f22}	-0.2175	0.1983	0.0396	2.5208	42.7268	0.9886	0.0745	0.0721	0.9338	0.0717	0.0190	0.9406	0.1472	1.5203	0.9087	0.9786	0.9985	0.9105	0.9571	0.9185	0.9628	0.8817	0.9357
β_3^{tc}	-0.4838	0.1702	0.0292	0.3843	6.9942	0.9772	0.0149	0.0055	0.8973	0.0162	0.0035	0.9269	0.0244	0.1427	0.8585	0.9632	0.9912	0.8688	0.9257	0.9026	0.9513	0.8258	0.8911
d_3^{11}	1.5307	0.5601	0.3186	8.9892	169.9516	0.9680	0.5734	0.6095	0.9155	0.5852	0.3787	0.9361	4.7480	71.5084	0.8836	0.9516	0.9845	0.8895	0.9416	0.9132	0.9590	0.8535	0.9136
σ_3^2	1.9026	2.3503	5.5649	5.4821	90.3545	0.9612	0.2565	0.6671	0.8516	0.3398	0.4408	0.9269	4.0031	76.5999	0.8950	0.9431	0.9793	0.8183	0.8849	0.9026	0.9513	0.8663	0.9237
ϕ_3^1	0.6198	0.0506	0.0026	3.0714	58.4278	0.9795	0.0481	0.0817	0.9361	0.0425	0.0074	0.9384	0.5188	9.4269	0.9566	0.9662	0.9927	0.9132	0.9590	0.9158	0.9609	0.9375	0.9757
ϕ_3^2	0.1752	0.0530	0.0028	5.4044	101.0815	0.9703	0.0543	0.0399	0.9018	0.0491	0.0083	0.9361	0.2964	4.3587	0.8950	0.9544	0.9862	0.8740	0.9297	0.9132	0.9590	0.8663	0.9237
ϕ_3^3	-0.0307	0.0517	0.0027	4.1993	79.6273	0.9909	0.0602	0.2328	0.9269	0.0455	0.0080	0.9452	1.3766	26.5517	0.9498	0.9820	0.9998	0.9026	0.9513	0.9239	0.9665	0.9293	0.9702
π_1	0.3413	0.0565	0.0033	0.0480	0.0103	0.9543	0.0471	0.0021	0.9521				0.0472	0.0021	0.9521	0.9348	0.9739	0.9321	0.9721			0.9321	0.9721
π_2	0.3224	0.0464	0.0023	0.0474	0.0099	0.9498	0.0465	0.0018	0.9452				0.0483	0.0309	0.9498	0.9293	0.9702	0.9239	0.9665			0.9293	0.9702
π_3	0.3364	0.0567	0.0032	0.0477	0.0087	0.9543	0.0469	0.0023	0.9498				0.0486	0.0305	0.9543	0.9348	0.9739	0.9293	0.9702			0.9348	0.9739

Table A.2 continued

parm	CI1L avg	CI1L std	CI1L n	CI2L avg	CI2L std	CI2L n	CI3L avg	CI3L std	CI3L n	CI4L avg	CI4L std	CI4L n
β_1^0	17.6135	1.5998	438	17.4826	0.8876	438	17.4809	0.8843	438	17.4912	0.8668	438
β_1^{c1}	1.4430	0.3013	438	1.4285	0.1252	438	1.4284	0.1252	438	1.4288	0.1271	438
β_1^{c2}	1.0819	0.3215	438	1.0777	0.2769	438	1.0778	0.2763	438	1.0788	0.2764	438
β_1^{f11}	2.9442	1.2132	438	2.8735	0.3442	438	2.8733	0.3436	438	2.8769	0.3641	438
β_1^{f21}	3.4509	0.7024	438	3.4186	0.3486	438	3.4184	0.3468	438	3.4220	0.3701	438
β_1^{f22}	1.6748	1.6298	438	1.5729	0.1698	437	1.5754	0.1736	438	1.5877	0.2499	438
β_1^{tc}	0.7373	0.5105	438	0.7135	0.1284	438	0.7134	0.1277	438	0.7159	0.1262	438
d_1^{11}	4.6921	32.6725	438	2.7670	0.9487	437	2.6445	1.0671	438	5.1699	23.2097	438
σ_1^2	2.8902	11.7502	438	2.2799	3.2666	434	2.2110	3.4094	438	3.6370	8.4030	437
ϕ_1^1	0.9182	0.7367	438	0.8622	0.0732	434	0.8610	0.0731	438	0.8926	0.2182	437
ϕ_1^2	0.2752	0.5683	438	0.2147	0.0646	431	0.2025	0.0507	438	0.3252	0.8484	438
ϕ_1^3	0.2767	1.3054	438	0.1644	0.0449	430	0.1535	0.0378	438	0.2713	0.7986	438
β_2^0	8.6542	0.3884	438	8.6276	0.4215	437	8.6244	0.3890	438	8.9836	7.4303	438
β_2^{c1}	2.1222	0.0191	438	2.1218	0.0190	437	2.1219	0.0191	438	2.1303	0.1696	438
β_2^{c2}	4.2447	0.0186	438	4.2446	0.0186	437	4.2446	0.0186	438	4.2451	0.0216	438
β_2^{f11}	5.6578	0.0890	438	5.6565	0.0917	437	5.6553	0.0890	438	5.7565	2.0796	438
β_2^{f21}	7.0858	0.1032	438	7.0834	0.1032	438	7.0832	0.1035	438	7.1216	0.7716	438
β_2^{f22}	2.8476	0.1085	438	2.8417	0.1079	438	2.8412	0.1088	438	2.8915	0.9967	438
β_2^{tc}	2.1218	0.0215	438	2.1214	0.0215	438	2.1214	0.0215	438	2.1256	0.0901	438
d_2^{11}	3.6857	1.1683	438	3.1500	0.8744	437	3.2046	0.9637	438	3.6786	7.1074	437
σ_2^2	2.3165	0.4134	438	2.2344	0.3901	437	2.2077	0.3516	438	2.9733	7.9759	438
ϕ_2^1	0.8417	0.0556	438	0.8326	0.0560	435	0.8323	0.0561	438	0.8698	0.4639	438
ϕ_2^2	0.3616	0.0660	438	0.3411	0.0677	434	0.3347	0.0656	438	0.4223	0.8683	438
ϕ_2^3	0.2148	0.0504	438	0.1794	0.0452	433	0.1722	0.0397	438	0.2592	0.6249	438
β_3^0	39.6409	690.6849	438	4.9651	0.7308	438	4.9628	0.7218	438	5.4030	7.0282	438
β_3^{c1}	5.7267	27.9762	438	4.2200	0.2110	438	4.2200	0.2112	438	4.2707	0.9208	438
β_3^{c2}	3.3279	38.2971	438	1.4269	0.2550	438	1.4271	0.2541	438	1.4430	0.4075	438
β_3^{f11}	23.8986	315.7046	438	8.4562	0.3291	437	8.4514	0.3432	438	8.6246	2.9290	438
β_3^{f21}	8.7257	155.1261	438	0.6429	0.5373	438	0.6425	0.5323	438	0.8531	3.2582	438
β_3^{f22}	7.1212	118.4232	438	0.3779	0.3362	438	0.3732	0.2760	438	0.5766	4.2148	437
β_3^{tc}	1.7106	19.3538	438	0.7149	0.1304	438	0.7153	0.1285	438	0.7388	0.3759	438
d_3^{11}	25.8489	471.0285	438	2.7758	1.7323	437	2.7648	1.1839	438	14.3788	198.1363	437
σ_3^2	16.9264	250.3707	438	2.8630	3.7478	437	2.8588	3.5347	438	13.0977	212.2493	436
ϕ_3^1	9.2411	161.9131	438	0.8953	0.2030	434	0.8848	0.0695	438	2.1703	26.0937	438
ϕ_3^2	15.1091	280.1716	438	0.2955	0.1181	432	0.2861	0.0653	438	0.9589	12.0742	438
ϕ_3^3	11.6568	220.7107	438	0.1880	0.6449	432	0.1482	0.0404	438	3.8345	73.5953	438
π_1	0.5020	0.0770	999	0.5003	0.0784	438				0.5005	0.0772	438
π_2	0.4753	0.0650	999	0.4740	0.0644	438				0.4779	0.0950	438
π_3	0.4955	0.0721	999	0.4943	0.0761	437				0.4983	0.0988	438

Table A.3: EM2nd variant simulation results for CON

parm	est avg	est std	est mse	SE1 avg	SE1 std	CI1 CP	SE2 avg	SE2 std	CI2 CP	SE3 avg	SE3 std	CI3 CP	SE4 avg	SE4 std	CI4 CP	CI1BLCL	CI1BUCL	CI2BLCL	CI2BUCL	CI3BLCL	CI3BUCL	CI4BLCL	CI4BUCL
β_1^0	11.736	3.388	11.917	205.107	3390.437	0.940	0.818	0.226	0.870	0.823	0.744	0.860	1.385	5.733	0.847	0.913	0.967	0.832	0.908	0.821	0.899	0.806	0.887
β_1^{c1}	-0.915	0.449	0.209	1.866	27.343	0.933	0.015	0.012	0.897	0.014	0.010	0.900	0.044	0.348	0.900	0.905	0.962	0.862	0.931	0.866	0.934	0.866	0.934
β_1^{c2}	-0.612	0.652	0.445	40.633	697.792	0.950	0.112	0.031	0.890	0.116	0.199	0.887	0.185	0.621	0.887	0.925	0.975	0.855	0.925	0.851	0.923	0.851	0.923
β_1^{f11}	2.115	0.813	0.674	244.460	4127.110	0.963	0.511	0.132	0.880	0.496	0.232	0.887	0.885	3.302	0.897	0.942	0.985	0.843	0.917	0.851	0.923	0.862	0.931
β_1^{f21}	-2.487	0.837	0.708	177.184	2910.036	0.983	0.642	0.185	0.897	0.672	1.017	0.903	1.390	8.252	0.897	0.969	0.998	0.862	0.931	0.870	0.937	0.862	0.931
β_1^{f22}	-1.173	0.820	0.677	173.631	2903.450	0.967	0.681	0.199	0.877	0.668	0.517	0.887	1.364	8.257	0.893	0.946	0.987	0.839	0.914	0.851	0.923	0.858	0.928
β_1^{tc}	0.535	0.189	0.037	1.175	14.358	0.940	0.016	0.010	0.923	0.014	0.008	0.913	0.062	0.650	0.920	0.913	0.967	0.893	0.953	0.882	0.945	0.889	0.951
d_1^{11}	1.330	0.537	0.425	20.706	303.175	0.870	0.582	0.890	0.797	0.827	3.441	0.730	5.714	68.623	0.857	0.832	0.908	0.751	0.842	0.680	0.780	0.817	0.896
σ_1^2	2.347	7.536	57.887	15.423	246.842	0.940	0.246	0.336	0.920	0.592	3.563	0.903	1.379	6.084	0.943	0.913	0.967	0.889	0.951	0.870	0.937	0.917	0.970
ϕ_1^1	0.585	0.075	0.006	2.793	38.067	0.950	0.045	0.010	0.893	0.043	0.013	0.907	0.097	0.323	0.930	0.925	0.975	0.858	0.928	0.874	0.940	0.901	0.959
ϕ_1^2	0.103	0.066	0.004	2.230	33.590	0.963	0.053	0.023	0.893	0.049	0.013	0.923	0.131	0.608	0.910	0.942	0.985	0.858	0.928	0.893	0.953	0.878	0.942
ϕ_1^3	-0.050	0.059	0.003	3.020	35.776	0.977	0.049	0.015	0.913	0.046	0.013	0.943	0.149	0.879	0.937	0.960	0.994	0.882	0.945	0.917	0.970	0.909	0.964
β_2^0	-5.976	1.680	2.823	11516.677	197482.789	0.953	0.979	2.699	0.847	0.956	2.079	0.873	1.851	10.168	0.833	0.929	0.977	0.806	0.887	0.836	0.911	0.791	0.876
β_2^{c1}	1.440	0.511	0.264	518.426	8949.265	0.990	0.016	0.030	0.920	0.015	0.027	0.913	0.046	0.256	0.893	0.979	1.001	0.889	0.951	0.882	0.945	0.858	0.928
β_2^{c2}	2.990	0.269	0.072	5787.776	99889.785	0.960	0.145	0.563	0.880	0.139	0.404	0.897	0.236	1.210	0.853	0.938	0.982	0.843	0.917	0.862	0.931	0.813	0.893
β_2^{f11}	4.104	0.630	0.407	7749.796	133420.818	0.963	0.549	0.509	0.877	0.555	0.484	0.903	1.530	10.054	0.860	0.942	0.985	0.839	0.914	0.870	0.937	0.821	0.899
β_2^{f21}	-4.937	0.861	0.746	6870.866	118053.847	0.970	0.706	0.821	0.907	0.699	0.747	0.900	1.982	15.038	0.887	0.951	0.989	0.874	0.940	0.866	0.934	0.851	0.923
β_2^{f22}	-1.973	0.817	0.668	3816.613	65023.231	0.960	0.862	2.929	0.900	0.803	1.728	0.903	1.939	15.429	0.863	0.938	0.982	0.866	0.934	0.870	0.937	0.824	0.902
β_2^{tc}	1.474	0.222	0.050	589.962	10145.301	0.970	0.017	0.024	0.880	0.016	0.023	0.897	0.038	0.241	0.870	0.951	0.989	0.843	0.917	0.862	0.931	0.832	0.908
d_2^{11}	1.528	0.588	0.569	1078.910	18149.687	0.900	0.525	0.477	0.750	0.538	0.477	0.760	4.205	44.760	0.753	0.866	0.934	0.701	0.799	0.712	0.808	0.705	0.802
σ_2^2	1.521	0.291	0.085	1557.156	26359.877	0.963	0.286	0.691	0.857	0.270	0.599	0.910	4.986	65.263	0.903	0.942	0.985	0.817	0.896	0.878	0.942	0.870	0.937
ϕ_2^1	0.574	0.055	0.003	749.019	12915.754	0.983	0.047	0.031	0.903	0.046	0.045	0.917	0.240	2.469	0.943	0.969	0.998	0.870	0.937	0.885	0.948	0.917	0.970
ϕ_2^2	0.216	0.047	0.002	3138.610	54076.546	0.987	0.056	0.052	0.950	0.052	0.046	0.980	0.155	1.094	0.953	0.974	1.000	0.925	0.975	0.964	0.996	0.929	0.977
ϕ_2^3	-0.070	0.049	0.002	1009.292	17320.863	0.990	0.052	0.042	0.917	0.049	0.047	0.943	0.224	1.963	0.947	0.979	1.001	0.885	0.948	0.917	0.970	0.921	0.972
β_3^0	3.466	1.409	1.991	1.274	1.107	0.977	0.851	0.793	0.883	0.812	0.216	0.887	2.365	17.778	0.880	0.960	0.994	0.847	0.920	0.851	0.923	0.843	0.917
β_3^{c1}	-2.942	0.397	0.161	0.022	0.025	0.950	0.014	0.011	0.903	0.015	0.012	0.910	0.022	0.084	0.873	0.925	0.975	0.870	0.937	0.878	0.942	0.836	0.911
β_3^{c2}	0.986	0.226	0.051	0.160	0.065	0.960	0.122	0.214	0.883	0.111	0.030	0.887	0.474	4.567	0.870	0.938	0.982	0.847	0.920	0.851	0.923	0.832	0.908
β_3^{f11}	5.936	0.632	0.403	0.725	0.273	0.973	0.517	0.228	0.883	0.513	0.124	0.920	1.029	6.418	0.880	0.955	0.992	0.847	0.920	0.889	0.951	0.843	0.917
β_3^{f21}	-0.457	0.786	0.620	1.004	1.085	0.963	0.694	1.108	0.907	0.642	0.166	0.900	2.446	23.156	0.873	0.942	0.985	0.874	0.940	0.866	0.934	0.836	0.911
β_3^{f22}	-0.222	0.769	0.591	1.093	1.103	0.980	0.700	0.534	0.923	0.682	0.177	0.923	1.344	8.900	0.910	0.964	0.996	0.893	0.953	0.893	0.953	0.878	0.942
β_3^{tc}	-0.473	0.180	0.033	0.028	0.038	0.957	0.016	0.009	0.887	0.017	0.009	0.907	0.020	0.064	0.843	0.934	0.980	0.851	0.923	0.874	0.940	0.802	0.884
d_3^{11}	1.249	0.595	0.477	1.329	3.209	0.950	0.610	1.201	0.790	0.640	0.789	0.810	7.245	106.577	0.823	0.925	0.975	0.744	0.836	0.766	0.854	0.780	0.866
σ_3^2	3.802	12.068	150.058	1.257	4.681	0.920	0.279	0.758	0.757	0.491	1.079	0.887	6.170	89.926	0.833	0.889	0.951	0.708	0.805	0.851	0.923	0.791	0.876
ϕ_3^1	0.602	0.078	0.006	0.064	0.032	0.973	0.043	0.007	0.897	0.042	0.004	0.927	0.068	0.104	0.940	0.955	0.992	0.862	0.931	0.897	0.956	0.913	0.967
ϕ_3^2	0.175	0.054	0.003	0.071	0.034	0.973	0.050	0.013	0.903	0.048	0.005	0.923	0.074	0.148	0.900	0.955	0.992	0.870	0.937	0.893	0.953	0.866	0.934
ϕ_3^3	-0.020	0.057	0.003	0.069	0.044	0.977	0.048	0.011	0.907	0.045	0.004	0.913	0.080	0.163	0.937	0.960	0.994	0.874	0.940	0.882	0.945	0.909	0.964
π_1	0.336	0.077	0.006	0.058	0.121	0.907	0.047	0.004	0.897				0.057	0.155	0.907	0.874	0.940	0.862	0.931			0.874	0.940
π_2	0.323	0.062	0.004	0.050	0.019	0.930	0.047	0.004	0.907				0.063	0.173	0.927	0.901	0.959	0.874	0.940			0.897	0.956
π_3	0.340	0.075	0.006	0.059	0.120	0.937	0.047	0.002	0.920				0.053	0.078	0.927	0.909	0.964	0.889	0.951			0.897	0.956

Table A.3 continued

parm	CI1L avg	CI1L std	CI1L n	CI2L avg	CI2L std	CI2L n	CI3L avg	CI3L std	CI3L n	CI4L avg	CI4L std	CI4L n
β_1^0	582.743	9396.783	300	17.261	2.452	300	17.372	2.475	300	18.467	14.989	300
β_1^{c1}	6.521	75.698	300	1.427	0.209	300	1.427	0.208	300	1.492	0.892	300
β_1^{c2}	113.392	1934.102	300	1.193	0.529	300	1.214	0.731	300	1.334	1.745	300
β_1^{f11}	679.234	11439.457	300	3.322	1.035	299	3.337	1.195	300	4.217	9.069	300
β_1^{f21}	493.000	8065.945	300	3.976	1.061	299	4.114	2.876	300	5.928	22.729	300
β_1^{f22}	481.898	8047.767	300	2.683	0.879	299	2.679	1.550	300	4.527	22.840	300
β_1^{tc}	3.941	39.744	300	0.753	0.256	299	0.758	0.267	300	0.862	1.782	300
d_1^{11}	58.269	840.285	300	2.659	2.404	296	3.380	9.430	300	16.780	190.141	300
σ_1^2	44.692	684.166	300	3.485	10.782	294	3.903	14.478	300	6.116	19.651	300
ϕ_1^1	8.419	105.467	300	0.844	0.083	290	0.837	0.100	300	0.943	0.824	299
ϕ_1^2	6.238	93.100	300	0.214	0.084	289	0.211	0.077	300	0.426	1.679	300
ϕ_1^3	8.399	99.161	300	0.169	0.055	290	0.164	0.053	300	0.444	2.434	299
β_2^0	31927.910	547383.963	300	9.405	7.206	300	9.307	5.539	300	11.657	27.600	300
β_2^{c1}	1439.042	24805.522	300	2.143	0.217	298	2.149	0.243	300	2.201	0.586	300
β_2^{c2}	16046.373	276875.050	300	4.332	1.361	300	4.307	0.936	300	4.533	3.024	300
β_2^{f11}	21485.026	369816.608	300	6.064	1.420	300	6.061	1.382	300	8.667	27.458	300
β_2^{f21}	19049.583	327222.272	300	7.398	2.120	300	7.368	1.994	300	10.752	41.257	300
β_2^{f22}	10580.183	180231.814	300	3.958	8.067	300	3.777	4.769	300	6.893	42.654	300
β_2^{tc}	1637.302	28120.705	300	2.104	0.157	300	2.104	0.157	300	2.147	0.543	300
d_2^{11}	2991.485	50307.385	300	2.716	1.389	295	2.679	1.469	300	12.876	123.968	300
σ_2^2	4317.545	73064.424	300	2.387	1.824	295	2.336	1.631	300	15.272	180.792	299
ϕ_2^1	2076.796	35799.948	300	0.826	0.083	293	0.827	0.106	300	1.325	6.796	300
ϕ_2^2	8699.799	149889.774	300	0.351	0.141	292	0.344	0.130	300	0.612	3.016	300
ϕ_2^3	2797.600	48010.098	300	0.188	0.118	294	0.180	0.134	300	0.660	5.438	300
β_3^0	6.327	3.141	300	5.678	2.480	300	5.559	1.590	300	9.804	49.021	300
β_3^{c1}	4.191	0.275	300	4.190	0.280	300	4.190	0.279	300	4.196	0.280	300
β_3^{c2}	1.481	0.285	300	1.467	0.600	300	1.437	0.286	300	2.411	12.578	300
β_3^{f11}	8.675	0.799	300	8.541	0.886	300	8.527	0.838	300	9.796	17.318	300
β_3^{f21}	3.036	3.037	300	2.256	3.083	299	2.120	0.735	300	7.115	64.161	300
β_3^{f22}	3.221	3.072	300	2.221	1.518	300	2.155	0.670	300	4.006	24.651	300
β_3^{tc}	0.716	0.128	300	0.701	0.152	300	0.703	0.146	300	0.712	0.201	300
d_3^{11}	4.394	8.787	300	2.660	3.275	297	2.681	2.138	300	21.067	295.349	300
σ_3^2	6.469	21.420	300	4.354	13.360	295	5.572	17.318	300	21.356	249.573	299
ϕ_3^1	0.878	0.074	300	0.868	0.089	294	0.860	0.109	300	0.897	0.227	300
ϕ_3^2	0.325	0.099	300	0.287	0.069	294	0.286	0.064	300	0.345	0.399	299
ϕ_3^3	0.208	0.124	300	0.154	0.040	296	0.149	0.032	300	0.243	0.448	300
π_1	0.524	0.320	300	0.493	0.107	300	999	999	0	0.521	0.418	300
π_2	0.483	0.076	300	0.476	0.083	300	999	999	0	0.520	0.455	300
π_3	0.524	0.326	300	0.499	0.105	300	999	999	0	0.514	0.217	300

Table A.4: EM2nd variant simulation results for NCON

parm	est avg	est std	est mse	SE1 avg	SE1 std	CI1 CP	SE2 avg	SE2 std	CI2 CP	SE3 avg	SE3 std	CI3 CP	SE4 avg	SE4 std	CI4 CP	CI1BLCL	CI1BUCL	CI2BLCL	CI2BUCL	CI3BLCL	CI3BUCL	CI4BLCL	CI4BUCL
β_1^0	12.1191	1.6870	2.9249	135.3773	2334.3439	0.9533	0.2750	0.0881	0.8867	0.2701	0.0922	0.8867	0.3095	0.2513	0.8900	0.9295	0.9772	0.8508	0.9225	0.8508	0.9225	0.8546	0.9254
β_1^{c1}	-1.0134	0.2626	0.0691	3.3359	57.3110	0.9533	0.0139	0.0073	0.9100	0.0152	0.0124	0.9167	0.0170	0.0396	0.8800	0.9295	0.9772	0.8776	0.9424	0.8854	0.9479	0.8432	0.9168
β_1^{c2}	-0.6978	0.3406	0.1187	33.2953	575.3343	0.9500	0.0144	0.0071	0.8967	0.0158	0.0128	0.9067	0.0175	0.0467	0.8800	0.9253	0.9747	0.8622	0.9311	0.8738	0.9396	0.8432	0.9168
β_1^{f11}	2.0758	0.3922	0.1595	54.5098	940.7595	0.9600	0.0634	0.0308	0.9133	0.0697	0.0561	0.9233	0.0790	0.2169	0.9033	0.9378	0.9822	0.8815	0.9452	0.8932	0.9534	0.8699	0.9368
β_1^{f21}	-2.3915	0.2905	0.0845	91.6237	1582.1815	0.9900	0.0774	0.0417	0.9033	0.0842	0.0687	0.9200	0.1160	0.6517	0.8967	0.9787	1.0013	0.8699	0.9368	0.8893	0.9507	0.8622	0.9311
β_1^{f22}	-1.0868	0.1418	0.0203	140.2946	2423.8176	0.9867	0.0763	0.0391	0.9267	0.0836	0.0684	0.9500	0.0953	0.3055	0.8967	0.9737	0.9997	0.8972	0.9562	0.9253	0.9747	0.8622	0.9311
β_1^{tc}	0.4920	0.1174	0.0139	47.3091	817.6708	0.9400	0.0157	0.0070	0.9333	0.0147	0.0086	0.9100	0.0231	0.0771	0.9400	0.9131	0.9669	0.9051	0.9616	0.8776	0.9424	0.9131	0.9669
d_1^{11}	1.5314	0.5786	0.3632	91.1720	1558.1570	0.9433	0.5283	0.1480	0.8700	0.5717	0.6120	0.8833	1.0155	1.9561	0.8967	0.9172	0.9695	0.8319	0.9081	0.8470	0.9197	0.8622	0.9311
σ_1^2	3.3973	11.5428	137.6347	96.4633	1655.5610	0.9433	0.2374	0.1311	0.9367	0.3133	0.9024	0.9100	0.9780	3.2067	0.9400	0.9172	0.9695	0.9091	0.9642	0.8776	0.9424	0.9131	0.9669
ϕ_1^1	0.5909	0.0846	0.0072	123.5390	2135.1262	0.9533	0.0451	0.0167	0.9033	0.0421	0.0049	0.9000	0.1146	0.6735	0.9000	0.9295	0.9772	0.8699	0.9368	0.8661	0.9340	0.8661	0.9340
ϕ_1^2	0.0975	0.0577	0.0033	288.0189	4979.0959	0.9900	0.0535	0.0203	0.8767	0.0485	0.0065	0.9067	0.1061	0.5043	0.9067	0.9787	1.0013	0.8395	0.9139	0.8738	0.9396	0.8738	0.9396
ϕ_1^3	-0.0427	0.0641	0.0042	106.2170	1835.5316	0.9867	0.0503	0.0223	0.9100	0.0450	0.0053	0.9333	0.1296	0.7251	0.9233	0.9737	0.9997	0.8776	0.9424	0.9051	0.9616	0.8932	0.9534
β_2^0	-6.0092	0.2880	0.0830	1.1988	14.2752	0.9633	0.2788	0.1115	0.9133	0.2808	0.0367	0.9200	0.5352	3.7396	0.8900	0.9421	0.9846	0.8815	0.9452	0.8893	0.9507	0.8546	0.9254
β_2^{c1}	1.4991	0.0130	0.0002	0.0355	0.3086	0.9800	0.0146	0.0176	0.9433	0.0127	0.0010	0.9533	0.0553	0.6159	0.9433	0.9642	0.9958	0.9172	0.9695	0.9295	0.9772	0.9172	0.9695
β_2^{c2}	2.9997	0.0141	0.0002	0.0867	1.1804	0.9800	0.0148	0.0115	0.9300	0.0133	0.0011	0.9300	0.0374	0.3027	0.9167	0.9642	0.9958	0.9011	0.9589	0.9011	0.9589	0.8854	0.9479
β_2^{f11}	3.9999	0.0627	0.0039	0.3039	3.8733	0.9733	0.0642	0.0389	0.9433	0.0585	0.0044	0.9400	0.1125	0.4887	0.9267	0.9551	0.9916	0.9172	0.9695	0.9131	0.9669	0.8972	0.9562
β_2^{f21}	-5.0006	0.0707	0.0050	0.3698	4.7402	0.9867	0.0787	0.0616	0.9433	0.0705	0.0057	0.9433	0.1996	1.6350	0.9200	0.9737	0.9997	0.9172	0.9695	0.9172	0.9695	0.8893	0.9507
β_2^{f22}	-2.0013	0.0701	0.0049	0.1970	1.7417	0.9900	0.0821	0.1209	0.9500	0.0701	0.0055	0.9667	0.3588	4.4231	0.9233	0.9787	1.0013	0.9253	0.9747	0.9464	0.9870	0.8932	0.9534
β_2^{tc}	1.5017	0.0151	0.0002	0.0481	0.4841	0.9767	0.0156	0.0112	0.9233	0.0146	0.0018	0.9333	0.0394	0.3752	0.8967	0.9596	0.9938	0.8932	0.9534	0.9051	0.9616	0.8622	0.9311
d_2^{11}	1.9382	0.5804	0.3406	1.8987	17.5174	0.9700	0.5920	0.4986	0.8967	0.5969	0.1452	0.9267	1.4293	8.4555	0.8500	0.9507	0.9893	0.8622	0.9311	0.8972	0.9562	0.8096	0.8904
σ_2^2	1.4884	0.2285	0.0524	1.9060	27.1291	0.9667	0.2168	0.1007	0.8867	0.2243	0.0860	0.9100	1.5213	13.1168	0.8967	0.9464	0.9870	0.8508	0.9225	0.8776	0.9424	0.8622	0.9311
ϕ_2^1	0.5799	0.0449	0.0020	0.7874	12.5734	0.9733	0.0444	0.0094	0.9233	0.0418	0.0031	0.9300	0.0685	0.1581	0.9467	0.9551	0.9916	0.8932	0.9534	0.9011	0.9589	0.9212	0.9721
ϕ_2^2	0.2145	0.0485	0.0024	0.5203	7.8185	0.9900	0.0532	0.0176	0.9333	0.0476	0.0037	0.9567	0.0772	0.0888	0.9500	0.9787	1.0013	0.9051	0.9616	0.9336	0.9797	0.9253	0.9747
ϕ_2^3	-0.0713	0.0459	0.0021	0.1755	1.9132	0.9867	0.0494	0.0145	0.9067	0.0450	0.0033	0.9367	0.0829	0.1980	0.9267	0.9737	0.9997	0.8738	0.9396	0.9091	0.9642	0.8972	0.9562
β_3^0	3.1058	1.7028	2.9860	13473.9724	181485.2663	0.9533	0.2815	0.0923	0.8833	0.2818	0.1090	0.8900	0.3188	0.3463	0.8700	0.9295	0.9772	0.8470	0.9197	0.8546	0.9254	0.8319	0.9081
β_3^{c1}	-2.8434	0.8020	0.6677	1422.8205	20153.9881	0.9467	0.0143	0.0104	0.8933	0.0142	0.0087	0.9033	0.0178	0.0390	0.8667	0.9212	0.9721	0.8584	0.9283	0.8699	0.9368	0.8282	0.9051
β_3^{c2}	1.0612	0.3648	0.1368	4377.3866	69808.0190	0.9600	0.0150	0.0118	0.8700	0.0149	0.0096	0.8967	0.0207	0.0687	0.8633	0.9378	0.9822	0.8319	0.9081	0.8622	0.9311	0.8245	0.9022
β_3^{f11}	5.9239	0.3922	0.1597	4126.5684	44496.7229	0.9700	0.0661	0.0475	0.9200	0.0648	0.0383	0.9033	0.0815	0.1883	0.9200	0.9507	0.9893	0.8893	0.9507	0.8699	0.9368	0.8893	0.9507
β_3^{f21}	-0.5522	0.8359	0.7220	2789.5902	31439.7435	0.9700	0.0815	0.0687	0.8800	0.0795	0.0566	0.8867	0.0974	0.2432	0.8700	0.9507	0.9893	0.8432	0.9168	0.8508	0.9225	0.8319	0.9081
β_3^{f22}	-0.2546	0.3074	0.0975	1890.3038	24439.2865	0.9700	0.0804	0.0676	0.9000	0.0787	0.0560	0.9100	0.0908	0.1803	0.9033	0.9507	0.9893	0.8661	0.9340	0.8776	0.9424	0.8699	0.9368
β_3^{tc}	-0.4275	0.3701	0.1422	4809.3874	80092.6282	0.9633	0.0156	0.0075	0.8800	0.0170	0.0077	0.9000	0.0152	0.0187	0.8533	0.9421	0.9846	0.8432	0.9168	0.8661	0.9340	0.8133	0.8934
d_3^{11}	1.4680	0.5962	0.3729	14156.0218	207531.8606	0.9833	0.5672	0.3141	0.9100	0.5914	0.3849	0.9167	1.2716	5.3498	0.9133	0.9689	0.9978	0.8776	0.9424	0.8854	0.9479	0.8815	0.9452
σ_3^2	2.1111	5.2175	27.3910	43237.5838	683846.0801	0.9600	0.2312	0.1341	0.7800	0.3499	0.4620	0.8933	0.4486	1.4205	0.8533	0.9378	0.9822	0.7331	0.8269	0.8584	0.9283	0.8133	0.8934
ϕ_3^1	0.6184	0.0646	0.0042	9659.6218	145145.8277	0.9933	0.0447	0.0116	0.9367	0.0450	0.0179	0.9533	0.0591	0.0761	0.9500	0.9841	1.0025	0.9091	0.9642	0.9295	0.9772	0.9253	0.9747
ϕ_3^2	0.1774	0.0656	0.0043	10376.8712	149325.8317	0.9867	0.0539	0.0234	0.8967	0.0522	0.0222	0.9333	0.0761	0.1518	0.9000	0.9737	0.9997	0.8622	0.9311	0.9051	0.9616	0.8661	0.9340
ϕ_3^3	-0.0424	0.0775	0.0062	4236.8058	58279.9530	0.9800	0.0490	0.0147	0.8867	0.0482	0.0189	0.9100	0.0674	0.0886	0.9000	0.9642	0.9958	0.8508	0.9225	0.8776	0.9424	0.8661	0.9340
π_1	0.3415	0.0804	0.0065	0.1941	2.3185	0.9333	0.0470	0.0058	0.9133				0.0639	0.2960	0.9133	0.9051	0.9616	0.8815	0.9452			0.8815	0.9452
π_2	0.3335	0.0456	0.0021	0.1859	2.3627	0.9767	0.0473	0.0057	0.9700				0.0654	0.2964	0.9700	0.9596	0.9938	0.9507	0.9893			0.9507	0.9893
π_3	0.3250	0.0754	0.0058	0.0633	0.1061	0.9500	0.0460	0.0053	0.9267				0.0474	0.0167	0.9300	0.9253	0.9747	0.8972	0.9562			0.9011	0.9589

Table A.4 continued

parm	CI1L avg	CI1L std	CI1L n	CI2L avg	CI2L std	CI2L n	CI3L avg	CI3L std	CI3L n	CI4L avg	CI4L std	CI4L n
β_1^0	391.4766	6469.4134	300	17.2713	1.3344	300	17.2712	1.3300	300	17.2917	1.2867	300
β_1^{c1}	10.6429	158.7745	300	1.4618	0.2395	299	1.4619	0.2402	300	1.4644	0.2479	300
β_1^{c2}	93.2746	1594.6590	300	1.0542	0.3108	300	1.0570	0.3028	300	1.0569	0.3285	300
β_1^{f11}	153.7443	2607.4510	300	2.9419	0.5596	299	2.9433	0.5688	300	2.9662	0.7325	300
β_1^{f21}	257.0316	4385.3254	300	3.3919	0.4012	299	3.3992	0.3787	300	3.4792	1.6936	300
β_1^{f22}	390.1369	6718.2829	300	1.5571	0.1874	299	1.5688	0.1752	300	1.6026	0.7879	300
β_1^{tc}	131.7746	2266.3891	300	0.6999	0.1556	300	0.7010	0.1505	300	0.7100	0.2397	300
d_1^{11}	253.7556	4318.8508	300	2.6871	0.7240	296	2.8420	1.6342	300	3.9099	5.2337	300
σ_1^2	270.6146	4588.7384	300	4.0273	13.5167	297	4.8875	16.5130	300	6.3173	18.3564	299
ϕ_1^1	343.1101	5918.1196	300	0.8548	0.0921	294	0.8439	0.1187	300	1.0065	1.8125	299
ϕ_1^2	798.3859	13801.0930	300	0.2113	0.0671	291	0.2028	0.0544	300	0.3537	1.3928	297
ϕ_1^3	294.4396	5087.7391	300	0.1714	0.0689	293	0.1609	0.0419	300	0.3885	2.0072	299
β_2^0	10.8240	39.1395	300	8.5386	0.4140	300	8.5345	0.4065	300	9.1721	9.8947	300
β_2^{c1}	2.1638	0.7444	300	2.1210	0.0219	299	2.1204	0.0184	300	2.2168	1.5871	300
β_2^{c2}	4.4182	3.0395	300	4.2425	0.0200	300	4.2423	0.0199	300	4.2807	0.6220	300
β_2^{f11}	6.2648	10.4275	300	5.6601	0.0876	299	5.6590	0.0887	300	5.7424	0.9810	300
β_2^{f21}	7.8150	12.7526	300	7.0773	0.1021	300	7.0746	0.0999	300	7.3471	4.1091	300
β_2^{f22}	3.1151	4.6820	300	2.8520	0.2246	300	2.8370	0.0988	300	3.5934	12.0998	300
β_2^{tc}	2.1955	1.2284	300	2.1244	0.0214	300	2.1241	0.0213	300	2.1792	0.9247	300
d_2^{11}	6.5337	48.4847	300	3.2928	1.3988	299	3.2093	0.8870	300	5.6519	23.2645	300
σ_2^2	6.6754	75.1159	300	2.1995	0.3482	295	2.1990	0.3785	300	5.7160	36.2145	299
ϕ_2^1	2.8524	34.8123	300	0.8288	0.0612	293	0.8283	0.0629	300	0.8661	0.3929	300
ϕ_2^2	1.6148	21.6614	300	0.3403	0.0656	293	0.3324	0.0612	300	0.3950	0.2169	300
ϕ_2^3	0.5213	5.3011	300	0.1792	0.0526	293	0.1688	0.0390	300	0.2679	0.5446	300
β_3^0	37351.1081	503041.9635	300	4.9958	0.8987	300	4.9972	0.9061	300	5.0502	1.1355	300
β_3^{c1}	3947.8452	55862.6917	300	4.1656	0.3888	299	4.1584	0.4073	300	4.1611	0.3948	300
β_3^{c2}	12134.6252	193494.3162	300	1.4929	0.4942	299	1.5019	0.5150	300	1.5058	0.5390	300
β_3^{f11}	11446.0624	123335.6142	300	8.3813	0.5456	300	8.3806	0.5482	300	8.4014	0.4819	300
β_3^{f21}	7732.5920	87144.8875	300	0.8189	1.1934	300	0.8185	1.1877	300	0.8425	1.3510	300
β_3^{f22}	5239.6961	67740.9906	300	0.4358	0.4621	300	0.4336	0.4498	300	0.4581	0.6497	300
β_3^{tc}	13331.3346	222001.3210	300	0.7528	0.2742	300	0.7533	0.2736	300	0.7532	0.2770	300
d_3^{11}	39238.6426	575238.3391	300	2.6851	1.0076	294	2.7209	1.2010	300	4.6360	14.6916	300
σ_3^2	119848.1251	1895489.6778	300	2.7864	5.2817	294	3.1809	7.4712	300	3.4429	8.2913	299
ϕ_3^1	26775.3183	402316.2685	300	0.8834	0.0679	293	0.8847	0.0922	300	0.8988	0.1756	299
ϕ_3^2	28762.8165	413902.4737	300	0.2975	0.0893	292	0.2971	0.0892	300	0.3521	0.4101	299
ϕ_3^3	11743.6303	161540.8181	300	0.1614	0.0557	291	0.1654	0.0940	300	0.2155	0.2471	299
π_1	0.8960	6.4055	300	0.5012	0.1106	300				0.5476	0.8067	300
π_2	0.8708	6.5285	300	0.4900	0.0622	300				0.5386	0.8025	300
π_3	0.5225	0.2585	300	0.4775	0.1051	300				0.4814	0.1018	300

A.4.2 Model 2

Table A.5: EM1st variant simulation results for constant(nmax=6)

parm	est avg	est std	est mse	SE1 avg	SE1 std	CI1 CP	SE2 avg	SE2 std	CI2 CP	SE3 avg	SE3 std	CI3 CP	SE4 avg	SE4 std	CI4 CP	CI1BLCL	CI1BUCL	CI2BLCL	CI2BUCL	CI3BLCL	CI3BUCL	CI4BLCL	CI4BUCL
β_1^0	5.473	6.903	53.547	3774.960	76860.496	0.694	0.863	0.539	0.553	0.935	2.561	0.561	4.803	54.499	0.589	0.665	0.722	0.522	0.583	0.530	0.591	0.558	0.619
β_1^{c1}	-1.835	0.382	0.173	167.038	4140.240	0.403	0.097	0.037	0.330	0.097	0.019	0.342	0.441	6.939	0.464	0.373	0.434	0.301	0.360	0.313	0.372	0.434	0.495
β_1^{c2}	3.022	0.872	0.760	1895.975	52563.732	0.598	0.185	0.121	0.432	0.182	0.095	0.431	1.163	13.114	0.542	0.567	0.628	0.402	0.463	0.401	0.462	0.511	0.572
β_1^{f11}	-0.141	6.316	41.200	6516.681	171735.976	0.692	0.611	0.403	0.513	0.602	0.313	0.532	3.888	45.422	0.556	0.663	0.720	0.482	0.544	0.501	0.562	0.525	0.586
β_1^{f21}	1.078	9.547	91.375	5267.889	121422.065	0.569	0.751	0.521	0.425	0.818	2.563	0.429	4.208	38.297	0.442	0.538	0.599	0.395	0.456	0.399	0.460	0.412	0.473
β_1^{f22}	0.865	10.484	110.446	10026.294	151021.976	0.533	0.876	1.998	0.379	0.923	3.053	0.401	10.737	163.272	0.446	0.502	0.563	0.349	0.409	0.371	0.432	0.416	0.477
β_1^{tc}	1.348	0.448	0.224	300.855	8300.024	0.796	0.224	0.087	0.651	0.222	0.062	0.647	1.363	27.175	0.732	0.771	0.821	0.621	0.680	0.617	0.676	0.704	0.759
d_1^{11}	4.561	8.252	80.773	1643.542	50703.220	0.740	1.210	3.317	0.575	1.158	1.942	0.562	49.011	969.498	0.647	0.713	0.767	0.544	0.605	0.531	0.592	0.617	0.676
d_1^{21}	0.949	1.109	1.431	898.779	27473.036	0.926	0.462	0.382	0.728	0.431	0.324	0.732	5.862	69.988	0.784	0.910	0.942	0.700	0.755	0.704	0.759	0.758	0.809
d_1^{22}	1.286	0.458	0.217	458.879	13760.138	0.954	0.386	0.220	0.857	0.381	0.162	0.903	3.583	78.954	0.866	0.941	0.967	0.835	0.879	0.885	0.921	0.845	0.887
σ_1^2	2.035	0.439	0.211	190.374	3683.937	0.936	0.275	0.093	0.771	0.289	0.070	0.819	1.087	13.807	0.750	0.921	0.951	0.745	0.797	0.795	0.843	0.723	0.777
β_2^0	10.630	4.947	27.967	5878.932	90967.346	0.878	0.998	3.104	0.704	0.928	1.137	0.728	12.368	241.697	0.729	0.858	0.898	0.675	0.732	0.700	0.755	0.701	0.756
β_2^{c1}	-2.831	0.332	0.139	169.896	3271.352	0.663	0.098	0.038	0.634	0.088	0.025	0.589	0.264	2.059	0.739	0.633	0.692	0.604	0.664	0.558	0.619	0.712	0.766
β_2^{c2}	1.391	0.674	0.491	4277.171	78565.944	0.845	0.192	0.249	0.691	0.195	0.181	0.730	1.206	21.536	0.722	0.822	0.867	0.662	0.719	0.702	0.757	0.694	0.750
β_2^{f11}	-1.384	4.311	18.734	6479.699	95806.666	0.868	0.636	1.058	0.690	0.660	0.828	0.728	2.870	28.287	0.721	0.847	0.889	0.661	0.718	0.700	0.755	0.693	0.749
β_2^{f21}	2.784	5.561	30.974	7460.734	147542.486	0.869	0.865	2.826	0.701	0.808	0.906	0.733	10.006	216.671	0.731	0.848	0.890	0.672	0.729	0.705	0.760	0.703	0.758
β_2^{f22}	2.839	6.571	44.524	14617.551	200783.926	0.853	1.043	4.664	0.670	1.035	3.477	0.695	10.104	217.862	0.688	0.831	0.875	0.641	0.699	0.666	0.723	0.659	0.716
β_2^{tc}	2.312	0.395	0.191	1602.507	45675.791	0.847	0.223	0.086	0.787	0.201	0.076	0.731	0.786	10.108	0.845	0.825	0.869	0.761	0.812	0.703	0.758	0.822	0.867
d_2^{11}	6.341	8.069	94.712	3069.371	65661.086	0.603	1.306	5.547	0.536	1.444	2.001	0.464	98.233	2733.341	0.618	0.572	0.633	0.505	0.566	0.434	0.495	0.587	0.648
d_2^{21}	1.243	1.045	1.387	3370.628	72765.168	0.922	0.481	0.970	0.696	0.461	0.356	0.713	19.059	525.300	0.774	0.905	0.939	0.667	0.724	0.685	0.741	0.748	0.800
d_2^{22}	1.184	0.447	0.207	3384.020	59852.659	0.959	0.387	0.271	0.905	0.336	0.208	0.898	4.035	99.323	0.905	0.947	0.971	0.887	0.923	0.879	0.917	0.887	0.923
σ_2^2	1.899	0.382	0.156	418.851	9046.747	0.919	0.274	0.067	0.849	0.253	0.067	0.822	0.619	4.866	0.847	0.902	0.936	0.827	0.871	0.798	0.846	0.825	0.869
β_3^0	-4.842	5.592	35.921	61.341	1880.896	0.825	0.871	0.627	0.766	0.619	0.251	0.742	4.780	33.347	0.781	0.801	0.848	0.740	0.792	0.715	0.769	0.755	0.806
β_3^{c1}	-1.098	0.182	0.043	0.123	0.044	0.759	0.098	0.030	0.695	0.093	0.014	0.713	0.297	1.485	0.756	0.732	0.785	0.666	0.723	0.685	0.741	0.729	0.782
β_3^{c2}	3.918	0.304	0.099	0.201	0.104	0.865	0.180	0.133	0.768	0.132	0.054	0.729	1.120	9.954	0.797	0.844	0.886	0.742	0.794	0.701	0.756	0.772	0.822
β_3^{f11}	4.160	5.103	27.386	0.779	3.447	0.848	0.606	0.442	0.785	0.430	0.178	0.758	4.100	38.864	0.808	0.826	0.870	0.759	0.810	0.731	0.784	0.783	0.832
β_3^{f21}	-1.852	8.523	72.662	62.084	1880.932	0.737	0.763	0.734	0.671	0.521	0.228	0.644	5.230	39.932	0.692	0.709	0.764	0.642	0.700	0.614	0.673	0.663	0.720
β_3^{f22}	-1.553	10.332	108.838	336.498	8659.802	0.681	0.883	1.827	0.613	0.582	0.329	0.589	16.350	355.053	0.642	0.652	0.710	0.582	0.643	0.558	0.619	0.612	0.671
β_3^{tc}	0.597	0.291	0.094	0.304	0.098	0.927	0.225	0.072	0.838	0.225	0.040	0.852	0.600	2.456	0.854	0.911	0.943	0.815	0.861	0.830	0.874	0.832	0.876
d_3^{11}	1.528	3.341	11.272	0.821	1.455	0.892	1.058	1.834	0.773	0.494	0.810	0.760	39.317	604.590	0.822	0.873	0.911	0.747	0.799	0.733	0.786	0.798	0.846
d_3^{21}	0.814	0.601	0.362	0.547	0.355	0.961	0.453	0.368	0.866	0.327	0.169	0.853	6.887	86.096	0.898	0.949	0.973	0.845	0.887	0.831	0.875	0.879	0.917
d_3^{22}	1.384	0.439	0.193	0.613	0.271	0.954	0.394	0.206	0.868	0.400	0.124	0.893	1.872	13.079	0.875	0.941	0.967	0.847	0.889	0.874	0.912	0.854	0.895
σ_3^2	1.862	0.346	0.132	0.371	0.132	0.955	0.276	0.084	0.865	0.265	0.055	0.884	0.651	4.974	0.842	0.942	0.968	0.844	0.886	0.864	0.904	0.819	0.864
π_1	0.324	0.120	0.014	0.051	0.010	0.619	0.047	0.014	0.616				0.382	5.228	0.709	0.589	0.649	0.585	0.646			0.681	0.737
π_2	0.369	0.129	0.018	0.051	0.009	0.583	0.048	0.008	0.576				0.173	1.207	0.665	0.552	0.613	0.545	0.606			0.635	0.694
π_3	0.306	0.059	0.004	0.047	0.004	0.837	0.047	0.012	0.831				0.315	4.920	0.854	0.814	0.860	0.808	0.854			0.832	0.876

Table A.5 continued

parm	CI1L avg	CI1L std	CI1L n	CI2L avg	CI2L std	CI2L n	CI3L avg	CI3L std	CI3L n	CI4L avg	CI4L std	CI4L n
β_1^0	10471.698	213042.106	999	12.026	4.391	989	12.180	7.998	999	22.077	150.546	999
β_1^{c1}	465.230	11475.858	999	2.615	0.531	993	2.610	0.533	999	3.445	19.147	999
β_1^{c2}	5258.871	145696.426	999	4.363	1.020	988	4.370	1.007	999	6.777	36.133	999
β_1^{f11}	18066.535	476018.958	999	5.912	6.996	990	5.952	6.925	999	14.512	125.842	999
β_1^{f21}	14608.174	336558.322	999	9.776	9.718	991	10.035	11.813	999	18.668	106.021	999
β_1^{f22}	27798.300	418603.381	999	11.283	11.453	990	11.396	13.024	999	37.810	452.201	999
β_1^{tc}	835.180	23006.025	999	2.037	0.586	992	2.033	0.559	999	5.074	75.274	999
d_1^{11}	4558.619	140539.475	999	7.939	13.442	945	7.276	12.812	999	140.403	2687.060	999
d_1^{21}	2491.755	76149.955	999	2.067	1.572	972	1.912	1.685	999	16.947	193.946	999
d_1^{22}	1272.768	38140.428	999	2.156	0.749	988	2.114	0.759	999	10.919	218.807	999
σ_1^2	529.633	10211.065	999	2.989	0.618	995	2.992	0.625	999	5.184	38.150	999
β_2^0	16307.887	252143.171	998	16.717	8.694	996	16.505	4.369	999	47.369	669.346	999
β_2^{c1}	474.588	9067.368	998	4.016	0.448	997	4.015	0.444	999	4.324	5.538	999
β_2^{c2}	11856.871	217769.609	998	2.115	1.018	994	2.112	0.927	999	4.797	59.635	999
β_2^{f11}	17961.820	265557.560	998	3.904	6.115	991	3.950	5.832	999	9.837	78.455	999
β_2^{f21}	20682.808	408959.236	998	6.851	9.820	994	6.718	6.599	999	31.691	600.438	999
β_2^{f22}	40520.982	556534.136	999	8.511	14.344	996	8.413	11.523	999	33.094	603.699	999
β_2^{tc}	4444.458	126604.415	998	3.346	0.488	996	3.337	0.475	999	4.715	27.902	999
d_2^{11}	8512.203	181999.692	998	10.324	18.340	943	9.899	12.626	999	278.694	7576.077	999
d_2^{21}	9343.470	201691.016	998	2.397	2.897	981	2.244	1.688	999	53.742	1455.999	999
d_2^{22}	9380.672	165900.012	999	2.030	0.869	993	1.928	0.826	999	12.035	275.274	999
σ_2^2	1162.883	25075.754	998	2.796	0.523	998	2.780	0.548	999	3.645	13.383	999
β_3^0	177.851	5213.231	999	10.598	2.449	993	10.374	2.289	999	20.291	91.744	999
β_3^{c1}	1.595	0.250	999	1.580	0.255	990	1.574	0.256	999	2.038	3.994	999
β_3^{c2}	5.576	0.431	999	5.575	0.449	994	5.555	0.426	999	7.709	27.245	999
β_3^{f11}	6.937	11.598	999	6.673	6.875	988	6.431	6.856	999	15.671	107.585	999
β_3^{f21}	178.043	5213.397	999	8.702	9.215	990	8.317	9.244	999	20.374	110.446	999
β_3^{f22}	940.619	24003.001	999	11.232	10.989	994	10.806	10.246	999	53.083	983.861	999
β_3^{tc}	1.239	0.365	999	1.091	0.348	994	1.085	0.328	999	2.091	6.756	999
d_3^{11}	3.181	6.191	999	4.157	6.707	958	2.571	5.225	999	109.967	1675.751	999
d_3^{21}	1.955	1.224	999	1.813	1.172	978	1.497	0.920	999	19.555	238.609	999
d_3^{22}	2.621	0.894	999	2.286	0.712	989	2.252	0.701	999	6.278	36.138	999
σ_3^2	2.842	0.539	999	2.752	0.495	996	2.735	0.507	999	3.684	13.671	999
π_1	0.484	0.157	999	0.481	0.165	992				1.384	14.471	999
π_2	0.546	0.172	999	0.542	0.177	992				0.859	3.315	999
π_3	0.453	0.081	999	0.453	0.085	997				1.179	13.620	999

Table A.6: EM1st variant simulation results for constant(nmax=15)

parm	est avg	est std	est mse	SE1 avg	SE1 std	CI1 CP	SE2 avg	SE2 std	CI2 CP	SE3 avg	SE3 std	CI3 CP	SE4 avg	SE4 std	CI4 CP	CI1BLCL	CI1BUCL	CI2BLCL	CI2BUCL	CI3BLCL	CI3BUCL	CI4BLCL	CI4BUCL
β_1^0	7.745	1.859	3.478	1.369	9.787	0.947	0.493	0.317	0.866	0.495	0.288	0.892	1.058	7.409	0.852	0.933	0.961	0.845	0.887	0.873	0.911	0.830	0.874
β_1^{c1}	-1.992	0.099	0.010	0.076	0.331	0.957	0.048	0.008	0.900	0.049	0.006	0.919	0.054	0.064	0.889	0.944	0.970	0.881	0.918	0.902	0.936	0.869	0.908
β_1^{c2}	3.014	0.215	0.047	0.275	3.048	0.958	0.114	0.072	0.852	0.115	0.052	0.887	0.193	0.688	0.827	0.945	0.970	0.830	0.874	0.867	0.906	0.803	0.850
β_1^{f11}	1.014	1.583	2.507	1.323	16.277	0.955	0.332	0.207	0.848	0.337	0.150	0.880	0.839	6.166	0.839	0.942	0.968	0.825	0.870	0.860	0.900	0.816	0.862
β_1^{f21}	0.362	2.546	6.538	1.200	8.858	0.946	0.403	0.275	0.863	0.412	0.264	0.887	0.936	5.502	0.853	0.932	0.960	0.841	0.884	0.867	0.906	0.831	0.875
β_1^{f22}	1.338	2.726	7.502	8.740	154.317	0.952	0.413	0.284	0.848	0.425	0.302	0.862	1.561	24.721	0.832	0.939	0.965	0.825	0.870	0.840	0.883	0.808	0.855
β_1^{tc}	1.495	0.214	0.046	0.388	4.183	0.972	0.192	0.038	0.922	0.190	0.034	0.913	0.238	0.342	0.925	0.962	0.982	0.905	0.939	0.895	0.930	0.908	0.941
d_1^{11}	1.254	3.723	13.926	2.393	44.195	0.882	0.406	1.810	0.831	0.346	0.899	0.777	21.026	467.933	0.827	0.862	0.902	0.807	0.854	0.751	0.802	0.803	0.850
d_1^{21}	0.540	0.468	0.221	1.393	31.047	0.965	0.267	0.309	0.914	0.225	0.164	0.893	3.255	73.835	0.902	0.954	0.976	0.896	0.931	0.874	0.912	0.883	0.920
d_1^{22}	1.171	0.313	0.099	1.181	23.541	0.954	0.303	0.101	0.908	0.295	0.091	0.904	0.778	10.835	0.883	0.941	0.967	0.890	0.926	0.886	0.922	0.863	0.903
σ_1^2	1.897	0.151	0.023	0.254	2.003	0.965	0.133	0.020	0.908	0.138	0.018	0.917	0.164	0.776	0.876	0.954	0.976	0.890	0.926	0.900	0.934	0.855	0.896
β_2^0	12.452	0.906	0.824	0.773	4.929	0.964	0.474	0.236	0.920	0.432	0.224	0.908	0.821	3.101	0.919	0.952	0.976	0.903	0.937	0.890	0.926	0.902	0.936
β_2^{c1}	-2.991	0.094	0.009	0.061	0.015	0.949	0.048	0.005	0.918	0.047	0.004	0.923	0.056	0.131	0.914	0.935	0.963	0.901	0.935	0.906	0.939	0.896	0.931
β_2^{c2}	1.228	0.188	0.036	0.140	0.078	0.962	0.110	0.056	0.898	0.098	0.052	0.883	0.195	0.692	0.901	0.950	0.974	0.879	0.917	0.863	0.903	0.882	0.919
β_2^{f11}	-0.971	0.455	0.208	0.395	0.241	0.963	0.322	0.156	0.916	0.286	0.154	0.887	0.553	1.726	0.920	0.951	0.975	0.899	0.933	0.867	0.906	0.903	0.937
β_2^{f21}	2.959	0.867	0.753	0.667	4.961	0.958	0.388	0.200	0.915	0.344	0.192	0.885	0.627	1.454	0.924	0.945	0.970	0.898	0.932	0.865	0.905	0.907	0.940
β_2^{f22}	3.839	1.255	1.600	9662.872	256052.361	0.960	0.394	0.230	0.912	0.377	0.691	0.892	0.637	1.506	0.924	0.948	0.972	0.894	0.929	0.873	0.911	0.907	0.940
β_2^{tc}	2.480	0.202	0.041	0.228	0.052	0.967	0.190	0.031	0.936	0.180	0.027	0.920	0.226	0.109	0.943	0.956	0.978	0.921	0.951	0.903	0.937	0.928	0.957
d_2^{11}	1.351	2.728	7.643	0.555	1.081	0.885	0.356	0.751	0.865	0.348	0.572	0.793	3.126	45.255	0.861	0.865	0.905	0.844	0.886	0.767	0.818	0.839	0.882
d_2^{21}	0.747	0.432	0.189	0.351	0.229	0.968	0.250	0.123	0.912	0.233	0.122	0.883	0.529	2.833	0.912	0.957	0.979	0.894	0.929	0.863	0.903	0.894	0.929
d_2^{22}	1.075	0.281	0.080	0.376	0.149	0.952	0.299	0.087	0.932	0.266	0.073	0.902	0.419	0.477	0.916	0.939	0.965	0.916	0.948	0.883	0.920	0.899	0.933
σ_2^2	1.805	0.147	0.022	0.167	0.040	0.960	0.132	0.016	0.925	0.130	0.014	0.926	0.148	0.345	0.903	0.948	0.972	0.908	0.941	0.910	0.942	0.884	0.921
β_3^0	-6.957	1.187	1.410	3.068	75.239	0.961	0.484	0.360	0.893	0.487	0.274	0.913	1.121	12.120	0.864	0.949	0.973	0.874	0.912	0.895	0.930	0.842	0.885
β_3^{c1}	-1.002	0.052	0.003	0.060	0.013	0.959	0.048	0.006	0.934	0.047	0.004	0.924	0.061	0.295	0.929	0.947	0.971	0.918	0.949	0.907	0.940	0.913	0.945
β_3^{c2}	3.987	0.198	0.039	0.161	0.100	0.961	0.110	0.056	0.875	0.113	0.067	0.894	0.160	0.709	0.854	0.949	0.973	0.854	0.895	0.875	0.913	0.832	0.876
β_3^{f11}	2.988	1.112	1.237	2.840	75.250	0.954	0.322	0.159	0.877	0.330	0.186	0.886	0.507	3.150	0.875	0.941	0.967	0.856	0.897	0.866	0.906	0.854	0.895
β_3^{f21}	-1.793	1.894	3.630	110.439	3465.311	0.952	0.397	0.349	0.873	0.397	0.238	0.888	0.902	10.351	0.871	0.939	0.965	0.852	0.893	0.868	0.907	0.850	0.892
β_3^{f22}	-2.913	1.399	1.966	0.627	1.372	0.957	0.432	0.755	0.881	0.401	0.232	0.899	2.482	43.784	0.874	0.944	0.970	0.861	0.901	0.880	0.918	0.853	0.894
β_3^{tc}	0.512	0.227	0.052	0.256	0.057	0.965	0.191	0.033	0.901	0.203	0.032	0.923	0.236	1.003	0.881	0.954	0.976	0.882	0.919	0.906	0.939	0.861	0.901
d_3^{11}	1.574	4.293	18.574	0.657	1.851	0.889	0.316	0.458	0.782	0.411	0.953	0.799	1.021	8.610	0.759	0.869	0.908	0.756	0.807	0.774	0.823	0.732	0.785
d_3^{21}	0.812	0.418	0.175	0.409	0.244	0.955	0.249	0.115	0.845	0.277	0.145	0.890	0.401	1.803	0.820	0.942	0.968	0.822	0.867	0.870	0.909	0.796	0.843
d_3^{22}	1.368	0.354	0.126	0.479	0.172	0.945	0.302	0.092	0.860	0.340	0.094	0.890	0.344	0.789	0.809	0.931	0.959	0.838	0.881	0.870	0.909	0.784	0.833
σ_3^2	1.756	0.133	0.018	0.166	0.038	0.982	0.132	0.016	0.941	0.127	0.014	0.942	0.152	0.380	0.914	0.974	0.990	0.926	0.956	0.927	0.956	0.896	0.931
π_1	0.330	0.057	0.003	0.047	0.003	0.913	0.047	0.011	0.915				0.170	3.193	0.929	0.895	0.930	0.898	0.932			0.913	0.945
π_2	0.337	0.054	0.003	0.047	0.002	0.927	0.047	0.012	0.926				0.175	3.678	0.932	0.911	0.943	0.910	0.942			0.916	0.948
π_3	0.334	0.050	0.002	0.047	0.003	0.948	0.047	0.002	0.946				0.078	0.548	0.950	0.934	0.962	0.932	0.960			0.936	0.963

Table A.6 continued

parm	CI1L avg	CI1L std	CI1L n	CI2L avg	CI2L std	CI2L n	CI3L avg	CI3L std	CI3L n	CI4L avg	CI4L std	CI4L n
β_1^0	13.233	26.497	998	11.318	1.199	998	11.315	1.173	998	12.686	19.907	998
β_1^{c1}	2.853	0.841	998	2.821	0.139	998	2.821	0.139	998	2.827	0.150	998
β_1^{c2}	4.582	8.321	998	4.274	0.291	996	4.276	0.303	998	4.436	1.583	998
β_1^{f11}	4.497	45.120	998	1.947	2.056	997	1.965	2.062	998	3.278	17.142	998
β_1^{f21}	3.960	24.726	998	1.942	3.264	996	2.004	3.323	998	3.298	15.546	998
β_1^{f22}	25.693	427.674	998	3.083	3.187	996	3.114	3.292	998	6.170	68.515	998
β_1^{tc}	2.626	11.542	998	2.184	0.292	998	2.183	0.290	998	2.264	0.877	998
d_1^{11}	7.426	122.581	998	2.253	7.037	989	2.023	5.823	998	59.365	1296.982	998
d_1^{21}	4.145	86.048	998	1.108	1.033	996	0.999	0.788	998	9.368	204.643	998
d_1^{22}	4.141	65.224	998	1.872	0.461	998	1.848	0.504	998	3.157	29.994	998
σ_1^2	2.931	5.476	998	2.708	0.214	998	2.710	0.216	998	2.799	2.061	998
β_2^0	18.188	13.124	998	17.693	0.936	998	17.684	0.922	998	18.231	7.644	998
β_2^{c1}	4.233	0.132	998	4.231	0.133	996	4.231	0.133	998	4.241	0.270	998
β_2^{c2}	1.788	0.296	998	1.769	0.277	998	1.761	0.282	998	1.923	1.834	998
β_2^{f11}	1.861	0.691	998	1.721	0.564	998	1.677	0.546	998	2.263	4.734	998
β_2^{f21}	5.017	13.650	998	4.413	1.000	998	4.393	0.951	998	4.846	3.846	998
β_2^{f22}	26788.048	709727.726	998	5.752	1.069	998	5.786	1.980	998	6.157	3.906	998
β_2^{tc}	3.568	0.277	998	3.549	0.282	998	3.544	0.281	998	3.576	0.292	998
d_2^{11}	2.467	4.877	998	2.234	4.204	995	2.145	4.168	998	9.863	125.425	998
d_2^{21}	1.451	0.856	998	1.291	0.647	998	1.243	0.691	998	2.013	7.828	998
d_2^{22}	1.855	0.534	998	1.748	0.401	998	1.691	0.442	998	2.028	1.207	998
σ_2^2	2.596	0.209	998	2.579	0.207	998	2.578	0.209	998	2.613	0.901	998
β_3^0	16.786	208.284	997	10.058	1.106	997	10.053	0.967	998	11.655	33.197	998
β_3^{c1}	1.427	0.073	997	1.424	0.074	998	1.423	0.074	998	1.451	0.779	998
β_3^{c2}	5.668	0.218	997	5.649	0.268	998	5.651	0.264	998	5.725	1.772	998
β_3^{f11}	11.138	208.478	997	4.441	1.263	998	4.450	1.274	998	4.858	8.656	998
β_3^{f21}	308.002	9605.115	997	3.285	2.229	997	3.275	2.127	998	4.609	28.667	998
β_3^{f22}	4.841	3.859	997	4.606	2.341	998	4.523	1.441	998	10.192	121.214	998
β_3^{tc}	1.040	0.250	997	0.921	0.263	998	0.941	0.260	998	1.044	2.775	998
d_3^{11}	2.887	7.943	997	2.258	4.886	991	2.505	6.620	998	4.389	24.496	998
d_3^{21}	1.639	0.842	997	1.376	0.594	997	1.398	0.685	998	1.808	4.960	998
d_3^{22}	2.360	0.643	997	2.123	0.506	997	2.154	0.559	998	2.270	2.129	998
σ_3^2	2.528	0.192	997	2.511	0.187	998	2.509	0.189	998	2.553	0.988	998
π_1	0.485	0.077	999	0.485	0.080	995				0.822	8.836	998
π_2	0.494	0.075	999	0.495	0.080	997				0.845	10.181	998
π_3	0.490	0.068	999	0.490	0.069	996				0.570	1.501	998

Table A.7: EM1st variant simulation results for non-constant(nmax=6)

parm	est avg	est std	est mse	SE1 avg	SE1 std	CI1 CP	SE2 avg	SE2 std	CI2 CP	SE3 avg	SE3 std	CI3 CP	SE4 avg	SE4 std	CI4 CP	CI1BLCL	CI1BUCL	CI2BLCL	CI2BUCL	CI3BLCL	CI3BUCL	CI4BLCL	CI4BUCL
β_1^0	7.880	0.465	0.217	0.591	0.145	0.981	0.440	0.049	0.933	0.444	0.049	0.942	0.452	0.128	0.920	0.973	0.990	0.917	0.948	0.927	0.957	0.903	0.937
β_1^{c1}	-1.992	0.087	0.008	0.114	0.028	0.984	0.082	0.009	0.938	0.084	0.009	0.939	0.082	0.024	0.906	0.977	0.992	0.923	0.953	0.924	0.954	0.887	0.924
β_1^{c2}	3.000	0.096	0.009	0.122	0.030	0.983	0.088	0.010	0.925	0.090	0.010	0.941	0.088	0.025	0.891	0.975	0.991	0.909	0.942	0.926	0.956	0.871	0.911
β_1^{f11}	1.004	0.258	0.067	0.326	0.073	0.982	0.238	0.025	0.918	0.243	0.024	0.936	0.238	0.060	0.905	0.974	0.991	0.901	0.935	0.920	0.951	0.886	0.923
β_1^{f21}	0.585	0.319	0.102	0.406	0.092	0.976	0.300	0.032	0.924	0.305	0.031	0.934	0.300	0.081	0.902	0.967	0.986	0.908	0.941	0.918	0.949	0.884	0.921
β_1^{f22}	1.591	0.311	0.097	0.399	0.088	0.985	0.295	0.031	0.929	0.301	0.030	0.944	0.295	0.075	0.902	0.978	0.993	0.913	0.946	0.929	0.959	0.884	0.921
β_1^{tc}	1.496	0.206	0.043	0.256	0.055	0.983	0.204	0.031	0.932	0.201	0.029	0.937	0.222	0.075	0.917	0.975	0.991	0.916	0.947	0.921	0.952	0.900	0.934
d_1^{11}	0.943	0.345	0.122	0.500	0.182	0.954	0.358	0.106	0.923	0.340	0.091	0.888	0.431	0.328	0.892	0.941	0.968	0.906	0.940	0.868	0.908	0.873	0.912
d_1^{21}	0.480	0.247	0.061	0.363	0.122	0.978	0.270	0.072	0.941	0.247	0.057	0.934	0.331	0.238	0.914	0.969	0.987	0.926	0.956	0.918	0.949	0.896	0.932
d_1^{22}	1.157	0.333	0.113	0.471	0.163	0.954	0.342	0.096	0.924	0.328	0.086	0.896	0.403	0.284	0.890	0.941	0.968	0.908	0.941	0.877	0.916	0.870	0.910
σ_1^2	1.809	0.254	0.073	0.344	0.088	0.954	0.243	0.040	0.901	0.252	0.039	0.902	0.235	0.086	0.863	0.941	0.968	0.883	0.920	0.884	0.921	0.841	0.885
β_2^0	12.504	0.439	0.192	0.569	0.138	0.985	0.440	0.050	0.951	0.427	0.046	0.942	0.475	0.150	0.940	0.978	0.993	0.938	0.965	0.927	0.957	0.925	0.955
β_2^{c1}	-3.001	0.081	0.007	0.109	0.026	0.982	0.082	0.010	0.947	0.080	0.009	0.941	0.087	0.029	0.938	0.974	0.991	0.933	0.961	0.926	0.956	0.923	0.953
β_2^{c2}	1.200	0.086	0.007	0.115	0.027	0.978	0.088	0.010	0.948	0.086	0.009	0.948	0.094	0.029	0.932	0.969	0.987	0.934	0.962	0.934	0.962	0.916	0.947
β_2^{f11}	-1.010	0.237	0.056	0.311	0.067	0.985	0.238	0.026	0.944	0.232	0.024	0.943	0.251	0.072	0.919	0.978	0.993	0.929	0.959	0.928	0.958	0.902	0.936
β_2^{f21}	3.010	0.304	0.093	0.391	0.090	0.980	0.299	0.033	0.941	0.291	0.030	0.935	0.315	0.090	0.934	0.972	0.989	0.926	0.956	0.919	0.950	0.918	0.949
β_2^{f22}	4.008	0.298	0.089	0.384	0.089	0.984	0.294	0.032	0.944	0.288	0.030	0.944	0.308	0.085	0.925	0.977	0.992	0.929	0.959	0.929	0.959	0.909	0.942
β_2^{tc}	2.506	0.198	0.039	0.244	0.054	0.967	0.202	0.031	0.935	0.192	0.029	0.932	0.228	0.078	0.936	0.956	0.978	0.919	0.950	0.916	0.947	0.920	0.951
d_2^{11}	0.865	0.339	0.116	0.457	0.163	0.955	0.354	0.100	0.927	0.314	0.088	0.891	0.460	0.335	0.908	0.942	0.968	0.911	0.944	0.871	0.911	0.889	0.926
d_2^{21}	0.687	0.259	0.067	0.353	0.129	0.961	0.266	0.070	0.926	0.245	0.064	0.911	0.334	0.213	0.907	0.948	0.973	0.910	0.943	0.893	0.929	0.888	0.925
d_2^{22}	1.061	0.319	0.103	0.429	0.158	0.938	0.335	0.095	0.921	0.302	0.084	0.885	0.432	0.314	0.905	0.923	0.953	0.904	0.938	0.865	0.905	0.886	0.923
σ_2^2	1.728	0.248	0.066	0.329	0.090	0.955	0.243	0.041	0.910	0.240	0.039	0.908	0.246	0.095	0.876	0.942	0.968	0.892	0.928	0.889	0.926	0.855	0.896
β_3^0	-6.983	0.456	0.208	0.582	0.141	0.975	0.440	0.049	0.936	0.439	0.048	0.936	0.456	0.139	0.912	0.965	0.985	0.920	0.951	0.920	0.951	0.894	0.930
β_3^{c1}	-0.999	0.086	0.007	0.110	0.027	0.980	0.082	0.010	0.938	0.081	0.009	0.943	0.086	0.026	0.914	0.972	0.989	0.923	0.953	0.928	0.958	0.896	0.932
β_3^{c2}	4.004	0.089	0.008	0.118	0.029	0.983	0.089	0.010	0.941	0.087	0.010	0.938	0.093	0.028	0.922	0.975	0.991	0.926	0.956	0.923	0.953	0.905	0.939
β_3^{f11}	3.002	0.241	0.058	0.310	0.067	0.981	0.239	0.026	0.939	0.233	0.024	0.933	0.251	0.075	0.935	0.973	0.990	0.924	0.954	0.917	0.948	0.919	0.950
β_3^{f21}	-2.020	0.308	0.095	0.393	0.091	0.971	0.300	0.032	0.935	0.294	0.031	0.942	0.313	0.087	0.917	0.960	0.982	0.919	0.950	0.927	0.957	0.900	0.934
β_3^{f22}	-3.010	0.287	0.082	0.384	0.088	0.981	0.295	0.032	0.953	0.289	0.030	0.953	0.307	0.084	0.937	0.973	0.990	0.940	0.967	0.940	0.967	0.921	0.952
β_3^{tc}	0.503	0.219	0.048	0.272	0.056	0.979	0.203	0.031	0.914	0.215	0.030	0.942	0.203	0.059	0.888	0.970	0.988	0.896	0.932	0.927	0.957	0.868	0.908
d_3^{11}	1.173	0.405	0.165	0.569	0.209	0.951	0.350	0.098	0.871	0.389	0.105	0.898	0.351	0.206	0.822	0.938	0.965	0.850	0.893	0.879	0.917	0.797	0.846
d_3^{21}	0.800	0.308	0.095	0.436	0.144	0.978	0.266	0.071	0.879	0.302	0.074	0.927	0.255	0.124	0.827	0.969	0.987	0.858	0.899	0.911	0.944	0.803	0.851
d_3^{22}	1.372	0.387	0.151	0.543	0.184	0.954	0.339	0.096	0.880	0.378	0.098	0.908	0.327	0.185	0.817	0.941	0.968	0.859	0.900	0.889	0.926	0.793	0.842
σ_3^2	1.682	0.256	0.070	0.321	0.086	0.947	0.245	0.041	0.904	0.233	0.039	0.881	0.258	0.100	0.899	0.933	0.961	0.885	0.922	0.860	0.901	0.880	0.918
π_1	0.332	0.048	0.002	0.047	0.002	0.947	0.047	0.002	0.947				0.047	0.002	0.947	0.933	0.961	0.933	0.961			0.933	0.961
π_2	0.334	0.047	0.002	0.047	0.002	0.960	0.047	0.002	0.960				0.047	0.002	0.960	0.947	0.972	0.947	0.972			0.947	0.972
π_3	0.334	0.046	0.002	0.047	0.002	0.961	0.047	0.002	0.961				0.047	0.002	0.961	0.948	0.973	0.948	0.973			0.948	0.973

Table A.7 continued

parm	CI1L avg	CI1L std	CI1L n	CI2L avg	CI2L std	CI2L n	CI3L avg	CI3L std	CI3L n	CI4L avg	CI4L std	CI4L n
β_1^0	11.271	0.655	964	11.211	0.654	964	11.212	0.653	964	11.219	0.655	964
β_1^{c1}	2.836	0.122	964	2.827	0.123	964	2.827	0.122	964	2.827	0.123	964
β_1^{c2}	4.258	0.136	964	4.250	0.136	964	4.251	0.136	964	4.251	0.136	964
β_1^{f11}	1.704	0.326	964	1.576	0.327	964	1.583	0.324	964	1.581	0.334	964
β_1^{f21}	1.452	0.330	964	1.223	0.302	964	1.233	0.300	964	1.231	0.340	964
β_1^{f22}	2.525	0.403	964	2.400	0.412	964	2.406	0.410	964	2.407	0.419	964
β_1^{tc}	2.238	0.282	964	2.192	0.284	964	2.190	0.282	964	2.213	0.294	964
d_1^{11}	1.941	0.655	964	1.693	0.473	964	1.638	0.536	964	1.902	0.805	964
d_1^{21}	1.240	0.414	964	1.046	0.297	964	0.986	0.324	964	1.222	0.607	964
d_1^{22}	2.106	0.609	964	1.912	0.462	964	1.874	0.524	964	2.081	0.659	964
σ_1^2	2.738	0.387	964	2.648	0.358	964	2.652	0.372	964	2.652	0.351	964
β_2^0	17.757	0.620	964	17.725	0.619	964	17.723	0.619	964	17.737	0.619	964
β_2^{c1}	4.255	0.115	964	4.250	0.115	964	4.250	0.115	964	4.252	0.115	964
β_2^{c2}	1.729	0.121	964	1.716	0.121	964	1.715	0.121	964	1.719	0.122	964
β_2^{f11}	1.685	0.301	964	1.581	0.302	964	1.573	0.304	964	1.605	0.316	964
β_2^{f21}	4.401	0.421	964	4.338	0.422	964	4.334	0.421	964	4.353	0.424	964
β_2^{f22}	5.773	0.414	964	5.728	0.417	964	5.725	0.416	964	5.738	0.419	964
β_2^{tc}	3.611	0.275	964	3.589	0.276	964	3.584	0.277	964	3.606	0.277	964
d_2^{11}	1.775	0.616	964	1.599	0.459	964	1.505	0.525	964	1.876	0.833	964
d_2^{21}	1.395	0.462	964	1.241	0.342	964	1.190	0.388	964	1.410	0.543	964
d_2^{22}	1.927	0.589	964	1.787	0.442	964	1.718	0.503	964	2.029	0.729	964
σ_2^2	2.616	0.383	964	2.538	0.347	964	2.533	0.363	964	2.552	0.338	964
β_3^0	10.014	0.640	964	9.951	0.640	964	9.951	0.640	964	9.963	0.643	964
β_3^{c1}	1.448	0.119	964	1.432	0.120	964	1.431	0.120	964	1.435	0.120	964
β_3^{c2}	5.672	0.126	964	5.668	0.126	964	5.668	0.126	964	5.669	0.126	964
β_3^{f11}	4.335	0.335	964	4.297	0.336	964	4.295	0.336	964	4.307	0.340	964
β_3^{f21}	3.070	0.420	964	2.979	0.419	964	2.974	0.420	964	2.996	0.424	964
β_3^{f22}	4.396	0.396	964	4.336	0.399	964	4.333	0.399	964	4.348	0.403	964
β_3^{tc}	1.066	0.238	964	0.930	0.245	964	0.952	0.237	964	0.936	0.265	964
d_3^{11}	2.306	0.764	964	1.946	0.556	964	1.981	0.633	964	1.998	0.603	964
d_3^{21}	1.675	0.532	964	1.372	0.412	964	1.416	0.454	964	1.379	0.431	964
d_3^{22}	2.470	0.700	964	2.172	0.546	964	2.206	0.607	964	2.199	0.560	964
σ_3^2	2.547	0.393	964	2.477	0.358	964	2.466	0.375	964	2.502	0.348	964
π_1	0.487	0.066	999	0.487	0.066	964				0.487	0.066	964
π_2	0.491	0.065	999	0.491	0.065	964				0.491	0.065	964
π_3	0.490	0.065	999	0.490	0.065	964				0.490	0.065	964

Table A.8: EM2nd variant simulation results for constant(nmax=6)

parm	est avg	est std	est mse	SE1 avg	SE1 std	CI1 CP	SE2 avg	SE2 std	CI2 CP	SE3 avg	SE3 std	CI3 CP	SE4 avg	SE4 std	CI4 CP	CI1BLCL	CI1BUCL	CI2BLCL	CI2BUCL	CI3BLCL	CI3BUCL	CI4BLCL	CI4BUCL
β_1^0	7.862	1.620	2.625	368.727	11627.008	0.965	0.556	0.199	0.870	0.552	0.180	0.872	0.712	1.152	0.849	0.954	0.976	0.849	0.891	0.851	0.893	0.827	0.871
β_1^{c1}	-1.993	0.121	0.015	0.123	0.084	0.956	0.088	0.011	0.902	0.089	0.011	0.912	0.095	0.152	0.889	0.943	0.969	0.884	0.920	0.894	0.930	0.870	0.908
β_1^{c2}	3.005	0.380	0.144	7150.563	226000.074	0.954	0.142	0.056	0.828	0.143	0.049	0.850	0.179	0.287	0.805	0.941	0.967	0.805	0.851	0.828	0.872	0.780	0.830
β_1^{f11}	1.013	1.405	1.974	297.964	9396.971	0.958	0.376	0.140	0.879	0.384	0.124	0.898	0.457	0.586	0.873	0.946	0.970	0.859	0.899	0.879	0.917	0.852	0.894
β_1^{f21}	0.432	2.019	4.103	0.800	3.078	0.943	0.454	0.163	0.850	0.459	0.135	0.876	0.598	1.122	0.860	0.929	0.957	0.828	0.872	0.856	0.896	0.838	0.882
β_1^{f22}	1.455	2.416	5.860	218.351	6877.149	0.934	0.457	0.164	0.856	0.468	0.142	0.864	0.596	1.058	0.835	0.919	0.949	0.834	0.878	0.843	0.885	0.812	0.858
β_1^{tc}	1.502	0.219	0.048	0.260	0.088	0.966	0.203	0.034	0.931	0.200	0.035	0.922	0.232	0.122	0.937	0.955	0.977	0.915	0.947	0.905	0.939	0.922	0.952
d_1^{11}	1.020	1.803	3.251	46.416	1444.899	0.888	0.389	0.566	0.809	0.372	0.683	0.755	0.989	6.090	0.805	0.868	0.908	0.785	0.833	0.728	0.782	0.780	0.830
d_1^{21}	0.521	0.477	0.228	431.816	13634.796	0.966	0.281	0.143	0.901	0.250	0.170	0.886	0.484	1.674	0.900	0.955	0.977	0.882	0.920	0.866	0.906	0.881	0.919
d_1^{22}	1.153	0.367	0.137	0.490	0.334	0.935	0.338	0.104	0.889	0.328	0.101	0.874	0.454	0.828	0.872	0.920	0.950	0.870	0.908	0.853	0.895	0.851	0.893
σ_1^2	1.883	0.268	0.072	0.358	0.113	0.976	0.251	0.045	0.919	0.260	0.045	0.931	0.251	0.209	0.868	0.967	0.985	0.902	0.936	0.915	0.947	0.847	0.889
β_2^0	12.440	0.664	0.444	368.597	11627.012	0.957	0.550	0.140	0.908	0.498	0.158	0.891	0.733	0.686	0.918	0.944	0.970	0.890	0.926	0.872	0.910	0.901	0.935
β_2^{c1}	-2.995	0.100	0.010	0.113	0.027	0.969	0.088	0.012	0.926	0.084	0.009	0.916	0.097	0.053	0.917	0.958	0.980	0.910	0.942	0.899	0.933	0.900	0.934
β_2^{c2}	1.211	0.199	0.040	7150.516	226000.076	0.945	0.142	0.042	0.897	0.127	0.045	0.866	0.196	0.198	0.891	0.931	0.959	0.878	0.916	0.845	0.887	0.872	0.910
β_2^{f11}	-0.948	0.455	0.210	297.786	9396.976	0.971	0.374	0.110	0.918	0.338	0.117	0.876	0.505	0.347	0.923	0.961	0.981	0.901	0.935	0.856	0.896	0.906	0.940
β_2^{f21}	2.938	1.030	1.064	0.601	0.341	0.962	0.450	0.124	0.920	0.407	0.139	0.903	0.604	0.444	0.919	0.950	0.974	0.903	0.937	0.885	0.921	0.902	0.936
β_2^{f22}	3.878	1.258	1.598	218.236	6877.152	0.958	0.450	0.139	0.908	0.413	0.166	0.884	0.596	0.404	0.914	0.946	0.970	0.890	0.926	0.864	0.904	0.897	0.931
β_2^{tc}	2.496	0.208	0.043	0.244	0.053	0.966	0.202	0.034	0.942	0.191	0.029	0.919	0.239	0.120	0.936	0.955	0.977	0.928	0.956	0.902	0.936	0.921	0.951
d_2^{11}	1.088	1.964	3.891	46.282	1444.899	0.881	0.361	0.270	0.833	0.358	0.420	0.760	0.554	1.181	0.840	0.861	0.901	0.810	0.856	0.734	0.786	0.817	0.863
d_2^{21}	0.725	0.385	0.149	431.772	13634.797	0.954	0.276	0.100	0.885	0.251	0.104	0.861	0.401	0.419	0.899	0.941	0.967	0.865	0.905	0.840	0.882	0.880	0.918
d_2^{22}	1.068	0.331	0.111	0.436	0.153	0.951	0.336	0.105	0.914	0.302	0.084	0.883	0.457	0.524	0.879	0.938	0.964	0.897	0.931	0.863	0.903	0.859	0.899
σ_2^2	1.781	0.251	0.063	0.326	0.090	0.970	0.249	0.045	0.930	0.244	0.039	0.928	0.260	0.124	0.903	0.959	0.981	0.914	0.946	0.912	0.944	0.885	0.921
β_3^0	-6.961	0.979	0.961	0.745	0.329	0.969	0.560	0.206	0.903	0.532	0.144	0.896	0.855	3.420	0.898	0.958	0.980	0.885	0.921	0.877	0.915	0.879	0.917
β_3^{c1}	-1.000	0.092	0.008	0.114	0.027	0.969	0.088	0.011	0.941	0.085	0.009	0.934	0.098	0.082	0.926	0.958	0.980	0.926	0.956	0.919	0.949	0.910	0.942
β_3^{c2}	3.995	0.166	0.028	0.198	0.083	0.956	0.144	0.058	0.883	0.136	0.042	0.876	0.202	0.384	0.865	0.943	0.969	0.863	0.903	0.856	0.896	0.844	0.886
β_3^{f11}	2.968	0.724	0.525	0.506	0.168	0.957	0.383	0.153	0.895	0.366	0.091	0.885	0.623	2.856	0.895	0.944	0.970	0.876	0.914	0.865	0.905	0.876	0.914
β_3^{f21}	-1.949	1.370	1.880	0.630	0.309	0.953	0.456	0.169	0.879	0.441	0.116	0.885	0.702	3.199	0.854	0.940	0.966	0.859	0.899	0.865	0.905	0.832	0.876
β_3^{f22}	-2.941	1.512	2.291	0.634	0.460	0.955	0.462	0.182	0.886	0.444	0.126	0.901	0.721	3.134	0.886	0.942	0.968	0.866	0.906	0.882	0.920	0.866	0.906
β_3^{tc}	0.500	0.228	0.052	0.270	0.056	0.973	0.205	0.035	0.901	0.214	0.031	0.925	0.229	0.295	0.886	0.963	0.983	0.882	0.920	0.909	0.941	0.866	0.906
d_3^{11}	1.069	1.262	1.610	0.556	0.555	0.904	0.409	0.558	0.820	0.363	0.329	0.807	1.677	13.759	0.796	0.886	0.922	0.796	0.844	0.783	0.831	0.771	0.821
d_3^{21}	0.790	0.327	0.107	0.435	0.178	0.965	0.294	0.172	0.876	0.288	0.088	0.890	0.540	2.330	0.866	0.954	0.976	0.856	0.896	0.871	0.909	0.845	0.887
d_3^{22}	1.360	0.381	0.147	0.527	0.181	0.946	0.345	0.106	0.889	0.375	0.099	0.899	0.446	1.751	0.837	0.932	0.960	0.870	0.908	0.880	0.918	0.814	0.860
σ_3^2	1.736	0.245	0.060	0.319	0.091	0.973	0.250	0.044	0.950	0.239	0.039	0.931	0.267	0.148	0.911	0.963	0.983	0.936	0.964	0.915	0.947	0.893	0.929
π_1	0.331	0.054	0.003	1310.866	41430.963	0.923	0.047	0.002	0.918				0.199	4.704	0.928	0.906	0.940	0.901	0.935			0.912	0.944
π_2	0.336	0.052	0.003	1310.866	41430.963	0.942	0.047	0.002	0.934				0.198	4.704	0.940	0.928	0.956	0.919	0.949			0.925	0.955
π_3	0.333	0.047	0.002	0.047	0.002	0.956	0.047	0.002	0.955				0.049	0.036	0.956	0.943	0.969	0.942	0.968			0.943	0.969

Table A.8 continued

parm	CI1L avg	CI1L std	CI1L n	CI2L avg	CI2L std	CI2L n	CI3L avg	CI3L std	CI3L n	CI4L avg	CI4L std	CI4L n
β_1^0	1031.271	32227.539	1000	11.3645	1.53251	997	11.36	1.552	1000	11.59544	2.9174	1000
β_1^{c1}	2.846	0.207	999	2.82935	0.16971	999	2.8296	0.1689	1000	2.8419	0.3798	1000
β_1^{c2}	19823.683	626428.617	1000	4.28143	0.43917	997	4.2819	0.4381	1000	4.32955	0.694	1000
β_1^{f11}	826.642	26046.573	1000	1.9549	1.84795	999	1.9725	1.8334	1000	2.13023	2.3893	1000
β_1^{f21}	2.808	8.850	999	1.96253	2.54854	996	1.9702	2.5296	1000	2.31438	3.9477	1000
β_1^{f22}	606.750	19062.085	1000	3.04032	2.80983	997	3.0767	2.8783	1000	3.36431	3.9896	1000
β_1^{tc}	2.254	0.316	999	2.20119	0.29175	998	2.1988	0.2973	1000	2.23949	0.345	1000
d_1^{11}	129.199	4004.966	1000	1.89475	2.93867	993	1.7847	3.1686	1000	3.52181	16.99	1000
d_1^{21}	1197.163	37793.024	1000	1.12651	0.69877	995	1.0405	0.7866	1000	1.67507	4.6402	1000
d_1^{22}	2.148	1.007	999	1.90387	0.51509	999	1.8684	0.5845	1000	2.21002	2.2109	1000
σ_1^2	2.851	0.428	999	2.75481	0.37797	1000	2.7591	0.3938	1000	2.77913	0.5745	1000
β_2^0	1037.380	32227.345	1000	17.6651	0.92829	997	17.654	0.9166	1000	17.78676	1.3339	1000
β_2^{c1}	4.249	0.140	999	4.2439	0.13887	999	4.2424	0.1417	1000	4.2472	0.1395	1000
β_2^{c2}	19821.128	626428.698	1000	1.761	0.28064	997	1.751	0.2873	1000	1.83694	0.4806	1000
β_2^{f11}	826.059	26046.589	1000	1.76788	0.50566	999	1.7121	0.5135	1000	2.05449	0.933	1000
β_2^{f21}	4.649	1.203	999	4.45811	1.08771	998	4.4318	1.08	1000	4.6822	1.3348	1000
β_2^{f22}	609.201	19062.008	1000	5.80842	1.10214	999	5.7923	1.1041	1000	5.97656	1.2313	1000
β_2^{tc}	3.599	0.290	999	3.5769	0.28939	999	3.5712	0.2916	1000	3.60623	0.3096	1000
d_2^{11}	128.930	4004.964	1000	1.89562	2.72047	992	1.8461	3.0022	1000	2.41045	4.1642	1000
d_2^{21}	1197.219	37793.022	1000	1.30891	0.53975	995	1.2467	0.5991	1000	1.61182	1.1535	1000
d_2^{22}	1.949	0.587	999	1.79669	0.47441	997	1.7271	0.5187	1000	2.09723	1.3465	1000
σ_2^2	2.683	0.382	999	2.61523	0.35396	1000	2.6083	0.3673	1000	2.64002	0.3744	1000
β_3^0	10.148	0.982	1000	10.0348	0.93495	1000	10.019	0.8885	1000	10.6009	9.0534	1000
β_3^{c1}	1.451	0.128	1000	1.43589	0.12765	1000	1.4345	0.1279	1000	1.44915	0.206	1000
β_3^{c2}	5.681	0.239	1000	5.6664	0.23411	1000	5.6637	0.2348	1000	5.73461	0.7343	1000
β_3^{f11}	4.498	0.793	1000	4.39873	0.79207	999	4.3783	0.7725	1000	4.92737	7.7496	1000
β_3^{f21}	3.527	1.640	1000	3.28044	1.55334	999	3.2513	1.5414	1000	3.88106	8.8708	1000
β_3^{f22}	4.790	1.909	1000	4.59747	1.62388	999	4.5705	1.619	1000	5.1874	8.6269	1000
β_3^{tc}	1.062	0.244	1000	0.93594	0.24805	1000	0.9509	0.2413	1000	0.99803	0.8242	1000
d_3^{11}	2.177	2.340	1000	1.97828	2.28995	997	1.8207	1.9993	1000	5.50228	38.096	1000
d_3^{21}	1.667	0.617	1000	1.42085	0.57993	999	1.3837	0.4977	1000	2.09544	6.4062	1000
d_3^{22}	2.429	0.686	1000	2.1658	0.54493	1000	2.1865	0.6002	1000	2.46702	4.7945	1000
σ_3^2	2.617	0.384	1000	2.55446	0.34415	1000	2.5435	0.3605	1000	2.58882	0.4067	1000
π_1	3633.821	114838.650	1000	0.48643	0.07424	999				0.9049	13.027	1000
π_2	3633.828	114838.650	1000	0.49302	0.07277	999				0.90687	13.027	1000
π_3	0.489	0.065	1000	0.48902	0.06521	1000				0.49317	0.1065	1000

Table A.9: EM2nd variant simulation results for constant(nmax=15)

parm	est avg	est std	est mse	SE1 avg	SE1 std	CI1 CP	SE2 avg	SE2 std	CI2 CP	SE3 avg	SE3 std	CI3 CP	SE4 avg	SE4 std	CI4 CP	CI1BLCL	CI1BUCL	CI2BLCL	CI2BUCL	CI3BLCL	CI3BUCL	CI4BLCL	CI4BUCL
β_1^0	7.921	0.527	0.278	0.643	0.200	0.966	0.453	0.110	0.895	0.463	0.084	0.910	0.565	2.109	0.875	0.955	0.977	0.876	0.914	0.892	0.928	0.855	0.896
β_1^{c1}	-1.998	0.048	0.002	0.062	0.014	0.981	0.047	0.005	0.938	0.048	0.004	0.948	0.048	0.071	0.905	0.973	0.989	0.923	0.953	0.934	0.962	0.887	0.923
β_1^{c2}	3.001	0.137	0.019	0.160	0.052	0.968	0.108	0.028	0.865	0.112	0.023	0.902	0.131	0.469	0.831	0.957	0.979	0.844	0.886	0.884	0.920	0.808	0.854
β_1^{f11}	0.979	0.369	0.137	0.431	0.113	0.960	0.301	0.072	0.876	0.316	0.051	0.896	0.369	1.457	0.853	0.948	0.972	0.856	0.896	0.877	0.915	0.831	0.875
β_1^{f21}	0.618	0.466	0.217	7.832	229.379	0.972	0.385	0.089	0.878	0.406	0.077	0.898	0.456	1.368	0.843	0.962	0.982	0.858	0.898	0.879	0.917	0.820	0.866
β_1^{f22}	1.589	0.423	0.179	0.504	0.153	0.967	0.351	0.074	0.883	0.369	0.061	0.908	0.403	0.913	0.862	0.956	0.978	0.863	0.903	0.890	0.926	0.841	0.883
β_1^{tc}	1.503	0.182	0.033	0.237	0.054	0.980	0.190	0.030	0.944	0.187	0.028	0.943	0.212	0.098	0.936	0.971	0.989	0.930	0.958	0.929	0.957	0.921	0.951
d_1^{11}	0.847	0.280	0.102	0.369	0.144	0.888	0.269	0.085	0.852	0.244	0.072	0.786	0.359	0.420	0.842	0.868	0.908	0.830	0.874	0.761	0.811	0.819	0.865
d_1^{21}	0.485	0.237	0.057	0.311	0.120	0.956	0.237	0.070	0.919	0.205	0.053	0.891	0.338	0.753	0.896	0.943	0.969	0.902	0.936	0.872	0.910	0.877	0.915
d_1^{22}	1.156	0.295	0.089	0.408	0.148	0.953	0.299	0.088	0.918	0.288	0.078	0.891	0.365	0.280	0.901	0.940	0.966	0.901	0.935	0.872	0.910	0.882	0.920
σ_1^2	1.887	0.140	0.020	0.175	0.043	0.982	0.130	0.016	0.921	0.136	0.014	0.945	0.132	0.216	0.875	0.974	0.990	0.904	0.938	0.931	0.959	0.855	0.896
β_2^0	12.508	0.452	0.205	0.541	0.155	0.959	0.454	0.097	0.937	0.392	0.064	0.907	0.655	1.270	0.935	0.947	0.971	0.922	0.952	0.889	0.925	0.920	0.950
β_2^{c1}	-2.998	0.046	0.002	0.060	0.014	0.982	0.047	0.004	0.941	0.047	0.004	0.937	0.047	0.012	0.928	0.974	0.990	0.926	0.956	0.922	0.952	0.912	0.944
β_2^{c2}	1.202	0.103	0.011	0.130	0.039	0.972	0.108	0.026	0.940	0.091	0.017	0.915	0.158	0.249	0.941	0.962	0.982	0.925	0.955	0.898	0.932	0.926	0.956
β_2^{f11}	-0.991	0.298	0.089	0.352	0.091	0.959	0.302	0.057	0.943	0.255	0.039	0.904	0.426	0.241	0.956	0.947	0.971	0.929	0.957	0.886	0.922	0.943	0.969
β_2^{f21}	2.983	0.366	0.134	0.464	0.183	0.971	0.386	0.080	0.934	0.325	0.057	0.909	0.553	0.510	0.944	0.961	0.981	0.919	0.949	0.891	0.927	0.930	0.958
β_2^{f22}	3.994	0.339	0.115	0.406	0.107	0.967	0.352	0.069	0.943	0.297	0.047	0.909	0.487	0.260	0.949	0.956	0.978	0.929	0.957	0.891	0.927	0.935	0.963
β_2^{tc}	2.503	0.184	0.034	0.226	0.049	0.971	0.191	0.031	0.950	0.180	0.028	0.945	0.231	0.355	0.956	0.961	0.981	0.936	0.964	0.931	0.959	0.943	0.969
d_2^{11}	0.809	0.257	0.074	0.351	0.125	0.918	0.272	0.134	0.900	0.234	0.067	0.826	0.466	3.177	0.885	0.901	0.935	0.881	0.919	0.803	0.850	0.865	0.905
d_2^{21}	0.685	0.236	0.056	0.312	0.108	0.948	0.239	0.107	0.917	0.213	0.058	0.888	0.403	2.579	0.916	0.934	0.962	0.900	0.934	0.868	0.908	0.899	0.933
d_2^{22}	1.067	0.287	0.084	0.377	0.137	0.948	0.300	0.106	0.917	0.267	0.075	0.889	0.453	1.999	0.908	0.934	0.962	0.900	0.934	0.870	0.908	0.890	0.926
σ_2^2	1.794	0.129	0.017	0.169	0.037	0.987	0.130	0.015	0.945	0.129	0.013	0.948	0.130	0.046	0.925	0.980	0.994	0.931	0.959	0.934	0.962	0.909	0.941
β_3^0	-7.000	0.537	0.289	0.637	0.187	0.971	0.447	0.087	0.889	0.464	0.082	0.907	0.488	0.218	0.865	0.961	0.981	0.870	0.908	0.889	0.925	0.844	0.886
β_3^{c1}	-1.001	0.046	0.002	0.059	0.012	0.979	0.047	0.004	0.951	0.046	0.004	0.945	0.047	0.011	0.935	0.970	0.988	0.938	0.964	0.931	0.959	0.920	0.950
β_3^{c2}	3.999	0.127	0.016	0.157	0.046	0.963	0.106	0.023	0.888	0.110	0.021	0.916	0.117	0.059	0.859	0.951	0.975	0.868	0.908	0.899	0.933	0.837	0.881
β_3^{f11}	2.984	0.347	0.121	0.423	0.110	0.970	0.298	0.054	0.888	0.308	0.050	0.902	0.330	0.138	0.872	0.959	0.981	0.868	0.908	0.884	0.920	0.851	0.893
β_3^{f21}	-1.995	0.452	0.204	0.561	0.207	0.971	0.381	0.080	0.885	0.394	0.068	0.913	0.423	0.240	0.849	0.961	0.981	0.865	0.905	0.896	0.930	0.827	0.871
β_3^{f22}	-2.995	0.421	0.177	0.491	0.135	0.972	0.346	0.065	0.875	0.359	0.061	0.905	0.380	0.167	0.851	0.962	0.982	0.855	0.896	0.887	0.923	0.829	0.873
β_3^{tc}	0.496	0.206	0.042	0.254	0.054	0.978	0.189	0.030	0.926	0.202	0.030	0.938	0.190	0.059	0.893	0.969	0.987	0.910	0.942	0.923	0.953	0.874	0.912
d_3^{11}	1.068	0.327	0.125	0.439	0.157	0.917	0.264	0.082	0.806	0.295	0.083	0.823	0.269	0.164	0.762	0.900	0.934	0.781	0.831	0.799	0.847	0.736	0.788
d_3^{21}	0.787	0.287	0.083	0.382	0.137	0.962	0.233	0.067	0.867	0.260	0.069	0.902	0.241	0.131	0.816	0.950	0.974	0.846	0.888	0.884	0.920	0.792	0.840
d_3^{22}	1.364	0.361	0.131	0.475	0.184	0.947	0.296	0.089	0.853	0.337	0.093	0.900	0.292	0.182	0.790	0.933	0.961	0.831	0.875	0.881	0.919	0.765	0.815
σ_3^2	1.736	0.129	0.017	0.160	0.035	0.963	0.130	0.014	0.944	0.124	0.013	0.931	0.134	0.037	0.922	0.951	0.975	0.930	0.958	0.915	0.947	0.905	0.939
π_1	0.333	0.049	0.002	0.047	0.002	0.950	0.047	0.002	0.950				0.047	0.009	0.950	0.936	0.964	0.936	0.964			0.936	0.964
π_2	0.331	0.047	0.002	0.047	0.002	0.954	0.047	0.002	0.952				0.047	0.009	0.952	0.941	0.967	0.939	0.965			0.939	0.965
π_3	0.336	0.048	0.002	0.047	0.002	0.957	0.047	0.002	0.957				0.047	0.002	0.957	0.944	0.970	0.944	0.970			0.944	0.970

Table A.9 continued

parm	CI1L avg	CI1L std	CI1L n	CI2L avg	CI2L std	CI2L n	CI3L avg	CI3L std	CI3L n	CI4L avg	CI4L std	CI4L n
β_1^0	11.357	0.740	1000	11.276	0.744	1000	11.278	0.740	1000	11.480	5.555	1000
β_1^{c1}	2.831	0.068	1000	2.828	0.068	1000	2.829	0.068	1000	2.832	0.146	1000
β_1^{c2}	4.270	0.194	1000	4.256	0.194	1000	4.256	0.193	1000	4.297	1.188	1000
β_1^{f11}	1.874	0.448	1000	1.648	0.458	1000	1.671	0.432	1000	1.823	4.027	1000
β_1^{f21}	22.037	635.783	1000	1.485	0.439	1000	1.529	0.417	1000	1.685	3.787	1000
β_1^{f22}	2.683	0.584	1000	2.469	0.549	1000	2.487	0.546	1000	2.598	2.512	1000
β_1^{tc}	2.230	0.252	1000	2.192	0.250	1000	2.190	0.250	1000	2.217	0.293	1000
d_1^{11}	1.585	0.532	1000	1.431	0.393	1000	1.377	0.440	1000	1.665	1.079	1000
d_1^{21}	1.124	0.416	1000	0.981	0.301	1000	0.905	0.332	1000	1.256	2.059	1000
d_1^{22}	2.001	0.540	1000	1.848	0.422	1000	1.820	0.466	1000	2.014	0.646	1000
σ_1^2	2.715	0.204	1000	2.694	0.198	1000	2.696	0.200	1000	2.710	0.561	1000
β_2^0	17.758	0.639	1000	17.736	0.638	1000	17.724	0.638	1000	17.892	2.982	1000
β_2^{c1}	4.244	0.065	1000	4.243	0.065	1000	4.242	0.065	1000	4.243	0.065	1000
β_2^{c2}	1.741	0.145	1000	1.728	0.144	1000	1.720	0.144	1000	1.789	0.615	1000
β_2^{f11}	1.739	0.367	1000	1.655	0.363	1000	1.587	0.368	1000	1.900	0.611	1000
β_2^{f21}	4.433	0.564	1000	4.358	0.508	1000	4.318	0.507	1000	4.575	1.217	1000
β_2^{f22}	5.767	0.475	1000	5.736	0.476	1000	5.710	0.475	1000	5.848	0.528	1000
β_2^{tc}	3.598	0.257	1000	3.580	0.258	1000	3.576	0.258	1000	3.625	0.911	1000
d_2^{11}	1.511	0.474	1000	1.395	0.452	1000	1.316	0.405	1000	1.901	8.777	1000
d_2^{21}	1.310	0.415	1000	1.194	0.386	1000	1.136	0.363	1000	1.610	7.127	1000
d_2^{22}	1.847	0.520	1000	1.743	0.429	1000	1.681	0.454	1000	2.124	5.496	1000
σ_2^2	2.582	0.186	1000	2.563	0.182	1000	2.562	0.184	1000	2.565	0.186	1000
β_3^0	10.069	0.759	1000	9.980	0.752	1000	9.985	0.754	1000	10.010	0.759	1000
β_3^{c1}	1.426	0.065	1000	1.422	0.065	1000	1.422	0.065	1000	1.422	0.065	1000
β_3^{c2}	5.673	0.179	1000	5.663	0.179	1000	5.664	0.179	1000	5.667	0.180	1000
β_3^{f11}	4.391	0.480	1000	4.303	0.482	1000	4.308	0.481	1000	4.334	0.491	1000
β_3^{f21}	3.266	0.668	1000	3.028	0.602	1000	3.041	0.591	1000	3.107	0.729	1000
β_3^{f22}	4.467	0.572	1000	4.349	0.581	1000	4.357	0.575	1000	4.389	0.600	1000
β_3^{tc}	1.022	0.225	1000	0.899	0.224	1000	0.922	0.221	1000	0.906	0.245	1000
d_3^{11}	1.951	0.596	1000	1.694	0.461	1000	1.718	0.513	1000	1.737	0.485	1000
d_3^{21}	1.553	0.511	1000	1.305	0.389	1000	1.331	0.436	1000	1.346	0.413	1000
d_3^{22}	2.350	0.672	1000	2.109	0.514	1000	2.144	0.568	1000	2.146	0.534	1000
σ_3^2	2.497	0.186	1000	2.482	0.182	1000	2.480	0.184	1000	2.486	0.179	1000
π_1	0.489	0.068	1000	0.489	0.068	1000				0.490	0.070	1000
π_2	0.486	0.065	1000	0.486	0.065	1000				0.486	0.067	1000
π_3	0.493	0.067	1000	0.493	0.067	1000				0.493	0.067	1000

Table A.10: EM2nd variant simulation results for non-constant(nmax=6)

parm	est avg	est std	est mse	SE1 avg	SE1 std	CI1 CP	SE2 avg	SE2 std	CI2 CP	SE3 avg	SE3 std	CI3 CP	SE4 avg	SE4 std	CI4 CP	CI1BLCL	CI1BUCL	CI2BLCL	CI2BUCL	CI3BLCL	CI3BUCL	CI4BLCL	CI4BUCL
β_1^0	7.925	0.471	0.223	0.607	0.145	0.980	0.444	0.050	0.933	0.450	0.048	0.944	0.452	0.125	0.911	0.971	0.989	0.918	0.949	0.930	0.958	0.893	0.929
β_1^{c1}	-2.007	0.089	0.008	0.119	0.030	0.982	0.086	0.010	0.930	0.087	0.010	0.941	0.085	0.024	0.911	0.974	0.990	0.914	0.946	0.926	0.956	0.893	0.929
β_1^{c2}	3.008	0.094	0.009	0.124	0.032	0.984	0.088	0.011	0.929	0.090	0.010	0.928	0.087	0.025	0.893	0.976	0.992	0.913	0.945	0.912	0.944	0.874	0.912
β_1^{f11}	1.004	0.269	0.072	0.342	0.077	0.974	0.249	0.027	0.925	0.254	0.026	0.933	0.249	0.065	0.895	0.964	0.984	0.909	0.941	0.918	0.949	0.876	0.914
β_1^{f21}	0.605	0.320	0.103	0.403	0.092	0.972	0.290	0.032	0.917	0.298	0.031	0.922	0.286	0.073	0.894	0.962	0.982	0.900	0.934	0.905	0.939	0.875	0.913
β_1^{f22}	1.586	0.318	0.102	0.405	0.094	0.977	0.298	0.033	0.931	0.305	0.032	0.942	0.297	0.077	0.909	0.968	0.986	0.915	0.947	0.928	0.956	0.891	0.927
β_1^{tc}	1.493	0.212	0.045	0.255	0.055	0.972	0.204	0.030	0.933	0.201	0.030	0.929	0.221	0.076	0.920	0.962	0.982	0.918	0.949	0.913	0.945	0.903	0.937
d_1^{11}	0.982	0.371	0.138	0.507	0.187	0.954	0.357	0.101	0.915	0.350	0.095	0.909	0.412	0.287	0.888	0.941	0.967	0.898	0.932	0.891	0.927	0.868	0.908
d_1^{21}	0.494	0.265	0.070	0.367	0.124	0.972	0.269	0.068	0.925	0.252	0.059	0.926	0.324	0.243	0.901	0.962	0.982	0.909	0.941	0.910	0.942	0.882	0.920
d_1^{22}	1.164	0.351	0.124	0.477	0.168	0.935	0.342	0.095	0.901	0.331	0.090	0.888	0.402	0.337	0.881	0.920	0.950	0.882	0.920	0.868	0.908	0.861	0.901
σ_1^2	1.820	0.252	0.070	0.347	0.093	0.960	0.242	0.041	0.908	0.253	0.041	0.907	0.232	0.086	0.866	0.948	0.972	0.890	0.926	0.889	0.925	0.845	0.887
β_2^0	12.490	0.474	0.224	0.573	0.151	0.972	0.444	0.047	0.936	0.429	0.046	0.929	0.478	0.134	0.922	0.962	0.982	0.921	0.951	0.913	0.945	0.905	0.939
β_2^{c1}	-3.000	0.091	0.008	0.114	0.029	0.975	0.086	0.010	0.929	0.084	0.010	0.931	0.090	0.027	0.917	0.965	0.985	0.913	0.945	0.915	0.947	0.900	0.934
β_2^{c2}	1.196	0.093	0.009	0.117	0.030	0.982	0.088	0.010	0.935	0.086	0.010	0.938	0.093	0.028	0.922	0.974	0.990	0.920	0.950	0.923	0.953	0.905	0.939
β_2^{f11}	-0.999	0.266	0.071	0.328	0.079	0.975	0.250	0.026	0.919	0.243	0.025	0.914	0.263	0.071	0.912	0.965	0.985	0.902	0.936	0.897	0.931	0.894	0.930
β_2^{f21}	3.010	0.302	0.091	0.382	0.093	0.967	0.292	0.032	0.938	0.284	0.030	0.924	0.308	0.086	0.926	0.956	0.978	0.923	0.953	0.908	0.940	0.910	0.942
β_2^{f22}	4.001	0.319	0.102	0.391	0.094	0.978	0.300	0.031	0.920	0.292	0.030	0.917	0.313	0.084	0.912	0.969	0.987	0.903	0.937	0.900	0.934	0.894	0.930
β_2^{tc}	2.499	0.206	0.042	0.248	0.057	0.971	0.203	0.031	0.940	0.194	0.028	0.931	0.227	0.077	0.939	0.961	0.981	0.925	0.955	0.915	0.947	0.924	0.954
d_2^{11}	0.876	0.334	0.112	0.466	0.183	0.967	0.350	0.098	0.939	0.319	0.088	0.906	0.433	0.304	0.912	0.956	0.978	0.924	0.954	0.888	0.924	0.894	0.930
d_2^{21}	0.673	0.249	0.063	0.360	0.127	0.975	0.265	0.070	0.938	0.248	0.062	0.915	0.319	0.202	0.921	0.965	0.985	0.923	0.953	0.898	0.932	0.904	0.938
d_2^{22}	1.073	0.320	0.103	0.440	0.157	0.950	0.338	0.095	0.927	0.307	0.082	0.890	0.421	0.288	0.905	0.936	0.964	0.911	0.943	0.871	0.909	0.887	0.923
σ_2^2	1.704	0.242	0.068	0.330	0.090	0.957	0.245	0.040	0.920	0.238	0.038	0.896	0.254	0.101	0.887	0.944	0.970	0.903	0.937	0.877	0.915	0.867	0.907
β_3^0	-7.002	0.486	0.236	0.584	0.145	0.979	0.441	0.047	0.926	0.440	0.045	0.923	0.461	0.125	0.912	0.970	0.988	0.910	0.942	0.906	0.940	0.894	0.930
β_3^{c1}	-1.000	0.086	0.007	0.113	0.027	0.972	0.085	0.010	0.944	0.084	0.009	0.938	0.089	0.026	0.917	0.962	0.982	0.930	0.958	0.923	0.953	0.900	0.934
β_3^{c2}	4.000	0.093	0.009	0.116	0.028	0.975	0.087	0.010	0.925	0.086	0.010	0.930	0.091	0.027	0.910	0.965	0.985	0.909	0.941	0.914	0.946	0.892	0.928
β_3^{f11}	3.007	0.264	0.070	0.328	0.075	0.978	0.249	0.027	0.929	0.244	0.025	0.931	0.259	0.072	0.909	0.969	0.987	0.913	0.945	0.915	0.947	0.891	0.927
β_3^{f21}	-1.989	0.310	0.096	0.382	0.090	0.977	0.291	0.032	0.923	0.285	0.030	0.913	0.303	0.084	0.898	0.968	0.986	0.906	0.940	0.896	0.930	0.879	0.917
β_3^{f22}	-3.005	0.314	0.099	0.391	0.097	0.972	0.298	0.033	0.925	0.292	0.031	0.923	0.311	0.089	0.911	0.962	0.982	0.909	0.941	0.906	0.940	0.893	0.929
β_3^{tc}	0.497	0.221	0.049	0.272	0.058	0.967	0.204	0.031	0.923	0.214	0.031	0.940	0.206	0.062	0.894	0.956	0.978	0.906	0.940	0.925	0.955	0.875	0.913
d_3^{11}	1.137	0.398	0.162	0.547	0.191	0.951	0.360	0.100	0.891	0.380	0.102	0.900	0.373	0.226	0.849	0.938	0.964	0.872	0.910	0.881	0.919	0.827	0.871
d_3^{21}	0.766	0.306	0.095	0.425	0.147	0.960	0.271	0.070	0.881	0.297	0.074	0.914	0.270	0.137	0.833	0.948	0.972	0.861	0.901	0.897	0.931	0.810	0.856
d_3^{22}	1.356	0.389	0.153	0.534	0.189	0.952	0.341	0.098	0.871	0.376	0.101	0.896	0.339	0.201	0.807	0.939	0.965	0.850	0.892	0.877	0.915	0.783	0.831
σ_3^2	1.667	0.239	0.064	0.320	0.085	0.950	0.242	0.041	0.918	0.233	0.038	0.892	0.253	0.102	0.905	0.936	0.964	0.901	0.935	0.873	0.911	0.887	0.923
π_1	0.334	0.047	0.002	0.047	0.002	0.952	0.047	0.002	0.952				0.047	0.002	0.952	0.939	0.965	0.939	0.965			0.939	0.965
π_2	0.331	0.049	0.002	0.047	0.002	0.939	0.047	0.002	0.939				0.047	0.002	0.939	0.924	0.954	0.924	0.954			0.924	0.954
π_3	0.335	0.045	0.002	0.047	0.002	0.962	0.047	0.002	0.962				0.047	0.002	0.962	0.950	0.974	0.950	0.974			0.950	0.974

Table A.10 continued

parm	CI1L avg	CI1L std	CI1L n	CI2L avg	CI2L std	CI2L n	CI3L avg	CI3L std	CI3L n	CI4L avg	CI4L std	CI4L n
β_1^0	11.340	0.662	1000	11.276	0.663	1000	11.278	0.663	1000	11.283	0.665	1000
β_1^{c1}	2.858	0.126	1000	2.848	0.126	1000	2.849	0.126	1000	2.849	0.126	1000
β_1^{c2}	4.269	0.132	1000	4.261	0.132	1000	4.262	0.132	1000	4.262	0.132	1000
β_1^{f11}	1.731	0.329	1000	1.591	0.336	1000	1.598	0.333	1000	1.596	0.348	1000
β_1^{f21}	1.463	0.336	1000	1.224	0.308	1000	1.238	0.306	1000	1.222	0.338	1000
β_1^{f22}	2.527	0.419	1000	2.397	0.419	1000	2.404	0.419	1000	2.403	0.425	1000
β_1^{tc}	2.234	0.285	1000	2.189	0.290	1000	2.187	0.290	1000	2.209	0.297	1000
d_1^{11}	1.992	0.690	1000	1.732	0.510	1000	1.698	0.572	1000	1.896	0.738	1000
d_1^{21}	1.264	0.428	1000	1.059	0.314	1000	1.013	0.345	1000	1.220	0.634	1000
d_1^{22}	2.126	0.635	1000	1.922	0.480	1000	1.885	0.550	1000	2.097	0.821	1000
σ_1^2	2.755	0.392	1000	2.663	0.356	1000	2.668	0.371	1000	2.665	0.351	1000
β_2^0	17.740	0.667	1000	17.707	0.668	1000	17.704	0.668	1000	17.717	0.668	1000
β_2^{c1}	4.255	0.128	1000	4.249	0.128	1000	4.249	0.128	1000	4.251	0.128	1000
β_2^{c2}	1.725	0.130	1000	1.710	0.130	1000	1.709	0.130	1000	1.713	0.130	1000
β_2^{f11}	1.701	0.339	1000	1.584	0.335	1000	1.575	0.336	1000	1.608	0.347	1000
β_2^{f21}	4.394	0.421	1000	4.335	0.419	1000	4.330	0.421	1000	4.348	0.420	1000
β_2^{f22}	5.767	0.447	1000	5.720	0.447	1000	5.717	0.447	1000	5.729	0.449	1000
β_2^{tc}	3.604	0.285	1000	3.580	0.286	1000	3.576	0.287	1000	3.596	0.288	1000
d_2^{11}	1.807	0.646	1000	1.600	0.461	1000	1.526	0.519	1000	1.819	0.774	1000
d_2^{21}	1.396	0.449	1000	1.223	0.335	1000	1.180	0.371	1000	1.359	0.525	1000
d_2^{22}	1.961	0.582	1000	1.804	0.446	1000	1.741	0.501	1000	2.014	0.670	1000
σ_2^2	2.585	0.375	1000	2.507	0.340	1000	2.499	0.355	1000	2.527	0.333	1000
β_3^0	10.042	0.681	1000	9.978	0.682	1000	9.978	0.682	1000	9.990	0.684	1000
β_3^{c1}	1.451	0.120	1000	1.434	0.121	1000	1.434	0.120	1000	1.437	0.122	1000
β_3^{c2}	5.666	0.132	1000	5.662	0.132	1000	5.662	0.132	1000	5.663	0.132	1000
β_3^{f11}	4.354	0.367	1000	4.309	0.369	1000	4.307	0.369	1000	4.318	0.371	1000
β_3^{f21}	3.019	0.417	1000	2.930	0.419	1000	2.925	0.420	1000	2.947	0.423	1000
β_3^{f22}	4.395	0.433	1000	4.331	0.436	1000	4.328	0.437	1000	4.344	0.440	1000
β_3^{tc}	1.062	0.240	1000	0.928	0.238	1000	0.947	0.233	1000	0.937	0.261	1000
d_3^{11}	2.229	0.715	1000	1.916	0.547	1000	1.926	0.617	1000	1.992	0.629	1000
d_3^{21}	1.621	0.538	1000	1.340	0.406	1000	1.369	0.451	1000	1.369	0.438	1000
d_3^{22}	2.438	0.708	1000	2.155	0.550	1000	2.184	0.612	1000	2.198	0.586	1000
σ_3^2	2.526	0.369	1000	2.455	0.335	1000	2.445	0.351	1000	2.478	0.328	1000
π_1	0.490	0.065	1000	0.490	0.065	1000				0.490	0.065	1000
π_2	0.486	0.068	1000	0.486	0.068	1000				0.486	0.068	1000
π_3	0.491	0.063	1000	0.491	0.063	1000				0.491	0.063	1000

Appendix B

Vector and matrix differential calculus

The purpose of this section is to introduce the ideas of Magnus and Neudecker (1999) which are concerned with calculating the matrix and vector equivalents of derivatives of scalar functions. The method is based upon the differential operator $\boldsymbol{d}(\cdot)$, so $\boldsymbol{d}\,(f(\boldsymbol{x}))$ is the differential of $f(\boldsymbol{x})$. The term $\boldsymbol{dx}$ is actually $\boldsymbol{d}\,(\boldsymbol{x})$, the differential of $\boldsymbol{x}$, but for functions of $\boldsymbol{x}$, $\boldsymbol{dx}$ has special meaning in the sense that expressions involving this term in a certain way can be used to identify the score vector and Hessian matrix of f as a function of $\boldsymbol{x}$. These identification methods can be summarised for different functions, and appear in tables (B.1) and (B.2), which are taken from the first and second identification tables on page 198 and 215 respectively of Magnus and Neudecker (1999). The term $\boldsymbol{dx}$ is in fact a notational convenience to make the multivariate differential theory appear, at least notationally, the same as its scalar counterpart. In reality $\boldsymbol{dx}$ is an increment vector, usually with small components, and alongside Taylor series expansions, is prominent in the definition of what it means for vector functions to be differentiable.

In order to use tables (B.1) and (B.2), we write $\boldsymbol{A}$, $\boldsymbol{a}$ or α to denote a constant matrix, vector or scalar respectively, and $\boldsymbol{X}$, $\boldsymbol{x}$ or δ depending on whether the argument to the function in question is a matrix, vector or scalar respectively. For dimensions we will use $\boldsymbol{f} : S \longrightarrow \mathbb{R}^{m}$, $S \subseteq \mathbb{R}^{n}$ for a vector function of a vector argument, $f : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^{n}$ for a scalar function of a vector argument, $\boldsymbol{f} : S \longrightarrow \mathbb{R}^{m}$, $S \subseteq \mathbb{R}^{n\times q}$ for a vector function of a matrix argument, and $\boldsymbol{F} : S \longrightarrow \mathbb{R}^{m\times p}$, $S \subseteq \mathbb{R}^{n}$, for a matrix function of a vector argument. These tables will be used extensively in chapter C to derive score vectors and Hessian matrices of log-likelihood functions. Furthermore since $\boldsymbol{\theta}$ for LMMs and Θ for MLMMs are comprised of vector components, for simplicity we apply the theory of Magnus and Neudecker componentwise: that is in turn we assume the log-likelihood to be a function of only one component, with the other components considered fixed. Differentials of any expressions involving just the fixed components are zero, which leads to much shorter expressions. Thus the components of the score vectors and Hessian matrices are identified with the tables. At the end of this section we present some useful rules for computing differentials.

Table B.1: First identification table

Function	Differential	Derivative	Order of D
$f(\delta)$	$\boldsymbol{d}(f(\delta)) = \alpha \boldsymbol{d}\delta$	$\boldsymbol{D}_\delta(f(\delta)) = \alpha$	1×1
$f(\boldsymbol{x})$	$\boldsymbol{d}(f(\boldsymbol{x})) = \boldsymbol{a}^\top \boldsymbol{d}\boldsymbol{x}$	$\boldsymbol{D}_{\boldsymbol{x}}(f(\boldsymbol{x})) = \boldsymbol{a}^\top$	$1 \times n$
$f(\boldsymbol{X})$	$\boldsymbol{d}(f(\boldsymbol{X})) = \boldsymbol{a}^\top \boldsymbol{d}(\mathrm{vec}(\boldsymbol{X}))$ $= \{\mathrm{vec}(\boldsymbol{a})\}^\top \mathrm{vec}(\boldsymbol{d}(\boldsymbol{X})$ $= \mathrm{tr}(\boldsymbol{a}^\top \boldsymbol{d}(\boldsymbol{X}))$	$\boldsymbol{D}_{\mathrm{vec}(\boldsymbol{X})}(f(\boldsymbol{X})) = \boldsymbol{a}^\top$	$1 \times nq$
$\boldsymbol{f}(\delta)$	$\boldsymbol{d}(\boldsymbol{f}(\delta)) = \boldsymbol{a}\boldsymbol{d}\delta$	$\boldsymbol{D}_\delta(\boldsymbol{f}(\delta)) = \boldsymbol{a}$	$m \times 1$
$\boldsymbol{f}(\boldsymbol{x})$	$\boldsymbol{d}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{A}\boldsymbol{d}\boldsymbol{x}$	$\boldsymbol{D}_{\boldsymbol{x}}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{A}$	$m \times n$
$\boldsymbol{f}(\boldsymbol{X})$	$\boldsymbol{d}(\boldsymbol{f}(\boldsymbol{X})) = \boldsymbol{A}\boldsymbol{d}(\mathrm{vec}(\boldsymbol{X}))$	$\boldsymbol{D}_{\mathrm{vec}(\boldsymbol{X})}(\boldsymbol{f}(\boldsymbol{X})) = \boldsymbol{A}$	$m \times nq$

Table B.2: Second identification table

Function	Differential	Derivative	Order of H
$f(\delta)$	$\boldsymbol{d}^2(f(\delta)) = \beta \boldsymbol{d}(\delta)^2$	$\boldsymbol{H}_\delta\left(f(\delta)\right) = \beta$	1×1
$f(\boldsymbol{x})$	$\boldsymbol{d}^2(f(\boldsymbol{x})) = \left\{\boldsymbol{d}\left(\boldsymbol{x}\right)\right\}^{\intercal} \boldsymbol{B} \left\{\boldsymbol{d}\left(\boldsymbol{x}\right)\right\}$	$\boldsymbol{H}_{\boldsymbol{x}}\left(f(\boldsymbol{x})\right) = \frac{1}{2}\left(\boldsymbol{B} + \boldsymbol{B}^{\intercal}\right)$	$n \times n$
$f(\boldsymbol{X})$	$\boldsymbol{d}^2(f(\boldsymbol{X})) = \left\{\boldsymbol{d}\left(\mathrm{vec}(\boldsymbol{X})\right)\right\}^{\intercal} \boldsymbol{B} \left\{\boldsymbol{d}\left(\mathrm{vec}(\boldsymbol{X})\right)\right\}$	$\boldsymbol{H}_{\mathrm{vec}(\boldsymbol{X})}\left(f(\boldsymbol{X})\right) = \frac{1}{2}\left(\boldsymbol{B} + \boldsymbol{B}^{\intercal}\right)$	$nq \times nq$
$\boldsymbol{F}(\boldsymbol{x})$	$\mathrm{vec}[\boldsymbol{d}^2\left(\boldsymbol{F}(\boldsymbol{x})\right)] = (\boldsymbol{I}_{mp} \otimes \boldsymbol{dx})^{\intercal} \boldsymbol{B} \boldsymbol{dx}$	$\boldsymbol{H}_{\boldsymbol{x}}\left(\mathrm{vec}(F(\boldsymbol{x}))\right) = \frac{1}{2}\left(\boldsymbol{B} + (\boldsymbol{B}^{\intercal})_v\right)$	$mnp \times n$

where $(\boldsymbol{B}^{\intercal})_v = (\boldsymbol{B}_{1,1}, ..., \boldsymbol{B}_{m,1}, ..., \boldsymbol{B}_{1,p}, ..., \boldsymbol{B}_{m,p})^{\intercal}$, and $\boldsymbol{B}_{j,k}$ is a $n \times n$ matrix for all $j = 1, ..., m$ and $k = 1, ..., p$.

Here we present without proof some important rules which permit easy manipulation of differentials. We shall use these repeatedly without reference to them, and unless otherwise stated they are taken from Magnus and Neudecker (1999, Ch. 8). For differentiable functions $\boldsymbol{f} : S \longrightarrow \mathbb{R}$ and $\boldsymbol{g} : S \longrightarrow \mathbb{R}$ where $S \subseteq \mathbb{R}^n$, and for $\alpha \in \mathbb{R}$, then the following rules hold

$$\boldsymbol{d}(\alpha) = 0, \tag{B.1}$$

$$\boldsymbol{d}(\alpha \boldsymbol{f}(\boldsymbol{x})) = \alpha \boldsymbol{d}(\boldsymbol{f}(\boldsymbol{x})), \tag{B.2}$$

$$\boldsymbol{d}((\boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})) = \boldsymbol{d}(\boldsymbol{f}(\boldsymbol{x})) + \boldsymbol{d}(\boldsymbol{g}(\boldsymbol{x})), \tag{B.3}$$

$$\boldsymbol{d}((\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x})) = \boldsymbol{d}(\boldsymbol{f}(\boldsymbol{x})) - \boldsymbol{d}(\boldsymbol{g}(\boldsymbol{x})), \tag{B.4}$$

$$\boldsymbol{d}(\boldsymbol{f}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})) = \boldsymbol{d}(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{f}(\boldsymbol{x})\boldsymbol{d}(\boldsymbol{g}(\boldsymbol{x})), \tag{B.5}$$

$$\boldsymbol{d}(\boldsymbol{f}(\boldsymbol{x})^{\alpha}) = \alpha \boldsymbol{f}(\boldsymbol{x})^{\alpha-1}\boldsymbol{d}(\boldsymbol{f}(\boldsymbol{x})), \tag{B.6}$$

$$\boldsymbol{d}(log(\boldsymbol{f}(\boldsymbol{x}))) = \boldsymbol{f}(\boldsymbol{x})^{-1}\boldsymbol{d}(\boldsymbol{f}(\boldsymbol{x})), \tag{B.7}$$

$$\boldsymbol{d}\left(e^{\boldsymbol{f}(\boldsymbol{x})}\right) = e^{\boldsymbol{f}(\boldsymbol{x})}\boldsymbol{d}(\boldsymbol{f}(\boldsymbol{x})) \tag{B.8}$$

The above rules also apply if $\boldsymbol{x}$ is a scalar and give the familiar results of single variable differential calculus. There are also many useful rules for differentials of matrix functions of matrix arguments, a few of which are analogues of some of the rules above. For differentiable functions $\boldsymbol{F} : S \longrightarrow \mathbb{R}^{m \times p}$ and $\boldsymbol{G} : S \longrightarrow \mathbb{R}^{m \times p}$ where $S \subseteq \mathbb{R}^{n \times q}$, and for a matrix $\boldsymbol{A}$ of real constants, then the following rules hold

$$\boldsymbol{d}(\boldsymbol{A}) = 0, \tag{B.9}$$

$$\boldsymbol{d}(\boldsymbol{A}\boldsymbol{F}(\boldsymbol{X})) = \boldsymbol{A}\boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X})), \tag{B.10}$$

$$\boldsymbol{d}((\boldsymbol{F}(\boldsymbol{X}) + \boldsymbol{G}(\boldsymbol{X})) = \boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X})) + \boldsymbol{d}(\boldsymbol{G}(\boldsymbol{X})), \tag{B.11}$$

$$\boldsymbol{d}((\boldsymbol{F}(\boldsymbol{X}) - \boldsymbol{G}(\boldsymbol{X})) = \boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X})) - \boldsymbol{d}(\boldsymbol{G}(\boldsymbol{X})), \tag{B.12}$$

$$\boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X})\boldsymbol{G}(\boldsymbol{X})) = \boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X}))\boldsymbol{G}(\boldsymbol{X}) + \boldsymbol{F}(\boldsymbol{X})\boldsymbol{d}(\boldsymbol{G}(\boldsymbol{X})), \tag{B.13}$$

$$\boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X}) \otimes \boldsymbol{G}(\boldsymbol{X})) = \boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X})) \otimes \boldsymbol{G}(\boldsymbol{X}) + \boldsymbol{F}(\boldsymbol{X}) \otimes \boldsymbol{d}(\boldsymbol{G}(\boldsymbol{X})), \tag{B.14}$$

$$\boldsymbol{d}\left(\boldsymbol{F}(\boldsymbol{X})^{\intercal}\right) = (\boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X})))^{\intercal}, \tag{B.15}$$

$$\boldsymbol{d}(vec(\boldsymbol{F}(\boldsymbol{X}))) = vec(\boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X}))), \tag{B.16}$$

$$\boldsymbol{d}(tr(\boldsymbol{F}(\boldsymbol{X}))) = tr(\boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X}))). \tag{B.17}$$

The following three results are concerned with functions, ϕ say, of a matrix function $\boldsymbol{F}$. In all cases the subset $S \subseteq \mathbb{R}^{n \times q}$ is open, ϕ is differentiable, and the matrix function $\boldsymbol{F}$ is k times either continuously differentiable or differentiable ($k \geq 1$). Continuous differentiability means that each partial derivative of $\boldsymbol{F}$ exists and is a continuous function, and in turn this implies differentiability. For details see page 103 and Theorem 7 (page 101) in (Magnus and Neudecker (1999, Ch. 5)).

Let $\boldsymbol{F} : S \longrightarrow \mathbb{R}^{m \times m} (m \geq 2)$ be a matrix function and $|\boldsymbol{F}| : S \longrightarrow \mathbb{R}$ be a scalar function where $|\boldsymbol{F}|(\boldsymbol{X}) = |\boldsymbol{F}(\boldsymbol{X})|$, then if $\boldsymbol{F}(\boldsymbol{X})$ is non-singular

$$\boldsymbol{d}(|\boldsymbol{F}(\boldsymbol{X})|) = |\boldsymbol{F}(\boldsymbol{X})| \operatorname{tr}\left(\boldsymbol{F}(\boldsymbol{X})^{-1}\boldsymbol{dF}(\boldsymbol{X})\right). \tag{B.18}$$

Let $\boldsymbol{F} : S \longrightarrow T_+$ be a matrix function where $T_+ = \left\{\boldsymbol{Y} : \boldsymbol{Y} \in \mathbb{R}^{m \times m}, |\boldsymbol{Y}| > 0\right\}$ and $log(|\boldsymbol{F}|) : S \longrightarrow \mathbb{R}$ be a scalar function where $(\log(|\boldsymbol{F}|)(\boldsymbol{X})) = \log(|\boldsymbol{F}(\boldsymbol{X})|)$ then

$$\boldsymbol{d}(\log(|\boldsymbol{F}(\boldsymbol{X})|)) = \operatorname{tr}\left(\boldsymbol{F}(\boldsymbol{X})^{-1}\boldsymbol{dF}(\boldsymbol{X})\right). \tag{B.19}$$

Let $\boldsymbol{F} : S \longrightarrow T$ be a matrix function where T is the set of non-singular real $m \times m$ matrices and $\boldsymbol{F}^{-1} : S \longrightarrow T$ be a matrix function where $\boldsymbol{F}^{-1}(\boldsymbol{X}) = (\boldsymbol{F}(\boldsymbol{X}))^{-1}$ then

$$\boldsymbol{d}\left(\boldsymbol{F}(\boldsymbol{X})^{-1}\right) = -\boldsymbol{F}(\boldsymbol{X})^{-1}(\boldsymbol{d}(\boldsymbol{F}(\boldsymbol{X})))\boldsymbol{F}(\boldsymbol{X})^{-1}. \tag{B.20}$$

Appendix C

Derivatives of log-likelihood functions

In this chapter we will firstly derive the information matrix for the LMM in section C.1, and then building on these results we will derive the score vector and Hessian matrix for MLMMs in section C.2.2. Both of these sections necessarily involve taking derivatives of the log-likelihood functions associated with a LMM and a MLMM respectively, where these are scalar functions with vector arguments. The score vector of such a function is then a vector function of a vector argument, and we shall need to derive the vector derivative of this function to obtain the Hessian matrices. The approach we take to deriving these derivatives is to work with the concept of the derivatives of vector valued functions of vector arguments, and sometimes the derivatives of matrix valued functions of vector arguments. This approach avoids the necessity of taking the "ordinary" derivatives of each element of the vector and matrix functions separately (these are derivatives of scalar functions of scalar arguments) but instead provides methods to identify the whole vector of derivatives simultaneously, and in doing so we believe this approach is more elegant than the ordinary approach since there are less steps to take in the derivations. Notwithstanding this, even if the reader accepts this elegance argument it is evident from this section that the derivations when given in full are still very lengthy, and so it is clear this approach offers no economy of effort for either the reader or writer.

The definition of derivatives of vector and matrix functions rely instrumentally on the definition of ordinary derivatives, but with some extra features. Accordingly it comes as no surprise that the familiar raft of mathematical conditions that can be satisfied in order for the derivatives to exist also apply in some way to the existence or not of these vector and matrix derivatives. We adopt our methods from the book length exposition on this topic by Magnus and Neudecker (1999), but for brevity we omit discussions of these mathematical considerations and instead assume all derivatives exist everywhere in the parameter space. In Appendix B we present a very brief summary based on the more practical sections from Magnus and Neudecker (1999).

C.1 Information matrix for weighted LMMs

This section is concerned with deriving the information matrix $I_N(\boldsymbol{\theta})$ for a LMM using a weighted log-likelihood function given by

$$L(\boldsymbol{\theta}|\boldsymbol{y}_i) = -\frac{1}{2}\sum_{i=1}^{N} w_i n_i \log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} w_i \log|\boldsymbol{V}_i(\boldsymbol{\zeta})| - \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i, \qquad \text{(C.1)}$$

where $\tilde{\boldsymbol{e}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}$. We shall compute minus the Hessian matrix for this model by finding the second order differentials of (C.1) by considering it to be a function in turn of only one component of $\boldsymbol{\theta}$. The negative of the Hessian in partitioned form is given by

$$-\boldsymbol{H}_{\boldsymbol{\theta}}\left(L(\boldsymbol{\theta})\right) = -\begin{bmatrix} \boldsymbol{H}_{\boldsymbol{\beta}}\left(L(\boldsymbol{\beta})\right) & \boldsymbol{D}^2_{(\sigma^2)(\boldsymbol{\beta})}\left(L(\boldsymbol{\beta})\right) & \boldsymbol{D}^2_{(\boldsymbol{\psi})(\boldsymbol{\beta})}\left(L(\boldsymbol{\beta})\right) & \boldsymbol{D}^2_{(\boldsymbol{\phi})(\boldsymbol{\beta})}\left(L(\boldsymbol{\beta})\right) \\ & \boldsymbol{H}_{\sigma^2}\left(L(\sigma^2)\right) & \boldsymbol{D}^2_{(\boldsymbol{\psi})(\sigma^2)}\left(L(\sigma^2)\right) & \boldsymbol{D}^2_{(\boldsymbol{\phi})(\sigma^2)}\left(L(\sigma^2)\right) \\ & & \boldsymbol{H}_{\boldsymbol{\psi}}\left(L(\boldsymbol{\psi})\right) & \boldsymbol{D}^2_{(\boldsymbol{\phi})(\boldsymbol{\psi})}\left(L(\boldsymbol{\psi})\right) \\ & \text{symm} & & \boldsymbol{H}_{\boldsymbol{\phi}}\left(L(\boldsymbol{\phi})\right) \end{bmatrix}, \tag{C.2}$$

and we shall compute each line of this matrix in turn, taking expectations of each component matrix in each line in order to obtain $I_N(\boldsymbol{\theta})$. For LMMs with AR errors we will need to take derivatives of the ACF in order to obtain the matrices in the right hand column of C.2, and we derive these in the next subsection.

C.1.1 Derivatives of the autocorrelation function

This section is concerned with deriving closed form equations for the elements of $\boldsymbol{D}_{\boldsymbol{\phi}}\left(\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) = \left\{{}_r\,\boldsymbol{D}_{\boldsymbol{\phi}}[(\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk}]\right\}_{j=1,\,k=1}^{n \quad n}$ which is the $n^2 \times r$ derivative matrix of the vector $\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))$, and $i \in I_N$. From (A.6) we see this involves calculating the derivatives of $(\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk} = \rho_{|j-k|}(\boldsymbol{\phi})$, for $j, k = 1, ..., n$ where

$$\rho_s(\boldsymbol{\phi}) = \sum_{v=1}^{r} \phi_v \rho_{s-v}, \qquad s = r+1, ..., n, \tag{C.3}$$

and $\rho_0(\boldsymbol{\phi}) = 1$, and $\rho_1(\boldsymbol{\phi}), ..., \rho_r(\boldsymbol{\phi})$ are given in subsection (**??**) for $r = 1, 2, 3$. Thus we need to calculate

$$\begin{aligned}\boldsymbol{D}_{\boldsymbol{\phi}}\left(\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) &= \Big\{_c\, \boldsymbol{D}_{\boldsymbol{\phi}}[(\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk}]\Big\}_{j=1,\, k=1}^{n \quad n} \\ &= \Big\{_c\, \boldsymbol{D}_{\boldsymbol{\phi}}[\rho_{|j-k|}(\boldsymbol{\phi})]\Big\}_{j=1,\, k=1}^{n \quad n} \\ &= \Big\{_c\, \boldsymbol{D}_{\boldsymbol{\phi}}\left[\sum\nolimits_{v=1}^{r} \phi_v \rho_{|j-k|-v}(\boldsymbol{\phi})\right]\Big\}_{j=1,\, k=1}^{n \quad n}, \end{aligned} \tag{C.4}$$

where $\rho_{|j-k|-v}(\boldsymbol{\phi})$ is given by (C.3) when $|j-k| - v \geq r + 1$, and again $\rho_1(\boldsymbol{\phi}), ..., \rho_r(\boldsymbol{\phi})$ are given in subsection (**??**) for $r = 1, 2, 3$. So for any $j, k = 1, ..., n$ we have

$$\begin{aligned}\boldsymbol{D}_{\boldsymbol{\phi}}[\rho_{|j-k|}(\boldsymbol{\phi})] &= \left(\frac{\partial}{\partial \phi_1}\left[\rho_{|j-k|}(\boldsymbol{\phi})\right], ..., \frac{\partial}{\partial \phi_r}\left[\rho_{|j-k|}(\boldsymbol{\phi})\right]\right) \\ &= \left(\frac{\partial}{\partial \phi_1}\left[\sum\nolimits_{v=1}^{r} \phi_v \rho_{|j-k|-v}(\boldsymbol{\phi})\right], ..., \frac{\partial}{\partial \phi_r}\left[\sum\nolimits_{v=1}^{r} \phi_v \rho_{|j-k|-v}(\boldsymbol{\phi})\right]\right).\end{aligned} \tag{C.5}$$

We also want to derive closed form equations for the elements of the $n^2 r \times r$ Hessian

$$\boldsymbol{H}_{\boldsymbol{\phi}}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right) = \Big\{_c\, \boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i)_{jk})\Big\}_{j=1,\, k=1}^{n \quad n} \tag{C.6}$$

where for any $j, k = 1, ..., n$ we have

$$\boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i)_{jk}) =$$

$$\begin{bmatrix} \frac{\partial^2}{\partial \phi_1^2}[\rho_{|j-k|}(\boldsymbol{\phi})] & \frac{\partial^2}{\partial \phi_2 \partial \phi_1}[\rho_{|j-k|}(\boldsymbol{\phi})] & \cdots & \frac{\partial^2}{\partial \phi_r \partial \phi_1}[\rho_{|j-k|}(\boldsymbol{\phi})] \\ \vdots & & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2}{\partial \phi_1 \partial \phi_r}[\rho_{|j-k|}(\boldsymbol{\phi})] & \frac{\partial^2}{\partial \phi_2 \partial \phi_r}[\rho_{|j-k|}(\boldsymbol{\phi})] & \cdots & \frac{\partial^2}{\partial \phi_r^2}[\rho_{|j-k|}(\boldsymbol{\phi})] \end{bmatrix} =$$

$$\begin{bmatrix} \frac{\partial^2}{\partial \phi_1^2}\left[\sum_{v=1}^r \phi_v \rho_{|j-k|-v}(\boldsymbol{\phi})\right] & \frac{\partial^2}{\partial \phi_2 \partial \phi_1}\left[\sum_{v=1}^r \phi_v \rho_{|j-k|-v}(\boldsymbol{\phi})\right] & \cdots & \frac{\partial^2}{\partial \phi_r \partial \phi_1}\left[\sum_{v=1}^r \phi_v \rho_{|j-k|-v}(\boldsymbol{\phi})\right] \\ \vdots & & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2}{\partial \phi_1 \partial \phi_r}\left[\sum_{v=1}^r \phi_v \rho_{|j-k|-v}(\boldsymbol{\phi})\right] & \frac{\partial^2}{\partial \phi_r \partial \phi_1}\left[\sum_{v=1}^r \phi_v \rho_{|j-k|-v}(\boldsymbol{\phi})\right] & \cdots & \frac{\partial^2}{\partial \phi_r^2}\left[\sum_{v=1}^r \phi_v \rho_{|j-k|-v}(\boldsymbol{\phi})\right] \end{bmatrix}, \tag{C.7}$$

using (C.5). Now for $l = 1, ..., r$ we have

$$\begin{aligned} \frac{\partial}{\partial \phi_l}\left[\rho_{|j-k|}(\boldsymbol{\phi})\right] &= \frac{\partial}{\partial \phi_l}\left[\sum\nolimits_{v=1}^r \phi_v \rho_{|j-k|-v}(\boldsymbol{\phi})\right] \\ &= \rho_{|j-k|-l}(\boldsymbol{\phi}) + \sum_{v=1}^r \phi_v \left(\frac{\partial}{\partial \phi_l}\left[\rho_{|j-k|-v}(\boldsymbol{\phi})\right]\right), \end{aligned} \tag{C.8}$$

and for $l, m = 1, ..., r$ we have

$$\begin{aligned} \frac{\partial^2}{\partial \phi_l \partial \phi_m}\left[\rho_{|j-k|}(\boldsymbol{\phi})\right] &= \frac{\partial}{\partial \phi_m}\left[\rho_{|j-k|-l}(\boldsymbol{\phi}) + \sum_{v=1}^r \phi_v \left(\frac{\partial}{\partial \phi_l}\left[\rho_{|j-k|-v}(\boldsymbol{\phi})\right]\right)\right] \\ &= \rho_{|j-k|-l}(\boldsymbol{\phi}) + \sum_{v=1}^r \phi_v \left(\frac{\partial}{\partial \phi_l}\left[\rho_{|j-k|-m-v}(\boldsymbol{\phi})\right]\right) + \rho_{|j-k|-m}(\boldsymbol{\phi}) \\ &+ \sum_{v=1}^r \phi_v \left(\frac{\partial}{\partial \phi_m}\left[\rho_{|j-k|-l-v}(\boldsymbol{\phi})\right]\right) + \sum_{v=1}^r \phi_v \left(\frac{\partial^2}{\partial \phi_l \partial \phi_m}\left[\rho_{|j-k|-v}(\boldsymbol{\phi})\right]\right), \end{aligned} \tag{C.9}$$

where for $x \in \mathbb{Z}$ we replace ρ_x with $\rho_{|x|}$ in the equations (C.8) and (C.9) if $x < 0$.

Thus equations (C.8) and (C.9) allow us to calculate the elements of (C.5) and (C.7) respectively.

Since for all r we have $\rho_0(\boldsymbol{\phi}) = 1$ then $\frac{\partial}{\partial \phi_1}\left[\rho_0(\boldsymbol{\phi})\right] = 0$, and $\frac{\partial^2}{\partial \phi_l \partial \phi_m}\left[\rho_0(\boldsymbol{\phi})\right] = 0$ for all $l = 1, ..., r$ and $l, m = 1, ..., r$. However for $s = 1, ..., r$, and as with the autocorrelation function itself (C.3), for all $l = 1, ..., r$ and $l, m = 1, ..., r$ we need to first calculate $\frac{\partial}{\partial \phi_l}\left[\rho_s(\boldsymbol{\phi})\right]$ and $\frac{\partial^2}{\partial \phi_l \partial \phi_m}\left[\rho_s(\boldsymbol{\phi})\right]$ for $s = 1, ..., r$ before being able to use (C.8) and (C.9) recursively to calculate the partial derivatives for $s = r + 1, ..., n$. We now give these partial derivatives for $r = 1, 2, 3$. For an AR(1) process we have

$$\frac{\partial \rho_1(\boldsymbol{\phi})}{\partial \phi_1} = 1, \tag{C.10}$$

and so

$$\frac{\partial^2 \rho_1(\boldsymbol{\phi})}{\partial \phi_1^2} = 0. \tag{C.11}$$

For an AR(2) process we have

$$\begin{aligned}
\frac{\partial \rho_1(\boldsymbol{\phi})}{\partial \phi_1} &= \frac{1}{1-\phi_2}, \\
\frac{\partial \rho_1(\boldsymbol{\phi})}{\partial \phi_2} &= \frac{\phi_1}{(1-\phi_2)^2}, \\
\frac{\partial \rho_2(\boldsymbol{\phi})}{\partial \phi_1} &= \frac{2\phi_1}{1-\phi_2}, \\
\frac{\partial \rho_2(\boldsymbol{\phi})}{\partial \phi_2} &= \frac{\phi_1^2}{(1-\phi_2)^2} + 1,
\end{aligned} \tag{C.12}$$

and so

$$\begin{aligned}
\frac{\partial^2 \rho_1(\boldsymbol{\phi})}{\partial \phi_1^2} &= 0,\\
\frac{\partial^2 \rho_1(\boldsymbol{\phi})}{\partial \phi_2 \partial \phi_1} &= \frac{\partial^2 \rho_1(\boldsymbol{\phi})}{\partial \phi_1 \partial \phi_2} = \frac{1}{(1-\phi_2)^2},\\
\frac{\partial^2 \rho_1(\boldsymbol{\phi})}{\partial \phi_2^2} &= \frac{2\phi_1}{(1-\phi_2)^3},
\end{aligned} \tag{C.13}$$

and

$$\begin{aligned}
\frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_1^2} &= \frac{2}{1-\phi_2},\\
\frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_2 \partial \phi_1} &= \frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_1 \partial \phi_2} = \frac{2\phi_1}{(1-\phi_2)^2},\\
\frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_2^2} &= \frac{2\phi_1^2}{(1-\phi_2)^3}.
\end{aligned} \tag{C.14}$$

For an AR(3) process let

$$\begin{aligned}
h_1 &= \phi_2^2 + \phi_2\phi_3^2 + \phi_1\phi_2\phi_3 - 2\phi_2 - \phi_3^2 - \phi_1\phi_3 + 1,\\
h_2 &= \phi_1 - \phi_1\phi_2 + \phi_2\phi_3 - \phi_2^2\phi_3,\\
h_3 &= \phi_3^2 + \phi_1\phi_3 + 2\phi_2 - 2,\\
h_4 &= \phi_1 + 2\phi_3 - \phi_1\phi_2 - 2\phi_2\phi_3,\\
h_5 &= \phi_3^2 + \phi_1\phi_3 + \phi_2 - 1,\\
h_6 &= \phi_1^2 + \phi_1\phi_3 - \phi_2^2 + \phi_2,\\
h_7 &= \phi_1\phi_2 - \phi_1\phi_2^2 + \phi_2^2\phi_3 - \phi_2^3\phi_3,\\
h_8 &= \phi_1^3 + \phi_1^2\phi_3 - \phi_1\phi_2^2 + \phi_1\phi_2,
\end{aligned} \tag{C.15}$$

Then for $\rho_1(\boldsymbol{\phi})$ we have

$$
\begin{aligned}
\frac{\partial\rho_1(\boldsymbol{\phi})}{\partial\phi_1} &= \frac{(\phi_3-\phi_2\phi_3)(\phi_1-\phi_1\phi_2+\phi_2\phi_3-\phi_2^2\phi_3)}{h_1^2}-\frac{\phi_2-1}{h_1},\\
\frac{\partial\rho_1(\boldsymbol{\phi})}{\partial\phi_2} &= -\frac{(\phi_3^2+\phi_1\phi_3+2\phi_2-2)(\phi_1-\phi_1\phi_2+\phi_2\phi_3-\phi_2^2\phi_3)}{h_1^2}-\frac{\phi_1-\phi_3+2\phi_2\phi_3}{h_1},\\
\frac{\partial\rho_1(\boldsymbol{\phi})}{\partial\phi_3} &= \frac{(\phi_1+2\phi_3-\phi_1\phi_2-2\phi_2\phi_3)(\phi_1-\phi_1\phi_2+\phi_2\phi_3-\phi_2^2\phi_3)}{h_1^2}+\frac{(\phi_2-\phi_2^2)}{h_1},
\end{aligned}
\tag{C.16}
$$

and so

$$
\begin{aligned}
\frac{\partial^2\rho_1(\boldsymbol{\phi})}{\partial\phi_1^2} &= \frac{2(\phi_3-\phi_2\phi_3)^2(\phi_1-\phi_1\phi_2+\phi_2\phi_3-\phi_2^2\phi_3)}{h_1^3}-\frac{2(\phi_3-\phi_2\phi_3)(\phi_2-1)}{h_1^2},\\
\frac{\partial^2\rho_1(\boldsymbol{\phi})}{\partial\phi_2\partial\phi_1} &= \frac{\partial^2\rho_1(\boldsymbol{\phi})}{\partial\phi_1\partial\phi_2}=\frac{(\phi_2-1)h_3}{h_1^2}-\frac{(\phi_3-\phi_2\phi_3)(\phi_1-\phi_3+2\phi_2\phi_3)}{h_1^2},\\
&\quad-\frac{1}{h_1}-\frac{\phi_3h_2}{h_1^2}-\frac{2(\phi_3-\phi_2\phi_3)h_3h_2}{h_1^3},\\
\frac{\partial^2\rho_1(\boldsymbol{\phi})}{\partial\phi_3\partial\phi_1} &= \frac{\partial^2\rho_1(\boldsymbol{\phi})}{\partial\phi_1\partial\phi_3}=\frac{(\phi_3-\phi_2\phi_3)(\phi_2-\phi_2^2)}{h_1^2}-\frac{(\phi_2-1)h_4}{h_1^2}-\frac{(\phi_2-1)h_2}{h_1^2},\\
&\quad+\frac{2(\phi_3-\phi_2\phi_3)h_4h_2}{h_1^3},\\
\frac{\partial^2\rho_1(\boldsymbol{\phi})}{\partial\phi_3\partial\phi_2} &= \frac{\partial^2\rho_1(\boldsymbol{\phi})}{\partial\phi_2\partial\phi_3}=\frac{2\phi_2-1}{h_1}-\frac{(\phi_1+2\phi_3)h_2}{h_1^2}-\frac{(\phi_1-\phi_3+2\phi_2\phi_3)h_4}{h_1^2},\\
&\quad-\frac{(\phi_2-\phi_2^2)h_3}{h_1^2}-\frac{2h_4h_3h_2}{h_1^3},\\
\frac{\partial^2\rho_1(\boldsymbol{\phi})}{\partial\phi_2^2} &= \frac{2h_3^2h_2}{h_1^3}-\frac{2\phi_3}{h_1}-\frac{2h_2}{h_1^2}+\frac{2(\phi_1-\phi_3+2\phi_2\phi_3)h_3}{h_1^2},\\
\frac{\partial^2\rho_1(\boldsymbol{\phi})}{\partial\phi_3^2} &= \frac{2(\phi_2-\phi_2^2)h_4}{h_1^2}-\frac{(2\phi_2-2)h_2}{h_1^2}+\frac{2h_4^2h_2}{h_1^3}.
\end{aligned}
\tag{C.17}
$$

For $\rho_2(\boldsymbol{\phi})$ we have

$$\begin{aligned}
\frac{\partial \rho_2(\boldsymbol{\phi})}{\partial \phi_1} &= \frac{\phi_3(\phi_1^2 + \phi_1\phi_3 - \phi_2^2 + \phi_2)}{h_5^2} - \frac{(2\phi_1 + \phi_3)}{h_5},\\
\frac{\partial \rho_2(\boldsymbol{\phi})}{\partial \phi_2} &= \frac{\phi_1^2 + \phi_1\phi_3 - \phi_2^2 + \phi_2}{h_5^2} + \frac{2\phi_2 - 1}{h_5},\\
\frac{\partial \rho_2(\boldsymbol{\phi})}{\partial \phi_3} &= \frac{(\phi_1 + 2\phi_3)(\phi_1^2 + \phi_1\phi_3 - \phi_2^2 + \phi_2)}{h_5^2} - \frac{\phi_1}{h_5},
\end{aligned}\tag{C.18}$$

and so

$$\begin{aligned}
\frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_1^2} &= \frac{2\phi_3(2\phi_1 + \phi_3)}{h_5^2} - \frac{2\phi_3^2(\phi_1^2 + \phi_1\phi_3 - \phi_2^2 + \phi_2)}{h_5^3} - \frac{2}{h_5},\\
\frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_2 \partial \phi_1} &= \frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_1 \partial \phi_2} = \frac{2\phi_1 + \phi_3}{h_5^2} - \frac{2\phi_3 h_6}{h_5^3} - \frac{\phi_3(2\phi_2 - 1)}{h_5^2}\\
\frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_3 \partial \phi_1} &= \frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_1 \partial \phi_3} = \frac{h_6}{h_5^2} - \frac{1}{h_5} + \frac{\phi_1\phi_3}{h_5^2} + \frac{(\phi_1 + 2\phi_3)(2\phi_1 + \phi_3)}{h_5^2}\\
&\quad - \frac{2\phi_3(\phi_1 + 2\phi_3)(\phi_1^2 + \phi_1\phi_3 - \phi_2^2 + \phi_2)}{h_5^3},\\
\frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_2 \partial \phi_3} &= \frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_3 \partial \phi_2} = \frac{\phi_1}{h_5^2} - \frac{2(\phi_1 + 2\phi_3)h_6}{h_5^3} - \frac{(2\phi_2 - 1)(\phi_1 + 2\phi_3)}{h_5^2}\\
\frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_2^2} &= \frac{2}{h_5} - \frac{2(\phi_1^2 + \phi_1\phi_3 - \phi_2^2 + \phi_2)}{h_5^3} - \frac{2(2\phi_2 - 1)}{h_5^2},\\
\frac{\partial^2 \rho_2(\boldsymbol{\phi})}{\partial \phi_3^2} &= \frac{2h_6}{h_5^2} - \frac{2(\phi_1 + 2\phi_3)^2 h_6}{h_5^3} + \frac{2\phi_1(\phi_1 + 2\phi_3)}{h_5^2}.
\end{aligned}\tag{C.19}$$

For $\rho_3(\boldsymbol{\phi})$ we have

$$\begin{aligned}
\frac{\partial \rho_3(\boldsymbol{\phi})}{\partial \phi_1} &= \frac{\phi_2 - \phi_2^2}{h_1} - \frac{(3\phi_1^2 + 2\phi_1\phi_3 - \phi_2^2 + \phi_2)}{h_5} \\
&\quad + \frac{(\phi_3 - \phi_2\phi_3)(\phi_1\phi_2 - \phi_1\phi_2^2 + \phi_2^2\phi_3 - \phi_2^3\phi_3)}{h_1^2} + \frac{\phi_3(\phi_1^3 + \phi_3\phi_1^2 - \phi_1\phi_2^2 + \phi_1\phi_2)}{h_5^2} \\
\frac{\partial \rho_3(\boldsymbol{\phi})}{\partial \phi_2} &= \frac{\phi_1^3 + \phi_1^2\phi_3 - \phi_1\phi_2^2 + \phi_1\phi_2}{h_5^2} + \frac{\phi_1 - 2\phi_1\phi_2 + 2\phi_2\phi_3 - 3\phi_2^2\phi_3}{h_1} \\
&\quad - \frac{(\phi_1 - 2\phi_1\phi_2}{h_5)} - \frac{(\phi_3^2 + \phi_1\phi_3 + 2\phi_2 - 2)(\phi_1\phi_2 - \phi_1\phi_2^2 + \phi_2^2\phi_3 - \phi_2^3\phi_3)}{h_1^2} \\
\frac{\partial \rho_3(\boldsymbol{\phi})}{\partial \phi_3} &= \frac{\phi_2^2 - \phi_2^3}{h_1} - \frac{\phi_1^2}{h_5} + \frac{(\phi_1 + 2\phi_3 - \phi_1\phi_2 - 2\phi_2\phi_3)(\phi_1\phi_2 - \phi_1\phi_2^2 + \phi_2^2\phi_3 - \phi_2^3\phi_3)}{h_1^2} \\
&\quad + \frac{(\phi_1 + 2\phi_3)(\phi_1^3 + \phi_1^2\phi_3 - \phi_1\phi_2^2 + \phi_1\phi_2)}{h_5^2} + 1, \qquad \text{(C.20)}
\end{aligned}$$

and so

$$\begin{aligned}
\frac{\partial^2 \rho_3(\boldsymbol{\phi})}{\partial \phi_1^2} &= \frac{2(\phi_3 - \phi_2\phi_3)^2(\phi_1\phi_2 - \phi_1\phi_2^2 + \phi_2^2\phi_3 - \phi_2^3\phi_3)}{h_1^3} - \frac{2\phi_3^2(\phi_1^3 + \phi_3\phi_1^2 - \phi_1\phi_2^2 + \phi_1\phi_2)}{h_5^3} \\
&\quad - \frac{(6\phi_1 + 2\phi_3)}{h_5} + \frac{2\phi_3(3\phi_1^2 + 2\phi_1\phi_3 - \phi_2^2 + \phi_2)}{h_5^2} \\
&\quad + \frac{2(\phi_3 - \phi_2\phi_3)(\phi_2 - \phi_2^2)}{h_1^2}, \\
\frac{\partial^2 \rho_3(\boldsymbol{\phi})}{\partial \phi_2 \partial \phi_1} &= \frac{\partial^2 \rho_3(\boldsymbol{\phi})}{\partial \phi_1 \partial \phi_2} = \frac{3\phi_1^2 + 2\phi_1\phi_3 - \phi_2^2 + \phi_2}{h_5^2} + \frac{\phi_3(\phi_1 - 2\phi_1\phi_2)}{h_5^2} - \frac{\phi_3 h_7}{h_1^2} \\
&\quad + \frac{(\phi_3 - \phi_2\phi_3)(\phi_1 - 2\phi_1\phi_2 + 2\phi_2\phi_3 - 3\phi_2^2\phi_3)}{h_1^2} - \frac{(\phi_2 - \phi_2^2)h_3}{h_1^2} \\
&\quad - \frac{2\phi_3(\phi_1^3 + \phi_1^2\phi_3 - \phi_1\phi_2^2 + \phi_1\phi_2)}{h_5^3} - \frac{2(\phi_3 - \phi_2\phi_3)h_3h_7}{h_1^3}, \\
\frac{\partial^2 \rho_3(\boldsymbol{\phi})}{\partial \phi_3 \partial \phi_1} &= \frac{\partial^2 \rho_3(\boldsymbol{\phi})}{\partial \phi_1 \partial \phi_3} = \frac{h_8}{h_5^2} - \frac{2\phi_1}{h_5} + \frac{(\phi_2 - \phi_2^2)h_4}{h_1^2} - \frac{(\phi_2 - 1)h_7}{h_1^2} + \frac{\phi_1^2\phi_3}{h_5^2} \\
&\quad + \frac{(\phi_1 + 2\phi_3)(3\phi_1^2 + 2\phi_1\phi_3 - \phi_2^2 + \phi_2)}{h_5^2} + \frac{(\phi_3 - \phi_2\phi_3)(\phi_2^2 - \phi_2^3)}{h_1^2} \\
&\quad + \frac{2(\phi_3 - \phi_2\phi_3)h_4h_7}{h_1^3} - \frac{2\phi_3(\phi_1 + 2\phi_3)h_8}{h_6^3}, \\
\frac{\partial^2 \rho_3(\boldsymbol{\phi})}{\partial \phi_3 \partial \phi_2} &= \frac{\partial^2 \rho_3(\boldsymbol{\phi})}{\partial \phi_2 \partial \phi_3} = \frac{\phi_1^2}{h_5^2} + \frac{2\phi_2 - 3\phi_2^2}{h_1} - \frac{(\phi_2^2 - \phi_2^3)h_3}{h_1^2} + \frac{(\phi_1 - 2\phi_1\phi_2)(\phi_1 + 2\phi_3)}{h_5^2} \\
&\quad + \frac{h_4(\phi_1 - 2\phi_1\phi_2 + 2\phi_2\phi_3 - 3\phi_2^2\phi_3)}{h_1^2} - \frac{(\phi_1 + 2\phi_3)h_7}{h_1^2} \\
&\quad - \frac{2(\phi_1 + 2\phi_3)(\phi_1^3 + \phi_1^2\phi_3 - \phi_1\phi_2^2 + \phi_1\phi_2)}{h_5^3} - \frac{2h_4h_3h_7}{h_1^3}, \\
\frac{\partial^2 \rho_3(\boldsymbol{\phi})}{\partial \phi_2^2} &= \frac{2\phi_1}{h_5} - \frac{2(\phi_1^3 + \phi_1^2\phi_3 - \phi_1\phi_2^2 + \phi_1\phi_2)}{h_5^3} + \frac{2(\phi_1 - 2\phi_1\phi_2)}{h_5^2} - \frac{(2\phi_1 - 2\phi_3 + 6\phi_2\phi_3)}{h_1} \\
&\quad - \frac{2h_7}{h_1^2} + \frac{2h_3^2h_7}{h_1^3} - \frac{2h_3(\phi_1 - 2\phi_1\phi_2 + 2\phi_2\phi_3 - 3\phi_2^2\phi_3)}{h_1^2}, \\
\frac{\partial^2 \rho_3(\boldsymbol{\phi})}{\partial \phi_3^2} &= \frac{2h_8}{h_5^2} + \frac{2h_1^2(1 + 2\phi_3)}{h_5^2} - \frac{(2\phi_2 - 2)h_7}{h_1^2} - \frac{2(\phi_1 + 2\phi_3)^2h_8}{h_5^3} \\
&\quad + \frac{2h_4^2h_7}{h_1^3} + \frac{2(\phi_2^2 - \phi_2^3)h_4}{h_1^2}.
\end{aligned} \tag{C.21}$$

Thus to calculate $\boldsymbol{D}_{\boldsymbol{\phi}}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)$ given in equation C.4 we need the first-order partial derivatives $\boldsymbol{D}_{\boldsymbol{\phi}}[\rho_{|j-k|}(\boldsymbol{\phi})]$ for $j, k = 1, ..., n$ given by equation C.5 whose r elements are defined recursively by equation C.8. In order to start the recursive process off, for $r \leq 3$ the initial derivatives are given in equations C.10,C.12,C.16. Similarly to calculate $\boldsymbol{H}_{\boldsymbol{\phi}}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)$ given in equation (C.6) we need the second-order derivatives $\boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i)_{jk})$ for $j, k = 1, ..., n$ given in (C.7) whose elements are defined recursively by equation C.9. For $r \leq 3$ the initial derivatives are given in equations C.11,C.13C.14,C.17,C.19,C.21.

C.1.2 Line 1 of the information matrix

$\boldsymbol{H}_{\boldsymbol{\beta}}\left(L(\boldsymbol{\beta})\right)$:

We firstly derive $\boldsymbol{D}_{\boldsymbol{\beta}}\left(L(\boldsymbol{\beta})\right)$ which will in turn be used to derive all the elements of the top row of C.2. Now $L(\boldsymbol{\beta}) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^p$, so that by the first identification table (Table B.1) we have that if $\boldsymbol{d}\left(L(\boldsymbol{\beta})\right) = \boldsymbol{a}^\intercal \boldsymbol{d}\left(\boldsymbol{\beta}\right)$ for $\boldsymbol{a} \in \mathbb{R}^p$ then $\boldsymbol{D}_{\boldsymbol{\beta}}\left(L(\boldsymbol{\beta})\right) = \boldsymbol{a}^\intercal$. Now from (C.1) we have $\boldsymbol{d}\left(L(\boldsymbol{\beta})\right) = -\sum_{i=1}^N w_i \tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{d}\left(\tilde{\boldsymbol{e}}_i\right)$, so that

$$\begin{aligned}\boldsymbol{d}\left(L(\boldsymbol{\beta})\right) &= -\sum_{i=1}^N w_i \tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\left(-\boldsymbol{X}_i \boldsymbol{d}\left(\boldsymbol{\beta}\right)\right) \\ &= \left\{\sum_{i=1}^N w_i \boldsymbol{y}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{X}_i - \boldsymbol{\beta}^\intercal \sum_{i=1}^m w_i \boldsymbol{X}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{X}_i\right\} \boldsymbol{d}\left(\boldsymbol{\beta}\right). \qquad \text{(C.22)}\end{aligned}$$

We see that the $1 \times p$ vector $\boldsymbol{D}_{\boldsymbol{\beta}}\left(L(\boldsymbol{\beta})\right)$ is given by the expression in curly parentheses in (C.22). Now from the second identification table (Table B.2) we have that if $\boldsymbol{d}^2\left(L(\boldsymbol{\beta})\right) = \left(\boldsymbol{d}\beta\right)^\intercal \boldsymbol{B}\left(\boldsymbol{d}\beta\right)$, where $\boldsymbol{B}$ is a $p \times p$ matrix, then $\boldsymbol{H}_{\boldsymbol{\beta}}\left(l(\boldsymbol{\beta})\right) = (1/2)\left(\boldsymbol{B} + \boldsymbol{B}^\intercal\right)$. Now from (C.22) we have

$$\begin{aligned}\boldsymbol{d}^2\left(L(\boldsymbol{\beta})\right) &= \sum_{i=1}^{N} w_i \boldsymbol{y}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{X}_i \boldsymbol{d}^2\boldsymbol{\beta} - \sum_{i=1}^{N} w_i (\boldsymbol{d\beta})^{\intercal} \boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{X}_i \boldsymbol{d\beta} \\ &- \sum_{i=1}^{N} w_i \boldsymbol{\beta}^{\intercal} \boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{X}_i \boldsymbol{d}^2\boldsymbol{\beta} \\ &= (\boldsymbol{d\beta})^{\intercal} \left\{ - \sum_{i=1}^{N} w_i \boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{X}_i \right\} \boldsymbol{d\beta}, \end{aligned} \tag{C.23}$$

since $\boldsymbol{d}^2\boldsymbol{\beta} = 0$. The $p \times p$ matrix in curly brackets is $\boldsymbol{B}$. This matrix is symmetrical so that $(1/2)\left(\boldsymbol{B} + \boldsymbol{B}^{\intercal}\right) = \boldsymbol{B}$. We then have

$$\boldsymbol{H}_{\boldsymbol{\beta}}\left(L(\boldsymbol{\beta})\right) = - \sum_{i=1}^{N} w_i \boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{X}_i, \tag{C.24}$$

which is a $p \times p$ matrix as required. Thus we have

$$- \boldsymbol{E}\left[\boldsymbol{H}_{\boldsymbol{\beta}}\left(L(\boldsymbol{\beta})\right)\right] = \sum_{i=1}^{N} w_i \boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{X}_i. \tag{C.25}$$

$\boldsymbol{D}^2_{(\boldsymbol{\alpha})(\boldsymbol{\beta})}\left(L(\boldsymbol{\beta})\right)$ for $\boldsymbol{\alpha} \in \{\sigma^2, \boldsymbol{\phi}^{\intercal}, \boldsymbol{\psi}^{\intercal}\sigma^2\}^{\intercal}$:

To derive the cross partial derivatives we let $\boldsymbol{\alpha} \in \{\sigma^2, \boldsymbol{\phi}^{\intercal}, \boldsymbol{\psi}^{\intercal}\sigma^2\}^{\intercal}$ be a $n_{\boldsymbol{\alpha}} \times 1$ vector, where $\boldsymbol{\psi} = \mathrm{v}(\boldsymbol{D})$. We define the function $g(\boldsymbol{\alpha}) = \boldsymbol{D}_{\boldsymbol{\beta}}\left(L(\boldsymbol{\beta})\right)^{\intercal}$ so that $g(\boldsymbol{\alpha}) : S \longrightarrow \mathbb{R}^p$, $S \subseteq \mathbb{R}^{n_{\boldsymbol{\alpha}}}$. From the first identification table (Table B.1) we have that if $\boldsymbol{d}\left(g(\boldsymbol{\alpha})\right) = \boldsymbol{A}\boldsymbol{d\alpha}$, where $\boldsymbol{A}$ is a $p \times n_{\boldsymbol{\alpha}}$ matrix, then $\boldsymbol{D}_{\boldsymbol{\alpha}}\left(g(\boldsymbol{\alpha})\right) = \boldsymbol{D}^2_{(\boldsymbol{\alpha})(\boldsymbol{\beta})}\left(L(\boldsymbol{\beta})\right) = \boldsymbol{A}$. Now

$$
\begin{aligned}
\boldsymbol{d}\left(g(\boldsymbol{\alpha})\right) &= \sum_{i=1}^{N} w_i \boldsymbol{X}_i^{\intercal} \boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) \boldsymbol{y}_i - \sum_{i=1}^{N} w_i \boldsymbol{X}_i^{\intercal} \boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) \boldsymbol{X}_i \boldsymbol{\beta} \\
&= -\sum_{i=1}^{N} w_i \boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right) \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{y}_i + \sum_{i=1}^{N} w_i \boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right) \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{X}_i \boldsymbol{\beta} \\
&= -\sum_{i=1}^{N} w_i \mathrm{vec}\left[\boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right) \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{y}_i\right] \\
&+ \sum_{i=1}^{N} w_i \mathrm{vec}\left[\boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right) \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \boldsymbol{X}_i \boldsymbol{\beta}\right] \\
&= \sum_{i=1}^{N} w_i \left((\boldsymbol{\beta}^{\intercal} \boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right) \mathrm{vec}[\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)] \\
&- \sum_{i=1}^{N} w_i \left((\boldsymbol{y}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right) \mathrm{vec}[\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)] \\
&= \left\{ \sum_{i=1}^{N} w_i \left((\boldsymbol{\beta}^{\intercal} \boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right) \boldsymbol{D}_{\boldsymbol{\alpha}}(\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}))) \right. \\
&\left. - \sum_{i=1}^{N} w_i \left((\boldsymbol{y}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right) \boldsymbol{D}_{\boldsymbol{\alpha}}(\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}))) \right\} \boldsymbol{d\alpha},
\end{aligned}
\tag{C.26}
$$

so that the $p \times n_{\boldsymbol{\alpha}}$ matrix $\boldsymbol{D}^2_{(\boldsymbol{\alpha})(\boldsymbol{\beta})}\left(l(\boldsymbol{\beta})\right)$ is given by

$$
\begin{aligned}
\boldsymbol{D}^2_{(\boldsymbol{\alpha})(\boldsymbol{\beta})}\left(L(\boldsymbol{\beta})\right) &= \sum_{i=1}^{N} w_i \left((\boldsymbol{\beta}^{\intercal} \boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right) \boldsymbol{D}_{\boldsymbol{\alpha}}(\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}))) \\
&- \sum_{i=1}^{N} w_i \left((\boldsymbol{y}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right) \boldsymbol{D}_{\boldsymbol{\alpha}}(\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}))) \\
&= \sum_{i=1}^{N} w_i \left[\left(-\left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}\right)^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) \otimes \left(\boldsymbol{X}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right] \boldsymbol{D}_{\boldsymbol{\alpha}}(\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}))),
\end{aligned}
\tag{C.27}
$$

and

$$
\begin{aligned}
-\boldsymbol{E}\left[\boldsymbol{D}^2_{(\boldsymbol{\alpha})(\boldsymbol{\beta})}\left(l(\boldsymbol{\beta})\right)\right] &= -\sum_{i=1}^{N} w_i \left((\boldsymbol{\beta}^\intercal \boldsymbol{X}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{X}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right) \boldsymbol{D}_{\boldsymbol{\alpha}}(\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})))\\
&\quad + \sum_{i=1}^{N} w_i \left((\boldsymbol{E}\left[\boldsymbol{y}_i\right]^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{X}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right) \boldsymbol{E}\left[\boldsymbol{D}_{\boldsymbol{\alpha}}(\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})))\right]\\
&= -\sum_{i=1}^{N} w_i \left((\boldsymbol{\beta}^\intercal \boldsymbol{X}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{X}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right) \boldsymbol{E}\left[\boldsymbol{D}_{\boldsymbol{\alpha}}(\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})))\right]\\
&\quad + \sum_{i=1}^{N} w_i \left((\boldsymbol{\beta}^\intercal \boldsymbol{X}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{X}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right) \boldsymbol{E}\left[\boldsymbol{D}_{\boldsymbol{\alpha}}(\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})))\right]\\
&= \boldsymbol{0}. \qquad (C.28)
\end{aligned}
$$

C.1.3 Line 2 of the information matrix

$\boldsymbol{H}_{\sigma^2}\left(L(\sigma^2)\right)$:

We shall first derive $\boldsymbol{D}_{\sigma^2}\left(L(\sigma^2)\right)$, which will also yield $\boldsymbol{d}\left(L(\sigma^2)\right)$ that is required in order to calculate $\boldsymbol{d}^2\left(L(\sigma^2)\right)$. Now $L(\sigma^2) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}$, so that from the first identification table (Table B.1) we have that if $\boldsymbol{d}\left(L(\sigma^2)\right) = \alpha \boldsymbol{d}\sigma^2$ for $\alpha \in \mathbb{R}$, then $\boldsymbol{D}_{\sigma^2}\left(L(\sigma^2)\right) = \alpha$. We have

$$
\begin{aligned}
\boldsymbol{d}\left(L(\sigma^2)\right) &= -\frac{1}{2}\sum_{i=1}^{N} w_i \boldsymbol{d}\left(\log|\boldsymbol{V}_i(\boldsymbol{\zeta})|\right) - \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\tilde{\boldsymbol{e}}_i\\
&= -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{d}\sigma^2\right] + \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{d}\sigma^2\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\\
&= \left\{-\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right] + \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right\}\boldsymbol{d}\sigma^2,
\end{aligned}
\qquad (C.29)
$$

so we see that

$$\boldsymbol{D}_{\sigma^2}\left(L(\sigma^2)\right) = -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right] + \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i, \tag{C.30}$$

which is a scalar as required. Now $L(\sigma^2) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}$, so that from the second identification table (Table B.2), we have that if $\boldsymbol{d}^2\left(L(\sigma^2)\right) = \beta(\boldsymbol{d}\left(\sigma^2\right))^2$ for $\beta \in \mathbb{R}$, then $\boldsymbol{H}_{\sigma^2}\left(L(\sigma^2)\right) = \beta$. We will also need the result

$$\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{d}\sigma^2\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) = \boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\sigma^2 + \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{d}\sigma^2. \tag{C.31}$$

So from the second line of (C.29), and using $\boldsymbol{d}^2\sigma^2 = 0$ we have

$$
\begin{aligned}
\boldsymbol{d}^2\left(L(\sigma^2)\right) &= -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{C}_i(\boldsymbol{\phi})\right]\boldsymbol{d}\sigma^2 + \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{d}\sigma^2\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\tilde{\boldsymbol{e}}_i \\
&= -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{C}_i(\boldsymbol{\phi})\right]\boldsymbol{d}\sigma^2 + \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\boldsymbol{d}\sigma^2 \\
&+ \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\tilde{\boldsymbol{e}}_i\boldsymbol{d}\sigma^2 \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right]\boldsymbol{d}\sigma^2 \\
&- \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\boldsymbol{d}\sigma^2 \\
&- \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\boldsymbol{d}\sigma^2 \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right]\left(\boldsymbol{d}\sigma^2\right)^2 \\
&- \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\left(\boldsymbol{d}\sigma^2\right)^2 \\
&- \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\left(\boldsymbol{d}\sigma^2\right)^2 \\
&= \left\{\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right]\right. \\
&\left. -\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right\}\left(\boldsymbol{d}\sigma^2\right)^2, \qquad \text{(C.32)}
\end{aligned}
$$

so we see that the scalar Hessian $\boldsymbol{H}_{\sigma^2}\left(L(\sigma^2)\right)$ is given by the expression within curly brackets in (C.32). Taking expectations of this involves calculating $\boldsymbol{E}[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{A}\tilde{\boldsymbol{e}}_i] = \mathrm{tr}(\boldsymbol{A}\mathrm{Var}[\tilde{\boldsymbol{e}}_i])$, where $\boldsymbol{A} = \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}$. Since $\boldsymbol{A}\mathrm{Var}[\tilde{\boldsymbol{e}}_i] = \boldsymbol{A}\mathrm{Var}[\boldsymbol{Y}_i] = \boldsymbol{A}\boldsymbol{V}_i(\boldsymbol{\zeta}) = \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})$ we have

$$
\begin{aligned}
-\boldsymbol{E}\left[\boldsymbol{H}_{\sigma^2}\left(L(\sigma^2)\right)\right] &= -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right] \\
&\quad + \sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right] \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right]. \qquad \text{(C.33)}
\end{aligned}
$$

$\boldsymbol{D}^2_{(\mathrm{v}(\boldsymbol{D}))(\sigma^2)}\left(L(\sigma^2)\right)$:

Let $g(\mathrm{v}(\boldsymbol{D})) = \boldsymbol{D}_{\sigma^2}\left(L(\sigma^2)\right)$, so that $\boldsymbol{D}_{\mathrm{v}(\boldsymbol{D})}\left(g(\mathrm{v}(\boldsymbol{D}))\right) = \boldsymbol{D}^2_{(\mathrm{v}(\boldsymbol{D}))(\sigma^2)}\left(L(\sigma^2)\right)$. Now $g(\mathrm{v}(\boldsymbol{D})) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^{q(q+1)/2}$, so that from the first identification table (Table B.1) we have that if $\boldsymbol{d}\left(g(\mathrm{v}(\boldsymbol{D}))\right) = \boldsymbol{a}^{\intercal}\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right)$ for $\boldsymbol{a} \in \mathbb{R}^{q(q+1)/2}$, then $\boldsymbol{D}^2_{(\mathrm{v}(\boldsymbol{D}))(\sigma^2)}\left(L(\sigma^2)\right) = \boldsymbol{a}^{\intercal}$. We will need two preliminary results for what follows. Firstly when $\boldsymbol{V}_i(\boldsymbol{\zeta})$ is viewed as a function of $\boldsymbol{D}$ we have

$$
\begin{aligned}
\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) &= -\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\boldsymbol{d}\left(\boldsymbol{D}\right)\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \\
&\quad -\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\boldsymbol{d}\left(\boldsymbol{D}\right)\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}. \qquad \text{(C.34)}
\end{aligned}
$$

Secondly consider $(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{A} \otimes \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{B})$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are $n_i \times c_1$ and $n_i \times c_2$ matrices respectively. Then $\boldsymbol{A}^{\intercal}\tilde{\boldsymbol{e}}_i \in \mathbb{R}^{c_1}$, $\boldsymbol{B}^{\intercal}\tilde{\boldsymbol{e}}_i \in \mathbb{R}^{c_2}$, and so we have

$$\begin{aligned}
\boldsymbol{E}\left[(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{A}\otimes\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{B})\right] &= \boldsymbol{E}\left[\left\{\boldsymbol{A}^{\intercal}\tilde{\boldsymbol{e}}_i\otimes\boldsymbol{B}^{\intercal}\tilde{\boldsymbol{e}}_i\right\}^{\intercal}\right] \\
&= \left\{\boldsymbol{E}\left[\boldsymbol{A}^{\intercal}\tilde{\boldsymbol{e}}_i\otimes\boldsymbol{B}^{\intercal}\tilde{\boldsymbol{e}}_i\right]\right\}^{\intercal} \\
&= \left\{\boldsymbol{E}\left[\mathrm{vec}\left(\boldsymbol{B}^{\intercal}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{A}\right)\right]\right\}^{\intercal} \\
&= \left\{\mathrm{vec}\left(\boldsymbol{B}^{\intercal}\boldsymbol{E}\left[\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i\right]^{\intercal}\boldsymbol{A}\right)\right\}^{\intercal} \\
&= \left\{\mathrm{vec}\left(\boldsymbol{B}^{\intercal}\mathrm{Var}[\tilde{\boldsymbol{e}}_i]\boldsymbol{A}\right)\right\}^{\intercal} \\
&= \left\{\mathrm{vec}\left(\boldsymbol{B}^{\intercal}\mathrm{Var}[\boldsymbol{Y}_i]\boldsymbol{A}\right)\right\}^{\intercal} \\
&= \left\{\mathrm{vec}\left(\boldsymbol{B}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})\boldsymbol{A}\right)\right\}^{\intercal}, \qquad \text{(C.35)}
\end{aligned}$$

which is a $1\times c_1c_2$ vector. Now from (C.30) we have

$$
\begin{aligned}
\boldsymbol{d}\left(g(\mathrm{v}(\boldsymbol{D}))\right) &= -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{C}_i(\boldsymbol{\phi})\right] + \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\tilde{\boldsymbol{e}}_i \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right] \\
&\quad - \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{vec}\left[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\boldsymbol{d}\left(\boldsymbol{D}\right)\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right] \\
&\quad - \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{vec}\left[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\boldsymbol{d}\left(\boldsymbol{D}\right)\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right] \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\boldsymbol{d}\left(\boldsymbol{D}\right)\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right] \\
&\quad - \frac{1}{2}\sum_{i=1}^{N} w_i\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{D}\right)\right] \\
&\quad - \frac{1}{2}\sum_{i=1}^{N} w_i\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{D}\right)\right] \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i\left[\mathrm{vec}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right]^{\intercal}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i\right)\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{D}\right)\right] \\
&\quad - \frac{1}{2}\sum_{i=1}^{N} w_i\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right) \\
&\quad - \frac{1}{2}\sum_{i=1}^{N} w_i\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right) \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i\left[\mathrm{vec}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right]^{\intercal}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right) \\
&\quad - \frac{1}{2}\sum_{i=1}^{N} w_i\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right) \\
&\quad - \frac{1}{2}\sum_{i=1}^{N} w_i\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right) \\
&= \left\{\frac{1}{2}\sum_{i=1}^{N} w_i\left[\mathrm{vec}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right]^{\intercal}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q\right. \\
&\qquad - \frac{1}{2}\sum_{i=1}^{N} w_i\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q \\
&\qquad \left.-\frac{1}{2}\sum_{i=1}^{N} w_i\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q\right\}\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right).
\end{aligned}
\tag{C.36}
$$

Thus the expression in the curly brackets in the last line of (C.36) is $\boldsymbol{D}^2_{(\mathrm{v}(\boldsymbol{D}))(\sigma^2)}\left(L(\sigma^2)\right)$, which is a $1 \times (q(q+1)/2)$ vector as required. Now taking expectations of this derivative vector and using (C.35), we have

$\boldsymbol{E}[\boldsymbol{e}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{e}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i]$ and $\boldsymbol{E}[\boldsymbol{e}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{e}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i]$ are both equal to $[\mathrm{vec}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)]^{\intercal}$. Thus we have

$$\begin{aligned}
-\boldsymbol{E}\left[\boldsymbol{D}^2_{(\mathrm{v}(\boldsymbol{D}))(\sigma^2)}\left(L(\sigma^2)\right)\right] &= -\frac{1}{2}\sum_{i=1}^{N} w_i\left[\mathrm{vec}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right]^{\intercal}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q \\
&\quad + \sum_{i=1}^{N} w_i[\mathrm{vec}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)]^{\intercal}\widetilde{\boldsymbol{D}}_q \\
&= -\frac{1}{2}\sum_{i=1}^{N} w_i[\mathrm{vec}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)]^{\intercal}\widetilde{\boldsymbol{D}}_q \\
&\quad + \sum_{i=1}^{N} w_i[\mathrm{vec}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)]^{\intercal}\widetilde{\boldsymbol{D}}_q \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i[\mathrm{vec}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)]^{\intercal}\widetilde{\boldsymbol{D}}_q. \qquad \text{(C.37)}
\end{aligned}$$

Let $g(\phi) = \boldsymbol{D}_{\sigma^2}\left(L(\sigma^2)\right)$, so that $\boldsymbol{D}_{\phi}\left(g(\phi)\right) = \boldsymbol{D}^2_{(\boldsymbol{\phi})(\sigma^2)}\left(L(\sigma^2)\right)$. Now $g(\boldsymbol{\phi}) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^r$, so that from the first identification table (Table B.1) we have that if $\boldsymbol{d}\left(g(\boldsymbol{\phi})\right) = \boldsymbol{a}^{\intercal}\boldsymbol{d\phi}$ for $\boldsymbol{a} \in \mathbb{R}^r$, then $\boldsymbol{D}_{\boldsymbol{\phi}}\left(g(\boldsymbol{\phi})\right) = \boldsymbol{a}^{\intercal}$. In order to identify this vector of partial derivatives we shall need the following result. When $\boldsymbol{V}_i(\boldsymbol{\zeta})$ is viewed as a function of only $\boldsymbol{\phi}$ then

$$\begin{aligned}
&\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) = -\sigma^2\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \\
&-\sigma^2\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} + \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}. \qquad \text{(C.38)}
\end{aligned}$$

So from (C.30) we have

$$
\begin{aligned}
\boldsymbol{d}\left(g(\boldsymbol{\phi})\right) &= -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{C}_i(\boldsymbol{\phi}) + \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right] \\
&\quad + \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\tilde{\boldsymbol{e}}_i \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}) - \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right] \\
&\quad + \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\tilde{\boldsymbol{e}}_i \\
&= \frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\right] - \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right] \\
&\quad - \frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \mathrm{vec}\left[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right] \\
&\quad + \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{vec}\left[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right] \\
&\quad - \frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \mathrm{vec}\left[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right] \\
&= \frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \left[\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right]^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right] \\
&\quad - \frac{1}{2}\sum_{i=1}^{N}\left[\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right]^{\intercal}\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right] \\
&\quad - \frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right] \\
&\quad + \frac{1}{2}\sum_{i=1}^{N} w_i \left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right] \\
&\quad - \frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right] \\
&= \Bigg\{\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \left[\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right]^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \\
&- \frac{1}{2}\sum_{i=1}^{N} w_i \left[\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right]^{\intercal}\boldsymbol{D}_{\boldsymbol{\phi}}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \\
&\quad - \frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}}\left(vec(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \\
&\quad + \frac{1}{2}\sum_{i=1}^{N} w_i \left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}}\left(vec(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \\
&\quad -\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\Bigg\}\boldsymbol{d\phi},
\end{aligned}
\tag{C.39}
$$

where the last expression in curly brackets is an $1 \times r$ vector and so is equal to $\boldsymbol{D}^2_{(\boldsymbol{\phi})(\sigma^2)}\left(L(\sigma^2)\right)$. Now using (C.35) we get that $\boldsymbol{E}[(\tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \otimes \tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})] = \text{vec}[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}]^\intercal$, $\boldsymbol{E}[(\tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \otimes \tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})] = \text{vec}[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}]^\intercal$, and $\boldsymbol{E}[(\tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \otimes \tilde{\boldsymbol{e}}_i^\intercal \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})] = \text{vec}[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}]^\intercal$. Thus from (C.39) we have that

$$\begin{aligned}
-\boldsymbol{E}\left[\boldsymbol{D}^2_{(\boldsymbol{\phi})(\sigma^2)}\left(L(\sigma^2)\right)\right] &= -\frac{\sigma^2}{2}\sum_{i=1}^N w_i\left[\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right]^\intercal\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}}\left(\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \\
&\quad + \frac{1}{2}\sum_{i=1}^N w_i\left[\text{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right]^\intercal\boldsymbol{D}_{\boldsymbol{\phi}}\left(\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \\
&\quad - \frac{1}{2}\sum_{i=1}^N w_i\left[\text{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right]^\intercal\boldsymbol{D}_{\boldsymbol{\phi}}\left(\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \\
&\quad + \sigma^2\sum_{i=1}^N w_i\left[\text{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{C}_i(\boldsymbol{\phi})\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right]^\intercal\boldsymbol{D}_{\boldsymbol{\phi}}\left(\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \\
&= -\frac{\sigma^2}{2}\sum_{i=1}^N w_i\left[\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right]^\intercal\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}}\left(\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \\
&\quad + \sigma^2\sum_{i=1}^N w_i\left[\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right]^\intercal\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}}\left(\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \\
&= \frac{\sigma^2}{2}\sum_{i=1}^N w_i\left[\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right]^\intercal\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}}\left(\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right),
\end{aligned} \tag{C.40}$$

where $\boldsymbol{D}_{\boldsymbol{\phi}}\left(\text{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)$ is given in equation C.4, and where a summary of all the necessary equations required to calculate this can be found at the end of subsection C.1.1.

C.1.4 Line 3 of the information matrix

$\boldsymbol{H}_{\boldsymbol{\psi}}\left(L(\boldsymbol{\psi})\right)$ where $\boldsymbol{\psi} = \text{v}(\boldsymbol{D})$:

To account for the symmetry of $\boldsymbol{D}$ we take derivatives of $L(\boldsymbol{\theta})$ with respect to $\mathrm{v}(\boldsymbol{D})$ rather than $\mathrm{vec}(\boldsymbol{D})$, and so we want to derive $\boldsymbol{H}_{\mathrm{v}(\boldsymbol{D})}\left(L(\mathrm{v}(\boldsymbol{D}))\right)$. Now $L(\mathrm{v}(\boldsymbol{D})) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^{(q(q+1)/2)}$, so that from the second identification table (Table B.2) we have that if $\boldsymbol{d}^2\left(L(\mathrm{v}(\boldsymbol{D}))\right) = [\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right)]^{\intercal}\boldsymbol{B}[\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right)]$ for a $(q(q+1)/2) \times (q(q+1)/2)$ matrix $\boldsymbol{B}$ then $\boldsymbol{H}_{\mathrm{v}(\boldsymbol{D})}\left(L(\mathrm{v}(\boldsymbol{D}))\right) = (1/2)\left(\boldsymbol{B} + \boldsymbol{B}^{\intercal}\right)$. Computing the differential of (C.1) we get

$$\boldsymbol{d}\left(L(\mathrm{v}(\boldsymbol{D}))\right) = -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)] + \frac{1}{2}\sum_{i=1}^{N} w_i(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i). \tag{C.41}$$

Now computing the differential of this expression can be greatly simplified by noting that $\mathrm{vec}(\boldsymbol{d}^2\left(\boldsymbol{D}\right)) = \boldsymbol{d}^2\left(\mathrm{vec}(\boldsymbol{D})\right) = \widetilde{\boldsymbol{D}}_q\boldsymbol{d}^2\left(\mathrm{v}(\boldsymbol{D})\right) = 0$. This implies that $\boldsymbol{d}^2\left(\boldsymbol{D}\right) = 0$. Accordingly when $\boldsymbol{V}_i(\boldsymbol{\zeta})$ is viewed as a function of $\mathrm{v}(\boldsymbol{D})$, since $\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right) = \boldsymbol{Z}_i\boldsymbol{d}^2\left(\boldsymbol{D}\right)\boldsymbol{Z}_i^{\intercal}$, we have that $\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right) = 0$. Using this result, and again when $\boldsymbol{V}_i(\boldsymbol{\zeta})$ is viewed as a function of $\mathrm{v}(\boldsymbol{D})$, we also get the result that

$$\boldsymbol{d}\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right) = -2\tilde{\boldsymbol{e}}_i\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i. \tag{C.42}$$

Using both of these results we have

$$
\begin{aligned}
\boldsymbol{d}^2\left(L(\mathrm{v}(\boldsymbol{D}))\right) &= \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left\{\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right\} \\
&\quad -\sum_{i=1}^{N} w_i\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right) \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left\{\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\boldsymbol{d}\left(\boldsymbol{D}\right)\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\left(\boldsymbol{d}\left(\boldsymbol{D}\right)\boldsymbol{Z}_i^{\intercal}\right)\right\} \\
&\quad -\sum_{i=1}^{N} w_i \mathrm{tr}\left\{\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\boldsymbol{d}\left(\boldsymbol{D}\right)\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\left(\boldsymbol{d}\left(\boldsymbol{D}\right)\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right)\right\} \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left\{\boldsymbol{d}\left(\boldsymbol{D}\right)\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\boldsymbol{d}\left(\boldsymbol{D}\right)\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\right\} \\
&\quad -\sum_{i=1}^{N} w_i \mathrm{tr}\left\{\boldsymbol{d}\left(\boldsymbol{D}\right)\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\boldsymbol{d}\left(\boldsymbol{D}\right)\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\right\} \\
&= \frac{1}{2}\left[\mathrm{vec}(\boldsymbol{d}\left(\boldsymbol{D}\right))\right]^{\intercal}\sum_{i=1}^{N} w_i\left(\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\otimes\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\right)\left[\mathrm{vec}(\boldsymbol{d}\left(\boldsymbol{D}\right))\right] \\
&- \left[\mathrm{vec}(\boldsymbol{d}\left(\boldsymbol{D}\right))\right]^{\intercal}\sum_{i=1}^{N} w_i\left(\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\otimes\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\right)\left[\mathrm{vec}(\boldsymbol{d}\left(\boldsymbol{D}\right))\right] \\
&= \left[\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right)\right]^{\intercal}\left\{\frac{1}{2}\boldsymbol{D}_q^{\intercal}\sum_{i=1}^{N} w_i\left(\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\otimes\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\right)\boldsymbol{D}_q\right. \\
&\left.-\boldsymbol{D}_q^{\intercal}\sum_{i=1}^{N} w_i\left(\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\otimes\left(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right)\right)\boldsymbol{D}_q\right\}\left[\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right)\right],
\end{aligned}
\tag{C.43}
$$

so that the $(q(q+1)/2)\times(q(q+1)/2)$ matrix $\boldsymbol{B}$ is given by the expression within curly brackets in (C.43). This matrix is symmetrical so that $(1/2)\left(\boldsymbol{B}+\boldsymbol{B}^{\intercal}\right) = \boldsymbol{B}$. Thus $\boldsymbol{H}_{\mathrm{v}(\boldsymbol{D})}\left(L(\mathrm{v}(\boldsymbol{D}))\right)$ is

$$\begin{aligned}\boldsymbol{H}_{\mathrm{v}(\boldsymbol{D})}\left(L(\mathrm{v}(\boldsymbol{D}))\right) &= \frac{1}{2}\widetilde{\boldsymbol{D}}_q^{\intercal}\sum_{i=1}^{N} w_i\left((\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\right)\widetilde{\boldsymbol{D}}_q \\ &- \widetilde{\boldsymbol{D}}_q^{\intercal}\sum_{i=1}^{N} w_i\left((\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\right)\widetilde{\boldsymbol{D}}_q\end{aligned} \tag{C.44}$$

and so

$$\begin{aligned}-\boldsymbol{E}\left[\boldsymbol{H}_{\mathrm{v}(\boldsymbol{D})}\left(L(\mathrm{v}(\boldsymbol{D}))\right)\right] &= -\frac{1}{2}\widetilde{\boldsymbol{D}}_q^{\intercal}\sum_{i=1}^{N} w_i\left((\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\right)\widetilde{\boldsymbol{D}}_q \\ &\quad + \widetilde{\boldsymbol{D}}_q^{\intercal}\sum_{i=1}^{N} w_i\left((\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\mathrm{Var}[\boldsymbol{Y}_i]\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\right)\widetilde{\boldsymbol{D}}_q \\ &= -\frac{1}{2}\widetilde{\boldsymbol{D}}_q^{\intercal}\sum_{i=1}^{N} w_i\left((\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\right)\widetilde{\boldsymbol{D}}_q \\ &+ \widetilde{\boldsymbol{D}}_q^{\intercal}\sum_{i=1}^{N} w_i\left((\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\right)\widetilde{\boldsymbol{D}}_q \\ &= \frac{1}{2}\widetilde{\boldsymbol{D}}_q^{\intercal}\sum_{i=1}^{N} w_i\left((\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i)\right)\widetilde{\boldsymbol{D}}_q.\end{aligned} \tag{C.45}$$

$\boldsymbol{D}^2_{(\boldsymbol{\phi})(\boldsymbol{\psi})}\left(L(\boldsymbol{\psi})\right) = \boldsymbol{D}^2_{(\boldsymbol{\phi})(\mathrm{v}(\boldsymbol{D}))}\left(L(\mathrm{v}(\boldsymbol{D}))\right)$ for $\boldsymbol{\psi} = \mathrm{v}(\boldsymbol{D})$:

First we derive $\boldsymbol{D}_{\mathrm{v}(\boldsymbol{D})}\left(L(\mathrm{v}(\boldsymbol{D}))\right)$. Now $L(\mathrm{v}(\boldsymbol{D})) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^{q(q+1)/2}$, so that from the first identification table (Table B.1) we have that if $\boldsymbol{d}\left(L(\mathrm{v}(\boldsymbol{D}))\right) = \boldsymbol{a}^{\intercal}\boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right)$ for $\boldsymbol{a} \in \mathbb{R}^{q(q+1)/2}$, then $\boldsymbol{D}_{\mathrm{v}(\boldsymbol{D})}\left(L(\mathrm{v}(\boldsymbol{D}))\right) = \boldsymbol{a}^{\intercal}$. We have

$$\begin{aligned}\boldsymbol{d}\left(L(\mathrm{v}(\boldsymbol{D}))\right) &= -\frac{1}{2}\sum_{i=1}^{N} w_i \boldsymbol{d}\left(\log\left(|\boldsymbol{V}_i(\boldsymbol{\zeta})|\right)\right) - \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal} \boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\tilde{\boldsymbol{e}}_i \\ &= -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right] + \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{vec}\left[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right] \\ &= -\frac{1}{2}\sum_{i=1}^{N} w_i \left[\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right]^{\intercal}(\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i)\boldsymbol{d}\left(\mathrm{vec}(\boldsymbol{D})\right) \\ &\quad + \frac{1}{2}\sum_{i=1}^{N} w_i ((\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i) \otimes (\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i))\boldsymbol{d}\left(\mathrm{vec}(\boldsymbol{D})\right) \\ &= \Bigg\{ -\frac{1}{2}\sum_{i=1}^{N} w_i \left[\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right]^{\intercal}(\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i)\widetilde{\boldsymbol{D}}_q \\ &\quad + \frac{1}{2}\sum_{i=1}^{N} w_i ((\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i) \otimes (\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i))\widetilde{\boldsymbol{D}}_q \Bigg\} \boldsymbol{d}\left(\mathrm{v}(\boldsymbol{D})\right). \end{aligned} \tag{C.46}$$

So we see that

$$\begin{aligned}\boldsymbol{D}_{\mathrm{v}(\boldsymbol{D})}\left(L(\mathrm{v}(\boldsymbol{D}))\right) &= -\frac{1}{2}\sum_{i=1}^{N} w_i \left[\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\right]^{\intercal}(\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i)\widetilde{\boldsymbol{D}}_q \\ &\quad + \frac{1}{2}\sum_{i=1}^{N} w_i ((\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i) \otimes (\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i))\widetilde{\boldsymbol{D}}_q, \end{aligned} \tag{C.47}$$

which is a $1 \times q(q+1)/2$ vector as required. Letting $\boldsymbol{g}(\boldsymbol{\phi}) = \boldsymbol{D}_{\mathrm{v}(\boldsymbol{D})}\left(L(\mathrm{v}(\boldsymbol{D}))\right)^{\intercal}$ we have $\boldsymbol{g}(\boldsymbol{\phi}) : S \longrightarrow \mathbb{R}^{q(q+1)/2}$, $S \subseteq \mathbb{R}^{r}$. From the first identification table (Table B.1) we have that if $\boldsymbol{d}\left(\boldsymbol{g}(\boldsymbol{\phi})\right) = \boldsymbol{A}\boldsymbol{d}\boldsymbol{\phi}$ for a $q(q+1)/2 \times r$ matrix $\boldsymbol{A}$, then $\boldsymbol{D}^2_{(\boldsymbol{\phi})(\mathrm{v}(\boldsymbol{D}))}\left(L(\mathrm{v}(\boldsymbol{D}))\right) = \boldsymbol{A}$. We have

$$
\begin{aligned}
\boldsymbol{d}\left(\boldsymbol{g}(\boldsymbol{\phi})\right) = &-\frac{1}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal} \otimes \boldsymbol{Z}_i^{\intercal})\mathrm{vec}\big((\boldsymbol{d}\big(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\big)\big)\\
&+\frac{1}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}\left\{\big(\boldsymbol{Z}_i^{\intercal}\boldsymbol{d}\big(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\big)\tilde{\boldsymbol{e}}_i\big) \otimes \big(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\big)\right.\\
&\left.+\big(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\big) \otimes \big(\boldsymbol{Z}_i^{\intercal}\boldsymbol{d}\big(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\big)\tilde{\boldsymbol{e}}_i\big)\right\}\\
=&\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal} \otimes \boldsymbol{Z}_i^{\intercal})\mathrm{vec}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}(\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right]\\
-&\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}\left\{\big(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\big) \otimes \big(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\big)\right.\\
+&\left.\big(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\big) \otimes \big(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\big)\right\}\\
=&\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal} \otimes \boldsymbol{Z}_i^{\intercal})(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \otimes \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right]\\
-&\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}\left\{\mathrm{vec}\left[\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right]\right.\\
+&\left.\mathrm{vec}\left[\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{Z}_i\right]\right\}\\
=&\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \otimes \boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right]\\
-&\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}\left\{(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right]\right.\\
+&\left.(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\mathrm{vec}\left[\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\right]\right\}\\
=&\Bigg\{\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \otimes \boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\\
-&\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\\
-&\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}) \otimes (\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\Bigg\}\boldsymbol{d}\phi.
\end{aligned}
\tag{C.48}
$$

So we see that the $q(q+1)/2 \times r$ matrix $\boldsymbol{D}^2_{(\boldsymbol{\phi})(\mathrm{v}(\boldsymbol{D}))}\left(L(\mathrm{v}(\boldsymbol{D}))\right)$ is given by the expression within parentheses in (C.48). Now $-\boldsymbol{E}[\boldsymbol{D}^2_{(\boldsymbol{\phi})(\mathrm{v}(\boldsymbol{D}))}\left(L(\mathrm{v}(\boldsymbol{D}))\right)]$ involves calculating only $\boldsymbol{E}[\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}] = \mathrm{Var}[\boldsymbol{Y}_i] = \boldsymbol{V}_i(\boldsymbol{\zeta})$, so that $\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{E}[\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}]\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} = \boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}$. Thus

$$\begin{aligned}-\boldsymbol{E}[\boldsymbol{D}^2_{(\boldsymbol{\phi})(\mathrm{v}(\boldsymbol{D}))}\left(L(\mathrm{v}(\boldsymbol{D}))\right)] &= -\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i\widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\\ &\quad+\sigma^2\sum_{i=1}^{N} w_i\widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\\ &= \frac{\sigma^2}{2}\sum_{i=1}^{N} w_i\widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right).\end{aligned} \tag{C.49}$$

C.1.5 Line 4 of the information matrix

$\boldsymbol{H}_{\boldsymbol{\phi}}\left(L(\boldsymbol{\phi})\right)$:

We will need a few preliminary results. Firstly for a matrix function $\boldsymbol{F}(\boldsymbol{x})$ where $\boldsymbol{F} : S \longrightarrow \mathbb{R}^{m\times p}$, $S \subseteq \mathbb{R}^{n}$, and where $\boldsymbol{F}$ is twice differentiable, we have that $\mathrm{vec}[\boldsymbol{d}^2\left(\boldsymbol{F}(\boldsymbol{x})\right)] = (\boldsymbol{I}_{mp}\otimes\boldsymbol{dx})^{\intercal}\boldsymbol{H}_{\boldsymbol{x}}\left(\boldsymbol{F}(\boldsymbol{x})\right)\boldsymbol{dx}$, where $\boldsymbol{H}_{\boldsymbol{x}}\left(\boldsymbol{F}(\boldsymbol{x})\right) = \{_c\boldsymbol{H}_{\boldsymbol{x}}\left(\boldsymbol{F}_{jk}(\boldsymbol{x})\right)\}_{j=1,\,k=1}^{m\quad p}$ is a $mnp \times n$ matrix where the $(j,k)^{th}$ element is the $n \times n$ Hessian matrix $\boldsymbol{H}_{\boldsymbol{x}}\left(\boldsymbol{F}_{jk}(\boldsymbol{x})\right)$. This comes directly from B.2. With a bit of simple algebra the $mp \times 1$ second differential vector $\mathrm{vec}[\boldsymbol{d}^2\left(\boldsymbol{F}(\boldsymbol{x})\right)]$ can be written

$$\begin{aligned}\mathrm{vec}[\boldsymbol{d}^2\left(\boldsymbol{F}(\boldsymbol{x})\right)] &= (\boldsymbol{I}_{mp}\otimes\boldsymbol{dx})^{\intercal}\,\boldsymbol{H}_{\boldsymbol{x}}\left(\boldsymbol{F}(\boldsymbol{x})\right)\boldsymbol{dx}\\ &= \left\{_c\,(\boldsymbol{dx})^{\intercal}\,\boldsymbol{H}_{\boldsymbol{x}}\left(\boldsymbol{F}_{jk}(\boldsymbol{x})\right)\boldsymbol{dx}\right\}_{j=1,\,k=1}^{m\quad p}.\end{aligned} \tag{C.50}$$

Now let $\boldsymbol{A} = \{_m\boldsymbol{a}_{jk}\}_{j=1,\,k=1}^{m\quad p}$ be a $m \times p$ matrix and let $\boldsymbol{a} \in \mathbb{R}^{mp}$ be a vector such that

$\boldsymbol{a} = \{{}_r\boldsymbol{a}_{jk}\}_{j=1,\,k=1}^{m\quad p}$. Then from (C.50) we have

$$\boldsymbol{a}^{\intercal}\left(\boldsymbol{I}_{mp}\otimes\boldsymbol{dx}\right)^{\intercal}\boldsymbol{H_x}\left(\boldsymbol{F}(\boldsymbol{x})\right)\boldsymbol{dx} = \left(\boldsymbol{dx}\right)^{\intercal}\left\{\sum_{j=1}^{m}\sum_{k=1}^{p}\boldsymbol{a}_{jk}\boldsymbol{H_x}\left(\boldsymbol{F}_{jk}(\boldsymbol{x})\right)\right\}\boldsymbol{dx}, \tag{C.51}$$

In particular from (C.51) we have

$$\begin{aligned}&\left[\operatorname{vec}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right]^{\intercal}\left(\boldsymbol{I}_{n_i^2}\otimes\boldsymbol{d\phi}\right)^{\intercal}\boldsymbol{H_\phi}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{d\phi} = \\ &\qquad\qquad(\boldsymbol{d\phi})^{\intercal}\left\{\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1})_{jk}(\boldsymbol{H_\phi}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right))_{jk}\right\}\boldsymbol{d\phi},\end{aligned} \tag{C.52}$$

and

$$\begin{aligned}&\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\left(\boldsymbol{I}_{n_i^2}\otimes\boldsymbol{d\phi}\right)^{\intercal}\boldsymbol{H_\phi}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{d\phi} = \\ &\qquad(\boldsymbol{d\phi})^{\intercal}\left\{\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\sum_{t=1}^{n_i}\sum_{s=1}^{n_i}\tilde{\boldsymbol{e}}_{it}\boldsymbol{V}_{itj}(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_{is}\boldsymbol{V}_{isk}(\boldsymbol{\zeta})^{-1}(\boldsymbol{H_\phi}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right))_{jk}\right\}\boldsymbol{d\phi}.\end{aligned} \tag{C.53}$$

The last result we need is that when $\boldsymbol{V}_i(\boldsymbol{\zeta})$ is viewed as a function of $\boldsymbol{\phi}$ that

$$\begin{aligned}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) &= -2\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} + \\ &\qquad\qquad\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}.\end{aligned} \tag{C.54}$$

Now $L(\boldsymbol{\phi}) : S \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^{r}$, so that from the second identification table (Table B.2) we have that if $\boldsymbol{d}^2\left(L(\boldsymbol{\phi})\right) = \{\boldsymbol{d}\left(\boldsymbol{\phi}\right)\}^{\intercal}\boldsymbol{B}\{\boldsymbol{d}\left(\boldsymbol{\phi}\right)\}$ for a $r \times r$ matrix $\boldsymbol{B}$ then $\boldsymbol{H_\phi}\left(L(\boldsymbol{\phi})\right) = (1/2)\left(\boldsymbol{B} + \boldsymbol{B}^{\intercal}\right)$. We have

$$
\begin{aligned}
\boldsymbol{d}\left(L(\boldsymbol{\phi})\right) &= -\frac{1}{2}\sum_{i=1}^{N} w_i \boldsymbol{d}\left(\log|\boldsymbol{V}_i(\boldsymbol{\zeta})|\right) - \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal} \boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) \tilde{\boldsymbol{e}}_i \\
&= -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)] + \frac{1}{2}\sum_{i=1}^{N} w_i \tilde{\boldsymbol{e}}_i^{\intercal} \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i,
\end{aligned}
\tag{C.55}
$$

so that

$$
\begin{aligned}
\boldsymbol{d}^2\left(L(\boldsymbol{\phi})\right) &= -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right) + \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right] \\
&\quad + \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\tilde{\boldsymbol{e}}_i\right] \\
&= -\frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[-\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right) + \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right] \\
&\quad - \sum_{i=1}^{N} w_i \mathrm{tr}\left[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right] \\
&\quad + \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right] - \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right] \\
&\quad - \sum_{i=1}^{N} w_i \mathrm{tr}\left[\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right] \\
&\quad + \frac{1}{2}\sum_{i=1}^{N} w_i \mathrm{tr}\left[\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\right] \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \left[\mathrm{vec}\left(\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right)\right]^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\left[\mathrm{vec}\left(\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right)\right] \\
&\quad - \sum_{i=1}^{N} w_i \left[\mathrm{vec}\left(\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right)\right]^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right)\left[\mathrm{vec}\left(\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right)\right] \\
&\quad - \frac{1}{2}w_i\sum_{i=1}^{N}\left[\mathrm{vec}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right]^{\intercal}\mathrm{vec}\left(\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right) \\
&\quad + \frac{1}{2}\sum_{i=1}^{N} w_i \left(\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\otimes\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right)\mathrm{vec}\left(\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right). \qquad \text{(C.56)}
\end{aligned}
$$

Now when $\boldsymbol{V}_i(\boldsymbol{\zeta})$ is viewed as a function of $\boldsymbol{\phi}$, we have $\mathrm{vec}(\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)) = \sigma^2\mathrm{vec}(\boldsymbol{d}^2\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)) = \sigma^2(\boldsymbol{I}_{n_i^2}\otimes\boldsymbol{d\phi})^{\intercal}\boldsymbol{H}_{\boldsymbol{\phi}}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)\boldsymbol{d\phi}$. If we let $f(\boldsymbol{\phi})$ be the function defined by the last two terms in the right-hand side of C.56, then we have

$$
\begin{aligned}
f(\boldsymbol{\phi}) &= -\frac{1}{2}\sum_{i=1}^{N} w_i \Big\{ \left[\operatorname{vec}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right]^{\intercal} \operatorname{vec}\left(\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right) \\
&\quad - w_i \left(\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) \otimes \left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right) \operatorname{vec}\left(\boldsymbol{d}^2\left(\boldsymbol{V}_i(\boldsymbol{\zeta})\right)\right) \Big\} \\
&= -\frac{\sigma^2}{2}\sum_{i=1}^{N} w_i \Big\{ \left[\operatorname{vec}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right]^{\intercal} \left(\boldsymbol{I}_{n_i^2} \otimes \boldsymbol{d\phi}\right)^{\intercal} \boldsymbol{H}_{\boldsymbol{\phi}}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right) \boldsymbol{d\phi} \\
&\quad - w_i \left(\left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) \otimes \left(\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right) \left(\boldsymbol{I}_{n_i^2} \otimes \boldsymbol{d\phi}\right)^{\intercal} \boldsymbol{H}_{\boldsymbol{\phi}}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right) \boldsymbol{d\phi} \Big\} \\
&= \left(\boldsymbol{d\phi}\right)^{\intercal} \left\{ \frac{\sigma^2}{2}\sum_{i=1}^{N} \left[-w_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} \left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{jk} \boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk}) \right.\right. \\
&\quad \left.\left. + w_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\sum_{t=1}^{n_i}\sum_{s=1}^{n_i} (\tilde{\boldsymbol{e}}_i)_t \left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{tj} (\tilde{\boldsymbol{e}}_i)_s \left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{sk} \boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk}) \right] \right\} \boldsymbol{d\phi},
\end{aligned}
\tag{C.57}
$$

where the last line follows from (C.52) and (C.53). If we let the function inside the curly brackets in (C.60) be $g(\boldsymbol{\phi})$, so that $f(\boldsymbol{\phi}) = (\boldsymbol{d\phi})^{\intercal} g(\boldsymbol{\phi})\boldsymbol{d\phi}$, then (C.56) becomes

$$
\begin{aligned}
\boldsymbol{d}^2\left(L(\boldsymbol{\phi})\right) &= \left(\boldsymbol{d\phi}\right)^{\intercal} \left\{ \frac{\boldsymbol{\sigma}^4}{2}\sum_{i=1}^{N} \Big[w_i \left(\boldsymbol{D}_{\phi}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\right)^{\intercal} \left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \otimes \boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right) \boldsymbol{D}_{\phi}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \right. \\
&\left. -2w_i \left(\boldsymbol{D}_{\phi}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\right)^{\intercal} \left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1} \otimes \left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right) \boldsymbol{D}_{\phi}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right) \Big] + g(\phi) \right\} \boldsymbol{d\phi}.
\end{aligned}
\tag{C.58}
$$

Now $g(\boldsymbol{\phi})$ is a symmetric $r \times r$ matrix where the symmetry follows from the fact that it is the sum of the $n_i \times n_i$ symmetric Hessian matrices $\boldsymbol{H}_{\boldsymbol{\phi}}\left(\boldsymbol{C}_i(\boldsymbol{\phi})\right)$. Each of the other two terms in (C.58) define $r \times r$ matrices which are symmetric, and so the whole expression in curly brackets in (C.58) is a symmetric $r \times r$ matrix. Thus this expression is $\boldsymbol{H}_{\boldsymbol{\phi}}\left(L(\boldsymbol{\phi})\right)$, so that

$$\boldsymbol{H}_{\boldsymbol{\phi}}\left(L(\boldsymbol{\phi})\right)=\frac{\boldsymbol{\sigma}^4}{2}\sum_{i=1}^{N}\Big[w_i\left(\boldsymbol{D}_\phi\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\right)^\intercal\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_\phi\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)$$

$$-2w_i\left(\boldsymbol{D}_\phi\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\right)^\intercal\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^\intercal\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right)\boldsymbol{D}_\phi\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\Big]+g(\phi),\tag{C.59}$$

where

$$g(\boldsymbol{\phi})=\frac{\sigma^2}{2}\sum_{i=1}^{N}\left[-w_i\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{jk}\boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk})\right.$$
$$\left.+w_i\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\sum_{t=1}^{n_i}\sum_{s=1}^{n_i}(\tilde{\boldsymbol{e}}_i)_t\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{tj}(\tilde{\boldsymbol{e}}_i)_s\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{sk}\boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk})\right].\tag{C.60}$$

Now taking expectations of (C.59) involves taking expectations of $g(\boldsymbol{\phi})$ which in turn only involves calculating $\boldsymbol{E}[\tilde{\boldsymbol{e}}_{it}\tilde{\boldsymbol{e}}_{is}]=(\boldsymbol{V}_i(\boldsymbol{\zeta}))_{ts}$. Thus we have

$$\begin{aligned}\boldsymbol{E}[g(\boldsymbol{\phi})]&=\frac{\sigma^2}{2}\sum_{i=1}^{N}\left(-w_i\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\boldsymbol{V}_{ijk}(\boldsymbol{\zeta})^{-1}\boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk})\right.\\&\left.+w_i\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\sum_{t=1}^{n_i}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{tj}\sum_{s=1}^{n_i}\left[(\boldsymbol{V}_i(\boldsymbol{\zeta}))_{ts}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{sk}\right]\boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk})\right)\\&=\frac{\sigma^2}{2}\sum_{i=1}^{N}\left(-w_i\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{jk}\boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk})\right.\\&\left.+w_i\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\sum_{t=1}^{n_i}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{tj}(\boldsymbol{V}_i(\boldsymbol{\zeta}))_{t.}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{.k}\boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk})\right)\\&=\frac{\sigma^2}{2}\sum_{i=1}^{N}\left(-w_i\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{jk}\boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk})\right.\\&\left.+w_i\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)_{jk}\boldsymbol{H}_{\boldsymbol{\phi}}((\boldsymbol{C}_i(\boldsymbol{\phi}))_{jk})\right)\\&=0\end{aligned}\tag{C.61}$$

Thus from (C.59) we have

$$\begin{aligned}
-\boldsymbol{E}\left[\boldsymbol{H}_{\boldsymbol{\phi}}\left(L(\boldsymbol{\phi})\right)\right] = -\frac{\boldsymbol{\sigma}^4}{2}\sum_{i=1}^{N}\Big[&w_i\left(\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\right)^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\\
&-2w_i\left(\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\right)^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\mathrm{Var}[\boldsymbol{Y}_i]\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\right)\times\\
&\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)]\\
= -\frac{\boldsymbol{\sigma}^4}{2}\sum_{i=1}^{N}\Big[&w_i\left(\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\right)^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\\
&-2w_i\left(\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\right)^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\Big]\\
= \frac{\boldsymbol{\sigma}^4}{2}\sum_{i=1}^{N}&w_i\left(\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right)\right)^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta})^{-1}\right)\boldsymbol{D}_{\phi}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}))\right).
\end{aligned} \tag{C.62}$$

C.2 Score vector and Hessian matrix for MLMMs

We want to obtain the score vector $\boldsymbol{D}_{\boldsymbol{\theta}}\left(L(\boldsymbol{\theta})\right)^{\intercal}$, and Hessian $\boldsymbol{H}_{\boldsymbol{\theta}}\left(L(\boldsymbol{\theta})\right)$ of the ordinary or incomplete log-likelihood given in (2.10) where for brevity we will write $L_i(\boldsymbol{\theta})$ for $L(\boldsymbol{y}_i|\boldsymbol{\theta})$. Then we have $\boldsymbol{d}\left(L(\boldsymbol{y}|\boldsymbol{\theta})\right) = \sum_{i=1}^{N}\boldsymbol{d}\left(L(\boldsymbol{y}_i|\boldsymbol{\theta})\right)$ which implies $\boldsymbol{D}_{\boldsymbol{\theta}}\left(L(\boldsymbol{\theta})\right) = \sum_{i=1}^{N}\boldsymbol{D}_{\boldsymbol{\theta}}\left(L_i(\boldsymbol{\theta})\right)$, and $\boldsymbol{H}_{\boldsymbol{\theta}}\left(L(\boldsymbol{\theta})\right) = \sum_{i=1}^{N}\boldsymbol{H}_{\boldsymbol{\theta}}\left(L_i(\boldsymbol{\theta})\right)$. All the MLMMs we are considering are the same for each unit, and so each Score vector and Hessian will have the same form. Thus to obtain the Score vector and Hessian for the sample we need only compute the differential of $L(\boldsymbol{y}_i|\boldsymbol{\theta})$ for an arbitrary i in order to identify $\boldsymbol{D}_{\boldsymbol{\theta}}\left(L_i(\boldsymbol{\theta})\right)^{\intercal}$, and $\boldsymbol{H}_{\boldsymbol{\theta}}\left(L_i(\boldsymbol{\theta})\right)$.

Since the component probabilities are constrained to sum to one, we will consider each $L_i(\boldsymbol{\theta})$ to be a function of $\tilde{\boldsymbol{\pi}} = (\boldsymbol{\pi}_1, ...\boldsymbol{\pi}_{G-1})^{\intercal}$, and obtain $\boldsymbol{\pi}_G$ as $\boldsymbol{\pi}_G = 1 - \sum_{l=1}^{G-1}\boldsymbol{\pi}_l$.

We will also partition $\boldsymbol{\theta}$ with respect to the component density parameters and mixing proportions as $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^\intercal, ..., \boldsymbol{\theta}_g^\intercal, \tilde{\boldsymbol{\pi}}^\intercal]^\intercal$. Let $L_i(\boldsymbol{\theta}_g)$ and $L_i(\tilde{\boldsymbol{\pi}})$ denote the log-likelihood function for unit i considered to be a function of only $\boldsymbol{\theta}_g$ and $\tilde{\boldsymbol{\pi}}$ respectively, with all other parameters considered fixed. Using this partition of $\boldsymbol{\theta}$ we can write the Score vector and Hessian in partitioned form as

$$\boldsymbol{D}_{\boldsymbol{\theta}} \left(L_i(\boldsymbol{\theta})\right)^\intercal = \left[\boldsymbol{D}_{\boldsymbol{\theta}_1} \left(L_i(\boldsymbol{\theta}_1)\right), ..., \boldsymbol{D}_{\boldsymbol{\theta}_G} \left(L_i(\boldsymbol{\theta}_G)\right), \boldsymbol{D}_{\tilde{\boldsymbol{\pi}}} \left(L_i(\tilde{\boldsymbol{\pi}})\right)\right]^\intercal, \tag{C.63}$$

and

$$\boldsymbol{H}_{\boldsymbol{\theta}} \left(L_i(\boldsymbol{\theta})\right) = \begin{bmatrix} \boldsymbol{H}_{\boldsymbol{\theta}_1} \left(L_i(\boldsymbol{\theta}_1)\right) & \boldsymbol{D}^2_{(\boldsymbol{\theta}_2)(\boldsymbol{\theta}_1)} \left(L_i(\boldsymbol{\theta}_1)\right) & \cdots & \boldsymbol{D}^2_{(\boldsymbol{\theta}_G)(\boldsymbol{\theta}_1)} \left(L_i(\boldsymbol{\theta}_1)\right) & \boldsymbol{D}^2_{(\tilde{\boldsymbol{\pi}})(\boldsymbol{\theta}_1)} \left(L_i(\boldsymbol{\theta}_1)\right) \\ & \boldsymbol{H}_{\boldsymbol{\theta}_2} \left(L_i(\boldsymbol{\theta}_2)\right) & \cdots & \boldsymbol{D}^2_{(\boldsymbol{\theta}_G)(\boldsymbol{\theta}_2)} \left(L_i(\boldsymbol{\theta}_2)\right) & \boldsymbol{D}^2_{(\tilde{\boldsymbol{\pi}})(\boldsymbol{\theta}_2)} \left(L_i(\boldsymbol{\theta}_2)\right) \\ & & \ddots & \vdots & \vdots \\ & \text{symm} & & \boldsymbol{H}_{\boldsymbol{\theta}_G} \left(L_i(\boldsymbol{\theta}_G)\right) & \boldsymbol{D}^2_{(\tilde{\boldsymbol{\pi}})(\boldsymbol{\theta}_G)} \left(L_i(\boldsymbol{\theta}_G)\right) \\ & & & & \boldsymbol{H}_{\tilde{\boldsymbol{\pi}}} \left(L_i(\tilde{\boldsymbol{\pi}})\right) \end{bmatrix}. \tag{C.64}$$

For each unit, the way in which $L(\boldsymbol{y}_i|\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}_g$ and $\tilde{\boldsymbol{\pi}}$ is the same. Specifically, since there are no parameters that are shared across components, $L(\boldsymbol{y}_i|\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}_g$ only through the component density $f_{ig}(\boldsymbol{y}|\boldsymbol{\theta}_g)$. Thus the form of $\boldsymbol{D}_{\boldsymbol{\theta}_j} \left(L_i(\boldsymbol{\theta}_j)\right)$ and $\boldsymbol{H}_{\boldsymbol{\theta}_j} \left(L_i(\boldsymbol{\theta}_j)\right)$ will be the same for all $i = 1, ..., N$, and $j = 1, ..., G$, and the form of

$\boldsymbol{D}^2_{(\boldsymbol{\theta}_j)(\boldsymbol{\theta}_k)}\left(L_i(\boldsymbol{\theta}_k)\right)$ will also be the same for all $i = 1, .., N$, and for all $j, k = 1, ..., G$. The form of $\boldsymbol{D}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right)$ and $\boldsymbol{H}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right)$ will also be the same for all i. Thus we need only derive two score vectors, two Hessians and two cross-product matrices in order to find the score vector and Hessian of our sample given in (C.63) and (C.64), and in this section we will derive, in general form, all the necessary equations we need to compute these. In sections (C.2.1), and (C.2.2) we will use the general equations derived here to give in explicit form these equations for the classes of MLMMs we are concerned with.

To derive in general form the two score vectors, two Hessians and two cross-product matrices we require, we will consider $\boldsymbol{\theta}_g$ to be a partitioned vector with four components $\boldsymbol{\theta}_g = [\boldsymbol{\beta}_g^{\intercal}, \sigma_g^2, \boldsymbol{\psi}_g^{\intercal}, \boldsymbol{\phi}_g^{\intercal}]^{\intercal}$, where $\boldsymbol{\psi}_g = \mathrm{v}(\boldsymbol{D}_g)$, and we will write $\boldsymbol{\theta}_g^s$ for $s = 1, 2, 3, 4$, to index these components in this order. We will also write $T_s \subseteq \mathbb{R}^{n_s}$ for the domain sets of $\boldsymbol{\theta}_g^s$ where n_s is the number of parameters in $\boldsymbol{\theta}_g^s$. Thus the derivative vector of L_i with respect to $\boldsymbol{\theta}_g$ is

$$\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i(\boldsymbol{\theta}_g)\right) = \Big[\boldsymbol{D}_{\boldsymbol{\beta}_g}\left(L_i(\boldsymbol{\beta}_g)\right), \boldsymbol{D}_{\sigma_g^2}\left(L_i(\sigma_g^2)\right), \boldsymbol{D}_{\boldsymbol{\psi}_g}\left(L_i(\boldsymbol{\psi}_g)\right), \boldsymbol{D}_{\boldsymbol{\phi}_g}\left(L_i(\boldsymbol{\phi}_g)\right)\Big], \quad \text{(C.65)}$$

where the dimensions of the components are $(1 \times p)$,(1×1),$(1 \times q(q+1)/2)$, and $(1 \times r)$ respectively. The Hessian of L_i with respect to $\boldsymbol{\theta}_g$ is

$\boldsymbol{H}_{\boldsymbol{\theta}_g}\left(L_i(\boldsymbol{\theta}_g)\right) =$

$$\begin{bmatrix} \boldsymbol{H}_{\boldsymbol{\beta}_g}\left(L_i(\boldsymbol{\beta}_g)\right) & \boldsymbol{D}^2_{\left(\sigma_g^2\right)(\boldsymbol{\beta}_g)}\left(L_i(\boldsymbol{\beta}_g)\right) & \boldsymbol{D}^2_{(\boldsymbol{\psi}_g)(\boldsymbol{\beta}_g)}\left(L_i(\boldsymbol{\beta}_g)\right) & \boldsymbol{D}^2_{(\boldsymbol{\phi}_g)(\boldsymbol{\beta}_g)}\left(L_i(\boldsymbol{\beta}_g)\right) \\ & \boldsymbol{H}_{\sigma_g^2}\left(L_i(\sigma_g^2)\right) & \boldsymbol{D}^2_{(\boldsymbol{\psi}_g)\left(\sigma_g^2\right)}\left(L_i(\sigma_g^2)\right) & \boldsymbol{D}^2_{(\boldsymbol{\phi}_g)\left(\sigma_g^2\right)}\left(L_i(\sigma_g^2)\right) \\ & & \boldsymbol{H}_{\boldsymbol{\psi}_g}\left(L_i(\boldsymbol{\psi}_g)\right) & \boldsymbol{D}^2_{(\boldsymbol{\phi}_g)(\boldsymbol{\psi}_g)}\left(L_i(\boldsymbol{\psi}_g)\right) \\ & \text{symm} & & \boldsymbol{H}_{\boldsymbol{\phi}_g}\left(L_i(\boldsymbol{\phi}_g)\right) \end{bmatrix}. \tag{C.66}$$

Using the approach taken by Boldea and Magnus (2009), we introduce the following notation

$$\upsilon_{ig} = \boldsymbol{\pi}_g f_g(\boldsymbol{y}_i|\lambda^g, \boldsymbol{\theta}_g), \tag{C.67}$$

and

$$\alpha_{ig} = \frac{\upsilon_{ig}}{\sum_{k=1}^{G} \upsilon_{ik}}, \tag{C.68}$$

for $g = 1, ..., G$, and $i = 1, ..., N$. We will also use $f_{ig} = f_g(\boldsymbol{y}_i|\lambda^g, \boldsymbol{\theta}_g)$ for brevity. Using this notation we have

$$\begin{aligned}
\boldsymbol{d}[L(\boldsymbol{\theta}|\boldsymbol{y}_i)] &= \frac{\boldsymbol{d}[f(\boldsymbol{y}_i|\boldsymbol{\theta})]}{f(\boldsymbol{y}_i|\boldsymbol{\theta})} \\
&= \frac{\sum_{j=1}^{G} \boldsymbol{d}[\boldsymbol{\pi}_j f_{ij}]}{\sum_{k=1}^{G} \boldsymbol{\pi}_k f_{ik}} \\
&= \sum_{j=1}^{G} \left(\frac{\boldsymbol{\pi}_j f_{ij}}{\sum_{k=1}^{G} \boldsymbol{\pi}_k f_{ik}} \frac{\boldsymbol{d}[\boldsymbol{\pi}_j f_{ij}]}{\boldsymbol{\pi}_j f_{ij}} \right) \\
&= \sum_{j=1}^{G} \left(\frac{\boldsymbol{\pi}_j f_{ij}}{\sum_{k=1}^{G} \boldsymbol{\pi}_k f_{ik}} \boldsymbol{d}[\log(\boldsymbol{\pi}_j f_{ij})] \right) \\
&= \sum_{j=1}^{G} \frac{\upsilon_{ij}}{\sum_{k=1}^{G} \upsilon_{ik}} \boldsymbol{d}[\log \upsilon_{ij}] \\
&= \sum_{j=1}^{G} \alpha_{ij} \boldsymbol{d}[\log \upsilon_{ij}], \qquad \text{(C.69)}
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{d}^2 \left(L(\boldsymbol{\theta}|\boldsymbol{y}_i) \right) &= \boldsymbol{d} \left(\frac{\boldsymbol{d} \left(f(\boldsymbol{y}_i|\boldsymbol{\theta}) \right)}{f(\boldsymbol{y}_i|\boldsymbol{\theta})} \right) \\
&= \frac{\boldsymbol{d}^2 \left(f(\boldsymbol{y}_i|\boldsymbol{\theta}) \right)}{f(\boldsymbol{y}_i|\boldsymbol{\theta})} - \left(\frac{\boldsymbol{d} \left(f(\boldsymbol{y}_i|\boldsymbol{\theta}) \right)}{f(\boldsymbol{y}_i|\boldsymbol{\theta})} \right)^2 . \qquad \text{(C.70)}
\end{aligned}$$

Now

$$\begin{aligned}
\frac{\boldsymbol{d}^2 \left(f(\boldsymbol{y}_i|\boldsymbol{\theta}) \right)}{f(\boldsymbol{y}_i|\boldsymbol{\theta})} &= \frac{\boldsymbol{d} \left(\sum_{j=1}^{G} \boldsymbol{d} \left(\boldsymbol{\pi}_j f_{ij} \right) \right)}{f(\boldsymbol{y}_i|\boldsymbol{\theta})} \\
&= \frac{\sum_{j=1}^{G} \boldsymbol{d}^2 \left(\boldsymbol{\pi}_j f_{ij} \right)}{\sum_{k=1}^{G} \boldsymbol{\pi}_k f_{ik}} \\
&= \sum_{j=1}^{G} \left(\frac{\boldsymbol{\pi}_j f_{ij}}{\sum_{k=1}^{G} \boldsymbol{\pi}_k f_{ik}} \frac{\boldsymbol{d}^2 \left(\boldsymbol{\pi}_j f_{ij} \right)}{\boldsymbol{\pi}_j f_{ij}} \right) \\
&= \sum_{j=1}^{G} \alpha_{ij} \frac{\boldsymbol{d}^2 \upsilon_{ij}}{\upsilon_{ij}} \qquad \text{(C.71)}
\end{aligned}$$

So using (C.69) and (C.71) we have that (C.70) becomes

$$\boldsymbol{d}^2 \left(L(\boldsymbol{\theta}|\boldsymbol{y}_i) \right) = \sum_{j=1}^{G} \alpha_{ij} \frac{\boldsymbol{d}^2 \upsilon_{ij}}{\upsilon_{ij}} - \left(\sum_{j=1}^{G} \alpha_{ij} \frac{\boldsymbol{d} \upsilon_{ij}}{\upsilon_{ij}} \right)^2 . \qquad \text{(C.72)}$$

Now for any g we have

$$\begin{aligned}
\boldsymbol{d}^2\left(\log(\boldsymbol{\pi}_g f_{ig})\right) + \left[\boldsymbol{d}\left(\log(\boldsymbol{\pi}_g f_{ig})\right)\right]^2 &= \boldsymbol{d}\left[\frac{\boldsymbol{d}\left(\boldsymbol{\pi}_g f_{ig}\right)}{\boldsymbol{\pi}_g f_{ig}}\right] + \left(\frac{\boldsymbol{d}\left(\boldsymbol{\pi}_g f_{ig}\right)}{\boldsymbol{\pi}_g f_{ig}}\right)^2 \\
&= \frac{\boldsymbol{d}^2\left(\boldsymbol{\pi}_g f_{ig}\right)}{\boldsymbol{\pi}_g f_{ig}} + \left(\frac{\boldsymbol{d}\left(\boldsymbol{\pi}_g f_{ig}\right) - \boldsymbol{d}\left(\boldsymbol{\pi}_g f_{ig}\right)}{\boldsymbol{\pi}_g f_{ig}}\right)^2 \\
&= \frac{\boldsymbol{d}^2\left(\boldsymbol{\pi}_g f_{ig}\right)}{\boldsymbol{\pi}_g f_{ig}} \\
&= \frac{\boldsymbol{d}^2 \upsilon_{ig}}{\upsilon_{ig}}, \qquad \text{(C.73)}
\end{aligned}$$

and so using this in (C.72) we have

$$\boldsymbol{d}^2\left(L(\boldsymbol{\theta}|\boldsymbol{y}_i)\right) = \sum_{j=1}^{G} \alpha_{ij}\left[\boldsymbol{d}^2\left(\log \upsilon_{ij}\right) + \left(\boldsymbol{d}\left(\log \upsilon_{ij}\right)\right)^2\right] - \left[\sum_{j=1}^{G} \alpha_{ij} \boldsymbol{d}\left(\log \upsilon_{ij}\right)\right]^2. \qquad \text{(C.74)}$$

Equations (C.69) and (C.74) are the same as those given in Boldea and Magnus (2009) in equations (A.1) and (A.2) respectively. From this point onwards the results of our derivations are different since Boldea and Magnus use a mixture of normal densities with covariance matrices that do not depend on any data, and are thus specified directly (as opposed to our approach which is to use random effects to induce a covariance structure). Furthermore they specify means that do not include a regression component which also do not depend on the data. For reasons previously described, we will consider $L_i(\boldsymbol{\theta})$ to be a function of the s^{th} component of $\boldsymbol{\theta}$ for an arbitrary $g \in \{1, ..., G\}$. We now show that determining the Score vector and Hessian of $L_i(\boldsymbol{\theta}_g^s)$ reduces to determining the Score vector and Hessian of $\log f_{ig}$ (considered as a function of only $\boldsymbol{\theta}_g^s$).

Now since $\boldsymbol{d}[L_i(\boldsymbol{\theta}_g^s)] = \boldsymbol{D}_{\boldsymbol{\theta}_g^s}\left(L_i(\boldsymbol{\theta}_g^s)\right) d\boldsymbol{\theta}_g^s$, and $\boldsymbol{d}^2[L_i(\boldsymbol{\theta}_g^s)] = (\boldsymbol{d}\boldsymbol{\theta}_g^s)^{\intercal} \boldsymbol{H}_{\boldsymbol{\theta}_g^s}\left(L_i(\boldsymbol{\theta}_g^s)\right) \boldsymbol{d}\boldsymbol{\theta}_g^s$, and under the assumption that $L_i(\boldsymbol{\theta})$ is a function of only $\boldsymbol{\theta}_g^s$, from (C.69) we see that

$$\begin{aligned}\boldsymbol{d}[L_i(\boldsymbol{\theta}_g^s)] &= \boldsymbol{D}_{\boldsymbol{\theta}_g^s}\left(L_i(\boldsymbol{\theta}_g^s)\right)\boldsymbol{d\theta}_g^s \\ &= \alpha_{ig}\boldsymbol{d}[\log \upsilon_{ig}] \\ &= \alpha_{ig}\boldsymbol{d}[\log(\boldsymbol{\pi}_g f_{ig})] \\ &= \alpha_{ig}\boldsymbol{d}[\log\boldsymbol{\pi}_g] + \alpha_{ig}\boldsymbol{d}[\log f_{ig}] \\ &= \alpha_{ig}\boldsymbol{d}[\log f_{ig}] \\ &= \alpha_{ig}\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(\log f_{ig})\boldsymbol{d\theta}_g^s \\ &= \alpha_{ig}\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s))\boldsymbol{d\theta}_g^s, \end{aligned} \tag{C.75}$$

where $L_i^1(\boldsymbol{\theta}_g^s) = \log f_{ig}(\boldsymbol{y}|\lambda^g, \boldsymbol{\theta}_g^s)$ denotes we are considering $\log f_{ig} = \log f_{ig}(\boldsymbol{y}|\lambda^g, \boldsymbol{\theta}_g)$ to be a function of only the s^{th} component of $\boldsymbol{\theta}_g$. The superscript "1" simply denotes that $\log f_{ig}$ is just the log likelihood for the i^{th} unit of a 1-component MLMM (conditional on that unit belonging to component g). Thus (C.75) implies

$$\boldsymbol{D}_{\boldsymbol{\theta}_g^s}\left(L_i(\boldsymbol{\theta}_g^s)\right) = \alpha_{ig}\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s)). \tag{C.76}$$

We note that (C.76) has the same form as $(\boldsymbol{q}_i^g)^\intercal$ given in Theorem 1 of Boldea and Magnus (swapping the indexes t and i of Boldea and Magnus to i and g respectively to match the notation we use here), although their score vector is computed with respect to the whole of $\boldsymbol{\theta}_g$. Computing the second differential, and using (C.75) and (C.74), we have

$$\begin{aligned}\boldsymbol{d}^2[L_i(\boldsymbol{\theta}_g^s)] &= (\boldsymbol{d\theta}_g^s)^\top \boldsymbol{H}_{\boldsymbol{\theta}_g^s}\left(L_i(\boldsymbol{\theta}_g^s)\right)\boldsymbol{d\theta}_g^s \\ &= \alpha_{ig}\left[\boldsymbol{d}^2\log\upsilon_{ig} + (\boldsymbol{d}\log\upsilon_{ig})^2\right] - [\alpha_{ig}\boldsymbol{d}\log\upsilon_{ig}]^2 \\ &= \alpha_{ig}\boldsymbol{d}^2\log f_{ig} + \alpha_{ig}(1-\alpha_{ig})(\boldsymbol{d}\log f_{ig})^2 \\ &= \alpha_{ig}(\boldsymbol{d\theta}_g^s)^\top\boldsymbol{H}_{\boldsymbol{\theta}_g^s}(\log f_{ig})\boldsymbol{d\theta}_g^s + \alpha_{ig}(1-\alpha_{ig})\left(\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(\log f_{ig})\boldsymbol{d\theta}_g^s\right)^2 \\ &= \alpha_{ig}(\boldsymbol{d\theta}_g^s)^\top\boldsymbol{H}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s))\boldsymbol{d\theta}_g^s + \alpha_{ig}(1-\alpha_{ig})\left(\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s)\boldsymbol{d\theta}_g^s\right)^2.\end{aligned} \tag{C.77}$$

Using the fact that for any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same dimension we have $(\boldsymbol{x}^\top\boldsymbol{y})^2 = (\boldsymbol{x}^\top\boldsymbol{y})(\boldsymbol{x}^\top\boldsymbol{y}) = (\boldsymbol{y}^\top\boldsymbol{x})(\boldsymbol{x}^\top\boldsymbol{y})$, then (C.77) becomes

$$\begin{aligned}\boldsymbol{d}^2[L_i(\boldsymbol{\theta}_g^s)] &= (\boldsymbol{d\theta}_g^s)^\top \boldsymbol{H}_{\boldsymbol{\theta}_g^s}\left(L_i(\boldsymbol{\theta}_g^s)\right)\boldsymbol{d\theta}_g^s \\ &= \alpha_{ig}(\boldsymbol{d\theta}_g^s)^\top\boldsymbol{H}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s))\boldsymbol{d\theta}_g^s + \alpha_{ig}(1-\alpha_{ig})(\boldsymbol{d\theta}_g^s)^\top\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s))^\top\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s))\boldsymbol{d\theta}_g^s \\ &= (\boldsymbol{d\theta}_g^s)^\top\left\{\alpha_{ig}\boldsymbol{H}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s)) + \alpha_{ig}(1-\alpha_{ig})\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s))^\top\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s))\right\}\boldsymbol{d\theta}_g^s.\end{aligned} \tag{C.78}$$

Thus

$$\boldsymbol{H}_{\boldsymbol{\theta}_g^s}\left(L_i(\boldsymbol{\theta}_g^s)\right) = \alpha_{ig}\boldsymbol{H}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s)) + \alpha_{ig}(1-\alpha_{ig})\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s))^\top\boldsymbol{D}_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s)), \tag{C.79}$$

which is a symmetric $n_s \times n_s$ matrix as required. This is the same form as $\boldsymbol{Q}_i^{gg}$ given in Theorem 1 of Boldea and Magnus (this can be seen by noting that $-\boldsymbol{C}_{gi} = \boldsymbol{H}_{\boldsymbol{\theta}_g}\left(L_i(\boldsymbol{\theta}_g)\right)$) although their Hessian and score vector are computed with respect to the whole of $\boldsymbol{\theta}_g$. Again we swap the indexes t and i of Boldea and Magnus to i and g

respectively to match the notation we use here.

We now derive the cross products $\boldsymbol{D}^2_{(\boldsymbol{\theta}^s_g)(\boldsymbol{\theta}^t_g)}\left(L_i(\boldsymbol{\theta}^t_g)\right)$ in $(C.66)$, for $s, t \in \{1, 2, 3, 4\}$, $s \neq t$. Let $g(\boldsymbol{\theta}^s_g) = \boldsymbol{D}_{\boldsymbol{\theta}^t_g}\left(L_i(\boldsymbol{\theta}^t_g)\right)^{\intercal}$, and recalling that $\boldsymbol{\theta}^s_g \in T^s \subseteq \mathbb{R}^{n_s}$ for $s \in \{1, 2, 3, 4\}$, we have $g : T^s \to \mathbb{R}^{n_t}$. Then by (B.1) we have that if $\boldsymbol{d}\left(g(\boldsymbol{\theta}^s_g)\right) = \boldsymbol{A}\boldsymbol{d}\boldsymbol{\theta}^s_g$ for a $n_t \times n_s$ matrix $\boldsymbol{A}$, then $\boldsymbol{D}_{\boldsymbol{\theta}^s_g}\left(g(\boldsymbol{\theta}^s_g)\right) = \boldsymbol{D}^2_{(\boldsymbol{\theta}^s_g)(\boldsymbol{\theta}^t_g)}\left(L_i(\boldsymbol{\theta}^t_g)\right) = \boldsymbol{A}$. Now from (C.76) we have $g(\boldsymbol{\theta}^s_g) = \alpha_{ig} g^1(\boldsymbol{\theta}^s_g)$, where $g^1(\boldsymbol{\theta}^s_g) = \boldsymbol{D}_{\boldsymbol{\theta}^t_g}\left(L^1_i(\boldsymbol{\theta}^t_g)\right)^{\intercal}$. Noting that $\boldsymbol{d}\left(g^1(\boldsymbol{\theta}^s_g)\right) = \boldsymbol{D}^2_{(\boldsymbol{\theta}^s_g)(\boldsymbol{\theta}^t_g)}\left(L^1_i(\boldsymbol{\theta}^t_g)\right) d\boldsymbol{\theta}^s_g$ we have

$$\boldsymbol{d}\left(g(\boldsymbol{\theta}^s_g)\right) = \boldsymbol{d}\left(\alpha_{ig}\right) g^1(\boldsymbol{\theta}^s_g) + \alpha_{ig}\boldsymbol{D}^2_{(\boldsymbol{\theta}^s_g)(\boldsymbol{\theta}^t_g)}\left(L^1_i(\boldsymbol{\theta}^t_g)\right) d\boldsymbol{\theta}^s_g. \tag{C.80}$$

Now remembering that we are considering $g(\boldsymbol{\theta}^s_g)$, and hence α_{ig}, to be a function of only $\boldsymbol{\theta}^s_g$, we have

$$\begin{aligned}
\boldsymbol{d}\left(\alpha_{ig}\right) &= \frac{\boldsymbol{d}\left(\upsilon_{ig}\right)}{\sum_{j=1}^{G}\upsilon_{ij}} - \frac{\upsilon_{ig}\boldsymbol{d}\left(\upsilon_{ig}\right)}{\left(\sum_{j=1}^{G}\upsilon_{ij}\right)^2} \\
&= \frac{\upsilon_{ig}\boldsymbol{d}\left(\upsilon_{ig}\right)}{\upsilon_{ig}\sum_{j=1}^{G}\upsilon_{ij}} - \frac{\upsilon_{ig}^2\boldsymbol{d}\left(\upsilon_{ig}\right)}{\upsilon_{ig}\left(\sum_{j=1}^{G}\upsilon_{ij}\right)^2} \\
&= \alpha_{ig}\frac{\boldsymbol{d}\left(\upsilon_{ig}\right)}{\upsilon_{ig}} - \alpha_{ig}^2\frac{\boldsymbol{d}\left(\upsilon_{ig}\right)}{\upsilon_{ig}} \\
&= \alpha_{ig}(1-\alpha_{ig})(\boldsymbol{d}\left(\log\upsilon_{ig}\right)) \\
&= \alpha_{ig}(1-\alpha_{ig})(\boldsymbol{d}\left(L^1_i(\boldsymbol{\theta}^s_g)\right)) \\
&= \alpha_{ig}(1-\alpha_{ig})\boldsymbol{D}_{\boldsymbol{\theta}^s_g}\left(L^1_i(\boldsymbol{\theta}^s_g)\right)\boldsymbol{d}\boldsymbol{\theta}^s_g.
\end{aligned} \tag{C.81}$$

Thus (C.80) becomes

$$\begin{aligned}\boldsymbol{d}\left(g(\boldsymbol{\theta}_g^s)\right) &= \alpha_{ig}(1-\alpha_{ig})\left[\boldsymbol{D}_{\boldsymbol{\theta}_g^s}\left(L_i^1(\boldsymbol{\theta}_g^s)\right)\boldsymbol{d\theta}_g^s\right]\boldsymbol{D}_{\boldsymbol{\theta}_g^t}\left(L_i^1(\boldsymbol{\theta}_g^t)\right)^{\intercal} + \alpha_{ig}\boldsymbol{D}^2_{(\boldsymbol{\theta}_g^s)(\boldsymbol{\theta}_g^t)}\left(L_i^1(\boldsymbol{\theta}_g^t)\right)\boldsymbol{d\theta}_g^s \\ &= \alpha_{ig}(1-\alpha_{ig})\boldsymbol{D}_{\boldsymbol{\theta}_g^t}\left(L_i^1(\boldsymbol{\theta}_g^t)\right)^{\intercal}\left[\boldsymbol{D}_{\boldsymbol{\theta}_g^s}\left(L_i^1(\boldsymbol{\theta}_g^s)\right)\boldsymbol{d\theta}_g^s\right] + \alpha_{ig}\boldsymbol{D}^2_{(\boldsymbol{\theta}_g^s)(\boldsymbol{\theta}_g^t)}\left(L_i^1(\boldsymbol{\theta}_g^t)\right)\boldsymbol{d\theta}_g^s \\ &= \left\{\alpha_{ig}(1-\alpha_{ig})\boldsymbol{D}_{\boldsymbol{\theta}_g^t}\left(L_i^1(\boldsymbol{\theta}_g^t)\right)^{\intercal}\boldsymbol{D}_{\boldsymbol{\theta}_g^s}\left(L_i^1(\boldsymbol{\theta}_g^s)\right) + \alpha_{ig}\boldsymbol{D}^2_{(\boldsymbol{\theta}_g^s)(\boldsymbol{\theta}_g^t)}\left(L_i^1(\boldsymbol{\theta}_g^t)\right)\right\}\boldsymbol{d\theta}_g^s,\end{aligned} \tag{C.82}$$

and so

$$\boldsymbol{D}^2_{(\boldsymbol{\theta}_g^s)(\boldsymbol{\theta}_g^t)}\left(L_i(\boldsymbol{\theta}_g^t)\right) = \alpha_{ig}(1-\alpha_{ig})\boldsymbol{D}_{\boldsymbol{\theta}_g^t}\left(L_i^1(\boldsymbol{\theta}_g^t)\right)^{\intercal}\boldsymbol{D}_{\boldsymbol{\theta}_g^s}\left(L_i^1(\boldsymbol{\theta}_g^s)\right) + \alpha_{ig}\boldsymbol{D}^2_{(\boldsymbol{\theta}_g^s)(\boldsymbol{\theta}_g^t)}\left(L_i^1(\boldsymbol{\theta}_g^t)\right), \tag{C.83}$$

which is a $n_t \times n_s$ matrix as required. Equations (C.83) and (C.79) will allow us to calculate all the elements of (C.2) in general form.

We now derive the cross products $\boldsymbol{D}^2_{(\boldsymbol{\theta}_j)(\boldsymbol{\theta}_k)}\left(L_i(\boldsymbol{\theta}_k)\right)$ in $(C.64)$, for $j,k \in \{1,...,G\}$, $j \neq k$. Let $g(\boldsymbol{\theta}_j) = \boldsymbol{D}_{\boldsymbol{\theta}_k}\left(L_i(\boldsymbol{\theta}_k)\right)^{\intercal}$ so that we have $g: T \to \mathbb{R}^{n_\theta}$. Then by (B.1) we have that if $\boldsymbol{d}\left(g(\boldsymbol{\theta}_j)\right) = \boldsymbol{A}\boldsymbol{d\theta}_j$ for a $n_\theta \times n_\theta$ matrix $\boldsymbol{A}$, then $\boldsymbol{D}_{\boldsymbol{\theta}_j}\left(g(\boldsymbol{\theta}_j)\right) = \boldsymbol{D}^2_{(\boldsymbol{\theta}_j)(\boldsymbol{\theta}_k)}\left(L_i(\boldsymbol{\theta}_k)\right) = \boldsymbol{A}$. Now in the same way we derived (C.76), we have $\boldsymbol{D}_{\boldsymbol{\theta}_k}\left(L_i(\boldsymbol{\theta}_k)\right) = \alpha_{ik}\boldsymbol{D}_{\boldsymbol{\theta}_k}(L_i^1(\boldsymbol{\theta}_k))$, and so $g(\boldsymbol{\theta}_j) = \alpha_{ik}g^1(\boldsymbol{\theta}_j)$, where $g^1(\boldsymbol{\theta}_j) = \boldsymbol{D}_{\boldsymbol{\theta}_k}\left(L_i^1(\boldsymbol{\theta}_k)\right)^{\intercal}$. Now $g^1(\boldsymbol{\theta}_j)$ is in fact not a function of $\boldsymbol{\theta}_j$ at all, and so $\boldsymbol{d}\left(g^1(\boldsymbol{\theta}_j)\right) = 0$ when we consider $g^1(\boldsymbol{\theta}_j)$ to be a function of only $\boldsymbol{\theta}_j$. Thus

$$\boldsymbol{d}\left(g(\boldsymbol{\theta}_j)\right) = \boldsymbol{d}\left(\alpha_{ik}\right)g^1(\boldsymbol{\theta}_j). \tag{C.84}$$

Now remembering that we are considering $g(\boldsymbol{\theta}_j)$, and hence α_{ik}, to be a function of only $\boldsymbol{\theta}_j$, we have

$$\begin{aligned}
\boldsymbol{d}\left(\alpha_{ik}\right) &= -\frac{\upsilon_{ik}\boldsymbol{d}\left(\upsilon_{ij}\right)}{\left(\sum_{l=1}^{G}\upsilon_{il}\right)^2} \\
&= -\frac{\upsilon_{ij}\upsilon_{ik}\boldsymbol{d}\left(\upsilon_{ij}\right)}{\upsilon_{ij}\left(\sum_{l=1}^{G}\upsilon_{il}\right)^2} \\
&= -\alpha_{ik}\alpha_{ij}\frac{\boldsymbol{d}\left(\upsilon_{ij}\right)}{\upsilon_{ij}} \\
&= -\alpha_{ik}\alpha_{ij}\boldsymbol{d}\left(\log\upsilon_{ij}\right) \\
&= -\alpha_{ik}\alpha_{ij}\boldsymbol{d}\left(L_i^1(\boldsymbol{\theta}_j)\right) \\
&= -\alpha_{ik}\alpha_{ij}\boldsymbol{D}_{\boldsymbol{\theta}_j}\left(L_i^1(\boldsymbol{\theta}_j)\right)\boldsymbol{d\theta}_j. \qquad \text{(C.85)}
\end{aligned}$$

Thus from (C.84) we have

$$\begin{aligned}
\boldsymbol{d}\left(g(\boldsymbol{\theta}_j)\right) &= \left[-\alpha_{ik}\alpha_{ij}\boldsymbol{D}_{\boldsymbol{\theta}_j}\left(L_i^1(\boldsymbol{\theta}_j)\right)\boldsymbol{d\theta}_j\right]\boldsymbol{D}_{\boldsymbol{\theta}_k}\left(L_i^1(\boldsymbol{\theta}_k)\right)^{\intercal} \\
&= -\alpha_{ik}\alpha_{ij}\boldsymbol{D}_{\boldsymbol{\theta}_k}\left(L_i^1(\boldsymbol{\theta}_k)\right)^{\intercal}\boldsymbol{D}_{\boldsymbol{\theta}_j}\left(L_i^1(\boldsymbol{\theta}_j)\right)\boldsymbol{d\theta}_j, \qquad \text{(C.86)}
\end{aligned}$$

and so

$$\begin{aligned}
\boldsymbol{D}^2_{(\boldsymbol{\theta}_j)(\boldsymbol{\theta}_k)}\left(L_i(\boldsymbol{\theta}_k)\right) &= -\alpha_{ik}\alpha_{ij}\boldsymbol{D}_{\boldsymbol{\theta}_k}\left(L_i^1(\boldsymbol{\theta}_k)\right)^{\intercal}\boldsymbol{D}_{\boldsymbol{\theta}_j}\left(L_i^1(\boldsymbol{\theta}_j)\right) \\
&= -\boldsymbol{D}_{\boldsymbol{\theta}_k}\left(L_i(\boldsymbol{\theta}_k)\right)^{\intercal}\boldsymbol{D}_{\boldsymbol{\theta}_j}\left(L_i(\boldsymbol{\theta}_j)\right). \qquad \text{(C.87)}
\end{aligned}$$

The last line of (C.87) follows since

$$\begin{aligned}
\alpha_{ig}\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i^1(\boldsymbol{\theta}_g)\right) &= \left[\alpha_{ig}\boldsymbol{D}_{\boldsymbol{\beta}_g}\left(L_i^1(\boldsymbol{\beta}_g)\right), \alpha_{ig}\boldsymbol{D}_{\sigma_g^2}\left(L_i^1(\sigma_g^2)\right), \alpha_{ig}\boldsymbol{D}_{\boldsymbol{\psi}_g}\left(L_i^1(\boldsymbol{\psi}_g)\right), \alpha_{ig}\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(L_i^1(\boldsymbol{\phi}_g)\right)\right] \\
&= \left[\boldsymbol{D}_{\boldsymbol{\beta}_g}\left(L_i(\boldsymbol{\beta}_g)\right), \boldsymbol{D}_{\sigma_g^2}\left(L_i(\sigma_g^2)\right), \boldsymbol{D}_{\boldsymbol{\psi}_g}\left(L_i(\boldsymbol{\psi}_g)\right), \boldsymbol{D}_{\boldsymbol{\phi}_g}\left(L_i(\boldsymbol{\phi}_g)\right)\right] \\
&= \boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i(\boldsymbol{\theta}_g)\right) \qquad \text{(C.88)}
\end{aligned}$$

from (C.65) and (C.76). We note that (C.87) is a $n_\theta \times n_\theta$ matrix as required. This is the same form as $\boldsymbol{Q}_i^{kj}$ given in Theorem 1 of Boldea and Magnus, since to match with our notation we have $\boldsymbol{Q}_i^{kj} = \boldsymbol{D}^2_{(\boldsymbol{\theta}_j)(\boldsymbol{\theta}_k)}\left(L_i(\boldsymbol{\theta}_k)\right)$.

It is convenient to derive $\boldsymbol{D}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right)$ here rather than in section (C.2.1), and we shall do this componentwise, that is we will take the derivative of $L(\boldsymbol{y}_i|\boldsymbol{\theta})$ with respect to $\boldsymbol{\pi}_g$ for $g = 1, ..., G-1$. Then we will obtain the $1 \times (G-1)$ derivative vector of $L(\boldsymbol{y}_i|\boldsymbol{\theta})$ with respect to $\tilde{\boldsymbol{\pi}}$ as $\boldsymbol{D}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right) = \left(\boldsymbol{D}_{\boldsymbol{\pi}_1}\left(L_i(\boldsymbol{\pi}_1)\right), ..., \boldsymbol{D}_{\boldsymbol{\pi}_{G-1}}\left(L_i(\boldsymbol{\pi}_{G-1})\right)\right)$. Accordingly let $L(\boldsymbol{\pi}_g)$ be $L(\boldsymbol{y}_i|\boldsymbol{\theta})$ considered to be a function of only $\boldsymbol{\pi}_g$ for any $g = 1, ..., G-1$. Since $\log \upsilon_{ij}$ is a function of $\boldsymbol{\pi}_g$ when $j = g$ or $j = G$ (i.e. $\boldsymbol{\pi}_G = 1 - \sum_{l=1}^{G-1} \boldsymbol{\pi}_l$), from (C.69) we have

$$d[L(\boldsymbol{\pi}_g)] = \alpha_{ig}\boldsymbol{d}\left(\log \upsilon_{ig}\right) + \alpha_{iG}\boldsymbol{d}\left(\log \upsilon_{iG}\right). \tag{C.89}$$

Now $\boldsymbol{d}[\log \upsilon_{ij}] = \boldsymbol{D}_{\boldsymbol{\pi}_g}\left(\log \upsilon_{ij}(\boldsymbol{\pi}_g)\right) d\boldsymbol{\pi}_g$ is equal to $(1/\boldsymbol{\pi}_g)\boldsymbol{d\pi}_g$ when $j = g$, and $-(1/\boldsymbol{\pi}_G)d\boldsymbol{\pi}_g$ when $j = G$. Thus (C.89) becomes

$$d[L(\boldsymbol{\pi}_g)] = \left\{\alpha_{ig}\left(\frac{1}{\boldsymbol{\pi}_g}\right) - \alpha_{iG}\left(\frac{1}{\boldsymbol{\pi}_G}\right)\right\} d\boldsymbol{\pi}_g. \tag{C.90}$$

So the scalar derivative $\boldsymbol{D}_{\boldsymbol{\pi}_g}\left(L_i(\boldsymbol{\pi}_g)\right)$ is given by the expression in curly parentheses in (C.90), and so $\boldsymbol{D}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right) = \big(\alpha_{i1}(\boldsymbol{\pi}_1^{-1}) - \alpha_{iG}(\boldsymbol{\pi}_G^{-1}), ..., \alpha_{i(G-1)}(\boldsymbol{\pi}_{G-1}^{-1}) - \alpha_{iG}(\boldsymbol{\pi}_G^{-1})\big)$. Now introducing the following notation of Boldea and Magnus

$$\boldsymbol{a}_j = \begin{cases} (1/\boldsymbol{\pi}_j)\boldsymbol{e}_j & j = 1, ..., G-1, \\ -(1/\boldsymbol{\pi}_G)\mathbf{1}_{G-1} & j = G, \end{cases} \tag{C.91}$$

where $\boldsymbol{e}_j$ is the j^{th} column of the identity matrix I_{G-1}, and $\mathbf{1}_{G-1}$ is a $G-1$ dimensional vector of ones, we see that $\boldsymbol{D}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right)$ can be written

$$\boldsymbol{D}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right) = \sum_{j=1}^{G} \alpha_{ij}\boldsymbol{a}_j^{\intercal}. \tag{C.92}$$

We note that $C.92$ is the same as $(\boldsymbol{q}_i^{\pi})^{\intercal}$ given in Theorem 1 of Boldea and Magnus.

We now derive $\boldsymbol{D}^2_{(\tilde{\boldsymbol{\pi}})(\boldsymbol{\theta}_g)}\left(L_i(\boldsymbol{\theta}_g)\right)$ in (C.64). For convenience we will derive $\boldsymbol{D}^2_{(\boldsymbol{\theta}_g)(\tilde{\boldsymbol{\pi}})}\left(L_i(\tilde{\boldsymbol{\pi}})\right)$ and transpose it to get $\boldsymbol{D}^2_{(\tilde{\boldsymbol{\pi}})(\boldsymbol{\theta}_g)}\left(L_i(\boldsymbol{\theta}_g)\right)$. Let $g(\boldsymbol{\theta}_g) = \boldsymbol{D}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right)^{\intercal} = \sum_{j=1}^{G}\alpha_{ij}\boldsymbol{a}_j$, so that $g : T \to \mathbb{R}^{G-1}$. Then from (B.1), if $\boldsymbol{d}\left(g(\boldsymbol{\theta}_g)\right) = \boldsymbol{A}\boldsymbol{d}\boldsymbol{\theta}_g$ for a $(G-1)\times n_\theta$ matrix then $\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(g_i(\boldsymbol{\theta}_g)\right) = \boldsymbol{D}^2_{(\boldsymbol{\theta}_g)(\tilde{\boldsymbol{\pi}})}\left(L_i(\tilde{\boldsymbol{\pi}})\right) = \boldsymbol{A}$. From (C.91) we get

$$\boldsymbol{d}\left(g(\boldsymbol{\theta}_g)\right) = \sum\nolimits_{j=1}^{G-1}\boldsymbol{d}\left(\alpha_{ij}\right)\boldsymbol{\pi}_j^{-1}\boldsymbol{e}_j - \boldsymbol{d}\left(\alpha_{iG}\right)\boldsymbol{\pi}_G^{-1}\mathbf{1}_{G-1}. \tag{C.93}$$

Noting that the derivations in (C.81) apply with $\boldsymbol{\theta}_g$ instead of $\boldsymbol{\theta}_g^s$, then from (C.81) and (C.85) we have for $j \neq g$ that $\boldsymbol{d}\left(\alpha_{ij}\right) = -\alpha_{ij}\alpha_{ig}\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i^1(\boldsymbol{\theta}_g)\right)\boldsymbol{d}\boldsymbol{\theta}_g$, whilst for $j = g$ we have $\boldsymbol{d}\left(\alpha_{ij}\right) = \alpha_{ig}(1-\alpha_{ig})\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i^1(\boldsymbol{\theta}_g)\right)\boldsymbol{d}\boldsymbol{\theta}_g$. Thus for $g = G$ we have

$$\begin{aligned}
\boldsymbol{d}\left(g(\boldsymbol{\theta}_G)\right) &= -\sum_{j=1}^{G-1}\alpha_{ij}\alpha_{iG}\left[\boldsymbol{D}_{\boldsymbol{\theta}_G}\left(L_i^1(\boldsymbol{\theta}_G)\right)\boldsymbol{d}\boldsymbol{\theta}_G\right]\boldsymbol{\pi}_j^{-1}\boldsymbol{e}_j \\
&\quad - \alpha_{iG}(1-\alpha_{iG})\left[\boldsymbol{D}_{\boldsymbol{\theta}_G}\left(L_i^1(\boldsymbol{\theta}_G)\right)\boldsymbol{d}\boldsymbol{\theta}_G\right]\boldsymbol{\pi}_G^{-1}\mathbf{1}_{G-1} \\
&= -\sum_{j=1}^{G-1}\alpha_{ij}\alpha_{iG}\boldsymbol{\pi}_j^{-1}\boldsymbol{e}_j\left[\boldsymbol{D}_{\boldsymbol{\theta}_G}\left(L_i^1(\boldsymbol{\theta}_G)\right)\boldsymbol{d}\boldsymbol{\theta}_G\right] \\
&\quad - \alpha_{iG}(1-\alpha_{iG})\boldsymbol{\pi}_G^{-1}\mathbf{1}_{G-1}\left[\boldsymbol{D}_{\boldsymbol{\theta}_G}\left(L_i^1(\boldsymbol{\theta}_G)\right)\boldsymbol{d}\boldsymbol{\theta}_G\right] \\
&= \left\{\alpha_{iG}\left(-\boldsymbol{\pi}_G^{-1}\mathbf{1}_{G-1} - \left[\sum_{j=1}^{G-1}\alpha_{ij}\boldsymbol{\pi}_j^{-1}\boldsymbol{e}_j - \alpha_{iG}\boldsymbol{\pi}_G^{-1}\mathbf{1}_{G-1}\right]\right)\boldsymbol{D}_{\boldsymbol{\theta}_G}\left(L_i^1(\boldsymbol{\theta}_G)\right)\right\}\boldsymbol{d}\boldsymbol{\theta}_G \\
&= \left\{\alpha_{iG}\left(\boldsymbol{a}_G - \sum_{j=1}^{G}\alpha_{ij}\boldsymbol{a}_j\right)\boldsymbol{D}_{\boldsymbol{\theta}_G}\left(L_i^1(\boldsymbol{\theta}_G)\right)\right\}\boldsymbol{d}\boldsymbol{\theta}_G,
\end{aligned} \tag{C.94}$$

and for $g \neq G$ we have

$$\begin{aligned}
\boldsymbol{d}\left(g(\boldsymbol{\theta}_g)\right) &= -\sum_{j=1, j\neq g}^{G-1} \alpha_{ij}\alpha_{ig}\left[\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i^1(\boldsymbol{\theta}_g)\right)\boldsymbol{d\theta}_g\right]\boldsymbol{\pi}_j^{-1}\boldsymbol{e}_j \\
&\quad + \alpha_{ig}(1-\alpha_{ig})\left[\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i^1(\boldsymbol{\theta}_g)\right)\boldsymbol{d\theta}_g\right]\boldsymbol{\pi}_g^{-1}\boldsymbol{e}_g \\
&\quad + \alpha_{iG}\alpha_{ig}\left[\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i^1(\boldsymbol{\theta}_g)\right)\boldsymbol{d\theta}_g\right]\boldsymbol{\pi}_G^{-1}\boldsymbol{1}_{G-1} \\
&= \left\{\alpha_{ig}\left(\boldsymbol{\pi}_g^{-1}\boldsymbol{e}_g - \left[\sum_{j=1}^{G-1}\alpha_{ij}\boldsymbol{\pi}_j^{-1}\boldsymbol{e}_j - \alpha_{iG}\boldsymbol{\pi}_G^{-1}\boldsymbol{1}_{G-1}\right]\right)\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i^1(\boldsymbol{\theta}_g)\right)\right\}\boldsymbol{d\theta}_g \\
&= \left\{\alpha_{ig}\left(\boldsymbol{a}_g - \sum_{j=1}^{G}\alpha_{ij}\boldsymbol{a}_j\right)\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i^1(\boldsymbol{\theta}_g)\right)\right\}\boldsymbol{d\theta}_g. \qquad \text{(C.95)}
\end{aligned}$$

So for all $g \in \{1, ..., G\}$ we have

$$\boldsymbol{D}^2_{(\boldsymbol{\theta}_g)(\tilde{\boldsymbol{\pi}})}\left(L_i(\tilde{\boldsymbol{\pi}})\right) = \alpha_{ig}\left(\boldsymbol{a}_g - \sum_{j=1}^{G}\alpha_{ij}\boldsymbol{a}_j\right)\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i^1(\boldsymbol{\theta}_g)\right), \qquad \text{(C.96)}$$

which is a $(G-1)\times n_\theta$ matrix as required. Thus

$$\boldsymbol{D}^2_{(\tilde{\boldsymbol{\pi}})(\boldsymbol{\theta}_g)}\left(L_i(\boldsymbol{\theta}_g)\right) = \alpha_{ig}\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i^1(\boldsymbol{\theta}_g)\right)^{\intercal}\left(\boldsymbol{a}_g - \sum_{j=1}^{G}\alpha_{ij}\boldsymbol{a}_j\right)^{\intercal}, \qquad \text{(C.97)}$$

which is a $n_\theta \times (G-1)$ matrix. We note for all $i \in I_N$ that if the estimated posterior probabilities $\hat{\alpha}_{ij}$ for all $j = 1, ..., G$ are all precisely zero for all but one $g \in I_G$ (and so $\hat{\alpha}_{ig} = 1$), and regardless of whether this precise classification of units to components is correct, we have that $\boldsymbol{D}^2_{(\tilde{\boldsymbol{\pi}})(\hat{\boldsymbol{\theta}}_j)}\left(L_i(\hat{\boldsymbol{\theta}}_j)\right) = \boldsymbol{0}$ for all $j = 1, ..., G$. This shows when we use $\boldsymbol{H}_{\boldsymbol{\theta}}\left(L_i(\hat{\boldsymbol{\theta}})\right)$ in equation (C.64) to calculate $I_N(\hat{\boldsymbol{\theta}})$ that when we have precise classification of units to components (but not necessarily correct), we would conclude that there is no covariances between the estimators of the mixing proportions contained in $\hat{\boldsymbol{\pi}}$, and those of the component density parameters $\hat{\boldsymbol{\theta}}_j$, $j = 1, ..., G$. Now to compare to Boldea and Magnus we need to account for the different ordering of $\boldsymbol{\theta}$ they use, which is to put $\tilde{\boldsymbol{\pi}}$ before the vectors $\boldsymbol{\theta}_g$. Thus $Q_i^{g\pi} = \boldsymbol{D}^2_{(\tilde{\boldsymbol{\pi}})(\boldsymbol{\theta}_g)}\left(L_i(\boldsymbol{\theta}_g)\right)$ in our notation. We then see that the transpose of $Q_i^{\pi g} = \boldsymbol{D}^2_{(\boldsymbol{\theta}_g)(\tilde{\boldsymbol{\pi}})}\left(L_i(\tilde{\boldsymbol{\pi}})\right)$ given in their Theorem 1 is the same form as (C.97).

C.2.1 Score vector for MLMMs

We derive $\boldsymbol{D}_{\boldsymbol{\theta}}\left(L_i(\boldsymbol{\theta})\right)$ in (C.63) by firstly deriving $\boldsymbol{D}_{\boldsymbol{\theta}_g}\left(L_i(\boldsymbol{\theta}_g)\right)$ in (C.65). We use the notation $\tilde{\boldsymbol{e}}_{ig} = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g$. Now from (C.76) we see that the derivative vector of $L_i(\boldsymbol{\theta}_g^s)$ is a simple function of the derivative vector of $L_i^1(\boldsymbol{\theta}_g^s)$. This means if we set the weights to be one, the results in section (C.1) can be used (using only the i^{th} summand for unit i). So we get

$$\boldsymbol{D}_{\boldsymbol{\beta}_g}\left(L_i(\boldsymbol{\beta}_g)\right) = \alpha_{ig}\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{X}_i \tag{C.98}$$

from (C.22),

$$\boldsymbol{D}_{\sigma_g^2}\left(L_i(\sigma_g^2)\right) = -\frac{1}{2}\alpha_{ig}\mathrm{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right] + \frac{1}{2}\alpha_{ig}\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\tilde{\boldsymbol{e}}_{ig} \tag{C.99}$$

from (C.30),

$$\begin{aligned}\boldsymbol{D}_{\mathrm{v}(\boldsymbol{D}_g)}\left(L_i(\mathrm{v}(\boldsymbol{D}_g))\right) = &-\frac{1}{2}\alpha_{ig}\left[\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1})\right]^{\intercal}(\boldsymbol{Z}_i\otimes\boldsymbol{Z}_i)\widetilde{\boldsymbol{D}}_q \\ &+\frac{1}{2}\alpha_{ig}\left[(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i)\otimes(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i)\right]\widetilde{\boldsymbol{D}}_q\end{aligned} \tag{C.100}$$

from (C.47) for when $\boldsymbol{\psi}_g = \mathrm{v}(\boldsymbol{D}_g)$. We need to derive the $1 \times r$ derivative vector $\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(L_i(\boldsymbol{\phi}_g)\right)$ since it was not done so in section (C.1). From (C.55) we have

$$\begin{aligned}
\boldsymbol{d}\left(L_i^1(\boldsymbol{\phi}_g)\right) &= -\frac{1}{2}\mathrm{tr}[\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)\right)] + \frac{1}{2}\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{d}\left(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)\right)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\tilde{\boldsymbol{e}}_{ig} \\
&= -\frac{\sigma_g^2}{2}\mathrm{tr}[\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right)] + \frac{\sigma_g^2}{2}\mathrm{vec}\left[\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{d}\left(\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\tilde{\boldsymbol{e}}_{ig}\right] \\
&= -\frac{\sigma_g^2}{2}\mathrm{vec}\left[\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right]^{\intercal}\boldsymbol{d}\left(\mathrm{vec}[\boldsymbol{C}_i(\boldsymbol{\phi}_g)]\right) + \frac{\sigma_g^2}{2}\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\otimes\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\boldsymbol{d}\left(\mathrm{vec}[\boldsymbol{C}_i(\boldsymbol{\phi}_g)]\right) \\
&= \Bigg\{-\frac{\sigma_g^2}{2}\mathrm{vec}\left[\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right]^{\intercal}\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right) \\
&\quad+\frac{\sigma_g^2}{2}\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\otimes\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)\Bigg\}\boldsymbol{d}\left(\boldsymbol{\phi}\right), \qquad \text{(C.101)}
\end{aligned}$$

where for $v = 1, ..., r$ by changing every occurrence of ϕ_v to $(\boldsymbol{\phi}_g)_v$, $\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)$ is given in equation C.4, and where a summary of all the necessary equations required to calculate this can be found at the end of subsection C.1.1. So the $1 \times r$ derivative vector $\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(L_i^1(\boldsymbol{\phi}_g)\right)$ is given by the expression in curly brackets in (C.101). Thus

$$\begin{aligned}
\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(L_i(\boldsymbol{\phi}_g)\right) = &-\frac{\sigma_g^2}{2}\alpha_{ig}\mathrm{vec}\left[\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right]^{\intercal}\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right) \\
&+\frac{\sigma_g^2}{2}\alpha_{ig}\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\otimes\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right). \qquad \text{(C.102)}
\end{aligned}$$

C.2.2 Hessian for MLMMs

In order to derive $\boldsymbol{H}_{\boldsymbol{\theta}}\left(L_i(\boldsymbol{\theta})\right)$ in (C.64) we firstly derive $\boldsymbol{H}_{\boldsymbol{\theta}_g}\left(L_i(\boldsymbol{\theta}_g)\right)$ in (C.66), and we will derive the diagonal entries $\boldsymbol{H}_{\boldsymbol{\theta}_g^s}\left(L_i(\boldsymbol{\theta}_g^s)\right)$ first. We see from (C.79) that the Hessian of $L_i(\boldsymbol{\theta}_g^s)$ is a simple function of the Hessian and derivative vectors of $L_i^1(\boldsymbol{\theta}_g^s)$. So as in section (C.2.1), by setting the weights equal to one, the results in section (C.1) can be used (using only the i^{th} summand for unit i). Thus from (C.24) and (C.22) we get

$$\boldsymbol{H}_{\boldsymbol{\beta}_g}\left(L_i(\boldsymbol{\beta}_g)\right) = -\alpha_{ig}\boldsymbol{X}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{X}_i + \alpha_{ig}(1-\alpha_{ig})\boldsymbol{X}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\tilde{\boldsymbol{e}}_{ig}\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{X}_i. \qquad \text{(C.103)}$$

For the other components of $\boldsymbol{\theta}_g$, the expressions given by (C.79) do not simplify appreciably, and so we omit them for brevity. Instead we give the equation numbers for the derivative vectors and Hessians needed to compute it. So using (C.30), and the expression within curly brackets in (C.32), for $\boldsymbol{H}_{\sigma_g^2}\left(L_i(\sigma_g^2)\right)$ we need the following equations

$$\boldsymbol{D}_{\sigma_g^2}(L_i^1(\sigma_g^2)) = -\frac{1}{2}\text{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right] + \frac{1}{2}\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\tilde{\boldsymbol{e}}_{ig}, \quad \text{(C.104)}$$

and

$$\begin{aligned}\boldsymbol{H}_{\sigma_g^2}\left(L_i^1(\sigma_g^2)\right) = &\frac{1}{2}\text{tr}\left[\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\right] \\ &- \tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\tilde{\boldsymbol{e}}_{ig}. \qquad \text{(C.105)}\end{aligned}$$

Using (C.47), and (C.44), for $\boldsymbol{H}_{\text{v}(\boldsymbol{D}_g)}\left(L_i(\text{v}(\boldsymbol{D}_g))\right)$ we need

$$\begin{aligned}\boldsymbol{D}_{\text{v}(\boldsymbol{D}_g)}(L_i^1(\text{v}(\boldsymbol{D}_g))) = &-\frac{1}{2}\left[\text{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1})\right]^{\intercal}(\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i)\boldsymbol{D}_q \\ &+ \frac{1}{2}((\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i) \otimes (\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i))\boldsymbol{D}_q, \qquad \text{(C.106)}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{H}_{\text{v}(\boldsymbol{D}_g)}(L_i^1(\text{v}(\boldsymbol{D}_g))) = &\frac{1}{2}\widetilde{\boldsymbol{D}}_q^{\intercal}\left((\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i) \otimes (\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i)\right)\widetilde{\boldsymbol{D}}_q \\ &- \widetilde{\boldsymbol{D}}_q^{\intercal}\left((\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i) \otimes (\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\tilde{\boldsymbol{e}}_{ig}\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i)\right)\widetilde{\boldsymbol{D}}_q. \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(C.107)}\end{aligned}$$

The final Hessian we require is $\boldsymbol{H}_{\boldsymbol{\phi}_g}\left(L_i(\boldsymbol{\phi}_g)\right)$, for which we need $\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(L_i^1(\boldsymbol{\phi}_g)\right)$ from within curly brackets in (C.101), and

$$\boldsymbol{H}_{\boldsymbol{\phi}_g}\left(L_i^1(\boldsymbol{\phi}_g)\right) = \frac{\boldsymbol{\sigma}_g^4}{2}\left(\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)\right)^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)$$
$$-\boldsymbol{\sigma}_g^4\left(\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)\right)^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\otimes\left(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\tilde{e}_{ig}\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\right)\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)+g(\boldsymbol{\phi}_g), \tag{C.108}$$

from (C.59), where

$$\begin{aligned} g(\boldsymbol{\phi}_g) = &-\frac{\sigma_g^2}{2}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\left(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)_{jk}\boldsymbol{H}_{\boldsymbol{\phi}_g}((\boldsymbol{C}_i(\boldsymbol{\phi}_g))_{jk}) \\ &+\frac{\sigma_g^2}{2}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\sum_{t=1}^{n_i}\sum_{s=1}^{n_i}(\tilde{\boldsymbol{e}}_{ig})_t\left(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)_{tj}(\tilde{\boldsymbol{e}}_{ig})_s\left(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)_{sk}\boldsymbol{H}_{\boldsymbol{\phi}_g}((\boldsymbol{C}_i(\boldsymbol{\phi}_g))_{jk}), \end{aligned} \tag{C.109}$$

from (C.60), and where for $v = 1, ..., r$ by replacing every occurrence of ϕ_v with $(\boldsymbol{\phi}_g)_v$ $\boldsymbol{H}_{\boldsymbol{\phi}_g}((\boldsymbol{C}_i(\boldsymbol{\phi}_g))_{jk})$ is given in equation C.7, and where a summary of all the necessary equations required to calculate this can be found at the end of subsection C.1.1.

We now derive the cross-products in (C.66), which are given by $(C.83)$. For brevity we do not write $(C.83)$ out in full for each $s, t \in \{1, 2, 3, 4\}$, $s \neq t$, but rather give the equation numbers, or the equations, for the constituent parts. For $\boldsymbol{D}^2_{\left(\sigma_g^2\right)(\boldsymbol{\beta}_g)}\left(L_i(\boldsymbol{\beta}_g)\right)$ we need

$$\boldsymbol{D}_{\boldsymbol{\beta}_g}\left(L_i^1(\boldsymbol{\beta}_g)\right) = \tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{X}_i \tag{C.110}$$

from (C.22), and $\boldsymbol{D}_{\sigma_g^2}\left(L_i^1(\sigma_g^2)\right)$ from (C.104). The cross product $\boldsymbol{D}^2_{\left(\sigma_g^2\right)(\boldsymbol{\beta}_g)}\left(L_i^1(\boldsymbol{\beta}_g)\right)$ has not been derived in section (C.1), although much of the work has been done. Noting that when we view $\boldsymbol{V}_i(\boldsymbol{\zeta}_g)$ as a function of only σ_g^2 we have $\boldsymbol{D}_{\sigma_g^2}\left(\mathrm{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}_g))\right) = \mathrm{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))$. Then using (C.27) we have

$$\boldsymbol{D}^2_{\left(\sigma_g^2\right)(\boldsymbol{\beta}_g)}\left(L_i^1(\boldsymbol{\beta}_g)\right) = -\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1} \otimes \boldsymbol{X}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g)). \tag{C.111}$$

For $\boldsymbol{D}^2_{(\mathrm{v}(\boldsymbol{D}_g))(\boldsymbol{\beta}_g)}\left(L_i(\boldsymbol{\beta}_g)\right)$ we need (C.110) and (C.106). To derive $\boldsymbol{D}^2_{(\mathrm{v}(\boldsymbol{D}_g))(\boldsymbol{\beta}_g)}\left(L_i^1(\boldsymbol{\beta}_g)\right)$ we again use (C.27), which requires us to calculate $\boldsymbol{D}_{\mathrm{v}(\boldsymbol{D}_g)}\left(\operatorname{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}_g))\right)$. When $\boldsymbol{V}_i(\boldsymbol{\zeta}_g)$ is viewed as a function of only $\mathrm{v}(\boldsymbol{D}_g)$ we have $\boldsymbol{D}_{\mathrm{v}(\boldsymbol{D}_g)}\left(\operatorname{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}_g))\right) = (\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i)\widetilde{\boldsymbol{D}}_q$, and so

$$\boldsymbol{D}^2_{(\mathrm{v}(\boldsymbol{D}_g))(\boldsymbol{\beta}_g)}\left(L_i^1(\boldsymbol{\beta}_g)\right) = -\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{X}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q. \tag{C.112}$$

For $\boldsymbol{D}^2_{(\boldsymbol{\phi}_g)(\boldsymbol{\beta}_g)}\left(L_i(\boldsymbol{\beta}_g)\right)$ we need (C.110), and $\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(L_i^1(\boldsymbol{\phi}_g)\right)$ which is given by the expression within curly brackets in (C.101). Now when $\boldsymbol{V}_i(\boldsymbol{\zeta}_g)$ is viewed as a function of only $\boldsymbol{\phi}_g$ we have $\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}_g))\right) = \sigma_g^2\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)$. Thus using (C.27) we have

$$\boldsymbol{D}^2_{(\boldsymbol{\phi}_g)(\boldsymbol{\beta}_g)}\left(L_i^1(\boldsymbol{\beta}_g)\right) = -\sigma_g^2\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1} \otimes \boldsymbol{X}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right). \tag{C.113}$$

For $\boldsymbol{D}^2_{(\mathrm{v}(\boldsymbol{D}_g))\left(\sigma_g^2\right)}\left(L_i(\sigma_g^2)\right)$ we need (C.104), (C.100), and from (C.36)

$$\begin{aligned}\boldsymbol{D}^2_{(\mathrm{v}(\boldsymbol{D}_g))\left(\sigma_g^2\right)}\left(L_i^1(\sigma_g^2)\right) = {} & \frac{1}{2}\left[\operatorname{vec}\left(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\right]^{\intercal}\left(\boldsymbol{C}_i(\boldsymbol{\phi}_g)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q \\ & - \frac{1}{2}\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i \otimes \tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q \\ & - \frac{1}{2}\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i \otimes \tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{Z}_i\right)\widetilde{\boldsymbol{D}}_q.\end{aligned} \tag{C.114}$$

For $\boldsymbol{D}^2_{(\boldsymbol{\phi}_g)\left(\sigma_g^2\right)}\left(L_i(\sigma_g^2)\right)$ we need (C.104), $\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(L_i^1(\boldsymbol{\phi}_g)\right)$ which is given by the expression in curly parentheses in (C.101), and from (C.39)

$$
\begin{aligned}
\boldsymbol{D}^2_{(\boldsymbol{\phi}_g)(\sigma_g^2)}\left(L_i^1(\sigma_g^2)\right) = {} & \frac{\sigma_g^2}{2}\left[\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right]^{\intercal}\left(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\otimes\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)\\
& -\frac{1}{2}\left[\operatorname{vec}(\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1})\right]^{\intercal}\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)\\
& -\frac{\sigma_g^2}{2}\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\otimes\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)\\
& +\frac{1}{2}\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\otimes\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)\\
& -\frac{\sigma_g^2}{2}\left(\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\otimes\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\boldsymbol{C}_i(\boldsymbol{\phi}_g)\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\right)\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right).
\end{aligned}
\tag{C.115}
$$

For $\boldsymbol{D}^2_{(\boldsymbol{\phi}_g)(\mathrm{v}(\boldsymbol{D}_g))}\left(L_i(\mathrm{v}(\boldsymbol{D}_g))\right)$ we need $\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(L_i^1(\boldsymbol{\phi}_g)\right)$ which is given by the expression within curly brackets in (C.101), and (C.106). From the expression within curly brackets in (C.48), we also need

$$
\begin{aligned}
\boldsymbol{D}^2_{(\boldsymbol{\phi}_g)(\mathrm{v}(\boldsymbol{D}_g))}\left(L_i^1(\mathrm{v}(\boldsymbol{D}_g))\right) = {} & \frac{\sigma_g^2}{2}\widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\otimes\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1})\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)\\
& -\frac{\sigma_g^2}{2}\widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1})\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\tilde{\boldsymbol{e}}_{ig}\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1})\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right)\\
& -\frac{\sigma_g^2}{2}\widetilde{\boldsymbol{D}}_q^{\intercal}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1}\tilde{\boldsymbol{e}}_{ig}\tilde{\boldsymbol{e}}_{ig}^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1})\otimes(\boldsymbol{Z}_i^{\intercal}\boldsymbol{V}_i(\boldsymbol{\zeta}_g)^{-1})\boldsymbol{D}_{\boldsymbol{\phi}_g}\left(\operatorname{vec}(\boldsymbol{C}_i(\boldsymbol{\phi}_g))\right).
\end{aligned}
\tag{C.116}
$$

The cross products $\boldsymbol{D}^2_{(\boldsymbol{\theta}_j)(\boldsymbol{\theta}_k)}\left(L_i(\boldsymbol{\theta}_k)\right)$ in $(C.64)$, for $j,k\in\{1,...,G\}$ are given by (C.87). This can be calculated from (C.65), which in turn is given by the equations in section (C.2.1). For brevity we do not write these expressions out explicitly.

The final quantity of (C.64) we need to derive is $\boldsymbol{H}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right)$ which is given by

$$
\begin{aligned}
\boldsymbol{H}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right) &= \left\{{}_m\ \frac{\partial}{\boldsymbol{\pi}_k}\left[\frac{\partial L_i(\tilde{\boldsymbol{\pi}})}{\partial\boldsymbol{\pi}_j}\right]\right\}_{j=1,\,k=1}^{G-1\;G-1}\\
&= \left\{{}_m\ \frac{\partial}{\boldsymbol{\pi}_k}\left[\frac{\alpha_{ij}}{\boldsymbol{\pi}_j}-\frac{\alpha_{iG}}{\boldsymbol{\pi}_G}\right]\right\}_{j=1,\,k=1}^{G-1\;G-1}
\end{aligned}
\tag{C.117}
$$

using (C.92), where

$$\frac{\partial}{\boldsymbol{\pi}_k}\left[\frac{\alpha_{ij}}{\boldsymbol{\pi}_j}-\frac{\alpha_{iG}}{\boldsymbol{\pi}_G}\right]=\frac{\partial\alpha_{ij}}{\partial\boldsymbol{\pi}_k}\boldsymbol{\pi}_j^{-1}-\alpha_{ij}\boldsymbol{\pi}_j^{-2}\frac{\partial\boldsymbol{\pi}_j}{\partial\boldsymbol{\pi}_k}-\frac{\partial\alpha_{iG}}{\partial\boldsymbol{\pi}_k}\boldsymbol{\pi}_G^{-1}-\alpha_{iG}\boldsymbol{\pi}_G^{-2}. \tag{C.118}$$

Remembering that $\boldsymbol{\pi}_G = 1-\sum_{j=1}^{G-1}\boldsymbol{\pi}_j$, we see that $\partial/\partial\boldsymbol{\pi}_k[\sum_{l=1}^{G}\boldsymbol{\pi}_l f_{il}] = f_{ik}-f_{iG}$ for $k=1,...,G-1$, and so for $j,k=1,...,G-1$, we have

$$\frac{\partial\alpha_{ij}}{\boldsymbol{\pi}_k}=\begin{cases}\frac{f_{ik}}{\sum_{l=1}^{G}\boldsymbol{\pi}_l f_{il}}-\frac{\boldsymbol{\pi}_k f_{ik}^2}{\left(\sum_{l=1}^{G}\boldsymbol{\pi}_l f_{il}\right)^2}+\frac{\boldsymbol{\pi}_k f_{ik}f_{iG}}{\left(\sum_{l=1}^{G}\boldsymbol{\pi}_l f_{il}\right)^2} & \text{if } k=j\\ -\frac{\boldsymbol{\pi}_j f_{ij}f_{ik}}{\left(\sum_{l=1}^{G}\boldsymbol{\pi}_l f_{il}\right)^2}+\frac{\boldsymbol{\pi}_j f_{ij}f_{iG}}{\left(\sum_{l=1}^{G}\boldsymbol{\pi}_l f_{il}\right)^2} & \text{if } k\neq j.\end{cases} \tag{C.119}$$

Thus using (C.118) and (C.119) we have

$$\frac{\partial}{\boldsymbol{\pi}_k}\left[\frac{\alpha_{ij}}{\boldsymbol{\pi}_j}-\frac{\alpha_{iG}}{\boldsymbol{\pi}_G}\right]=\begin{cases}-\frac{\alpha_{ik}^2}{\boldsymbol{\pi}_k^2}+\frac{2\alpha_{ik}\alpha_{iG}}{\boldsymbol{\pi}_k\boldsymbol{\pi}_G}-\frac{\alpha_{iG}^2}{\boldsymbol{\pi}_G^2} & \text{if } k=j\\ -\frac{\alpha_{ij}\alpha_{ik}}{\boldsymbol{\pi}_j\boldsymbol{\pi}_k}+\frac{\alpha_{ij}\alpha_{iG}}{\boldsymbol{\pi}_j\boldsymbol{\pi}_G}+\frac{\alpha_{ik}\alpha_{iG}}{\boldsymbol{\pi}_k\boldsymbol{\pi}_G}-\frac{\alpha_{iG}^2}{\boldsymbol{\pi}_G^2} & \text{if } k\neq j.\end{cases} \tag{C.120}$$

So (C.118) becomes

$$\begin{aligned}\boldsymbol{H}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right)&=\left\{-\frac{\alpha_{ij}\alpha_{ik}}{\boldsymbol{\pi}_j\boldsymbol{\pi}_k}+\frac{\alpha_{ij}\alpha_{iG}}{\boldsymbol{\pi}_j\boldsymbol{\pi}_G}+\frac{\alpha_{ik}\alpha_{iG}}{\boldsymbol{\pi}_k\boldsymbol{\pi}_G}-\frac{\alpha_{iG}^2}{\boldsymbol{\pi}_G^2}\right\}_{j=1,\,k=1}^{G-1\,G-1}\\&=-\boldsymbol{D}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right)^{\intercal}\boldsymbol{D}_{\tilde{\boldsymbol{\pi}}}\left(L_i(\tilde{\boldsymbol{\pi}})\right),\end{aligned} \tag{C.121}$$

which agrees with $\boldsymbol{Q}_i^{\tilde{\pi}\tilde{\pi}}$ in Boldea and Magnus (2009, Theorem 1).

C.3 An alternative justification for componentwise inference

In Sub-subsection 3.4.3.1 we argued using intuition that our assumption of a well behaved MLMM implies $(I(\boldsymbol{\theta}_0))^{-1} \approx \text{diag}\{\boldsymbol{A}_1, ..., \boldsymbol{A}_G, \boldsymbol{A}_{G+1}\}$, and this "result" then formed the core assumption underpinning the use of componentwise inference. Given the importance of this result it is useful to derive it more mathematically, which is what we do now in this sub-subsection. In this respect, and again using the assumption of a well behaved MLMM, we firstly show that the Hessian of the MLMM log-likelihood function $H_{\boldsymbol{\theta}}(L_i(\boldsymbol{\theta}))$ for the i^{th} unit is given by

$H_{\boldsymbol{\theta}}(L_i(\boldsymbol{\theta})) \approx \text{diag}\{\mathbf{0}, ..., H_{\boldsymbol{\theta}_g}(L_i^1(\boldsymbol{\theta}_g)), ..., \mathbf{0}, H_{\tilde{\boldsymbol{\pi}}}(L_i(\boldsymbol{\theta}))\}$ if $\hat{p}_{ig} \approx 1$ for $g \in I_G$.

Consider the contribution of the i^{th} unit to $J_N(\hat{\boldsymbol{\theta}})$ which is given by the negative of the matrix in (C.64). We might want to use this matrix because as we described in Section 3.4 there are difficulties calculating $I_N(\hat{\boldsymbol{\theta}})$, however both are consistent estimators of $I(\boldsymbol{\theta}_0)$ given certain assumptions. We will show in subsection C.2 for the i^{th} unit, that that the off-diagonal sub-matrices of $J_N(\hat{\boldsymbol{\theta}})$, which we denote by $\boldsymbol{D}^2_{(\boldsymbol{\theta}_j)(\boldsymbol{\theta}_k)}(L_i(\boldsymbol{\theta}_k))$, $j, k \in I_G$, $j = 1, ..., G$, $j \neq k$, and $\boldsymbol{D}^2_{(\tilde{\boldsymbol{\pi}})(\boldsymbol{\theta}_j)}(L_i(\boldsymbol{\theta}_j))$, $j = 1, ..., G$, are given by (C.87) and (C.97) respectively. We will also show that the first G diagonal sub-matrices of $J_N(\hat{\boldsymbol{\theta}})$ contain the component specific Hessians $H_{\boldsymbol{\theta}_j}(L_i(\boldsymbol{\theta}_j))$, $j = 1, ..., G$, which are given by (C.66), whose diagonal elements are given by (C.79), and off-diagonal elements by (C.83). The terms α_{ij}, $j = 1, ..., G$ that feature in the aforementioned equations are equal to the $\hat{p}_{ij}$, $j = 1, ..., G$, when we replace the MLMM parameters in the terms α_{ij} with their estimates. Now if the classification problem for an MLMM is easy, then for all $i \in I_N$ we would expect the estimates of the posterior probabilities $\hat{p}_{ij}$, $j = 1, ..., G$ to be close to either one or zero. Consequently from the equations (C.87), (C.97) for all $j, k = 1, ..., G$, $j \neq k$ we see that $\boldsymbol{D}^2_{(\boldsymbol{\theta}_j)(\boldsymbol{\theta}_k)}(L_i(\boldsymbol{\theta}_k)) \approx \mathbf{0}$, and for all $j = 1, ..., G$ that $\boldsymbol{D}^2_{(\tilde{\boldsymbol{\pi}})(\boldsymbol{\theta}_j)}(L_i(\boldsymbol{\theta}_j)) \approx \mathbf{0}$.

Furthermore for any $g \in I_G$ when we look at the s sub-vectors in each $\boldsymbol{\theta}_g$, $s \in$

$\{1, 2, 3, 4\}$ (these correspond to the fixed effects, random effects and within-unit error covariance parameters, and the autoregressive parameter), from (C.79), and (C.83) we see for all $s, t = 1, ..., 4$ that we have $H_{\boldsymbol{\theta}_g^s}(L_i(\boldsymbol{\theta}_g^s)) \approx H_{\boldsymbol{\theta}_g^s}(L_i^1(\boldsymbol{\theta}_g^s))$, and $\boldsymbol{D}^2_{(\boldsymbol{\theta}_g^s)(\boldsymbol{\theta}_g^t)}(L_i(\boldsymbol{\theta}_g)) \approx \boldsymbol{D}^2_{(\boldsymbol{\theta}_g^s)(\boldsymbol{\theta}_g^t)}(L_i^1(\boldsymbol{\theta}_g))$ if unit i is in component g, and $H_{\boldsymbol{\theta}_g}(L_i(\boldsymbol{\theta}_g)) \approx \mathbf{0}$ and $\boldsymbol{D}^2_{(\boldsymbol{\theta}_g^s)(\boldsymbol{\theta}_g^t)}(L_i(\boldsymbol{\theta}_g)) \approx \mathbf{0}$ if not. From (C.66) we see this implies that $H_{\boldsymbol{\theta}_g}(L_i(\boldsymbol{\theta}_g)) \approx H_{\boldsymbol{\theta}_g}(L_i^1(\boldsymbol{\theta}_g))$ if unit i is in component g, whilst $H_{\boldsymbol{\theta}_g}(L_i(\boldsymbol{\theta}_g)) \approx \mathbf{0}$ if not.

Accordingly if unit i is in component g we see that $H_{\boldsymbol{\theta}}(L_i(\boldsymbol{\theta}))$ which is given by (C.64) will have all zero sub-matrices apart from the g^{th} diagonal one which will be approximately equal to the Hessian matrix of a LMM with just unit i in the sample. Thus the information contributed by the i^{th} unit to $I_N(\hat{\boldsymbol{\theta}})$ about the MLMM parameter estimator $\hat{\boldsymbol{\theta}}$, and approximated by $J_N(\hat{\boldsymbol{\theta}})$, will be approximately equal to the information that unit i contributes to $I_N^1(\hat{\boldsymbol{\theta}}_g)$.

Thus given a well behaved MLMM we have shown that the Hessian of the MLMM log-likelihood function $H_{\boldsymbol{\theta}}(L_i(\boldsymbol{\theta}))$ for the i^{th} unit is given by

$H_{\boldsymbol{\theta}}(L_i(\boldsymbol{\theta})) \approx \text{diag}\{\mathbf{0}, ..., H_{\boldsymbol{\theta}_g}(L_i^1(\boldsymbol{\theta}_g)), ..., \mathbf{0}, H_{\tilde{\boldsymbol{\pi}}}(L_i(\boldsymbol{\theta}))\}$ if $\hat{p}_{ig} \approx 1$, where $H_{\tilde{\boldsymbol{\pi}}}(L(\boldsymbol{\theta})) = \sum_{i=1}^N H_{\tilde{\boldsymbol{\pi}}}(L_i(\boldsymbol{\theta}))$. Thus summing over the N units we get

$\sum_{i=1}^N H_{\boldsymbol{\theta}}(L_i(\boldsymbol{Y}_i|\boldsymbol{\theta})) \approx \text{diag}\{\sum_{k=1}^{N_1} H_{\boldsymbol{\theta}_1}(L_k^1(\boldsymbol{Y}_k^{(1)}|\boldsymbol{\theta}_1)), ..., \sum_{k=1}^{N_G} H_{\boldsymbol{\theta}_G}(L_k^1(\boldsymbol{Y}_k^{(G)}|\boldsymbol{\theta}_G)), H_{\tilde{\boldsymbol{\pi}}}(L(\boldsymbol{\theta}))\}$. Letting $J^1_{N_g}(\boldsymbol{\theta}_g) = -\sum_{k=1}^{N_g} H_{\boldsymbol{\theta}_g}(L_k^1(\boldsymbol{Y}_k^{(g)}|\boldsymbol{\theta}_g))$ for $g \in I_G$ be the observed information matrix for the g^{th} 1-component model - i.e. using the 1-component log-likelihood function, we then get

$$\begin{aligned} J_N(\boldsymbol{\theta}) &= -H_{\boldsymbol{\theta}}(L(\boldsymbol{Y}|\boldsymbol{\theta})) \\ &= -\sum_{i=1}^{N} H_{\boldsymbol{\theta}}(L_i(\boldsymbol{Y}_i|\boldsymbol{\theta})) \\ &\approx \text{diag}\left\{J^1_{N_1}(\boldsymbol{\theta}_1), ..., J^1_{N_G}(\boldsymbol{\theta}_G), H_{\tilde{\boldsymbol{\pi}}}\right\}. \end{aligned} \tag{C.122}$$

Noting that $N^{-1}J^1_{N_g}(\boldsymbol{\theta}_g) = \left(\frac{N_g}{(N)(N_g)}\right) J^1_{N_g}(\boldsymbol{\theta}_g) = \boldsymbol{\pi}_g N_g^{-1} J^1_{N_g}(\boldsymbol{\theta}_g)$, and because for any $g \in I_G$ we are assuming that $N_g^{-1} J_{N_g}(\hat{\boldsymbol{\theta}}^1_g(\boldsymbol{Y}^{(g)}))$ is a consistent estimator of $I^1(\boldsymbol{\theta}^0_g)$ (see Section 3.3), then from (C.122) and (C.121) we have

$$\lim_{N\to\infty} N^{-1} J_N(\hat{\boldsymbol{\theta}}(\boldsymbol{Y})) \approx \text{diag}\left\{\boldsymbol{\pi}_1 I^1(\boldsymbol{\theta}^0_1), ..., \boldsymbol{\pi}_G I^1(\boldsymbol{\theta}^0_G), \text{diag}\{-\boldsymbol{\pi}_j^{-1}\}_{j=1}^{G-1}\right\}, \tag{C.123}$$

where again we use the fact that $N_g/N = \boldsymbol{\pi}_g$ for all N as $N \to \infty$. Since from Subsection 3.4.2 we are assuming naively that $N^{-1}J_N(\hat{\boldsymbol{\theta}}(\boldsymbol{Y}))$ is a consistent estimator of $I(\boldsymbol{\theta}_0)$ then C.123 shows that

$$\begin{aligned} N^{-1} J_N(\hat{\boldsymbol{\theta}}(\boldsymbol{Y})) &\xrightarrow{P} I(\boldsymbol{\theta}_0) \\ &\approx \text{diag}\left\{\boldsymbol{\pi}_1 I^1(\boldsymbol{\theta}^0_1), ..., \boldsymbol{\pi}_G I^1(\boldsymbol{\theta}^0_G), \text{diag}\{-\boldsymbol{\pi}_j^{-1}\}_{j=1}^{G-1}\right\}. \end{aligned} \tag{C.124}$$

As required equation C.124 agrees with the result $(I(\boldsymbol{\theta}_0))^{-1} \approx \text{diag}\{\boldsymbol{A}_1, ..., \boldsymbol{A}_G, \boldsymbol{A}_{G+1}\}$ we derived in Sub-subsection 3.4.3.1 - i.e. $\boldsymbol{A}_g^{-1} = \boldsymbol{\pi}_g I^1(\boldsymbol{\theta}^0_g)$ for $g \in I_G$, and $\boldsymbol{A}_{G+1} = \text{diag}\{-\boldsymbol{\pi}_j^{-1}\}_{j=1}^{G-1}$.

Bibliography

J. Aitchison and S.D. Silvey. Maximum-Likelihood Estimation of Parameters Subject to Restraints. *Annals of Mathematical Statistics*, 29(3):813–828, 1958.

F Bartolucci, S Bacci, and F Pennoni. Mixture latent autoregressive models for longitudinal data, August 2011. URL `http://arxiv.org/abs/1108.1498`.

P. Billingsley. *Probability and Measure.* John Wiley & Sons, Chichester, UK, third edition, 1995.

K.G. Binmore. *Topological Ideas.* Cambridge University Press, London, UK, 1981.

O Boldea and J.R Magnus. Maximum Likelihood Estimation of the Multivariate Normal Mixture Model. *Journal of the American Statistical Association*, 104(488):1539–1549, 2009.

Jenkins G.M. Box, G.E.P. and G.C. Reinsel. *Time Series Analysis.* Pearson Education, Singapore, third edition, 1994.

G Celeux, O Martin, and C Lavergne. Mixture of Linear Mixed Models for Clustering Gene Expression Profiles from Repeated Microarray Experiments. *Statistical Modelling*, 5(3):243–267, 2005.

K.C. Chanda. A Note on the Consistency and Maxima of the Roots of Likelihood Equations. *Biometrika*, 41(1-2):56–61, 1954.

Laird N. Cnaan, A. and P Slasor. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statist. Med.*, 16:2349–2380, 1997.

Geoffrey Coke and Min Tsao. Random effects mixture models for clustering electrical load series. *Journal of Time Series Analysis*, 31(6):451–464, 2010.

H. Cramér. *Mathematical Methods of Statistics.* Princeton University Press, Princeton, USA, 1946.

N.E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474, 1969.

E. Demidenko. *Mixed Models, Theory and Applications.* John Wiley & Sons, Chichester, UK, 2004.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via EM Algorithm. *Journal of the Royal Statisical Society Series B-Methodological*, 39(1):1–38, 1977.

Sylvia Fruehwirth-Schnatter and Sylvia Kaufmann. Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26(1):78–89, 2008.

Bettina Grün. Fitting finite mixtures of linear mixed models with the EM algorithm.

In Paula Brito, editor, *Compstat 2008—Proceedings in Computational Statistics*, volume II, pages 165–173. Physica Verlag, Heidelberg, Germany, 2008.

Bettina Grün and Kurt Hornik. Modelling HIV RNA levels with finite mixtures for censored longitudinal data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2011.

R.J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.*, 13(2):795–800, 1985.

C. Hennig. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296, 2000.

R.I. Jennrich. Asymptotic Properties of Non-Linear Least Squares Estimators. *Ann. Math. Statist*, 40(2):633–643, 1969.

M.I Jordan. Lecture 12: Stochastic equicontinuity and chaining, 2007. URL `https://www.koofers.com/files/notes-mbkafeyvda/`.

N.M. Kiefer. Discrete Parameter Variation - Efficient Estimation of a Switching Regression-Model. *Econometrica*, 46(2):427–434, 1978.

E.L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer-Verlag, New York, USA, second edition, 1998.

J Li. Uniform convergence and stochastic equicontinuity lecture notes. URL `http://econ.duke.edu/uploads/media_items/`

`uniform-convergence-and-stochastic-equicontinuity.original.pdf`, `accessed:2014-08-30`.

B.G. Lindsay. *Mixture Models: Theory, Geometry and Applications.* Institute of Mathematical Statistics, Hayward, USA, 1995.

T.A. Louis. Finding the Observed Information Matrix when using the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233, 1982.

Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons, Chichester, UK, third edition, 1999.

X.L. Meng and D.B Rubin. Maximum likelihood estimation via EM Algorithm: A general framework. *Biometrika*, 80(2):267–78, 1993.

B.C. Peters and H.F Walker. An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM Review*, 35 (2):362–378, 1978.

R.A Redner and H.F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984.

E. San Martin and F Quintana. Consistency and identifiability revisited. *Brazilian Journal of Probability and Statistics*, 16:99–106, 2002.

M. Schervish. *Theory of Statistics.* Springer-Verlag, New York, USA, 1995.

James R. Schott. *Matrix analysis for statistics.* John Wiley & Sons, Chichester, UK, second edition, 2005.

George A.F. Seber and A.J. Lee. *Linear Regression Analysis.* John Wiley & Sons, Chichester, UK, second edition, 2003.

R.J. Serfling. *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, Chichester, UK, 1980.

M.L. Siow and A. Hackshaw. A potential new enriching trial design for selecting non-small-cell lung cancer patients with no predictive biomarker for trials based on both histology and early tumor response: further analysis of a thalidomide trial. *Cancer Medicine.*, 2(3):360–365, 2013.

M.L. Siow, R. Rudd, P.J. Woll, C. Ottensmeier, D. Gilligan, A. Price, S. Spiro, N. Gower, M. Jitlal, and A. Hackshaw. Randomized double-blind placebo-controlled trial of thalidomide in combination with gemcitabine and carboplatin in advanced non-small-cell lung cancer. *Journal of Clinical Oncology.*, 27(31):5248–5254, 2009.

R. Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.*, 26(2):195–239, 1974.

K. Tanaka and A. Takemura. Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when the scale parameters are exponentially small. *Bernoulli*, 12(6):1003–1017, 2006.

R.E. Tarone and G. Gruenhage. Note on Uniqueness of Roots of Likelihood Equations for Vector-Valued Parameters. *Journal of the American Statistical Association*, 70 (352):903–904, 1975.

A.W. Van Der Vaart. *Asymptotic Statistics.* Cambridge University Press, Cambridge, UK, 1998.

G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data.* Springer Verlag, New York, USA, 2009.

W. Wang and T. Fan. ECM-based Maximum Likelihood Inference for Multivariate Linear Mixed Models with Autoregressive Errors. *Computational Statistics and Data Analysis*, 54():1328–1341, 2009.

H. White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50: 1–26, 1982.

W Xu and D Hedeker. A random-effetcs mixture model for classifying treatment response in longitudinal clinical trials. *Journal of Biopharmaceutical Statistics*, 11(4): 253–273, 2002.

S.J. Yakowitz and J.D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.