UNIVERSITY COLLEGE LONDON

Contributions to Inference without Likelihoods

by

João Jesus

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Mathematical & Physical Sciences
Department of Statistical Science

October 2011

Declaration of Authorship

I declare that this dissertation represents my own work, except where due acknowledgment is made.

Part of the work presented in the thesis has been published:

Jesus, J. and Chandler, R.E. Estimating functions and the generalized method of moments. *Interface Focus* (2011) (On-line Early Release, doi: 10.1098/rsfs.2011.0057).

Signed: _____

Date: _____

UNIVERSITY COLLEGE LONDON

Abstract

Faculty of Mathematical & Physical Sciences
Department of Statistical Science

Doctor of Philosophy

by João Jesus

This thesis is concerned with statistical inference in situations where one is unwilling or unable to formulate a likelihood function. The theory of estimating functions (EFs) provides an alternative inference framework in such settings.

The research was motivated by problems arising in the application of a class of stochastic models for rainfall based on point processes. These models are often used by hydrologists to produce synthetic rainfall sequences for risk assessment purposes, notably in the UKCP09 climate change projections for the UK. In the absence of a likelihood function, the models are usually fitted by minimizing some measure of disagreement between theoretical properties and the observed counterparts.

In general situations of this type, two "subjective" decisions are required: what properties to use, and how to weight their contribution to the objective function. The choice of weights can be formalised by defining a minimum variance criterion for the estimator. This is equivalent to the Generalized Method of Moments estimator which is widely used in econometrics. The first contribution of this thesis is to translate the problem to an EF framework which is much more familiar to statisticians. Simulations show that the theory has poor finite sample performance for point process rainfall models. This is associated with inaccurate estimation of the covariance matrix of observed properties. A two-stage approach is developed to overcome this problem.

The second main contribution is to apply EF theory to the Whittle likelihood, which is based on the periodogram of the data. A problem here is that the covariance matrix of the estimators depends on fourth-order properties which are often intractable. An EF approach provides a feasible alternative in practical applications. After establishing the conditions under which EF theory can be applied to Whittle estimation, simulations are once again used to explore the finite sample performance.

Acknowledgements

First and most importantly, I would like to express my gratitude to my primary supervisor, Dr. Richard Chandler, for his guidance and support throughout the work for this thesis. I feel very fortunate to have had the opportunity to work with such a knowledgeable and inspiring person who shared his ideas and time with me during the entire PhD. I would also like to thank my subsidiary supervisor, Dr. Paul Northrop, for his reviews, comments and suggestions, particularly during the first year. This thesis has undoubtedly benefited from the comments and suggestions from staff and fellow students during departmental seminars and presentations as well as more informal discussions.

I am deeply grateful to the Department of Statistical Science and its staff, past and present, for the opportunities that led to the present work. This thesis would not have been possible without the inspiring experience from my first spell in the department, when the desire to do research in the field of statistical inference was formed.

The work in this thesis was made possible through a studentship from the Engineering and Physical Sciences Research Council.

Finally, I would like to show my deepest gratitude to my wife, Ana, who has been a constant source of encouragement and support, even during the most difficult times.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The likelihood function is fundamental to most modern methods of statistical inference. Suppose a data vector $\mathbf{y}$ is considered as the realized value of a random vector $\mathbf{Y}$ with joint density $f(\mathbf{y};\theta)$, for a parameter vector θ in some set Θ. Then the likelihood function is defined as $L(\theta|\mathbf{y}) \propto f(\mathbf{y};\theta)$ for $\theta \in \Theta$, and can be used to make inference about the value of θ on the basis of the data $\mathbf{y}$.

Likelihood-based inference yields point estimates of θ, as well as assessments of uncertainty such as confidence regions. Unfortunately, inference based on the likelihood function is not always feasible. This may be because the structure of a model may be too complex to derive the joint density $f(\mathbf{y};\theta)$; or because the full probability structure of a statistical model has not been specified, either because it is not possible or not desirable. Some examples can be found in Fuentes (2007), Li and Yin (2009), Mikosch et al. (1993).

An important example is the class of models based on point processes (Cox and Isham, 1980) which in fact provided a motivation for the work presented here. This class of processes is often used by hydrologists and meteorologists to model rainfall, their use is widespread across UK institutions from which we highlight the Department for Environment Food and Rural Affairs (DEFRA). The UK Climate Projections, and in particular UKCP09, is based on a weather generator that models rainfall using a Neyman-Scott Rectangular Pulses Model(NSRPM) (Burton et al., 2008), these models are described later in the thesis. The NSRPM and the Poisson Rectangular Pulses Model(PRPM) will be used in this thesis to show the application of the different methods of estimation, also with the aim of evaluating finite sample performance of asymptotic theory.

The theory of estimating functions provides a general framework that is useful in the absence of a likelihood function as it allows for consistent estimation of θ as well as characterization of uncertainty. The groundwork of the estimating functions theory was introduced by Godambe (1960) and Durbin (1960), where a major share of their work has been devoted to finding optimal estimating functions, in the sense that the resulting estimators have the smallest possible standard errors within some class. In Chapter 2 we present definitions and results that are paramount to the remainder of the thesis. We will describe the theory of estimating functions in detail, discussing the concept of optimality and deriving some asymptotic results, with focus on the statement of conditions for consistency and asymptotic distribution in a way that these can be checked in practical applications. Towards the end of Chapter 2 we show how some likelihood based estimators can be fitted in the estimating functions framework.

In many applications (e.g. Rodriguez-Iturbe et al. (1987), Wheater et al. (2005) , Anderson and Sørensen (1996) and Li and Yin (2009)) the model can be specified in terms of a restricted set of constraints representing the relationship between the parameters and the data. These commonly involve summary statistics, $T(\mathbf{y})$, (e.g. means, variances and autocorrelations) computed from the observations $\mathbf{y}$ which are somehow matched with their expectations under the model, $\tau(\theta)$. Methods for dealing with these situations have been developed via separate frameworks in the fields of statistics and econometrics, see Bera and Bilias (2002) for a review. In the statistics literature, the treatment is usually via estimating functions, where in econometrics inference is often carried out using a generalized method of moments (GMM). The foundations of the GMM methodology for econometric problems were set out by Hansen (1982). In the GMM context the constraints representing the relationship between the parameters and the data are called moment conditions, and estimation is done by minimizing a weighted sum of squares or quadratic form as a measure of disagreement between observed and theoretical properties. In Chapter 3 we establish the parallels between the econometrics and the statistics building blocks of moment based inference, within the framework of estimating functions. We translate the requirements for consistency and asymptotic distribution shown in Chapter 2 into the GMM settings, and by further exploring the concept of optimal estimating function within a certain class we show the optimal way of combining any given set of moment conditions - optimal weighting. In Chapter 4 a simulation study is performed with data generated using PRPM and NSRPM; The aim is to look at finite sample performance when inference is done using asymptotic approximations (asymptotic distribution and optimal weights), and show that even for a modest sample size the use of such approximations can prove useful.

There are also situations where the features of interest of the model can be defined in terms of its spectral representation, effectively the absence of $L(\theta|\mathbf{y})$ is surrounded by the existence of the second order spectral density, $h(w, \theta)$. This leads to estimation in the spectral domain where the Whittle likelihood (Whittle, 1953) is the most common method. The rationale behind such method is that the Fourier coefficients of a stationary process are approximately Gaussian in large samples, and this allows for the construction of an approximate likelihood for the transformed data. Although the original work in Whittle (1953) was based on the assumption that the original data was Gaussian, Hannan (1973) and Robinson (1978) proved the consistency and asymptotic normality of the Whittle estimator for the wider classes of stationary processes with no long range dependence. One of the weaknesses of the Whittle likelihood approach is that the covariance matrix of the estimator depends on the fourth order spectrum which for many processes of interest is very difficult to obtain. The translation of the Whittle likelihood principles into the estimating functions framework can provide a solution for this problem in practical applications. In Chapter 5 we start by giving an overview of the developments in parametric estimation in the frequency domain, we then present some steps involved in deriving the Whittle likelihood as this is important to understand the approximations involved in the Whittle method. We then proceed to prove that applying asymptotic results from the estimating functions theory makes it possible to obtain an asymptotic distribution for the estimator that does not depend on the expression of fourth order spectrum. The finite sample properties of the Whittle estimator are studied using simulations of PRPM and NSRPM in Chapter 6.

Chapter 2

Theory of Estimating Functions

In this chapter we describe in some detail the theory of estimating functions, we start by presenting some notation and a definition of estimating functions. Then through the treatment of the particular case where the estimating functions are linear, we show some concepts that are later generalized to a broader class of estimating functions. In this chapter we will cover issues like optimality, lower bounds, consistency and asymptotic distribution for a very generic class of estimating functions. The final part of the chapter focus on the broad character of this theory by showing that some more traditional approaches can be seen as particular cases of this theory. The theory of estimating functions provides a general framework for when the investigator wants to estimate one or several parameters of interest that belong to the representation of some statistical model. Assume that there is available a data vector of length n, regarded as the realized value of a vector $\mathbf{Y}$ of random variables. The distribution, or process, generating the data $\mathbf{Y}$ is considered to be a member of some family of distributions (or processes) indexed by the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$. We denote by $\boldsymbol{\theta}_0$ the value of the parameter corresponding to the data on hand, i.e., the estimation target. In cases where the functional form of the distribution is fully specified, we denote the density of $\mathbf{Y}$ by $f(\mathbf{y}; \boldsymbol{\theta})$.

Suppose we have a vector-valued function $\mathbf{g}(\boldsymbol{\theta}; \mathbf{Y})$, $\mathbb{R}^p \times \mathbb{R}^n \longmapsto \mathbb{R}^k$, $k \geq p$, such that:

$$E_{\boldsymbol{\theta}_0}[\mathbf{g}(\boldsymbol{\theta}_0; \mathbf{Y})] = \mathbf{0} \tag{2.1}$$

where $\mathbf{0}$ is a $k \times 1$ vector of zeros, and $E_{\boldsymbol{\theta}_0}$ is the expectation with respect to the true distribution of $\mathbf{Y}$. Such function is called an *estimating function* (EF), and the

corresponding equations

$$\mathbf{g}(\boldsymbol{\theta}; \mathbf{Y}) = \mathbf{0} \tag{2.2}$$

are called *estimating equations* (EE) (Godambe, 1960). The statistic $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ is an estimator of $\boldsymbol{\theta}$ if equation (2.2) is satisfied by $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\mathbf{Y})$. Some authors (e.g. Durbin (1960)) use the terminology unbiased EF alluding to the fact that the estimator $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ is asymptotically unbiased for $\boldsymbol{\theta}$; however, more conditions are required on the EF for that to hold, as will become clear in Section 2.4 below. Furthermore if the following conditions are satisfied

$$\partial g/\partial \boldsymbol{\theta} \text{ exists for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}; \tag{2.3}$$

$$E_{\boldsymbol{\theta}_0}[\partial g/\partial \boldsymbol{\theta}]^2 > 0 \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}; \tag{2.4}$$

$$\int \mathbf{g}(\boldsymbol{\theta}; \mathbf{y}) f(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} \text{ is differentiable with respect to } \boldsymbol{\theta} \text{ under the integral sign ;} \tag{2.5}$$

then $\mathbf{g}(\boldsymbol{\theta}; \mathbf{y})$ is called a *regular estimating function* (Godambe, 1960). For the remainder of this section and so that the following results are valid it will be assumed that $\mathbf{g}(\boldsymbol{\theta}; \mathbf{y})$ is a regular estimating function with the same dimension as the parameter (i.e. $k = p$) and its corresponding EE has solution $\hat{\boldsymbol{\theta}}(\mathbf{y})$, which is assumed to be unique. We start by looking at the case where $\mathbf{g}(\boldsymbol{\theta}; \mathbf{y})$ is linear and then move to a more general framework.

2.1 Linear Estimating Functions

The theory of estimating functions was originally developed by Durbin (1960) and Godambe (1960, 1976). Durbin considered in particular the case where $\mathbf{g}(\boldsymbol{\theta}; \mathbf{y})$ is linear in $\boldsymbol{\theta}$ and $k = p$. In this section we examine this class of linear estimating functions, as a means of establishing some fundamental ideas that will be required later. For this class the estimating equation takes the following form:

$$\mathbf{g}(\boldsymbol{\theta};\mathbf{y}) = T_1(\mathbf{y})\boldsymbol{\theta} + T_2(\mathbf{y}) = \mathbf{0} \tag{2.6}$$

where $T_1(\mathbf{y})$ is a $p \times p$ matrix and $T_2(\mathbf{y})$ is a $p \times 1$ vector, both depending only on the observations, and $\mathbf{0}$ is a $p \times 1$ vector of zeros. If

- $T_1(\mathbf{y})$ is non-singular with probability one
- $E_{\boldsymbol{\theta}_0}[T_1(\mathbf{Y})\boldsymbol{\theta}_0 + T_2(\mathbf{Y})] = 0$

then (2.6) is called a *set of linear estimating equations*, and conditions (2.3-2.5) hold.

We now introduce the idea of an optimal estimating function within the class of linear estimating functions. The principle is that the optimality of the estimator defined as the solution of an estimating equation is closely related to the optimality of the estimating function, as will be shown in the following sections. For any class of regular estimating functions, and in particular the linear case, desirable properties of $\mathbf{g}(\boldsymbol{\theta};\mathbf{y})$ are as follows:

1. Evaluating the estimating function at the true value of the parameter given any set of observations y should result in a value as close to zero as possible.
2. For values of $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{\delta}$, say, different from the true parameter value the expectation $E_{\boldsymbol{\theta}_0}[\mathbf{g}(\boldsymbol{\theta}_0 + \boldsymbol{\delta};\mathbf{Y})]$ should be as large as possible for an arbitrary small $\boldsymbol{\delta}$.

In the one parameter case (i.e. when $\boldsymbol{\theta}$ is scalar), the first criterion can be summarized as minimizing the variance of $\mathbf{g}(\boldsymbol{\theta}_0;\mathbf{Y})$, which is the same as minimizing $E_{\boldsymbol{\theta}_0}[\mathbf{g}(\boldsymbol{\theta}_0;\mathbf{Y})^2]$, from (2.1). The second criteria can be achieved by maximizing $E_{\boldsymbol{\theta}_0}\left[\frac{\partial \mathbf{g}(\boldsymbol{\theta};\mathbf{Y})}{\partial \boldsymbol{\theta}}|\boldsymbol{\theta}_0\right]^2$.

For the multi-parameter case the notion of variance is extended to the concept of covariance matrix,

$$\mathrm{Var}[\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})] = E[\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})^T] \ . \tag{2.7}$$

Minimizing $\mathrm{Var}[\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})]$ in the matrix sense conserves the meaning set out in point 1. above, similarly for the parameter vector case the criterion defined in point 2. above can be achieved by maximizing $E_{\boldsymbol{\theta}_0}\left[\frac{\partial \mathbf{g}(\boldsymbol{\theta};\mathbf{Y})}{\partial \boldsymbol{\theta}}|\boldsymbol{\theta}_0\right]^2$ in the matrix sense. A rigorous definition of matrix comparison is that $A > B \Leftrightarrow A - B = C$, where C is a positive-definite matrix;

if C is positive semi-definite then we can claim $A \geq B$, and this definition will be used to define the concept of optimal estimating function. From Definition (2.6) and (2.7), given a particular choice of $T_1(\mathbf{Y})$and $T_2(\mathbf{Y})$ it is possible to find $T_1^*(\mathbf{Y})$ and $T_2^*(\mathbf{Y})$ leading to arbitrarily small elements of the covariance matrix without changing the solution of (2.6). This can be done by multiplication of $T_1(\mathbf{Y})$ and $T_2(\mathbf{Y})$ by an appropriate p x p matrix of constants. This shows the need to further restrict our class of estimating functions. Durbin (1960) suggested the comparison to be made between standardized versions of the estimating functions, with standardization constant being $E\left[\frac{\partial \mathbf{g}(\boldsymbol{\theta};\mathbf{Y})}{\partial \boldsymbol{\theta}}\right]^{-1}$. It is clear that for estimating functions of the form in (2.6), $\frac{\partial \mathbf{g}(\boldsymbol{\theta};\mathbf{y})}{\partial \boldsymbol{\theta}} = T_1(\mathbf{y})$. The standardized version of (2.6) is therefore

$$\tilde{T}_1\boldsymbol{\theta} + \tilde{T}_2 = 0$$

where $\tilde{T}_1 = E[T_1(\mathbf{Y})]^{-1}T_1(\mathbf{Y})$ (so that $E[\tilde{T}_1] = I$, where I is the identity $p \times p$ matrix) and $\tilde{T}_2 = E[T_1(\mathbf{Y})]^{-1}T_2(\mathbf{Y})$.

Having determined an appropriate standardization, we can consider what might be the "best" set of estimating functions within the linear class considered in this section. According to criteria 1. above, these will be the estimating functions with the lowest variance. Thus if there are $\tilde{T}_1^*$ and $\tilde{T}_2^*$ such that for all other linear estimating functions having $E[\tilde{T}_1] = I$

$$\mathrm{Var}[\tilde{T}_1\boldsymbol{\theta}_0 + \tilde{T}_2] - \mathrm{Var}[\tilde{T}_1^*\boldsymbol{\theta}_0 + \tilde{T}_2^*]$$

is positive (semi-)definite then the equations $\tilde{T}_1^*\hat{\boldsymbol{\theta}} + \tilde{T}_2^* = 0$ are called a *set of best linear estimating equations.*

This means that given any $p \times p$ matrix λ, the variance of the product $\lambda(\tilde{T}_1^*\boldsymbol{\theta}_0 + \tilde{T}_2^*)$ is never larger than that of the corresponding product $\lambda(\tilde{T}_1\boldsymbol{\theta}_0 + \tilde{T}_2)$. Note that

$$\mathrm{Var}[\lambda(\tilde{T}_1\boldsymbol{\theta}_0 + \tilde{T}_2)] - \mathrm{Var}[\lambda(\tilde{T}_1^*\boldsymbol{\theta}_0 + \tilde{T}_2^*)] = \lambda(\mathrm{Var}[\tilde{T}_1\boldsymbol{\theta}_0 + \tilde{T}_2] - \mathrm{Var}[\tilde{T}_1^*\boldsymbol{\theta}_0 + \tilde{T}_2^*])\lambda^T$$

is also positive (semi-)definite, which means that $\mathrm{Var}[\lambda(\tilde{T}_1^*\boldsymbol{\theta}_0 + \tilde{T}_2^*)]$ is never greater than $\mathrm{Var}[\lambda(\tilde{T}_1\boldsymbol{\theta}_0 + \tilde{T}_2)]$, in the matrix sense.

At this point we provide a simple example of how the variance of a linear estimating function can be related to the variance of the resulting estimator. Let $\mu = E[Y]$ and $\sigma^2 = \text{Var}[Y]$ be the parameters of interest, given a sample of size n of realizations of Y, we can define the following linear estimating function

$$\mathbf{g}(\boldsymbol{\theta}; \mathbf{Y}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} - \begin{bmatrix} \overline{Y} \\ \frac{\sum_{t=1}^{n} (Y_t - \overline{Y})^2}{n-1} \end{bmatrix} \tag{2.8}$$

where $\overline{Y} = \frac{\sum_{t=1}^{n} Y_t}{n}$ which gives the estimate,

$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{t=1}^{n} y_t}{n} \\ \frac{\sum_{t=1}^{n} (y_t - \overline{y})^2}{n-1} \end{bmatrix}$$

In this case the variance of the estimator is the same as the variance of the estimating function, this idea will be pursued later in more general cases.

2.2 Non-linear case

The theory of estimating equations described above can be extended to the case where $\mathbf{g}(\boldsymbol{\theta}; \mathbf{y})$ is non-linear in $\boldsymbol{\theta}$. Consider a vector-valued regular estimating function, having $k = p$. In this case the definition of covariance matrix of an estimating function (2.7) applies in the same way. Furthermore define

$$\mathbf{D}_{\mathbf{g}}(\boldsymbol{\theta}) = E\left[\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}}\right] \tag{2.9}$$

so that we can define the class of standardized estimating functions in the non-linear case as

$$g_s(\boldsymbol{\theta}; \mathbf{y}) = {D_{\mathbf{g}}}^{-1}(\boldsymbol{\theta}) \mathbf{g}(\boldsymbol{\theta}; \mathbf{y}) \tag{2.10}$$

The idea of comparing covariance matrices is also valid for non-linear estimating functions, and as in the linear case it is possible to define a class of standardized estimating functions. The next section develops the idea of optimal covariance matrices within a given class of EFs further.

2.3 Optimality and lower bound

Based on the idea of ranking covariance matrices to compare estimators, it follows naturally given the principle shown in the example of the previous section, (2.8), to ask if there is an optimal covariance matrix for estimating functions. The first step in this direction is to establish lower bound for the covariance matrix (Godambe, 1960, 1976; Chandrasekar and Kale, 1984; Mukhopadhyay, 2007). We will look at the lower bound for a class of estimating functions, $G_p^{(s)}$, which consists of all the p-dimensional standardized estimating functions $\mathbf{g}_s(\boldsymbol{\theta};\mathbf{y})$ with covariance matrix (2.7) that is finite and positive definite. Another assumption needed for this result is that the information matrix $\mathbf{J}(\boldsymbol{\theta})$,

$$\mathbf{J}(\boldsymbol{\theta}) = E\left[\frac{\partial \log f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^T \frac{\partial \log f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$$

where $f(\mathbf{y};\boldsymbol{\theta})$ is the density of $\mathbf{Y}$, exists and is positive definite for all $\boldsymbol{\theta}$. To simplify notation, let

$$\begin{aligned}\mathbf{S}(\boldsymbol{\theta}) &= \frac{\partial \log f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ \mathbf{N}(\boldsymbol{\theta}) &= E\left[\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})\mathbf{S}(\boldsymbol{\theta})^T\right] = E\left[\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})\frac{\partial \log f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right].\end{aligned} \tag{2.11}$$

Starting from (2.1),

$$E[\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})] = \int \mathbf{g}(\boldsymbol{\theta};\mathbf{y}) f(\mathbf{y};\boldsymbol{\theta}) \mathbf{dy}$$

This is only true at the true parameter value, elsewhere denoted $\boldsymbol{\theta}_0$, but in this abstract setting we have to consider that true parameter value can be any value in $\boldsymbol{\Theta}$, this is dealt by considering that the $\boldsymbol{\theta}$ in $\mathbf{g}(\boldsymbol{\theta};\mathbf{y})$ is the same as in $f(\mathbf{y};\boldsymbol{\theta})$ and that is the case considered here. Using assumption (2.5), and differentiating with respect to $\boldsymbol{\theta}$,

$$\int \frac{\partial \mathbf{g}(\boldsymbol{\theta};\mathbf{y})}{\partial \boldsymbol{\theta}} f(\mathbf{y};\boldsymbol{\theta})\mathbf{dy} + \int \mathbf{g}(\boldsymbol{\theta};\mathbf{y}) \frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{dy} = 0 \quad .$$

If we rewrite the second term using $\frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = f(\mathbf{y};\boldsymbol{\theta}) \frac{\partial \log f(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, and simplify by using (2.9) and (2.11), we find

$$\mathbf{D}_{\mathbf{g}}(\boldsymbol{\theta}) + \mathbf{N}(\boldsymbol{\theta}) = 0 \Rightarrow \mathbf{D}_{\mathbf{g}}(\boldsymbol{\theta}) = -\mathbf{N}(\boldsymbol{\theta}) \tag{2.12}$$

Now let $\mathbf{u} = (u_1, \ldots, u_k)$ and $\mathbf{v} = (v_1, \ldots, v_k)$ be two arbitrary real vectors. Applying the Cauchy-Schwarz inequality to, $\mathbf{u}^T\mathbf{S}(\boldsymbol{\theta})$ and $\mathbf{v}^T\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})$ we obtain (Chandrasekar and Kale, 1984),

$$\left(\mathbf{u}^T\mathbf{N}(\boldsymbol{\theta})^T\mathbf{v}\right)^2 \leq \left(\mathbf{u}^T\mathbf{J}(\boldsymbol{\theta})\mathbf{u}\right)\left(\mathbf{v}^T\mathbf{V}(\mathbf{g}(\boldsymbol{\theta};\mathbf{Y}))\mathbf{v}\right)$$

where $\mathbf{J}(\boldsymbol{\theta})$ is the information matrix and $\mathbf{V}(\mathbf{g}(\boldsymbol{\theta};\mathbf{y}))$ is the covariance matrix, defined in (2.7).

By taking the particular vector $\mathbf{u} = \mathbf{J}^{-1}(\boldsymbol{\theta})N^T(\boldsymbol{\theta})\mathbf{v}$

$$\left(\mathbf{v}^T\mathbf{N}(\boldsymbol{\theta})\mathbf{J}^{-1}(\boldsymbol{\theta})\mathbf{N}(\boldsymbol{\theta})^T\mathbf{v}\right)^2 \leq \left(\mathbf{v}^T\mathbf{N}(\boldsymbol{\theta})\mathbf{J}^{-1}(\boldsymbol{\theta})\mathbf{N}(\boldsymbol{\theta})^T\mathbf{v}\right)\left(\mathbf{v}^T\mathbf{V}(\mathbf{g}(\boldsymbol{\theta};\mathbf{Y}))\mathbf{v}\right)$$

Since $\mathbf{J}(\boldsymbol{\theta})$ is positive definite the term $\left(\mathbf{v}^T\mathbf{N}(\boldsymbol{\theta})\mathbf{J}^{-1}(\boldsymbol{\theta})\mathbf{N}(\boldsymbol{\theta})^T\mathbf{v}\right)$ is either positive or zero. In the first case we can divide both sides of the inequality by this term yielding

$$\mathbf{v}^T\mathbf{N}(\boldsymbol{\theta})\mathbf{J}^{-1}(\boldsymbol{\theta})\mathbf{N}(\boldsymbol{\theta})^T\mathbf{v} \leq \mathbf{v}^T\mathbf{V}(\mathbf{g}(\boldsymbol{\theta};\mathbf{Y}))\mathbf{v}. \tag{2.13}$$

Although we cannot perform the same step if $\left(\mathbf{v}^T\mathbf{N}(\boldsymbol{\theta})\mathbf{J}^{-1}(\boldsymbol{\theta})\mathbf{N}(\boldsymbol{\theta})^T v\right)$ is zero, (2.13) still holds because $V[\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})]$ is positive definite by definition of $G_p^{(s)}$. Therefore (2.13) is true for all $\mathbf{v} \in R^k$ and $\boldsymbol{\theta} \in \Theta$, which means that the matrix

$$\mathbf{V}(\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})) - \mathbf{N}(\boldsymbol{\theta})\mathbf{J}^{-1}(\boldsymbol{\theta})\mathbf{N}(\boldsymbol{\theta})^T \tag{2.14}$$

is positive definite or semi-definite. In fact by plugging (2.12) in (2.14) we obtain

$$\mathbf{V}(\mathbf{g}(\boldsymbol{\theta};\mathbf{Y})) - \mathbf{D}_{\mathbf{g}}\mathbf{J}^{-1}(\boldsymbol{\theta})\mathbf{D}_{\mathbf{g}}{}^T$$

which is also positive definite or semi-definite. Furthermore if we consider the class of all standardized estimating functions as defined in (2.10), meaning that $\mathbf{D}_{\mathbf{g}_s} = I$ and $\mathbf{V}(\mathbf{g}_s(\boldsymbol{\theta};\mathbf{Y})) - \mathbf{J}^{-1}(\boldsymbol{\theta})$ is positive definite or semi-definite, we can state that the inverse of the information matrix, $\mathbf{J}^{-1}(\boldsymbol{\theta})$, is a minimal variance matrix for the class of standardized regular estimating functions. This result is equivalent to the Cramér-Rao lower bound for unbiased estimators but applied to unbiased estimating functions.

2.4 Consistency and asymptotic distribution

In this section we will present some sufficient conditions for consistency of the EF estimator, and use the result on consistency to present some results on its limiting distribution. In order to show that the estimators $\hat{\boldsymbol{\theta}}$ defined as the solution of an estimating equation are consistent estimators, we need to impose further assumptions. Let

$$\mathbf{g}_n^*(\boldsymbol{\theta};\mathbf{y}_n) = \eta_n \mathbf{g}(\boldsymbol{\theta};\mathbf{y}_n)$$

be a sequence of estimating function resulting from considering the increasing sequence $\ldots \mathbf{y}_{n-1} \subseteq \mathbf{y}_n \subseteq \mathbf{y}_{n+1} \ldots$ where η_n is a matrix that does not depend on $\boldsymbol{\theta}$ and converges as $n \to \infty$ to a matrix of constants. In this case $\mathbf{g}_n^*(\boldsymbol{\theta};\mathbf{y}_n)$ is called a normalised estimating function. The first of these assumptions is existence of a normalizing matrix

η_n such that $\mathbf{g}_n^*(\boldsymbol{\theta}; \mathbf{y}_n)$ converges uniformly in probability to a limiting deterministic function that has a unique root at $\boldsymbol{\theta}_0$ This is formalized as

$$\sup_{\boldsymbol{\theta}\in\Theta} |\mathbf{g}_n^*(\boldsymbol{\theta}; \mathbf{Y}_n) - \mathbf{g}_\ell(\boldsymbol{\theta})| \xrightarrow{p} \mathbf{0} \ ,$$

$$\lim_{n\to\infty} P[\sup_{\boldsymbol{\theta}\in\Theta} |\mathbf{g}_n^*(\boldsymbol{\theta}; \mathbf{Y}_n) - \mathbf{g}_\ell(\boldsymbol{\theta})| < \boldsymbol{\epsilon}] = 1 \ \ , \text{ for any } \boldsymbol{\theta} \in \Theta, \boldsymbol{\epsilon} > 0$$

which means that the difference between $\mathbf{g}_n^*(\boldsymbol{\theta}; \mathbf{y}_n)$ and $\mathbf{g}_\ell(\boldsymbol{\theta})$ disappears as $n \to \infty$. Note that for $\hat{\boldsymbol{\theta}}_n$, defined as the unique solution in Θ of the equation $\mathbf{g}_n^*(\hat{\boldsymbol{\theta}}_n; \mathbf{Y}_n) = 0$, we have

$$|\mathbf{g}_n^*(\hat{\boldsymbol{\theta}}_n; \mathbf{Y}_n) - \mathbf{g}_\ell(\hat{\boldsymbol{\theta}}_n)| \leq \sup_{\boldsymbol{\theta}\in\Theta} |\mathbf{g}_n^*(\boldsymbol{\theta}; \mathbf{Y}_n) - \mathbf{g}_\ell(\boldsymbol{\theta})| \ \ .$$

So

$$\begin{aligned} \lim_{n\to\infty} P[|\mathbf{g}_\ell(\hat{\boldsymbol{\theta}}_n)| < \boldsymbol{\epsilon}] = 1 \ \ , \boldsymbol{\epsilon} > \mathbf{0} \ . \\ \lim_{n\to\infty} P[-\boldsymbol{\epsilon} < \mathbf{g}_\ell(\hat{\boldsymbol{\theta}}_n) < \boldsymbol{\epsilon}] = 1 \ \ , \boldsymbol{\epsilon} > \mathbf{0} \end{aligned} \tag{2.15}$$

For the remaining of the proof we make the assumption that the parameter space Θ is compact. This follows Hall (2005). However alternative proofs replace this by some conditions on $\mathbf{g}_\ell(\boldsymbol{\theta})$, see Jesus and Chandler (2011).

Define N as any open subset of Θ containing $\boldsymbol{\theta}_0$, and N^c the complement of N relative to Θ. This implies that N^c is a closed subset of a compact set, therefore compact; since $\mathbf{g}_\ell(\boldsymbol{\theta})$ is continuous and nonzero except at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, it must be that $\inf_{\boldsymbol{\theta}\in N_c} |\mathbf{g}_\ell(\boldsymbol{\theta})| > 0$. Using this quantity to replace $\boldsymbol{\epsilon}$ in (2.15)

$$\lim_{n\to\infty} P[-\inf_{\boldsymbol{\theta}\in N_c} |\mathbf{g}_\ell(\boldsymbol{\theta})| < \mathbf{g}_\ell(\hat{\boldsymbol{\theta}}_n) < \inf_{\boldsymbol{\theta}\in N_c} |\mathbf{g}_\ell(\boldsymbol{\theta})|] = 1 \ .$$

This means that as $n \to \infty$, $\hat{\boldsymbol{\theta}}_n$ cannot belong to N^c:

$$\begin{aligned} \lim_{n\to\infty} P[\hat{\boldsymbol{\theta}}_n \notin N^c] &= 1 \ , \text{ and hence} \\ \lim_{n\to\infty} P[\hat{\boldsymbol{\theta}}_n \in N] &= 1 \ . \end{aligned}$$

We can set N to be any neighbourhood of $\boldsymbol{\theta}_0$ obtaining:

$$\lim_{n\to\infty} P[|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0| < \boldsymbol{\epsilon}] = 1 \ , \text{ for any } \boldsymbol{\epsilon} > 0 \ .$$

Therefore $\hat{\boldsymbol{\theta}}_n$, defined as the estimator resulting from the estimating function $g_n(\boldsymbol{\theta}; \mathbf{y})$, is a consistent estimator of $\boldsymbol{\theta}$, under all the assumptions added in this section.

The remainder of this section is used to discuss the asymptotic distribution of the estimator. Define,

$$\tilde{\mathbf{g}}_n(\boldsymbol{\theta}; \mathbf{y}) = \gamma_n \mathbf{g}_n(\boldsymbol{\theta}; \mathbf{y})$$

and assume the existence of (γ_n), a sequence of p x p normalizing matrices that do not depend on $\boldsymbol{\theta}$ that are such that,

$$\lim_{n\to\infty} \mathbf{V}(\tilde{\mathbf{g}}_n(\boldsymbol{\theta}; \mathbf{Y})) = \boldsymbol{\Sigma}(\boldsymbol{\theta}) \ , \tag{2.16}$$

and

$$\tilde{\mathbf{g}}_n(\boldsymbol{\theta}; \mathbf{y})) \xrightarrow{d} g_\infty$$

where g_∞ is a random variable with mean zero and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Note that $\tilde{\mathbf{g}}_n(\boldsymbol{\theta}; \mathbf{y})$ retains all the relevant properties of $\mathbf{g}_n(\boldsymbol{\theta}; \mathbf{y})$ as an estimating function, including the solution for the estimating equation. From (2.4) we can write, using the mean value theorem,

$$\tilde{\mathbf{g}}_n(\hat{\boldsymbol{\theta}}_n;\mathbf{y}) = \tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y}) + \left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_n^\dagger}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \tag{2.17}$$

for some $\boldsymbol{\theta}_n^\dagger : |\boldsymbol{\theta}_n^\dagger - \boldsymbol{\theta}_0| < |\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|$. The left-hand side of (2.17) is zero by definition, and re-arranging we obtain

$$(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]^{-1}_{\boldsymbol{\theta}_n^\dagger}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y}) \ . \tag{2.18}$$

Assume there is a sequence, $\boldsymbol{\delta}_n$ of p x p invertible matrices such that

$$\left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]\boldsymbol{\delta}_n \xrightarrow{p} \mathbf{M}(\boldsymbol{\theta}) \tag{2.19}$$

for any $\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < c$ where c is a positive constant and $\mathbf{M}(\boldsymbol{\theta})$ is an invertible positive definite matrix. Assume also that for any sequence $\boldsymbol{\psi}_n$ with $\lim_{n\to\infty}\boldsymbol{\psi}_n = \boldsymbol{\theta}_0$,

$$\left(\left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\psi}_n} - \left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\right)\boldsymbol{\delta}_n \xrightarrow{p} 0 \ . \tag{2.20}$$

If we pre-multiply both sides of (2.18) by $\boldsymbol{\delta}_n^{-1}$

$$\begin{aligned}
\boldsymbol{\delta}_n^{-1}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &= -\boldsymbol{\delta}_n^{-1}\left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]^{-1}_{\boldsymbol{\theta}_n^\dagger}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y}) \\
&= -\left(\left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_n^\dagger}\boldsymbol{\delta}_n\right)^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y}) \\
&= -\left[\left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\boldsymbol{\delta}_n + \left(\left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_n^\dagger} - \left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\right)\boldsymbol{\delta}_n\right]^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y}) \\
&= -\left[\left[\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\boldsymbol{\delta}_n + o_p(1)\right]^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y})
\end{aligned}$$

where this last step is due to (2.20) coupled with consistency of $\hat{\boldsymbol{\theta}}_n$. Therefore, we have

$$\boldsymbol{\delta}_n^{-1}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} -\mathbf{M}_0^{-1} g_\infty \tag{2.21}$$

where $\mathbf{M}_0 = \mathbf{M}(\boldsymbol{\theta}_0)$. A common situation is the case where the estimating function converges in distribution to the multivariate normal distribution,

$$\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0; \mathbf{y}) \xrightarrow{d} MVN(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.22}$$

giving,

$$\boldsymbol{\delta}_n^{-1}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} MVN(\mathbf{0}, \mathbf{M}_0^{-1}\boldsymbol{\Sigma}\mathbf{M}_0^{-T}) \quad . \tag{2.23}$$

In some of the examples in the next section, the properties (2.19) and (2.22) can be directly deduced from the way the estimating functions $g_n(\boldsymbol{\theta}; \mathbf{y})$ are built. For example (2.22) is usually shown by invoking the central limit theorem.

Having an approximate distribution for the estimator allows us to calculate confidence intervals or confidence regions. The simple approach is to treat each element of the estimator vector as having a marginal normal distribution, in this case confidence intervals for the individual parameters can be constructed by selecting the appropriate percentile of the normal distribution. Another possibility is described in some detail in Jesus and Chandler (2011), and it can be used when the estimating function is a gradient vector,

$$g_n(\boldsymbol{\theta}; \mathbf{y}) = \frac{\partial Q_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}}$$

where $Q_n(\boldsymbol{\theta}; \mathbf{y})$ is a differentiable function that is minimized to obtain the estimator $\hat{\boldsymbol{\theta}}$. The result in Jesus and Chandler (2011) states that a simultaneous confidence region at the level $(1-\alpha)\%$ consists of the $\boldsymbol{\theta}$ such that,

$$a^{-1}\left[2\left(Q_n(\boldsymbol{\theta}; \mathbf{y}) - Q_n(\hat{\boldsymbol{\theta}}; \mathbf{y})\right) - c\right] \tag{2.24}$$

is less than the $(1-\alpha)^{\text{th}}$ percentile of the χ^2_b distribution where

$$a = \frac{|\kappa_3|}{4\kappa_4}, \quad b = \frac{8\kappa_2^3}{\kappa_3^2}, \quad c = \kappa_1 - ab,$$

and

$$\kappa_r = 2^{r-1}(r-1)!\text{tr}\left[\left(\mathbf{M}_0^{-1}\boldsymbol{\Sigma}\right)^r\right] \ .$$

2.5 Examples - Likelihood as estimating function

2.5.1 Maximum Likelihood Estimator

The maximum likelihood method for parametric estimation provides a way of building the function $\mathbf{g}(\boldsymbol{\theta};\mathbf{y})$. Define the likelihood function for $\boldsymbol{\theta}$ given y as

$$L(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta})$$

where, as before, $\mathbf{y}$ is the vector of observations, with density $f(\mathbf{y};\boldsymbol{\theta})$ and $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)$ is the parameter of interest. The likelihood function becomes the objective function, and our optimum is the maximum, which is attained for the same $\boldsymbol{\theta}$ as the log-likelihood

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \log\ L(\boldsymbol{\theta}|\mathbf{y}) \ .$$

The vector-valued function $\mathbf{g}(\boldsymbol{\theta};\mathbf{y})$ is formed by differentiating $l(\boldsymbol{\theta}|\mathbf{y})$ with respect to $\boldsymbol{\theta}$, this function is called the score function and is denoted as

$$\mathbf{g}(\boldsymbol{\theta};\mathbf{y}) = \frac{\partial\ell(\boldsymbol{\theta}|\mathbf{y})}{\partial\boldsymbol{\theta}} \ .$$

To determine whether the score function satisfies condition (2.1), we can take expectations. Consider $E_{\boldsymbol{\theta}_0}[\mathbf{g}(\boldsymbol{\theta}_0;\mathbf{Y})]$ which has i^{th} component

$$\int \frac{1}{f(\mathbf{y};\boldsymbol{\theta}_0)} \left.\frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} f(\mathbf{y};\boldsymbol{\theta}_0)\mathbf{dy} = \int \left.\frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \mathbf{dy}$$

Provided that the order of the integration and differentiation can be changed, we have

$$\frac{\partial}{\partial\theta_i}\int f(\mathbf{y};\boldsymbol{\theta})\mathbf{dy}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 0 \tag{2.25}$$

so that (2.1) does indeed hold.

The maximum likelihood estimator (MLE) is a particular case where the covariance matrix of the estimating function can be computed at the true parameter value using the expected value of its partial derivatives. The result is shown in two steps as follows, first we write the general form of the partial derivative of the i^{th} component of the score in order of the j^{th} component of the parameter vector, in terms of the partial derivatives of $f(\mathbf{y};\boldsymbol{\theta})$,

$$\frac{\partial g_i(\boldsymbol{\theta}|\mathbf{y})}{\partial\theta_j} = \frac{\partial^2 \log f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} = \frac{1}{f(\mathbf{y};\boldsymbol{\theta})}\frac{\partial^2 f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} - \frac{1}{f(\mathbf{y};\boldsymbol{\theta})^2}\frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_j}$$

Secondly we take expectations,

$$\begin{aligned}
E\left[\frac{\partial g_i(\boldsymbol{\theta};\mathbf{y})}{\partial\theta_j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] &= \int \frac{1}{f(\mathbf{y};\boldsymbol{\theta}_0)}\frac{\partial^2 f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} f(\mathbf{y};\boldsymbol{\theta}_0)\mathbf{dy} \\
&\quad - \int \frac{1}{f(\mathbf{y};\boldsymbol{\theta}_0)^2}\frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_i}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} f(\mathbf{y};\boldsymbol{\theta}_0)\mathbf{dy} \\
&= \frac{\partial^2}{\partial\theta_i\partial\theta_j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\int f(\mathbf{y};\boldsymbol{\theta}_0)\mathbf{dy} \\
&\quad - \int \frac{\partial \log f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_i}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\frac{\partial \log f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} f(\mathbf{y};\boldsymbol{\theta}_0)\mathbf{dy} \\
&= 0 - E\left[g_i(\boldsymbol{\theta}_0;\mathbf{y})g_j(\boldsymbol{\theta}_0;\mathbf{y})\right]
\end{aligned} \tag{2.26}$$

where the last step is similar to (2.25).

Using this result the covariance matrix of the score vector can be written in terms of its partial derivatives,

$$V[\mathbf{g}(\boldsymbol{\theta}_0;\mathbf{Y})] = E[\mathbf{g}(\boldsymbol{\theta}_0;\mathbf{Y})\mathbf{g}(\boldsymbol{\theta}_0;\mathbf{Y})^T] = -E\left[\frac{\partial \mathbf{g}(\boldsymbol{\theta};\mathbf{Y})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = -E[\mathbf{H}_{\boldsymbol{\theta}_0}]$$

where $\mathbf{H}_{\boldsymbol{\theta}}$ is the Hessian matrix of second derivatives of the log-likelihood. If all the diagonal elements of $\mathbf{H}_{\boldsymbol{\theta}}$ have non-zero expectation then conditions (2.3) and (2.4) are satisfied. Furthermore by noting that the standardized version of this particular

estimating equation at the true parameter value is, from (2.10),

$$\mathbf{g}_s(\boldsymbol{\theta}_0|\mathbf{Y}) = E[\mathbf{H}_{\boldsymbol{\theta}_0}]^{-1}\mathbf{g}(\boldsymbol{\theta}_0|\mathbf{Y})$$

$$V[\mathbf{g}_s(\boldsymbol{\theta}_0|\mathbf{Y})] = E[\mathbf{H}_{\boldsymbol{\theta}_0}]^{-1}V[\mathbf{g}(\boldsymbol{\theta}_0|\mathbf{Y})](E[\mathbf{H}_{\boldsymbol{\theta}_0}]^{-1})^T = -E[\mathbf{H}_{\boldsymbol{\theta}_0}]^{-1}$$

$$= E\left[\left.\frac{\partial \log f}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \left.\frac{\partial \log f}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]^{-1}, \quad \text{from (2.26)}$$

it is shown that the score function is an estimating function that achieves the minimal asymptotic covariance matrix $\mathbf{J}^{-1}$.

In the likelihood case, the multivariate normality of the estimator can be justified by application of the central limit theorem when observations are independent. Even if the observations are not independent multivariate normality can hold under mild regularity conditions (Sweeting, 1990). This means that for large samples

$$\mathbf{g}(\boldsymbol{\theta}_0; \mathbf{Y}) \sim MVN(\mathbf{0}, -E[\mathbf{H}_{\boldsymbol{\theta}_0}])$$

From the distribution of the score we can obtain the distribution of the estimator itself. Comparison with (2.19) shows that we need to assume the existence of a sequence $(\boldsymbol{\delta}_n)$ such that as $n \to \infty$

$$\left[\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0} \boldsymbol{\delta}_n \xrightarrow{p} \mathbf{M}(\boldsymbol{\theta}_0)$$

where $\mathbf{M}$ is a matrix of constants. We can use the results in (2.19-2.23) to obtain the approximation for large n,

$$\boldsymbol{\delta}_n^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim MVN(\mathbf{0}, -E[\mathbf{H}_{\boldsymbol{\theta}_0}]^{-1})$$

2.5.2 Marginal and Conditional Likelihood

It is not uncommon for situations where it is difficult to write down a full likelihood function, or even where its expression is known the existence of many nuisance parameters make it difficult to maximize. In some of these cases it is possible to find a transformation of the data such that its likelihood does not depend on nuisance parameters. Suppose that the parameter vector can be partitioned as $\boldsymbol{\theta} = (\psi, \lambda)$, such that the likelihood can be factorized in the following way (Northrop, 2006)

$$L(\psi, \lambda|\mathbf{y}) = f_y(\mathbf{y}; \boldsymbol{\theta}) = f_z(z(\mathbf{y}); \psi) f_x(x(\mathbf{y})|z; \psi, \lambda) \tag{2.27}$$

From (2.27), and as long as there is a transformation $z(\mathbf{y})$ that does not depend on the parameters, we can use $z(\mathbf{y})$ to perform likelihood-based inference for the parameter vector ψ without having to estimate parameters in λ. Thus the marginal log-likelihood becomes,

$$\ell_z(\psi|z) = \log(f_z(z(\mathbf{y}); \psi))$$

The score function resulting from the likelihood $\ell_z(\psi|z)$, is a regular unbiased estimating function for ψ with covariance matrix achieving the lower bound for the class of estimating functions based on z. This lower bound is $\mathbf{J}_z(\psi)^{-1}$, where

$$\mathbf{J}_z(\psi) = -E\left[\frac{\partial^2 \log f_z(Z; \psi)}{\partial\psi\partial\psi}\right]$$

with $\mathbf{Z} = z(\mathbf{Y})$. Intuitively, the information matrix $\mathbf{J}_z(\psi)$ should be smaller (meaning less information) than the one based on f_y. The difference between the two depends on $\partial \log(f_x)/\partial\psi$ as we see next

$$\begin{aligned}\mathbf{J}_y(\psi) &= -E\left[\frac{\partial^2 \log f_y(\mathbf{Y}; \psi, \lambda)}{\partial\psi^2}\right] \\ &= -E\left[\frac{\partial^2 \log f_z(\mathbf{Z}; \psi)}{\partial\psi^2} + \frac{\partial^2 \log f_x(\mathbf{X}; \psi, \lambda)}{\partial\psi^2}\right]\end{aligned}$$

and therefore we obtain for a fixed value of λ,

$$\mathbf{J}_y(\psi) = \mathbf{J}_z(\psi) + \mathbf{J}_x(\psi) \tag{2.28}$$

where

$$\mathbf{X} \;=\; z(\mathbf{X})$$

which shows that if $E\left[\frac{\partial^2 \log f_x(x;\psi)}{\partial \psi^2}\right] = 0$, then $\mathbf{J}_y(\psi) = \mathbf{J}_z(\psi)$. Note that the density of X may depend on ψ and therefore we are not making full use of the information available about the parameter. The main problem of this method is that by ignoring the data x, we may be losing information about the parameter of interest. Because in general, $E\left[\frac{\partial^2 \log f_x(x;\psi)}{\partial \psi^2}\right] = 0$ is only zero if the density of X does not depend on ψ; thus if $f_x(x;\psi,\lambda)$ does depend on ψ we get $\mathbf{J}_y(\psi) > \mathbf{J}_z(\psi)$ and inference based on $z(\mathbf{y})$ is not fully efficient.

A similar situation may arise when the factorization, instead of (2.27), has the form,

$$L(\psi,\lambda|\mathbf{y}) = f_y(\mathbf{y};\boldsymbol{\theta}) = f_x(x(\mathbf{y});\psi,\lambda) f_z(z(\mathbf{y})|x;\psi)$$

and in this case

$$\ell_z(\psi|z) = \log(f_z(z(\mathbf{y})|x;\psi))$$

is called the conditional log-likelihood. The analogue of (2.28) also applies to the conditional likelihood, so that in general the conditional likelihood is not fully efficient. Both the marginal and conditional likelihood approaches provide a way of building estimating functions, in fact they can be treated in exactly the same way as the full likelihood although based on a partition of the data.

2.6 Summary

In this chapter we described the theory of estimating functions focusing on their asymptotic properties and how these are related to the properties of the estimator. We have shown that under mild conditions the estimating function estimator is consistent and is asymptotically normally distributed. We also explored the concept of optimal estimating function within a certain class. In the description of the estimator several assumptions were introduced, which may leave the impression that these assumptions restrict the class of problems to which this theory can be applied. However most of the assumptions are standard in the inference literature and may even be relaxed or replaced in specific inference problems. Hence, the results from this chapter are applicable to a wide class of

estimation problems including the particular field of moment based inference, this will be pursued in the following chapter. In a later chapter we will apply this theory to an estimator based on the frequency domain of the data. We have also shown two examples of how the estimating function may arise based on likelihood theory. We conclude that given a set of estimating functions, and potentially regardless of how they were derived, it is possible to use the results in this section to comment on consistency and even approximate on the asymptotic distribution of the estimator resulting from such estimating functions. These results are useful to obtain some measure of uncertainty about the estimator in the form of confidence intervals or regions.

Chapter 3

Generalized Method of Moments

In this chapter we explore the relationships between the estimating functions framework described in the previous chapter and the Generalized Method of Moments widely used in econometrics. In the first part of this chapter we describe the GMM method and show that it can be seen as a way of building estimating functions. We translate the sufficient conditions for consistency and asymptotic distribution of the estimator from the previous chapter into the GMM framework. In section 3.2 we revisit the optimality concepts of section 2.3, and by exploring some of the specific properties of the estimating functions built from the GMM method, we present a lower bound for the variance of this particular class of estimating function.

3.1 Generalized method of moments as estimating functions

In the econometrics literature the application of estimating functions is often referred to as the Generalized Method of Moments (Hansen, 1982; Bera and Bilias, 2002; Hall, 2005). This comes from the fact that one possible way of building the estimating function $g(\boldsymbol{\theta}; \mathbf{y})$ is by minimizing the distance between some theoretical properties of the distribution of y and their sample/observed counterparts. These properties are usually formalized in terms of moments of the distribution, and are called *population moment conditions*. For example, we might have a model which specifies

$$E[\mathbf{h}_n(\boldsymbol{\theta}; \mathbf{y})] = \mathbf{0} \tag{3.1}$$

where $\mathbf{h}_n(\boldsymbol{\theta};\mathbf{y})$ is a k-vector valued function depending on the data, $\mathbf{y}$, and the parameter of interest, $\boldsymbol{\theta}$. Here we allow $k > p$, which was not the case in the previous chapter. This situation often arises when considering a vector $\mathbf{y}$ of k summary statistics considered as the realization of a $k \times 1$ random vector $T(\mathbf{Y})$. If we denote by $\tau(\boldsymbol{\theta})$ its expectation $E_{\boldsymbol{\theta}}[T(\mathbf{Y})]$ as a function of $\boldsymbol{\theta}$, then (3.1) can be obtained by putting $\mathbf{h}_n(\boldsymbol{\theta};\mathbf{y}) = T(\mathbf{y}) - \tau(\boldsymbol{\theta})$.

We now state some assumptions on the asymptotic behaviour of $\mathbf{h}_n(\boldsymbol{\theta};\mathbf{y})$ which will be shown equivalent to the assumptions made on the previous section on a generic estimating function. Define,

$$\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y}) = \gamma_n \mathbf{h}_n(\boldsymbol{\theta};\mathbf{y})$$

where γ_n is a k x k normalizing matrix that does not depend on $\boldsymbol{\theta}$ and is such that,

$$\lim_{n\to\infty} \mathrm{Var}(\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})) = \mathbf{S}(\boldsymbol{\theta}) \tag{3.2}$$

$$\lim_{n\to\infty} E(\tilde{\mathbf{h}}_n(\boldsymbol{\theta}_0;\mathbf{y})) = \mathbf{0}$$

$$\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y}) \xrightarrow{d} \tilde{\mathbf{h}}_\infty(\boldsymbol{\theta}) \tag{3.3}$$

where $\mathbf{S}(\boldsymbol{\theta})$ is the limiting covariance matrix of $\tilde{\mathbf{h}}_\infty(\boldsymbol{\theta})$ and is finite positive definite matrix for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. This condition is equivalent to 2.16.

Assume also that there is a sequence of $p \times p$ invertible matrices, $\boldsymbol{\delta}_n$, such that for some $c > 0$,

$$\lim_{n\to\infty} \boldsymbol{\delta}_n = \mathbf{0} \tag{3.4}$$

$$\frac{\partial \tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial \boldsymbol{\theta}} \boldsymbol{\delta}_n \xrightarrow{p} \mathbf{H}(\boldsymbol{\theta}) \tag{3.5}$$

for $\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < c$, where $\mathbf{H}(\boldsymbol{\theta})$ is an invertible positive definite matrix. Condition 3.4 is more restrictive than 2.19. Define

$$\boldsymbol{\Omega}_n^{(r)}(\boldsymbol{\theta}) = \frac{\partial^2 \tilde{h}_r(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$$

where $\tilde{h}_r(\boldsymbol{\theta}; \mathbf{y})$ denotes the r^{th} element of $\tilde{\mathbf{h}}_n(\boldsymbol{\theta}; \mathbf{y})$; we omit the n for convenience of notation. We will assume that

$$\sup_{\boldsymbol{\theta}:|\boldsymbol{\theta}-\boldsymbol{\theta}_0|<c} \left| \Omega_n^{(r)}(\boldsymbol{\theta})\boldsymbol{\delta}_n - \Omega_n^{(r)}(\boldsymbol{\theta_0}) \right| \to \mathbf{0} \quad . \tag{3.6}$$

The final assumption stated here is that for any sequence $\boldsymbol{\psi}_n$ with $\boldsymbol{\psi}_n \to \boldsymbol{\theta}_0$

$$\left(\left[\frac{\partial \tilde{\mathbf{h}}_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\psi}_n} - \left[\frac{\partial \tilde{\mathbf{h}}_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\theta}_0} \right) \boldsymbol{\delta}_n \xrightarrow{p} 0 \quad . \tag{3.7}$$

The next step in the method consists of building a p-dimensional estimating function from the k-dimensional moment condition, that can be used in estimation. This is done by minimizing a quadratic form with respect to the parameter $\boldsymbol{\theta}$ (Hall, 2005):

$$Q(\boldsymbol{\theta}) = \tilde{\mathbf{h}}_n(\boldsymbol{\theta}; \mathbf{y})^T \ \mathbf{W}_n \ \tilde{\mathbf{h}}_n(\boldsymbol{\theta}; \mathbf{y}) \quad , \tag{3.8}$$

where $\mathbf{W}_n$ is a positive semi-definite $k \times k$ matrix which may depend on the data but converges in probability to a positive definite matrix of constants $\mathbf{W}$, say. In many applications (Wheater et al., 2005) $\mathbf{W}_n$ is chosen to be a diagonal matrix, so that (3.8) is a weighted sum of squares of elements of $\tilde{\mathbf{h}}_n(\boldsymbol{\theta}; \mathbf{y})$; in this case the minimiser of (3.8) can be called a weighted least squares estimator. Strictly speaking weighted least squares estimation refers to the case where the moment conditions are defined as $h_i(\boldsymbol{\theta}) = h(\boldsymbol{\theta}, y_i)$ and the h_i are uncorrelated and have different variances, the diagonal elements of $\mathbf{W}_n$ are inversely proportional to these variances. It is clear that $Q(\boldsymbol{\theta}) \geq 0$

and if $\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y}) = \mathbf{0}$ then $Q(\boldsymbol{\theta}) = 0$. The first order condition to solve for the minimum becomes,

$$\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]^T_{|\hat{\boldsymbol{\theta}}} \mathbf{W}_n\ \tilde{\mathbf{h}}_n(\hat{\boldsymbol{\theta}};\mathbf{y}) = \mathbf{0} \quad ,$$

so that the estimating function is

$$\mathbf{g}_n(\boldsymbol{\theta};\mathbf{y};\mathbf{W}_n) = \left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]^T \mathbf{W}_n\ \tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y}) \quad .$$

The theory of Chapter 2 can now be applied to the regular estimating functions $\mathbf{g}_n(\boldsymbol{\theta};\mathbf{y};\mathbf{W}_n)$, including the limiting distribution of the resulting estimator under suitable normalization. We now show that the assumptions required by that theory are satisfied under the previous conditions on $\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})$.

If we define

$$\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y};\mathbf{W}_n) = \boldsymbol{\delta}_n^T\mathbf{g}_n(\boldsymbol{\theta};\mathbf{y};\mathbf{W}_n) = \left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\boldsymbol{\delta}_n\right]^T \mathbf{W}_n\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y}) \tag{3.9}$$

then, from (3.2) and (3.5)

$$\lim_{n\to\infty} \mathrm{Var}(\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y};\mathbf{W}_n)) \quad = \quad \Sigma(\boldsymbol{\theta}) = \mathbf{H}^T(\boldsymbol{\theta})\mathbf{W}\mathbf{S}(\boldsymbol{\theta})\mathbf{W}^T\mathbf{H}(\boldsymbol{\theta})$$

which means that (2.16) holds. To establish (2.19) and (2.20) is more difficult. We start by establishing (2.19). Note that

$$\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y};\mathbf{W}_n)}{\partial\boldsymbol{\theta}} = \boldsymbol{\delta}_n^T\left(\boldsymbol{\Lambda}_n + \left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]^T \mathbf{W}_n\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]\right)$$

where $\mathbf{\Lambda}_n$ is a $p \times p$ matrix with $(i,j)^{th}$ element

$$\begin{aligned}\Lambda_{(i,j)} &= \sum_{r=1}^{k}\sum_{s=1}^{k} W_{rs}\frac{\partial^2 \tilde{h}_r(\boldsymbol{\theta};\mathbf{y})}{\partial\theta_i\partial\theta_j}\tilde{h}_s(\boldsymbol{\theta};\mathbf{y}) \\ &= \sum_{r=1}^{k}\sum_{s=1}^{k} W_{rs}\Omega_{i,j}^{(r)}(\boldsymbol{\theta})\tilde{h}_s(\boldsymbol{\theta};\mathbf{y})\end{aligned}$$

Next, consider the matrix

$$\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y};\mathbf{W}_n)}{\partial\boldsymbol{\theta}}\boldsymbol{\delta}_n = \left(\boldsymbol{\delta}_n^T\mathbf{\Lambda}_n\boldsymbol{\delta}_n + \left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\boldsymbol{\delta}_n\right]^T\mathbf{W}_n\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]\boldsymbol{\delta}_n\right) \tag{3.10}$$

where $\boldsymbol{\delta}_n^T\mathbf{\Lambda}_n\boldsymbol{\delta}_n$ has $(u,v)^{th}$ element

$$\sum_{i=1}^{p}\sum_{j=1}^{p}\delta_{iu}\delta_{vj}\Lambda_{i,j} = \sum_{r=1}^{k}\sum_{s=1}^{k}W_{rs}\tilde{h}_s(y;\boldsymbol{\theta})\sum_{i=1}^{p}\sum_{j=1}^{p}\delta_{iu}\delta_{vj}\Omega_{i,j}^{(r)}(\boldsymbol{\theta})$$

Note that $\sum_{i=1}^{p}\sum_{j=1}^{p}\delta_{iu}\delta_{vj}\Omega_{i,j}^{(r)}(\boldsymbol{\theta})$ is the $(u,v)^{th}$ element of $\boldsymbol{\delta}_n^T\mathbf{\Omega}_n^{(r)}(\boldsymbol{\theta})\boldsymbol{\delta}_n$, which tends to zero in a neighbourhood of $\boldsymbol{\theta}_0$ from (3.4) and (3.6). From (3.3) we have that $\sum_{r=1}^{k}\sum_{s=1}^{k}W_{rs}\tilde{h}_s(\boldsymbol{\theta};\mathbf{y})$ converges to a random variable with expectation zero and finite variance and we can state that $\boldsymbol{\delta}_n^T\mathbf{\Lambda}_n\boldsymbol{\delta}_n \xrightarrow{p} 0$. The second part of the right hand side of (3.10) converges in probability to $\mathbf{H}(\boldsymbol{\theta})^T\mathbf{W}\mathbf{H}(\boldsymbol{\theta})$, from (3.5) and because the sequence $\mathbf{W}_n$ converges to a matrix of constants $\mathbf{W}$, and we can write,

$$\frac{\partial\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y};\mathbf{W}_n)}{\partial\boldsymbol{\theta}}\boldsymbol{\delta}_n \xrightarrow{p} \mathbf{H}(\boldsymbol{\theta})^T\mathbf{W}\mathbf{H}(\boldsymbol{\theta}) \tag{3.11}$$

which means that (2.19) holds with $\mathbf{M}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta})^T\mathbf{W}\mathbf{H}(\boldsymbol{\theta})$.

We turn now to the condition on continuity (2.20), therefore we are interested in studying the behaviour of

$$\left(\left[\frac{\partial \tilde{\mathbf{g}}_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\psi}_n} - \left[\frac{\partial \tilde{\mathbf{g}}_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\theta}_0} \right) \boldsymbol{\delta}_n$$

where $\boldsymbol{\psi}_n$ is any sequence $\boldsymbol{\psi}_n$ with $\boldsymbol{\psi}_n \to \boldsymbol{\theta}_0$, as $n \to \infty$. From 3.10 the expression above can be written as,

$$\begin{aligned}
&\boldsymbol{\delta}_n^T \left[\boldsymbol{\Lambda}_n(\boldsymbol{\psi}_n) - \boldsymbol{\Lambda}_n(\boldsymbol{\theta}_0) \right] \boldsymbol{\delta}_n \\
&+ \left(\left[\frac{\partial \tilde{\mathbf{h}}_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\psi}_n} \boldsymbol{\delta}_n \right)^T \mathbf{W}_n \left[\frac{\partial \tilde{\mathbf{h}}_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\psi}_n} \boldsymbol{\delta}_n \\
&- \left(\left[\frac{\partial \tilde{\mathbf{h}}_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\theta}_0} \boldsymbol{\delta}_n \right)^T \mathbf{W}_n \left[\frac{\partial \tilde{\mathbf{h}}_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\theta}_0} \boldsymbol{\delta}_n
\end{aligned}$$

The $(u, v)^{th}$ element of $\boldsymbol{\delta}_n^T \left[\boldsymbol{\Lambda}_n(\boldsymbol{\psi}_n) - \boldsymbol{\Lambda}_n(\boldsymbol{\theta}_0) \right] \boldsymbol{\delta}_n$ is

$$\sum_{r=1}^{k} \sum_{s=1}^{k} \left(W_{rs} \tilde{h}_s(y; \boldsymbol{\psi}_n) \sum_{i=1}^{p} \sum_{j=1}^{p} \delta_{iu} \delta_{vj} \Omega_{i,j}^{(r)}(\boldsymbol{\psi}_n) - W_{rs} \tilde{h}_s(y; \boldsymbol{\theta}_0) \sum_{i=1}^{p} \sum_{j=1}^{p} \delta_{iu} \delta_{vj} \Omega_{i,j}^{(r)}(\boldsymbol{\theta}_0) \right)$$

since

$$\left| \Omega_n^{(r)}(\boldsymbol{\psi}_n) \boldsymbol{\delta}_n - \Omega_n^{(r)}(\boldsymbol{\theta_0}) \right| < \sup_{\boldsymbol{\theta}: |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < c} \left| \Omega_n^{(r)}(\boldsymbol{\theta}) \boldsymbol{\delta}_n - \Omega_n^{(r)}(\boldsymbol{\theta_0}) \right|$$

the uniform convergence assumption stated in (3.6) is suficient to show that

$$\boldsymbol{\delta}_n^T \left[\boldsymbol{\Lambda}_n(\boldsymbol{\psi}_n) - \boldsymbol{\Lambda}_n(\boldsymbol{\theta}_0) \right] \boldsymbol{\delta}_n \xrightarrow{p} \mathbf{0} \tag{3.12}$$

We now proceed with the evaluation of the remaining terms in (3.12), performing some simple algebraic manipulations we obtain,

$$
\begin{aligned}
&\left[\left(\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}+\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\psi}_n}-\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\right)\boldsymbol{\delta}_n\right]^T\mathbf{W}_n\\
&\left[\left(\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}+\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\psi}_n}-\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\right)\boldsymbol{\delta}_n\right]\\
&-\left(\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\boldsymbol{\delta}_n\right)^T\mathbf{W}_n\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\boldsymbol{\delta}_n\\
&=\left[\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\boldsymbol{\delta}_n\right]^T\mathbf{W}_n\left(\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\psi}_n}-\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\right)\boldsymbol{\delta}_n\\
&+\left[\left(\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\psi}_n}-\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\right)\boldsymbol{\delta}_n\right]^T\mathbf{W}_n\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\boldsymbol{\delta}_n\\
&+\left[\left(\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\psi}_n}-\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\right)\boldsymbol{\delta}_n\right]^T\mathbf{W}_n\left(\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\psi}_n}-\left[\frac{\partial\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_0}\right)\boldsymbol{\delta}_n
\end{aligned}
$$

This expression also converges in probability to $\mathbf{0}$, form (3.7). This means that (2.20) holds.

The results from this section allow us the use of (2.18). For the class of EFs constructed using the above procedure, i.e. minimizing (3.8), (2.18) takes the specific form,

$$\boldsymbol{\delta}_n^{-1}(\boldsymbol{\theta}_n-\boldsymbol{\theta}_0)=\mathbf{H}(\boldsymbol{\theta}_0)^T\mathbf{W}\mathbf{H}(\boldsymbol{\theta}_0)\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y}) \tag{3.13}$$

3.2 Lower bound - Optimal weighting

Up to now we only required that the weighting matrix ($\mathbf{W}_n$) in (3.8) is positive semi-definite and, in case it depends on the data, that the sequence ($\mathbf{W}_n$) converges in probability to a positive definite matrix of constants, $\mathbf{W}$. However it is clear that the choice of $\mathbf{W}_n$ will have an important impact on the covariance matrix of the estimating function and estimator. From Section 2.3 we know that there is an optimal covariance matrix for regular estimating functions. In the present context therefore, it seems reasonable to try to find an optimal weighting matrix.

A discussion of the optimal choice of weights can be found in Hansen (1982) and Hall (2005). The weighting matrix that gives a lower bound for the variance matrix of the estimator is $\mathbf{S}^{-1}$, where $\mathbf{S}$ is the covariance matrix of the limiting distribution of $\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})$ as defined in (3.2). Some steps of the proof in Hall (2005, Section 3.6) will be shown next. Define $\hat{\boldsymbol{\theta}}_n(\mathbf{W})$ as the GMM estimator based on the weighting matrix $\mathbf{W}$, and $\mathbf{V}(\mathbf{W})$ as the variance of the limiting distribution of $\boldsymbol{\delta}_n^{-1}(\hat{\boldsymbol{\theta}}_n(\mathbf{W}) - \boldsymbol{\theta}_0)$,

$$\mathbf{V}(\mathbf{W}) \quad = \quad \mathbf{M}_0(\mathbf{W})^{-1}\mathbf{H}^T(\boldsymbol{\theta}_0)\mathbf{W}\mathbf{S}(\boldsymbol{\theta}_0)\mathbf{W}^T\mathbf{H}(\boldsymbol{\theta}_0)\mathbf{M}_0(\mathbf{W})^{-1}.$$

where $\mathbf{M}_0(\mathbf{W}) = \mathbf{M}(\theta_0) = \mathbf{H}(\boldsymbol{\theta}_0)^T\mathbf{W}\mathbf{H}(\boldsymbol{\theta}_0)$, the change of notation here aims at emphasizing the dependency on the weighting matrix.

Using the fact that $\tilde{\mathbf{g}}_n(\boldsymbol{\theta};\mathbf{y};\mathbf{W})$ is a regular estimating function, from (3.13) we have,

$$\begin{aligned}\boldsymbol{\delta}_n^{-1}\left[\hat{\boldsymbol{\theta}}_n(\mathbf{W}) - \boldsymbol{\theta}_0\right] &= -\mathbf{M}_0(\mathbf{W})^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{W}) + o_p(1)\\ \boldsymbol{\delta}_n^{-1}\left[\hat{\boldsymbol{\theta}}_n(\mathbf{S}^{-1}) - \boldsymbol{\theta}_0\right] &= -\mathbf{M}_0(\mathbf{S}^{-1})^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{S}^{-1}) + o_p(1)\end{aligned}$$

Note that although $\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{W})$ is different from $\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{S}^{-1})$, the matrix $\boldsymbol{\delta}_n^{-1}$ is the same since it is based on the choice of $\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})$; see (3.4) and (3.5).

Notice next that we can write $\boldsymbol{\delta}_n^{-1}\left[\hat{\boldsymbol{\theta}}_n(\mathbf{W}) - \boldsymbol{\theta}_0\right]$ as

$$\boldsymbol{\delta}_n^{-1}\left[\hat{\boldsymbol{\theta}}_n(\mathbf{W}) - \boldsymbol{\theta}_0\right] = \boldsymbol{\delta}_n^{-1}\left[\hat{\boldsymbol{\theta}}_n(\mathbf{S}^{-1}) - \boldsymbol{\theta}_0\right] + \boldsymbol{\delta}_n^{-1}\left[\hat{\boldsymbol{\theta}}_n(\mathbf{W}) - \hat{\boldsymbol{\theta}}_n(\mathbf{S}^{-1})\right] \tag{3.14}$$

Calculating the limiting variance of each side of (3.14) we obtain

$$\mathbf{V}(\mathbf{W}) = \mathbf{V}(\mathbf{S}^{-1}) + \mathbf{V}_1 - \mathbf{C} \tag{3.15}$$

where

$$
\begin{aligned}
\mathbf{V}_1 &= \lim_{n\to\infty} \mathrm{Var}\left[\mathbf{M}_0(\mathbf{W})^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{W}) - \mathbf{M}_0(\mathbf{S}^{-1})^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{S}^{-1})\right] \\
\mathbf{C} &= \lim_{n\to\infty} \mathrm{Cov}\left[\mathbf{M}_0(\mathbf{W})^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{W}) - \mathbf{M}_0(\mathbf{S}^{-1})^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{S}^{-1}), \mathbf{M}_0(\mathbf{S}^{-1})^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{S}^{-1})\right] \\
&= \lim_{n\to\infty} E\left[\mathbf{M}_0(\mathbf{W})^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{W}) - \mathbf{M}_0(\mathbf{S}^{-1})^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{S}^{-1})[\mathbf{M}_0(\mathbf{S}^{-1})^{-1}\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{S}^{-1})]^T\right] \\
&= \lim_{n\to\infty} \mathbf{M}_0(\mathbf{W})^{-1}E[\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{W})\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{S}^{-1})^T]\left[\mathbf{M}_0(\mathbf{S}^{-1})^{-1}\right]^T \\
&\quad -\mathbf{M}_0(\mathbf{S}^{-1})^{-1}E[\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{S}^{-1})\tilde{\mathbf{g}}_n(\boldsymbol{\theta}_0;\mathbf{y};\mathbf{S}^{-1})^T]\left[\mathbf{M}_0(\mathbf{S}^{-1})^{-1}\right]^T
\end{aligned}
$$

Now we use (3.10), together with the fact that both $\mathbf{S}$ and $[\mathbf{H}(\boldsymbol{\theta}_0)^T\mathbf{S}^{-1}\mathbf{H}(\boldsymbol{\theta}_0)]$ are symmetric to obtain,

$$
\begin{aligned}
\mathbf{C} &= [\mathbf{H}(\boldsymbol{\theta}_0)^T\mathbf{S}^{-1}\mathbf{H}(\boldsymbol{\theta}_0)]^{-1} - [\mathbf{H}(\boldsymbol{\theta}_0)^T\mathbf{S}^{-1}\mathbf{H}(\boldsymbol{\theta}_0)]^{-1} \\
&= 0
\end{aligned}
$$

and (3.15) becomes,

$$
\mathbf{V}(\mathbf{W}) - \mathbf{V}(\mathbf{S}^{-1}) = \mathbf{V}_1
$$

Since $\mathbf{V}_1$ is a covariance matrix it must be positive semi-definite. Therefore $\mathbf{V}(\mathbf{S}^{-1})$ provides an achievable lower bound for the covariance matrix of $\boldsymbol{\delta}_n^{-1}(\hat{\boldsymbol{\theta}}_n(\mathbf{W}) - \boldsymbol{\theta}_0)$, and $\mathbf{S}^{-1}$ is the optimal weighting matrix.

In practice $\mathbf{S}$ is unknown and it should be replaced by a consistent estimator of $\mathbf{S}$. Usually this requires at least a further step in the estimation procedure. Starting with a sub-optimal weighting matrix, an estimator $\hat{\boldsymbol{\theta}}$ is obtained, this original estimator is plugged into an analytical representation of $\mathbf{S}$ as a function of $\boldsymbol{\theta}$, this way we have a consistent estimator for $\mathbf{S}$. The second step is to obtain a new GMM estimator for $\boldsymbol{\theta}$ using $\mathbf{W} = \hat{S}^{-1}$. Clearly this iterative procedure can be performed several times to obtain an estimator with better finite sample properties. If one proceeds with the algorithm until suitable convergence is achieved the estimator is called *iterative GMM estimator*.

3.3 Summary

In this chapter we used the estimation function framework set out in the previous chapter to discuss consistency and asymptotic distribution of the GMM estimator, this is an alternative approach to the one in Hansen (1982). The GMM approach to building estimating functions is particularly useful when the set of moment conditions is larger than the set of unknown parameters, and the concept of optimal weighting matrix provides an answer to the practical problem of which moment condition to use or how to weight them when combining them in an objective function. Once the estimating function is defined as the gradient vector of the quadratic form (3.8), we are in the setting from the previous section. Moreover we can make use of the result regarding the construction of confidence regions based on the objective function itself (2.24), which we will apply in the next chapter. In the statistical science literature there have been some extensions to the GMM methodology, namely the generalized empirical likelihood which allows also for semi-parametric inference (Owen, 2000; Imbens et al., 1998; Lindsay and Qu, 2003).

Chapter 4

Simulation Study - Application of GMM to rainfall models

This work on estimating functions has been largely motivated by work with a class of stochastic models for rainfall time series. These models date back to Rodriguez-Iturbe et al. (1984) and are widely used in the hydrological community to generate artificial rainfall time series, for purposes such as the assessment of flood risk and the impacts of climate change. A model of this type is used in the "weather generator" provided as part of the latest suite of UK national climate change projections (Burton et al., 2008). In the initial part of this chapter we describe their structure, and highlight some of the difficulties that they pose for inference. In section 4 we present a simulation based finite sample study where we apply the GMM methods to the Poisson Rectangular Pulses Model and Neyman-Scott Rectangular Pulses Model. The description of the study is made in section 4.2.1, and is followed by the results and their discussion in the subsequent sections.

4.1 Model description

These models are based on a simplified conceptual representation of the physical strucure of the rainfall proces, in which rainfall is considered to be generated from "cells" of activity in the atmosphere. This conceptual representation is illustrated in Figure 4.1. The core of such models is a point process, each event of which marks the arrival time of a cell. Many arrival processes can be included in this class of models, from the Poisson and Poisson cluster processes of Rodriguez-Iturbe et al. (1987) to the more

recent development of Markov modulated Poisson processes of Ramesh (1998). Each cell has a random duration, during which it deposits rain with a constant intensity (also random) so that its temporal profile is rectangular. At any point in time, the rainfall intensity at a particular location is the sum of the intensities of all cells that are currently active, rainfall is thus represented in continuous time. However, rainfall data are usually recorded discretely using a raingauge with an automated recording system attached to it (nowadays usually a digital data logger) and data from such devices are routinely archived at an hourly resolution. Because of the aggregated nature of the data, and also because of the complex dependencies induced by the model structures, it is not feasible to write down a likelihood for the data. Therefore these models are traditionally fitted by matching theoretical moments with the observed counterparts. There are however alternatives, namely the marginal likelihood approach of Northrop (2006) where by building a likelihood based on only certain aspects of the data it is possible to estimate a subset of the parameters defining these models. A different approach has been taken by Chandler (1997) where by transforming the data into the frequency domain it is possible to build an approximate likelihood for the parameters of these models, we pursue this idea further in chapter 5.

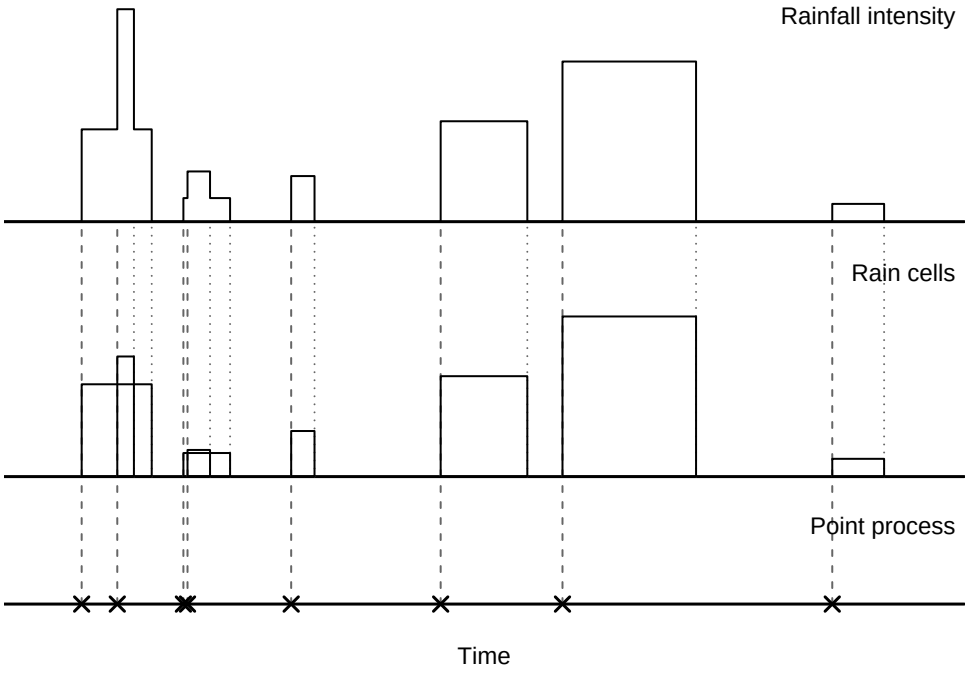

FIGURE 4.1: Schematic diagram of a generic point-process rainfall model. Vertical dashed and dotted lines mark the start and end times of rain cells respectively.

4.1.1 Poisson Rectangular Pulses Model - PRPM

One of the simplest models of this type is the Poisson Rectangular Pulses model, for which the driving point process, say $N(t)$, is Poisson. Formally the model can be described using the following random variables:

- T_i is the arrival time of the i^{th} cell, with cells occurring in a Poisson process of rate λ
- L_i is the duration of the i^{th} cell, exponentially distributed with parameter η. We will refer to the mean cell duration as $\mu_L = \eta^{-1}$.
- X_i is the cell intensity that remains constant over the duration of the cell. At this stage we assume no distribution and simply define its moments, X_i are i.i.d. as random variables X with,

$$\mu_{X^k} = E\left[X^k\right]$$

These three random variables (arrival time, duration and intensity) are assumed to be independent of each other and of the corresponding variables for other cells. The model allows for cell overlapping, i.e. $T_i < T_j \leq T_i + L_i$ for some $j \neq i$.

The rainfall intensity at time t, denoted $Y(t)$, is then the sum of the contributions from all cells active at time t, and can be formally defined as

$$Y(t) = \int_{u=0}^{\infty} X_{t-u}(u) dN(t-u)$$

where $X_u(\tau)$ is the random intensity of the cell that arrived at time u, observed τ units of time later. As the intensity of a cell is constant over its lifetime, the random variable $X_{t-u}(u)$ can be written as,

$$X_{t-u}(u) \quad = \quad \begin{cases} X(t-u) & \text{with probability } \mathcal{F}_{\mathrm{L}}(\mathrm{u}). \\ 0 & \text{with probability } 1 - \mathcal{F}_{\mathrm{L}}(\mathrm{u}), \end{cases}$$

where $\mathcal{F}_L(l) = P[L > l]$ is the survivor function of the cell duration L and $X(t-u)$ is the initial intensity of a cell born at time $t-u$. From this formulation of the process

$Y(t)$ it is possible to derive some of its moments (Rodriguez-Iturbe et al., 1987), we show the steps for the mean only as the technique is similar for higher order moments.

$$\begin{aligned} E[Y(t)] &= E\left[\int_{u=0}^{\infty} X_{t-u}(u)dN(t-u)\right] \\ &= \int_{u=0}^{\infty} E\left[X_{t-u}(u)\right] E\left[dN(t-u)\right]. \end{aligned} \tag{4.1}$$

The last step is justified by independence between the cell arrival process and the intensity process. Now, treating each expectation separately

$$E[dN(t-u)] = \lambda du$$

and

$$E[X_{t-u}(u)] = E_{\chi_{(t,u)}}\left[E[X_{t-u}(u)|\chi_{(t,u)}]\right]$$

where

$$\chi_{(t,u)} = \begin{cases} 1 & \text{if } L(t-u) > u \\ 0 & \text{otherwise} \end{cases}$$

and $L(t-u)$ is the duration of a cell born at $(t-u)$. Now clearly we have

$$P(\chi_{(t,u)}=1) = P(L(t-u)>u) = P(L>u) = \mathcal{F}_L(u) = 1 - P(\chi_{(t,u)}=0)\ .$$

Moreover,

$$\begin{aligned} E[X_{t-u}(u)|\chi_{(t,u)}=0] &= 0 \\ E[X_{t-u}(u)|\chi_{(t,u)}=1] &= \mu_X\ . \end{aligned}$$

It follows that

$$E[X_{t-u}(u)] = \mu_X \mathcal{F}_L(u)$$

and (4.1) becomes

$$\begin{aligned} E[Y(t)] &= \lambda\mu_X \int_{u=0}^{\infty} \mathcal{F}_L(u)du \\ &= \lambda\mu_X\mu_L \ . \end{aligned} \tag{4.2}$$

The variance and autocovariance can be derived similarly to obtain

$$\mathrm{Var}[Y(t)] = \lambda\mu_{X^2} \int_0^{\infty} \mathcal{F}_L(v)dv = \lambda\mu_{X^2}\mu_L \tag{4.3}$$

$$c_y(\tau) = \mathrm{Cov}[Y(t), Y(t+\tau)] = \lambda\mu_{X^2} \int_{\tau}^{\infty} \mathcal{F}_L(v)dv \tag{4.4}$$

An exponential distribution is often assumed for the cell duration (Yoo et al., 2008; Northrop, 2006), in such case the expressions for the moments can be further simplified. Denote the parameter of the exponentially distributed cell duration as η, then expressions (4.2 - 4.4) can be written as,

$$\begin{aligned} E[Y(t)] &= \lambda\mu_X\eta^{-1} \\ \mathrm{Var}[Y(t)] &= \lambda\mu_{X^2}\eta^{-1} \\ c_y(\tau) &= \lambda\mu_{X^2}\eta^{-1}e^{-\eta\tau} \end{aligned}$$

This can be further simplified in the case where X is also exponentially distributed, using $\mu_{X^2} = 2\mu_X^2$.

The model described above is for rainfall in continuous time, however we can only observe the aggregated totals over disjoint time intervals of fixed length h, hence define the random variable

$$Y_i^{(h)} = \int_{(i-1)h}^{ih} Y(v)dv \tag{4.5}$$

It is very difficult to write down a joint distribution for these aggregated totals, meaning that traditional likelihood inference is not feasible. A well established alternative is to perform parametric estimation through a generalised method of moments using a set of model properties that can be derived from (4.5) together with the properties derived earlier for $Y(t)$ (Rodriguez-Iturbe et al., 1987). As before we will only show how the first moment is derived, and note that we are only interested the particular case of exponentially distributed cell duration. The first moment is obtained by taking expectations on both sides of (4.5),.

Properties		
Mean	$\mu_Y(h;\boldsymbol{\theta})$	$h\lambda\mu_X\eta^{-1}$
Variance	$\sigma^2_Y(h;\boldsymbol{\theta})$	$2\lambda\mu_{X^2}\eta^{-3}(h\eta - 1 + e^{-h\eta})$
Lag k autocovariance	$c_Y(h;\boldsymbol{\theta};k)$	$\lambda\mu_{X^2}\eta^{-3}\left(e^{h\eta} + e^{-h\eta} - 2\right)e^{-kh\eta}$
Probability of dry	$p_Y(h;\boldsymbol{\theta})$	$e^{-\lambda\eta^{-1}}e^{-\lambda h}$
Parameters		
$\boldsymbol{\theta}$	λ, μ_X, μ_{X^2}, η	
λ	Parameter for the Poisson process of cell arrivals	
μ_X	Mean cell intensity	
μ_{X^2}	Second moment of cell intensity process	
η	Parameter of exponential cell duration	

TABLE 4.1: Properties and parameters for the PRPM with exponential cell duration aggregated over time intervals of length h (Rodriguez-Iturbe et al., 1987)

$$E\left[Y_i^{(h)}\right] = \int_{(i-1)h}^{ih} E[Y(v)]dv = h\lambda\mu_X\eta^{-1}.$$

Similarly the autocovariance function can be derived as

$$c_Y(h;\boldsymbol{\theta};k) = \mathrm{Cov}[Y_i^{(h)}, Y_{i+k}^{(h)}] = \lambda\mu_{X^2}\eta^{-3}\left(e^{h\eta} + e^{-h\eta} - 2\right)e^{-kh\eta}\ ,\ k \geq 1.$$

Another property of interest for the aggregated process is the probability $p_Y(h;\boldsymbol{\theta})$ that an interval of length h is dry (i.e. experiences zero rainfall). For this to occur it is required that there are no cells active at the beginning of the interval (i.e. $Y((i-1)h) = 0$) and no cells arrive during the interval. Thus

$$p_Y(h;\boldsymbol{\theta}) = P[Y_i^{(h)} = 0] = e^{-\lambda\eta^{-1}}e^{-\lambda h}$$

Where $e^{-\lambda\eta^{-1}}$ is the probability that no cells are active at any given time.

Table 4.1. summarizes the results of this section.

4.1.2 Poisson-Cluster Rectangular Pulses Models

The PRPM described above is a simple way to describe rainfall at a particular level of aggregation. However one of the main criticisms is that it fails to replicate the observed properties of rainfall when other levels of aggregation are considered. Extensive reported experience suggests, for example, that if the PRPM is calibrated to reproduce the hourly

rainfall variance at some location then it will tend not to reproduce the daily variance at the same location (Rodriguez-Iturbe et al., 1987). In fact, it is known that rain events do not occur with constant intensities, but with randomly varying intensities within the same rain event. A suggestion in the literature is to improve the modelling of rainfall by considering clusters of cells instead of a single cell. Several models have been developed that are defined in terms of clusters of cells, introducing the notion of a storm. Instead of using a Poisson process for the arrival of cells, these models use a clustered point process. This type of model in general shows better consistency across timescales when compared to the PRPM (Cowpertwait et al., 2007).

We now extend the notation of the previous section to accommodate the extra complexity of these models

- λ - Storm(Cluster) arrival rate
- β - Parameter of the distribution of the displacement of cells within a storm
- C - Number of cells per storm

There are several clustering mechanisms however we will focus on two of these that deserve particular attention due to their extensive use in hydrology (Cowpertwait, 1991; Onof et al., 2000; Wheater et al., 2005; Burton et al., 2008). These are the Bartlett-Lewis and the Neyman-Scott mechanisms. A common feature of these models is that the storms arrive following a Poisson process with rate λ. What distinguishes these two point processes is the way the cell arrival times relate to storm origins.

4.1.2.1 Neyman-Scott Rectangular Pulses Model - NSRPM

The cell origin times under the NSRPM are defined by a set of independent and identically distributed random variables, representing displacements from the storm origin. To each storm origin there is a random number C of cells associated to it. Some authors assume the existence of a cell associated with the storm origin, however in our treatment of the NSRPM we assume no cell starts at the storm origin. A common choice for the distribution defining the displacement of the cell arrival times relative to the storm origins is the exponential distribution. In this text we will assume this distribution, and denote its parameter by β. Different discrete distributions can be used for the number of cells per storm, the Poisson and Geometric being the most common.

The expressions for the first and second order moments of the aggregated process are given here without assuming a specific distribution for C (Rodriguez-Iturbe et al., 1987):

$$\begin{aligned}
E[Y_i^{(h)}] &= h\lambda\mu_X\mu_C\eta^{-1} \\
\mathrm{Var}[Y_i^{(h)}] &= \lambda\eta^{-3}(\eta h - 1 + e^{-\eta h})\left[2\mu_C\mu_{x^2} + E[C^2 - C]\mu_X^2\beta^2/(\beta^2 - \eta^2)\right] \\
&\quad -\lambda(\beta h - 1 + e^{\beta h})E[C^2 - C]\mu_X^2/[\beta(\beta^2 - \eta^2)] \\
\mathrm{Cov}[Y_i^{(h)}, Y_{i+k}^{(h)}] &= \lambda\eta^{-3}(1 - e^{-\eta h})^2 e^{\eta(k-1)h}\left[\mu_C\mu_{X^2} + \frac{1}{2}E[C^2 - C]\mu_X^2\beta^2/(\beta^2 - \eta^2)\right] \\
&\quad -\lambda(1 - e^{-\beta h})^2 e^{-\beta(k-1)h}\frac{1}{2}E(C^2 - C)\mu_X^2/[\beta(\beta^2 - \eta^2)],\ k \geq 1.
\end{aligned}$$

The expression for the probability that an interval of length h is dry is (Cowpertwait, 1991)

$$p_y(h;\boldsymbol{\theta}) = \exp\left\{-\lambda h + \frac{\lambda}{\beta\mu_C}\left[1 - \exp(-\mu_C + \mu_C e^{-\beta h})\right] - \lambda\int_0^\infty [1 - p_t(h)]dt\right\}$$

where $p_t(h)$ is the probability of a dry period of length h assuming no overlapping of storms within the interval of interest, and is given by

$$\begin{aligned}
p_t(h) &= \left(1 - e^{-\beta t} + e^{-\beta(t+h)}\right)\left(1 - \frac{\beta(e^{-\beta t} - e^{-\eta t})}{\eta - \beta}\right) \\
&\quad \exp\left\{-\frac{\mu_C\beta\left(e^{-\beta t} - e^{-\eta t}\right)}{\eta - \beta} - \mu e^{-\beta t} + \mu e^{-\beta(t+h)}\right\}
\end{aligned}$$

4.1.2.2 Bartlett-Lewis Rectangular Pulses Model - BLRPM

One possible alternative to the NSRPM is the model where cells arrive following a Bartlett-Lewis process. In the BLRPM the cells arrive following a Poisson process of rate β starting with one cell at the storm origin, this process of cell origins is terminated after an exponentially distributed time from the storm origin. Given the similarities between the BLRPM and the NSRPM, however we will not include this model in our finite sample performance study.

4.1.3 Final remark on point process rainfall models

There is extensive literature regarding the use of these models in hydrology, Onof and Wheater (1994); Wheater et al. (2005); Cowpertwait et al. (2007) use the BLRPM to model rainfall. Modelling of rainfall using the NSRPM can be found in Cowpertwait

(1991), and more recently in Burton et al. (2008) which set out the basis for the use of NSRPM to model rainfall within the weather generator in the UK Climate Projections, namely the UKCP09. We will focus our finite sample study on the PRPM and the NSRPM; although the main objective of the study is to verify the asymptotic theory in practice the application of such theory to a model that is known to be used in practical applications can introduce an extra level of contribution.

4.2 Simulation study

In chapter 3 we saw that given a set of m moment conditions and a parameter vector, $\boldsymbol{\theta}$, of length $p \leq n$ we can consistently estimate $\boldsymbol{\theta}$ by combining these moment condition in a quadratic form (3.8) to be minimized. Suppressing n we have,

$$Q(\boldsymbol{\theta}) = \tilde{\mathbf{h}}(\boldsymbol{\theta}; \mathbf{y})^T \ \mathbf{W} \ \tilde{\mathbf{h}}(\boldsymbol{\theta}; \mathbf{y}) \tag{4.6}$$

We now illustrate the application of this method to the four-parameter Poisson and to the six-parameter Neyman-Scott models described in section 4.1. Such models are often fitted by minimizing an expression of the form (4.6), where often $\mathbf{W}_n$ is assumed diagonal by design. This simulation study will provide an opportunity to determine whether the asymptotic results provide a reasonable approximation for a modest sized sample, which can be useful to evaluate the scope for asymptotic theory to be used in practice. As well as examining the validity of asymptotics, we compare different choices of weighting matrix $\mathbf{W}$, one of the choices is clearly the theoretical optimum, but this weighting matrix will be compared with simpler and current practice weighting schemes. The main objectives in this study are: validity of approximations derived from asymptotic theory in finite samples, and performance of different weighting schemes. The results will be looked at from the perspective of the two desirable properties of the estimator: bias and minimal variance. It's not our aim for example to evaluate the choice of moment conditions, the properties from the model chosen to match with sample counterparts are based on current practice and are assumed as given for the purpose of this study.

4.2.1 Study setup

4.2.1.1 Data

Given that we will be using simulations we should try to ensure that the simulated time series is as similar as possible to real data, given the model chosen. In this study we will try to respect and replicate some features of rainfall data:

- Seasonality - It is a known fact that rainfall behaves differently depending on the time of the year. For the class of models considered here this can be handled by estimating different parameter values for each calendar month (Rodriguez-Iturbe et al., 1988; Wheater et al., 2005). To mimic this, each of our simulations will consist of independent sets of 30 days worth of rainfall, representing data for the same calendar month in each of the years.
- Frequency/Resolution - These models assume that the data available consists of aggregate rainfall over disjoint h-hourly periods. We need to decide on the frequency of our simulated time-series. We assume that hourly totals are available, which is usually the case in the applications where this type of model is used.
- Length of time series - Wheater et al. (2005) suggest on the basis of empirical experience that 20 years of data may be required for reliable calibration of this type of model; this guideline was also based on the extent of data availability in the UK. Accordingly we adopt $n = 20$ years as the basic simulation period.

So each simulation consists of 20 sets of 30 days worth of hourly rainfall totals. The number of simulations in this experiment is 1000.

Minimization of the objective function must be carried out numerically: our experience is that this can be challenging, and that working with the logarithm of the model parameters can stabilise the procedure. Therefore the parameter vector of interest is $\boldsymbol{\theta} = (\log(\lambda), \log(\mu_X), \log(\sigma_X/\mu_X), \log(\mu_L))$. The values used to generate the simulation data were based on real data. In particular hourly data for January from Birmingham airport covering the period from 1950 to 1996; to obtain the parameter values we rounded the estimates obtained using the moment estimator obtained from (3.8) with equal weights $\mathbf{W} = I$. The resulting values were $\boldsymbol{\theta} = (-3.5, 0, 0, 1.1)$, where the parameter $\log(\sigma_X/\mu_X)$ was fixed at zero prior to estimating the remaining parameters meaning we are assuming that cell intensity is exponentially distributed. This is an assumption frequently made in applications (Rodriguez-Iturbe et al., 1987; Cowpertwait,

1991). These estimates are consistent with values found in the literature for this model (Northrop, 2006).

4.2.1.2 Moment Conditions

In order to estimate the parameters using the results in section 3.1 we need to find a set of moment conditions. In section 4.1.1 we have shown the expressions for some properties in terms of the parameters of interest. We will now present a vector of statistics $\mathbf{T}(\mathbf{y})$ whose expectations can be expressed using the properties from Table 4.1. These will allow us to define the moment conditions as $E[\mathbf{T}(\mathbf{y}) - \boldsymbol{\tau}(\boldsymbol{\theta})] = \mathbf{0}$. Although the intuition behind the use of GMM is that the moment conditions are exactly unbiased, the asymptotic results from section 3.1 show that the method is still justified provided that the moment conditions are asymptotically unbiased after suitable normalization. In the present context of finite samples we should try to choose moment conditions that minimize the finite sample bias. In fact one may expect that finite sample performance could be improved by using moment conditions that are exactly unbiased; however as previously stated the aim of this work is not to discuss the optimal choice of moment conditions, but rather the relevance of the asymptotic theory regarding the choice of weights. The choice of the properties is therefore based on common practice according to the literature (Rodriguez-Iturbe et al., 1988; Wheater et al., 2005). The properties are mean; variance, lag 1 autocorrelation and proportion of dry intervals at hourly resolution; variance at 6-hourly resolution; variance, lag 1 autocorrelation and proportion of dry intervals at 24-hourly resolution. The sample mean, variances, autocorrelation and proportion of dry periods are formally defined as for a contiguous time series,

$$
\begin{aligned}
\overline{y}^{(h)} &= \frac{1}{N^{(h)}} \sum_{i=1}^{N^{(h)}} y_i^{(h)} \\
s^2(h;\mathbf{y}) &= \frac{1}{N^{(h)}-1} \sum_{i=1}^{N^{(h)}} \left(y_i^{(h)} - \overline{y}^{(h)} \right)^2 \quad , h = 1, 6, 24. \\
z(h;\mathbf{y};l;u) &= \frac{\sum_{i=l+1}^{u} \left(y_i^{(h)} - \overline{y}^{(h)} \right) \left(y_{i-1}^{(h)} - \overline{y}^{(h)} \right)}{\sum_{i=l}^{u} \left(y_i^{(h)} - \overline{y}^{(h)} \right)^2} \\
p_{\text{dry}} &= \frac{1}{N^{(h)}} \sum_{i=1}^{N^{(h)}} \chi_i^{(h)}
\end{aligned}
$$

where $N^{(h)}$ is the total number of h-hourly periods in the series, and χ_i is an indicator function that takes the value one if $y_i = 0$ and zero otherwise.

To apply the methods described in chapter 3 we also need to estimate the covariance matrix of the moment conditions. For this we follow a suggestion by Rodriguez-Iturbe et al. (1988), also used in Wheater et al. (2005) where treating data from different years as independent allows us to calculated a separate set of moment conditions for each year and to use the resulting sample of moment conditions to calculate both a mean set of moments and an estimate of the covariance matrix of its mean; call this covariance estimate $\hat{\mathbf{S}}$. Thus we change the moment conditions slightly, i.e., we first calculate $T_i(\mathbf{y})$ for each set of 30 days (year), say $T_i^l(\mathbf{y})$. In our simulations we have 20 years, and the sample moments become,

$$\overline{T}_i(\mathbf{y}) \quad = \quad \frac{1}{20}\sum_{l=1}^{20} T_i^l(\mathbf{y}) \quad . \tag{4.7}$$

Strictly speaking these new moment conditions do not satisfy (3.1) because the finite sample bias of $s^2(h;\mathbf{y})$ and $z(h;\mathbf{y};l;u)$ from using 30 days is not reduced when increasing the number of years. However for practical purposes it may be that this bias is small compared to sampling variation. In particular, for the autocorrelation we attempt to minimize such effect by using Quenouille's bias-reduced estimator (Kendall and Ord, 1990), defined as

$$r(h;\mathbf{y};N^{(h)}) \quad = \quad 2z(h;\mathbf{y};1;N^{(h)}) - \frac{z(h;\mathbf{y};1;\lceil N^{(h)}/2\rceil) + z(h;\mathbf{y};\lfloor N^{(h)}/2\rfloor;N^{(h)})}{2} \quad , \; h = 1,24$$

The collection $\left(\overline{Y}^{(1)}, s^2(1;\mathbf{Y}), s^2(6;\mathbf{Y}), s^2(24;\mathbf{Y}), r(1;\mathbf{Y}), r(24;\mathbf{Y}), p_{\text{dry}}(1;\mathbf{Y}), p_{\text{dry}}(24;\mathbf{Y})\right)$ is used to obtain the moment conditions $\tilde{\mathbf{h}}(\boldsymbol{\theta};\mathbf{y}) = E[\mathbf{T}(\mathbf{y}) - \boldsymbol{\tau}(\boldsymbol{\theta})]$ in the notation of section 3.1, and according to the theory the asymptotically unbiased moment condition requires that for consistent estimators you need

$$\begin{aligned} E\left[\overline{Y}^{(h)}\right] &= \mu_Y(h;\boldsymbol{\theta}) \\ E\left[s^2(h;\mathbf{Y})\right] &\to \sigma_Y^2(h;\boldsymbol{\theta}) \\ E\left[r(h;\mathbf{Y})\right] &\to \frac{c_Y(h;\boldsymbol{\theta};1)}{\sigma_Y^2(h;\boldsymbol{\theta})} \\ E\left[p_{\text{dry}}(h;\mathbf{Y})\right] &= p_Y(h;\boldsymbol{\theta}) \end{aligned} \tag{4.8}$$

In this case the normalizing matrix is the identity matrix. The matrix $\hat{\mathbf{S}}$, the covariance matrix of the new sample moments has $(i,j)^{th}$ element

$$\frac{1}{20}\,\frac{1}{19}\,\sum_{l=1}^{20}\left(T_i^l(\mathbf{y})-\overline{T}_i(\mathbf{y})\right)\left(T_j^l(\mathbf{y})-\overline{T}_j(\mathbf{y})\right) \tag{4.9}$$

Here the subscript n, corresponding to the sample size has been suppressed for simplicity of notation.

4.2.1.3 Estimation of Estimator Variance

We are interested in comparing the performance of different weights in terms of the variance of the resulting estimator. To assess the variability of each estimator we will use two different measures. One is the empirical covariance matrix obtained by treating the set of simulated estimates as a sample from the distribution of the estimator. In this section we denote by $\hat{\Theta}$ the matrix of estimates obtained, clearly there will be $K = 1000$ estimates, as many as the number of simulations in the study. Therefore $\hat{\Theta}$ is the $p \times K$ matrix where each row includes the estimates for each component of the parameter vector. Denote

$$\overline{\boldsymbol{\theta}}_i \;=\; \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}_{ik}\ ,\ i=1,...,p\ .$$

where $\hat{\boldsymbol{\theta}}_{ik}$ is the $(i,k)^{th}$ element of $\hat{\Theta}$.

The empirical covariance matrix will be denoted by $\overline{\text{Var}}(\hat{\boldsymbol{\theta}})$ and it has $(i,j)^{th}$ element

$$\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\boldsymbol{\theta}}_{ik}-\overline{\boldsymbol{\theta}}_i\right)\left(\hat{\boldsymbol{\theta}}_{jk}-\overline{\boldsymbol{\theta}}_j\right)\ ,\quad i,j=1,...,p\ , \tag{4.10}$$

In practical applications, however, it is necessary to estimate the covariance matrix of the estimator using data from a single realization. The asymptotic results from section 3.1 provide a means of achieving this as

$$\mathrm{Var}(\hat{\boldsymbol{\theta}})=\mathbf{M}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{M}(\boldsymbol{\theta}_0).$$

For each simulation therefore, we compute this quantity, with $\mathbf{M}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ replaced by estimates. We refer to the resulting estimate as the "theoretical covariance matrix", and denote it $\hat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})=\hat{\mathbf{M}}^{-1}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{M}}^{-T}$. To investigate the accuracy of the asymptotic results we can compare the average of the theoretical covariance matrices with the empirical covariance matrix $\overline{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$.

4.2.1.4 Weighting schemes considered

In order to perform inference using (4.6) we need to have a vector $\tilde{\mathbf{h}}_n(\boldsymbol{\theta};\mathbf{y})$ which was fully defined in the previous section, and a matrix $\mathbf{W}$ whose choice we discuss in this section. The main focus of this study will be then to compare different weighting schemes, $\mathbf{W}$. From section 3.1 we know that the matrix which gives the minimal asymptotic variance is $\mathbf{W}_{Opt}=\mathbf{S}^{-1}$, where $\mathbf{S}$ is the limiting variance of the sample moments. Since $\mathbf{S}$ is unknown we will use $\mathbf{W}_0=\hat{\mathbf{S}}^{-1}$.

We will compare the performance of this theoretical best with some alternatives:

- Equal weights($\mathbf{W}_1$) - $\mathbf{W}$ is a matrix with 1's on the diagonal and 0's everywhere else, i.e. the identity matrix, meaning all properties will have the same contribution to the objective function.
- Unequal weights($\mathbf{W}_2$) - $\mathbf{W}$ is a diagonal matrix with diagonal elements such that the properties 1-h mean, variance and proportion of dry have weights 100 times higher than the rest. This is an example where the weights are set by the investigator, based on the idea of having higher weights associated with properties that are considered more important (Wheater et al., 2005).

- GLS($\mathbf{W}_3$) - also a diagonal matrix where each element of the diagonal is the inverse of the estimated variance of the corresponding moment condition. This represents an approximation to the principle of giving higher weight to more stable properties.

We will use the notation $\hat{\boldsymbol{\theta}}(\mathbf{W}_i)$, $i = 0, 1, 2, 3$ to refer to the estimator obtained using the specified weighting matrix.

4.2.1.5 Performance Measurement Criteria

We will use several measures to evaluate the performance of estimators obtained using the four different weighting schemes. Boxplots will be used for the initial analysis of the distribution of the estimator, including bias, which will also be assessed by looking at its standard deviation.

The comparison of covariance matrices is not straightforward and there are different ways that such comparison can be performed. In section 2.3 we used a rigorous definition to derive the weighting matrix $\mathbf{W}_{Opt}$. For purposes of comparing different estimators (i.e. weighting schemes) in a finite sample simulation study, it is unlikely that such a strict criterion can be achieved in general; therefore it is necessary to consider alternative comparison methods.

However in practical applications this is a very strict criterion, and its application can lead most of the time to inconclusive results.

Another possible way of measuring the relative "size" of a covariance matrix is by using their determinants. This makes use of a generalization of the result in Draper and Guttman (1995). An approximate $100(1-\alpha)\%$ confidence region can be defined by the points $\boldsymbol{\theta}$ that satisfy

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T (\hat{\text{Var}}(\hat{\boldsymbol{\theta}}))^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) < C(\alpha, n, p) \tag{4.11}$$

where $C(\alpha, n, p)$ is a constant depending on the confidence level, the number of observations n, the number of parameters p and the distribution of the estimator, which is assumed to be multivariate normal. The volume of the ellipsoid defined by (4.11) is

$$C(\alpha, n, p) D(p) \left[\det(\hat{\text{Var}}(\hat{\boldsymbol{\theta}})) \right]^{1/2}$$

where $D(p)$ is a constant that depends only on the number of parameters p. This means that the volume of the confidence region defined by the covariance matrix of the estimator is proportional to the square root of its determinant. Moreover, the best estimator is the one that for the same α, n and p has the smaller volume. Since we are not interested in quantifying the difference in volume, but just comparing relative volumes, we can use any monotonic transformation of the quantity $det(\hat{\text{Var}}(\hat{\boldsymbol{\theta}}))^{1/2}$. In this case we have four parameters and correspondingly a four by four matrix, by using the fourth root of $det(\hat{\text{Var}}(\hat{\boldsymbol{\theta}}))^{1/2}$ we have a quantity on the standard error scale.

A third and possibly simpler way is to look at the individual components of the parameter vector, and compare the standard deviations. If on the one hand this seems too simplistic, on the other hand we know that practitioners tend to use these together with a normality assumption to build confidence intervals. Therefore we will also analyse the normality of the estimators.

We will check the accuracy of the asymptotic theory from section 2.4 by comparing $\hat{\text{Var}}(\hat{\boldsymbol{\theta}})$ and $\overline{\text{Var}}(\hat{\boldsymbol{\theta}})$; and by looking at the coverage of confidence intervals built using $\hat{\text{Var}}(\hat{\boldsymbol{\theta}})$, as well as confidence regions built from the objective function itself, as defined in Chapter 2.

4.2.2 Poisson Rectangular Pulses Model - Results and discussion

4.2.2.1 Boxplot

The simplest and most straightforward approach to start the analysis of the results is to look at a boxplot of the deviations from the true parameter value, $\boldsymbol{\theta}_0$. This is shown in Figure 4.2.

From Figure 4.2, is clear that the estimates corresponding to the weighting scheme $\mathbf{W}_1$ are very variable, except for $\log(\lambda)$. Moreover for the parameter $\log(\sigma_x/\mu_x)$ the distribution of the estimates seems strongly skewed, which indicates that inference based on asymptotic normality of the estimator is not suitable for this weighting scheme. The estimates using the weighting scheme labeled $\mathbf{W}_2$ show a large variation for the components $\log(\lambda)$ and $\log(\mu_L)$, but they behave reasonably well for the two remaining components, at least under this simple graphical analysis. The variance based weights, $\mathbf{W}_3$ and $\mathbf{W}_0$, seem to lead to estimators that perform much better than the other two especially for $\log(\mu_L)$, however it is not clear which of these has the best finite sample performance.

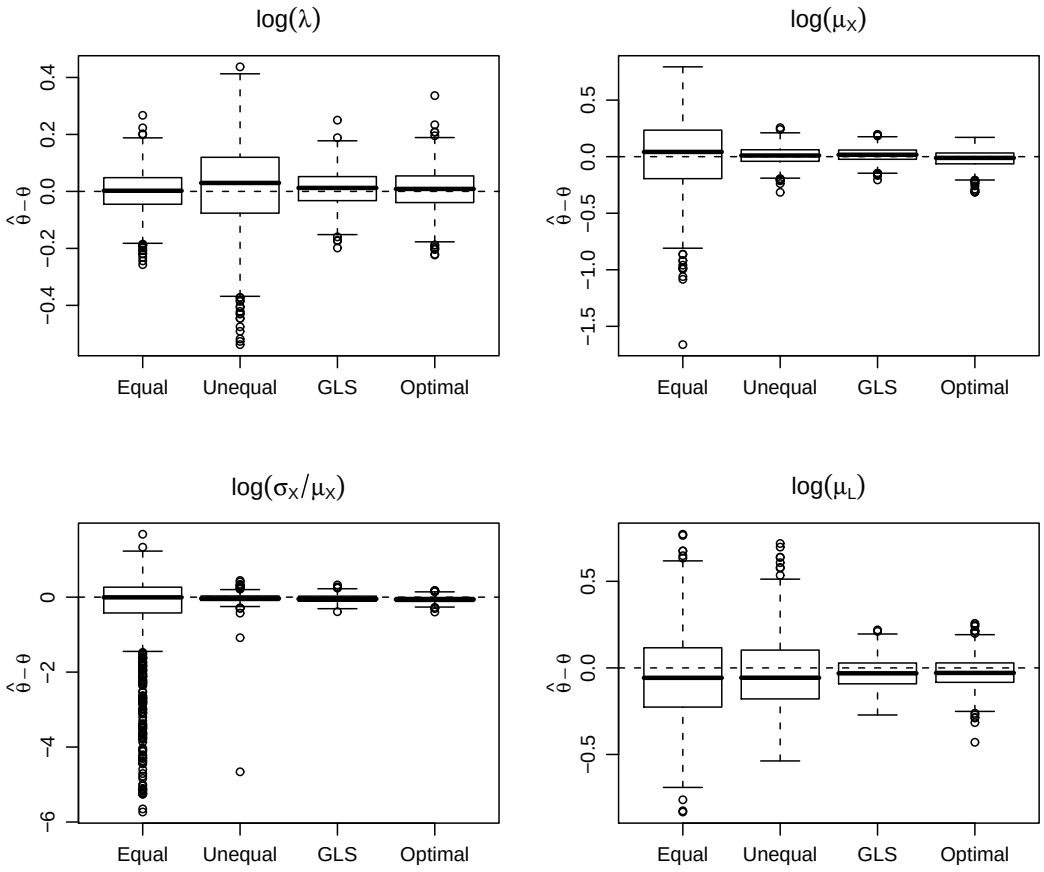

FIGURE 4.2: Distribution of estimation errors for each PRPM parameter, obtained using different weighting matrices in the GMM estimator. All distributions are obtained from 1000 simulated data sets, each containing 20 independent 30-day sequences and generated using parameter values $\log(\lambda) = -3.5, \log(\mu_X) = 0, \log(\sigma_X/\mu_X) = 0, \log(\mu_L) = 1.1$

4.2.2.2 Bias

In Table 4.2 we can see the estimated biases for each parameter from the 1000 simulations, together with their standard errors. From the study setting we can identify two potential sources of bias. One is that the moment conditions themselves are not exactly unbiased, in particular the sample variances and autocorrelations. The other is the finite sample bias, since the estimator defined by (3.8) is only asymptotically unbiased even when exactly unbiased moment conditions are used. In this particular case we have strong evidence to reject the hypothesis of the estimator being unbiased; however from Figure 4.2 we can see that for all weighting matrices, any bias is negligible compared with sampling variability.

		Parameters			
		$\log(\lambda)$	$\log(\mu_X)$	$\log(\sigma_X/\mu_X)$	$\log(\mu_L)$
Weighting Scheme	$\mathbf{W}_1$	0.000(0.00007)	0.008(0.0003)	-0.362(0.001)	-0.044(0.0003)
	$\mathbf{W}_2$	0.012(0.0001)	0.009(0.00008)	-0.031(0.0002)	-0.028(0.0002)
	$\mathbf{W}_3$	0.010(0.00006)	0.016(0.00006)	-0.041(0.0001)	-0.033(0.00009)
	$\mathbf{W}_0$	0.007(0.00007)	-0.017(0.00008)	-0.062(0.00008)	-0.028(0.00009)

TABLE 4.2: Estimated bias for each parameter under different weighting schemes, together with their standard errors.

		B			
		$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_0$
A	$\mathbf{W}_1$	—	-0.028	0.000	-0.001
	$\mathbf{W}_2$	-1.443	—	0.000	-0.002
	$\mathbf{W}_3$	-1.465	-0.057	—	-0.002
	$\mathbf{W}_0$	-1.468	-0.056	-0.004	—

TABLE 4.3: Minimum of the eigenvalues of the matrix resulting from the difference $(\overline{\text{Var}}(\hat{\boldsymbol{\theta}}_A) - \overline{\text{Var}}(\hat{\boldsymbol{\theta}}_B))$, for each combination of weighting schemes

4.2.2.3 Efficiency

In Table 4.2, the standard errors are obtained as the square roots of the diagonal elements of $\overline{\text{Var}}(\hat{\boldsymbol{\theta}})/1000$ so as to provide standard errors for the mean of the simulated estimates. Clearly however, their relative magnitudes are unaffected by this scaling and therefore the standard errors in Table 4.2 also provide a means of comparing the variability of the estimates themselves (rather than their means).

At a first glance we can see that the naive estimator, $\hat{\boldsymbol{\theta}}(\mathbf{W}_1)$, has generally larger standard errors, particularly when compared to $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$. However for, $\log(\lambda)$, the standard error of the estimator $\hat{\boldsymbol{\theta}}(\mathbf{W}_1)$ is lower than of the estimator $\hat{\boldsymbol{\theta}}(\mathbf{W}_2)$. A similar analysis can be made between $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$, where $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ is only better for two of the parameters. The most useful conclusion is perhaps that either of $\mathbf{W}_3$ and $\mathbf{W}_0$ perform better than both $\mathbf{W}_1$ and $\mathbf{W}_2$, and that these figures agree with the boxplot in Figure 4.2.

A first comparison of the full empirical covariance matrices, $\overline{\text{Var}}(\hat{\boldsymbol{\theta}})$, under different weighting schemes is done by calculating the difference between each pair and checking if the resulting matrix is positive definite by calculating the eigenvalues: if all eigenvalues are positive we have a positive definite matrix, while if some are positive and some are zero then we have a positive semi-definite matrix. Table 4.3 shows the minimum of the eigenvalues for each pair of matrices $(\mathbf{A}, \mathbf{B})$.

	$\log(\lambda)$	$\log(\mu_X)$	$\log(\sigma_X/\mu_X)$	$\log(\mu_L)$
$\mathbf{W}_1$	0.070	0.175	306.439	0.211
$\mathbf{W}_2$	0.129	0.075	16.128	0.175
$\mathbf{W}_3$	0.063	0.064	0.083	0.085
$\mathbf{W}_0$	0.054	0.052	0.057	0.060

TABLE 4.4: Standard errors obtained by averaging the theoretical covariance matrices obtained in each simulation, for each parameter and weighting scheme

There are only two non-negative elements in Table 4.3, corresponding to the matrices $\overline{\text{Var}}(\hat{\boldsymbol{\theta}}_{\mathbf{W}_1}) - \overline{\text{Var}}(\hat{\boldsymbol{\theta}}_{\mathbf{W}_3})$ and $\overline{\text{Var}}(\hat{\boldsymbol{\theta}}_{\mathbf{W}_2}) - \overline{\text{Var}}(\hat{\boldsymbol{\theta}}_{\mathbf{W}_3})$. Nevertheless the positive definite ordering is a very strong requirement, although the optimal choice should satisfy it asymptotically as shown in section 3.2.

4.2.2.4 Variance estimation

As previously mentioned the empirical covariance matrices used in Tables 4.2 and 4.3 cannot be obtained in practical applications where a single realization is available. Therefore we now examine the theoretical covariance matrices estimated for each simulation, as these are the ones used for inference in practice. A first step is to obtain the average of the covariance matrix estimates, and compare the resulting standard errors across weighting schemes with the ones obtained from the empirical covariance matrix.

Table 4.4 shows that on average the estimated standard deviation of the estimator is smaller when optimal weights are used, followed closely by $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$; the values corresponding to the equal weights and the unequal fixed weights are significantly larger which is not surprising and agrees with our initial analysis using Figure 4.2. If we compare these theoretical standard errors with the empirical ones, we can see that the use of asymptotics to build a covariance matrix of the estimator leads to underestimation of estimator variability in the cases $\hat{\boldsymbol{\theta}}(\mathbf{W}_1)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_2)$ for all parameters except $\log(\sigma_X/\mu_X)$. For this parameter the theoretical standard errors are much higher than their empirical counterparts. This high value is explained by the presence of outlying estimates that do not appear in the other weighting schemes, showing that these weights may carry some instability in the algorithm meaning that inference based on fixed weights can lead to estimates that are very far from the true value. The theoretical standard errors for $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ are very similar to the empirical ones. For $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ the theoretical standard errors are lower than their empirical counterparts: thus for this weighting scheme the theory seems to underestimate estimator variability, making the estimator looking more precise than it actually is.

To analyse the behaviour of the estimated standard errors obtained across simulations we plot estimates of their densities in Figure 4.3. We can see that the standard errors for $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ (bottom two rows of Figure 4.3) are not just smaller overall: they are also more consistent from simulation to simulation. The standard error for the estimator of $\log(\sigma_X/\mu_X)$ is less variable when using $\mathbf{W}_0$ instead of $\mathbf{W}_3$. For $\mathbf{W}_1$ and $\mathbf{W}_2$ (top two rows of Figure 4.3) the densities are substantially skewed, and have a dispersion not comparable with $\mathbf{W}_3$ or $\mathbf{W}_0$.

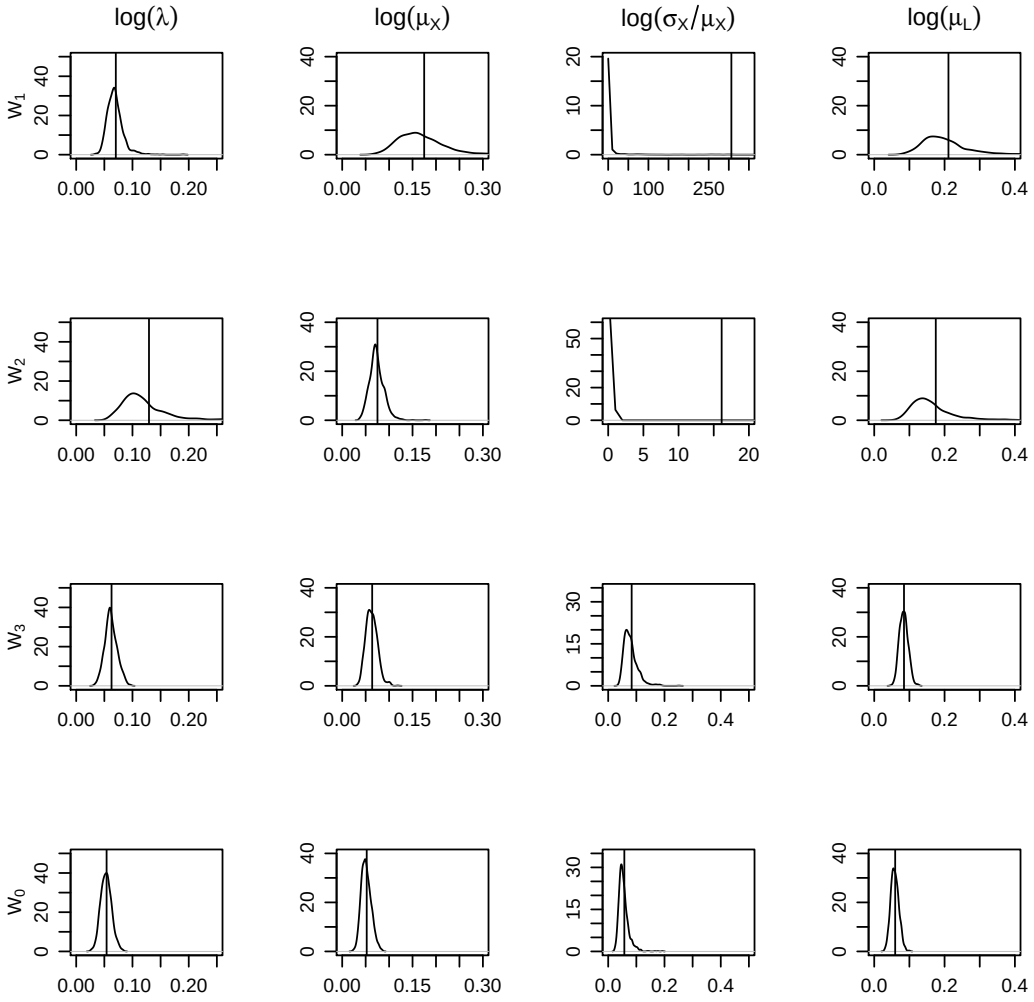

FIGURE 4.3: Estimated densities of theoretical standard errors from 1000 simulations together with the "average" standard errors (vertical lines). For each parameter except $\log(\sigma_X/\mu_X)$, the axis scales are the same for each weighting scheme.

In Table 4.5, overall estimator variability is compared in terms of the volume of the confidence region defined by the covariance matrix of the estimator. Since we have two different estimators for this matrix we also compare these. The figures in this table agree with our previous analysis that the weighting schemes, $\mathbf{W}_3$ and $\mathbf{W}_0$, obtained from the

	$det\left(\overline{\text{Var}}(\hat{\boldsymbol{\theta}})\right)^{1/8}$	$det\left(\overline{\hat{\text{Var}}}(\hat{\boldsymbol{\theta}})\right)^{1/8}$
$\mathbf{W}_1$	0.23	0.77
$\mathbf{W}_2$	0.11	0.33
$\mathbf{W}_3$	0.072	0.07
$\mathbf{W}_0$	0.075	0.054

TABLE 4.5: $det\left(\overline{\text{Var}}(\hat{\boldsymbol{\theta}})\right)^{1/8}$ and $det\left(\overline{\hat{\text{Var}}}(\hat{\boldsymbol{\theta}})\right)^{1/8}$ for different weighting schemes

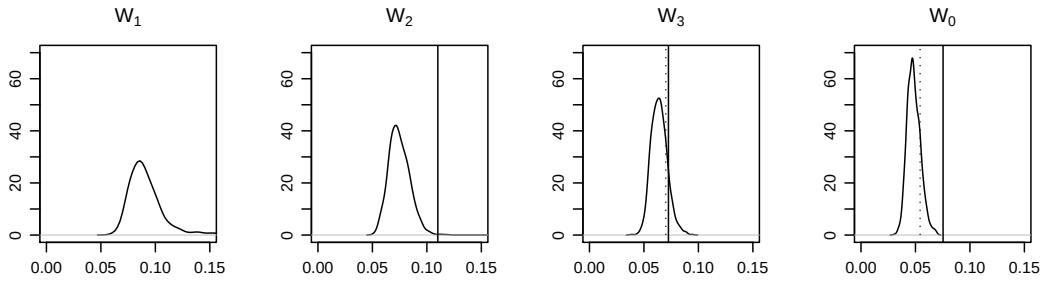

FIGURE 4.4: Estimated densities for the determinants of the theoretical covariances from 1000 simulations, together with $det\left(\overline{\text{Var}}(\hat{\boldsymbol{\theta}})\right)^{1/8}$ (solid line) and $det\left(\overline{\hat{\text{Var}}}(\hat{\boldsymbol{\theta}})\right)^{1/8}$ (dashed line) for different weighting schemes

variance of the estimating functions have a better performance. However comparison between the determinants of the empirical and theoretical covariance matrices shows that the theoretical covariance matrix overestimates estimator variability for $\hat{\boldsymbol{\theta}}(\mathbf{W}_1)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_2)$, and clearly underestimates this variability for $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$.

Figure 4.4 shows the estimated densities of the 8^{th}root of the determinant of the theoretical covariance matrix together with vertical lines at the values from Table 4.5. These lines are included for reference only: their interpretation cannot be associated with the mean of the distribution of determinants. We can see from Figure 4.4 that the determinants of the covariance matrices are more concentrated for $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ than for $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$, however Table 4.5 shows that the 8^{th}root of the determinant of the theoretical covariance is a biased estimator of the 8^{th}root of the determinant of the empirical covariance matrix in the $\mathbf{W}_0$ case, this suggests underestimation of estimator variability and we will pursue this further later in this chapter.

4.2.2.5 Confidence Intervals and Regions

One way to check the quality and accuracy of the inference, namely under/over estimation of estimator variability, is to analyse the coverage of confidence regions built

Weights	Conf. Level	$\log(\lambda)$	$\log(\mu_x)$	$\log(\sigma_x/\mu_x)$	$\log(\mu_d)$
$\mathbf{W}_1$	95%	0.93	0.68	0.55	0.86
	99%	0.98	0.80	0.65	0.94
$\mathbf{W}_2$	95%	0.87	0.92	0.84	0.84
	99%	0.95	0.97	0.93	0.93
$\mathbf{W}_3$	95%	0.94	0.92	0.83	0.92
	99%	0.98	0.97	0.92	0.98
$\mathbf{W}_0$	95%	0.84	0.81	0.68	0.82
	99%	0.93	0.90	0.80	0.90

TABLE 4.6: Coverage of confidence intervals built using normality assumption for each parameter and weighting scheme, for two different confidence levels.

according to the theory. Table 4.6 shows the proportion of simulations for which the confidence intervals built using the normality assumption and asymptotic standard errors included the true parameter value.

The results from Table 4.6 show that in general the coverage is well below its expectation. This may be due either to a failure of the normal approximation, and/or to poorly estimated standard errors. The first possibility is straightforward to investigate informally using normal probability plots; these are shown in Figure 4.5. From these plots it is clear that for this sample size neither of the estimators for $\log(\sigma_x/\mu_x)$ using $\mathbf{W}_1$ and $\mathbf{W}_2$ are normally distributed. This is not a surprise considering that the boxplots in Figure 4.2 showed some skewness for these estimators. For $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ the estimates do indeed appear to be close to normally distributed with the exception of $\log(\mu_X)$ using $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$. These results suggest that the poor coverages for $\hat{\boldsymbol{\theta}}(\mathbf{W}_1)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_2)$ in Table 4.6 may be partially due to lack of normality; but those for $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$, which are closer to the nominal levels, are primarily due to poor estimation of the standard errors. In fact we have already seen from Table 4.4 that the theoretical covariance matrix used in computing these confidence intervals, tends to underestimate the real variability of the estimator.

Table 4.7 shows coverages for the confidence regions built using a quadratic approximation to the objective function (3.8). The results for $\hat{\boldsymbol{\theta}}(\mathbf{W}_1)$,$\hat{\boldsymbol{\theta}}(\mathbf{W}_2)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ are similar, although the coverages are too low; the results for $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ are much worse, however. This reinforces the idea that the variability of $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ is significantly underestimated, using asymptotic theory, at least as implemented here.

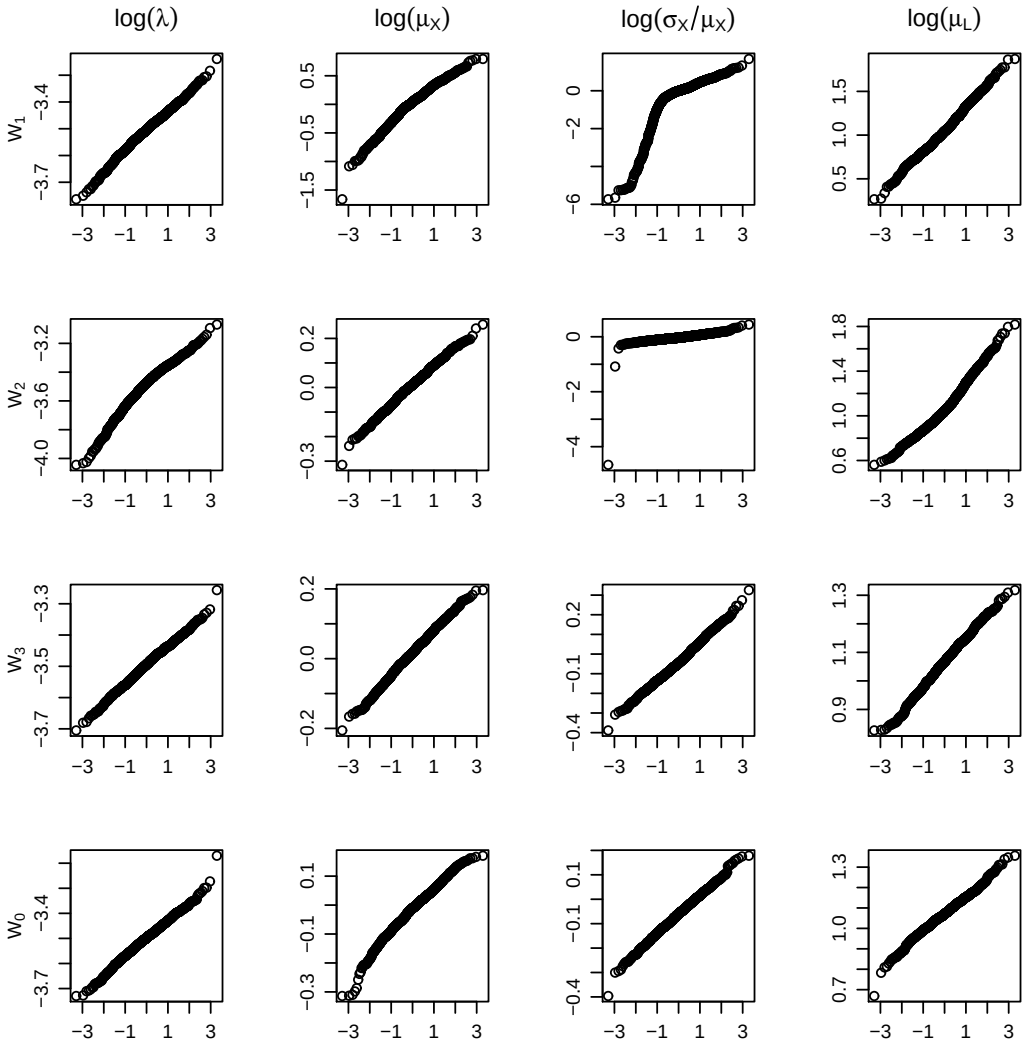

FIGURE 4.5: Normal probability plot of the estimates for each parameter under different weights.

	95%	99%
$\mathbf{W}_1$	0.85	0.92
$\mathbf{W}_2$	0.85	0.92
$\mathbf{W}_3$	0.89	0.95
$\mathbf{W}_0$	0.40	0.53

TABLE 4.7: Coverage of confidence region based on objective function threshold for each weighting scheme and two confidence levels

4.2.2.6 Some conclusions

From the analysis in this section we can conclude that $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ lead to a smaller estimator variance, however the use of the theory from section 3.2 does not lead to accurate confidence regions for the parameters using this estimators. Moreover there is clear underestimation of variability when using the theoretical results for $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$. With the elements at hand in this section, and for this particular model and simulation setup $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ seems to be the wise choice. The theoretical optimum seems to perform in terms of obtaining the estimates themselves but it fails in the estimation of sampling error, i.e. a poor estimation of the covariance matrix of the estimator, therefore later in this chapter we will try to improve the estimation of such matrix focusing on the weighting schemes $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$. In the next section we investigate different ways of improving inference for this model, this includes analysing the effect of the bias in the EFs and looking into estimation of the covariance of the EFs.

4.2.2.7 Improved performance

One of the reasons for the poor coverage performance could be that 20 years of data are not enough to apply the asymptotic approximations, or the bias in the EFs due to the 30 days per year method, in this section we investigate these potential issues.

We first looked at increasing the sample size by simulating more years, keeping the method the same but using 100 yrs instead of 20 yrs; the results improved as one would expect from applying asymptotic approximations to a larger dataset. The coverage of the confidence regions at 99% based on the objective function reached 0.98 and 0.92 for $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$ and $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ cases respectively, an improvement from the values of 0.95 and 0.53 in the original setting. In particular the coverage corresponding to $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ improves significantly, however it is still lower than that obtained using $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$. For the confidence intervals at 99% for the individual parameters the highest coverage was approximately 0.98 for both weighting schemes, for the parameter $\log(\lambda)$. These results may be more useful as a reference for the remainder of this section than for practical application, as 100 yrs of hourly rainfall data may be difficult to find in practice.

Among the potential causes for the low coverage shown in the previous section is the fact that we are using slightly biased estimating functions. We have seen that the resulting bias in the estimator is not significant, however the bias in the estimating functions may have an effect on the coverage of confidence intervals or regions. One possible way of investigating this is to perform a similar study as the one above but increase

	100 yrs × 30 days		30 yrs × 100 days	
	95%	99%	95%	99%
$\mathbf{W}_3$	0.92	0.98	0.92	0.97
$\mathbf{W}_0$	0.83	0.92	0.70	0.81

TABLE 4.8: Coverage of confidence region based on objective function threshold under two different settings of sample size, for two weighting schemes and two confidence levels

the number of days in the month, i.e., using the same amount of data but re-arranged in a different way, instead of 100 yrs x 30 days, we calculated the moment conditions and their covariance matrix using 30 yrs x 100 days. This way we can have an idea of how the inference improves by reducing the bias in the EFs. The results improved slightly compared with those from the previous section, but were actually worse than those obtained by simply increasing the number of years. For example, the coverage for the confidence region obtained from the objective function at 95% level, was 0.69 compared with 0.83 in the "100 years × 30 days " setting.

This suggests that the bias present in the EFs that were defined in this study is not the main cause for the poor coverages, however we look at a different way of evaluating that using bootsrapping. The methods described in section 3.1 can be applied using the original moment conditions (4.8), in which case we have to obtain the covariance matrix of the moment conditions using a different technique. One possible way to do that is through bootstrapping. The application of bootstrapping in order to avoid the "30 days" bias will allow us to keep the original moment conditions but will add extra computational burden. The method consists of sampling with replacement 20 elements from the set of 20 independent years, i.e. each period of 30 days is an element; a certain number of times, B. For each resampled sequence we calculate the moment conditions,

$$\tilde{\mathbf{h}}_i^{(b)}(\boldsymbol{\theta};\mathbf{y}) \quad b = 1, ..., B$$

The matrix $\hat{\mathbf{S}}^b$, the covariance matrix of the original sample moments estimated using bootstrapping has $(i,j)th$ element

$$\frac{1}{B-1}\sum_{b=1}^{B}\left(\tilde{\mathbf{h}}_i^{(b)}(\boldsymbol{\theta};\mathbf{y}) - \overline{h}(\boldsymbol{\theta};\mathbf{y})\right)\left(\tilde{\mathbf{h}}_j^{(b)}(\boldsymbol{\theta};\mathbf{y}) - \overline{h}(\boldsymbol{\theta};\mathbf{y})\right)$$

	$\log(\lambda)$	$\log(\mu_x)$	$\log(\sigma_x/\mu_x)$	$\log(\mu_d)$
$\mathbf{W}_3(95\%)$	0.90	0.92	0.82	0.86
$\mathbf{W}_3(99\%)$	0.96	0.98	0.91	0.93
$\mathbf{W}_0(95\%)$	0.84	0.81	0.68	0.76
$\mathbf{W}_0(99\%)$	0.91	0.90	0.82	0.86

TABLE 4.9: Coverage of confidence intervals built using normality assumption for each parameter and weighting scheme, for two different confidence levels, when bootstrapping is used

	95%	99%
$\mathbf{W}_3$	0.88	0.94
$\mathbf{W}_0$	0.36	0.49

TABLE 4.10: Coverage of confidence region based on objective function threshold for each weighting scheme and two confidence levels, when bootstrapping is used

The coverages resulting from the simulation study with bootstrapping ($B = 1000$) are shown in Tables 4.9 and 4.10. When comparing these with the original study in the previous section (Tables 4.6 and 4.7), we cannot see any clear improvement. Moreover for $\hat{\boldsymbol{\theta}}(\mathbf{W}_0)$ the objective function based confidence region has clearly lower coverage.

The results in this section suggest that the reduction of bias in the estimating functions does not lead to improved coverage performance, and hence that the presence of bias is not the main cause of low coverages.

4.2.2.8 Improved estimation of S

In section 4.2.2.4 it was noted that for the weighting matrix $\mathbf{W}_0$, the theoretical covariance matrix underestimates the real estimator variance. Therefore we look in more detail into estimation of $\text{Var}(\hat{\boldsymbol{\theta}})$ we focus on the estimation of $\mathbf{S}$ the covariance matrix of the moment conditions. Since this matrix is used in a slightly different way in depending on the weighting scheme considered, in particular for the weighting scheme $\mathbf{W}_0$ it defines the whole weighting matrix. An initial approach was to calculate $\hat{\mathbf{S}}$ from a long simulation (1000 years) using the true parameter value and then use it as a fixed value weighting matrix in the simulation routine. This in fact led to much improved coverages; however this approach would not be feasible in practical application where less than 1000 years of data is available. As described in section 3.2, the application of optimal GMM weights in econometrics is usually done by an iterative procedure, where the estimation of $\boldsymbol{\theta}$ by a consistent (but sub-optimal) estimator in a first step will provide an initial estimate of $\mathbf{S}(\boldsymbol{\theta})$ if an analytical expression for it is available. In the present context this is not feasible since an expression for $\mathbf{S}(\boldsymbol{\theta})$ is not available. For example,

	$\log(\lambda)$		$\log(\mu_x)$		$\log(\sigma_x/\mu_x)$		$\log(\mu_d)$	
	Emp	Theo	Emp	Theo	Emp	Theo	Emp	Theo
$\mathbf{W}_3^*$	0.064	0.065	0.063	0.063	0.096	0.084	0.086	0.088
$\mathbf{W}_0^*$	0.063	0.062	0.060	0.059	0.071	0.066	0.069	0.068

TABLE 4.11: Standard errors obtained from the empirical and average theoretical covariance matrices calculated using a two step procedure, for each parameter and for two different weighting schemes

the variance of the h-hour sample variance requires knowledge of the fourth order properties of the aggregated rainfall process; even for the relatively simple models considered here the results for the first and second order moments in section 4.1. suggest that the amount of work involved would be prohibitive. We therefore propose an alternative that can be used for any model where analytical evaluation of $\mathbf{S}(\boldsymbol{\theta})$ is impractical but where simulation is possible.

From the analysis in the previous section, where we used 100 years of data and the asymptotic approximations seemed to be more accurate, we concluded that one potential reason for the inaccuracy in estimating $\mathbf{S}(\boldsymbol{\theta})$ was the fact that we were using only 20 years of data, meaning 20 independent observation vectors to estimate an 8 by 8 symmetrical matrix. A possible solution is to use the 20 years of data to obtain an initial estimate of $\boldsymbol{\theta}$ and then use this estimate to simulate a longer series (1000 years, say), from which an improved estimate of $\mathbf{S}(\boldsymbol{\theta})$ can be obtained using (4.9). Denote by $\tilde{S}$ the estimator for $\mathbf{S}(\boldsymbol{\theta})$ using the simulated series. This is in fact equivalent to parameterizing the covariance matrix as a function of $\boldsymbol{\theta}$. The initial estimate of $\boldsymbol{\theta}$ in our case is obtained by using $\hat{\boldsymbol{\theta}}(\mathbf{W}_3)$, which was shown in the previous section to be a reasonable estimator. We then estimate $\boldsymbol{\theta}$ using a new set of weights $\mathbf{W}_3^*$ similar to $\mathbf{W}_3$ but calculated from the simulated series, and using the optimal weights based on $\tilde{S}$; the empirical and theoretical standard errors are shown in Table 4.11. This two step estimator has much better finite sample properties than the original setting for the weighting scheme $\mathbf{W}_0$; whereas for the weighting scheme $\mathbf{W}_3$ there are no significant improvements. Moreover, the values in Table 4.11 show that under this setting $\mathbf{W}_0^*$ gives smaller standard errors than $\mathbf{W}_3$, for all the parameters. The agreement between theoretical and empirical standard errors seem to be much better.

As before we use the coverages of confidence intervals/regions obtained using asymptotic theory to validate the inference, in particular the accuracy in estimating the estimator variability. The results in Tables 4.12 and 4.13 show an improvement in the $\mathbf{W}_0$ estimator, where the $\mathbf{W}_3$ estimator shows similar coverages to the original study. Although the coverage can be improved by estimating $\mathbf{S}(\boldsymbol{\theta})$ in this way they are still below expected.

	$\log(\lambda)$	$\log(\mu_x)$	$\log(\sigma_x/\mu_x)$	$\log(\mu_d)$
$\mathbf{W}_3^*(95\%)$	0.93	0.92	0.85	0.92
$\mathbf{W}_3^*(99\%)$	0.98	0.97	0.91	0.98
$\mathbf{W}_0^*(95\%)$	0.93	0.92	0.87	0.93
$\mathbf{W}_0^*(99\%)$	0.98	0.98	0.94	0.98

TABLE 4.12: Coverage of confidence intervals built using normality assumption for each parameter and weighting scheme, for two different confidence levels, when a two step procedure is used

	95%	99%
$\mathbf{W}_3^*$	0.91	0.97
$\mathbf{W}_0^*$	0.91	0.97

TABLE 4.13: Coverage of confidence region based on objective function threshold for each weighting scheme and two confidence levels, when a two step procedure is used

The other point to notice is that $\hat{\boldsymbol{\theta}}(\mathbf{W}_0^*)$ shows better coverages than $\hat{\boldsymbol{\theta}}(\mathbf{W}_3^*)$, which together with the finding already mentioned that $\hat{\boldsymbol{\theta}}(\mathbf{W}_0^*)$ has lower variance than $\boldsymbol{\theta}(\mathbf{W}_3^*)$ makes this two step procedure an important improvement for inference using this model with relatively modest sample sizes. The main drawback is the computing time which can be several times more than a simpler one step estimation using, say $\mathbf{W}_3$.

4.2.3 Extension to Neyman-Scott Rectangular Pulses Model

We also performed a similar study using a Neyman-Scott Rectangular Pulses Model, this model is one of the existing extensions to the PRPM in the previous section and the particular choice of this model is due to its use in practical applications, namely the UK Climate Impact Programme. The study of finite sample performance applied to this model tries to show that the use of asymptotic results like the ones in chapter 3. can improve inference in practical application and contribute to improve current practice. We used the same eight moment conditions (mean, variance, lag 1 autocorrelation and proportion of dry intervals at hourly resolution; variance at 6-hourly resolution; variance, lag 1 autocorrelation and proportion of dry intervals at 24-hourly resolution), and generated the data in the same way ,i.e., 20 sets of 30 days. The parameter vector of interest is $\boldsymbol{\theta} = (\log(\lambda), \log(\mu_x), \log(\sigma_x/\mu_x), \log(\mu_C), \log(\beta), \log(\eta))$, and the specific values used for simulation were $\boldsymbol{\theta} = (-4, -0.44, 0, 2.46, -1.8, 0.37)$, which were obtained from hourly rainfall data for January from Birmingham airport, as in the PRPM study of the previous section. In this section we skipped the use of $\mathbf{W}_0$ since we have seen that even for the simpler PRPM the calculation of the optimal weights without the two step procedure led to poor inference on estimator uncertainty. We also omitted the

results from $\mathbf{W}_3$ as this is used only for the first step of the two step procedure. We now apply the two step procedure to the estimation of the Neyman-Scott Rectangular Pulses Model.

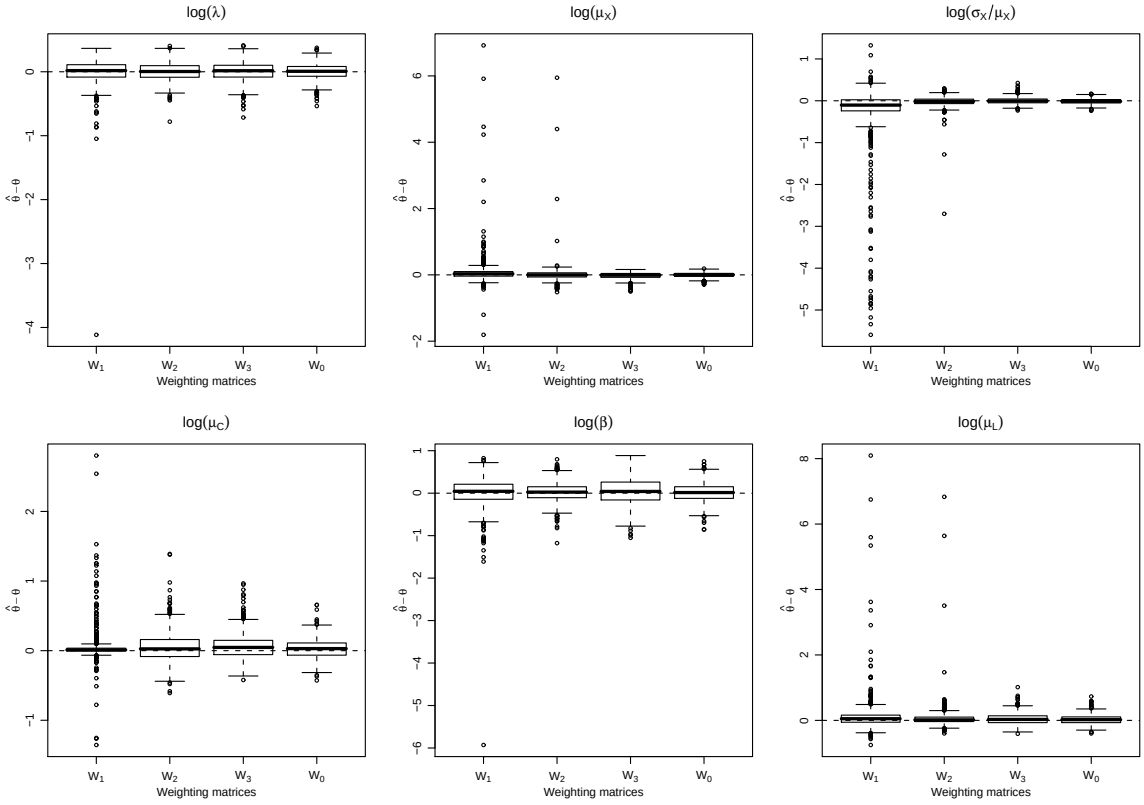

FIGURE 4.6: Boxplot of estimation errors for different weighting matrices

The analysis of Figure 4.6 leads to similar conclusions to the PRPM case in the sense that the estimates using $\mathbf{W}_3^*$ and $\mathbf{W}_0^*$ show less variation than the estimates obtained using the two data-independent weighting schemes. The reduction in variability from $\mathbf{W}_1$ to $\mathbf{W}_0^*$ can also be seen in Table 4.14, by noticing that the standard errors in brackets become smaller when looking down each individual parameter. In Table 4.15 we can see the reduction in volume of the confidence regions: for $\mathbf{W}_1$ we had some simulations returning a theoretical covariance matrix with non-finite elements which is in fact a typical problem in this type of models, for $\mathbf{W}_2$ the numbers in Table 4.15 show that the theory overestimates estimator uncertainty; and finally we can see in Tables 4.16. and 4.17.that both $\mathbf{W}_3^*$ and $\mathbf{W}_0^*$ lead to the smallest confidence regions and that the agreement between theoretical and empirical standard errors mean that the two step procedure works for this model in the same way it worked for the simpler model in the previous section. The performance of the different weights evaluated in terms of coverages of confidence intervals/regions has similar results to the PRPM case, although

	Parameters					
	$\log(\lambda)$	$\log(\mu_x)$	$\log(\sigma_x/\mu_x)$	$\log(\mu_C)$	$\log(\beta)$	$\log(\eta)$
$\mathbf{W}_1$	0	0.068	-0.224	0.043	0.013	0.103
	(0.000205)	(0.000401)	(0.000719)	(0.000214)	(0.000357)	(0.000507)
$\mathbf{W}_2$	0.001	0.005	-0.019	0.049	0.025	0.055
	(0.000136)	(0.000266)	(0.000128)	(0.000208)	(0.000216)	(0.000331)
$\mathbf{W}_3^*$	0.006	-0.024	0.001	0.062	0.046	0.046
	(0.000146)	(0.000098)	(0.000076)	(0.000175)	(0.0003)	(0.000166)
$\mathbf{W}_0^*$	0.002	-0.004	-0.011	0.027	0.015	0.025
	(0.000122)	(0.000072)	(0.000064)	(0.000137)	(0.000211)	(0.000135)

TABLE 4.14: Estimated bias for each parameter under different weighting schemes, together with their standard errors.

	Empirical	Theoretical
$\mathbf{W}_1$	0.097	—
$\mathbf{W}_2$	0.045	0.162
$\mathbf{W}_3^*$	0.026	0.028
$\mathbf{W}_0^*$	0.022	0.022

TABLE 4.15: $det\left(\overline{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\right)^{1/8}$ and $det\left(\overline{\hat{\mathrm{Var}}}(\hat{\boldsymbol{\theta}})\right)^{1/8}$ for different weighting schemes

Weights	Conf. Level	$\log(\lambda)$	$\log(\mu_x)$	$\log(\sigma_x/\mu_x)$	$\log(\mu_C)$	$\log(\beta)$	$\log(\eta)$
$\mathbf{W}_1$	95%	0.937	0.973	0.988	0.993	0.976	0.971
	99%	0.975	0.985	0.994	0.994	0.995	0.979
$\mathbf{W}_2$	95%	0.924	0.903	0.902	0.952	0.926	0.977
	99%	0.977	0.966	0.964	0.988	0.972	0.992
$\mathbf{W}_3^*$	95%	0.921	0.945	0.931	0.96	0.937	0.96
	99%	0.977	0.982	0.978	0.989	0.978	0.988
$\mathbf{W}_0^*$	95%	0.938	0.942	0.919	0.941	0.92	0.942
	99%	0.985	0.98	0.971	0.985	0.97	0.985

TABLE 4.16: Coverage of confidence intervals built using normality assumption for each parameter and weighting scheme, for two different confidence levels

the coverage rates for the confidence intervals built for individual parameters have similar values for all the weighting schemes the confidence regions based on the objective functions show a better coverage for the data-dependent weights when compared with the data-independent weights, in particular the theoretical optimum is also the best performing scheme for this set of simulations.

	95%	99%
$\mathbf{W}_1$	0.892	0.955
$\mathbf{W}_2$	0.893	0.956
$\mathbf{W}_3^*$	0.915	0.969
$\mathbf{W}_0^*$	0.938	0.984

TABLE 4.17: Coverage of confidence region based on objective function threshold for each weighting scheme and two confidence levels

4.3 Summary

In this chapter we looked at finite sample performance of the asymptotic theory described in Chapter 3, and compared the performance of different weighting matrices within the GMM framework. We compared estimators by measuring their variability via standard errors and volume of confidence region, and assessed the performance of the asymptotic theory regarding confidence region estimation by looking at the coverages of those confidence regions. We concluded that the asymptotic approximations can be used even for modest sample sizes, although in the case of data-dependent weights a two step procedure may be needed to obtain acceptable results. Moreover we can conclude that by using weights that are related to the variability of the moment condition we can perform better inference in terms of estimator variability. Here we considered that $\mathbf{W}_2$ is an example of current practice, and it was shown that even without a two step procedure the data-dependent weighting scheme $\mathbf{W}_3$ is an improvement for the PRPM case. The other objective of this study was to check the finite sample performance of the asymptotically optimal weights. The initial setting showed that the finite sample properties of the theoretical optimum can be quite poor, especially if moderate sample size are used. We can also conclude that the main reason for that poor performance is the difficulty in estimating $\mathbf{S}$, the covariance matrix of the moment conditions. We presented one possible way of improving the finite sample performance of the theoretical optimum by improving the estimation of $\mathbf{S}$. Furthermore we have shown that the theoretical optimum can be the finite sample optimum if we use a two step procedure to improve the estimation of $\mathbf{S}$. For the particular models studied here the fact that some of the moment conditions suffer from a finite sample bias does not seem to affect the inference outcome, this can be justified as bias being small compared to the variability of the moment conditions. It may be argued that, even with the use of the two step procedure, the coverages from $\mathbf{W}_0^*$ still tend to be less than their nominal values. However the difficulty of the inference task on the basis of a relatively small sample of data should not be underestimated, and for this particular area of application even ball-park figures are useful for current practice.

Chapter 5

Spectral Likelihood

In section 2.5.1 it was shown that maximum likelihood estimation is optimal for well behaved problems. The estimating equations framework applied to GMM presented in Chapter 2 shows that unbiased and consistent point estimation can be performed simply by finding a suitable moment matching condition. The main problem with this approach is that in general the estimates are strongly dependent on the properties chosen to build the estimating functions.

In this section, a spectral likelihood method known as the Whittle likelihood (Whittle, 1953) is presented. The motivation here is that the Fourier coefficients of a stationary stochastic process are approximately normally distributed and independent, therefore apparently solving the problem of specifying the joint density of the data; and since the Fourier transform is a one-to-one transformation of the data no information is lost, which seems to overcome the subjective choice of moments in the GMM framework. Nevertheless the use of the Fourier coefficients and their distribution also involves approximations resulting in some loss of efficiency.

Parametric estimation in the frequency domain has been widely studied, the main motivation being that for numerous stochastic processes it is difficult to write down an exact likelihood, where the spectral density may be straightforward to derive (Chandler, 1997; Fuentes, 2002). The Whittle likelihood is not the only possible frequency-domain estimation method: once the Fourier coefficients are seen as data, other inference methods can also be used. Rice (1979) compared the asymptotic properties of the Whittle estimator with two other estimators based on the sum of squared differences between the periodogram and spectral density. His findings argue in favour of the use of the Whittle likelihood approach, one of the reasons being that estimators based on the sum

of squares approach have an asymptotic bias. More recent work on spectral estimation by least squares has been done by Chiu (1988), who suggests that the parameters of time series can be estimated in the frequency domain using a weighted sum of squares instead of a likelihood. The advantages of this procedure are mainly computational, although by choosing the "right" weighting scheme one can obtain an estimator with the same efficiency as the Whittle likelihood estimator.

Although the original work of Whittle (1953) was based on the assumption that the original data were Gaussian, there has been extensive work on generalizing his results to cases where some of the assumptions are relaxed; Hannan (1973) proved the consistency and asymptotic normality of the Whittle estimator for the class of linear processes. Fox and Taqqu (1986) and Giraitis and Taqqu (1999) studied the effect of relaxing the no long-range dependency assumption, and showed that both for long-memory Gaussian processes and non-linear processes the estimator is consistent but that multivariate normality does not necessarily hold. Chandler (1997) presents an alternative way of deriving the Whittle likelihood, and shows that even if the multivariate normality does not hold for the whole set of Fourier frequencies considered in the Whittle likelihood, it may hold for a subset of these. Some authors argue that parametric estimation of non-linear processes in the frequency domain can be improved by using higher-order spectral densities together with the second-order spectral density used in the Whittle method, namely Anh et al. (2004) and Anh et al. (2007). In recent years there has been substantial work in applying the Whittle likelihood principles to particular classes of non-stationary series such as ARCH (Giraitis and Robinson, 2001); the work of Dahlhaus (2000) and Dahlhaus (2009) look at extending the Whittle likelihood to locally stationary processes, both for the univariate and multi-variate cases.

Even though there is extensive work in the area of parametric estimation in the frequency domain, this work has been focused mainly on relaxing the original Whittle (1953) assumptions, and not much progress has been made in the area of assessing estimator uncertainty. The work of Robinson (1978) provides an approximation for the asymptotic distribution of the Whittle estimator, however the result relies on the availability of an expression for the 4^{th} order spectral density. This is a major weakness in practical application as for numerous processes of interest such expression is not available Chandler (1997); Giraitis and Robinson (2001). Another important step in the direction of deriving an asymptotic distribution for the Whittle estimator was taken by Heyde (1997), this author makes use of the concept of estimating functions to obtain an asymptotic distribution for zero mean stationary processes. We concentrate our development of the Whittle estimator in the ability to calculating confidence intervals and

regions for the parameters. An important difference between the present approach and the methods in Heyde (1997) is that we allow our class of processes to have non-zero mean.

In this chapter we start with an outline of the general spectral theory and definitions, which will allow us to show some of the steps in the construction of the Whittle likelihood. We then present a set of results regarding convergence of functions of the periodogram to the spectral density. These will be used in the last section to show how the Whittle likelihood can be used in conjunction with the estimating functions framework to obtain an approximate asymptotic distribution for the Whittle estimator.

5.1 Definition of spectral likelihood

Spectral analysis has numerous applications in diverse fields. The focus in the present context is upon stationary stochastic processes with mean μ. Let (Y_t) be the process of interest, for which we have equally spaced observations. A single realization of a stationary process sampled at a finite number of equally spaced intervals can be represented as (Priestley, 1981, p.247)

$$y_t = \sum_{p=0}^{\lfloor \frac{n}{2} \rfloor} G_p e^{iw_p t} \quad , \tag{5.1}$$

where

$$G_p = \sum_{t=1}^{n} y_t e^{-iw_p t} \tag{5.2}$$

and

$$w_p = \frac{2\pi p}{n} \qquad p = 0, ..., \left\lfloor \frac{n}{2} \right\rfloor \quad . \tag{5.3}$$

G_p is called the discrete Fourier transform of the sequence $y_1, \ldots, y_n$. Denote the sample Fourier coefficients as

$$A_p = \sum_{t=1}^{n} y_t \cos(w_p t) \quad B_p = \sum_{t=1}^{n} -y_t \sin(w_p t) \quad , \tag{5.4}$$

which are respectively the real and imaginary parts of G_p as defined in (5.2). The Fourier coefficients contain all the information present in the original series y_t, as the series can be reconstructed using (5.1).

5.1.1 Cumulants and spectral densities

For reasons that will become clear later in this chapter we need to assume that, besides Y_t being stationary, it must not have long-range dependence. The long-range dependence assumption can be formulated in terms of cumulants. The cumulants of a random vector $\mathbf{Y}$ are defined by the cumulant generating function,

$$K(\mathbf{z}) = \log E[e^{\mathbf{z}^T \mathbf{Y}}]$$

where $\mathbf{z}$ is a vector the same length as $\mathbf{Y}$. In the case of stationary processes the cumulants can be defined as functions of the lags. The κth order cumulant, $c_\kappa(r_1, \ldots, r_{\kappa-1})$, of $\mathbf{Y}' = (Y_t, Y_{t+r_1}, \ldots, Y_{t+r_\kappa-1})$ is

$$c_\kappa(r_1, \ldots, r_{\kappa-1}) = \left. \frac{\partial^\kappa K(\mathbf{z})}{\partial z_1 \ldots \partial z_\kappa} \right|_{\mathbf{z}=\mathbf{0}}$$

where the r's are integers, representing lags. For $\kappa = 2$ we obtain $c_2(r) = \mathrm{Cov}(Y_t, Y_{t+r})$ which is the autocovariance at lag r.

The formulation of the assumption of no long-range dependence of (Y_t), is defined, as in Assumption I of Brillinger and Rosenblatt (1967)

$$\sum_{r_1=-\infty}^{\infty} \cdots \sum_{r_{\kappa-1}=-\infty}^{\infty} |c_\kappa(r_1, ..., r_{\kappa-1})| < \infty \tag{5.5}$$

for $\kappa \geq 2$, where $c_\kappa(r_1, ..., r_{\kappa-1})$ is the $\kappa - th$ order cumulant function of Y_t as defined above.

The stationarity assumption and (5.5) imply that the κth-order cumulant spectral density exists and is bounded. For all k Brillinger and Rosenblatt (1967), define the κth order spectral density as,

$$h^{(\kappa)}(\mathbf{w}) = \sum_{r_1=-\infty}^{\infty} \cdots \sum_{r_{\kappa-1}=-\infty}^{\infty} c_\kappa(r_1, \ldots, r_{\kappa-1}) e^{-i\sum_{j=1}^{\kappa-1} w_j r_j}$$

where $\mathbf{w} = (w_1, ..., w_{\kappa-1})^T \in (-\pi, \pi)^{\kappa-1}$. In particular let us write the second-order spectral density, defined as the Fourier transform of the autocovariance function, as it will be used frequently in this chapter

$$h(w) = \sum_{r=-\infty}^{\infty} c_2(r) e^{-iwr} \tag{5.6}$$

where w is any frequency in $(-\pi, \pi)$.

Estimation of the spectral density is usually based on the periodogram which represents the contribution of each of the frequencies w_p to the process Y_t, and is defined here as

$$I(w_p) = \frac{1}{n}\left(A_p^2 + B_p^2\right)$$

where A_p and B_p are the sample Fourier coefficients defined at (5.4). This definition corresponds to that in Hauser (1998). The estimation of the spectral density based on the periodogram can be justified by the properties of the sample Fourier coefficients.

5.1.2 Properties of the sample Fourier coefficients

The expected values, variances and asymptotic distribution of the random quantities A_p and B_p may be derived from the assumptions of stationarity of Y_t and (5.5). The derivations are standard and are included here to give a sense of the role of the various

assumptions. In this section we will use $c(r)$ rather than $c_2(r)$ to denote the autocovariance at lag r. For these derivations the following result is needed:

$$\sum_{t=1}^{n} e^{iw_p t} = \begin{cases} n & p = 0 \text{ (mod n)} \\ 0 & \text{otherwise.} \end{cases} \tag{5.7}$$

from which the implications on the real and imaginary parts of G_p follow. A straightforward application of (5.7) allows us to obtain the expected values of A_p and B_p as

$$E[A_p] = \begin{cases} n\mu & p = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.8}$$

and

$$E[B_p] = 0 \quad \text{for any } p \ . \tag{5.9}$$

We turn now to the variances of the Fourier coefficients. Consider first the expected value of A_p^2

$$\begin{aligned} E[A_p^2] &= E\left[\sum_{t=1}^{n}\sum_{s=1}^{n} y_t y_s \cos(w_p t)\cos(w_p s)\right] \\ &= \frac{1}{2}\sum_{t=1}^{n}\sum_{s=1}^{n}\left[c(|t-s|)+\mu^2\right]\left[\cos(w_p(t-s))+\cos(w_p(t+s))\right] \\ &= \frac{1}{2}\sum_{t=1}^{n}\sum_{s=1}^{n} c(|t-s|)\left[\cos(w_p(t-s))+\cos(w_p(t+s))\right] \\ &\quad +\frac{\mu^2}{2}\sum_{t=1}^{n}\sum_{s=1}^{n}\left[\cos(w_p(t-s))+\cos(w_p(t+s))\right] \\ &= \frac{1}{2}\sum_{t=1}^{n}\sum_{s=1}^{n} c(|t-s|) Re\left[e^{iw_p(t-s)}+e^{iw_p(t+s)}\right] \qquad (5.10) \\ &\quad +\frac{\mu^2}{2} Re\left[\sum_{t=1}^{n} e^{iw_p t}\left(\sum_{s=1}^{n} e^{-iw_p s}+\sum_{s=1}^{n} e^{iw_p s}\right)\right] \ . \qquad (5.11) \end{aligned}$$

For $p \neq 0$ the second term becomes zero as a consequence of (5.7); when $p = 0$ it yields $(n\mu)^2$. The first term can be evaluated by performing a change of variable ($v = t, r = t - s$), to obtain

$$
\begin{aligned}
& \frac{1}{2}\sum_{v=1}^{n}\sum_{r=v-n}^{v-1} c(|r|)\left[\cos(w_p r) + \cos(w_p(2v-r))\right] \\
= \; & \frac{1}{2}\sum_{r=1-n}^{0}\sum_{v=1}^{r+n} c(|r|)\cos(w_p r) + \frac{1}{2}\sum_{r=1}^{n-1}\sum_{v=r+1}^{n} c(|r|)\cos(w_p r) \\
& + \frac{1}{2}\sum_{r=1-n}^{0}\sum_{v=1}^{r+n} c(|r|)\cos(w_p(2v-r)) + \frac{1}{2}\sum_{r=1}^{n-1}\sum_{v=r+1}^{n} c(|r|\cos(w_p(2v-r)) \\
= \; & \frac{1}{2}\sum_{r=1-n}^{0}(n-|r|)c(|r|)\cos(w_p r) + \frac{1}{2}\sum_{r=1}^{n-1}(n-|r|)c(|r|)\cos(w_p r) \\
& + \frac{1}{2}\sum_{r=1-n}^{0}\sum_{v=1}^{r+n} c(|r|)Re\left[e^{iw_p(2v-r)}\right] + \frac{1}{2}\sum_{r=1}^{n-1}\sum_{v=r+1}^{n} c(|r|)\mathrm{Re}\left[e^{iw_p(2v-r)}\right] \\
= \; & \frac{1}{2}\sum_{r=1-n}^{n-1} n(1-\left|\frac{r}{n}\right|)c(|r|)\cos(w_p r) \qquad (5.12) \\
& + \frac{1}{2}\sum_{r=1-n}^{0}\sum_{v=1}^{r+n} c(|r|)Re\left[e^{iw_p(2v-r)}\right] + \frac{1}{2}\sum_{r=1}^{n-1}\sum_{v=r+1}^{n} c(|r|)\mathrm{Re}\left[e^{iw_p(2v-r)}\right] \qquad (5.13)
\end{aligned}
$$

This expression is valid for any n and cannot be simplified in general. However, it seems reasonable to try to find some asymptotic results that we can use to approximate this expression for large n. We consider (5.12) and (5.13) separately. First we will analyse the asymptotic behaviour of (5.12) multiplied by $2n^{-1}$ (the reason for this normalization will soon become clear). This yelds

$$\lim_{n\to\infty}\sum_{r=1-n}^{n-1}(1-\left|\frac{r}{n}\right|)c(|r|)\cos(w_p r) \quad . \qquad (5.14)$$

If we can show that

$$\lim_{n\to\infty}\sum_{r=1-n}^{n-1}\left|\frac{r}{n}\right| c(|r|)\cos(w_p r) = 0 \quad , \qquad (5.15)$$

then we can state that (5.14) is in fact equal to

$$\lim_{n\to\infty} \sum_{r=1-n}^{n-1} c(|r|)\cos(w_p r) \quad = \quad \lim_{n\to\infty} \sum_{r=1-n}^{n-1} c(r)\cos(w_p r) \tag{5.16}$$

since $c(r)$ is an even function (Priestley, 1981, p 214). Comparison with (5.6) shows that this limit is equal to $h(w_p)$, the spectral density at frequency w_p. Thus (5.14) tends to $h(w_p)$ as $n \to \infty$.

We now show a proof for (5.15). From the short-range dependence assumption (5.5) we have,

$$\sum_{r=-\infty}^{\infty} c(|r|) \leq \infty \quad \Rightarrow \quad \lim_{r\to\infty} |r|c(|r|) = 0$$

which means that for any $\epsilon > 0$ we can find R_ϵ such that $\forall\ |r| > R_\epsilon,\ |r|c(|r|) < \epsilon$. For any $n > R_\epsilon$, we have

$$\begin{aligned} \sum_{r=1-n}^{n-1} \left|\frac{r}{n}\right| c(|r|) \quad &= \quad \sum_{r=-R_\epsilon}^{R_\epsilon} \left|\frac{r}{n}\right| c(|r|) \ + \sum_{R_\epsilon<|r|\leq n-1} \left|\frac{r}{n}\right| c(|r|) \\ &< \quad \frac{R_\epsilon}{n} \sum_{r=-R_\epsilon}^{R_\epsilon} c(|r|) + (n-1-R_\epsilon)\frac{\epsilon}{n} \end{aligned}$$

which tends to ϵ as $n \to \infty$. Since ϵ can be made as small as we like, we must have

$$\lim_{n\to\infty} \sum_{r=1-n}^{n-1} \left|\frac{r}{n}\right| c(|r|) = 0$$

which together with the fact that $\cos(w_p r)$ is bounded gives the required result, (5.15).

Now we turn to (5.13), which can be written as

$$\frac{1}{2}c(0)\sum_{v=1}^{n} Re\left[e^{iw_p(2v)}\right] + \frac{1}{2}\sum_{r=1}^{n-1}\sum_{v=1}^{n-r} c(|r|)Re\left[e^{iw_p(2v+r)}\right] + \frac{1}{2}\sum_{r=1}^{n-1}\sum_{v=r+1}^{n} c(|r|)\mathrm{Re}\left[e^{iw_p(2v-r)}\right] .$$

The first term is zero due to (5.7). Changing the range of the last summation in v to start at 1 instead of $r+1$, and combining the two terms, yields

$$\begin{aligned} & Re\left[\sum_{r=1}^{n-1} c(|r|)\sum_{v=1}^{n-r} e^{iw_p(2v+r)}\right] \quad (w_p \neq 0,\pi) = Re\left[\sum_{r=1}^{n-1} c(|r|)e^{iw_p r}\frac{1-e^{2iw_p(n-r)}}{1-e^{2iw_p}}\right] \\ = \;& Re\left[\sum_{r=1}^{n-1} c(|r|)\frac{e^{iw_p r}-e^{-iw_p r}}{1-e^{2iw_p}}\right] = 2Re\left[\sum_{r=1}^{n-1} c(|r|)\sin(w_p r)\frac{1-e^{-2iw_p}}{2+2\cos(2w_p)}\right] \\ = \;& \frac{2}{2+2\cos(2w_p)}\sum_{r=1}^{n-1} c(|r|)\sin(w_p r)(1-\cos(2w_p)) \quad . \end{aligned}$$

If we multiply by $2n^{-1}$ as in (5.14) we obtain an expression that goes to zero as $n \to \infty$. Therefore, for fixed w and defining $q_n = \frac{wn}{2\pi}$

$$\lim_{n\to\infty}\frac{2}{n}V\left[A_q\right] = h(w) \;, \quad w \in (0,\pi).$$

Now we analyse the case $w_q = \pi$ (strictly speaking this requires us to consider that n is even). (5.13) is equal to

$$\frac{1}{2}Re\left[\sum_{r=1-n}^{0} c(|r|)e^{-i\pi r}\sum_{v=1}^{n+r} e^{2i\pi v}\right] + \frac{1}{2}Re\left[\sum_{r=1}^{n-1} c(|r|)e^{-i\pi r}\sum_{v=r+1}^{n} e^{2i\pi v}\right] \tag{5.17}$$

and the term including v in (5.17), $e^{2i\pi v}$ is always one. Therefore the expression for the variance simplifies to

$$\frac{1}{2}\sum_{r=1-n}^{n-1} n(1-\left|\frac{r}{n}\right|)c(|r|)\cos(w_p r)$$
$$+\frac{1}{2}\sum_{r=1-n}^{0} n(1-\left|\frac{r}{n}\right|)c(|r|)\cos(w_p r)$$
$$+\frac{1}{2}\sum_{r=1}^{n-1} n(1-\left|\frac{r}{n}\right|)c(|r|)\cos(w_p r) \quad .$$

Using the results (5.15) and (5.16) we can state that

$$\lim_{n\to\infty}\frac{2}{n}V\left[A_{n/2}\right] = h(\pi)$$

The variance of A_0 can be derived using similar steps,

$$V[A_0] = E[A_0^2] - E[A_0]^2$$

Noting that the evaluation of $E[A_p^2]$ leading to (5.10) and (5.11) is valid for any p and using (5.8), we have

$$V[A_0] = \frac{1}{2}\sum_{t=1}^{n}\sum_{s=1}^{n} c(|t-s|)Re\left[e^{iw_p(t-s)} + e^{iw_p(t+s)}\right] \tag{5.18}$$

Using the fact that for $p = 0$ the terms in the complex exponentials in (5.18) are always 1, together with the manipulation done to (5.10) yelding (5.12) and (5.13) we obtain

$$V[A_0] = \frac{1}{2}\sum_{r=1-n}^{n-1} n(1-\left|\frac{r}{n}\right|)c(|r|)$$
$$+\frac{1}{2}\sum_{r=1-n}^{0}(n-|r|)c(|r|) + \frac{1}{2}\sum_{r=1}^{n-1}(n-|r|)c(|r|)$$

Using (5.15) and (5.16) we can state,

$$\lim_{n\to\infty} \frac{1}{n} V\left[A_0\right] = h(0)$$

The derivation of the asymptotic variance of B_p is done using essentially the same steps, except that for $w_p = 0, \pi$, B_0 is identically zero. Thus for $p \neq 0, n/2$

$$\begin{aligned} V[B_p] &= E[B_p^2] = E\left[\sum_{t=1}^{n}\sum_{s=1}^{n} y_t y_s \sin(w_p t)\sin(w_p s)\right] \\ &= \frac{1}{2}\sum_{v=1}^{n}\sum_{r=v-n}^{v-1} c(|r|)\left[\cos(w_p r) - \cos(w_p(2v-r))\right] \quad . \end{aligned}$$

The only difference from the case A_p, $p \neq 0$, is the sign affecting the terms that tend to zero.

If the Y_ts are Gaussian, then A_p and B_p being linear combinations of these will be Gaussian as well. If other or no distributional assumption on the Y_ts is made we can still argue in favour of asymptotic normality and pairwise independence of the Fourier coefficients, the proof of this is rather technical and relies on the properties of the complex multivariate normal distribution; see Brillinger (1975, p.404). Putting all these results together yields the following large-sample distributions for the Fourier coefficients:

$$A_0 \sim MVN\left(n\mu, nh(0)\right) \tag{5.19}$$

$$A_p \sim MVN\left(0, \frac{n}{2}h(w_p)\right) \qquad p \neq 0$$

$$B_p \sim MVN\left(0, \frac{n}{2}h(w_p)\right) \qquad w_p \neq 0, \pi$$

$$B_p \equiv 0 \qquad w_p = 0, \pi \tag{5.20}$$

where μ is the mean of the original process.

Having transformed the data in this way, and noting that each of these new observations has a specified distribution it is desirable that a joint distribution for the sampled coefficients can be specified as well. Due to the approximation done in the derivations of the variances and also the application of CLT (in case the original data are not Gaussian), it is not true that the joint distribution of the Fourier coefficients is necessarily multivariate Normal. Despite this it can be shown that even if the MVN approximation is bad for the set of all Fourier coefficients it may be appropriate for a smaller collection of them. From this point onwards any reference made to the frequencies w_p is implicit that these belong to a collection ($\mathbf{\Omega}$) of Fourier frequencies for which the MVN approximation is adequate. For further discussion on the adequacy of the MVN distribution for collections of Fourier coefficients see Chandler (1997).

Now write $h(w_p;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters of interest. From (5.19)-(5.20) a likelihood for $\boldsymbol{\theta}$ can be formulated in the following way,

$$L(\boldsymbol{\theta}) = \prod_{p\in\mathbf{\Omega}} L(\boldsymbol{\theta}|A_p)L(\boldsymbol{\theta}|B_p)$$

Then the log-likelihood is given by (Chandler, 1997),

$$\begin{aligned}\log L(\boldsymbol{\theta}) = &-\sum_p \left[1-\frac{1}{2}\delta_{p(n/2)}\right]\left[\frac{I(w_p)}{h(w_p;\boldsymbol{\theta})}+\log(h(w_p;\boldsymbol{\theta}))\right]\\ &-\chi_{\mathbf{\Omega}}(0)\left[\frac{1}{2}\log(h(0;\boldsymbol{\theta}))+\frac{(A_0-n\mu(\boldsymbol{\theta}))^2}{2h(0;\boldsymbol{\theta})}\right]+\text{constant}\quad,\end{aligned} \tag{5.21}$$

where $\chi_{\mathbf{\Omega}}(p)$ is an indicator set function that takes the value 1 if $w_p \in \mathbf{\Omega}$ and 0 otherwise, and δ_{ij} is the Kronecker delta. The most common application of this method consists of using the spectral likelihood including all Fourier frequencies different from zero and π. In this context (5.21) is called the Whittle log-likelihood (Hauser, 1998)

5.2 Some useful convergence results

In this section we present some results that will be used in the remainder of the chapter, to show how the estimating function theory can be applied to the Whittle likelihood. The results are given without proof because the technical details are lengthy and irrelevant in this context. Consider in this whole section that Y_t is a zero mean process with $h(w;\boldsymbol{\theta}) < \infty$ and $\phi(w) \in Lip\ \xi,\ \xi > 1/2$.

The first two results ((5.22) and (5.23)) are a consequence of Theorem 2 in Robinson (1978) and its proof. They state the convergence of a weighted sum of periodogram ordinates to an asymptotically equivalent integral form.

$$\lim_{n\to\infty} \left| \frac{1}{n} \sum_{p=-\lfloor \frac{n-1}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} \phi(w_p) I(w_p) - \int_{-\pi}^{\pi} \phi(w) I(w) dw \right| = 0 \qquad \text{a.s.} \tag{5.22}$$

$$\lim_{n\to\infty} \left| \frac{1}{n} \sum_{p=-\lfloor \frac{n-1}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} \phi(w_p) I(w_p) - \int_{-\pi}^{\pi} \phi(w) h(w;\boldsymbol{\theta}) dw \right| = 0 \qquad \text{a.s.} \tag{5.23}$$

Another useful result can be obtained from the Newton-Cotes formula for integral approximation (Abramowitz and Stegun, 1964)

$$\int_a^b f(x)dx = \frac{1}{n} \sum_{p=0}^{n} f(x_p) + O\left(n^{-3} \frac{\partial^2 f(x)}{\partial x^2} \right). \tag{5.24}$$

where $x_p = a + p * (b-a)/n$, and $O\left(g(x;n)\right)$ represents a quantity that is smaller in absolute value than $g(x;n)$ times a constant when $n \to \infty$ for x fixed. Applied to the integral $\int_{-\pi}^{\pi} \phi(w,\boldsymbol{\theta}) h(w;\boldsymbol{\theta}) dw$, this result enables us to claim that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{p=-\lfloor \frac{n-1}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} \phi(w_p,\boldsymbol{\theta}) h(w_p;\boldsymbol{\theta}) - \int_{-\pi}^{\pi} \phi(w,\boldsymbol{\theta}) h(w;\boldsymbol{\theta}) dw = O\left(n^{-3} \frac{\partial^2 \phi(w,\boldsymbol{\theta}) h(w;\boldsymbol{\theta})}{\partial w^2} \right) \tag{5.25}$$

We will also make use of a result from Brillinger and Rosenblatt (1967), which states that the κ-order cumulant of the discrete Fourier transform of a stationary process can be approximated by a function of the κ-order spectral density.

$$c_\kappa(G_{p_1},...,G_{p_\kappa}) = (2\pi)^{\kappa-1}\,\Delta^{(n)}\left(\sum_{j=1}^{\kappa} w_{p_j}\right) h^{(\kappa)}(w_{p_1},...,w_{p_{\kappa-1}}) + O(1) \tag{5.26}$$

where G_p is the discrete Fourier transform defined in (5.2), and

$$\Delta^{(n)}(w_p) = \sum_{t=1}^{n} e^{-iw_p} \quad = \quad \begin{cases} n & p = 0 \\ 0 & \text{otherwise.} \end{cases}$$

The last result presented in this subsection states the convergence of the expectation and covariance matrix of integrals of linear functions of the periodogram. Define

$$\hat{\psi}_i(\boldsymbol{\theta}) = \int_{-\pi}^{\pi} \phi_i(w,\boldsymbol{\theta}) I(w) dw$$
$$\psi_i(\boldsymbol{\theta}) = \int_{-\pi}^{\pi} \phi_i(w,\boldsymbol{\theta}) h(w;\boldsymbol{\theta}) dw$$

(Priestley, 1981, p 427) shows that,

$$\lim_{n\to\infty} E\left[\hat{\psi}_i(\boldsymbol{\theta})\right] = \psi_i(\boldsymbol{\theta})$$
$$\lim_{n\to\infty} n\ \mathrm{Cov}(\hat{\psi}_i(\boldsymbol{\theta}), \hat{\psi}_j(\boldsymbol{\theta})) = e^{(4)}\psi_i(\boldsymbol{\theta})\psi_j(\boldsymbol{\theta}) + 4\pi \int_{-\pi}^{\pi} \phi_i(w,\boldsymbol{\theta})\overline{\phi}_j(w,\boldsymbol{\theta}) h^2(w;\boldsymbol{\theta}) dw \tag{5.27}$$

where $e^{(4)} = (E[\epsilon_t^4] - 3)$ and $\overline{\phi}_j(w,\boldsymbol{\theta}) = 1/2[\phi_j(w,\boldsymbol{\theta}) + \phi_j(-w,\boldsymbol{\theta})]$. Here ϵ is the purely random part of the process as defined in Wold's decomposition theorem for covariance-stationary processes (Kendall and Ord, 1990). From this point onwards we assume that $E[\epsilon_t^4]$ exists and is finite for the processes considered.

In the following section we derive the estimating functions defined by the Whittle likelihood and study their properties in the light of the results above.

5.3 Spectral Likelihood and Estimating functions

5.3.1 Rewriting the spectral scores

As was shown in setion 2.5.1, the score functions obtained when maximizing a log-likelihood are examples of estimating functions, and here it is exactly the same. We will now derive the expression for the partial derivatives of the Whittle log-likelihood (5.21) which we will call spectral scores. Differentiating with respect to θ_i yields

$$\begin{aligned}
g_i(\boldsymbol{\theta};\mathbf{y}) &= \frac{\partial \log L}{\partial \theta_i} \\
&= -\sum_{p\neq 0}\left[1-\frac{1}{2}\delta_{p(n/2)}\right]\left[-\frac{I(w_p)}{h(w_p;\boldsymbol{\theta})^2}\frac{\partial h(w_p;\boldsymbol{\theta})}{\partial \theta_i}+\frac{\partial h(w_p;\boldsymbol{\theta})}{\partial \theta_i}\frac{1}{h(w_p;\boldsymbol{\theta})}\right] \\
&\quad -\chi_{\boldsymbol{\Omega}}(0)\frac{1}{2}\left[\frac{\partial h(0;\boldsymbol{\theta})}{\partial \theta_i}\frac{1}{h(0;\boldsymbol{\theta})}\right] \\
&\quad -\chi_{\boldsymbol{\Omega}}(0)\frac{1}{2}\left[\frac{2(A_0-n\mu(\boldsymbol{\theta}))\left(-n\frac{\partial \mu(\boldsymbol{\theta})}{\partial \theta_i}\right)nh(0;\boldsymbol{\theta})-(A_0-n\mu(\boldsymbol{\theta}))^2 n\frac{\partial h(0;\boldsymbol{\theta})}{\partial \theta_i}}{n^2h(0;\boldsymbol{\theta})^2}\right] \\
&= -\sum_{p\neq 0}\left[1-\frac{1}{2}\delta_{p(n/2)}\right]\left[\frac{\partial h(w_p;\boldsymbol{\theta})}{\partial \theta_i}\frac{1}{h(w_p;\boldsymbol{\theta})^2}\left(h(w_p;\boldsymbol{\theta})-I(w_p)\right)\right] \\
&\quad -\chi_{\boldsymbol{\Omega}}(0)\frac{1}{2}\left[\frac{\partial h(0;\boldsymbol{\theta})}{\partial \theta_i}\frac{1}{h(0;\boldsymbol{\theta})}-\frac{\partial \mu(\boldsymbol{\theta})}{\partial \theta_i}\frac{2(A_0-n\mu(\boldsymbol{\theta}))}{h(0;\boldsymbol{\theta})}\right. \\
&\quad \left.-\frac{\partial h(0;\boldsymbol{\theta})}{\partial \theta_i}\frac{(A_0-n\mu(\boldsymbol{\theta}))^2}{nh(0;\boldsymbol{\theta})^2}\right]
\end{aligned} \tag{5.28}$$

This expression is useful as it allows to easily identify which components to use in case the zero frequency is included. For the remainder of this chapter we will assume without loss of generality that the zero frequency is included, for the situations where it is to be excluded similar derivations can be done.

For reasons that will become clear later on we will perform some algebraic manipulation of the spectral score. The spectral score can be written as a function of the periodogram for the centred process

$$Y_t^* = Y_t - \mu_0 \ , \tag{5.29}$$

where $\mu_0 = \mu(\boldsymbol{\theta}_0)$ (here, $\boldsymbol{\theta}_0$ denotes the true value of $\boldsymbol{\theta}$). The periodogram for the centered process can be written in terms of the variables in (5.28),

$$\begin{aligned} A_0^* &= \sum Y_t^* = A_0 - n\mu_0 \\ I^*(0) &= \frac{A_0^{*^2}}{n} = I(0) - 2A_0\mu_0 + n\mu_0^2 \\ I^*(w_p) &= I(w_p) \quad , p \neq 0 \ . \end{aligned}$$

Equivalently,

$$\begin{aligned} A_0 &= A_0^* + n\mu_0 \\ I(0) &= I^*(0) + 2A_0^*\mu_0 - n\mu_0^2 \end{aligned}$$

Note that only the Fourier coefficient at zero frequency is affected by the centring. Plugging these in (5.28) we obtain,

$$\begin{aligned} &-\sum_{p\neq 0}\left[1-\frac{1}{2}\delta_{p(n/2)}\right]\left[a_i(w_p;\boldsymbol{\theta})\left(h(w_p;\boldsymbol{\theta})-I^*(w_p)\right)\right] \\ &-\frac{1}{2}\left[a_i(0;\boldsymbol{\theta})h(0;\boldsymbol{\theta}) - \frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_i}\frac{2(A_0^*+n\mu_0-n\mu(\boldsymbol{\theta}))}{h(0;\boldsymbol{\theta})}\right. \\ &\left.-a_i(0;\boldsymbol{\theta})\frac{(A_0^*+n\mu_0-n\mu(\boldsymbol{\theta}))^2}{n}\right] \end{aligned} \tag{5.30}$$

where

$$a_i(w;\boldsymbol{\theta}) = \frac{\partial h(w;\boldsymbol{\theta})}{\partial\theta_i}\frac{1}{h(w;\boldsymbol{\theta})^2}$$

is introduced to simplify the notation. We can now use the fact that all the components in the sum are even functions of w and write (5.30) as,

$$
\begin{aligned}
& -\frac{1}{2}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})\left(h(w_p;\boldsymbol{\theta})-I^*(w_p)\right)+\frac{1}{2}a_i(0;\boldsymbol{\theta})\left(h(0;\boldsymbol{\theta})-I^*(0)\right) \\
& -\frac{1}{2}\left[a_i(0;\boldsymbol{\theta})h(0;\boldsymbol{\theta})-\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_i}\frac{2(A_0^*+n\mu_0-n\mu(\boldsymbol{\theta}))}{h(0;\boldsymbol{\theta})}-a_i(0;\boldsymbol{\theta})\frac{(A_0^*+n\mu_0-n\mu(\boldsymbol{\theta}))^2}{n}\right] \\
= & -\frac{1}{2}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})\left(h(w_p;\boldsymbol{\theta})-I^*(w_p)\right)+\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_i}\frac{2(A_0^*+n\mu_0-n\mu(\boldsymbol{\theta}))}{h(0;\boldsymbol{\theta})} \\
& +\frac{1}{2}a_i(0;\boldsymbol{\theta})\left[\frac{A_0^{*^2}+(n\mu_0-n\mu(\boldsymbol{\theta}))^2+2A_0^*(n\mu_0-n\mu(\boldsymbol{\theta}))}{n}-\frac{A_0^{*^2}}{n}\right] \\
= & -\frac{1}{2}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})\left(h(w_p;\boldsymbol{\theta})-I^*(w_p)\right)+\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_i}\frac{2(A_0^*+n(\mu_0-\mu(\boldsymbol{\theta})))}{h(0;\boldsymbol{\theta})} \\
& +\frac{a_i(0;\boldsymbol{\theta})}{2}\left[n(\mu_0-\mu(\boldsymbol{\theta}))^2+2A_0^*(\mu_0-\mu(\boldsymbol{\theta}))\right]
\end{aligned}
$$

This expression is much more tractable than (5.28), as it allows a straightforward application of the results from the previous section via the properties of the periodogram of a zero mean process. We now want to check if this estimating function, the spectral score, satisfies the conditions set in Chapter 2 in order to establish the consistency and asymptotic distribution of the estimator. The verification of these conditions on the asymptotic behaviour of the estimating functions is in this case implied by the asymptotic behaviour of the periodogram and properties of the spectral density. The asymptotic behaviour of the periodogram can be analyzed for the general class of stationary processes with no long range dependence for which we derived the spectral likelihood, in particular using the convergence results stated earlier. Regarding the properties of the spectral density that are sufficient for the result (2.22) to hold it is not possible to claim they are verified for every stationary processes with no long range dependence, therefore we reduce the class of processes for which this theory applies and assume that

$$
\begin{aligned}
0<h(w;\boldsymbol{\theta}) &< \infty \\
\frac{\partial^k h(w;\boldsymbol{\theta})}{\partial\theta^k} &< \infty \text{ and continuous, } \quad \mathrm{k}=1,2 \\
\frac{\partial^k h(w;\boldsymbol{\theta})}{\partial w^k} &< \infty\ , \quad k=1,2\ .
\end{aligned}
\tag{5.31}
$$

Note that these regularity conditions still allow for a wide class of models, and they partially correspond to the regularity conditions assumed in Chapter 2,

The only extra requirement for consistency is that the estimating function $g_i(\boldsymbol{\theta};\mathbf{y})$,under suitable normalization, converges to a deterministic function that has similar properties to a regular estimating function. By choosing the normalizing matrix $\boldsymbol{\eta}_n = n^{-1}\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, we can analyze each component of the estimating function vector. By applying results (5.23) and (5.25) we obtain,

$$\begin{aligned}\lim_{n\to\infty}\left[\boldsymbol{\eta}_n\mathbf{g}(\boldsymbol{\theta};\mathbf{y})\right]_i &= -\frac{1}{2}\int_{-\pi}^{\pi} a_i(w;\boldsymbol{\theta})\left(h(w;\boldsymbol{\theta})-h(w;\boldsymbol{\theta}_0)\right)dw \\ &\quad +\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_i}\frac{(\mu_0-\mu(\boldsymbol{\theta}))}{h(0;\boldsymbol{\theta})}+\frac{a_i(0;\boldsymbol{\theta})}{2}(\mu_0-\mu(\boldsymbol{\theta}))^2 \quad .\end{aligned}$$

This limiting function clearly attains the value zero for $\boldsymbol{\theta}=\boldsymbol{\theta}_0$.

In order to obtain a limiting covariance matrix for the estimator we need further conditions to be satisfied. We need that the covariance matrix of the estimating functions, when suitably normalized, converges to a finite positive definite matrix (2.16). We now show that the choice $\boldsymbol{\gamma}_n = n^{-1/2}\mathbf{I}$ does the trick by calculating the (i,j) element of the matrix $\boldsymbol{\Sigma}$. In the notation of Chapter 2

$$\begin{aligned}&\mathrm{Cov}(\boldsymbol{\gamma}_n g_i(\boldsymbol{\theta};\mathbf{y}),\boldsymbol{\gamma}_n g_j(\boldsymbol{\theta};\mathbf{y})) = \\ &\frac{1}{n}\mathrm{Cov}\left(\frac{1}{2}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})I^*(w_p)+\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_i}\frac{2A_0^*}{h(0;\boldsymbol{\theta})}+a_i(0;\boldsymbol{\theta})A_0^*(\mu_0-\mu(\boldsymbol{\theta})),\right. \\ &\left.\frac{1}{2}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_j(w_p;\boldsymbol{\theta})I^*(w_p)+\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_i}\frac{2A_0^*}{h(0;\boldsymbol{\theta})}+a_j(0;\boldsymbol{\theta})A_0^*(\mu_0-\mu(\boldsymbol{\theta}))\right) \qquad (5.32)\end{aligned}$$

and a simple manipulation gives this r.h.s. to be

$$\frac{1}{4n}\mathrm{Cov}\left(\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})I^*(w_p), \sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_j(w_p;\boldsymbol{\theta})I^*(w_p)\right) \tag{5.33}$$

$$+\frac{b_{ij}(\boldsymbol{\theta})}{n}\mathrm{Var}\,(A_0^*) \tag{5.34}$$

$$+\frac{d_{ij}(\boldsymbol{\theta})}{n}\mathrm{Cov}\left(\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})I^*(w_p), A_0^*\right) \tag{5.35}$$

where $b_{ij}(\boldsymbol{\theta})$ and $d_{ij}(\boldsymbol{\theta})$ are quantities that do not depend on $I^*(w_p)$ or A_0^*, and can easily be derived from (5.32). We will deal with each of the components (5.33),(5.34 and (5.35) separately. First we use (5.22) to write (5.33) in the integral form, for large n

$$\begin{aligned}
&\frac{1}{4n}\mathrm{Cov}\left(\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})I^*(w_p), \sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_j(w_p;\boldsymbol{\theta})I^*(w_p)\right)\\
\approx\ &\frac{1}{4n}\mathrm{Cov}\left(n\int_{-\pi}^{\pi} a_i(w;\boldsymbol{\theta})I^*(w), n\int_{-\pi}^{\pi} a_j(w;\boldsymbol{\theta})I^*(w)\right)
\end{aligned}$$

This allows us to apply (5.27) and obtain

$$\begin{aligned}
\lim_{n\to\infty}\quad &\frac{1}{4n}n^2\mathrm{Cov}\left(\int_{-\pi}^{\pi} a_i(w;\boldsymbol{\theta})I^*(w), \int_{-\pi}^{\pi} a_j(w;\boldsymbol{\theta})I^*(w)\right)\\
=\ &\frac{e^{(4)}}{4}\int_{-\pi}^{\pi} a_i(w,\boldsymbol{\theta}_0)h(w;\boldsymbol{\theta}_0)dw\int_{-\pi}^{\pi} a_j(w,\boldsymbol{\theta}_0)h(w;\boldsymbol{\theta}_0)dw\\
&+\pi\int_{-\pi}^{\pi} a_i(w,\boldsymbol{\theta}_0)a_j(w,\boldsymbol{\theta}_0)h^2(w;\boldsymbol{\theta}_0)dw
\end{aligned} \tag{5.36}$$

The evaluation of (5.34) can be done using (5.19)

$$\lim_{n\to\infty}\quad \frac{1}{n}b_{ij}(\boldsymbol{\theta})\mathrm{Var}\,(A_0^*) = b_{ij}(\boldsymbol{\theta})h(0;\boldsymbol{\theta}) \tag{5.37}$$

The treatment of (5.35) is not so straightforward. We start by writing it in the form of an expectation

$$
\begin{aligned}
& \frac{d_{ij}}{n}\mathrm{Cov}\left[A_0^*, \sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})I^*(w_p)\right] \\
= & \frac{d_{ij}}{n}E\left[A_0^*\frac{1}{n}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})(A_p^{*2}+B_p^{*2})\right] \\
= & \frac{d_{ij}}{n^2}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})E\left[A_0^*(A_p^{*2}+B_p^{*2})\right] \\
= & \frac{d_{ij}}{n^2}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})E\left[G_0^*G_p^*G_{-p}^*\right] \\
= & \frac{d_{ij}}{n^2}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})c_3\left(G_0^*, G_p^*, G_{-p}^*\right)
\end{aligned}
$$

since $E[G_p^*] = 0$ from (5.8) and (5.9) together with (5.29)

Now, by applying (5.26), we obtain

$$
\frac{d_{ij}}{n^2}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor} a_i(w_p;\boldsymbol{\theta})\left[(2\pi)^2 nh^{(3)}(0,w_p,w_{-p}) + O(1)\right] \quad .
$$

Taking limits and assuming $\partial h^{(3)}(0,\mathbf{w},-\mathbf{w})/\partial w$ is bounded the application of (5.24) yields, as $n \to \infty$ this converges to

$$
d_{ij}\,(2\pi)^2 \int_{-\pi}^{\pi} a_i(w;\boldsymbol{\theta})h^{(3)}(0,w,-w)dw \tag{5.38}
$$

Finally, by putting (5.36), (5.37) and (5.38) together we obtain the desired result

$$\begin{aligned}\lim_{n\to\infty} \quad & \mathrm{Cov}(\boldsymbol{\gamma}_n g_i(\boldsymbol{\theta};\mathbf{y}), \boldsymbol{\gamma}_n g_j(\boldsymbol{\theta};\mathbf{y})) = \\ & \pi \int_{-\pi}^{\pi} a_i(w,\boldsymbol{\theta}) a_j(w,\boldsymbol{\theta}) h^2(w;\boldsymbol{\theta}) dw \\ + \quad & b_{ij}(\boldsymbol{\theta}) h(0;\boldsymbol{\theta}) \\ + \quad & d_{ij} (2\pi)^2 \int_{-\pi}^{\pi} a_i(w;\boldsymbol{\theta}) h^{(3)}(0,w,-w) dw \\ + \quad & d_{ij} (2\pi)^2 \int_{-\pi}^{\pi} a_j(w;\boldsymbol{\theta}) h^{(3)}(0,w,-w) dw \end{aligned} \tag{5.39}$$

Equation (5.39) shows that normalizing the covariance matrix of the estimating function by $\boldsymbol{\gamma}_n = n^{-1/2}\mathbf{I}$ leads to convergence to a finite matrix, however the positive (semi) definite requirement needs to be verified for the particular processes considered in applications. Note it will be unusual for a limit of a sequence of positive (semi) definite matrices not itself to be positive (semi) definite.

Another necessary condition on $\tilde{\mathbf{g}}(\boldsymbol{\theta};\mathbf{y}) = \boldsymbol{\gamma}_n \mathbf{g}(\boldsymbol{\theta};\mathbf{y})$ is that after suitable normalization its partial derivatives with respect to $\boldsymbol{\theta}$ converge in probability to a deterministic matrix which may depend on $\boldsymbol{\theta}$, $\mathbf{M}(\boldsymbol{\theta})$. This is formalized as follow

$$\left[\frac{\partial \tilde{\mathbf{g}}(\boldsymbol{\theta};\mathbf{y})}{\partial \boldsymbol{\theta}}\right] \boldsymbol{\delta}_n \xrightarrow{p} \mathbf{M}(\boldsymbol{\theta}).$$

Using some of the convergence results above we will now show that the choice of $\boldsymbol{\delta}_n = n^{-1/2}\mathbf{I}$ satisfies this condition. Substituting for $\boldsymbol{\gamma}_n = n^{-1/2}\mathbf{I}$ and $\boldsymbol{\delta}_n = n^{-1/2}\mathbf{I}$ we can then write the $(i,j)^{th}$ element of the normalized Jacobian, $\partial \boldsymbol{\delta}_n \tilde{\mathbf{g}}(\boldsymbol{\theta};\mathbf{y})/\partial \boldsymbol{\theta}$ as

$$
\begin{aligned}
& \frac{1}{n}\left[-\frac{1}{2}\sum_{p=-\lfloor\frac{n-1}{2}\rfloor}^{\lfloor\frac{n}{2}\rfloor}\frac{\partial a_i(w_p;\boldsymbol{\theta})}{\partial\theta_j}\left(h(w_p;\boldsymbol{\theta})-I^*(w_p)\right)+a_i(w_p;\boldsymbol{\theta})\frac{\partial h(w_p;\boldsymbol{\theta})}{\partial\theta_j}\right. \\
+ \quad & 2\frac{\partial^2\mu(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\frac{[A_0^*+n(\mu_0-\mu(\boldsymbol{\theta}))]}{h(0;\boldsymbol{\theta})} \\
+ \quad & 2\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_i}\frac{n\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_j}h(0;\boldsymbol{\theta})-\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_j}[A_0^*+n(\mu_0-\mu(\boldsymbol{\theta}))]}{h^2(0;\boldsymbol{\theta})} \\
+ \quad & \frac{1}{2}\frac{\partial a_i(0;\boldsymbol{\theta})}{\partial\theta_j}\left[n(\mu_0-\mu(\boldsymbol{\theta}))^2+2A_0^*(\mu_0-\mu(\boldsymbol{\theta}))\right] \\
+ \quad & \left.a_i(0;\boldsymbol{\theta})\left[n(\mu_0-\mu(\boldsymbol{\theta}))\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_j}-A_0^*\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_j}\right]\right] \quad .
\end{aligned}
$$

By taking limits and applying (5.23) and (5.25) we obtain

$$
\begin{aligned}
- \quad & \frac{1}{2}\int_{-\pi}^{\pi}a_i(w;\boldsymbol{\theta})\frac{\partial h(w;\boldsymbol{\theta})}{\partial\theta_j}dw+2\frac{\partial^2\mu(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\frac{n(\mu_0-\mu(\boldsymbol{\theta}))}{h(0;\boldsymbol{\theta})} \\
+ \quad & \frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_i}\left[\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_j}\frac{2}{h(0;\boldsymbol{\theta})}-\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_j}\frac{(\mu_0-\mu(\boldsymbol{\theta}))}{h^2(0;\boldsymbol{\theta})}\right] \\
+ \quad & \frac{\partial a_i(0;\boldsymbol{\theta})}{\partial\theta_j}\frac{(\mu_0-\mu(\boldsymbol{\theta}))^2}{2}+a_i(0;\boldsymbol{\theta})\frac{\partial\mu(\boldsymbol{\theta})}{\partial\theta_j}(\mu_0-\mu(\boldsymbol{\theta}))
\end{aligned}
$$

The last requirement is the continuity with respect to $\boldsymbol{\theta}$ of the normalized Jacobian above, which is guaranteed by the continuity assumption stated in (5.32).

Combining all of these results, we have shown that the Whittle likelihood estimator has all the properties of a regular estimating functions estimator. The theory of Chapter 2 can now be used directly to give a limiting distribution:

$$
n^{1/2}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\xrightarrow{d}MVN(\mathbf{0},\mathbf{M}_0\boldsymbol{\Sigma}\mathbf{M}_0^T)
$$

where

$$\begin{aligned} \mathbf{M}_0 &= \left[\frac{\partial \tilde{\mathbf{g}}(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}}\right] \boldsymbol{\delta}_n \Big|_{\boldsymbol{\theta}_0} \\ \boldsymbol{\Sigma} &= \mathrm{Var}(\mathbf{g}(\boldsymbol{\theta}_0; \mathbf{y})) \end{aligned}$$

This result can be used to build approximate confidence intervals and regions for the parameter vector $\boldsymbol{\theta}$. In practice the matrices $\mathbf{M}_0$ are calculated using numerical differentiation, and $\boldsymbol{\Sigma}$ is estimated from the data using an approach of the type defined in (4.9). The application of the general framework of estimating functions to the Whittle estimator allows us to obtain approximate confidence regions without having to calculate spectral densities of higher order than two. This result is therefore very useful for the scientific areas in which estimation is mainly done in the frequency domain.

Chapter 6

Simulation Study - Application of spectral likelihood to rainfall models

We now illustrate the theory from the previous chapter, using a simulation study similar to that from Chapter 4. The models considered here are the same that were included in Chapter 4, they are the Poisson rectangular pulses model and the Neyman-Scott rectangular pulses model. We start by deriving the spectral likelihood for the point-processes of interest, this essentially involves constructing the relevant spectral densities. We then investigate if the spectral densities verify the conditions that allow the application of the estimating functions as set out in the previous chapter. Sections 6.2 and 6.3 present the results of the study for the Poisson and Neyman-Scott models respectively, and section 6.4 provides a short summary of this chapter.

As before this study will involve both the Poisson and the Neyman-Scott rectangular pulses models, and each simulation represents 20 years worth of data for a single calendar month. To be precise, 20 independent sets of 30 days worth of hourly values are generated for each simulation, and 1000 simulations were generated in total. As in the GMM study of Chapter 4 we will use the concept of empirical and theoretical covariance matrices to evaluate the performance of the estimator itself and of the measures of estimator uncertainty. Using bias and the empirical covariance matrix, from a sample of 1000 simulations, we can comment on the performance of the estimator; and by comparing this matrix with the theoretical covariance matrix we can evaluate the finite sample performance of the asymptotic theory from the previous chapter, namely the

assessment of estimator uncertainty. In particular, assessment of estimator uncertainty will be analysed by comparing empirical and theoretical matrices and also by looking at the coverage rates of confidence intervals and regions built using the estimating functions framework. In order to apply the theory of estimating functions regarding the asymptotic distribution of the estimator we used the fact that for each simulation data were generated in 20 independent sets. The procedure here is equivalent to (4.7), where the estimating function becomes the average spectral score. This way we can estimate the covariance matrix of the estimating functions.

6.1 Derivation of spectral likelihood

One of the motivations for the work on spectral likelihood is the fact that for the particular class of point process rainfall models it is relatively easy to write down their spectral density, so we start by formulating the analytical expression for the spectral densities of the processes included in this study. These are as before the Poisson rectangular pulses model and the Neyman-Scott rectangular pulses model, and it was mentioned previously that the only formal difference between the two is the cell arrival process, therefore it is not surprising that the derivation of the spectral densities for either of these processes have many similarities. From the definition of spectral density and the fact that we observe accumulated totals over h-hourly periods we obtain (Chandler, 1997)

$$h(w) = \sum_{k=-\infty}^{\infty} s\left(w + \frac{2k\pi}{h}\right)\left(\frac{\sin([wh + 2k\pi]/2)}{[wh + 2k\pi]/2}\right)^2 \quad , \quad |w| \leq \pi/h \tag{6.1}$$

where $s(w)$ is the spectral density of the underlying continuous time process that gives rise to Y_t. For the general class of point-process rainfall models considered here, the functional form of $s(w)$ can be derived as a function of the parameters defined in Section 4.1 and the incomplete spectral density of the cell arrival process denoted $s^*_{\cdot}(w)$ (Chandler, 1997). This is the incomplete spectral density for a point process, which is derived from the conditional intensity, for more details see Cox and Isham (1980).

$$s(w) = \frac{1}{2\pi w^2}\left[2\pi s^*_{\cdot}(w)\mu_X^2|1 - \phi_L(w)|^2 + 2\rho\left(\mu_X^2 + \sigma_X^2\right)\left(1 - \Re\left(\phi_L(w)\right)\right)\right] \tag{6.2}$$

where μ_X and σ_X are the mean and standard deviation of the distribution of cell intensity, $\phi_L(w)$ is the characteristic function of the cell duration distribution and ρ is the cell arrival rate. In section 4.1 we defined the cells as having exponentially distributed durations. Noting that the characteristic function for the exponential distribution with parameter α is,

$$\phi(w) = \frac{i\alpha}{w + i\alpha}$$

it follows that,

$$\begin{aligned} \Re\left(\phi(w)\right) &= \frac{\alpha^2}{\alpha^2 + w^2} \\ |\phi(w)|^2 &= \frac{\alpha^2}{\alpha^2 + w^2} \\ |1 - \phi(w)|^2 &= \frac{w^2}{\alpha^2 + w^2} \quad . \end{aligned} \tag{6.3}$$

We next look into the spectral densities of the arrival processes.

For the Poisson process of arrivals the incomplete spectral density is identically equal to zero (Chandler, 1997), i.e.

$$s^*_{Po}(w) = 0 \tag{6.4}$$

and the cell arrival rate ρ is simply equal to λ, so that by plugging (6.3) and (6.4) in (6.2) we obtain

$$s_{Po}(w) = \frac{\lambda(\mu_X^2 + \sigma_X^2)}{\pi(\eta^2 + w^2)}$$

where η is the parameter of the exponentially distributed cell duration. For the Neyman-Scott point process of cell arrivals the incomplete spectral density is (Chandler, 1997),

$$\begin{aligned} s^*_{NS}(w) &= \frac{\lambda}{2\pi} E\left[C(C-1)\right] |\phi_D(w)|^2 \\ &= \frac{\lambda}{2\pi}(\mu_C^2 - 1)|\phi_D(w)|^2 \end{aligned} \tag{6.5}$$

where C is the number of cells per storm and $\phi_D(w)$ is the characteristic function of the cell displacement from the storm origin which is assumed to be exponential with parameter β. To evaluate $E\left[C(C-1)\right]$ we need to make some assumption about the distribution of C. The assumption here is that $C-1$ has a Poisson distribution, this is the same as in Cowpertwait (1991), and it guarantees the existence of at least one cell per storm. For this point-process the cell arrival rate is the product of the storm arrival rate and the expected number of cells per storm, $\rho = \lambda\mu_C$ so that when we plug both (6.5) and (6.3) into (6.2) we obtain

$$s_{NS}(w) = \frac{\lambda}{2\pi(\eta^2 + w^2)} \left(\frac{\mu_X^2 \beta(\mu_C^2 - 1)}{\beta^2 + w^2} + 2\mu_C(\mu_X^2 + \sigma_X^2) \right)$$

where η is the parameter of the exponentially distributed cell duration.

We now look at the application of the theory in the previous section to these particular processes. First we need to check if conditions (2.3) - (2.5) are satisfied by the spectral densities defined above. Given the simple functional form of the spectral density it is trivial to check that these conditions are verified.

Although these regularity conditions are verified, there is another condition that needs to be met so that the results in the previous section can be applied in their entirety. The condition

$$\left[\frac{\partial \tilde{\mathbf{g}}(\boldsymbol{\theta}; \mathbf{w})}{\partial \boldsymbol{\theta}} \right] \boldsymbol{\delta}_n \xrightarrow{p} \mathbf{M}(\boldsymbol{\theta})$$

where $\mathbf{M}(\boldsymbol{\theta})$ needs to be invertible, actually fails for the whole set of parameters for the Poisson model. It is straightforward to see why $\mathbf{M}(\boldsymbol{\theta})$ is singular if we notice that both μ_X and σ_X contribute to the spectral density in the same way. $\mathbf{M}(\boldsymbol{\theta})$ is effectively the information matrix for the log-likelihood defined in (5.21) and since its rank is less than the number of parameters to estimate we can conclude that this model is not identified.

A possible way around which implies simplifying our model is to assume an exponential distribution for the cell intensity, this assumption is often considered in this class of models (Rodriguez-Iturbe et al., 1987; Cowpertwait, 1991). In our implementation this implies setting $\sigma_X/\mu_X = 1$. For the Neyman-Scott model the matrix $\mathbf{M}$ is not exactly singular but it is certainly ill-conditioned. Therefore we apply the exponential distributed cell intensity assumption to both models in our study.

6.2 Results for the Poisson Rectangular Pulses model

The setting for this study is very similar to what was done in Chapter 4, data was simulated in the same way: 1000 simulations, each consisting of 20 sets of 30 days worth of hourly rainfall. The parameter values used in the simulations were the same as in the GMM study, $\boldsymbol{\theta} = (\log(\lambda), \log(\mu_X), \log(\mu_L)) = (-3.5, 0, 1.1)$. In the implementation of (6.1) the infinite sum needed to be approximated by a finite truncation. For the processes we studied we found that $K = 10$ was sufficient. Also similarly to what was done in the GMM study, we will use the concepts of empirical and theoretical covariance matrix to compare the performance of the estimation uncertainty assessment.

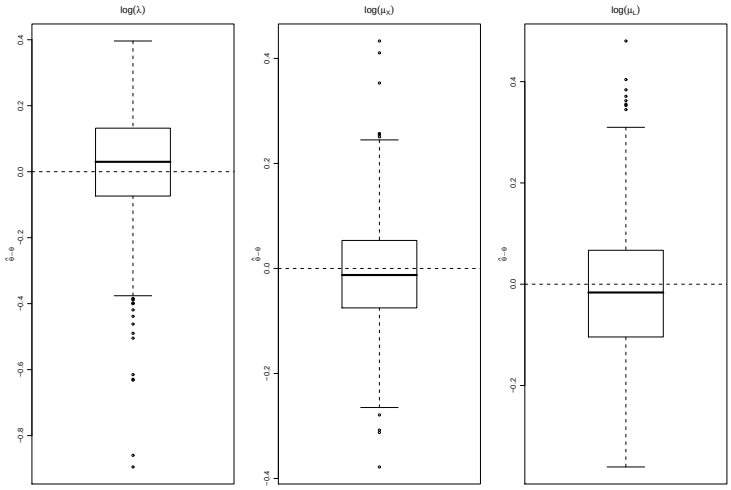

FIGURE 6.1: Distribution of estimation errors for each PRPM parameter. All distributions are obtained from 1000 simulated data sets, each containing 20 independent 30-day sequences and generated using parameter values $\log(\lambda) = -3.5, \log(\mu_X) = 0, \log(\mu_L) = 1.1$

Figure 6.1 shows the distribution of the 1000 estimates. From it we can see that there is no significant bias, or alternatively that the variance of the estimator is high relative to any existing bias. For λ and μ_L the resulting distribution has a negative and positive skewness respectively, this is not surprising as the parameters appear in the spectral density and consequently in the estimating function contributing in a similar way making

it difficult for the model/algorithm to distinguish between them. Furthermore if we calculate the correlation between the estimates for λ and μ_L we obtain a value of -92% for this set of simulations. Table 6.1 shows the bias for each parameter together with their standard errors. Although there is some evidence of bias, if we compare it with the actual estimator variability and from a mean squared error perspective, the squared bias is indeed smaller than the variance of the estimator, making it negligible for the purposes of parameter estimation for this particular model.

$\log(\lambda)$	$\log(\mu_X)$	$\log(\mu_L)$
0.016(0.005)	-0.009(0.003)	-0.014(0.004)

TABLE 6.1: Estimated bias for each parameter together with their standard errors.

We now proceed to evaluate the finite sample performance of the theory with respect to the assessment of estimator uncertainty. We first use the standard errors obtained from the empirical and the theoretical covariance matrix; from Table 6.2 we can see that the mean theoretical standard errors tend to overestimate the variability of the estimator, however if we consider the theoretical standard errors based on the median of the variances calculated for each simulation we obtain a better assessment of estimator variability. This means that for some particular simulations the theoretical standard errors obtained are much higher than the empirical standard error. This asymmetry can be confirmed in Figure 6.2 that shows estimates for the density of the theoretical standard errors. Table 6.3 shows the determinants of the empirical and theoretical covariance matrices, both based on the median and the mean of the theoretical covariance matrices, we conclude that confidence regions based on the theoretical covariance matrix will generally be larger compared to confidence regions based on the empirical covariance matrix. As in the GMM case we are calculating the covariance matrix of the estimating function using a relatively small amount of data, in this particular case we are estimating a 5×5 symmetrical matrix from only 20 observation, it is natural that this problem has repercussions in the subsequent estimation of a covariance matrix for the estimator.

	$\log(\lambda)$	$\log(\mu_X)$	$\log(\mu_L)$
Empirical	0.159	0.097	0.129
Median Theoretical	0.145	0.084	0.119
Mean Theoretical	0.253	0.118	0.207

TABLE 6.2: Empirical standard errors together with the standard errors obtained from the median and mean of the theoretical covariance matrices for each of the parameters.

Table 6.4 shows coverage rates for confidence intervals built using theoretical standard errors and asymptotic normality approximations, and also the coverage rates for confidence regions built based on the spectral likelihood itself, using (2.24). We can see that

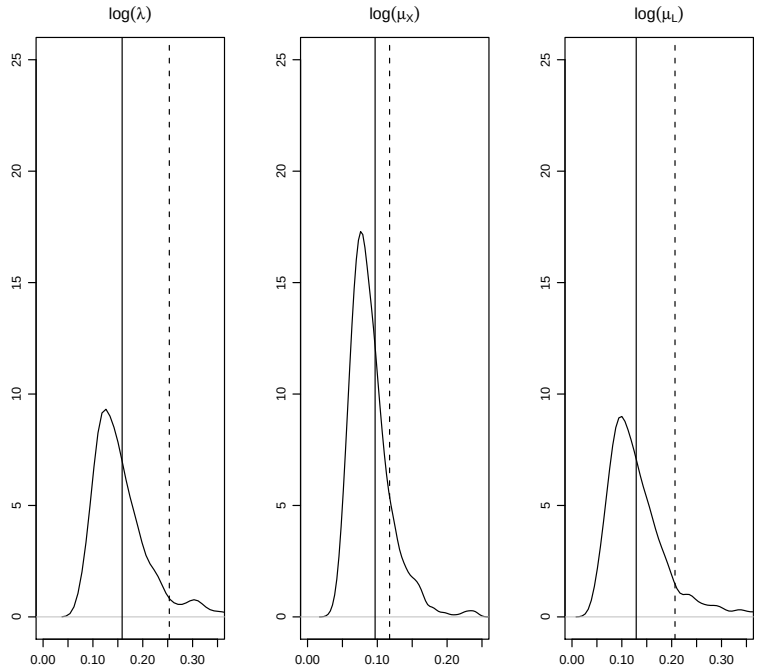

FIGURE 6.2: Estimated densities of theoretical standard errors from 1000 simulations together with the empirical standard errors (vertical lines) and the average theoretical standard errors (vertical dotted lines), for each parameter.

Empirical	0.171
Median Theoretical	0.163
Mean Theoretical	0.190

TABLE 6.3: Determinants of the empirical, median and mean theoretical covariance matrices.

despite the theoretical covariance matrix overestimating true estimator variability on average the coverage for the individual parameters is reasonable. The lower coverage of the confidence intervals for λ and μ_L can be explained, at least partially by the skewness seen in Figure 6.1.

	$\log(\lambda)$	$\log(\mu_X)$	$\log(\mu_L)$	$\boldsymbol{\theta}$
95%	0.93	0.94	0.90	0.72
99%	0.97	0.98	0.95	0.93

TABLE 6.4: Coverage rates for confidence intervals based on normality assumption for the individual parameters, and for the confidence region based on an objective function threshold.

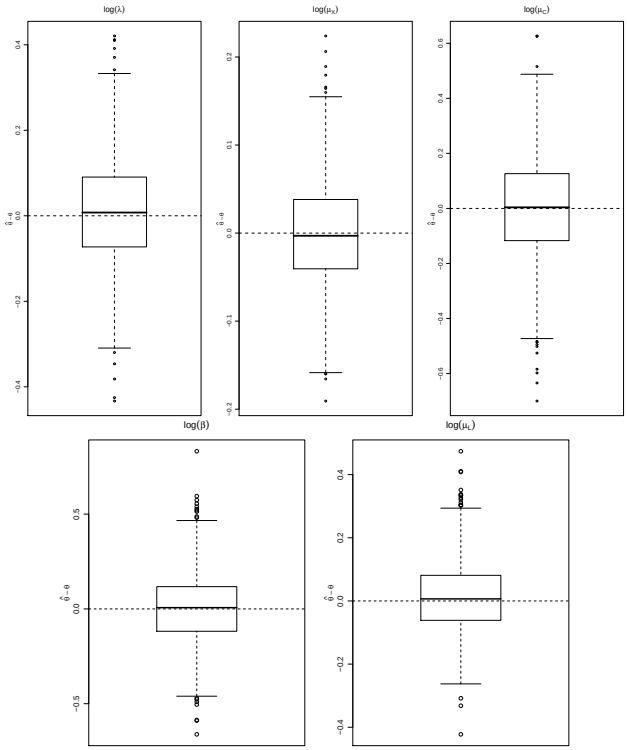

FIGURE 6.3: Distribution of estimation errors for each NS parameter, obtained using the spectral likelihood estimator. All distributions are obtained from 1000 simulated data sets, each containing 20 independent 30-day sequences and generated using parameter values $\log(\lambda) = -4, \log(\mu_X) = -0.44, \log(\mu_C) = 2.46, \log(\beta) = 1.8, \log(\mu_L) = -0.37$

6.3 Results for the Neyman-Scott Rectangular Pulses model

For the finite performance study with the Neyman-Scott model data was generated in a similar way to what was done in the Poisson case. As mentioned in the end of section 6.1 the parameters μ_X and σ_X contribute in different ways to the spectral density, and can in theory be identified using the Whittle likelihood approach. However, given the modest results for the Poisson we keep the assumption of exponentially distributed cell intensity for the Neyman-Scott simulations. The parameter values used here were the same that were used for the GMM study, $\boldsymbol{\theta} = (\log(\lambda), \log(\mu_X), \log(\mu_C), \log(\beta), \log(\mu_L)) = (-4, -0.44, 2.46, 1.8, -0.37)$.

Figure 6.3 shows the distribution of the 1000 estimates for each parameter, and from the analysis of the boxplots we cannot detect any signs of bias or skewness. In fact Table 6.5 suggests bias is absent for all the parameters apart from $\log(\mu_L)$, and even in this case bias is relatively smaller when compared to sample variability.

$\log(\lambda)$	$\log(\mu_X)$	$\log(\mu_C)$	$\log(\beta)$	$\log(\mu_L)$
0.008(0.004)	-0.003(0.002)	0.003(0.006)	-0.006(0.006)	0.014(0.004)

TABLE 6.5: Estimated bias for each parameter together with their standard errors.

A comparison of the empirical and mean theoretical standard errors present in Table 6.6 indicates that the estimating functions framework when applied to the spectral likelihood and for this particular models tends to overestimate the variance of the estimator quite significantly on an average basis. The lower value for median theoretical standard errors show that the high mean theoretical standard errors is due to a few high values for the theoretical standard errors of particular simulations, this suggests the method can be significantly unstable. This can be an effect from the small sample that is used to calculate the covariance matrix of the estimating functions as discussed in the results of the Poisson model.

	$\log(\lambda)$	$\log(\mu_X)$	$\log(\mu_C)$	$\log(\beta)$	$\log(\mu_L)$
Empirical	0.124	0.059	0.188	0.186	0.115
Median Theoretical	0.143	0.064	0.215	0.216	0.134
Mean Theoretical	0.516	0.302	1.301	1.407	0.840

TABLE 6.6: Empirical and theoretical standard errors for each of the parameters.

Empirical	0.049
Median Theoretical	0.064
Mean Theoretical	0.200

TABLE 6.7: Determinants of the Empirical and theoretical matrices.

The performance in terms of confidence intervals for the individual parameters based on normality approximation on its own is quite satisfactory as can be seen in Table 6.8. However given that there is an overestimation of estimator variability one could expect the coverages to be higher than the target value, which has not happened. For the coverage of the confidence regions based on objective function thresholds the results are not so good, in particular the 95% confidence region in which for only 77% of the simulations the true parameter value is inside the region.

	$\log(\lambda)$	$\log(\mu_X)$	$\log(\mu_C)$	$\log(\beta)$	$\log(\mu_L)$	$\boldsymbol{\theta}$
95%	0.96	0.96	0.92	0.93	0.92	0.77
99%	0.98	0.99	0.96	0.96	0.95	0.96

TABLE 6.8: Coverage rates for confidence interval based on normality assumption.

6.4 Summary

In this chapter we studied the finite sample performance of the spectral likelihood estimator, when applied to the Poisson and Neyman-Scott rectangular pulses model. We have seen that for these models it is relatively easy to write down a spectral likelihood, however that does not meant that the application of estimating functions results is straightforward. Some care is needed when verifying some of the conditions required by the theory, namely the conditions involving matrix classification. We concluded that for both models any potential bias was not significant when compared with estimator variability. The spectral likelihood estimator variability was found to be higher than the GMM estimators with data dependent weights from chapter 4, this can be seen by comparing the empirical standard errors of the estimators. For the Poisson rectangular pulses model when using the two step GMM the empirical standard error were between 0.06 and 0.071, when using a spectral likelihood approach the empirical standard errors are in the range 0.097 to 0.159. For the Neyman-Scott rectangular pulses model the difference is not so clear with standard errors in the interval 0.064 to 0.211 for the GMM estimation and between 0.059 and 0.188 for the spectral likelihood method, however when using the latter in were effectively estimating one less parameter. The finite sample performance of the estimating functions theory for the estimation of the variance of the spectral likelihood estimator seem quite poor, as the mean theoretical standard errors are significantly higher than the empirical standard errors. However, the reasonable level of coverage in the confidence intervals built based on normality assumptions suggest that when in the presence of a poor estimate the standard errors reflect that.

Chapter 7

Conclusion

In this thesis we presented work regarding inferential procedures that can be used in the absence of a likelihood function, in particular the framework of estimating functions. The theory of estimating functions permits to find a consistent estimator for θ as well as characterizing its uncertainty. We described the theory of estimating functions with increased focus on their asymptotic properties and the relationship between these and the properties of the estimator itself. We have given sufficient conditions for the consistency and asymptotic normality of the estimating functions estimator. These conditions were formulated in a way that allows for the investigator to check them for the particular application of interest, moreover it was shown in some detail the role of these conditions in obtaining the target asymptotic results. A consequence of this treatment of the estimating functions theory is that for specific applications such conditions can be relaxed or replaced by equivalent ones in the sense that they have the same impact on the derivation of the asymptotic results. The theory of estimating functions can accommodate some standard techniques such as maximum likelihood and least-squares estimation. Hence, the asymptotic results for the estimating functions are applicable to a wide class of estimation problems including the particular field of moment based inference, which we investigated in some detail in Chapter 3. We also presented an important result regarding the calculation of confidence regions for the class of problems where the estimating function can be seen as the gradient vector of some "objective function". This is clearly the case in the GMM setting.We were able to use the general results for estimating functions to particular techniques of building estimating functions, namely GMM and Whittle likelihood.

Application of the asymptotic results from the estimating functions theory to the GMM estimator allowed to derive a set of particular sufficient conditions for consistency of the

GMM estimator. These condition are set out in a different way to those usually included in the econometrics literature and are more intuitive for a statistical science focused audience, as well more straightforward to verify in practice. Also via the concept of optimal estimating function within a certain class we verify the existence of an asymptotically optimal weighting matrix for the case where the number of moment conditions is higher than the number of parameters to be estimated.

A significant amount of work was done in the analysis of finite sample performance of the asymptotic approximations derived in Chapter 3. The focus was on a class of models that provided some motivation for this thesis. These models can be specified in terms of a restricted set of constraints representing the relationship between the parameters and the data. The parameters are usually fitted by matching summary statistics (e.g. means, variances and autocorrelations) computed from the observations $\mathbf{y}$ with their expectations under the model, in a quadratic form. In the simulation based study of Chapter 4 we considered a relatively small sample and different ways of combining the properties to be matched, i.e. different weighting schemes. We included a scenario of current practice and compared it with 3 other scenarios. It was clear that the theoretical optimum provided an estimator with better properties. However the characterization of estimator uncertainty was not satisfactory for the case where optimal weights were used. The investigation of this mismatch between true estimator variance and the variance implied by the asymptotic theory led us to conclude that the problem was in the estimation of the covariance matrix of the estimating functions. This poor estimation of the covariance matrix of the estimating functions was due to the small sample size considered. A suggestion found in the econometrics literature is that if we have parametrization of this covariance matrix we can use an initial estimate of the parameters obtained from a consistent estimator to consistently estimate the covariance matrix. However in this case as in many others such explicit form for the covariance matrix is not available. We propose a generic solution for cases where simulation of the process being studied is possible; by simulating a big enough sample using an initial estimate, obtained from a consistent estimator, one can estimate the covariance matrix of the estimating functions accurately. This is in fact equivalent to the parametrization approach; However it adds an extra computational burden to the algorithm. From the study in Chapter 4 we concluded that using the asymptotic optimal weights based on the results from Chapter 3 is optimal in relatively small samples, this is an improvement over current practice; moreover the 2-step procedure suggested here is an important contribution for the area of moment based inference, as having explicit expressions for the covariance matrices of the estimating functions is only possible for relatively simple problems.

We also explored the possibility of transforming the data in order to obtain an approximate likelihood, more precisely we derived an approximate likelihood from the Fourier transform of the data. This method, known as the Whittle likelihood, has been widely used, however there is limited literature addressing the problem of characterizing estimator uncertainty. By considering the estimating function framework and the theory set out in Chapter 2 it was possible to show not only the sufficient conditions for the consistency of the spectral likelihood estimator, but more importantly to present an approximate distribution for this estimator. Hence, an important contribution of the present work to this field is to show that it is possible to derive a covariance matrix for the estimator from which does not depend on the knowledge of the expression for the 4^{th}order spectral density. It also provides generalization of the result in Heyde (1997) in the sense that under this setting it is not required that the process has zero mean. This is even more important for inference in models where the mean of the process depends on several parameters that also affect other properties of the data.

We also studied the finite sample properties of the spectral likelihood estimator using a similar setting to what was done for the GMM estimator in Chapter 4. Through simulations we studied the performance of the spectral likelihood estimator, namely bias and characterization of estimator uncertainty. In this case, although the results were satisfactory in terms of verifying that the theory holds even for modest size samples, we also found out that the performance was not as good as the GMM estimator for the two models considered. On the one hand we had to simplify our models by fixing one of the parameters, and on the other hand the variance of the spectral likelihood estimator turned out larger than for the GMM estimator with optimal weights. This provides evidence that for the class of models considered the use of moment based inference, in particular the 2-step GMM estimator, is more adequate than the data transformation approach using a Fourier transform.

The aim of this work was to explore different methods of performing inference in cases where a likelihood function is unavailable. The focus was on the estimating functions theory, and the first achievement was to combine the estimating functions framework which is significantly known in the field of statistics with the GMM methodology which is widely used in econometrics but not so much used in the area of statistics. An example of this lack of familiarity with the GMM methodology by part of the community of practitioners is the class of models that provided some motivation for this work, the class of point-process model for rainfall. Hence, another important outcome of this work was to show that by putting together the estimating functions and GMM theories one was able to suggest a straightforward improvement to the inference methods currently used

for estimating parameters of the Poisson and Neyman-Scott rectangular pulses model. One potential criticism to moment based inference is that the result of the inferential procedure may depend on the choice of properties that are used in the estimation, however as it was shown in this thesis this problem can be solved, if not entirely at least partially by having data dependent weights under a GMM setting. This problem of the estimator depending on the choice of moments was also one of the motivations to look into a data transformation approach. It was already mentioned that this problem can be solved within the framework of moment based inference. However it is also important to emphasize a last result included in this thesis, this is the ability to calculate confidence regions for the spectral likelihood estimator under mild conditions.

Some of the results in this thesis may raise some interesting research questions that can be the basis of further work, namely in the field of estimation using the spectral likelihood. One question regards the use of a 2-step approach in the estimation of the covariance matrix of the spectral scores, using simulations as suggested in the GMM optimal weights case. The other derives from the fact that to build our spectral scores we used the average of the periodogram, which is equivalent to the spectral estimate obtained when smoothing the raw periodogram using a Barlett's window (Priestley, 1981). This suggests that investigating the performance of other smoothed spectral estimators can be of interest, particularly in terms of finite sample performance.

The theory described here is general in nature, even if its development in this thesis has some focus in the particular class of point-processes for rainfall. However, for numerous models of complex systems it will often be optimal to develop specialised methods that are in general more difficult to implement, but by having specifically designed methods for the particular problem one can improve the inference.

Bibliography

Abramowitz, M. and I. A. Stegun, 1964: *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Washington, D.C. : U.S. Government Printing Office.

Anderson, T. G. and B. E. Sørensen, 1996: GMM estimation of a stochastic volatility model: a Monte Carlo study. *Journal Bus. Econ. Statist.*, **14**, 328–352.

Anh, V. V., N. N. Leonenko, and L. M. Sakhno, 2004: Quasi-likelihood-based higher-order spectral estimation of random fields with possible long-range dependence. *Journal of Applied Probability*, **41**, 35–53.

Anh, V. V., N. N. Leonenko, and L. M. Sakhno, 2007: Minimum contrast estimation of random processes based on information of second and third orders. *Journal of statistical planning and inference*, **137**, 1302–1331.

Bera, A. K. and Y. Bilias, 2002: The MM,ME,ML,EL,EF and GMM approaches to estimation: a synthesis. *J. Econometrics*, **107**, 51–86.

Brillinger, D. R., 1975: *Time Series - Data Analysis and Theory*. Holt, Rinehart and Winston, Inc.

Brillinger, D. R. and M. Rosenblatt, 1967: Asymptotic theory of estimates of k-th order spectra. *Proc Nat Acad Sciences USA*, **57 (2)**, 206–210.

Burton, A., C. G. Kilsby, H. J. Fowler, P. S. P. Cowpertwait, and P. E. O'Connel, 2008: Rainsim: A spatial-temporal stochastic rainfall modelling system. *Environmental Modelling & Software*, **23**, 1356–1369.

Chandler, R. E., 1997: A spectral method for estimating parameters in rainfall models. *Bernoulli*, **3 (3)**, 301–322.

Chandrasekar, B. and B. K. Kale, 1984: Unbiased statistical estimation functions for parameters in presence of nuisance parameters. *Journal of Statistical Planning and Inference*, **9**, 45–54.

Chiu, S.-T., 1988: Weighted least squares estimators on the frquency domain for the parameters of a time series. *The Annals of Statistics*, **16 (3)**, 1315–1326.

Cowpertwait, P. S. P., 1991: Further developments of the Neyman-scott clustered point process for modelling rainfall. *Water Resources Research*, **27 (7)**, 1431–1438.

Cowpertwait, P. S. P., V. Isham, and C. Onof, 2007: Point process models of rainfall: developments for fine-scale structure. *Proc. R. Soc. A*, **463**, 2569–2587.

Cox, D. R. and V. Isham, 1980: *Point Processes*. Chapman and Hall.

Dahlhaus, R., 2000: A likelihood approximation for locally stationary processes. *The Annals of Stat*, **28 (6)**, 1762–1794.

Dahlhaus, R., 2009: Local inference for locally stationary time series based on empirical spectral measure. *Journal of Econometrics*, **151**.

Draper, N. R. and I. Guttman, 1995: Confidence intervals versus regions. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **44 (3)**, 399–403.

Durbin, J., 1960: Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society. Series B*, **22 (1)**, 139–153.

Fox, R. and M. S. Taqqu, 1986: Large sample properties of parameter estimates for strongly dependent gaussian time series. *The Annals of Statistics*, **14**, 517–532.

Fuentes, M., 2002: Spectral methods for nonstationary spatial processes. *Biometrika*, **89**, 197–210.

Fuentes, M., 2007: Approximate likelihood for large irregularly spaced data. *Journal of the American Statistical Association*, **102**, 321–331.

Giraitis, L. and P. Robinson, 2001: Whittle estimation of ARCH models. *Econometric Theory*, **17**, 608–631.

Giraitis, L. and M. S. Taqqu, 1999: Whittle estimator for finite-variance non-gaussian time series with long memory. *The Annals of Statistics*, **27 (1)**, 178–203.

Godambe, V. P., 1960: An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, **31 (4)**, 1208–1211.

Godambe, V. P., 1976: Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, **63 (2)**, 277–284.

Hall, A. R., 2005: *Generalized Method of Moments*. Oxford University Press.

Hannan, E. J., 1973: The asymptotic theory of linear time series models. *Journal of Applied Probability*, **10**.

Hansen, L. R., 1982: Large sample properties of generalized method of moments estimator. *Econometrica*, **50**, 1029–1054.

Hauser, M., 1998: Maximum likelihood estimators for ARMA and ARFIMA models: a Monte Carlo study. *J. Statistical Planning and Inference*, **XX**.

Heyde, C. C., 1997: *Quasi-Likelihood and its applications*. New York: Springer.

Imbens, G. W., R. H. Spady, and P. Johnson, 1998: Information theoretic approaches to inference in moment conditions models. *Econometrica*, **66**, 333–357.

Jesus, J. and R. E. Chandler, 2011: Estimating functions and the generalized method of moments. *Interface Focus*.

Kendall, M. and J. K. Ord, 1990: *Time Series (third edition)*. Edward Arnold.

Li, H. and G. Yin, 2009: Generalized method of moments estimation for linear regression with clustered failure time data. *BIOMETRIKA*, **96 (2)**, 293–306, doi: {10.1093/biomet/asp005}.

Lindsay, B. G. and A. Qu, 2003: Inference functions and quadratic score tests. *Statistical Science*, **18**, 394–410.

Mikosch, T., T. Gadrich, C. Kluppelberg, and R. J. Adler, 1993: Parameter estimation for ARMA models with infinite variance innovations. *The Annals of Statistics*, **23 (1)**, 305–326.

Mukhopadhyay, P., 2007: On optimal estimating function in the presence of nuisance parameters. *Comm in Stat - Theory and Methods*, **36**, 1867–1876.

Northrop, P. J., 2006: Estimating the parameters of rainfall models using maximum marginal likelihood. *Student*, **5 (3/4)**, 173–183.

Onof, C., R. E. Chandler, A. kakou, P. Northrop, H. S. Wheater, and V. Isham, 2000: Rainfall modelling using Poisson-cluster processes: a review of developments. *Stoch. Env. Res. & Risk Ass*, **14**, 384–411.

Onof, C. and H. S. Wheater, 1994: Improvemnts to the modelling of british rainfall using a modified random parameter Bartlett-Lewis rectangular pulse model. *Journal of Hydrology*, **157**, 177–195.

Owen, A. B., 2000: *Empirical Likelihood*. Chapman & Hal.

Priestley, M. B., 1981: *Spectral analysis and time series*. Academic Press.

Ramesh, N. I., 1998: Temporal modelling of short-term rainfall using cox processes. *Environmetrics*, **9**, 629–643.

Rice, J., 1979: On the estimation of the parameters of a power spectrum. *Journal of Multivariate Analysis*, **9**, 378–392.

Robinson, P. M., 1978: Alternative models for stationary stochastic processes. *Stochastic Processes and their Applications*, **8**, 141–152.

Rodriguez-Iturbe, I., D. R. Cox, and V. Isham, 1987: Some models for rainfall based on stochastic point processes. *Proc R. Soc. Lond.*, **A410**, 269–288.

Rodriguez-Iturbe, I., D. R. Cox, and V. Isham, 1988: A point process model for rainfall: further developments. *Proc R. Soc. Lond.*, **A417**, 283–298.

Rodriguez-Iturbe, I., V. Gupta, and E. Waymire, 1984: Scale considerations in the modelling of temporal rainfall. *Water Resources Research*, **20**, 1611–1619.

Sweeting, T. J., 1990: Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, **8 (6)**, 1375–1381.

Wheater, H. S., et al., 2005: Spatial-temporal rainfall modelling for flood risk estimation. *Stoch. Env. Res. & Risk Ass*, **19**, 403–416.

Whittle, P., 1953: Estimation and information in stationary time series. *Ark. Mat.*, **2**, 423–434.

Yoo, C., D. Kim, T. W. Kim, and K. N. Hwang, 2008: Quantification of drought using a rectangular pulses poisson process model. *Journal of Hydrology*, **355**, 34–48.