University College London

Doctoral Thesis

# Bayesian Inference for Multiple Gaussian Graphical Models, with an Application to the SABRE Cohort Study

*Author*:
Marco Molinari

*Supervisor*:
Prof. Maria De Iorio

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy in the
Department of Statistical Science
Faculty of Mathematical and Physical Sciences

May, 2020

# Declaration of Authorship

I, Marco MOLINARI, declare that this thesis titled “Bayesian Inference for Multiple Gaussian Graphical Models, with an Application to the SABRE Cohort Study” and the work presented in it are my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Candidate　　　　　Date

# Abstract

The motivating application of this thesis is the Southall And Brent REvisited (SABRE) study, a tri-ethnic cohort study conducted in the UK. We analyse the metabolic and phenotypic data of SABRE, with a view to identifying potential ethnic differences in metabolite levels and associations, to gain a better understanding of different risk of cardio-metabolic disorders and diabetes development across ethnicities. Our first focus is on modelling the distribution of Homeostasis Model Assessment Insulin Resistance (HOMA IR), which is a frequent precursor to the development of type 2 diabetes. We adopt a Bayesian nonparametric random intercept/error model, which allows for data-driven clustering of patients, while adjusting for individual metabolite levels. The results highlight the presence of sub-populations in the data, with diverse levels of HOMA IR related to different metabolic profiles. The second stage of research is concerned with the development of Bayesian multiple graphical models, to infer the structure of association between metabolites, across ethnicities. In the first model we adopt a Dependent Generalised Dirichlet Process (DGDP) prior on the edge inclusion probabilities, allowing the estimation of multiple Gaussian Graphical Models (GGMs) in a sparse multivariate regression framework (i.e. the seemingly unrelated regression (SUR) model). The DGDP prior allows a convenient way to share information across edges and multiple graphs, while within the sparse SUR framework we impose sparsity on the precision matrices, through the Stochastic Search Structure Learning prior, and on the regression covariates, through the Horseshoe prior. In our final contribution, we propose a dynamic multiple groups extension of the Nodewise Regression technique. We allow multiple groups of different sample sizes to be analysed. We estimate dynamic multiple graphs adopting a dynamic shrinkage prior, which allows to share information across times and groups, while ensuring good computational scalability. Posterior inference is performed through Markov Chain Monte Carlo (MCMC).

# Impact Statement

The need to understand the mechanisms underpinning the development of diabetes and cardiovascular diseases is of paramount importance given the huge implications that these ailments can have on the quality of life of affected individuals and society as a whole. The purpose of this thesis is to provide novel statistical tools through which we can analyse a wide range of biostatistical data, in particular multivariate outcomes and covariates of interest. The novelty of the application lies in the approach that we take in analysing a wide range of metabolites, within a graphical modelling framework, elucidating ethnic differences and linking our findings to diseases of interest, such as diabetes and cardiovascular diseases. The proposed methodologies allow us to directly analyse multiple outcomes, conditionally to confouders of interest, and to account for the existence of multiple related sub-populations. The development of this work can be of great use inside the academic environment. First and foremost for the SABRE team who continue to study the SABRE datasets, of which the metabolomics data is a small part. Moreover, the application of the proposed models is favoured by the availability of computational routines that can be used in R. Other academic research bodies can benefit from our work, for example the UCLEB (UCL-LSHTM-Edinburgh-Bristol) consortium based at UCL, whose research is focused on genomics and metabolomics, could benefit from our contributions.

The impact of the thesis is not limited to the academic environment, in fact, the methodological developments, as well as the computational aspects, can be useful in industry. For example, biopharmaceutical companies that deal with metabolomics, and in general *'omics* data, could use the proposed models in the analysis of complex datasets that are readily available thanks to the technology advances in profiling genes and metabolites. The application of such methodology is not limited to the medical field, in fact it can also be employed for a variety of problems, such as in socio-economic disciplines, where multivariate data are commonplace.

# Dedication

To Marta and to my family.

# Acknowledgements

First and foremost I would like to thank my PhD advisor Maria De Iorio for her continuous support and trust during these three years. I am thankful for her extremely valuable advice, both about my work and about my career choices. I also would like to express my gratitude to my co-advisor Therese Tillin, who has been always present and helped me a lot in shaping my research work. It has been a pleasure and a satisfaction to work with them.

I would like to thank my collaborators: Alun Hughes and Nishi Chaturvedi, who gave me many useful inputs and advice, helping me improving the quality of my thesis. I am also grateful to Petros Dellaportas for his tips and availability and to Chakkapas Visavakul for always having the patience to solve my many IT problems.

These three years spent in London, and in particular at UCL, have been an invaluable experience and a great opportunity of personal growth. I have met many wonderful friends that contributed to make this experience much better and I am grateful for this to Marcel, Federico, Marta, Andrea, Xiaochen, Lifeng and William. A special thanks goes to Andrea Mainenti, who helped me from the very first day and treated me like a brother. It is also because of him if I managed to reach this point.

I am thankful to my family: Raul, Grazia, Fausto and Monica, who always supported my choices and believed in this opportunity.

Finally, my deepest gratefulness goes to Marta who always stood by me and constantly remembered me the importance of my values. Her support and love have always been there in the most difficult times.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 The SABRE study

The Southall And Brent REvisited (SABRE) study is a cohort research study involving nearly 5000 European, South-Asian (from the Indian subcontinent) and African-Caribbean men and women (Tillin et al., 2010). SABRE is funded by two registered charities, the Wellcome Trust and The British Heart Foundation, and is run by a team now based at University College London with support from experts at the Universities of Bristol, Cambridge, Edinburgh, Exeter, Glasgow, Newcastle, Oxford and Washington (Seattle, USA). The first measures available refer to the baseline study dating back to 1988-1991. A first follow-up is available from 2008-2011 and a second follow up is being currently completed. Recruited volunteer patients in SABRE have been analysed twice, first in a fasting state and a second time two hours after an oral glucose test (post oral glucose tolerance test, OGTT). Here we focus on the fasting state measurements for both the baseline and follow-up data.

### 1.1.1 The data

The SABRE study covers several research projects involving different kinds of data. In this work we concentrate on the metabolomics data, with a focus on ethnic differences and Insulin Resistance. Metabolomics is the large-scale study of small molecules, commonly known as metabolites, within cells, bio-fluids, tissues or organisms. Collectively, these small molecules and their interactions within a biological system are known as the *metabolome.* In Figure 1.1 we show a summary of the relationships occurring between different *'omics* fields. With the term *'omics* we refer to the collection of disciplines which includes genomics, metabolomics, proteomics, etc. Some

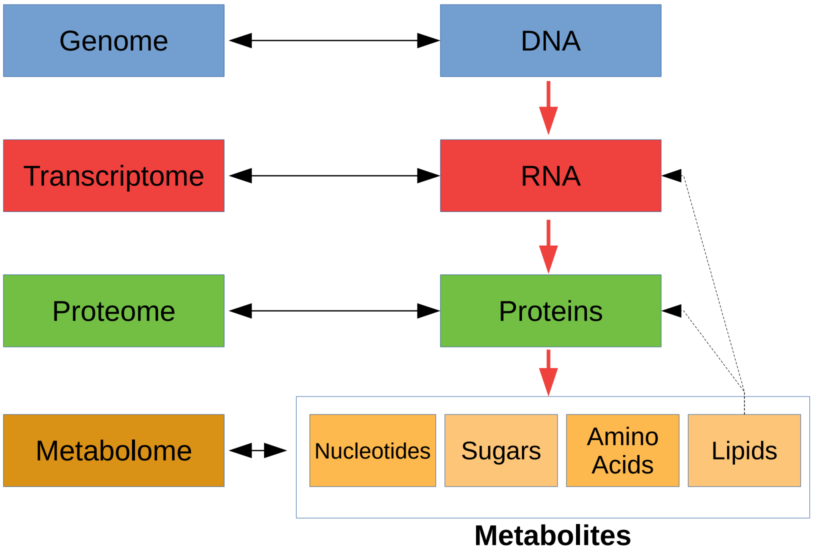


Figure 1.1: Examples of *'omics* fields of study and their relationship. The genome, the ensemble of every gene in the DNA, contains the instructions, which are translated and executed by the RNA to build proteins. Finally, the metabolome is the collection of all small molecules produced by the metabolic reactions occurring within an organism (the metabolism).

examples of small molecules include sugars, lipids, amino acids, fatty acids, phenolic compounds etc. The actual dataset contains over 200 metabolites and ratios of metabolites, and a major proportion of them is represented by lipoproteins. Lipoproteins are classified according to their density and we can distinguish *very-low density lipoprotein* (VLDL), *low density lipoprotein* (LDL), *intermediate density lipoprotein* (IDL) and *high density lipoprotein* (HDL). Each lipoprotein is further categorised according to its dimension, ranging from extra-extra large to extra small. The task of these lipoproteins is to carry through the blood stream lipids compounds which are *triglycerides*, *phospholipids*, *cholesterol esters* and *free cholesterol.* In this work we focus on a subset of all the metabolites, excluding from the analysis all the ratios of metabolites, reducing the total number of measurements to 88. All the metabolites concentrations are measured in millimoles per litre (mmol/L). Table D.1 contains the selected molecules along with the acronym used throughout the thesis. A considerable proportion of the selected metabolites has a high degree of correlation, in particular, lipoproteins compounds are very correlated (Figure E.1).

Insulin resistance and ethnic differences are of major interest in this research. Insulin resistance represents the "resistance"of the body towards the action of insulin aimed to decrease the level of glucose in the blood. In other terms, in a healthy individual, when the glucose concentration in the blood is too high, the beta cells in the pancreas produce insulin in order to facilitate the absorption of glucose by other cells for energy production, reducing its concentration. In the status of insulin resistance, the organism (liver, fat, muscles, etc.) does not respond properly to insulin production, diminishing the proportion of glucose absorbed, as a consequence the pancreas needs to increase the production of insulin to maintain the right glucose level in the blood. This disordered state can potentially lead to the development of type 2 diabetes. Diabetes poses an enormous individual and societal burden, with a high risk of major complications and diminished quality and length of life. Hence, it is imperative to understand causal mechanisms to identify those at highest risk and to tailor preventive and therapeutic measures for appropriate periods during the life course. The global epidemic of type 2 diabetes disproportionately affects non-European ethnic groups. South-Asians (from the Indian subcontinent) form the largest ethnic minority group in the UK with prevalence of diabetes in South-Asians estimated to be 2-4 times higher than that of the general population (Sproston and Mindell, 2006). Migrant populations of African-Caribbean origin, although smaller in number, are also at greater risk of developing type 2 diabetes, with prevalence also estimated at 2-4 times that of the general UK population (Sproston and Mindell, 2006). Research to date suggests that insulin resistance and differences in body fat distribution explain some of the ethnic differences in diabetes risk, but underlying mechanistic pathways are altogether poorly understood, although likely to involve a complex interplay between environmental, behavioural, metabolic, genetic and epigenetic disturbances.

The first part of the thesis is devoted to an exploratory analysis of the SABRE metabolic and anthropometric data. We exclude from the analysis the individuals with known diabetes at the time of the first visit. This choice is motivated by the fact that people with known diabetes were already receiving treatments that may alter their metabolite levels. We include control variables, such as the Homeostasis Model Assessment Insulin Resistance (HOMA IR), an index of Insulin Resistance (Matthews et al., 1985) proportional to the product of blood concentration of insulin and glucose (both measured in millimoles per liter (mmol/L)), three important enzymes, alanine aminotransferase, aspartate aminotransferase and gamma glutamyl transferase, which are indicators of the liver health. We include anthropometric variables, especially measures of body fat distribution, of which Waist to Hip Ratio (WHR) is the most relevant, being a measure of central adiposity. Age, sex and smoking status are also considered as standard control variables. The SABRE study specifically focuses on ethnic differences and this characteristic is reflected in our analysis, where we consider ethnicity as a potential factor for metabolic differences. In Table 1.1 we provide a brief description of the number of complete observations, proportion of females patients, mean level of HOMA IR, Age, Waist to Hip Ratio and proportion of current smokers, stratified by ethnicity. In Figure 1.2 we show the empirical distribution of HOMA IR for each ethnic group. We can see the difference, especially between Europeans and South-Asians, in the empirical distribution. The South-Asian density is slightly shifted to the right and has a heavier right tail, to indicate the presence of a greater number of less healthy individuals. In Table D.2 we report the full list of covariates that are used throughout the thesis, together with their acronyms (the total number of covariates is 21).

Table 1.1: Number of patients, percentage of females, mean level of HOMA IR, mean age, average Waist to Hip ratio and percentage of current smokers in each ethnic group, calculated on the baseline fasting sample data

| **Ethnicity** | **Obs.** | **Females** | **HOMA IR** | **Age** | **WHR** | **Current Smokers** |
|---|---|---|---|---|---|---|
| | | **percentage** | | | | **percentage** |
| Europeans | 1030 | 13.2% | 1.981 | 53.4 | 0.922 | 29.5% |
| South-Asians | 772 | 20.9% | 2.808 | 50.6 | 0.953 | 14.5% |
| Africans-Caribbean | 86 | 7% | 2.135 | 52.9 | 0.932 | 24.4% |

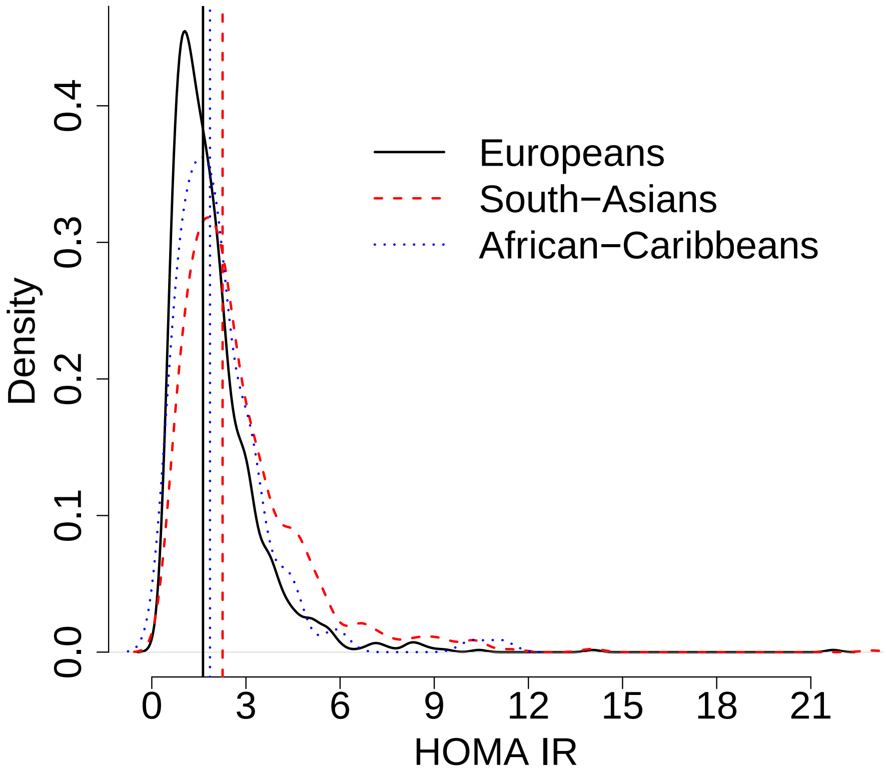


Figure 1.2: The plot shows the HOMA IR distribution, stratified by ethnicity (where the vertical line represents the sample median of each group). The plot is obtained through a kernel density estimate, with a Gaussian kernel with bandwidth equal to 1, as given by the standard function *density* employed in the `R` package *stats*.

## 1.2 Proposed methods

In this work, we address multiple questions about the impact that ethnic differences can have on the structure of associations among metabolites, the different risk of development of cardiovascular diseases and development of type 2 diabetes. We first address the question of how Insulin Resistance is affected by individual metabolic profiles and how its level differs across ethnicities. We use a Bayesian nonparametric random intercept/random error regression model to approach these questions. The adoption of Bayesian nonparametric methods, in particular the use of a Dirichlet Process type prior (Ferguson, 1973; Antoniak, 1974), allows data-driven clustering of individuals without the need to specify a priori the number of clusters. Using an extension of the Dirichlet Process, the Dependent Generalised Dirichlet Process, we allow the clustering to be ethnic specific, while borrowing information across groups. Moreover, the random error in the regression makes the model more robust to outliers.

To address the question of how ethnic differences can have an impact on the structure of associations of the metabolites, we adopt Bayesian Graphical Models, imposing sparsity to handle the great number of dependent and independent variables involved. We first analyse the baseline data, developing a model which is able to treat multiple groups of observations, corresponding to multiple graphs (defined by ethnicity). Under the Bayesian framework, we employ a nonparametric prior on the edge inclusion probabilities of the graph and we propose an extension of the model to estimate multiple graphs. A wealth of models is available to handle multivariate outcomes and possibly different sets of covariates. A pioneering method is the Seemingly Unrelated Regressions (SUR) Model introduced by Zellner (1971). The SUR model can be seen as a generalization of the linear regression model, where there are multiple regression equations, each one having its own dependent variable and possibly a specific set of regression covariates. The peculiarity of this model is that the regressions are not estimated independently, but they are linked together by the error terms that can be correlated. The SUR model offers flexibility but comes with a large number of parameters to be estimated. For this reason, we adopt a Sparse SUR approach (Wang, 2010; Billio et al., 2017), that is a SUR model where the coefficients associated with the regression covariates and the error precision matrix can shrink to zero. We treat the metabolites as the outcomes of a set of linear regressions where the covariates for each equation are represented by the variables previously introduced (Table D.2). The precision matrix between equations is specified conditionally on an underlying graphical model, which determines the patterns of conditional independences of the metabolites. Zeros in the error precision matrix indicate conditional independences between variables (Lauritzen, 1996). We use the Stochastic Search Structure Learning (SSSL) prior of Wang (2015), which allows efficient posterior inference of the underlying graph, and conditional on the graph, allows inference of the precision matrix. Sparsity in the vector of regression coefficients indicates a dependence of the mean of the response only on a subset of covariates. Different approaches have been proposed to handle variable selection in a linear regression context. See, among the others, O'Hara and Sillanpää (2009) for a review of methods for variable selection in a single linear regression framework. When moving to the context of simultaneous multiple regressions we need to perform the selection jointly on all the outcomes. Brown et al. (1998) uses an efficient MC3 method that exploits the marginal likelihood of the

Matrix Normal distribution. However, their model assumes that all the responses have the same predictors, which leads to the same selected subset for each response. The SUR model is broader and allows for different predictors in each outcome, hence a potential different selection. In order to maintain a good scalability with respect to the number of variables involved, we adopt the continuous Horseshoe shrinkage prior of Carvalho et al. (2010) which shrinks small or negligible coefficients to zero, while leaving important coefficients unaffected thanks to its heavy tails. We model the edges inclusion probabilities of the multiple GGMs with a Dependent Generalised Dirichlet Process (DGDP, Barcella et al. (2017)) prior, allowing clustering of the coefficients across multiple groups and sharing information about the possible common structure. The DGDP is a generalisation of the well-known Dirichlet Process (DP, Ferguson (1973) and Antoniak (1974)), which posses greater flexibility thanks to the richer parametrisation and enables the introduction of dependence between random measures through the weights of the process.

The last part of the project concerns the development of a Bayesian model to analyse temporal patterns of association among a set of metabolites over different groups of patients. We develop a model to jointly estimate multiple graphs, corresponding to the ethnicities, and multiple time points, which in our case correspond to the baseline and follow-up data of the SABRE study. We are interested in identifying potential ethnic differences in metabolite levels and associations as well as their evolution over time, with the aim of gaining a better understanding of different risk of development of cardio-metabolic disorders across ethnicities. Within a Bayesian framework, we employ Nodewise Regression technique (Meinshausen and Bühlmann, 2006) to infer the structure of the graphs, borrowing information across time and ethnicities. The response variables of interest are metabolite levels measured at two time points for two ethnic groups, Europeans and South-Asians. We use Nodewise Regression to estimate the high-dimensional precision matrix of the metabolites, by regressing each metabolite on the remaining and imposing sparsity on the regression coefficients through a dynamic extension of the Horseshoe shrinkage prior proposed by Kowal et al. (2017). The Horseshoe prior has some desirable characteristics as detailed above, moreover employing a continuous prior allows fast posterior inference. Graph selection is based on a functional of the posterior distribution as described in Carvalho et al. (2010).

### Acknowledgements

Alun Hughes and Nishi Chaturvedi receive support from the National Institute for Health Research University College London Hospitals Biomedical Research Centre, and work in a unit that receives support from the UK Medical Research Council (Programme Code $MC_UU_$12019/1). The SABRE study was funded at baseline by the Medical Research Council, Diabetes UK, and the British Heart Foundation. At follow-up the study was funded by the Wellcome Trust (067100, 37055891 and 086676/7/08/$Z$), the British Heart Foundation ($PG$/06/145, $PG$/08/103/26133, $PG$/12/29/29497 and $CS$/13/1/30327) and Diabetes UK (13/0004774). The SABRE study team also acknowledges the support of the National Institute of Health Research Clinical Research Network (NIHR CRN). We are extremely grateful to all the people who took part in the study, and past and present members of the SABRE team who helped to collect the data.

## 1.3 Outline of the thesis

The thesis is organised as follows.

- In Chapter 2, we introduce the fundamental ideas behind Bayesian inference and graphical models needed for the development of our models.

- In Chapter 3, we perform several exploratory analyses on the metabolites and the covariates previously listed. We perform network analysis on the metabolites and variables selection in a linear regression context considering each metabolite as outcome. We also build a simple change-point model aimed to discover the ideal cut-off to be used in the Differential Network analysis.

- In Chapter 4, we build a nonparametric random intercept / random error regression model to investigate the ethnic differences in the distribution of Insulin Resistance conditioning on the metabolites levels. Thanks to the clustering property of the Bayesian nonparametric prior, we can highlight clusters of individuals with different metabolic profiles and risk of development of type 2 diabetes.

- In Chapter 5, we present our approach to the estimation of multiple Gaussian Graphical Models. We specify the DP type prior distribution on the edge inclusion probabilities and we study in detail the graphical properties that this prior generates. We test the performance of the model in simulations and we analyse the SABRE dataset, providing in-depth analysis of the results.

- In Chapter 6, we propose an approach to infer dynamic multiple graphical models, providing evidence of the model performance on simulated datasets. We analyse the baseline and follow-up data from SABRE, providing a set of interpretable results with a pathway enrichment analysis.
- Finally, in Chapter 7,, we discuss the main findings and contributions of this project and we outline some open research questions.

# Chapter 2

# Methodological Background

In this research project, we adopt a Bayesian framework. The fundamental characteristic of the Bayesian approach is the way we treat an unknown parameter $\theta$, which is the object of inference and represents a characteristic of the *population* under study. Rather than consider it as an unknown constant value, as in the frequentist approach, $\theta$ is a random variable, so the object of inference is the distribution of $\theta$.

Most of this thesis is related to Graphical Models and in particular Gaussian Graphical Models (GGMs). A Graphical Model is a convenient way of defining a set of dependences over a multivariate random variable. If we further assume a multivariate Normal distribution, then we have a Gaussian Graphical Model (GGM). The resulting graph determines the structure of the precision matrix that characterises the distribution and defines a set of conditional independences between the coordinates of the multivariate random variable.

Here we provide an introduction to the central aspects of Bayesian inference and Graphical Models.

## 2.1 Introduction to Bayesian Inference

We use probabilities to express information beliefs, or uncertainty, about an unknown quantity of interest. Bayes' rule is a rational method through which we can update our beliefs about the unknown quantity, given some new information (Hoff, 2009). The process of learning through Bayes' rule is called Bayesian Inference. We define $\theta \in \Theta$ to be the random variable denoting a population characteristic, where $\Theta$ is the *parameter space*, that is the set of all possible values that $\theta$ can assume. We then call

$y$ the subset of observations, from the population, on which we measure the variable of interest. In order to make inference on $\theta$ we need to set up our model, which consists of a *prior distribution* and a distribution over the sampling model.

1. The probability distribution $p(\theta)$ describes our uncertainty about $\theta$ before the experiment takes place. We may have a vague idea about the true value of $\theta$, so the *prior* will reflect this information.

2. Given a specific value of $\theta$, the sampling model is $p(y \mid \theta)$. This represents the distribution of our observations $y$ given $\theta$ known.

Once a dataset $y$ is obtained we can update our uncertainty about $\theta$. Thus we need to update the *prior distribution* on $\theta$ with the newly gathered information. Bayes' theorem allows updating uncertainty through

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int_{\Theta} p(y \mid \tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} \tag{2.1}$$

Equation (2.1) is the *posterior distribution*, which is a combination of the likelihood $p(y \mid \theta)$ and the *prior* $p(\theta)$ in the numerator. The denominator is a normalizing constant (it does not contain $\theta$) which ensures that the new distribution integrates to one.

There are cases where the posterior is of simple derivation while in other cases it will require numerical approximations. We can have a *prior* which is conjugate or non-conjugate. Conjugacy refers to the fact that *prior* and *posterior* belong to the same family of probability distributions. When conjugacy is achieved, we can obtain the *posterior* distribution avoiding the computation of the integral in the denominator of (2.1), which can be problematic. There are several combinations of sampling models and *prior* that are conjugate. For example when $p(y \mid \theta)$ and $p(\theta)$ are both Gaussian distributions (where $\theta$ represents the mean), $p(y \mid \theta)$ Binomial and $p(\theta)$ Beta or $p(y \mid \theta)$ Poisson and $p(\theta)$ Gamma. When the *prior* is not conjugate to the data model, we can not simply ignore the denominator in Bayes' rule. Here lies the main reason that prevented the use of Bayesian inference to spread until the late 90's. Often in many applications, we have to deal with updates requiring complex integrals to be solved, so it is only with the development of efficient computational algorithms that Bayesian methods have become popular.

### 2.1.1 Monte Carlo

With a standard conjugate distribution, we can easily get the corresponding posterior distribution of $\theta$. However, we are often interested in a function of $\theta$, say $g(\theta)$, or a comparison between two or more populations, of which we want the posterior distribution. For example we can be interested in the posterior of $|\theta_1 - \theta_2|$, $\frac{\theta_1}{\theta_2}$ or $\log\left(\frac{\theta_1}{1-\theta_1}/\frac{\theta_2}{1-\theta_2}\right)$. To calculate such quantities, we can rely on Monte Carlo (MC) methods. Given a posterior distribution from which we can draw random values, we can get a numerical approximation of any function using these random samples. Let $\theta$ be a parameter of interest and let $y_1, \ldots, y_n$ be a sample from a distribution $p(y_1, \ldots, y_n \mid \theta)$. Suppose we can sample a number $S$ of independent random values of $\theta$ from the posterior distribution $p(\theta \mid y_1, \ldots, y_n)$:

$$\theta^{(1)}, \ldots, \theta^{(S)} \overset{iid}{\sim} p(\theta \mid y_1, \ldots, y_n)$$

Then, the empirical distribution of the samples $\left\{\theta^{(1)}, \ldots, \theta^{(S)}\right\}$ approximates $p(\theta \mid y_1, \ldots, y_n)$, with greater precision as $S$ increases. Moreover, by the law of large numbers we obtain that if $\theta^{(1)}, \ldots, \theta^{(S)}$ are *iid* samples from $p(\theta \mid y_1, \ldots, y_n)$, then

$$\frac{1}{S}\sum_{s=1}^{S} g\left(\theta^{(s)}\right) \longrightarrow \mathrm{E}\left[g(\theta) \mid y_1, \ldots, y_n\right] = \int g(\theta)p(\theta \mid y_1, \ldots, y_n)d\theta \qquad \text{for } S \longrightarrow \infty$$

This result implies that any by-product of the posterior distribution can be accurately approximated by a MC sample.

The previous framework works well when we can simulate from the posterior distribution of $\theta$. However, in many cases when we treat a model with multiple parameters, the joint posterior distribution may be non-standard and difficult to sample. In this event, a class of numerical approximation based on Markov Chain Monte Carlo (MCMC) is available. MCMC are a class of algorithms for sampling from a probability distribution, through the construction of a Markov Chain that has the desired joint posterior probability distribution as its equilibrium distribution. This method is based on sampling from a certain distribution (for example a factorization of the joint posterior), whose realisations are not *iid* as in a MC, but have a Markovian dependence, i.e. each sample is independent of the past given the most recent value. It is convenient to distinguish the two main categories of MCMC, the Gibbs sampler and the Metropolis-Hastings. See Robert and Casella (2011), among the others, for a discussion about the development of these algorithms. The Metropolis algorithm precedes the Gibbs sampler chronologically, but we will introduce them in a reverse order.

### 2.1.2 Gibbs Sampler

The Gibbs sampler is a technique for generating random samples from a target distribution without the need of calculating the density of the distribution itself. Instead of solving difficult, or often impossible calculations, we can rely on a sequence of easier computations. Among the most notable literature on Gibbs sampler, we have Geman and Geman (1984) and Gelfand and Smith (1990) with their pioneering works. We follow the work of Casella and George (1992) to explain the Gibbs sampler.
Suppose we have $p+1$ unknown parameters, with joint density $f(\phi, \theta_1, \ldots, \theta_p)$, and we want to make inference on them. We can be interested, for example, in the marginal density of the parameter $\phi$:

$$f(\phi) = \int_{\Theta_1} \cdots \int_{\Theta_p} f(\phi, \theta_1, \ldots, \theta_p) d\theta_1 \ldots d\theta_p$$

There are many cases where the analytical solution of such integral is very difficult or even impossible, and even a direct numerical solution may not work, such as a MC, when the multidimensional parametric space is too wide to be properly covered. Instead of computing directly $f(\phi)$, we can use the Gibbs sampler to iteratively generate samples from $f(\phi)$. Consider the case where we have a bivariate random variable $(\theta, \phi)$, through the Gibbs sampler we can generate samples of $f(\phi)$ by sampling from the *full conditional* distributions $f(\phi \mid \theta)$ and $f(\theta \mid \phi)$. The term *full conditional* refers to the distribution of a variable conditioned on all the other variables forming the joint distribution. Consequently, a condition to use the Gibbs sampler is the availability of these distributions and the ability to generate samples from them. The Gibbs sampling procedure starts by specifying an initial value for $\theta^{(0)} = \theta_0$ (we could equally start from $\phi$ instead) and then iteratively sample from the full conditionals alternating

$$\begin{aligned} \phi^{(s)} &\sim f(\phi \mid \theta = \theta^{(s)}) \\ \theta^{(s+1)} &\sim f(\theta \mid \phi = \phi^{(s)}) \end{aligned} \tag{2.2}$$

Under reasonably general conditions (Casella and George, 1992) the distribution of $\phi^{(s)}$ converges to the true marginal $f(\phi)$ as $s \longrightarrow \infty$, where $s$ is the number of iterations of the algorithm. Then, effectively after a certain number of iterations of the sampler, we are in a situation where we are sampling from the equilibrium distribution and we can regard the samples $(\phi^{(s)}, \theta^{(s)})$ as draws from the joint posterior distribution $f(\phi, \theta)$. The convergence of this algorithm is not always guaranteed. In Casella and George (1992) a sufficient condition for the convergence is that the marginal distribution is a proper density, i.e. $\int f_\phi(\phi)\ d\phi < \infty$. General convergence conditions for the Gibbs sampler are discussed in detail by Schervish and Carlin (1992).

### 2.1.3 Metropolis-Hastings algorithm

In order to use the Gibbs sampler, we must be able to sample from the full conditional distributions. This is not the only limitation of the Gibbs sampler, in fact, there are cases where the mixing of the sampler is very slow, which means that the algorithm remains stuck in a region of the parameter space with high density and it may take a very long time to explore all regions with significant probability mass. The Metropolis-Hastings sampler, in contrast, does not require a direct sampling from the full conditional distributions, but it rather makes use of a proposal distribution. This sampler was first introduced by Metropolis et al. (1953), who developed the Metropolis in the case of a symmetric proposal distribution. It was then extended to a more general type of proposal by Hastings (1970).

As before, we are interested in the joint posterior distribution $f(\theta, \phi)$, which is approximated through a Metropolis-Hastings sampler by drawing values from a proposal distribution alternately for each parameter or with a joint proposal. The proposal distribution is selected so that we can easily generate random values from it. There are two kinds of proposal, symmetric such as the Gaussian and Uniform distributions and asymmetric like the Gamma distribution. The proposed value is not automatically accepted and stored as in the Gibbs sampler, but it has to be accepted through a stochastic acceptance step. A Metropolis-Hastings algorithm for approximating $f(\theta, \phi)$, given a proposal distribution $q(\cdot)$ and a sampling model $p(y \mid \theta, \phi)$, runs as follows

- 1. sample $\theta^* \sim q_\theta(\theta \mid \theta^{(s)}, \phi^{(s)})$

  2. calculate the acceptance ratio:

$$r = \frac{p(y \mid \theta^*, \phi^{(s)}) p(\theta^*, \phi^{(s)})}{p(y \mid \theta^{(s)}, \phi^{(s)}) p(\theta^{(s)}, \phi^{(s)})} \times \frac{q(\theta^{(s)} \mid \theta^*, \phi^{(s)})}{q(\theta^* \mid \theta^{(s)}, \phi^{(s)})}$$

3. set $\theta^{(s+1)} = \theta^*$ with probability equal to $\min(1, r)$, otherwise $\theta^{(s+1)}$ remains equal to $\theta^{(s)}$

- 1. sample $\phi^* \sim q_\phi(\phi \mid \theta^{(s+1)}, \phi^{(s)})$

  2. calculate the acceptance ratio:

$$r = \frac{p(y \mid \theta^{(s+1)}, \phi^{(*)})p(\theta^{(s+1)}, \phi^{(*)})}{p(y \mid \theta^{(s+1)}, \phi^{(s)})p(\theta^{(s+1)}, \phi^{(s)})} \times \frac{q(\phi^{(s)} \mid \theta^{(s+1)}, \phi^{(*)})}{q(\phi^* \mid \theta^{(s+1)}, \phi^{(s)})}$$

  3. set $\phi^{(s+1)} = \phi^*$ with probability equal to $\min(1, r)$, otherwise $\phi^{(s+1)}$ remains equal to $\phi^{(s)}$

We can update each parameter individually or using block updates, i.e. updating more parameters together. It is sometimes the case that a multivariate update is required, over a subset of parameters, for mixing purposes and faster convergence. The Metropolis-Hastings algorithm is related to the Gibbs sampler, in fact the latter is a special case of the Metropolis, where each proposed new value is always accepted. When dealing with more complex models, we can mix the two algorithms, alternating steps of Gibbs with steps of Metropolis, depending on the posterior's complexity.

### Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is a particular type of Metropolis, whose proposal distribution is based on Hamiltonian dynamics. Hamiltonian dynamics is a well studied topic in physics, but the concept has been exploited in statistics to create an efficient way of constructing a proposal distribution for the Metropolis algorithm (Brooks et al. (2011), chapter 5). A Metropolis algorithm equipped with Hamiltonian proposals can quickly explore the state space, avoiding the random walk behaviour of independent Gaussian proposals and consequently, it can efficiently explore the posterior distribution of the parameters of interest.

Consider a parameter of interest $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$. The HMC works by adding an independent auxiliary random variable $\boldsymbol{\phi} \in \mathbb{R}^D$, with density $p(\boldsymbol{\phi}) = \mathrm{N}(0, M)$ which has the role of a *momentum* variable, where $M$ represents the mass matrix. The joint density $p(\boldsymbol{\theta}, \boldsymbol{\phi})$ can be factorised as $p(\boldsymbol{\theta})p(\boldsymbol{\phi})$ and has negative joint likelihood

$$H(\boldsymbol{\theta}, \boldsymbol{\phi}) = -\log p(\boldsymbol{\theta}) + \frac{1}{2}\log\left((2\pi)^D|M|\right) + \frac{1}{2}\boldsymbol{\phi}^T M^{-1}\boldsymbol{\phi} \tag{2.3}$$

In physics a Hamiltonian describes the sum of a potential energy function, which is $-\log p(\boldsymbol{\theta})$, defined as the position of $\boldsymbol{\theta}$, and a kinetic energy term $\boldsymbol{\phi}^T M^{-1}\boldsymbol{\phi}/2$. The gradient of (2.3) with respect to the position and momentum variables has a physical interpretation as the time evolution, with respect to a fictitious time $t$, of a dynamic system as given by Hamilton's equations

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial H}{\partial \boldsymbol{\phi}} = M^{-1}\boldsymbol{\phi} \qquad \frac{d\boldsymbol{\phi}}{dt} = -\frac{\partial H}{\partial \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) \tag{2.4}$$

In practice, the differential equations in (2.4) have to be approximated via numerical methods. The *leapfrog* method is a popular choice (Brooks et al., 2011) and it is adopted in `Stan` (Carpenter et al., 2017), a probabilistic programming language which allow the users to employ HMC to perform Bayesian inference. The approximation introduces a small bias which is then corrected introducing an accept-reject step.

### 2.1.4 Dirichlet Process

We provide a brief introduction of the Dirichlet Process, which is arguably the most widely used Bayesian nonparametric prior and is the starting point for many extensions developed in the literature. The DP can be viewed as a distribution of distributions (Ferguson, 1973; Antoniak, 1974). Sethuraman (1994) provides a constructive definition of this process, showing that if a random probability measure $P$ is distributed according to a DP, with mass parameter $\alpha$ and base measure $P_0$, then

$$P = \sum_{k=1}^{\infty} \psi_k \delta_{\theta_k} \tag{2.5}$$

where $\theta_1, \theta_2, \dots$ are *iid* realisations from $P_0$ and $\delta_{\theta_k}$ is the Dirac measure that assigns a mass probability of one in correspondence of the location $\theta_k$. The weights $\psi_k$ are generated according to the *stick breaking* construction (see Ishwaran and James (2001) for details):

$$\psi_k = \phi_k \prod_{j=1}^{k-1}(1-\phi_j), \qquad k = 2, 3, \dots \tag{2.6}$$

with the $\phi_k \overset{iid}{\sim} \text{Beta}(1, \alpha)$ and $\psi_1 = \phi_1$. By construction $0 \leq \psi_k \leq 1$ and $\sum_{k=1}^{\infty} \psi_k = 1$. In the following chapters the DP is extended to accommodate our modelling requirements.

### 2.1.5 Bayesian Shrinkage Priors

A wealth of literature is available about variable selection and regularisation in the Bayesian framework; see, for example, O'Hara and Sillanpää (2009) for a review on shrinkage priors. Two important examples are the class of two components discrete mixture priors, known as Spike and Slab (George and McCulloch, 1993) and the class of continuous shrinkage priors, of which examples are the Horseshoe prior and the Horseshoe+ prior (see Bhadra et al. (2017), among the others, for a review). The Spike and Slab approach implies a positive probability for the regression coefficient to be zero, but it can be computationally demanding with a relatively high number of parameters, due to the large state space. On the other hand, continuous priors are easier to implement and are usually more computationally efficient, although the probability for the coefficient to be exactly zero is zero. The Horseshoe prior (Carvalho et al., 2010) is characterised by an accentuated spike at zero to strongly shrink small or negligible coefficients, while leaving important coefficients unaffected thanks to its heavy tails (given by the tick tails of the half-Cauchy distribution). Therefore, it allows to effectively ignore spurious or redundant covariates (in the case of a regression), while retaining good scalability.

We introduce the Horseshoe prior for a simple Normal model with unknown mean parameter. Let $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{\theta}, \sigma^2 I_p)$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$ is an unknown mean parameter of dimension $p$ (assume $\sigma^2$ known). A Horseshoe prior on $\boldsymbol{\theta}$ is specified as follows

$$\begin{aligned} \theta_j \mid \lambda_j, \tau &\sim \mathrm{N}\left(0, \lambda_j^2 \tau^2\right) \\ \lambda_j &\sim \mathrm{C}^+\left(0, 1\right) \\ \tau &\sim \mathrm{C}^+\left(0, 1\right) \end{aligned} \tag{2.7}$$

with $j = 1, \ldots, p$. $\lambda_j^2$ is the local shrinkage parameter, specific for the $j-$th coordinate of $\boldsymbol{\theta}$, while $\tau^2$ represents the overall shrinkage level. The choice of a half-Cauchy distribution results in aggressive shrinkage over small or negligible coefficients and is therefore suitable for variable selection in a Bayesian context. $\mathrm{C}^+(0, 1)$ denotes the standard half-Cauchy distribution, with probability density function:

$$p(\lambda_j) = \frac{2}{\pi(1 + \lambda_j^2)}, \qquad \lambda_j > 0$$

The half-Cauchy is equivalent to a half-$t$ distribution with one degree of freedom, which corresponds to the distribution of the absolute value of a centred student-$t$

distribution. In particular, let $x$ a be random variable having a $t$ distribution with $\nu$ degrees of freedom, then $y = |x|$, where $|x|$ denotes the absolute value, is distributed as a half-$t$ distribution with $\nu$ degrees of freedom (Psarakis and Panaretoes, 1990). Carvalho et al. (2010) compare the performance of the variable selection based on (2.7) with that of a Spike and Slab prior (George and McCulloch, 1993), showing that the posterior selection given by the Horseshoe is consistent with that of the Spike and Slab. Moreover, the continuous nature of this prior allows efficient computations trough Gibbs sampler algorithms and it can be also easily employed with Bayesian software such as `Stan`.

## 2.2 Gaussian Graphical Models

Gaussian Graphical Models (GGMs) provide a fundamental tool to analyse the underlying dependence structure of a collection of random variables provided with a joint Normal distribution. It is worth to make a distinction between Network models and GGMs, the former are statistical models over an observable network while the latter are statistical models constructed over a non-observable network. GGMs were first introduced by Dempster (1972), exploiting the properties of the random variables that belong to the exponential family and combine them together with the covariance selection models. It is assumed that $M$ variables have a joint Normal distribution with probability density function

$$f(X) = (2\pi)^{-\frac{M}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}X^T\Sigma^{-1}X\right\} \tag{2.8}$$

where $|A|$ denotes the determinant of the matrix $A$. Let $\Sigma^{-1} = \Omega$ be the precision matrix or, as it is called in Dempster (1972), the concentration matrix, whose elements are *concentrations*. In order to introduce sparsity in the precision, we can restrict some parameters $\omega_{ij}$ to be zero (where $\omega_{ij}$ represents the element in position $(i, j)$ in the matrix $\Omega$). Dempster (1972) proved the existence of a unique estimate $\hat{\Sigma}$ of $\Sigma$ under such constraints. $\hat{\Sigma}$ is a maximum likelihood estimate and it is completely defined by the zero constraints on its inverse $\Omega$. Later, Wermuth (1976) showed that a zero in the precision matrix $\Omega$ corresponds to a zero partial correlation, under the assumption of Normality. This is a relevant result because a zero partial correlation implies a conditional independence of the corresponding coordinates of the multivariate Normal random variable.

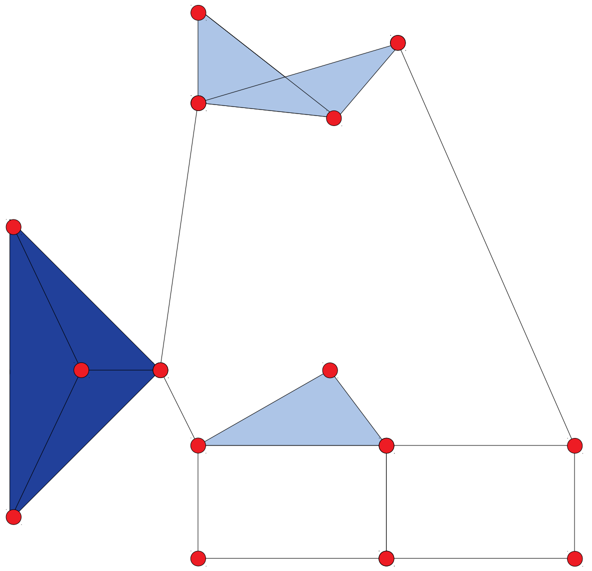


Figure 2.1: Example of Graph decomposition. The light blue triangles represent maximal cliques. The dark blue subgraphs are maximum cliques, i.e. the biggest cliques in the graph

### 2.2.1 General theory of graphical models

Here we provide some background on graph theory following the work of Lauritzen (1996). Let $G = (V, E)$, with $V = \{1, 2, \ldots, M\}$ and $E \subset \{(i, j) \in V \times V : i < j\}$, be an undirected graph whose vertices are associated with a $M$-dimensional vector $X$. From the vertex set $V$ we can extract a subset of vertices $A$, with $A \subseteq V$, which induces a subgraph $G_A$ that is defined by the subset of vertices $A$ and the implied subset of edges $E_A$. A graph, or a subgraph, is said to be *complete* if all the vertices are connected by an edge. The property of completeness is related to a key concept in graph theory, the *clique*. A *clique* is a subset of vertices that are fully connected, implying that the induced subgraph is complete (Lauritzen, 1996). In Figure 2.1 we show some examples of cliques, the light blue triangles are all *maximal cliques*, that is, a clique that cannot be extended including an adjacent vertex (because it will result in a non-complete subgraph); the two dark blue subgraphs are *maximum cliques*, that is, they are the biggest cliques in the graph. A *prime* component is a subgraph (or a graph) that does not admit a further decomposition. A subset $C$ is called a *separator* if it separate two subsets $A$ and $B$, i.e. every path from $A$ to $B$ has to go through $C$. An important aspect of graphs is their decomposability, which is also a property of major interest because it allows the factorization of the associated multivariate distribution. According to Lauritzen (1996), a triple of disjoint subsets $(A, B, C)$, of the vertex set $V$, generates a *weak decomposition* of the graph $G$, if $V = A \cup B \cup C$ and the two following conditions hold:

- $C$ separates $A$ from $B$
- $C$ is a complete subset of $V$

which is also a *strong decomposition* if all the nodes in the graph refer to distributions of the same nature (e.g. continuous distributions). Then, we can say that $(A, B, C)$ decomposes $G$ in two components, $G_{A \cup C}$ and $G_{B \cup C}$. For example, the graph in Figure 2.1 is non-decomposable because there are some cycles of more than three vertices that do not have a chord (a chord is an edge that is not part of the cycle, but connects two vertices of that cycle), and hence they are not decomposable. Indeed, every cycle made by four or more vertices that do not have a chord, is non-decomposable (see Lauritzen (1996) for the proof as well as further corollaries and definitions).

Now we introduce the Markov properties that apply to a probability distribution over a graph. Assume that we have a collection of random variables $(X_i)_{i \in V}$ taking values in the probability spaces $(\mathcal{X}_i)_{i \in V}$, meaning that there is a random variable associated with each element of the vertex set $V$. For a generic subset $A$ of $V$ we use the short notation $\mathcal{X}_A$ and $X_A$, with $x_A$ to denote an element of $\mathcal{X}_A$. Given an undirected graph $G = (V, E)$ and a correspondent collection of random variables $X_V$, for a probability measure $P$ on $\mathcal{X}_V$, Lauritzen (1996) defines the following three properties:

- Pairwise Markov property, if for any pair $(A, B)$ of non-adjacent vertices $A \perp\!\!\!\perp B \mid V \backslash \{A, B\}$
- Local Markov property, if for any vertex $A \in V$, $A \perp\!\!\!\perp V \backslash \mathrm{cl}(A) \mid \mathrm{bd}(A)$
- Global Markov property, if for any triple $(A, B, S)$ of disjoint subsets of $V$ such that $S$ separates $A$ from $B$, $A \perp\!\!\!\perp B \mid S$

where $A \perp\!\!\!\perp B$ means $A$ is independent of $B$, $\mathrm{bd}(A)$ denotes the boundary of $A$, i.e. the set of nodes that are neighbours to $A$, and $\mathrm{cl}(A)$ denotes the closure of $A$ which is the union between $A$ and its boundary. Various theorems about the relationships among these properties are discussed and proved in Lauritzen (1996), among which, the equivalence of the three Markov properties in a GGM. The global property is strictly connected to the concept of density factorization which is a fundamental aspect of

GGMs. Given a sample $x$ from a multivariate Gaussian distribution $X$, conditionally to a decomposable graph $G$, we can factorize the joint distribution as

$$p(X \mid \Omega, G) = \frac{\prod_{C \in \mathcal{C}} p(X_C \mid \Omega_C)}{\prod_{S \in \mathcal{S}} p(X_S \mid \Omega_S)} \tag{2.9}$$

where $\mathcal{C}$ and $\mathcal{S}$ denote the sets of cliques and separators, $\Omega_C$ and $\Omega_S$ represent the submatrices defined respectively by the clique $C$ and the separator $S$ (Giudici and Green, 1999). This decomposition gives an important gain in efficiency, as we do not need to conduct the calculations on the full joint distribution. This result has implications on the development of a conjugate family of distribution over the precision matrix of a multivariate Normal distribution, which includes the Hyper-Wishart distribution (Roverato, 2002) and the G-Wishart distribution (Atay-Kayis and Massam, 2005).

In the next chapter we introduce the SABRE metabolomics data and present an initial exploratory analysis and basic statistical models.

# Chapter 3

# Exploratory Data Analysis on SABRE

The complexity of the SABRE dataset and the large number of variables involved pose many challenges. We focus on the metabolites measurements, and try to infer some particular patterns in their relationship and look at the differences according to particular disease states or factors of interest. Moreover, we analyse the relation between the metabolites and the available covariates, for example, looking at the differences induced by a specific ethnicity. Also, we can infer the association between the metabolites and the development of diabetes, examining the different interactions for each ethnicity.

In this chapter, we restrict the analysis to a subset of the 88 metabolites, by merging together the lipoproteins sub-fractions. Instead of considering the finest distinction given by the compound content and dimension, we use aggregate measures for VLDL, IDL, LDL and HDL and we include total serum triglycerides and total serum cholesterol as extra metabolites. The decision to not analyse all the sub-fractions of each molecule of lipoproteins is motivated by the initial focus of the clinicians and epidemiologists, collaborating on this project, who are interested in understanding the overall dependencies between the major classes of lipoproteins and the other amino-acids. Moreover, given the complexity and high correlation of the data at hand we first perform an exploratory analysis of the metabolites on a reduced selection of variables. The results obtained with this first analysis give us additional evidence that a more detailed study of the lipoproteins sub-fractions would be needed, therefore in the following chapters we carry a more in depth analysis.

## 3.1 Network Analysis

The first analysis that we conduct is aimed to explore the space of metabolites and the interactions occurring among them. Individual Networks and Differential Networks are the techniques we use for this purpose (Valcárcel et al., 2011). Each analysis is conducted stratifying by ethnicity, in particular, we focus on the European and South-Asian sub-populations (the size of the African-Caribbean sample is not enough to allow a proper estimate of the networks). The next sub-sections show the results of the network analysis on the fasting dataset.

### Individual networks

Individual Networks are used to examine the patterns of association of a group of variables, exhibited under certain conditions. In our application, the two sub-samples, the Europeans and the South-Asians, have been divided into two subgroups according to the individual levels of Homeostasis Model Assessment Insulin Resistance (HOMA IR). Patients with a HOMA IR level under the first quantile (25% of HOMA IR cumulative distribution) and the patients with HOMA IR over the third quantile (75% of HOMA IR cumulative distribution). These values are calculated on the pooled sample of South-Asians and Europeans and are respectively 1.2 and 2.9. The construction of the network is based on a binary representation of the underlying partial correlations (significant and non-significant in a frequentist sense, according to a specified significance level). Instead of the standard partial correlation statistic, shrinkage methods for network estimation are adopted to decrease the number of connections in the graphs and avoid spurious associations. This implies a reduction in the network complexity and a corresponding more parsimonious biological interpretation (we refer to Valcárcel et al. (2011) for a similar application of this technique). The analysis is made with the `R` package *GeneNet*.

In Figure 3.1 we plot the two Individual Networks for the European ethnicity, originated according to the levels of HOMA IR. In Europeans, with both high and low levels of HOMA IR, the Individual Networks suggest, as expected, significant correlations between metabolites associated with the citric acid cycle (acetoacetate, glutamine, lactate and pyruvate), although the correlation with acetoacetate is lost in those with high HOMA IR, suggesting possible increase in energy supply from glycogenesis relative to ketogenesis and/or altered ketone metabolism. Europeans with low levels of HOMA IR have a well correlated group of "harmful"lipids, monounsaturated fatty acids (mufa), phosphatidylcholines and sphingomyelins- again this is a plausible grouping which corresponds to known pathways of lipid and phospholipid metabolism and which is less pronounced in those with high levels of HOMA IR, suggesting some level of dysregulation in this pathway in association with HOMA IR. We repeat the same procedure to estimate the Individual Networks in South-Asians. The results are reported in Figure 3.2 (note that some edges may overlap because we chose to fix the position of the nodes in order to better spot the differences between the different conditions and populations). South-Asians with low levels of HOMA IR have no significant metabolite correlations apart from that between beta-hydroxybutyrate and acetoacetate- both are ketone bodies. In contrast, South-Asians with high levels of HOMA IR have a large number of correlated metabolites and patterns of correlation appear remarkably similar to those seen in Europeans with low levels of HOMA IR. This is a surprising finding, which needs further study and points to ethnic differences in the mechanistic pathways underlying the development of Insulin Resistance.

### Differential Networks between low and high levels of HOMA IR

While with Individual Networks we can highlight the patterns of partial correlations for each individual level of HOMA IR, to find out which pairwise partial correlations are actually statistically different, we need to introduce another technique. Differential Networks can be used to find the main and more relevant differences in associations between low and high levels of HOMA IR. In our analysis we test the null hypothesis $H_0$:*The partial correlation between two metabolites across the two different sub-groups is the same* against the alternative hypothesis $H_1$:*The partial correlation between two metabolites across the two different sub-groups is different*, by means of a two-sample permutation test, as described in Valcárcel et al. (2011). Each connection in a Differential Network indicates a significant change in the partial correlation between two metabolite measures across the two conditions. The change in the structure can occur through either a significant increase or decrease in the pairwise partial correlation between two measures, or a significant alteration in the sign of partial correlations (here we choose to describe changes with respect to the low level of HOMA IR). Figures 3.3 and 3.4 show the estimated networks stratified by ethnicity, respectively for Europeans and South-Asians. In Europeans, compared

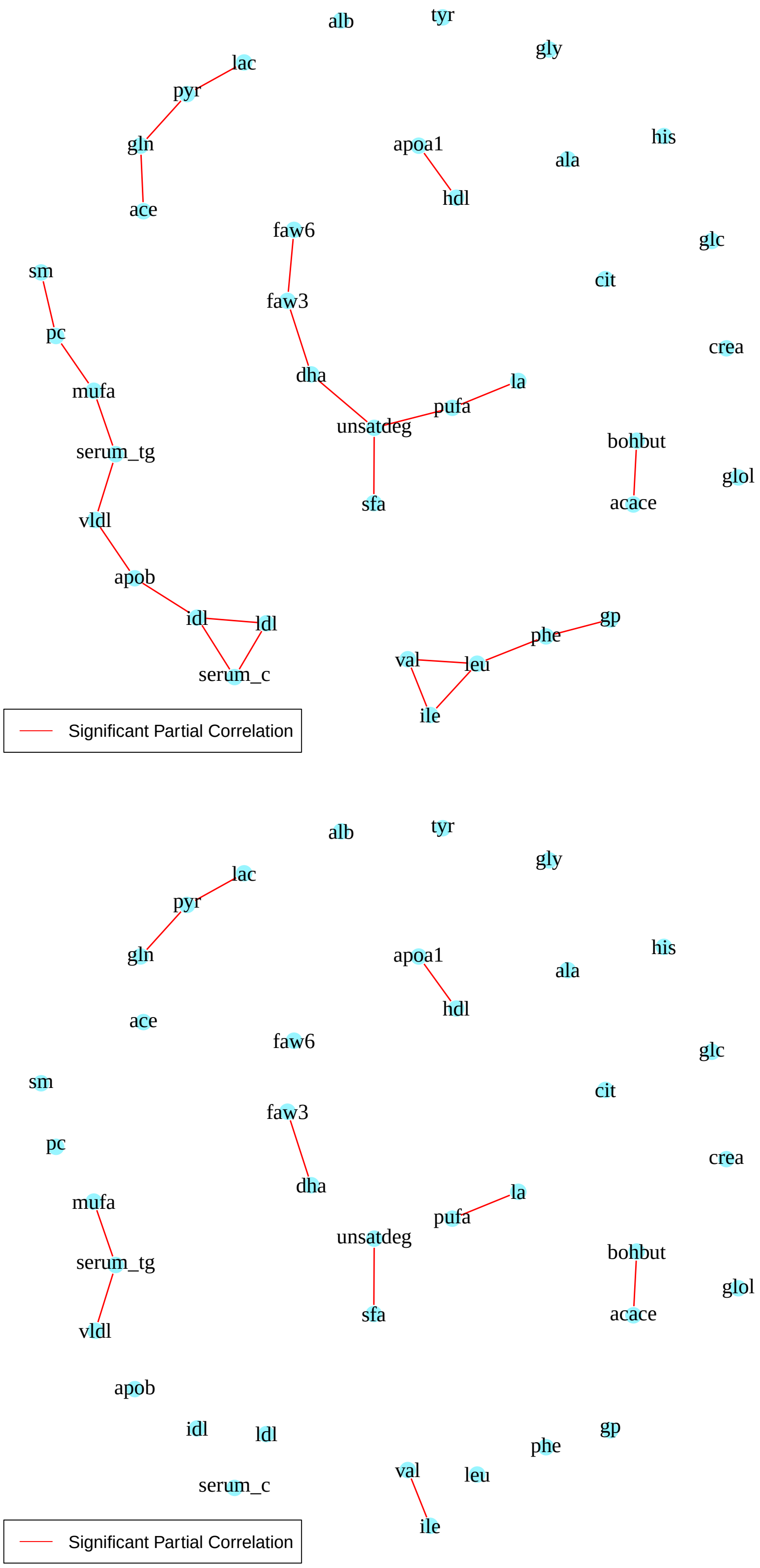


Figure 3.1: Individual Networks for Europeans. Top panel: low level of HOMA IR. Bottom Panel: High level of HOMA IR. Every edge included in the graph represents a partial correlation that is statistically significant at the 1% significance level.

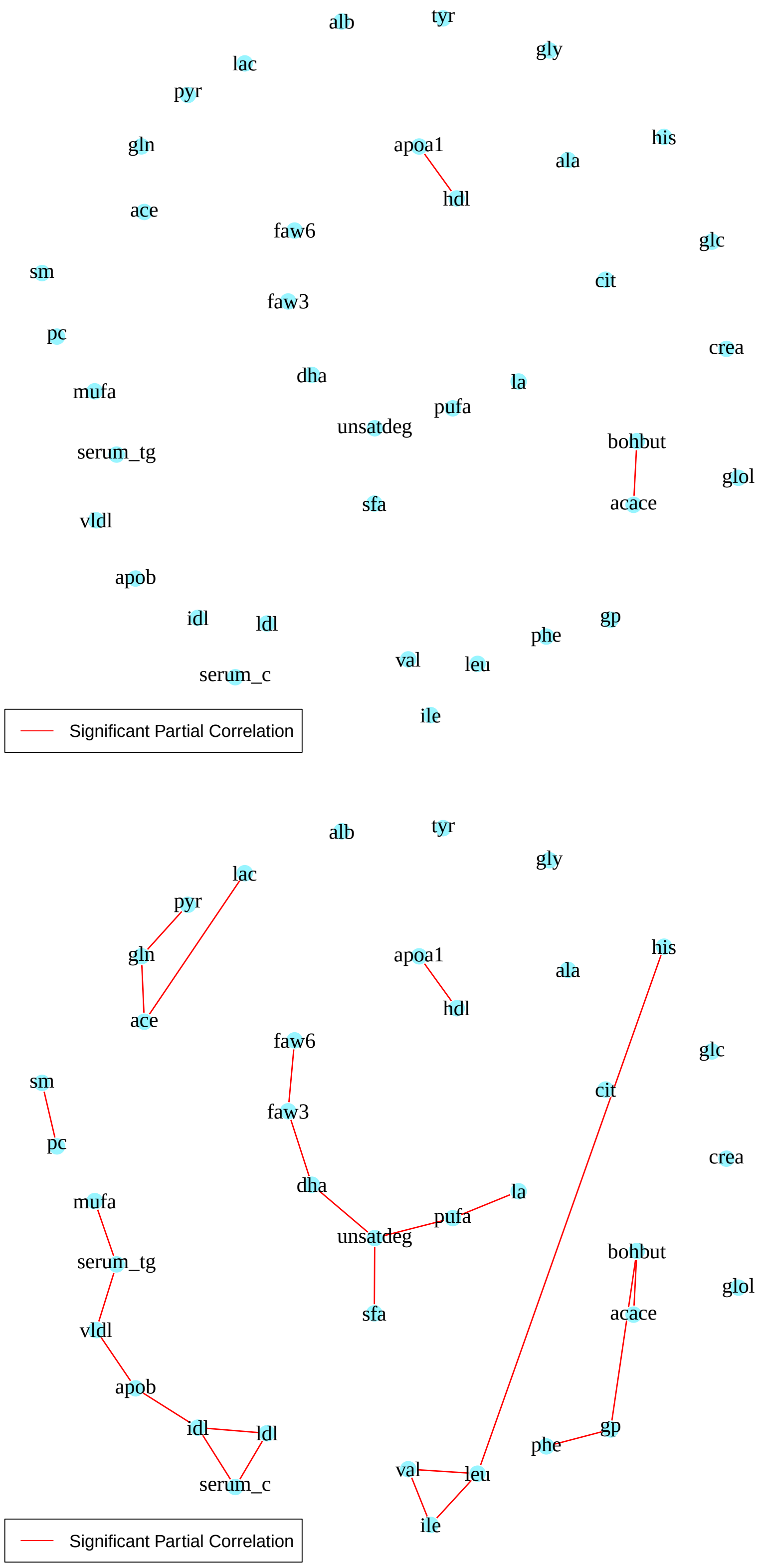


Figure 3.2: Individual Networks for South-Asians. Top panel: low level of HOMA IR. Bottom Panel: High level of HOMA IR. Every edge included in the graph represents a partial correlation that is statistically significant at the 1% significance level.

with those with low HOMA IR, individuals with high HOMA IR have different correlations for a number of metabolites –particularly noticeable for glycine (a glycogenic amino acid) where associations with histidine and citrate change from negative to positive and with acetoacetate from positive to negative. However, the correlation between acetoacetate and leucine (a ketogenic amino acid) also changes from positive to negative, likewise the correlation between tyrosine (both a ketogenic and glycogenic amino acid) and glucose. These differences perhaps reflect changes in the balance of amino acid glycogenesis and ketogenesis for energy provision in the fasting state in insulin resistant Europeans. In the insulin resistant state, phosphatidylcholines are positively correlated with omega 3 fatty acids and with total VLDL lipids, suggesting some alteration in structures of the phospholipids which make up the bulk of cell membranes, perhaps with implications for intracellular signalling processes.

In South-Asians, we see changes in many correlations on moving from low levels of HOMA IR to high levels. There are changes in correlations between glutamine and glucose (positive to negative) and between glutamine and pyruvate (stronger negative correlation). In contrast, the correlation between alanine and glucose changes from negative to positive in South-Asians with high HOMA IR. While alanine, like glutamine, is a glucogenic amino acid and abundant in muscle tissue, taken together, one interpretation could be that insulin resistant South-Asians may favour glutamine over alanine as a source of glucose in the fasting state (as glutamine breaks down, glucose levels increase). There is a change from positive to negative correlation between leucine, an essential branched chain and ketogenic amino acid, and LDL cholesterol suggesting an alteration in leucine's effect on lipid metabolism in insulin resistant South-Asians.

### Differential networks between Europeans and South-Asians

It is worth exploring the difference in the partial correlations between Europeans and South-Asians, adjusting the metabolites levels for the value of HOMA IR. We estimate two Individual Networks and one Differential Network, whose respective graphs are presented in Figures E.2 and Figure 3.5. There is a weaker negative correlation between glutamine and apoliprotein-B for South-Asians compared with Europeans, suggesting that glutamine may be less correlated with lipid profiles in South-Asians than in Europeans. There is some evidence to suggest a role for glutamine, not only as

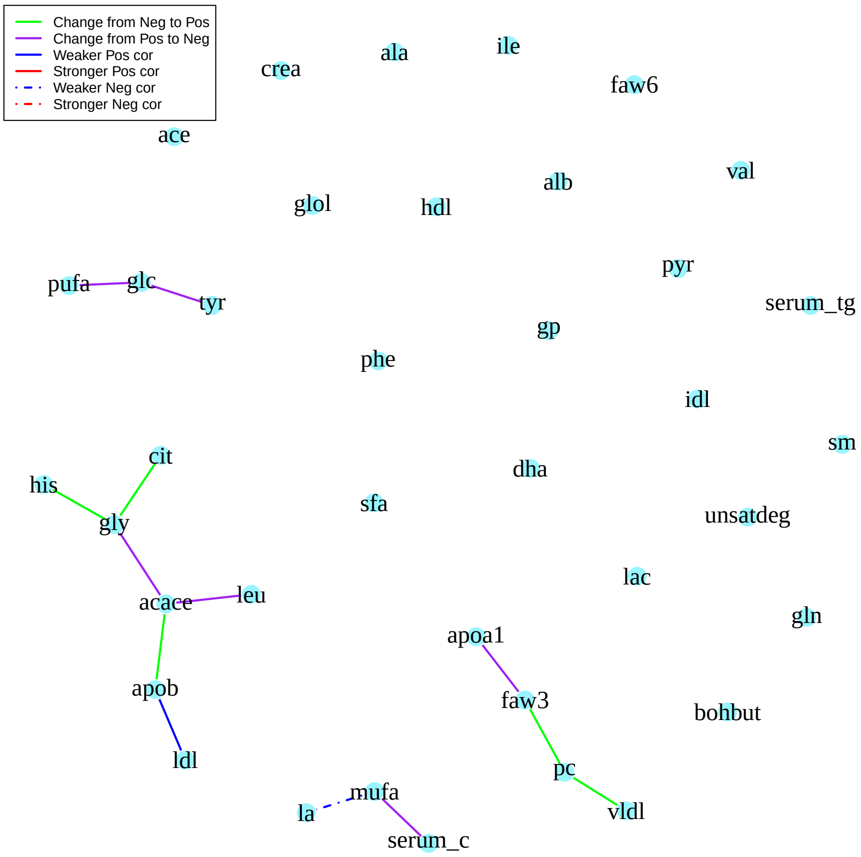


Figure 3.3: Differential Network for the European sub-population. Fasting dataset. The graph shows information about the sign of the partial correlations and whether there is a change in the sign going from a low level of HOMA IR to a high level of HOMA IR. Every edge included in the graph represents a partial correlation that is statistically significant at the 1% significance level.

a regulator of pancreatic beta cell function, but also as a substrate for lipid synthesis in adipose tissue and in fuelling inflammation – the latter could contribute to another potential pathway to explain excess HOMA IR in South-Asians. The change from negative to positive correlation between tyrosine and DHA in South-Asians is harder to explain, both are commonly found in high levels in fish, so there may be some dietary component. Both tyrosine and DHA can be used in the synthesis of dopamine, hence it is possible that there are some ethnic differences in dopamine metabolism and that its ensuing adrenergic effects might be connected with increased risk of HOMA IR in South-Asians. Changes in correlations between branched chain and aromatic amino acid correlations with glucose in South-Asians, although difficult to interpret, broadly support previously observed disturbances in metabolism involving these amino acids

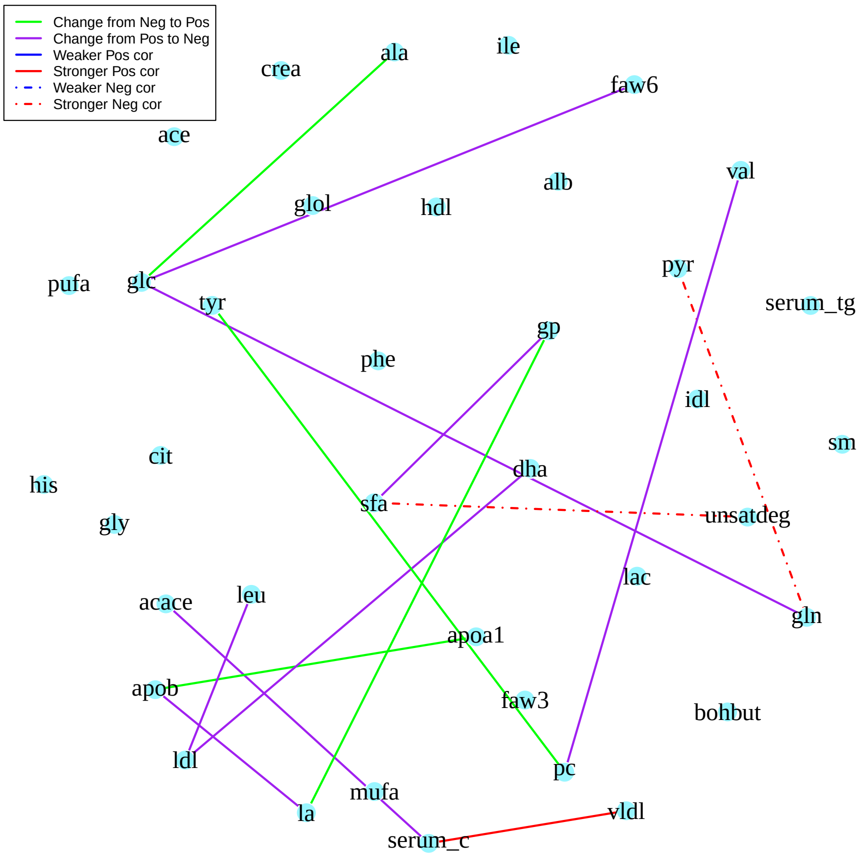


Figure 3.4: Differential Network for the South-Asian sub-population. Fasting dataset. The graph shows information about the sign of the partial correlations and whether there is a change in the sign going from a low level of HOMA IR to a high level of HOMA IR. Every edge included in the graph represents a partial correlation that is statistically significant at the 1% significance level.

in association with the development of insulin resistance and diabetes, both states being much more frequent in South Asian populations.

These associations are all cross-sectional in nature, so no causal inferences may be drawn and validation in other studies is required. However, the findings highlight areas for further research and also suggest areas of focus for further analyses using a more advanced platform for measurement of many thousands of metabolites. Throughout the previous analysis we used a cut-off on HOMA IR which is not entirely objective. In literature there is not a clear and unique value to distinct between individuals with *high* and *low* level of insulin resistance. The value we chose to discriminate the *higher* values of HOMA IR corresponds to the third quartile of the distribution, which correspond to 2.882. This threshold is not without foundation in the literature,

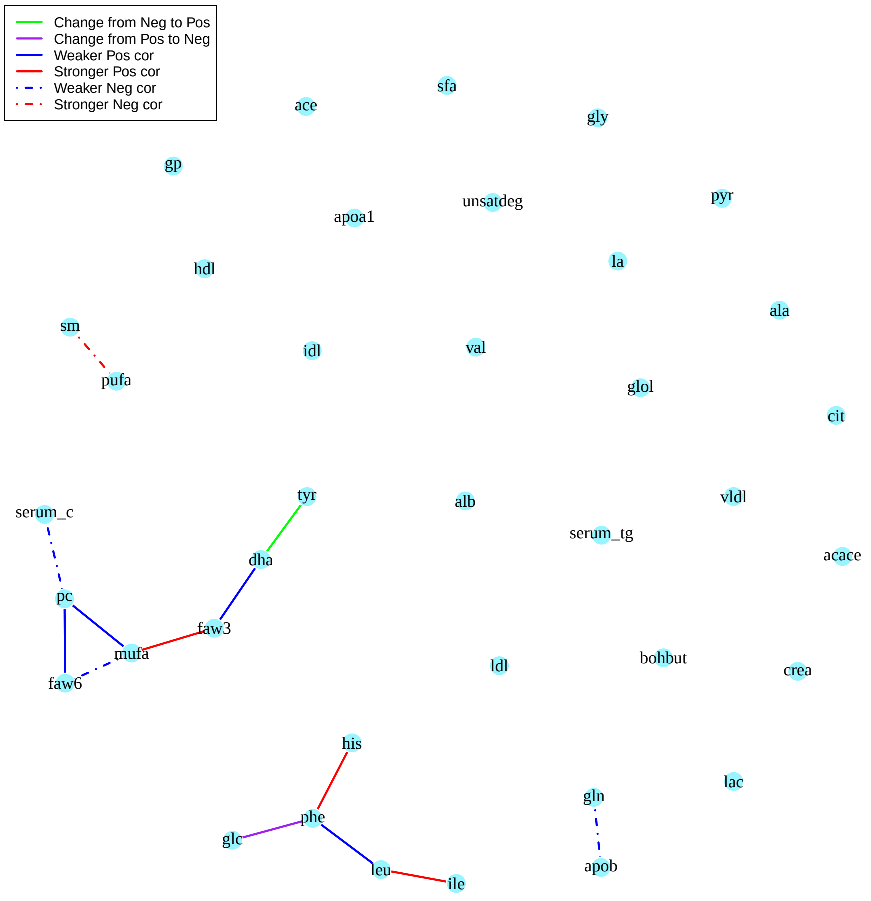


Figure 3.5: Differential Network between Europeans and South-Asians. The graph shows information about the sign of the partial correlations and whether there is a change in the sign comparing Europeans to South-Asians. Every edge included in the graph represents a partial correlation that is statistically significant at the 1% significance level.

because very similar values have been proposed by other authors, an example of which is Tam et al. (2012). In the next section we try to determine a cut-off which can be justified from a medical and statistical point of view. For this reason we build a Bayesian model to infer the value from the data, rather than fix it a priori.

## 3.2 HOMA IR cut-off selection

We estimate the ideal cut-off for a continuous variable, here HOMA IR, with a change-point model defined as follows. Given $M$ metabolites, where each metabolite represents a random variable, we assume the following model

$$\begin{aligned} Y &\sim \mathrm{N}\left(\mathbf{0}, \Sigma\right) \\ \Sigma &\sim \text{Inverse-Wishart}\left(\nu_0, S_0\right) \end{aligned}$$

where $Y$ represents the $n \times M$ matrix of observations, centred to have zero mean. Each row $\boldsymbol{y}_i$, for $i = 1, \ldots, n$, is an $M$-dimensional vector corresponding to the metabolite measurements for individual $i$. The prior over $\Sigma$ is an Inverse-Wishart random variable with the following probability density function

$$p(\Sigma \mid \nu_0, S_0) = \left[ 2^{\frac{\nu_0 M}{2}} \pi^{\frac{\binom{M}{2}}{2}} |S_0|^{-\frac{\nu_0}{2}} \prod_{j=1}^{M} \Gamma\left(\frac{\nu_0 + 1 - j}{2}\right) \right]^{-1} |\Sigma|^{-\frac{\nu_0 + M + 1}{2}} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(S_0 \Sigma^{-1}\right) \right\}$$

with expected value

$$\mathrm{E}\left[\Sigma\right] = \frac{S_0}{\nu_0 - M - 1}$$

where $\nu_0$ and $S_0$ are respectively the prior number of degrees of freedom and the prior base measure of the Inverse-Wishart distribution and $\mathrm{tr}(A)$ and $|A|$ represent the trace and the determinant of a matrix A respectively. The parameter of interest in this application is the covariance matrix $\Sigma$, whose inverse, defined as $\Omega = \Sigma^{-1}$, contains the concentrations, which imply conditional independences when equal to zero.

We search the ideal cut-off over a fine grid of values of HOMA IR and decide whether the proposed value produces a statistically acceptable split of the dataset by evaluating the ratio of marginal likelihoods. In particular, we assume that, conditioning on the proposed splitting value of HOMA IR, the full marginal likelihood can be factorised in the sum of the marginal likelihood of the two identified sub-samples.

Given the model likelihood

$$\begin{aligned} p\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \mid \Sigma\right) &= \prod_{i=1}^{n} (2\pi)^{-\frac{M}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \boldsymbol{y}_i^T \Sigma^{-1} \boldsymbol{y}_i \right\} \\ &= (2\pi)^{-\frac{nM}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \boldsymbol{y}_i^T \Sigma^{-1} \boldsymbol{y}_i \right\} \qquad (3.1) \\ &= (2\pi)^{-\frac{nM}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(S_y \Sigma^{-1}\right) \right\} \end{aligned}$$

where $S_y = \sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^T$, the marginal likelihood is given by

$$\begin{aligned}
p(Y) &= \int_{\Sigma} p(Y \mid \Sigma) p(\Sigma) \; d\Sigma \\
&= (2\pi)^{-\frac{nM}{2}} \left[ 2^{\frac{\nu_0 M}{2}} \pi^{\frac{\binom{M}{2}}{2}} |S_0|^{-\frac{\nu_0}{2}} \prod_{j=1}^{M} \Gamma\left( \frac{\nu_0 + 1 - j}{2} \right) \right]^{-1} \times \\
&\qquad \int_{\Sigma} |\Sigma|^{-\frac{n+\nu_0+M+1}{2}} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left[ (S_y + S_0)\, \Sigma^{-1} \right] \right\} \; d\Sigma \\
&= (2\pi)^{-\frac{nM}{2}} \left[ 2^{\frac{\nu_0 M}{2}} \pi^{\frac{\binom{M}{2}}{2}} |S_0|^{-\frac{\nu_0}{2}} \prod_{j=1}^{M} \Gamma\left( \frac{\nu_0 + 1 - j}{2} \right) \right]^{-1} \times \\
&\qquad \left[ 2^{\frac{(\nu_0+n)M}{2}} \pi^{\frac{\binom{M}{2}}{2}} |S_0 + S_y|^{-\frac{\nu_0+n}{2}} \prod_{j=1}^{M} \Gamma\left( \frac{\nu_0 + n + 1 - j}{2} \right) \right]
\end{aligned}$$

In passage two, the argument of the integral represents the kernel of the posterior Inverse-Wishart distribution with parameters $S_n = S_0 + S_y$ and $\nu_n = \nu_0 + n$. Therefore, the marginal likelihood reduces to the ratio between the posterior and prior normalizing constant of the Inverse-Wishart distribution, plus a residual term from the Normal likelihood.

Next, let indicate with $x$ the variable HOMA IR, then given a value $x_s$ of $x$, we can divide the full dataset $Y$ in $Y_{x<x_s}$ and $Y_{x \geq x_s}$, where $Y_{x<x_s}$ indicates the set of observations whose corresponding level of HOMA IR is lower than $x_s$ and the set of observations $Y_{x \geq x_s}$, whose corresponding level of HOMA IR is higher (or equal) than $x_s$. Conditional on the proposed value $x_s$, the marginal likelihood at iteration $s$, $p(Y)^{(s)}$, can be decomposed in the sum of two independent components, $p(Y)^{(s)} = p(Y_{x<x_s})^{(s)} + p(Y_{x \geq x_s})^{(s)}$, which are the marginal likelihood evaluated on the first sub-sample and on the second sub-sample respectively. The cut-off search proceeds iteratively as follows, at iteration $s$:

1. Propose a value $x_s$
2. Given $x_s$, divide the data $Y$ in $Y_{x<x_s}$ and $Y_{x \geq x_s}$
3. Evaluate the marginal likelihood ratio $\alpha = p(Y)^{(s)}/p(Y)^{(s-1)}$ and accept the new value of $x_s$ with probability equal to $\min(\alpha, 1)$

Ideally, we want to divide the dataset in such a way that one group of patients can be considered relatively healthy and the other group potentially less healthy. However, from the analysis of the metabolites we find that the algorithm converges to a threshold that is always located on the right extreme tail of the distribution of HOMA

IR, leading to a split of the dataset into a group with just a few individuals and a group with all the other patients. This is somehow expected, because the individuals with very high levels of HOMA IR experience a down-break of many connections at a metabolic level, but the groups identified are not useful for a proper network analysis. Even if the model works well in simulated settings, the structure of this real dataset is too complex to be examined in this way. For a pure exploratory purpose we examine the pairwise correlations for each couple of metabolites on a grid of ten values of HOMA IR. So, we equally divide the distribution of HOMA IR with the deciles of the distribution, and for each corresponding slice of Y we calculate all the pairwise correlations. The idea is to look at the correlation patterns and identify some recurring behaviours over the different metabolites. For example, a rapid change in the correlation for a certain decile, either an increase or a decrease, that can be observed for many metabolites. However, due to the variety of different measures involved, we can not find a clear pattern, nor a clear break point, except for a value of HOMA IR, around 1.7, where almost overall the pairwise correlations have a slight decrease.

## 3.3 Bayesian Variable Selection

In the previous section we have adjusted the metabolites levels for the set of covariates in Table D.2 before performing the analysis. Before introducing the more general Sparse SUR model, we can obtain some information about the covariates that are more important in predicting metabolites levels. We have a total of 684 regression coefficients, i.e. 18 for each of the 38 metabolites in the restricted selection, therefore, it is useful to reduce this number in order to have a clearer scenario of which factors are more associated with each metabolite.

Every metabolite represents the response variable of a regression model. We have $M$ independent linear regressions, with the same regressor matrix of dimension $n \times p$. The algorithm that we adopt to perform Bayesian variable selection is the MC$^3$ algorithm of Madigan et al. (1995), applied to each regression. The model is specified as follows:

$$\begin{aligned} \boldsymbol{y}_m \mid \boldsymbol{\beta}_m, \tau_m^2 &\sim \mathrm{N}\left(X\boldsymbol{\beta}_m, \tau_m^2\right) \\ \boldsymbol{\beta}_m \mid \boldsymbol{\beta}_0, \Omega_0, \tau_m^2 &\sim \mathrm{N}\left(\boldsymbol{\beta}_0, \Omega_0 \tau_m^2\right) \\ \tau_m^2 \mid \nu_0, s_0 &\sim \mathrm{Gamma}\left(\nu_0, s_0\right) \end{aligned} \tag{3.2}$$

for $m$ in $1, \dots, M$. $\boldsymbol{\beta}_0$ represents the prior mean of the regression coefficients and $\Omega_0$ represents the prior precision matrix. We use a semi-conjugate model, to be able to derive the marginal likelihood and use the MC$^3$ algorithm. We drop the subscript $m$ for ease of notation, noting that the same result applies to every regression. The likelihood of the model in Equation 3.2 is

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \tau^2) = (2\pi)^{-\frac{n}{2}} (\tau^2)^{\frac{n}{2}} \exp\left\{-\frac{\tau^2}{2} (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta})\right\}$$

and the prior distributions over $\beta$ and $\tau^2$ are respectively

$$\begin{aligned} p\left(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \Omega_0, \tau^2\right) &= (2\pi)^{-\frac{1}{2}} |\Omega_0 \tau^2|^{\frac{1}{2}} \exp\left\{-\frac{\tau^2}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \Omega_0 (\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} \\ p(\tau^2 \mid \nu_0, s_0) &= \frac{s_0^{\nu_0}}{\Gamma(\nu_0)} (\tau^2)^{\nu_0 - 1} \exp\left\{-s_0 \tau^2\right\} \end{aligned}$$

The marginal likelihood of the model is then

$$\begin{aligned} p(\boldsymbol{y}) &= \int_{\tau^2} \int_{\boldsymbol{\beta}} p(\boldsymbol{y} \mid \boldsymbol{\beta}, \tau^2) \times p\left(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \Omega_0 \tau^2\right) \times p\left(\tau^2 \mid \nu_0, s_0\right) \; d\boldsymbol{\beta} \; d\tau^2 \\ &= (2\pi)^{-\frac{n}{2}} |\Omega_0|^{\frac{1}{2}} |\Omega_n|^{-\frac{1}{2}} \frac{s_0^{\nu_0}}{\Gamma(\nu_0)} \int_{\tau^2} (\tau^2)^{\nu_0 - 1 + \frac{n}{2}} \exp\left\{-\tau^2 \left[\frac{1}{2}\boldsymbol{m}_0 - \frac{1}{2}\boldsymbol{\beta}_n \Omega_n \boldsymbol{\beta}_n^T + s_0\right]\right\} \; d\tau^2 \\ &= (2\pi)^{-\frac{n}{2}} |\Omega_0|^{\frac{1}{2}} |\Omega_n|^{-\frac{1}{2}} \frac{s_0^{\nu_0}}{\Gamma(\nu_0)} \left[\frac{s_n^{\nu_n}}{\Gamma(\nu_n)}\right]^{-1} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{m}_0 &= \boldsymbol{y}^T \boldsymbol{y} + \boldsymbol{\beta}_0 \Omega_0 \boldsymbol{\beta}_0 \\ \Omega_n &= \Omega_0 + X^T X \\ \boldsymbol{\beta}_n &= \Omega_n^{-1} \left(\boldsymbol{y}^T X + \boldsymbol{\beta}_0 \Omega_0\right)^T \\ \nu_n &= \nu_0 + \frac{n}{2} \\ s_n &= \frac{1}{2}\boldsymbol{m}_0 - \frac{1}{2}\boldsymbol{\beta}_n \Omega_n \boldsymbol{\beta}_n^T + s_0 \end{aligned}$$

The MC$^3$ algorithm proceeds iteratively proposing the inclusion of a new regressor or the removal of an old one. The new model is evaluated through a Metropolis step, where the probability of acceptance is the minimum between one and the ratio of the marginal likelihoods of the two models. The result of the variable selection is summarized in Table D.3. The approach used in this case does not take into account the possible influence that can arise across the regressions, but still can give an idea about the most important covariates. WHR, age and sex are the control variables that are selected more often, in particular WHR is the most important measure of body fat distribution. The ethnic group of origin is also important, with both the indicators for South-Asians and Africans-Caribbean selected for many metabolites, denoting the importance of the ethnic factor. HOMA IR is also quite important, together with the liver health indicators ALT and AST.

This exploratory analysis concludes the first part of the thesis. In the following chapters we present our major contributions, in particular in the next chapter we provide a formal analysis of the differences in the distribution of HOMA IR according to the ethnic group of origin, accounting for the individual metabolic profiles.

# Chapter 4

# Bayesian Nonparametric Modelling of Insulin Resistance

The causal mechanisms underlying the development of type 2 diabetes remain poorly understood, and no study has yet conclusively explained the reasons for the excess risk of diabetes experienced by South-Asian and African-Caribbean populations, suggesting that complex metabolic disturbances may underlie the ethnic differences (Tillin et al., 2012). Insulin resistance is a frequent precursors of type 2 diabetes in all populations and can be measured non-invasively using indices such as HOMA IR, which can be calculated from fasting blood glucose and insulin levels (Matthews et al., 1985). The main purpose of this work is to explore potential mechanisms underlying the marked ethnic differences in insulin resistance (Figure 1.2).

By employing Bayesian nonparametric statistical methods, we cluster individuals based on their HOMA IR levels. In doing so, we are able to account for the effect of covariates, in our case anthropometric measures and metabolites concentrations, and identify the most influential variables. Moreover, we are able to asses if the distribution of the selected covariates vary between clusters and if the clusters ethnic composition has an effect on the covariate distribution. We allow for clusters of individuals belonging to different ethnic groups. The full list of metabolites included in the analysis is available in Table D.1. We include three important enzymes: alanine aminotransferase, aspartate aminotransferase (that are liver health indicators) and gamma glutamyl transferase. Anthropometric variables are also included, in particular global measures of body fat distribution such as waist to hip ratio (WHR) and more specific adiposity measures, such as sagittal diameter and subscapular skinfold thickness. The list of anthropometric and clinic covariates can be found in Table D.2. We exclude from the analysis individuals with known diabetes since they were already receiving anti-diabetes medication or had undergone lifestyle modifications that might alter their metabolite levels and potentially the conclusions of the analysis. In this paper, we focus on the SABRE study baseline metabolic and phenotypic dataset. To address our research aims, we use a Bayesian nonparametric prior, the Dependent Generalized Dirichlet Process (DGDP, Barcella et al. (2017)) within a regression framework. The discrete nature of the DGDP allows for data-driven clustering of the observations. We specify the DGDP prior on the regression intercept and the error precision parameter, allowing for cluster specific locations and precisions. The choice of the DGDP allows a great flexibility, accounts for inter-subject variability and it does not fix a priori the number of clusters. When prior evidence is available, through the calibration of the DGDP hyper-parameters, we can favour a large number of clusters, allowing estimation of more heterogeneous groups. Moreover, to deal with the large number of clinical and anthropometric covariates and metabolites available, we adopt a Spike and Slab approach (George and McCulloch, 1993; George and McCulloch, 1997) in order to perform variable selection on the design matrix and highlight the most important determinant of the clinical outcome under study.

## 4.1 The model

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a continuous response variable observed over $n$ individuals. We assume a linear regression model:

$$y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \varepsilon_i \tag{4.1}$$

where $p$ is the number of independent variables (including the intercept). The error terms $\varepsilon_i$ are assumed to be normally distributed as

$$\varepsilon_i \overset{iid}{\sim} \mathrm{N}\left(0, \tau^2\right)$$

with mean 0 and precision $\tau^2$. The model in (4.1) assumes the same parameters for each observation. This assumption can be relaxed by allowing, for example, $\beta_0$ and $\tau^2$ to vary with $i$ (random effect model), accounting for inter-subject variability:

$$y_i = \beta_{i0} + \sum_{j=1}^{p-1} \beta_j x_{ij} + \varepsilon_i \tag{4.2}$$

where

$$\varepsilon_i \mid \tau_i^2 \overset{ind}{\sim} \mathrm{N}(0, \tau_i^2)$$

In this way, a subject-specific intercept and precision are introduced in the model, allowing for more flexibility. We now need to specify a prior on the model parameters. In particular, we need to choose a random effect distribution for $(\beta_{0i}, \tau_i^2)$. A traditional and computationally convenient choice is a Normal random effects model for $\beta_{i0}$ and a Gamma distribution for $\tau_i^2$. Instead, we opt for a nonparametric random effects distribution as, often, the parametric assumptions are too restrictive in applications. The random effects distribution needs to accommodate the heterogeneity in the population and to allow for outliers, clustering and over-dispersion. At the same time, the model should not be overly complex and should still allow computationally efficient implementation of full posterior inference. Ideally the model should be a natural generalization of a traditional random effects distribution. In the next section we describe our choice of prior distributions.

### 4.1.1 Prior distributions

The model in (4.2) requires the specification of prior distributions for the vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p-1})$, the intercept $\beta_{i0}$ and the precision parameter $\tau_i^2$. We adopt a nonparametric prior, the DGDP prior, on both the intercept and precision term $(\beta_{i0}, \tau_i^2)$. As explained below, this choice of prior distribution allows to cluster the observations. Moreover the use of the DGDP prior provides both flexibility and parsimony about the number of parameters that we introduce in the model. We now introduce the Generalised Dirichlet Process (GDP) and the extension to the DGDP.

Consider the DP described in Section 2.1.4. The discreteness of the nonparametric process induces clustering of the subjects in the sample based on the unique values of the random effects parameters (in our case $\theta_k = (\beta_{0k}, \tau_k^2)$), where the number $K$ of clusters is unknown and learned from the data. In this paper we are interested in modelling the distribution of HOMA IR in each of the three ethnic groups (i) allowing for borrowing information across groups (ii) highlighting differences and similarities (iii) accounting for the effect of covariates. To this end, we employ a generalisation of the DP proposed by Hjort (2000) and Ishwaran and James (2001), the Generalised Dirichlet Process (GDP). The GDP employees a richer parametrisation in the stick-breaking construction, allowing greater flexibility in the moments of the random distributions. Consider the stick-breaking in (2.6), where the elements $\phi_k$ are draws from a Beta$(1, \alpha)$. In the generalisation proposed by Hjort (2000) the $\phi_k$ are still draws from a Beta distribution, but the first hyper-parameter does not need to be fixed to one. In what follows we use an alternative parametrisation of the Beta distribution, where the hyper-parameters are specified in terms of the mean and the concentration parameter. In the GDP the $\{\phi_k\}$ are then independent draws from a Beta$(\mu_k \upsilon_k, (1 - \mu_k)\upsilon_k)$:

$$p(\phi_k \mid \upsilon_k, \mu_k) = \frac{\Gamma(\upsilon_k)}{\Gamma(\upsilon_k \mu_k)\Gamma(\upsilon_k(1-\mu_k))} \phi_k^{\upsilon_k \mu_k - 1}(1-\phi_k)^{\upsilon_k(1-\mu_k)-1} \tag{4.3}$$

where $\mathrm{E}[\phi_k] = \mu_k \in (0, 1)$ and $\mathrm{Var}[\phi_k] = \mu_k(1 - \mu_k)/(1 + \upsilon_k)$, with $\upsilon_k \in (0, \infty)$, are the expected value and the variance of the Beta random variable respectively. The weights of the GDP admit the same stick-breaking construction as for the DP. Hjort (2000) proposes a more parsimonious parametrisation of (4.3), setting $\mu_k = \mu$ and $\upsilon_k = \upsilon$. This simplification does not impose significant restriction in applications. We now explain how we introduce ethnicity information in the distribution of HOMA IR. The final model will contain two main components: one for the clinical covariates and one for the patients effect. The model for the covariates expresses prior information on how covariates influence the clinical outcome, while the nonparametric prior (GDP) is used as a random effect distribution to capture inter-patients variability. Moreover, it is desirable to specify a random effect distribution for each ethnicity in a way that the random effect distributions are related (similar or very different), but not necessarily identical. There is a wealth of literature on how to extend the DP to incorporate covariate information, for example, letting the weights and/or locations of the infinite mixture in (2.5) depend on a variable of interest that defines sub-groups in the observations. See the seminal paper of MacEachern (1999) on the Dependent Dirichlet Process (DDP). Similarly, also the GDP can be extended in presence of categorical covariates. Barcella et al. (2017) introduce the Dependent Generalised Dirichlet Process (DGDP) where the dependence among random distributions is introduced through the weights $\psi_k$ of the mixture in (2.5). The parameters $\psi_k$ are generated from the stick-breaking process, so the dependence is introduced directly on the parameters $\upsilon$ and $\mu$.

Consider $G$ groups defined by a covariate of interest $g \in \mathcal{G}$, where $\mathcal{G}$ is the covariate space. We let $\mu$, which represents the mean of the Beta random variables, depend on the particular value of $g$, while we assume the same $\upsilon$ across groups. We denote with $\mu_g$ the mean of the Beta random variable corresponding to group $g$. In our application groups are defined by the ethnicity. The random measure $P_g$, i.e. the random distribution associated to group $g$, is then defined as:

$$P_g = \sum_{k=1}^{\infty} \psi_{k,g} \delta_{\theta_k}$$

Here $\theta_k = (\beta_{0k}, \tau_k^2)$. In particular, dependence across the $\mu_g, g = 1, \ldots, G$, is obtained by specifying a Beta regression on $\mu_g$ and using a categorical predictor $\mathbf{z}_g$, i.e. an indicator variable which denotes to which group the observations are associated to. This strategy allows for group dependent clustering of the observations. See Barcella et al. (2017) for details and clustering properties.

Finally, the full model for HOMA IR is specified as follows

$$\begin{aligned} y_{1g}, \ldots, y_{ng} \mid g, P_g &\overset{ind}{\sim} \int \mathrm{N}\left(\beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j, \tau^2\right) P_g(d\beta_0 d\tau^2) \\ P_1, \ldots, P_G \mid \upsilon, \mu_g, G_0 &\sim \mathrm{DGDP}(\upsilon, \mu_g, G_0) \\ G_0(\beta_0, \tau^2) \mid m_0, \kappa_0^2, \tau_a, \tau_b &= \mathrm{N}\left(m_0, \kappa_0^2\right) \times \mathrm{Gamma}(\tau_a, \tau_b) \\ f(\mu_g) &= \mathbf{z}_g \boldsymbol{\eta} \\ \upsilon \mid a_\upsilon, b_\upsilon &\sim \mathrm{Gamma}(a_\upsilon, b_\upsilon) \end{aligned}$$

for $g = 1, \ldots, G$. The parameter $\mu$ is linked through a function $f$ (e.g. logit or probit), mapping from $(0, 1)$ into $(-\infty, \infty)$, to the linear predictor $\mathbf{z}_g\boldsymbol{\eta}$ where $\boldsymbol{\eta}$ is a vector of regression coefficients of appropriate dimension to which we assign a standard Normal prior:

$$\boldsymbol{\eta} \mid \boldsymbol{\eta}_\mu, \boldsymbol{\eta}_\Sigma \sim \mathrm{N}(\boldsymbol{\eta}_\mu, \boldsymbol{\eta}_\Sigma)$$

where $\boldsymbol{\eta}_\mu$ and $\boldsymbol{\eta}_\Sigma$ denote the prior mean and covariance matrix respectively. We assume independence a priori between the parameters $\beta_0$ and $\tau^2$, which is reflected in the choice of the base measure $G_0$, defined as the product of a Normal distribution and a Gamma distribution.

We specify a Spike and Slab prior on each of the $p - 1$ regression coefficients $\beta_j$. This prior specification provides an effective variable selection strategy (George and McCulloch, 1993; Malsiner-Walli and Wagner, 2018). We introduce indicator variables $\omega_j$:

$$\beta_j = \omega_j \mathrm{N}\left(\mu_\beta, \tau_\beta^2\right) + (1 - \omega_j)\delta_0(\beta_j)$$

where $p(\omega_j = 1|\pi) = \pi$ is the probability of the slab, i.e. the probability that a covariate is included in the model, while $1 - \pi$ represents the probability of the spike, i.e. the probability that the regression coefficient corresponding to the j-th covariate is equal to 0 and does not affect the response. As before, $\delta_0(\beta_j)$ is a mass point at zero, representing the spike of the mixture. $\mu_\beta$ represents the prior mean (usually set to 0) of the slab component and $\tau_\beta^2$ is the prior precision. The parameter $\pi$ is assigned a Beta prior:

$$\pi \sim \text{Beta}(\pi_a, \pi_b)$$

where $\pi_a$ and $\pi_b$ are the hyper-parameters of the Beta distribution (e.g. setting $\pi_a = \pi_b = 1$ gives a uniform distribution). Appropriate choices of these hyper-parameter allows us to impose sparsity in the variable selection.

### 4.1.2 Posterior Inference

Posterior inference is performed through Markov Chain Monte Carlo (MCMC) methods. A detailed description of the algorithm is provided in Appendix A. We run the MCMC for 30000 iterations, discarding a burn-in period of 15000, thinning every 5 iterations. We specify the following hyper-parameters: the truncation level of the stick-breaking is set to $L = 30$. The base measure parameters are set to $m_0 = 0$, $\kappa_0^2 = 0.1$, $\tau_a = 0.5$, $\tau_b = 0.5$. The DGDP concentration parameter $\upsilon$ has a Gamma prior with $a_\upsilon = 2$ and $a_\upsilon = 1$, with expected value $a_\upsilon/b_\upsilon = 2$. The regression coefficient $\boldsymbol{\eta}$ is parametrised with prior mean $\boldsymbol{\eta}_\mu = (0, 0, 0)$ and $\boldsymbol{\eta}_\Sigma = I_3$. The slab of the regression coefficient $\beta_j$ is a Normal distribution with prior mean $\mu_\beta = 0$ and prior precision $\tau_\beta^2 = 0.1$. The prior inclusion probability $\pi$ has a Beta prior with parameters $\pi_a = \pi_b = 1$. Details on the MCMC posterior updates can be found in Appendix A. We have employed the Binder loss function, as implemented in the `R` package *mcclust*, to provide a posterior estimate of the clustering allocation and overcome the problem of label switching that affects Bayesian mixture models. There is no consensus in the literature on how to report clustering output, but it is very common in Bayesian nonparametric models to report the clustering allocation that minimises the Binder loss function which assigns equal costs to each type of misclassification error. See Lau and Green, 2007 for a discussion.

## 4.2 Results on HOMA IR: Cluster Analysis

We employ the proposed model to analysis data from the SABRE study. The empirical distribution of the outcome of interest, the Homeostatic Model Assessment insulin resistance (HOMA IR) is shown for the three ethnic groups in Figure 1.2. Particularly noticeable is the difference between the distribution of HOMA IR in Europeans and South-Asians. The South-Asian distribution is slightly shifted to the right and has a heavier right tail, indicating a higher percentage of more insulin resistant individuals. The distribution shows multiple local modes, pointing towards the existence of multiple sub-populations in the sample.

Posterior inference for HOMA IR shows evidence of ten clusters. We use the Binder loss function (Binder, 1978) available in the `R` package *mcclust*, to estimate the number of clusters in the sample and the clustering allocation based on the MCMC output. In Figure 4.1 we show the number of active clusters at each iteration of the MCMC. From the plot we can see that the number of non-empty clusters is concentrated between 8 and 12. This number is in accord with the number of clusters estimated by the minimisation of *Binder* loss function. In Figure 4.2 we show the empirical distribution of the outcome HOMA IR in each of the ten estimated clusters. The overlap between

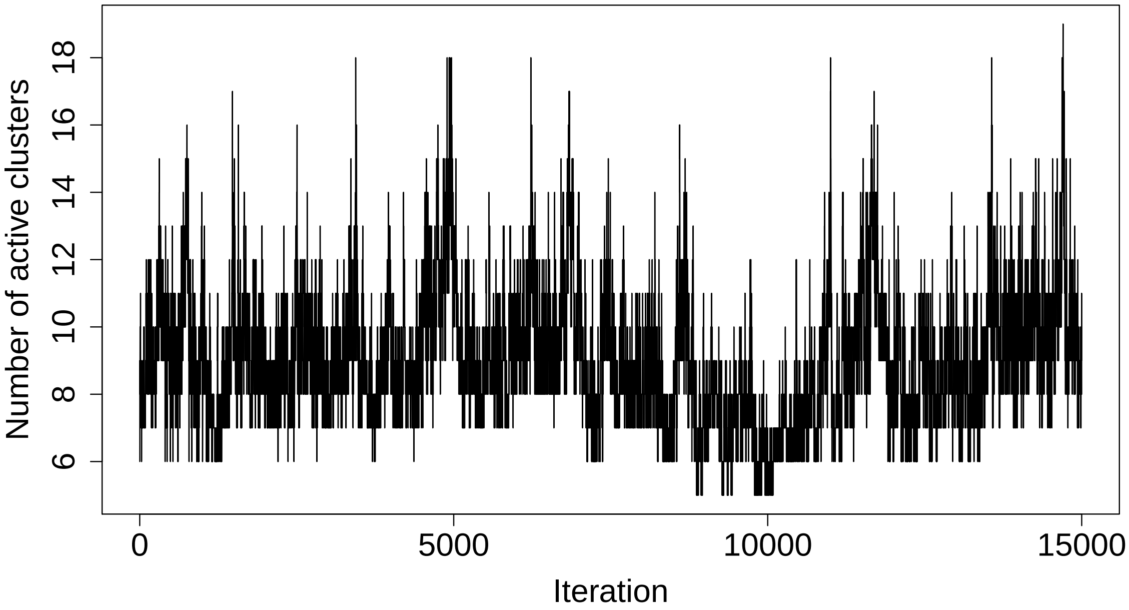


Figure 4.1: Number of active clusters (i.e. non-empty) identified at each iteration of the MCMC algorithm.

some of the curves is due to the fact that the clusters are estimated conditionally to the covariates and metabolite levels included in the regression model. Table 4.1 summarises the ethnic composition of each cluster, while Table 4.2 provides some basic information in terms of age, smoking habits, percentage of females and percentage of first generations (i.e. foreign-born) migrants in each cluster. It is worth noting that clusters 8, 9 and 10, the most insulin resistant clusters, are mostly composed of first generation migrants. In Table D.4 we report the ethnic composition for each cluster. The majority of South-Asians come from the Punjabi-Sikh minority, which represents the major South-Asian component in each cluster, with the exception of cluster 9, where there is a higher percentage of South-Asians of Muslim origin. To understand which covariates are the most important determinant of the response, we examine the posterior probability of each regression coefficient to be different from zero, $p(\omega_j = 1|rest, data)$. Ten predictors have the respective $p(\omega_j = 1|rest, data) > 0.5$ and are considered for further analysis (Figures E.3, E.4 and E.5). Cluster 1 is the largest and least insulin resistant group ($n = 1249, 57\%$ of participants). Its ethnic composition is: 71% of Europeans, 39% of South-Asians and 70% of Africans-Caribbean. The second largest group is cluster 6, comprising 243 participants (11% of the total, of which, 8% of Europeans, 13% of South-Asians and 24% of Africans-Caribbean). It is evident the clear distinction between clusters with a South-Asian

Table 4.1: Number of individuals in each ethnic group allocated to each cluster.

| **Cluster Number** | **Europeans** | **South-Asians** | **Africans-Caribbean** | **Total Number** |
|---|---|---|---|---|
| 1 | 784 | 382 | 83 | 1249 |
| 2 | 33 | 0 | 0 | 33 |
| 3 | 49 | 0 | 0 | 49 |
| 4 | 0 | 61 | 0 | 61 |
| 5 | 10 | 0 | 0 | 10 |
| 6 | 90 | 125 | 28 | 243 |
| 7 | 60 | 104 | 4 | 168 |
| 8 | 13 | 75 | 0 | 88 |
| 9 | 0 | 19 | 0 | 19 |
| 10 | 13 | 49 | 2 | 64 |

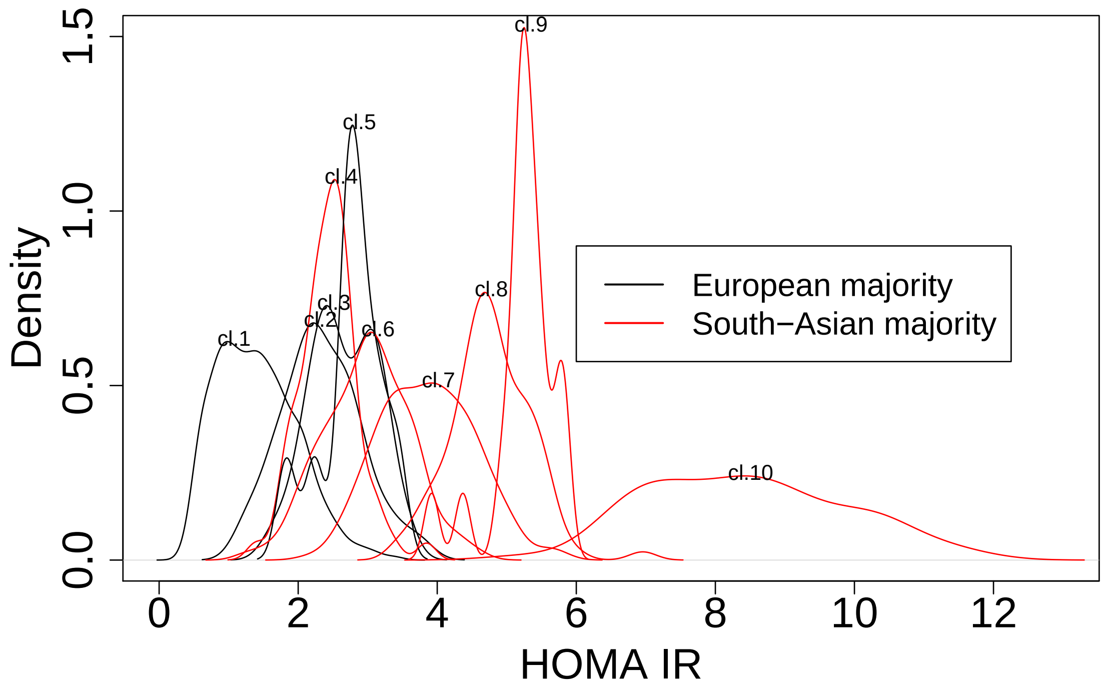


Figure 4.2: Empirical distribution of HOMA IR in each cluster. Black lines denote clusters with a higher proportion of Europeans, while red lines denote a higher proportion of South-Asians. A description of cluster main characteristics is given in Table 4.1. The numbers above each distribution denote the cluster. The plot is obtained through a kernel density estimate, with a Gaussian kernel with bandwidth equal to 1, as given by the standard function *density* employed in the `R` package *stats*.

majority, compared with Europeans, which are all characterised by higher levels of HOMA IR, with the exception of cluster 4. Cluster 5 (entirely Europeans) compared with cluster 1, presents both measures of adiposity(subscapular skinfold and sagittal diameter) are modestly higher, while levels of the amino acids tyrosine and isoleucine are significantly higher. Moreover, acetoacetate levels are lower compared to cluster 1, while the levels of alanine aminotransferase (ALT) in cluster 5 are higher than in cluster 1 (Figure E.5), the latter suggesting that raised HOMA IR levels may be characterised in this cluster by increased insulin levels with reduced clearance of insulin by the liver (Bonnet et al., 2011) implying relatively intact pancreatic beta cell function. The metabolite patterns for cluster 5 also indicate associations with both central and subcutaneous adiposity and amino acid perturbations.

Each of the ten clusters has a distinctive metabolic and phenotypic profile, consistent with suggestions that there are different pathways to type 2 diabetes (Udler et al.,

Table 4.2: Mean age, percentage of smoking habits, percentage of females and percentage of first generation migrants in each cluster.

| **Cluster Number** | **Mean Age** | **Ex-Smoker** | **Current Smoker** | **Females proportion** | **First Generation** |
|---|---|---|---|---|---|
| 1 | 52.51 | 27% | 25% | 18% | 46% |
| 2 | 53.27 | 33% | 45% | 3% | 21% |
| 3 | 52.02 | 41% | 31% | 10% | 10% |
| 4 | 49.54 | 11% | 16% | 33% | 98% |
| 5 | 52.00 | 20% | 30% | 0% | 10% |
| 6 | 51.93 | 23% | 19% | 12% | 67% |
| 7 | 51.43 | 17% | 19% | 11% | 67% |
| 8 | 51.25 | 11% | 12% | 9% | 86% |
| 9 | 51.89 | 11% | 37% | 11% | 95% |
| 10 | 50.95 | 17% | 25% | 9% | 80% |

2018) and that some pathways may be more strongly associated with a particular ethnic group. For example clusters 4 and 9 are entirely composed of South-Asians, while clusters 2, 3 and 5 are entirely Europeans. Of these clusters, 8, 9 and 10 are among the most insulin resistant with high levels of tyrosine, alanine, ALT and subcutaneous adiposity.

Some of the clusters identified are very small and will need replication in larger studies together with formal pathway analysis. However, these methods have generated intriguing, novel and persuasive clusters, which highlight the complexity and potential multiplicity of mechanisms underlying the development of insulin resistance and type 2 diabetes.

## 4.3 Model Fitting and Predictive Accuracy

We test the predictive accuracy, as well as the fitting of our model through a validation analysis performed on the real data. We split the dataset, randomly allocating 80% of the observations into a train set and the remaining 20% into the test set. Next, we asses the performance of the model through the classic Mean Squared Error (MSE), calculated as the average of the sum of the squared differences between the real outcome $y$ and the prediction given by our model. In Figure 4.3 we report the MSE value calculated at each iteration of the MCMC, for both the train and test set. From the MSE values for the test set we can see that the model is not overfitting, as the mean value of the MSE levels around 2.5 and at the same time the MSE values for the train set are steady around a mean value of 1.5, at convergence. The spike and slab prior allows to effectively perform variable selection, and therefore avoid a potential overfitting that a model with many covariate could incur in. We measure the goodness of fit of the model through the standard $R^2$ measure. The posterior mean of $R^2$ in our application is 0.8037 and, being the $R^2$ bounded between 0 and 1, the observed value of 0.8037 indicates a good fit of the model. Moreover in Figure 4.4 we show the distribution of the regression residuals. This plot is also showing the good fit of the model, with the residuals symmetrically concentrated around 0, with very few values higher than 3 in absolute value.

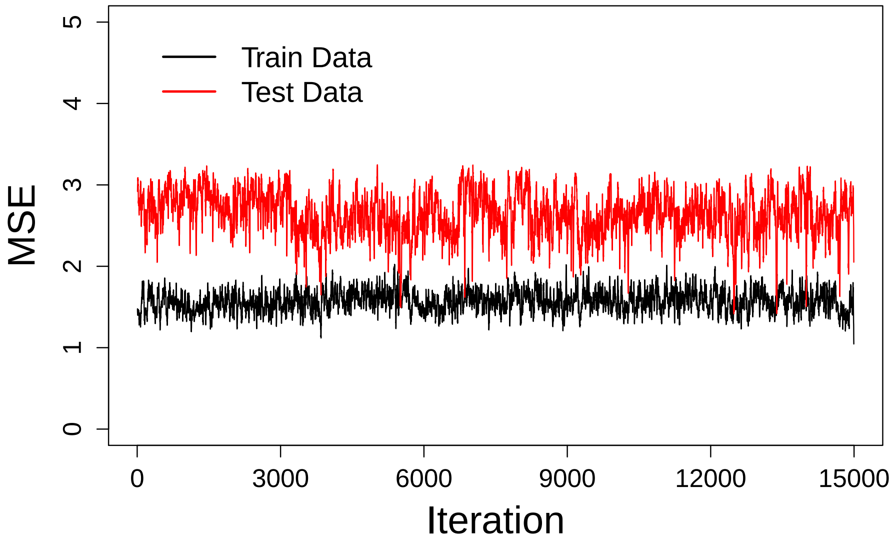


Figure 4.3: Mean Squared Error calculated at every iteration of the MCMC algorithm, respectively for the train set in black and the test set in red.

We then further asses the validity of our prior choices through a sensitivity analysis to determine the impact of different parametrisation of the spike and slab prior. In particular we tested different parametrisation of the Beta distribution over $\pi$, representing the prior probability of inclusion of a covariate in the model. We tested the following pair of values, respectively for $\pi_a$ and $\pi_b$: $(1, 1)$, $(1, 5)$, $(5, 1)$. The results of the analysis shows that the parametrisation with a Beta$(5, 1)$ allows a higher number of covariates to be included in the model, which is expected because such

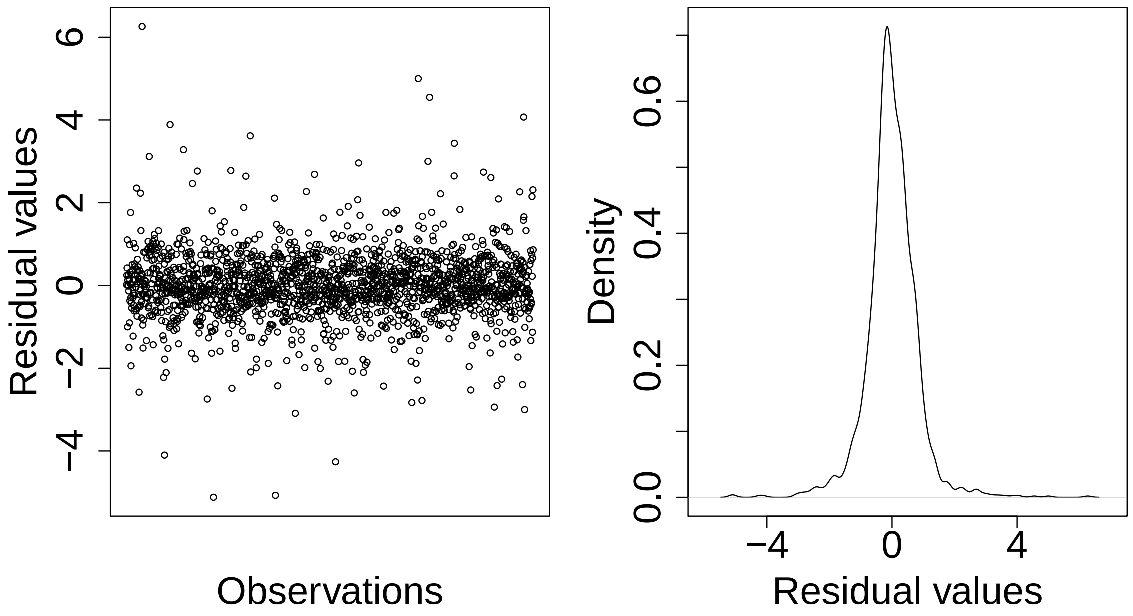


Figure 4.4: Regression residual values distribution calculated using the posterior mean of the fitted values. Each residual value is calculated as the difference between the real value of $y_i$ and the fitted value given by the model. The first plot on the left shows the residual value of each observation, while the second plot on the right shows a kernel density estimate of the empirical distribution of the residuals.

parametrisation put more probability mass close to 1 in the Beta density function, but critically the set of selected covariates always include the subset given by the stricter parametrisation used in the final application. Finally, the Beta$(1, 5)$ gives a prior distribution that favours smaller inclusion probabilities, but nevertheless the identified subset of covariates is selected anyway. We opt for the Beta$(1, 1)$ parametrisation, as this better represents our weak prior information about the number of covariates that should be included in the model. Moreover, the structure of the spike and slab allows for an effective selection of the variables, which is consistent with the other parametrisations.

## 4.4 Conclusions

The model presented in this chapter allows us to analyse multiple groups of patients and provides data-driven clustering of the observations thanks to the Bayesian non-parametric prior. We specify a Spike and Slab prior on the regression coefficients to effectively perform variable selection on the covariates, allowing us to understand which variables are important in predicting the dependent variable of interest, i.e. HOMA IR. We employ the proposed model to analyse data from the SABRE cohort study, a tri-ethnic information rich dataset on cardiovascular and metabolic diseases. Our clinical interest focuses on modelling the distribution of HOMA IR. We include anthropometric variables and metabolites concentrations as covariates in the regression framework. The results highlight the presence of sub-populations in the data, with a multi-ethnic composition, characterised by different levels of HOMA IR, which can lead to a different risk of developing type 2 diabetes. From the analysis, it is evident that clusters with higher levels of insulin resistance are composed mainly by the South-Asian ethnicity and, in particular, the more extreme clusters present a higher proportion of first-generation migrants. The results obtained from our analysis are promising and the proposed model has the potential to highlight areas for further research.

The next chapter introduces the problem of multiple Gaussian Graphical Models estimation. We provide a general multivariate regression framework, where the precision matrix of a multivariate Gaussian distribution is dependent on the graph realisation of a GGM.

# Chapter 5

# Bayesian Nonparametric Gaussian Graphical Models

In this chapter we discuss our approach to Bayesian modelling of multiple Gaussian Graphical Models (GGMs) and the application to the metabolomics SABRE data. We propose a novel approach to the estimation of multiple GGMs to analyse patterns of association among a set of metabolites, under different conditions, the ethnic origin in our case. We focus on the SABRE study baseline metabolic and phenotypic dataset, with a view to identify and elucidate potential mechanistic pathways to insulin resistance (and hence risk of developing type 2 diabetes), and to explore ethnic differences in these pathways. We model the relationship between a set of metabolites and a set of covariates through a Sparse Seemingly Unrelated Regressions model and we use GGMs to represent the conditional dependence structure among metabolites. We specify a Dependent Generalised Dirichlet Process prior on the edge inclusion probabilities to borrow strength across groups and we adopt the Horseshoe prior to identify important biomarkers. The statistical analysis poses several challenges: inter-subject variability, the high dimensionality of the dataset (due to the large number of variables under investigation) and the high correlation between metabolite levels. The statistical literature is rich in proposals on how to tackle these problems. We employ the Seemingly Unrelated Regressions (SUR, Zellner (1971)). To regularise posterior inference we adopt a Sparse SUR approach, assuming a local-global shrinkage prior for the regression coefficients, i.e. the Horseshoe prior (Carvalho et al., 2010), and we model association patterns among metabolites employing a Gaussian Graphical Model (GGM, Dempster (1972)). Zeros in the error precision matrix are obtained by imposing a set of conditional independence restrictions arising from an underlying graphical model (Lauritzen, 1996). Two common choices of prior distribution for the precision matrix are the G-Wishart prior of Lenkoski and Dobra (2011) and the Bayesian graphical Lasso of Wang (2012). The G-Wishart prior explicitly treats the graph as an unknown parameter leading to a direct inference of the underlying structure. However, the convergence of the posterior distribution can be slow due to the single edge update and the intractable normalizing constant that needs to be approximated. On the other hand, the Bayesian graphical Lasso is fast, thanks to the continuous priors, which enable a block Gibbs sampler that updates the precision matrix one column at time. However, this method does not explicitly provide a treatment of the underlying graphical structure. Here we use the Stochastic Search Structure Learning (SSSL) algorithm of Wang (2015) to specify the precision matrix prior distribution. The SSSL exploits the advantages of the G-Wishart and Bayesian graphical Lasso priors, enabling explicit structure learning while maintaining good scalability.

We specify a Generalised Dirichlet Process (GDP, Hjort (2000)) prior, previously exposed in Chapter 4 in a regression context, on the edge inclusion probabilities of a single GGM, allowing for clustering of the edges and sparsity in the graph. We briefly re-introduce the nonparametric prior in the context of graphical models, providing some results on the degree distribution induced by the GDP. Moving to a multiple GGMs scenario, we extend the GDP prior to multiple graphs, enabling borrowing information between graphs under different biological conditions introducing dependence through the Dependent GDP (DGDP). This strategy allows us to highlight common patterns and structural differences. In this context, each graph is characterised by the same set of nodes (that represent the dependent variables of the SUR model), connected by a set of group-specific edges. Thanks to the clustering property of the DGDP prior, we allow edges from different graphs to share the same edge probability and consequently to inform each other.

## 5.1 Methods

In this section we review the main properties of the SUR model and its generalisation to Sparse SUR. We also introduce the main properties of GGMs and we present our choice of prior distribution for the graph space based on the GDP. Finally, we generalise our modelling strategy to multiple GGMs.

### 5.1.1 Sparse SUR model

Consider $M$ response variables $\boldsymbol{y}_l,\ l = 1, \ldots, M$, each observed on $n$ individuals, i.e. $\boldsymbol{y}_l = (y_{l1}, \ldots, y_{ln})'$, modelled as individual linear regressions

$$\begin{cases} \boldsymbol{y}_1 = X_1\boldsymbol{\beta}_1 + \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{y}_l = X_l\boldsymbol{\beta}_l + \boldsymbol{u}_l \\ \vdots \\ \boldsymbol{y}_M = X_M\boldsymbol{\beta}_M + \boldsymbol{u}_M \end{cases} \tag{5.1}$$

where the $X_l$ is a $n \times p_l$ response specific matrix of explanatory variables, $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{lp_l})$ is a $p$-dimensional vector of regression coefficients and $\boldsymbol{u}_l = (u_{l1}, \ldots, u_{ln})$ is the $n$-dimensional vector of error terms, distributed as a Multivariate Normal, $\mathrm{N}\left(\mathbf{0}, I_n\right)$, where $I_n$ is the identity matrix of dimension $n \times n$.

The error terms are assumed to be correlated across equations. We denote with $\Omega$ the cross-equation precision matrix. We can rewrite the system of equations in a compact matrix form, as

$$\begin{aligned} &\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u} \\ &\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \Omega \otimes I_n) \end{aligned}$$

by concatenating the responses in a unique column vector $\boldsymbol{y}$ of dimension $Mn$. $X$ is now a block diagonal matrix of dimension $Mn \times Q$, where $Q = \sum_{l=1}^{M} p_l$ is the total number of parameters. $\boldsymbol{\beta}$ is an $Q$-dimensional vector containing all the regression coefficients. Here $\otimes$ denotes the Kronecker product. Note that the precision matrix of the concatenated error vectors implies that error terms within the same equation are independent (e.g. $u_{lj}$ and $u_{li}$ for $j \neq i$), but error terms corresponding to the same subject in different equations are assumed to be correlated (e.g. $u_{lj}$ and $u_{rj}$ for $l \neq r$). We shall denote the generic element of the regression coefficients vector $\boldsymbol{\beta}$ with $\boldsymbol{\beta}_{lj}$, which corresponds to the regression coefficient associated to the $j_{th}$ covariate in the $l_{th}$ equation.

### 5.1.2 Background on graphical models

We use the same terminology for graphical models introduced in Chapter 2. The graph $G$ can be represented by a set of $M(M-1)/2$ binary variables $Z = (z_{ij})_{i<j}$, where $z_{ij} = 1 \iff$ edge $(i,j) \in E$. There is a direct correspondence between the elements of the precision matrix $\Omega$ and the edges in the graph $G$. A missing edge in $E$ implies $\omega_{i,j} = 0$ (Wermuth, 1976), which in turn corresponds to a conditional independence assumption of $y_i$ and $y_j$ given the remaining variables $\boldsymbol{y}_{-ij}$, where $\boldsymbol{y}_{-ij}$ denotes the elements of the random vector $\boldsymbol{y}$ excluding the $i$ and $j$ coordinates. The parameter $\Omega$ is constrained to belong to the cone $PD_G$, i.e. the set of positive definite matrices with entries equal to zero for all $(i,j) \notin E$. We denote with $e_{ij}$ the edge between node $i$ and $j$ in the graph $G$, with $i,j \in \{1,\dots,M\}$ and let $r = M(M-1)/2$ be the total number of edges in the graph.

### 5.1.3 Prior Specification

We adopt the Horseshoe prior ((2.7)) of Carvalho et al. (2010) defined in Section 2.1.5, to impose regularisation on the regression coefficients $\boldsymbol{\beta}$. In the current regression framework the prior is specified as follows

$$\begin{aligned} \beta_{lj} \mid \lambda_{lj}, \tau_l &\sim \mathrm{N}\left(0, \lambda_{lj}^2 \tau_l^2\right) \\ \lambda_{lj} &\sim \mathrm{C}^+(0,1) \\ \tau_l &\sim \mathrm{C}^+(0,1) \end{aligned} \tag{5.2}$$

with $l = 1,\dots,M$ and $j = 1,\dots,p$. $\mathrm{C}^+$ denotes the standard half-Cauchy distribution, $\lambda_{lj}^2$ is the local shrinkage parameter, specific for the coefficient $\beta_{lj}$, while $\tau_l^2$ represents the overall shrinkage level for equation $l$. However, the original paper does not provide details for an efficient sampling scheme from the posterior distribution and a standard Gibbs sampling approach is difficult to implement due to the presence of the half-Cauchy prior. To overcome this problem, we adopt the conjugate sampler proposed by Makalic and Schmidt (2016), which allows a fast Gibbs sampling, avoiding to work directly with the half-Cauchy distribution. Makalic and Schmidt (2016) exploit the following relationship. Let $\kappa$ and $\rho$ be random variables such that

$$\kappa^2 \mid \rho \sim \mathrm{IG}(1/2, 1/a) \text{ and } \rho \sim \text{Inverse-Gamma}(1/2, 1/A^2) \tag{5.3}$$

then $\kappa \sim \mathrm{C}^{+}(0, A)$. Exploiting the scale mixture representation in (5.3) we can express (5.2) as

$$\begin{aligned}
\beta_{lj} \mid \lambda_{lj}, \tau_l &\sim \mathrm{N}\left(0, \lambda_{lj}^2 \tau_l^2\right) \\
\lambda_{lj}^2 \mid \nu_{lj} &\sim \text{Inverse-Gamma}\left(1/2, 1/\nu_{lj}\right) \\
\tau_l^2 \mid \xi_l &\sim \text{Inverse-Gamma}\left(1/2, 1/\xi_l\right) \\
\nu_{lj}, \xi_l &\sim \text{Inverse-Gamma}\left(1/2, 1\right)
\end{aligned} \tag{5.4}$$

We model the cross-equation precision matrix $\Omega$ with the SSSL prior of Wang (2015), specified as

$$p(\Omega) = \{C(\theta)\}^{-1} \prod_{i<j} \left\{(1-\pi)\mathrm{N}\left(\omega_{ij} \mid 0, v_0^2\right) + \pi \mathrm{N}\left(\omega_{ij} \mid 0, v_1^2\right)\right\} \prod_i \mathrm{Exp}\left(\omega_{ii} \mid \frac{\eta}{2}\right) 1_{\{\Omega \in PD_G\}} \tag{5.5}$$

where $\mathrm{Exp}(\omega \mid \eta)$ represents the Exponential density with expectation $1/\eta$ and $1_{\{.\}}$ is the indicator function. The normalising constant $C(\theta)$, with $\theta = \{v_0, v_1, \pi, \eta\}$, ensures that $p(\Omega)$ integrates to one over the space $PD_G$. The parameters $v_0, v_1$ are set to be small and large, respectively, in order to perform variable selection on the off-diagonal elements of the precision matrix. We do not impose regularisation on $\eta$, fixing its value to 1 as done in Wang (2015). The prior on $\pi$ is discussed later. The first product in (5.5) involving the off-diagonal elements of $\Omega$, involves a mixture of two Normal distributions and is similar to the Bayesian graphical Lasso. The second product multiplies $M$ Exponential densities for the diagonal elements of $\Omega$. Now, recalling the connection between the graph $G$ and its binary representation through the adjacency matrix $Z = (z_{ij})_{i<j}$, (5.5) can be rewritten as

$$p(\Omega \mid Z, \theta) = \{C(Z, v_0, v_1, \eta)\}^{-1} \prod_{i<j} \mathrm{N}\left(\omega_{ij} \mid 0, v_{z_{ij}}^2\right) \prod_i \mathrm{Exp}\left(\omega_{ii} \mid \frac{\eta}{2}\right) \tag{5.6}$$

$$p(Z \mid \theta) = \{C(\theta)\}^{-1} C(Z, v_0, v_1, \eta) \prod_{i<j} \left\{\pi_{ij}^{z_{ij}} (1-\pi_{ij})^{1-z_{ij}}\right\} \tag{5.7}$$

where $v_{z_{ij}}^2 = v_1^2$ if $z_{ij} = 1$ and $v_{z_{ij}}^2 = v_0^2$ if $z_{ij} = 0$, $C(Z, v_0, v_1, \eta)$ and $C(Z, v_0, v_1, \eta)$ are normalising constant for the respective densities. The joint distribution $p(\Omega, Z \mid \theta)$ admits (5.5) as a marginal distribution for $\Omega$. In the representation in (5.6)-(5.7) small values of $v_0$ give high probability to the event $z_{ij} = 0$, so that the distribution of $\omega_{ij}$ is concentrated around 0, implying that the correspondent edge will have a close-to-zero probability to be included in the graph $G$. Vice-versa for $v_1$ (Wang, 2015).

The choice of $v_0$ and $v_1 = v_0 \times h$ is important to ensure a good mixing of the MCMC and quick convergence to the true posterior distribution. The value of $v_0$ should be such that if the evidence is in support of $z_{ij} = 0$ then $\omega_{ij}$ is small enough to be replaced by zero. Wang (2015) discusses the choice of $v_0$ and $h$ and observe that, with standardised data, the MCMC converges quickly with $v_0 \geq 0.01$ and $h \leq 1000$. Finally choosing a value for $\eta$ is easier, as with standardised data, a choice of $\eta = 1$ assigns probability to the entire region of plausible values for the inverse variances $\omega_{ii}$.

There is a wealth of literature regarding the choice of the prior distribution for $\pi_{ij}$, the edge inclusion probability. See, for example, Carvalho and Scott (2009) and Tan et al. (2016) for a review of some popular methods. In this paper we adopt a nonparametric Bayesian approach to model the uncertainty about the inclusion probabilities, allowing for clustering of the edges and the possibility to impose sparsity on the graph. We specify a GDP prior on $\pi_{ij}$ as follows

$$\begin{aligned} \{\pi_{ij}\}_{i<j} \mid P &\overset{iid}{\sim} P \\ P \mid \alpha, \mu, P_0 &\sim \text{GDP}(\alpha, \mu, P_0) \\ P_0 \mid a_\pi, b_\pi &= \text{Beta}(a_\pi, b_\pi) \\ \alpha \mid \alpha_a, \alpha_b &\sim \text{Gamma}\left(\alpha_a, \alpha_b\right) \\ \mu &\sim \text{Beta}(a_\mu, b_\mu) \end{aligned} \tag{5.8}$$

We can tune the hyper-prior parameters characterising the base measure $P_0$ to achieve the desired level of sparsity. The parameters $\alpha$ and $\mu$ control the clustering structure of the GDP (note that posterior clustering depends also on the choice of the base measure). The choice of the hyper-parameters depends on the particular application. The model for $e_{ij}$ is then given by:

$$\begin{aligned} \{e_{ij} \mid \pi_{ij}\} &\overset{ind}{\sim} \text{Ber}(\pi_{ij}), \qquad i < j \\ \{\pi_{ij}\}_{i<j} \mid P &\overset{iid}{\sim} P \\ P \mid \alpha, \mu, P_0 &\sim \text{GDP}(\alpha, \mu, P_0) \end{aligned} \tag{5.9}$$

The above equations defines a GDP Mixture model (GDPM, see Lo (1984) and Barcella et al. (2017)) for $\{e_{ij}\}$. Recalling the discrete nature of the GDP. we can rewrite (5.9) as

$$\{e_{ij}\}_{i<j} \mid P \overset{iid}{\sim} \sum_{k=1}^{\infty} \psi_k \text{Ber}(e \mid \pi_k)$$

where the $\pi_k$ denote the (unique) locations of the GDP prior.

### 5.1.4 Degree Distribution

One of the main consequences of choosing a GDP prior is that the edges are clustered on the basis of their inclusion probability. A priori, the GDP does not constraint the number of clusters to a finite value, indeed their number can grow as new data become available. Only a posteriori, once we observe the data, the estimated number of clusters is finite, potentially equal to the number of edges. We now investigate the possible graphs structure supported by a GDP prior. We follow the framework of Tan et al. (2016) and describe some properties of the degree distribution.

The degree, $D_i$, of a node $i$ is the number of connections that involve node $i$, so $D_i = \sum_{j \neq i} e_{ij}$, where $e_{ij}$ is the edge connecting nodes $i$ and $j$. The degree $D_i$ is then bounded between 0 and $M-1$, the total number of nodes minus one. The following properties hold (proofs in Appendix B):

a Conditionally on $\pi_{ij}$, the probability that a node $i$ is connected to a node $j$ is $\pi_{ij}$.

b The degree of a node $i$ is distributed as a mixture of Binomial distributions, with mixing weights given by the GDP

$$D_i \mid P \sim \sum_{k=1}^{\infty} \psi_k \text{Binomial}(M-1, \pi_k)$$

where, once again for ease of notation, we have substituted the index $(ij)$ with $k$. We have that $\mathrm{E}\left[D_i \mid P\right] = \sum_{k=1}^{\infty} \psi_k (M-1)\pi_k$
$\mathrm{Var}\left[D_i \mid P\right] = (M-1)\sum_{k=1}^{\infty} \psi_k \pi_k \left[(1-\pi_k) + (M-1)\pi_k\right] - \left[\sum_{k=1}^{\infty} \psi_k (M-1)\pi_k\right]^2$

c Marginalising over the random measure, we obtain:

$$\begin{aligned}
\mathrm{E}\left[D_i\right] &= (M-1)\frac{a_\pi}{a_\pi + b_\pi} \\
\mathrm{E}\left[D_i^2\right] &= (M-1)\left\{\frac{a_\pi}{a_\pi + b_\pi} + (M-2)\frac{(a_\pi + 1)a_\pi}{(1 + a_\pi + b_\pi)(a_\pi + b_\pi)}\right\}
\end{aligned}$$

The shape of the degree distribution highlights structural characteristics of the graph implied by the prior choice, which are relevant in data analysis. In particular we focus on sparsity. In a *dense* graph each node is connected to many others and, as a consequence, there are few pairwise conditional independences, while a sparse graph presents fewer connections and hence the graph can be decomposed into subgraphs defined by conditional independence structures. A careful choice of prior hyper-parameters allows us to obtain the desired level of sparsity, retaining at the same time a good level of flexibility. To better understand the shape of the degree distributions implied by the GDP prior, (5.8), we perform a sensitivity analysis for different values of $\alpha$ and $\mu$ and different parametrizations of the base distribution $P_0$. In Figure E.6 and E.7 we present the resulting degree distribution for different combinations of hyper-parameters. It is evident that our prior choice is able to accommodate different shapes. However, simulations show that, by appropriate choice of hyper-parameters, we can obtain an exponential decay in the tails of the functions, but not a power law decay.

### 5.1.5 Multiple GGMs

Often in applications we observe groups of subjects under different experimental conditions. In the SABRE study, for example, we are interested in understanding how patterns of association between metabolites vary across three different ethnicities, in particular in relation with cardiovascular diseases and diabetes. In our application ethnicity defines three natural sub-samples, each characterised by its own graph. In general, we expect different groups to share some common structure as well as group specific connection patterns. Estimating a single graphical model would lead to an implicit assumption of homogeneity of the underlying graphs across the ethnicities, with a consequent loss of information about their heterogeneity and a consequent high risk of false positives. On the other hand, inferring each graph individually might lead to a loss of power given the reduction in sample size. There is a growing research interest in multiple graphical models. For example, in the Bayesian framework Peterson et al. (2015) estimate multiple GGMs through a Markov random field prior, which encourages common edges. Also relevant is the work of Tan et al. (2016), which propose a multiplicative prior to capture common and group-specific structures. We propose to model multiple graphs through an extension of the GDP prior, i.e. the Dependent Generalised Dirichlet Process (DGDP, Barcella et al. (2017)). Due to the discrete nature of the DGDP, each edge can be clustered together with any other edge, independently of the group $g$ of origin. This ensures sharing of structural information among groups, at the same time maintaining parsimony in the number of parameters to be estimated. This strategy also allows detecting group-specific connections.

Suppose we observe $R$ groups, for example, defined by ethnicity in the SABRE study.

Each sub-sample $g, g = 1, \ldots, R$, is characterised by a specific sample size $n_g$ and its own graph $G_g$. Here we assume that the vector of regression parameters $\boldsymbol{\beta}$ is common to all groups, although this assumption can be easily relaxed. The prior distributions in (5.6) and (5.7) are generalised to handle multiple precision matrices $\Omega_g$, and therefore multiple adjacency matrices $Z_g$ as follows:

$$p(\Omega_g \mid Z_g, \theta_g) = \{C(Z_g, v_0, v_1, \eta_g)\}^{-1} \prod_{i<j} \mathrm{N}\left(\omega_{g,ij} \mid 0, v^2_{z_{g,ij}}\right) \prod_i \mathrm{Exp}\left(\omega_{g,ii} \mid \frac{\eta_g}{2}\right) \tag{5.10}$$

$$p(Z_g \mid \theta_g) = \{C(\theta_g)\}^{-1} C(Z_g, v_0, v_1, \eta_g) \prod_{i<j} \left\{\pi_{g,ij}^{z_{g,ij}} (1 - \pi_{g,ij})^{1-z_{g,ij}}\right\} \tag{5.11}$$

The hyper-parameters $v_0^2$ and $v_1^2$ remain unchanged and are common to all groups. We can see that, conditional on the inclusion probabilities $\boldsymbol{\pi}_{g,ij}$, $\eta_g$, $v_0^2$ and $v_1^2$, (5.10) and (5.11) are independent across groups. The prior in (5.8) on $\pi_{g,ij}$ can be extended in presence of multiple groups, so that the random measures associated to each group are dependent. Dependence can be introduced in the weights of the stick-breaking representations, by allowing $\psi_k$ to be a function of a categorical $x$, identifying the group. Note that dependence on other group-specific covariates (when available) can be easily introduced. The resulting process is called Dependent GDP, which is defined as follows. Let

$$P_g = \sum_{k=1}^{\infty} \psi_{kg} \delta_{\theta_k}$$

be the random measure associated to group $g$. The locations are *iid* draws from a common base measure $P_0$, as before. The weights still admits the stick-breaking representation:

$$\psi_{kg} = \phi_{kg} \prod_{j=1}^{k-1} (1 - \phi_{jg}) \ , \ k = 2, 3, \ldots$$

$$\psi_{1g} = \phi_{1g}$$

Each $\phi_{kg}$ has a Beta distribution, $\mathrm{Beta}(\alpha\mu_g, \alpha(1 - \mu_g))$, but now $\mu_g$ is group-specific. (Barcella et al., 2017) propose to introduce dependence across the $\{\mu_g\}$ employing a Beta regression framework and letting the $\mu_g$ depend on a categorical covariates denoting group. Using the DGDP, the model in (5.8) can then be extended to the multiple graphs as follows

$$\begin{aligned}
\{\pi_{g,ij}\}_{i<j} \mid P_g &\overset{ind}{\sim} P_g \\
P_g \mid \alpha, \mu_1, \dots, \mu_R, P_0 &\sim \text{DGDP}(\alpha, \mu_1, \dots, \mu_R, P_0) \\
P_0 \mid a_\pi, b_\pi &= \text{Beta}(a_\pi, b_\pi) \\
\alpha \mid \alpha_a, \alpha_b &\sim \text{Gamma}(\alpha_a, \alpha_b) \\
\mu_g &= \text{logit}(\boldsymbol{x}_g \boldsymbol{\zeta}) \\
\boldsymbol{\zeta} \mid \boldsymbol{\zeta}_\mu, \boldsymbol{\zeta}_\Sigma &\sim \text{N}_R(\boldsymbol{\zeta}_\mu, \boldsymbol{\zeta}_\Sigma)
\end{aligned}$$

$\boldsymbol{x}_g$ is a categorical design vector of dimension $R$ which includes an intercept term and identifies the group from which the observations come from. $\boldsymbol{\zeta}$ is a vector of regression coefficients, to which we assign a Normal prior. In our application the European ethnicity is the reference group. The DGDP process offers a convenient way to share information across different groups and ensures a greater flexibility than the GDP thanks to the richer parametrization. Note that $\boldsymbol{x}_g$ can include other group specific covariates when available. The MCMC algorithm for posterior inference from a DGDP process is based on a truncation of the infinite mixture (Ishwaran and James, 2001). A discussion on how to choose the truncation level can be found in Ishwaran and James (2001) and Barcella et al. (2017). Details of the MCMC algorithm can be found in Appendix B.

## 5.2 Simulations results

In this section we compare the performance of the proposed model with respect to the baseline parametric version of the SSSL model. We then investigate the the efficacy of the proposed nonparametric model analysing synthetic datasets simulated from three different sparse SUR structures.

We start by providing a comparison between the nonparametric and the parametric version of the SSSL model. In Figures 5.1 we report the AUC values for two different simulated datasets, where for each simulation we generate data from a SUR model characterised by four graphs with 20 nodes each and a common linear regression term. In the first simulation we generate the first sparse graph with 20 edges and create the other three graphs as perturbation of the first by randomly removing and adding five edges. In the second simulation we generate the first sparse graph with 50 edges and create the other three graphs as perturbation of the first by randomly removing and adding ten edges. The aim of this comparison is to asses whether the nonparametric model performs better than its parametric counterpart under two different conditions, the first where the number of edges is low compared to the dimension of the graph and the second one where the number of edges included is about one third of the dimension of the graph. We can see from Figure 5.1 that the nonparametric model performs better in both cases, especially in the sparser scenario, providing overall higher AUC scores and also a lower variability over repeated simulations. We further observe that the nonparametric model is always performing at least as well as the parametric one, despite the higher complexity. This fact may be explained by the properties of the Bayesian nonparametric prior that allows the model to grow in complexity as needed, while at the same time retaining a small number of parameters thanks to the implicit clustering provided by the DP type priors. Next we compare the performance of another type of nonparametric prior in three different scenarios as described earlier. Each model is characterised by specific multiple graphs and a common linear regression term. We test our model on three different scenarios, investigating the ability of the proposed models to recover the true underlying graphical structure of each group through the Area Under the Curve (AUC) and the ability to correctly estimate the mean regression parameters. The purpose of the three simulation settings is to determine the models performance with different graph structures. For the first scenario we closely follow the work Peterson et al. (2015) in which every node in the first graph has two connections, following a second order autoregressive structure, and the other three dependent graphs are a perturbation of the first. This scenario is intended to check the performance of the models on a well-defined structure where the nodes have an ordering. The second scenario is based on a random allocation of the edges, and therefore we seek to assess the performance of the models on an opposite situation compared with the first scenario. Finally, in the third simulation study, we want to assess the ability of our models to estimate graphs that resemble the complex structure of the original data, to this end we select a subset of metabolites and we construct the graph from those observations. In the following paragraphs we give the details about the simulations and the respective results.

In the first simulation we generate four multiple graphs following the guidelines of Peterson et al. (2015). We construct four precision matrices $\Omega_1, \Omega_2, \Omega_3$ and $\Omega_4$ corresponding to graphs $G_1, G_2, G_3$ and $G_4$, of $M = 20$ nodes (for a total number of

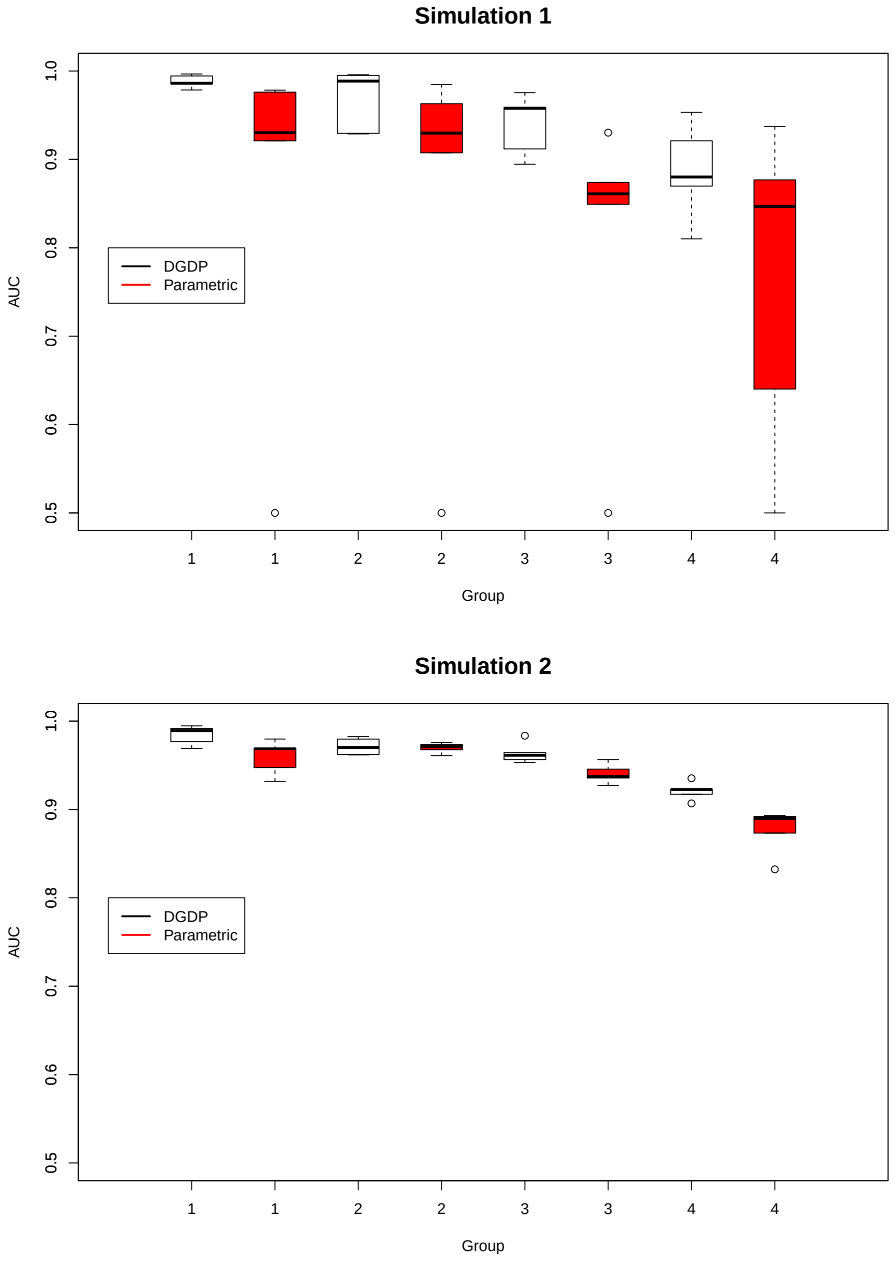


Figure 5.1: Comparison between DGDP and Parametric models. Boxplots of the AUC values for two simulated datasets. The AUC distribution is evaluated over 10 replicates for each scenario. The number on the x-axis refer to the group.

possible edges of $r = M(M-1)/2 \times 4 = 760$). We first define the precision matrix $\Omega_1$ and then we derive the others as a perturbation of the first. $\Omega_1$ is a $M \times M$ symmetric matrix with the main diagonal elements equal to one, first off-diagonal elements $\omega_{i,i+1} = \omega_{i+1,i} = 0.5$, for $i = 1, \ldots, 19$ and second off-diagonal elements $\omega_{i,i+2} = \omega_{i+2,i} = 0.4$, for $i = 1, \ldots, 18$, while the rest of the elements are set to zero. The total number of non-zero off-diagonal elements is 37. To construct $\Omega_2$, we remove ten edges at random from $\Omega_1$, setting the corresponding entries to zero. Then, we randomly add ten edges that are not present in $\Omega_1$, giving a value of 0.5 to the new precision coefficients. The procedure is repeated similarly for $\Omega_3$ and $\Omega_4$, avoiding the replacement of edges that were previously deleted. The newly created matrices are not necessarily positive definite, to this end, we compute the nearest positive-definite approximation through the `R` function *nearPD* (Higham, 2002), from the package Matrix). The precision matrices $\Omega_2, \Omega_3, \Omega_4$ constructed with this procedure are a perturbation of $\Omega_1$: as a result they exhibit some common edges and some group specific connections. The number of observations is fixed to $60, 50, 50, 40$, for group 1, 2, 3, and 4 respectively.

The second simulation scenario is similar to the first, but $\Omega_1$ has now a unit diagonal and we add 60 non-zero off-diagonal elements, chosen randomly from the $r$ possible edges. $\Omega_2$, $\Omega_3$ and $\Omega_4$ are constructed removing 10 edges and adding 10 new edges, randomly selected as before. Once again the number of observations is fixed to $60, 50, 50, 40$, for group 1, 2, 3, and 4 respectively.

The third simulated example reproduces the structure of the SABRE dataset. We randomly select 40 metabolites from the original dataset, keeping the original sample size of the three ethnic groups. For each group we fix precision matrices equal to the respective empirical partial correlation matrix. We set to zero the elements of the precision matrices in absolute value below 0.1, which implies missing edges in the associated graphs. The response for each group is then simulated by a multivariate Normal using the corresponding precision matrix.

Finally, all three simulation scenarios are characterised by the same linear regression term $X\boldsymbol{\beta}$. We simulate $p = 3$ independent covariates, each drawn from a Normal distribution, with mean 1, -1 and 1 and standard deviation 0.5, 0.5 and 1, respectively for $X_1$, $X_2$ and $X_3$. The regression coefficients $\boldsymbol{\beta}$ are fixed as follows: the intercept is fixed to 1 for every equation and the other three regression coefficients are chosen at random (with replacement) in the set $\{-1, 1, -2, 2, 0\}$. In this way, we introduce zeros in the regression coefficients vector, so that a covariate might not have an effect on a particular response. To evaluate the performance of our model in estimating the true underlying graph, we use the Area Under the Curve (AUC), which is a normalised measure of the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is created by plotting the true positive rate against the false positive rate at various thresholds. The AUC distributions for ten simulated datasets for the first and second scenarios are displayed in Figure 5.2 (top and middle panels) The distributions are concentrated between 0.9 and 1, denoting the ability of the model to recover the true graphical structure. We compare the DGDP with a Dependent DP (DDP, (MacEachern, 1999)), in particular with the ANOVA DDP (De Iorio et al., 2004). The performance is very similar for both algorithms. The third scenario presents similar results. The recovery of the true graph structures for groups one and two is excellent, while for group three (corresponding to the African-Caribbean ethnicity) the performance is slightly worse (with a median AUC level of about 0.8), which is, however, expected given the much smaller sample size. In Figure 5.3 we plot the posterior distributions of the regression coefficients $\beta_{l,2}$, $\beta_{l,3}$ and $\beta_{l,4}$, for each equation $l$ (we do not report the posterior distribution of the intercept) for 20 randomly selected responses and just one of the 10 replicates. The true value of each coefficient is represented by a dashed vertical red line. The posterior distribution are concentrated around the true values and the model is able to correctly identify covariates which have no effect in a particular equation.

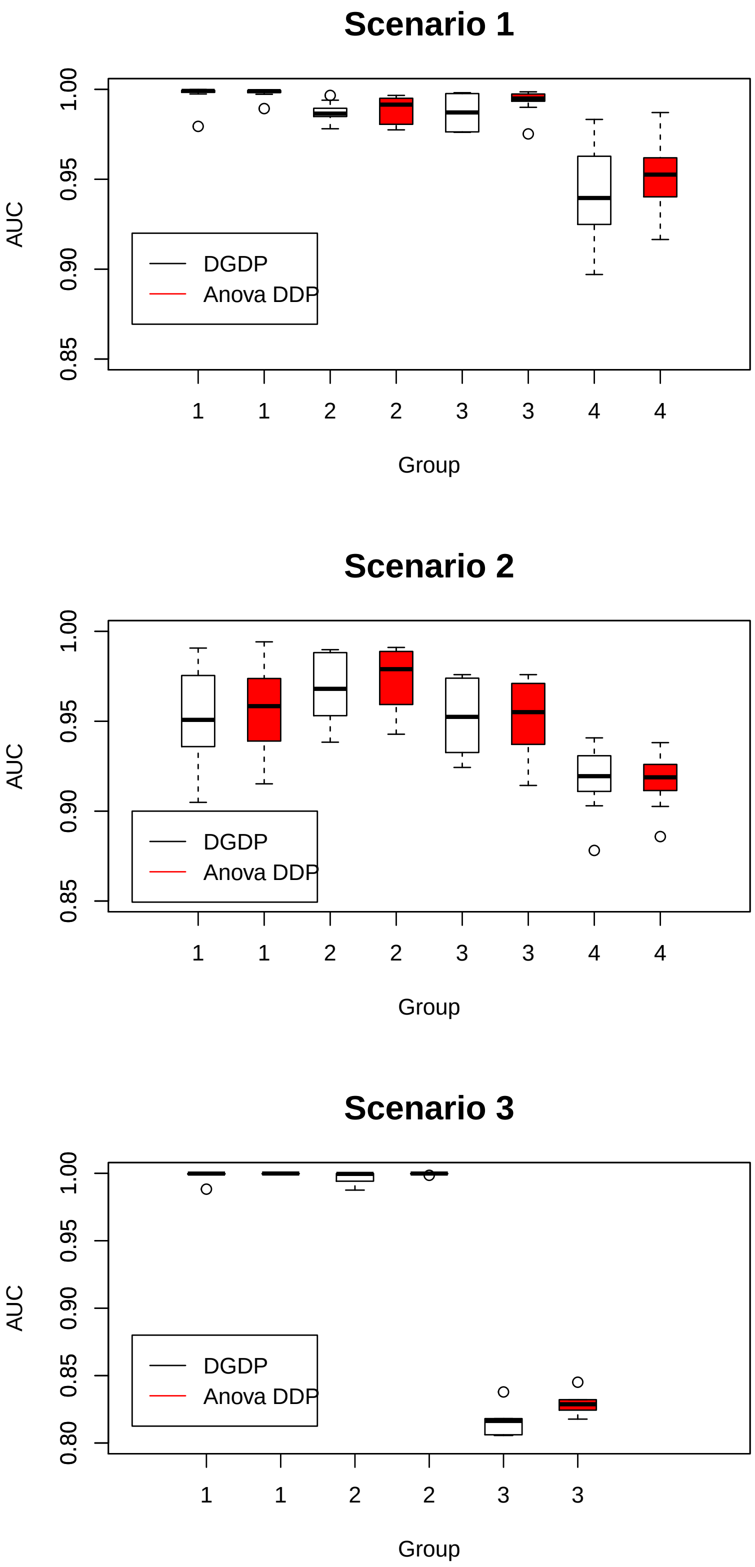


Figure 5.2: Comparison between DGDP and ANOVA DDP. Boxplots of the AUC values for the three simulation scenarios. The AUC distribution is evaluated over 10 replicates for each scenario. The number on the x-axis refer to the group.

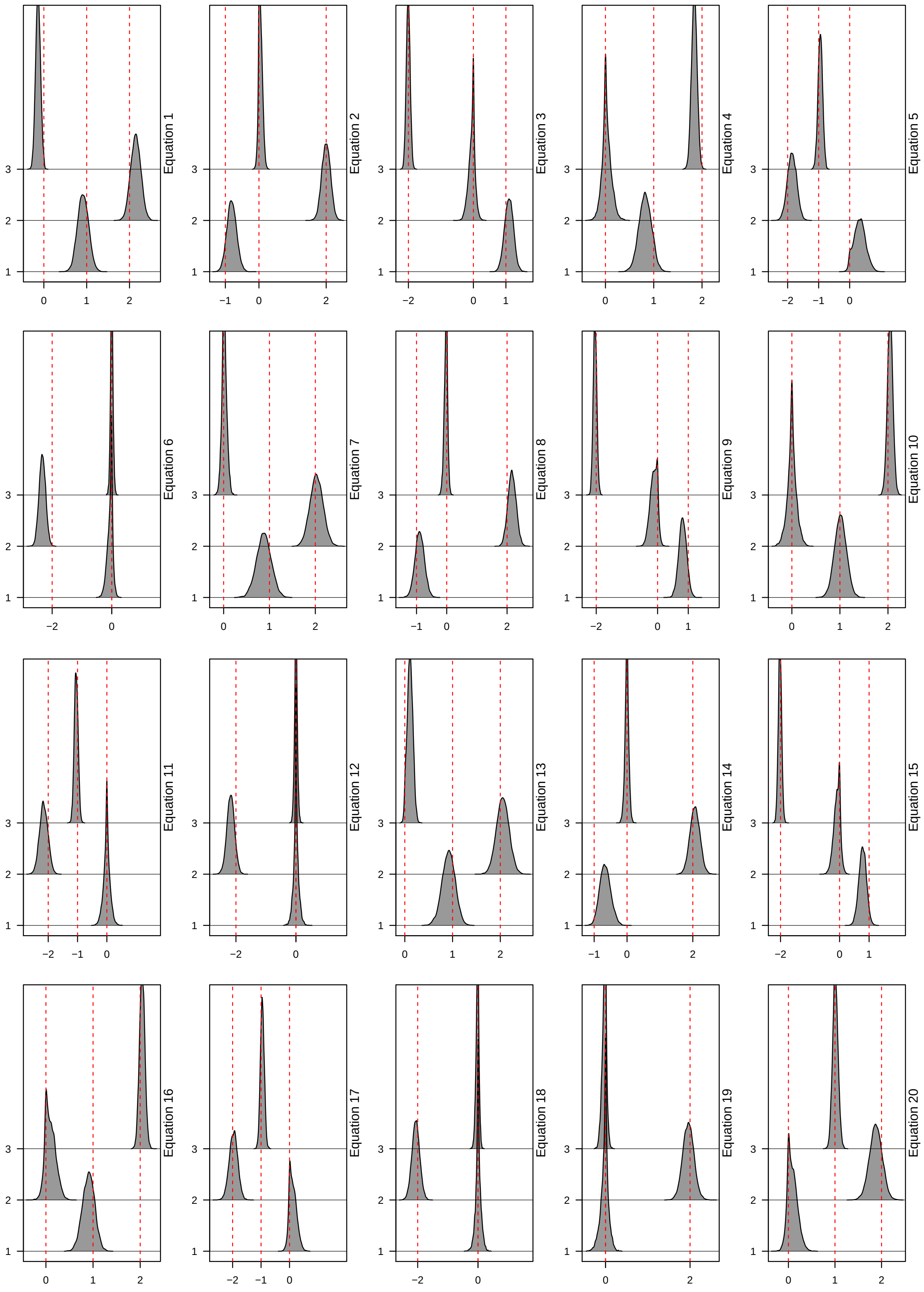


Figure 5.3: Posterior distributions of $\beta$ for 20 randomly selected equations and for one of the 10 simulation replicates. The red dashed vertical lines represent the true value of $\beta$ used in the simulations.

## 5.3 SABRE results

In this section we fit the proposed model for multiple GGMs to the SABRE metabolic dataset. The dataset described in Chapter 1 has a total of 2200 observations, stratified in three ethnicities, 1103 Europeans, 978 South-Asians and 119 Africans-Caribbean. The number of nodes (i.e. the number of equations in the SUR model) is $M = 88$, a list of which can be found in Table D.1. As predictors in the regression term of the mean, we include the covariates listed in Table D.2, consisting of measures of body-fat distributions, liver health functions, enzymes and control variables, such as smoking habits, sex and age (the total number of the covariates is $p = 18$ plus the intercept). All the covariates are included in each equation, but variable selection is equation-specific. We specify the following prior distributions. The scale parameters for the Normal mixture in (5.10) are chosen to ensure sparsity in the estimated graph, so that negligible and small off-diagonal coefficients of the precision matrix are set to zero. We choose $v_0 = 0.01$ and $h = 100$, while $\eta_g = 1$, following the recommendations of Wang (2015). The DGDP base measure $P_0$ is a Beta($a_\pi = 0.01, b_\pi = 0.01$). The concentration parameter $\alpha$ is assigned a Gamma($\alpha_a = 0.1, \alpha_a = 2$) prior, while the vector of coefficients $\boldsymbol{\zeta}$ in the Beta regression is given a Normal distribution with mean $\boldsymbol{\zeta}_\mu = \mathbf{0}$ and covariance matrix $\boldsymbol{\zeta}_\Sigma = 10 \times I_R$. We specify a Horseshoe prior for regression coefficients $\boldsymbol{\beta}$ as described in (5.4). We run the MCMC for 30000 iterations, comprising a burn-in period of 10000 iterations. In addition to the multiple graphs we also estimate the differential networks (Fuente, 2010; Valcárcel et al., 2011) arising from the pairwise comparison between the three ethnicities. A differential network includes all the edges that are present only in one of the two groups (i.e. present in one group and not the other and vice-versa), thus helping us to understand the main differences between two ethnicities. Here we focus mainly on the differences between Europeans and South-Asians, since the African-Caribbean ethnicity has a very small sample size that heavily affects the estimation of the latent graph.

In Figure 5.4 we show the differential network between Europeans and South-Asians, where an edge between two nodes is added to the differential graph if the probability of an edge being in one ethnic specific graph but not in the other is higher than 0.5. It is worth noting that there are no edges among the majority of lipoproteins subfractions, which implies that the presence or absence of those connections are shared by both of these ethnicities. On the other hand, the majority of the amino acids have some distinct connections, highlighting potential differences in underlying metabolic processes. For example, the amino acid *Histidine* features many connections with other amino acids and a subset of lipoproteins sub-fractions, with these edges only present in the South-Asian group. Other central nodes in the differential network are *Acetoacetate*, *Acetate*, *Pyruvate* and *Lactate*. In Appendix we report the individual networks for all three ethnicities in Figures E.8, E.9 and E.10, respectively for Europeans, South-Asians and Africans-Caribbean. These networks are characterised by a high number of edges, in particular, we can notice a very highly connected group of lipoproteins in the first two largest ethnicities. The African-Caribbean graph has less edges, but we can still identify clusters of connected lipoproteins.

To gain a better understanding of the estimated connections and to relate the estimated graph to known metabolic pathways, we conduct a pathway over-representation analysis (ORA) using the online software *MetaboAnalyst* (Chong et al., 2018). We include in the analysis all metabolites that have a connection in the differential network. ORA evaluates statistically the fraction of metabolites in a particular pathway found among the user-specified set of metabolites, in our case the metabolites with connections in the differential network. For each pathway, input metabolites that are part of the pathway are counted. Next, every pathway is tested for over or under-representation in the list of input metabolites using the hypergeometric test. The most represented pathways are the ones with smaller p-value levels and higher number of over-represented metabolites. Here we discuss the first four top-ranked pathways, *Pyruvate Dehydrogenase Deficiency (E3)*, *Pyruvate Carboxylase Deficiency*, *Diabetes Mellitus (MODY) Non-Insulin-Dependent* and *Chronic Progressive External Ophthalmoplegia (CPEO) and Kearns-Sayre Syndrom (KSS)*. Pyruvate dehydrogenase and pyruvate carboxylase deficiency are the most common disorders in pyruvate metabolism. Pyruvate dehydrogenase (PDH) is an enzyme complex made of three subunits, pyruvate dehydrogenase, dihydrolipoamide acyltransferase and dihydrolipoamide dehydrogenase (indicated as E3). The enzyme complex converts pyruvate into acetyl-CoA, that together with oxaloacetate, are two essential substrates in the production of citrate. PDH Complex Deficiency therefore leads to a limited production of citrate and because citrate is the first substrate in the tricarboxylic acid cycle, the cycle is blocked and other metabolic pathways need to be stimulated to produce acetyl-CoA. However, the most common deficiency involves the first subunit, while mutations in the second and E3 are less often the cause for PDH Complex Deficiency. The enzyme defect causes more pyruvate to be metabolized to lactate and leads to lactic acidosis (Bissonnette and Bissonnette, 2006). Overall, PDH Complex plays a key role in regulating the supply of adenosine triphosphate during the feed-fast cycle, where cells must select fatty acid or glucose as energy source, therefore it is important in regulating the glucose metabolism and, therefore, PDH deficiency is related to metabolic diseases, like type 2 diabetes and obesity (Lee, 2014). Of particular interest is pyruvate carboxylase deficiency. Lao-On et al. (2018) explore the role of pyruvate carboxylase in human diseases, such as diabetes. Pyruvate carboxylase (PC) is an anaplerotic enzyme which plays an essential role in various cellular metabolic pathways, including gluconeogenesis and and glucose-induced insulin secretion. Pyruvate originates as the final product of the pathway Pyruvate. In aerobic conditions, Pyruvate enters mitochondria via the mitochondrial pyruvate carrier, where it may be further metabolized in two different ways. In non-gluconeogenic tissues, like muscles and brain, Pyruvate is decarboxylated to form acetyl-CoA catalysed by the Pyruvate Dehydrogenase Complex. In gluconeogenic tissues, most of Pyruvate entering mitochondria is carboxylated by the enzyme pyruvate carboxylase to form oxaloacetate. Given the importance of oxaloacetate in various biochemical pathways, perturbation of oxaloacetate production by PC can produce serious diseases such as type 2 diabetes or neurological disorder.

Diabetes Mellitus (MODY) is an autosomal dominant monogenic disorder of pancreatic beta cells that is usually manifested before the age of 30 and accounts for $1-3\%$ of diabetes in this age group (Misra and Owen, 2018), although the prevalence of MODY in South-Asians is low, despite their increased risk of type 2 diabetes (Ehtisham et al., 2004). Finally, Chronic progressive external ophthalmoplegia (CPEO) is one of the most common mitochondrial disorders in adults. The main symptom is a slowly progressive extra-ocular muscle weakness. KSS and CPEO are probably the same disorder, but differ in the degree of severity (Gilman, 2011). In both CPEO and Kearns–Sayre syndrome, hearing loss and diabetes mellitus can precede the onset of muscle involvement by years (Shoffner et al., 1990). Additionally, involvement of systems other than muscle is common in CPEO. Multi-system involvement can cause functional impairments secondary to dysfunction of (proximal) skeletal muscles, retina, cochlea, cerebrum, cerebellum and heart (Smits et al., 2011). Ocular manifestations include, among the others, retinopathy and optic atrophy. Cardiac manifestations include cardiac conduction block and cardiomyopathy. Cerebral manifestations can include epilepsy, cerebellar ataxia, and dementia. Endocrine involvement includes diabetes mellitus, hypothyroidism, hypoparathyroidism, and hypogonadism. Sensorineural hearing loss and gastrointestinal involvement are also possible (Vorgerd and Deschauer, 2011).

In conclusion, these findings suggest plausible metabolic pathways which may be disordered to a greater extent in South-Asians and contribute to their excess risk of diabetes and cardiovascular disease.

In Figure 5.5 we plot the posterior mean of the regression coefficients $\beta_{lj}$, for each equation $l$ and covariate $j$. The only covariates that do not show association with any metabolite are waist to hip ratio (WHR) and diastolic blood pressure. It is worth noting, however, that WHR has a strong positive correlation with some of the other measures of adiposity (such as sagittal diameter), which can result in the selection of a variable over the other. The same scenario applies to the variable diastolic blood pressure, which is positively correlated with systolic blood pressure.

To evaluate the sensitivity of the model to the prior parametrisation, we perform a sensitivity analysis for different combinations of $v_0 \in \{0.01, 0.1\}$ and $a_\pi, b_\pi \in \{0.01, 0.5\}$. In particular, the choice of $v_0$ plays a major role in determining the level of sparsity of the graphs. A comparison of the estimates given $v_0 = 0.01$ and $v_0 = 0.1$ shows differences in the number of edges included in each of the three graphs. However, the peculiarities of the single graphs and the differential network are maintained. In particular, the differential network between Europeans and South-Asians highlights similar connectivity patterns, e.g., the amino acid Histidine presents connections with other amino acids and lipids sub-fractions, with these edges being only present in the South-Asian group, as before. Varying the prior values of $\pi_a, \pi_b$ between $\{0.01, 0.5\}$ does not lead to relevant differences.

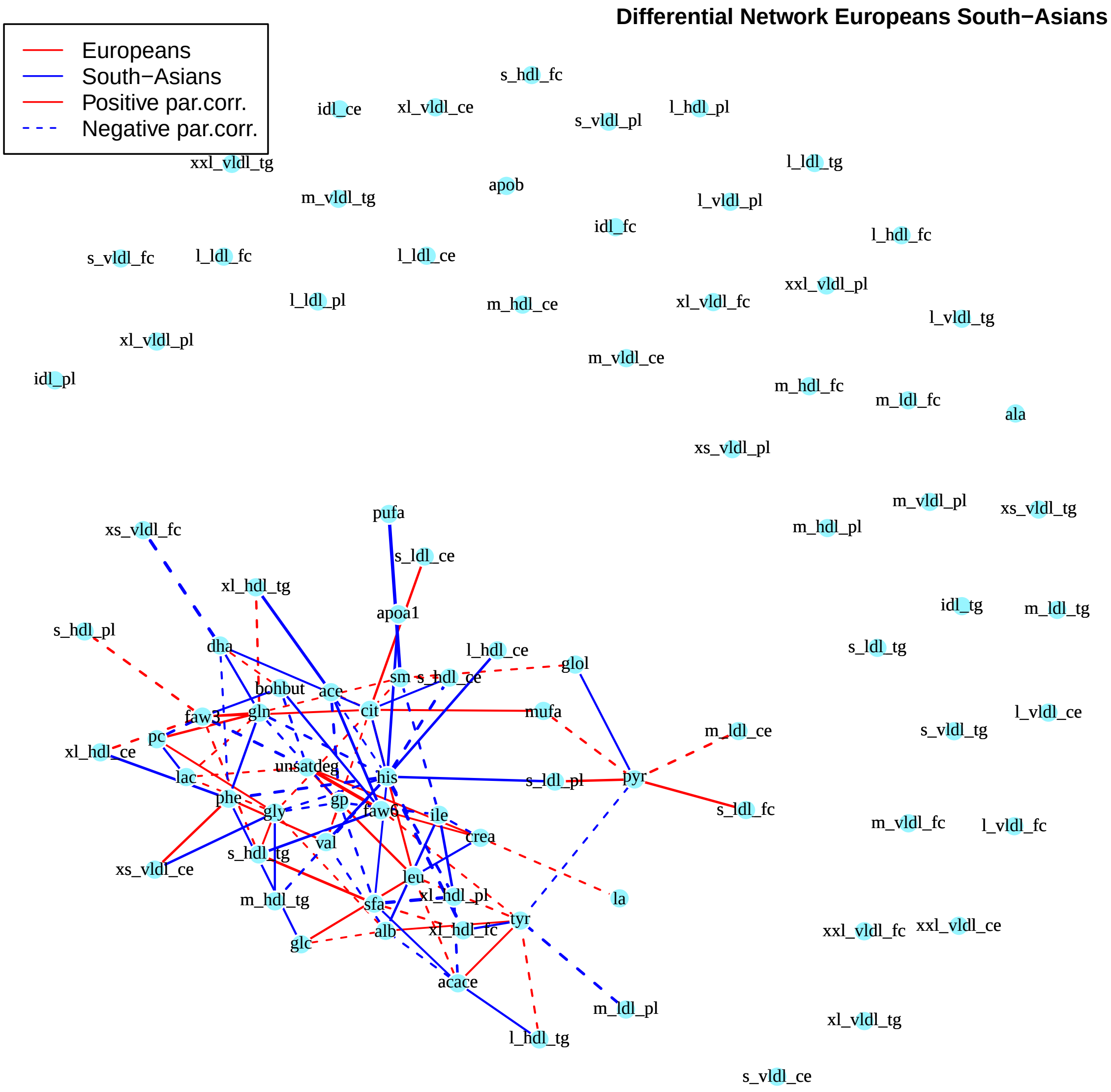


Figure 5.4: Differential network between Europeans and South-Asians. Red edges are only present in the European network, while blue edges are only present in the South-Asian network.

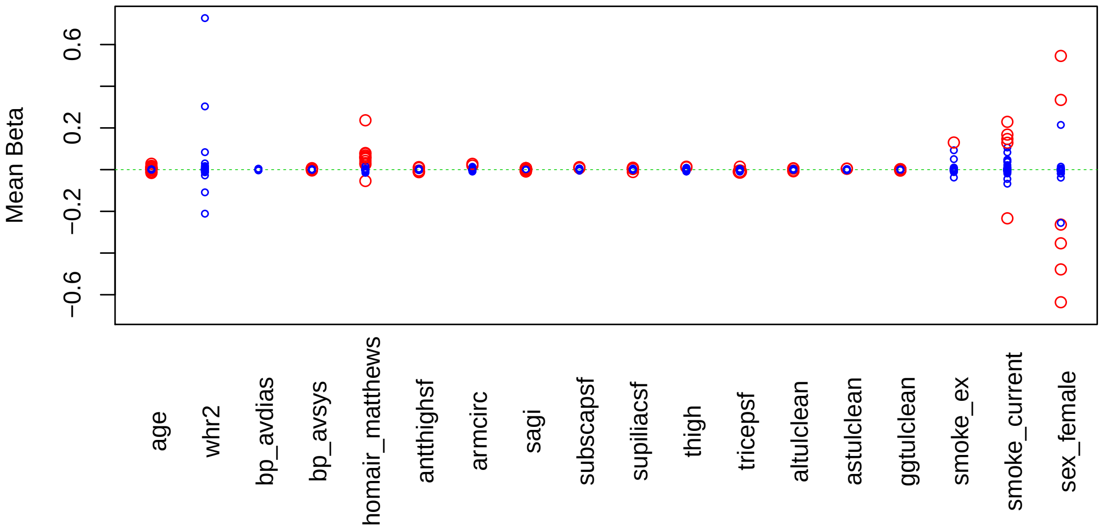


Figure 5.5: Each dot represents the mean of the posterior distribution of a coefficient $\beta_{lj}$. Red dots denote coefficients whose 95% credible interval does not contain the zero.

## 5.4 Conclusions

The model developed in this chapter allows the specification of a desired level of sparsity in the graph and the inclusion of prior information about specific connections between pairs of nodes, when prior knowledge is available, for example, from literature or from expert opinions. We analyse the properties induced on the Graph by the GDP prior in terms of the degree distribution. We demonstrate that this prior is able to capture a wide range of structures, from sparse to more dense graphs. The GDP prior allows us to cluster a posteriori the edges based on their inclusion probabilities. Using an extension of the GDP process, the DGDP, we develop a framework for inference on multiple GGMs. The DGDP offers a convenient way to share information across groups and allows for the possibility to include group specific information in the model. The SSSL prior ensures good scalability of the MCMC thanks to its efficient update scheme and good convergence rates. The SUR model is completed by specifying a global-local shrinkage prior on the coefficients in the mean regression term, allowing each equation to have its own vector of regression parameters and its variable selection. The Horseshoe prior effectively shrinks small and negligible coefficients to zero, while leaving important coefficients unaffected thanks to its heavy tails, as such performing (group specific) variable selection. We illustrate the performance

of the proposed model and compare it with an alternative nonparametric prior on the edge inclusion probabilities (the ANOVA DDP prior) in a simulation study. The results highlight the ability of the model to recover the true underlying structure of the graphs and to correctly identify association between covariates and response.
Finally, we employ the proposed sparse SUR model to analyse the SABRE metabolomics dataset. Our clinical interest focuses on different patterns of metabolite associations within the three ethnicities. Our approach allows us to provide an interpretable set of unique associations patterns which can aid mechanistic understanding of between-group differences in the development of insulin resistance and diabetes and can highlight areas for further research. In doing this, we still correct for potential confounders within the SUR framework. The results obtained from our analysis are promising.

# Chapter 6

# Bayesian Dynamic Multiple Graphical Models

In this final chapter we propose a novel approach for the estimation of multiple Graphical Models to analyse temporal patterns of association among a set of metabolites over different groups of patients. We are interested in identifying potential ethnic differences in metabolite levels and associations as well as their evolution over time, with the aim of gaining a better understanding of different risk of cardio-metabolic disorders across ethnicities. Within a Bayesian framework, we employ Nodewise Regression technique to infer the structure of the graphs, borrowing information across time and ethnicities. The response variables of interest are metabolite levels measured at two time points for two ethnic groups, Europeans and South-Asians. We use Nodewise Regression to estimate the high-dimensional precision matrix of the metabolites, by regressing each metabolite on the remaining and imposing sparsity on the regression coefficients. To this end we assume a time dependent global-local shrinkage prior, the dynamic horseshoe prior, on the regression parameters, which allows us to favour sparse graphs and to accommodate the temporal dimension. Moreover, we extend the prior to analyse multiple groups of observations. Posterior inference is performed through Markov Chain Monte Carlo (MCMC) methods. We also provide code to fit the proposed model using the Bayesian software `Stan`, which implements Hamiltonian Monte Carlo methods. The proposed approach is able to capture a wide range of graph topologies and identify common/group-specific structures across multiple graphs, in our case corresponding to different ethnicities, and allows us to describe temporal trends in metabolic associations.

Most cardiovascular diseases involve disturbances in cardiac metabolism, moreover, disturbances in metabolism that occur in diseases, such as diabetes mellitus, directly impact cardiac metabolism. It is also becoming more clear that heart disease can affect the metabolism, therefore, disturbances in metabolism may initiate a vicious cycle that causes and deteriorate cardiovascular diseases (McGarrah et al., 2018). Metabolomics has emerged as a powerful tool for defining changes in metabolism that occur across a range of cardiovascular disease states. Findings from metabolomics studies have contributed to a better understanding of the metabolic changes that occur in heart failure and ischaemic heart disease and have identified new cardiovascular disease biomarkers. With the advance of technologies, the metabolomics field continues to evolve rapidly and the use of metabolomics can help to better understand the evolution dynamics of specific diseases. Therefore, it is essential not to limit the analysis to differences in metabolite levels across ethnicities over time, but to consider as well patterns of variations in metabolic associations to gain better insight in molecular mechanisms of disease pathogenesis and to formulate novel data-driven scientific hypotheses. We represent association patterns among metabolites through a graph which is the main object of statistical inference. Within a Bayesian framework, our model is based on Nodewise Regression, originally introduced by Meinshausen and Bühlmann (2006). In Nodewise Regression estimating a graph is equivalent to estimating the precision matrix between variables, in our case metabolite levels. This is achieved by rewriting the problem in terms of $M$ independent linear regressions, where $M$ is the number of variables, and each variable is regressed on all the others. It is a local method, because it infers the neighbour structure of each node (i.e. the connections involving the node) independently, as opposed to global methods that aim to infer jointly the association patterns across all the nodes. In Nodewise Regression we model each variable, using the others as predictors. The element $j, l$ in the precision matrix is estimated to be non-zero if either the estimated coefficient of variable $j$ on $l$, or the estimated coefficient of variable $l$ on $j$, is non-zero.

A wealth of proposals is available in the literature to impose sparsity on the regression coefficients. See, Section 2.1.5 for a brief introduction to Bayesian variable selection and shrinkage priors. We opt the Horseshoe prior, which is characterised by an accentuated spike at zero to strongly shrink small or negligible coefficients, while leaving important coefficients unaffected thanks to its heavy tails. Moreover, this prior allows for efficient computations. Employing the dynamic extension of the Horseshoe prior proposed by Kowal et al. (2017), we are able to accurately estimate dynamic evolving complex precision matrices, across multiple groups of observations of different sample sizes at each time point.

A similar approach is taken by Lin et al., 2017, where the neighbour selection of each node is based on the Nodewise Regression method of Meinshausen and Bühlmann, 2006, but it is different from our approach in two ways: *a*) how shrinkage is imposed on the regression coefficients and hence how the edge selection is obtained, *b*) how information is shared across multiple groups and time points. Lin et al., 2017 use a Spike and Slab prior over the regression coefficients, while we adopt the Horseshoe prior which is fully continuous and does not require hyperparameter tuning (in its base version), moreover it allows us to use alternative inference approaches, such as Hamiltonian Monte Carlo, which is applicable only with continuous prior distributions. Lin et al., 2017 introduce additional parameters that force the graphs to be similar between groups and between times, while we share information across multiple groups directly through the hyperparameters of the horseshoe prior. Moreover, we can explicitly incorporate time dependence into the model through the schema proposed by Kowal et al., 2017, imposing a time structure on the linear predictor in the regression component. Both approaches are computationally efficient, being suitable for parallel computations on multicore machines, moreover the approach proposed by Kowal et al., 2017 scales linearly in the number of time points.

A limitation of the neighbourhood selection based on Nodewise Regression, which applies to both approaches, is the lack of a direct posterior estimate of the true precision matrix. As detailed in Lin et al., 2017, the Bayesian version of the Nodewise Regression allows to accurately estimate the true posterior distribution of the edge selection, but does not directly provide an estimate of the precision matrix. Nonetheless, we can use the estimated posterior distribution of the regression coefficients as an approximation of the true value and as a reference to understand the strength and direction of each specific connection in the graph.

## 6.1 Nodewise Regression for graphical models

In this section we explain how to estimate a Graphical Model through Nodewise Regression. We use the same terminology for graphical models introduced in Chapter 2. Let $G = (V, E)$ be an undirected graph, with vertex set $V = (1, \dots, M)$ and edge set $E \subset \{(j, l) \in V \times V : j < l\}$, whose vertices are associated with a $M$-dimensional vector of variables $\boldsymbol{y} = (y_1, \dots, y_M)$, which is assumed to follow a multivariate Normal distribution

$$\boldsymbol{y} \sim \mathrm{N}\left(\boldsymbol{0}, \Omega\right) \tag{6.1}$$

where $\Omega = (\omega_{jl})$ is the $M \times M$ precision matrix. An edge is present between nodes $V_j$ and $V_l$, that is $(j, l) \in E$, if and only if $\omega_{jl} \neq 0$ (Wermuth, 1976; Lauritzen, 1996). If $\omega_{jl} = 0$ (absence of an edge), then $y_j$ and $y_l$ are conditionally independent given the remaining variables $\boldsymbol{y}_{-jl}$, where $\boldsymbol{y}_{-jl}$ denotes the random vector $\boldsymbol{y}$ excluding the $(j, l)$ element. Therefore, estimating the graph $G$ corresponds to estimating the precision matrix $\Omega$. We address the problem of estimating the graph $G$ with the technique proposed by Meinshausen and Bühlmann (2006), Nodewise Regression, that exploits the relation between the partial correlation coefficients and the regression coefficients of a linear regression. Consider the following standard linear regression

$$y_l = \sum_{j \neq l} \beta_{jl} y_j + \varepsilon_l, \qquad \varepsilon_l \sim \mathrm{N}(0, \sigma_l^2) \tag{6.2}$$

where each $\beta_{jl}$ is the regression coefficient that encodes the effect of the variable $y_j$ on the dependent variable $y_l$, for $j \neq l$. Since $\beta_{jl} = \dfrac{-\omega_{jl}}{\omega_{ll}}$ and symmetrically $\beta_{lj} = \dfrac{-\omega_{lj}}{\omega_{jj}}$, then,

$$\omega_{jl} \neq 0 \iff \beta_{lj} \neq 0 \iff \beta_{jl} \neq 0$$

This result can be also derived from the moments of the conditional Normal distribution. Consider the partition where the scalar $y_l$ is the l-th coordinate of $\boldsymbol{y}$, and $\boldsymbol{y}_{-l}$ corresponds to the remaining coordinates. The conditional distribution of $y_l$ given $\boldsymbol{y}_{-l}$ is

$$y_l \mid \boldsymbol{y}_{-l}, \boldsymbol{\mu}, \Omega \sim \mathrm{N}\left(\mu_l - \sum_{j \neq l} \frac{\omega_{jl}}{\omega_{ll}}(y_j - \mu_j), \frac{1}{\omega_{ll}}\right) \tag{6.3}$$

where $\omega_{jl}/\omega_{ll} = \beta_{jl}$, for $j \neq l$. For ease of explanation, in the following sections we assume, without loss of generality, $\boldsymbol{\mu} = \boldsymbol{0}$, except in Section 6.3, where $\boldsymbol{\mu}$ is modelled through a linear predictor. The corresponding graph can then be constructed by applying either the *OR* rule or the *AND* rule. In the first case we draw an edge between nodes $j$ and $l$ if and only if $\beta_{jl} \neq 0$ *or* $\beta_{lj} \neq 0$. If using the *AND* rule, draw an edge between nodes $l$ and $j$ if and only if $\beta_{jl} \neq 0$ *and* $\beta_{lj} \neq 0$. Moreover, given $\beta_{jl}$ and $\sigma_l^2$, an estimate of $\Omega$ can be derived by setting the diagonal elements equal to $1/\sigma_l^2$ and off-diagonal elements equal to $-\beta_{jl}/\sigma_l^2$.

This framework allows us to express the problem of graphical model selection as $M$ independent linear regression problems. Alternative approaches to graphical models estimation are available in literature. Graphical Lasso is a popular global method in both the frequentist and Bayesian domain, based on a penalised maximum likelihood estimator (Friedman et al., 2008) or on double exponential prior (Wang, 2012), respectively. Graphical Lasso has the advantage of ensuring a positive definite estimate of $\Omega$, but requires a greater computational effort and it is less flexible in estimating the individual scaling levels (i.e. the diagonal elements of $\Omega$) compared to the nodewise approach (Janková and Geer, 2018). The possibility to treat each regression separately also allows a straightforward parallel programming strategy. Other Bayesian approaches to graphical model estimation rely on the specification of a suitable prior distribution over the graph space and, conditional on the graph, a prior for the precision matrix is selected. For example, Lenkoski and Dobra (2011) propose the G-Wishart prior, a generalisation of the Hyper-Wishart distribution, that allows to deal with non-decomposable graphs. However, the convergence of associated MCMC algorithms can be slow due to the single edge update and the intractable normalising constant in the marginal posterior that requires numerical approximations. Mohammadi and Wit (2015) develop a more efficient birth-death MCMC algorithm to estimate a single GGM using a G-Wishart prior. In Section 6.2 we compare our approach with the G-Wishart model proposed by Mohammadi and Wit (2015).

### 6.1.1 The Model

Often in the frequentist framework the Lasso (Meinshausen and Bühlmann, 2006) or square-root Lasso (Janková and Geer, 2018) are utilised, among the others, to impose sparsity on elements of the precision matrix, particularly in high-dimensional settings. In the Bayesian framework a wealth of sparse Bayesian regression techniques is available. In a regression context, when performing variable selection, a popular choice is to impose shrinkage priors on the regression coefficients. Typical examples are presented in 2, here we opt for the Horseshoe prior of Carvalho et al. (2010), here defined as

$$\begin{aligned}\beta_j \mid \lambda_j, \tau &\sim \mathrm{N}\left(0, \lambda_j^2 \tau^2\right) \\ \lambda_j, \tau &\sim \mathrm{C}^+(0,1)\end{aligned}$$

where $\tau$ is a global hyper-parameter that shrinks all the parameters towards zero, while $\lambda_j$ is a local hyper-parameter specific to $\beta_j$, a regression parameter in our case, that allows to counterbalance the global shrinkage.

Let $Y$ be a $n \times M$ matrix of observations, where $n$ is the sample size and each column $\boldsymbol{y}_l = (y_{1l}, y_{2l}, \ldots, y_{nl})^T$, for $l = 1, \ldots, M$, contains the measurements of the $l$-th variable. Let $p = M - 1$. The regression model for the $l$-th column can be written as

$$\begin{aligned}\boldsymbol{y}_l \mid \boldsymbol{\beta}_l, \sigma_l^2 &\sim \mathrm{N}\left(X\boldsymbol{\beta}_l, \sigma_l^2 I_n\right) \\ \beta_{jl} \mid \lambda_{jl}, \tau_l &\sim \mathrm{N}\left(0, \lambda_{jl}^2 \tau_l^2\right) \\ \sigma_l^2 \mid a_\sigma, b_\sigma &\sim \text{Inverse-Gamma}\left(a_\sigma, b_\sigma\right) \\ \lambda_{jl} &\sim \mathrm{C}^+\left(0,1\right) \\ \tau_l &\sim \mathrm{C}^+\left(0,1\right)\end{aligned} \tag{6.4}$$

where $X$ is the matrix of explanatory variables corresponding to $Y_{-l}$ (i.e. $Y$ excluding the $l$-th column) and $\boldsymbol{\beta}_l = (\beta_{1l}, \beta_{2l}, \ldots, \beta_{pl})$ is a vector of regression coefficients for the $l$-th regression. Carvalho et al. (2010) define a pseudo inclusion probability parameter $\kappa_{jl}$:

$$\kappa_{jl} = \frac{1}{1 + \mathrm{Var}(\beta_{jl} | \lambda_{jl}, \tau_l)} = \frac{1}{1 + \lambda_{jl}^2 \tau_l^2} \tag{6.5}$$

which is interpretable as the amount of shrinkage towards zero, with $\kappa_{jl} \approx 1$ yielding maximal shrinkage and $\kappa_{jl} \approx 0$ corresponding to minimal shrinkage. Carvalho et al. (2010) compare the performance of the variable selection based on (6.5) (with a threshold level of 0.5) with the explicit variable selection based on Spike and Slab, showing that the posterior selection given by $\kappa$ is consistent with that of the Spike and Slab.

To improve computational efficiency we use the following representation of the standard half-Cauchy distribution (employed by Gelman (2006) and Piironen and Vehtari (2017)). The standard half-Cauchy distribution can be expressed as the product of a standard half-Normal random variable times the square root of an Inverse-Gamma random variable. Let $z \sim \mathrm{N}^+(0,1)$ and $y \sim$ Inverse-Gamma$(1/2, 1/2)$ and define $x = z\sqrt{y}$, then $x \sim \mathrm{C}^+(0,1)$. $\mathrm{N}^+(0,1)$ denotes the standard half-Normal distribution, which is defined as the absolute value of a Normal distribution (Leone et al., 1961). This re-parametrisation can help to avoid divergent transitions in the HMC algorithm (a problem commonly encountered with funnel shaped distributions). Piironen and Vehtari (2017) also allow for a tunable global scale parameter $\tau_l^2$, which can help achieving the desired level of sparsity. Thus the prior distribution on the regression parameters becomes

$$\begin{aligned} \beta_{jl} \mid \lambda_{jl}, \tau_l &\sim \mathrm{N}(0, \lambda_{jl}^2 \tau_l^2) & \lambda_{jl}^a &\sim \mathrm{N}^+(0,1) \\ \sigma_l^2 \mid a_\sigma, b_\sigma &\sim \text{Inverse-Gamma}\left(a_\sigma, b_\sigma\right) & \lambda_{jl}^b &\sim \text{Inverse-Gamma}(1/2, 1/2) \\ \lambda_{jl} &= \lambda_{jl}^a \sqrt{\lambda_{jl}^b} & \tau_l^a &\sim \mathrm{N}^+(0,1) \\ \tau_l &= \tau_l^a \sqrt{\tau_l^b \tau_0} & \tau_l^b &\sim \text{Inverse-Gamma}(1/2, 1/2) \end{aligned} \tag{6.6}$$

where $\tau_0 = \frac{p_0}{p - p_0} \frac{\sigma}{\sqrt{n}}$ and $p_0$ is a prior guess about the number of non-zero coefficients. The choice of $p_0$ is extensively discussed by Piironen and Vehtari (2017) and is often fixed to be the reciprocal of the number of observations.

### Extension to multiple groups

We now extend the Horseshoe prior to allow borrowing information across multiple groups of observations. These group are usually defined by the problem under investigation, for example they might correspond to different biological conditions, disease status, spatial regions. Estimating a single graphical model would lead to an implicit assumption of homogeneity of the underlying graphs across the groups, with a consequent loss of information about their heterogeneity and a consequent high risk of false positives. On the other hand, inferring each graph individually might lead to a loss of power given the reduction in sample size. In our case, groups are defined by ethnicity. Let $R$ be the number of groups and let $Y_r$ be a matrix of dimension $n_r \times M$ containing only the observations belonging to group $r$, with $r = 1, \ldots, R$. We introduce dependence across groups through the global shrinkage parameter $\tau^2$ of the

Horseshoe prior. The model is now defined as follows

$$\begin{aligned} \boldsymbol{y}_{lr} \mid \boldsymbol{\beta}_{lr}, \sigma^2_{lr} &\sim \mathrm{N}\left(X_r \boldsymbol{\beta}_{lr}, \sigma^2_{lr} I_{n_r}\right) \\ \sigma^2_{lr} \mid a_\sigma, b_\sigma &\sim \text{Inverse-Gamma}\left(a_\sigma, b_\sigma\right) \\ \beta_{jlr} \mid \lambda_{jlr}, \tau_{lr} &\sim \mathrm{N}\left(0, \lambda^2_{jlr}\tau^2_{lr}\right) \\ \lambda_{jlr} &= \lambda^a_{jlr}\sqrt{\lambda^b_{jlr}} \\ \tau_{lr} &= \tau^a_l \sqrt{\tau^b_l \tau_{0r}} \end{aligned} \qquad \begin{aligned} \lambda^a_{jlr} &\sim \mathrm{N}^+(0,1) \\ \lambda^b_{jlr} &\sim \text{Inverse-Gamma}(1/2, 1/2) \\ \tau^a_l &\sim \mathrm{N}^+(0,1) \\ \tau^b_l &\sim \text{Inverse-Gamma}(1/2, 1/2) \end{aligned} \tag{6.7}$$

where $X_r$ is a $n_r \times p$ matrix corresponding to $Y_{-lr}$ (i.e. $Y_r$ excluding the $l$-th column) and $\boldsymbol{\beta}_{lr} = (\beta_{1lr}, \beta_{2lr}, \ldots, \beta_{plr})$ is a vector of regression coefficients specific to equation $l$ and group $r$. We exploit the structure of the Horseshoe prior, retaining group specific local shrinkage parameters $\lambda_{jlr}$, while we link together the global shrinkage parameters $\tau_{lr}$ through the common $\tau^b_l$ and $\tau^a_l$. The intuition behind our strategy is justified by the structure of the Horseshoe prior. The global shrinkage parameter $\tau_{lr}$ pulls all the coefficients globally towards zero, while the thick half-Cauchy tails for the local variances $\lambda^2_{jlr}$ allow the important coefficients to escape the global shrinkage (Carvalho et al., 2010). We expect the graphs to have group specific connection patterns, which is why we maintain group specific local shrinkage parameters $\lambda_{jlr}$, allowing edges to escape the global shrinkage independently in each group. We also expect groups to share some common structures and a similar number of connections, therefore we link together the global shrinkage parameters $\tau_{lr}$, allowing borrowing of information about the global level of sparsity of the graphs.

### Extension to multiple time points - Dynamic Horseshoe prior

A natural extension of the model above is to introduce a temporal dimension, which allows joint inference of time dependent data from multiple groups. In this work we consider the evolution over time of the patterns of metabolic associations for two ethnic groups. Here we provide an extension of the Nodewise Regression which enables estimation of time dependent graphs, along with the respective precision matrices. We allow for different sample sizes at each time point and observations stratified in multiple groups. Our goal is to estimate sparse multiple-graphs evolving over time. In summary we introduce two levels of dependence among graphs: between groups and between time points, with the aim to also understand how the differences in associations among groups evolve over time. To this end we extend the model in (6.7) by imposing a time structure over the shrinkage scale parameters of the Horseshoe prior following the approach proposed by Kowal et al. (2017), who introduce a general dynamic prior for sparse dynamic linear regressions, of which the dynamic Horseshoe prior is a special case. Time dependence is introduced specifying a stochastic volatility model on the log-variance of each regression coefficient. Let $t = 1, \ldots, T$ be the time index and let $Y_{rt}$ be a $n_{rt} \times M$ matrix containing only the observations belonging to group $r$ at time $t$. Given the time dependence of the log-variance $h_t$

$$h_{jrt} = \log\left(\tau_{jr}^2 \tau_0^2 / \sqrt{p n_{rt}}\right) + \phi_{jr}\left(h_{jrt-1} - \log\left(\tau_{jr}^2 \tau_0^2 / \sqrt{p n_{rt}}\right)\right) + \log\left(\lambda_{jrt}^2\right) \tag{6.8}$$

the time-dependent multiple-groups Nodewise Regression model becomes, (omitting the equation subscript $l$ for ease of notation):

$$\begin{aligned} \boldsymbol{y}_{rt} \mid \boldsymbol{\beta}_{rt}, \sigma^2_{rt} &\sim \mathrm{N}\left(X_{rt}\boldsymbol{\beta}_{rt}, \sigma^2_{rt} I_{n_r t}\right) & (\phi_{jr}+1)/2 \mid \phi_a, \phi_b &\sim \mathrm{Beta}(\phi_a, \phi_b) \\ \sigma^2_{rt} \mid a_\sigma, b_\sigma &\sim \text{Inverse-Gamma}\left(a_\sigma, b_\sigma\right) & \tau_0 &\sim \mathrm{C}^+\left(0,1\right) \\ \beta_{jrt} &= \beta_{jrt-1} + \gamma_{jrt}\exp(h_{jrt}/2) & \tau_{jr} &\sim \mathrm{C}^+(0,1) \\ \gamma_{jrt} &\sim \mathrm{N}(0,1) & \lambda_{jrt} &\sim \mathrm{C}^+(0,1) \end{aligned} \tag{6.9}$$

where $\phi_{jr}$ is an autoregressive coefficient specific to the $j$-th covariate and $r$-th group, $\tau_0^2$ is a global shrinkage parameter common to all predictors and shared by all groups, $\tau_{jr}^2$ is a predictor specific shrinkage parameter and $\lambda^2_{jrt}$ is a time and covariate specific local shrinkage parameter. $\boldsymbol{\beta}_{rt} = (\beta_{1rt}, \ldots, \beta_{prt})$ is a time and group specific vector of regression parameters and $X_{rt}$ is a $n_{rt} \times p$ matrix corresponding to $Y_{-lrt}$ (i.e. $Y_{rt}$ excluding the $l$-th column). The distribution of the logarithm of the square of a half-Cauchy random variable is a Z-distribution (Kowal et al., 2017). In particular, if $\eta = \log(\lambda^2)$, where $\lambda \sim \mathrm{C}^+(0,1)$, then $\eta$ has probability density function:

$$g(\eta) = \pi^{-1}\exp\left(\eta\right)\left[1 + \exp\left(\eta\right),\right]^{-1}, \quad \eta \in \mathbb{R}$$

The Z-distribution can be represented as a mean-variance mixture of Gaussian distributions (Barndorff-Nielsen et al., 1982) and thanks to the Polya-Gamma expansion proposed by Kowal et al. (2017) we can develop a multiple groups hierarchy similar to the one in model (6.7). To this end we define $\eta_{jrt} = \log\left(\lambda^2_{jrt}\right)$, $\mu_0 = \log\left(\tau_0^2\right)$ and $\mu_{jr} = \log\left(\tau_0^2\tau_{jr}^2\right)$ and we re-write the prior distributions for the parameters in log scale

$$\begin{aligned}
\eta_{jrt} \mid \xi_{\eta_{jrt}} &\sim \mathrm{N}\left(0, \xi_{\eta_{jrt}}^{-1}\right) & \xi_{\eta_{jrt}} &\sim \text{Polya-Gamma}(1,0) \\
\mu_{jr} \mid \mu_0, \xi_{\mu_{jr}} &\sim \mathrm{N}\left(\mu_0, \xi_{\mu_{jr}}^{-1}\right) & \xi_{\mu_{jr}} &\sim \text{Polya-Gamma}(1,0) \\
\mu_0 \mid \xi_{\mu_0} &\sim \mathrm{N}\left(0, \xi_{\mu_0}^{-1}\right) & \xi_{\mu_0} &\sim \text{Polya-Gamma}(1,0)
\end{aligned} \tag{6.10}$$

The global shrinkage parameters $\mu_0$ is shared by all groups, allowing borrowing of information about the global level of shrinkage. This modelling strategy allows us to propagate the shrinkage profile of each regression coefficient over time, allowing fast structural changes or slowly adjusting processes. Kowal et al. (2017) explore the theoretical properties of the dynamic Horseshoe prior and show its good performance when compared to alternative priors. Moreover, the Polya-Gamma expansion in (6.10) leads to efficient computations as it allows to design a fast block-Gibbs sampler.

### 6.1.2 Posterior Inference

For reasonable sized datasets, $M \leq 30$ and sample size $n \leq 1000$, posterior inference for the Nodewise Regression model with Horseshoe prior (static and dynamic) can be performed efficiently in Bayesian software like `Stan` (Carpenter et al., 2017) or `JAGS` (`http://mcmc-jags.sourceforge.net/`). In Appendix we provide sample code to implement the proposed approach in Stan, which implements Hamiltonian Monte Carlo (HMC, Brooks et al. (2011), chapter 5). For larger problems, implementation in a low level language is advisable. Here, we develop a block Gibbs sampling, extending the algorithm provided by Kowal et al. (2017), allowing for multiple groups of different sample sizes to be analysed. Details of the MCMC are provided in Appendix C.

## 6.2 Simulations

In this section we analyse the performance of the proposed models on synthetic datasets. We evaluate the ability of each model to recover the true graph structure $G$ using the Area Under the Curve (AUC), which is a normalised measure of the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is obtained by plotting the true positive rate against the false positive rate evaluated at different thresholds the for edge inclusion probability. We also use the posterior distributions of the Nodewise regression coefficients to give an estimate the precision matrix $\Omega$ and asses its accuracy through the Mean Absolute Error (MAE), calculated between the true matrix and the estimate given by the model. Although we cannot recover the true posterior distribution of the precision matrix, we show that the estimates given by the proposed models are very accurate, when compared to alternative methodologies.

In the first simulation we test the ability of Nodewise Regression to estimate the graph corresponding to a set of highly correlated variables. In order to test the robustness of the model to such scenario, we build a graph with 20 nodes with the following characteristics: $a$) the first ten nodes are all connected, with a correlation coefficient of 0.9, $b$) nodes from 16 to 20 are all connected, with a correlation coefficient of 0.7, $c$) we connect the two groups specifying a correlation of 0.4 between the pairs of nodes $(1, 16)$ and $(2, 18)$. The resulting precision matrix is a block matrix, where nodes 11 to 15 are disconnected, while the other 15 nodes form a dense sub-graph. The partial correlation is interpreted as the correlation between two variables, net of the effect of all the other remaining variables. Therefore, the partial correlation coefficient will decrease as the correlation value and the number of correlated nodes rise. We generate five datasets of $n = 100$ observations each, given the graph described above, and estimate the correspondent graphs and precision matrices with Nodewise Regression and with the G-Wishart model from the `R` package *BDgraph*. We use the AUC to measure the ability to retrieve the true conditional independences and the MAE to measure the actual discrepancy in the partial correlations (compared with the true one). In Figure 6.1 we report the boxplot summarising the AUC and MAE for the five simulations replicates. The AUC for the Nodewise regression is higher than that

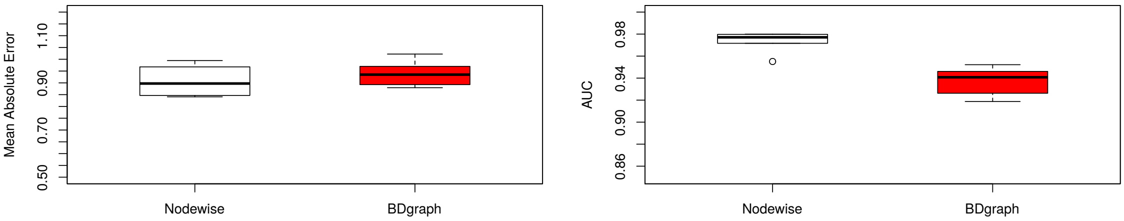


Figure 6.1: Mean Absolute Error (left figure) and AUC (right figure) comparison between the Horseshoe prior model in 6.6 and the G-Wishart model from the package *BDgraph*

of the *BDgraph* package, while, there is not a significant difference between the MAE boxplot. This simulation highlights the ability of the proposed Nodewise regression to be able to correctly estimate the true graph structure in the high correlation scenario described above.

In the second simulation we compare the estimate of the multiple groups model in (6.7) with that of the `R` package *BDgraph*, which implements a birth-death MCMC algorithm for Bayesian structure learning in graphical models. We construct three precision matrices $\Omega_1$, $\Omega_2$ and $\Omega_3$, corresponding to graphs $G_1$, $G_2$ and $G_3$, of $M = 20$ nodes. Following (Peterson et al., 2015), we first define the precision matrix $\Omega_1$ and then derive the others as a perturbation of the first. We set the main diagonal elements of $\Omega_1$ equal to 1, first off-diagonal elements $\omega_{i,i+1} = \omega_{i+1,i} = 0.5$, for $i = 1, \dots, 19$ and second off-diagonal elements $\omega_{i,i+2} = \omega_{i+2,i} = 0.5$, for $i = 1, \dots, 18$. Then we set all $\omega_{i,j} = 0.9$, for $i < j < 6$, while the rest of the elements are set to zero. $\Omega_2$ is derived from $\Omega_1$, setting the second off-diagonal elements $\omega_{i,i+2} = \omega_{i+2,i} = 0$, for $i = 1, \dots, 18$, all the remaining elements being equal. $\Omega_3$ is derived from $\Omega_1$, setting the first off-diagonal elements $\omega_{i,i+1} = \omega_{i+1,i} = 0$, for $i = 1, \dots, 19$, all the remaining elements being equal. The newly created matrices are not positive definite and, therefore, we compute the nearest positive-definite approximation through the `R` function *nearPD* (Higham (2002), from the package Matrix). The precision matrices $\Omega_2$ and $\Omega_3$ constructed with this procedure are a perturbation of $\Omega_1$: as a result they exhibit some common edges and some group specific connections. The number of observations is fixed to $60, 40, 30$, for group 1, 2 and 3 respectively. Each graph is characterised by a dense group of edges on nodes 1 to 6, representing a set of high partial correlations (absolute value of 0.9). In Figure 6.2 we display the boxplot of the MAE and AUC, calculated over ten simulations, respectively with the multiple groups Nodewise model (6.7) and with the *BDgraph* package. The Nodewise Regression model works better in terms of MAE, for which a value closer to 1 denotes an estimate of $\Omega$ close to the true one, and in terms of AUC, for which a value close to 1 denotes a better recovery of the true graph.

In the third simulation scenario we compare the multiple groups dynamic Nodewise model in (6.9) with the static multiple groups Nodewise model in (6.7) (where we assume the three times to be independent). We consider two groups and we construct two matrices $\Omega_{1t_1}$ and $\Omega_{2t_1}$, one for each group at time 1, of $M = 20$ nodes. First we

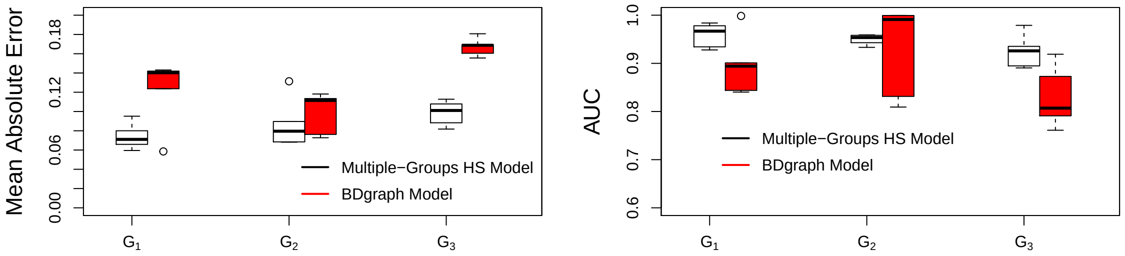


Figure 6.2: Mean Absolute Error (left panel) and AUC (right panel) for the comparison between the multiple groups Nodewise regression model in (6.7) and the package *BDgraph*

set the main diagonal elements of $\Omega_{1t_1}$ equal to 1 and we add 12 non-zero off-diagonal elements, chosen randomly from the $K$ possible edges and setting them equal to 0.5. Then $\Omega_{2t_1}$ is constructed removing 2 edges from $\Omega_{1t_1}$ at random and adding 3 new edges chosen randomly as before. These three new edges are set equal to 0.5. Finally, we simulate the evolution over time of the two precision matrices, removing 2 edges and adding a new edge randomly chosen for each time point (setting the corresponding elements in the precision matrix equal to 0.5) for a total of $T = 3$ time points. In Figure 6.3 we show the network generated with such procedure from which we simulate the dataset. The number of observations is fixed to $50, 40, 30$ respectively for $t_1, t_2$ and $t_3$ (where each group has half of the total sample size at each time point).

In the fourth simulation we construct the dynamic precision matrices following the same procedure as the second scenario, changing the number of time points to $T = 10$. The number of observations is fixed to 40 per time point (equally split between two groups). The generated graphs are characterised by a slowly changing pattern, where only one edge is added or removed at each time point. The results of the comparison between the dynamic and static Nodewise models is shown in Figure 6.5, where we report the boxplots of the AUC calculated over five simulations. The dynamic model has higher and less variable values of AUC in group 1, while there are fewer differences in performance between the models for group 2.

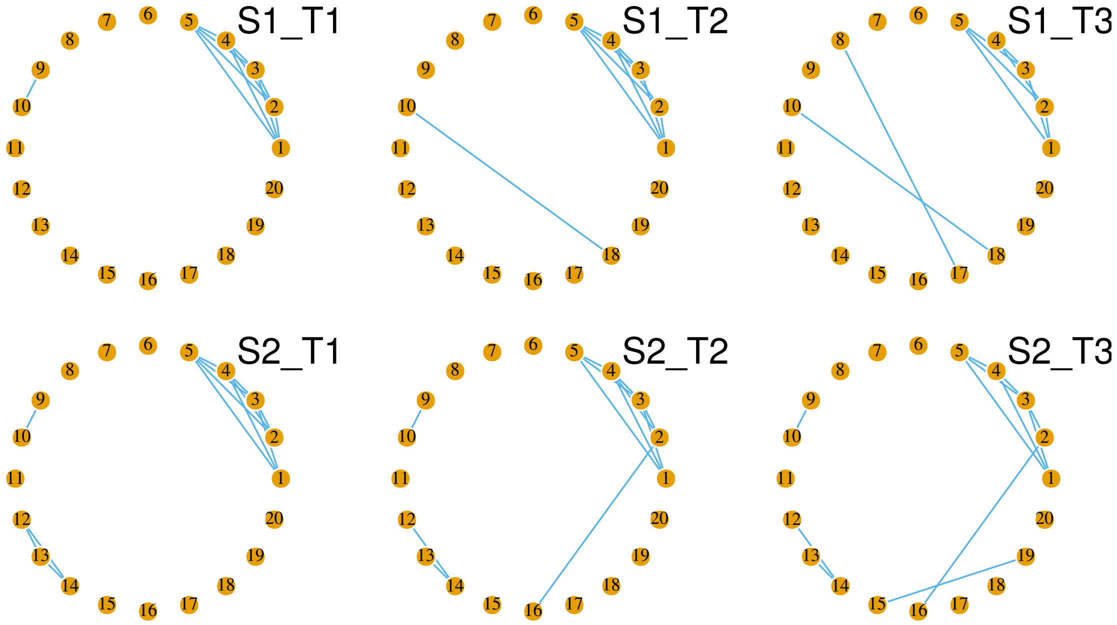


Figure 6.3: Simulation of time dependent graphs

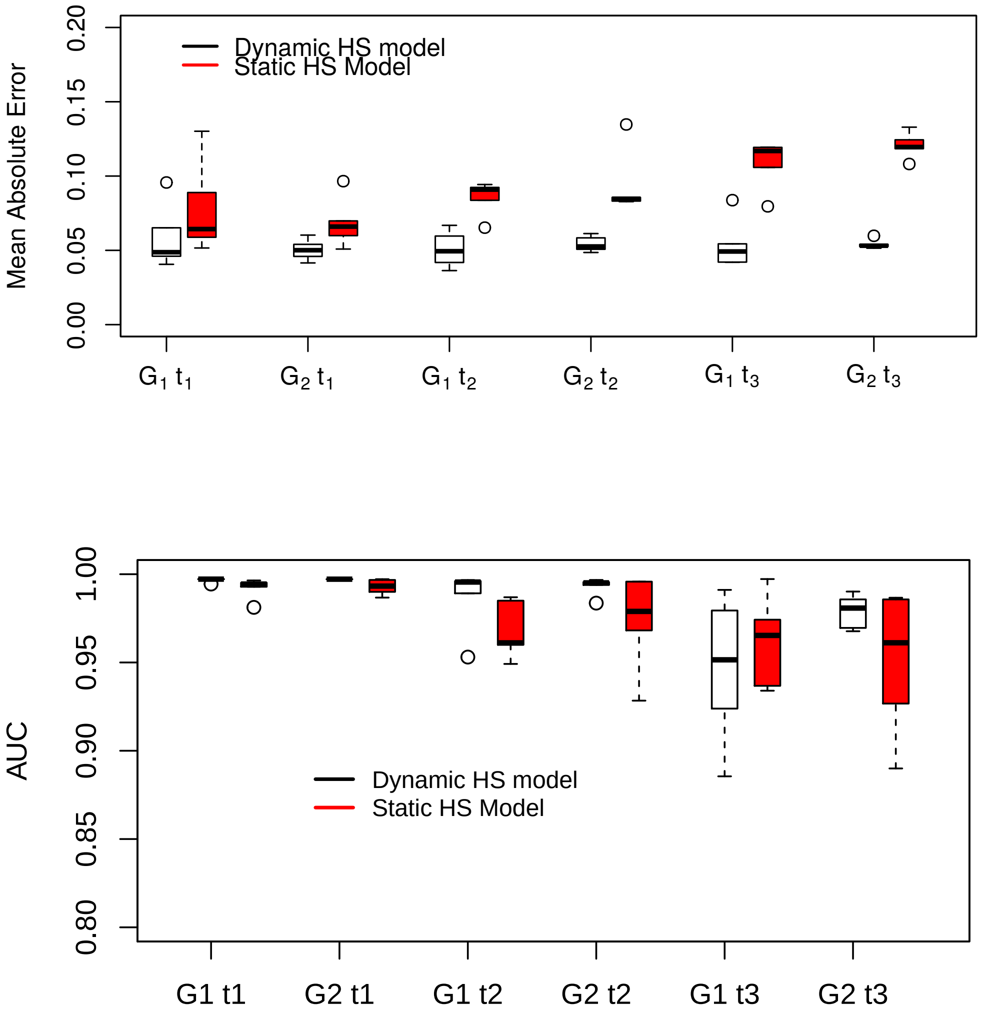


Figure 6.4: Mean Absolute Error (top panel) and AUC (bottom panel) comparison between the dynamic model in (6.9) and the static model in (6.7)

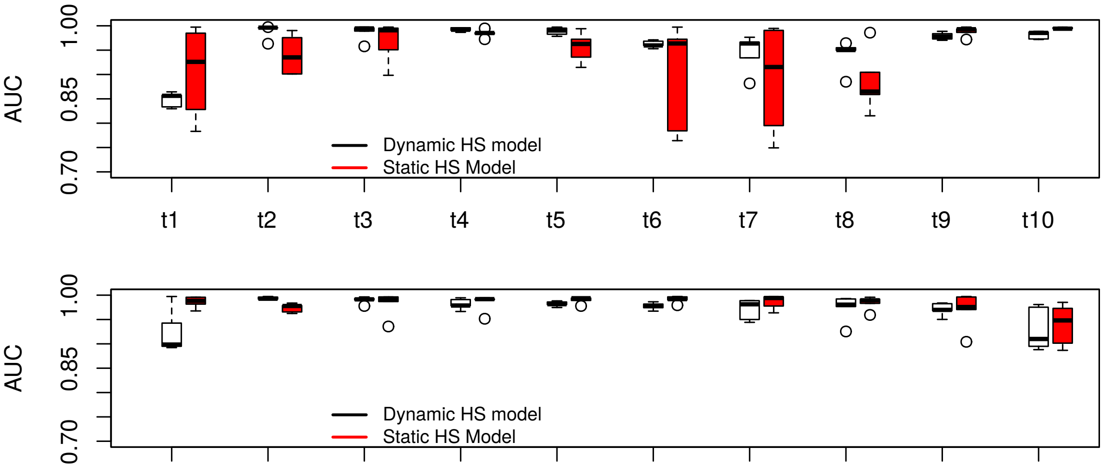


Figure 6.5: AUC comparison for groups 1 and 2 (top and bottom panel respectively) between the dynamic model in (6.9) and the static model in (6.7).

## 6.3 SABRE results

In this section we fit the Nodewise Regression for dynamic multiple graphical models to the SABRE metabolic data. The dataset described in Chapter 1 has a total of 1246 observations at baseline ($T_1$) and 875 at follow-up time ($T_2$). Individuals are stratified in two ethnicities at each time point, 690 Europeans and 556 South-Asians at $T_1$ and 503 Europeans and 372 South-Asians at $T_2$. Nodes on the graph correspond to metabolites and there are a total of $M = 88$ of nodes (i.e. the number of equations in the Nodewise Regression), a list of which can be found in Table D.1. When analysing this data, it is important to control for clinical events of interest (e.g. development of diabetes) that occur before $T_1$ and between $T_1$ and $T_2$. To this end, we model the mean $\mu_l$ of the conditional Normal distribution in (6.3) through a linear predictor and assume that $\mu_{lrt} = Z_{rt}\boldsymbol{\beta}^z_{lrt}$, where $Z$ is a matrix of predictors common to all equations and $\boldsymbol{\beta}^z_{lrt}$ is a $p_z$-dimensional vector of regression coefficients for the mean level. The model is completed by specifying a time dependent structure and a prior distribution on $\boldsymbol{\beta}^z_{lrt}$ as follows

$$
\begin{aligned}
\beta^z_{klrt} &= \beta^z_{klrt-1} + \gamma^z_{klrt} \\
\gamma^z_{klrt} &\sim \mathrm{N}(0, s_0)
\end{aligned}
$$

for $k = 1, \ldots, p_z$, where $s_0$ is the prior variance, here specified to induce a flat Normal distribution. Posterior inference is performed by sampling together the regression coefficients $\boldsymbol{\beta}_{lrt}$ and $\boldsymbol{\beta}^z_{lrt}$. We include the predictors listed in Table D.5, consisting of a measure of body-fat distribution, total blood lipids, blood pressure, control variables, such as smoking habits, sex and age and indicators of the occurrence of cardiovascular diseases and diabetes. We also include an intercept term so that the total number of covariates is $p_z = 20$. We run the MCMC for 10000 iterations, comprising a burn-in period of 2000 iterations and a thinning of 4. In addition to the individual networks we also estimate the differential networks (Fuente, 2010; Valcárcel et al., 2011) arising from the pairwise comparison between the two ethnicities for each time point and the pairwise comparison between $T_1$ and $T_2$ for each ethnicity. A differential network includes all the edges that are present only in one of the two groups/times (i.e. present in one group/time and not the other and vice-versa), thus helping us to understand the main differences between ethnicities and the evolution of the metabolic pathways over time.

In Figures 6.6 and 6.7 we show the differential networks between $T_1$ and $T_2$, respectively for Europeans and South-Asians, where an edge between two nodes is added to the differential graph if the probability of an edge being in one specific graph but not in the other is higher than 0.5. It is worth noticing that there are no edges among the majority of the metabolites in both differential networks, which implies that the presence or absence of those connections are shared by the respective ethnicity at $T_1$ and $T_2$. Moreover, for both ethnic groups, the edges in differential networks derive exclusively from edges present at baseline, but not at follow-up. The connected metabolites in the European differential network (Figure 6.6) belong predominantly to the groups of very low density lipoproteins and high density lipoproteins, with the addition of the metabolite 3-hydroxybutyrate, which is connected to cholesterol esters in xxl-VLDL. 3-hydroxybutyrate is a metabolic intermediate that constitutes about 70% of ketone bodies produced in the liver, mainly from the oxidation of fatty acids released from adipose tissue. Ketone body contribution to the overall energy metabolism in the heart and other tissues increases significantly, for example, after prolonged exercise, fasting periods and low carbohydrate diets. If the release of free fatty acids from adipose tissue exceeds the capacity of tissues to metabolize them, as occurs during insulin deficiency of type I diabetes or less commonly in the insulin-resistant of type II diabetes, severe and potentially fatal diabetic ketoacidosis can occur (Dedkova and Blatter, 2014). The connection between 3-hydroxybutyrate and cholesterol esters in xxl-VLDL is found also in the South-Asian differential network (Figure 6.7), where the cholesterol ester lipoprotein is also connected to the amino acid Alanine and other lipoproteins components. Alanine plays an important role in the Alanine-Glucose cycle, whose alterations that increase the levels of serum alanine aminotransferase are linked to the development of type II diabetes (Lehninger et al., 2005).

In Figures 6.8 and 6.9 we report the differential networks between Europeans and South-Asians, respectively at time $T_1$ and $T_2$, where an edge between two nodes is added to the graph if the probability of an edge being in one ethnic specific graph but not in the other is higher than 0.5. This network presents edges connecting amino-acids and lipoproteins sub-fractions, highlighting potential differences in underlying metabolic processes. To gain a better understanding of the estimated connections and to relate the estimated graph to known metabolic pathways, we conduct a pathway over-representation analysis (ORA) using the online software *MetaboAnalyst* (Chong et al., 2018). We include in the analysis all metabolites that have a connection in the differential network. ORA evaluates statistically the fraction of metabolites in a particular pathway found among the user-specified set of metabolites, in our case the metabolites with connections in the differential network. For each pathway, input metabolites that are part of the pathway are counted. Next, every pathway is tested for over or under-representation in the list of input metabolites using the hypergeometric test. The most represented pathways are the ones with smaller p-value and higher number of over-represented metabolites. Here we discuss the first two top-ranked pathways, *Pyruvate Dehydrogenase Deficiency (E3)* and *Pyruvate Carboxylase Deficiency.* Pyruvate dehydrogenase and pyruvate carboxylase deficiency are the most common disorders in pyruvate metabolism. Pyruvate dehydrogenase (PDH) is an enzyme complex consisting of three subunits, pyruvate dehydrogenase, dihydrolipoamide acyltransferase and dihydrolipoamide dehydrogenase (known as E3). This enzyme complex converts pyruvate into acetyl-CoA, an essential substrate in the production of citrate, whose limited production leads to a block in the tricarboxylic acid cycle and other metabolic pathways need to be stimulated to produce acetyl-CoA. The deficiency causes more pyruvate to be metabolized to lactate and leads to lactic acidosis (Bissonnette and Bissonnette, 2006). Overall, PDH Complex plays also a key role in regulating the supply of adenosine triphosphate during the feed-fast cycle, where cells must select fatty acid or glucose as energy source, therefore it is important in regulating the glucose metabolism, consequently PDH deficiency is related to metabolic diseases, e.g. type 2 diabetes and obesity (Lee, 2014). Of particular interest is pyruvate carboxylase deficiency. Lao-On et al. (2018) explore the roles of pyruvate carboxylase in human diseases, such as diabetes. Pyruvate carboxylase (PC) is an anaplerotic enzyme which plays an essential role in various cellular metabolic pathways, including gluconeogenesis, fatty acid synthesis, amino acid synthesis, and glucose-induced insulin secretion. In aerobic conditions, Pyruvate enters mitochondria via the mitochondrial pyruvate carrier, where may be further metabolized in two ways. In non-gluconeogenic tissues, such as muscles and brain, Pyruvate is decarboxylated to form acetyl-CoA catalysed by the Pyruvate Dehydrogenase Complex, while in gluconeogenic tissues, where pyruvate carboxylase is highly abundant, most of Pyruvate entering mitochondria is carboxylated by this enzyme to form oxaloacetate. Given the importance of oxaloacetate in various biochemical pathways, alterations of oxaloacetate production by PC can produce serious diseases such as type 2 diabetes and neurological disorder.

The third and fourth ranked pathways are related to seizures disorders, which are found to be associated with abnormal glucose levels, whether too high or too low. The problem is particularly relevant to individuals with diabetes, whose blood glucose levels can vary widely over the course of a day, as a result of the disease, variations in insulin levels, or other metabolic factors. Clinical studies show that adults with hyperglycaemia have an increased tendency to experiencing seizures (Stafstrom, 2003). These findings suggest plausible metabolic pathways which may be disordered to a greater extent in South-Asians and contribute to their excess risk of diabetes and cardiovascular disease.

The differential network between Europeans and South-Asians for the follow-up time has more edges compared to the baseline network, with new connections between amino-acids and lipoproteins sub-fractions, implying a greater difference between the ethnicities. To gain a better understanding of the network we proceed as before, conducting a pathway over-representation analysis including all nodes involved in connections in the differential network at $T_2$. The first ranked pathway is related to acute seizures, while the second and third pathways are again related to the disorders in the pyruvate metabolism (*Pyruvate Dehydrogenase Deficiency (E3)* and *Pyruvate Carboxylase Deficiency*).

In Appendix we report the individual networks for the two ethnicities at baseline and follow-up in Figures E.11, E.12, E.13 and E.14, respectively for Europeans at $T_1$ and $T_2$ and South-Asians at $T_1$ and $T_2$. These networks are characterised by a high number of edges, in particular, we can notice a very highly connected group of lipoproteins in all four graphs. In Figures E.15, E.16, E.17 and E.18 we plot the posterior mean of the regression coefficients $\beta_{jlrt}$, for each equation $l$ and covariate $j$, grouped by ethnicity and time. The measure of body-fat distribution WHR has a negative effect on many metabolites for both ethnicities, particularity at $T_1$, while a few metabolites are affected at $T_2$. Blood lipids (triglycerides, cholesterols) and HDL are important for both groups and time periods. The presence of diabetes, or diabetes treatment, also affects the mean level of some metabolites, in particular in Europeans. HOMA IR has an effect on an elevated number of metabolites at $T_1$ and $T_2$ in both Europeans and South-Asians. Overall, HOMA IR, blood lipids and serum

HDL are the control variables that have more significant effects (see the 95% credible region) on the metabolites. High blood triglycerides and low HDL are among the risk factors that determine the metabolic syndrome (Roberts et al., 2013), which can lead to the development of type 2 diabetes. In summary, these findings highlight the presence of complex interplays between metabolic processes, anthropometric factors and clinical markers, which can have different impacts on the risk of diabetes and other cardiovascular diseases across ethnicities and across time.

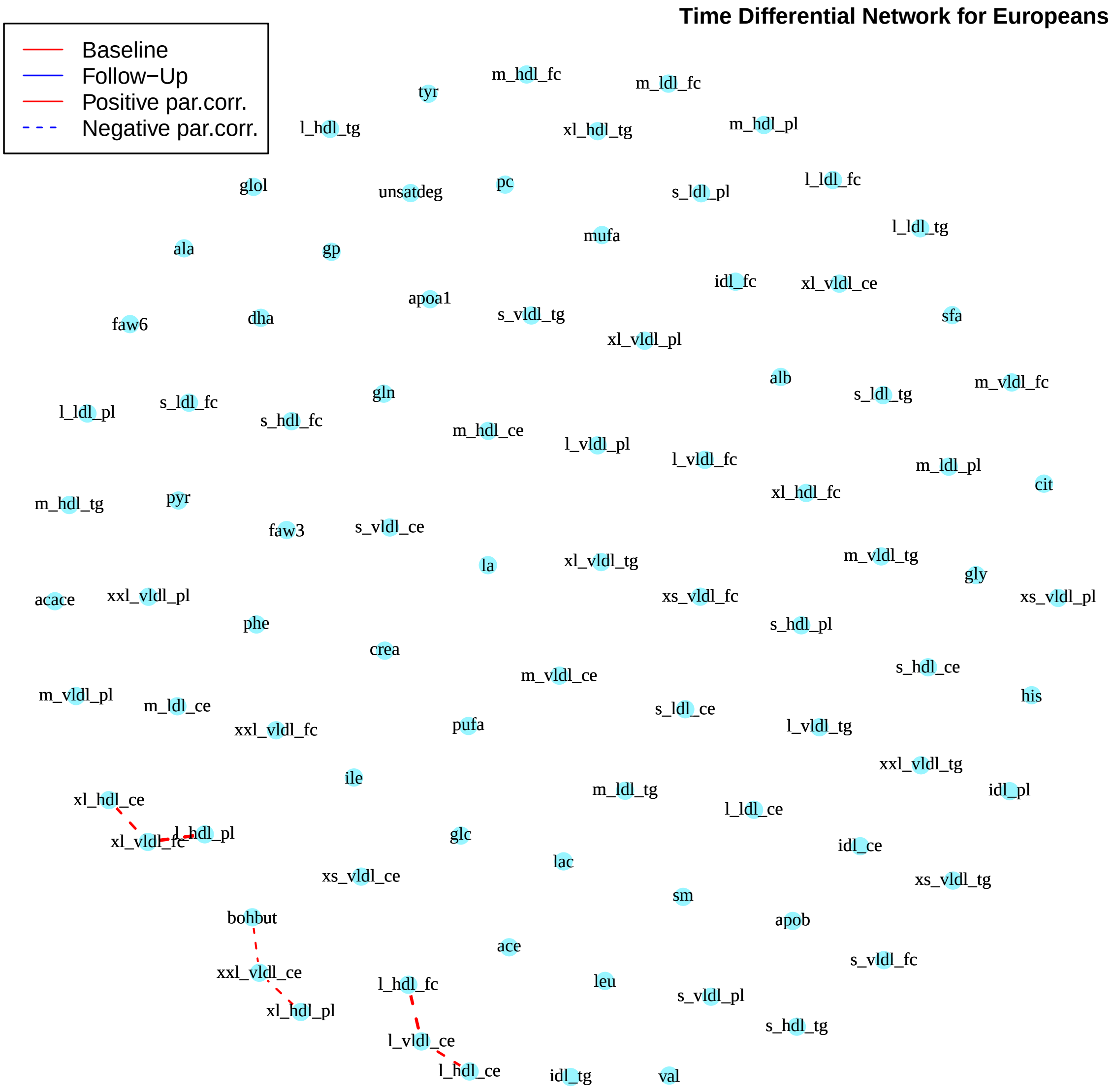


Figure 6.6: European Differential network between baseline and follow-up. Red edges are only present in the baseline network, while blue edges are only present in the follow-up network.

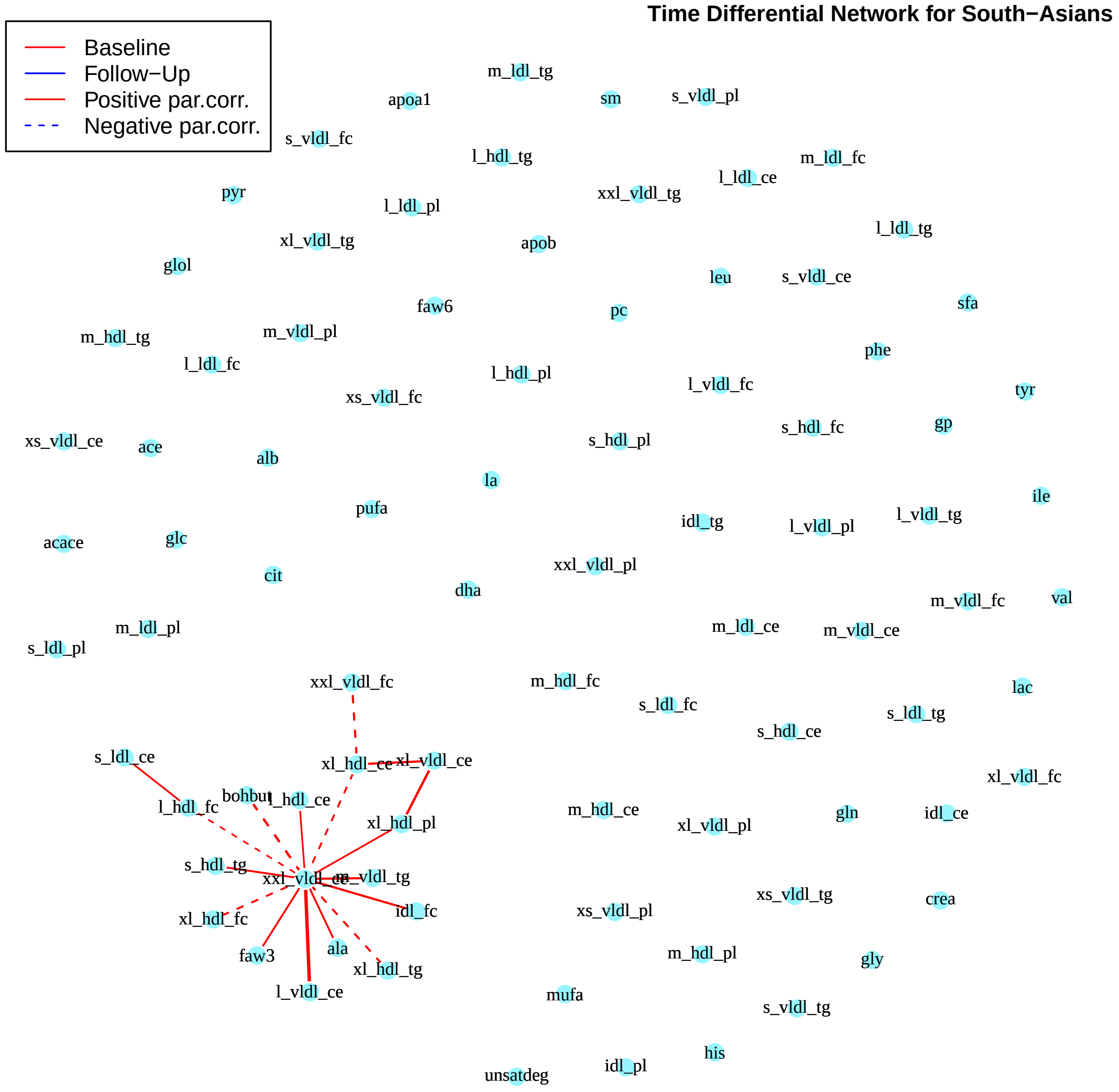


Figure 6.7: South-Asians Differential network between baseline and the follow-up. Red edges are only present in the baseline network, while blue edges are only present in the follow-up network.

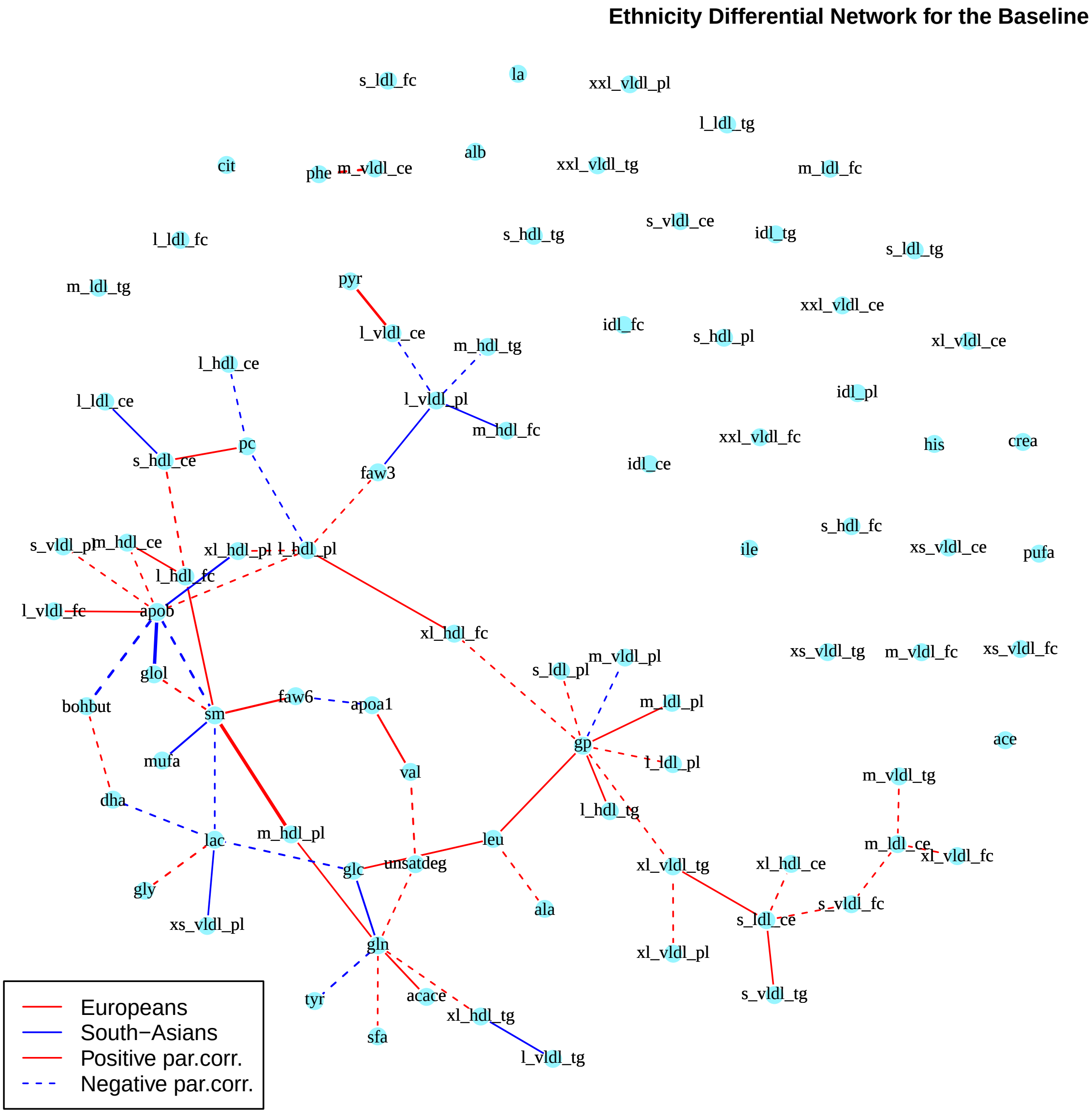


Figure 6.8: Differential network between Europeans and South-Asians at baseline. Red edges are only present in the European network, while blue edges are only present in the South-Asian network.

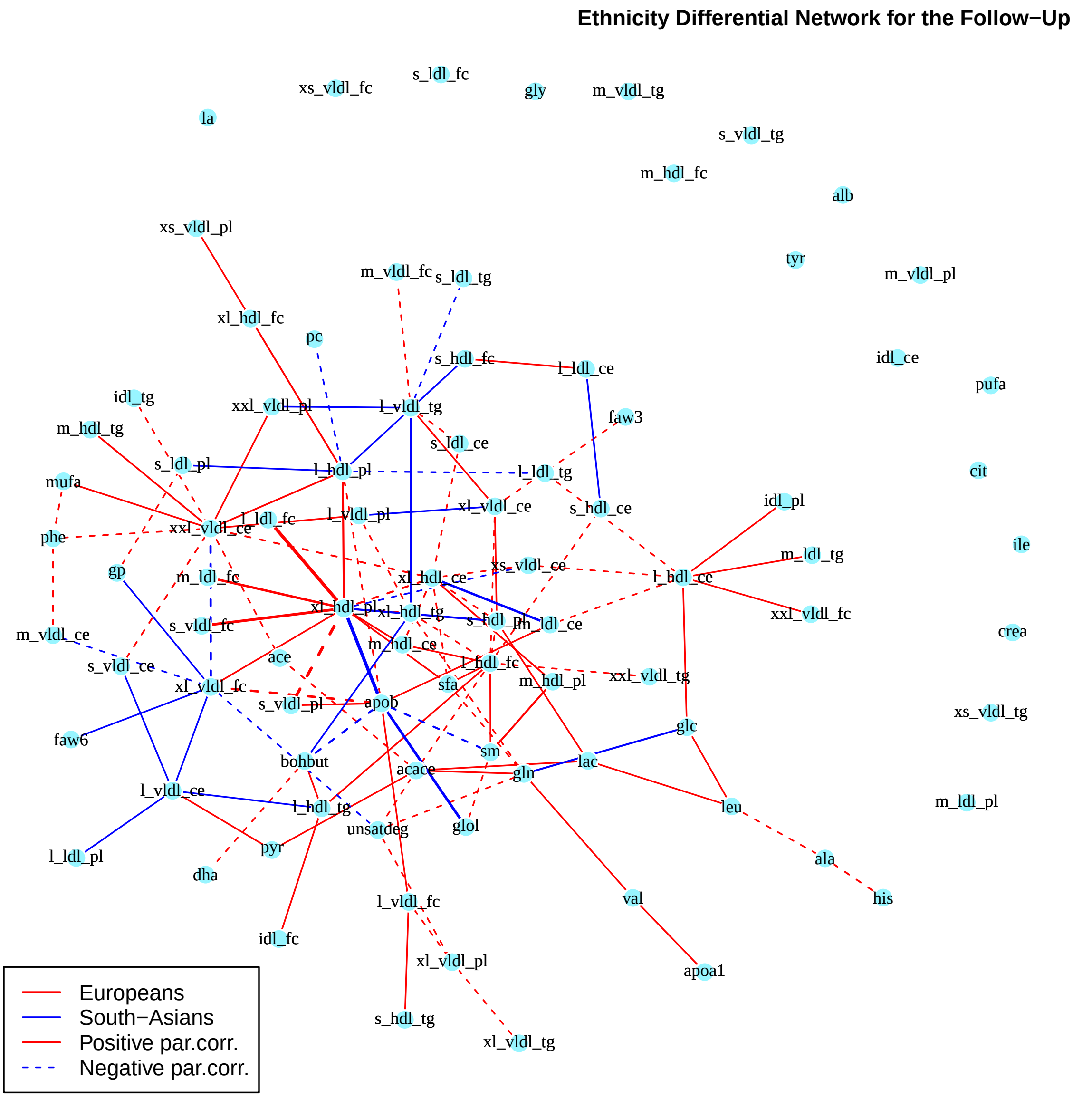


Figure 6.9: Differential network between Europeans and South-Asians at follow-up. Red edges are only present in the European network, while blue edges are only present in the South-Asian network.

## 6.4 Conclusions

In this final contribution we extend Nodewise Regression technique to infer dynamic evolving multiple graphs. The model allows to analyse multiple groups of different sample sizes observed at multiple time points, allowing borrowing of information across time and groups. The flexibility of the model permits to impose regularisation on the regression coefficients and the inclusion of prior information about specific connections between pairs of nodes, when prior knowledge is available. The structure of Nodewise Regression ensures good scalability of the MCMC thanks to the possibility to infer each regression independently. The Horseshoe prior effectively shrinks small and negligible coefficients to zero (inducing sparsity in the graph), while leaving important coefficients unaffected due to its heavy tails, as such performing (group and time specific) variable selection. We illustrate the performance of the proposed model in a simulation study and compare it with an alternative Bayesian model for graph estimation. The results highlight the ability of the model to recover the true underlying structure of the graphs and to accurately estimate the corresponding precision matrices. Finally, we employ the proposed dynamic model to analyse metabolic data from the SABRE cohort study, an information rich dataset on cardiovascular and metabolic diseases. Our clinical interest focuses on different patterns of metabolite associations which characterise the European and South-Asian ethnicities and their evolution over time, from the baseline visit to the follow-up. Our approach allows us to provide an interpretable set of unique associations patterns which can aid mechanistic understanding of between-group and between-times differences in the development of insulin resistance, diabetes and cardiovascular diseases and have the potential to highlight areas for further research. In doing this, we correct for potential confounders and clinical events that would alter the metabolites levels. The results obtained from our analysis are promising.

# Chapter 7

# Discussion and Conclusion

We discuss here the main findings and contributions of this work and we describe some open research questions and further developments of the proposed models.

## 7.1 Summary of the main findings and contributions

In Chapter 3 we present some preliminary network analysis as well as a variable selection on each metabolite. The differential networks highlight the diverse structure of connections for individuals with high levels of HOMA IR compared to those with low levels. These differences suggest changes in the balance of amino acid and in processes of glycogenesis and ketogenesis for energy provision in the fasting state in insulin-resistant Europeans. In South-Asians there is a significant change in the correlation between leucine, an essential branched chain, and ketogenic amino acid, and LDL cholesterol suggesting an alteration in leucine's effect on lipid metabolism in South-Asians affected by Insulin Resistance. The variable selection, performed for each metabolite independently, points out the importance of ethnic differences, in particular between Europeans and South-Asians and the relevance of HOMA IR as well as the liver health indicators ALT and AST.

The first contribution, presented in Chapter 4, involves the study of the distribution of HOMA IR over the three ethnicities, conditionally on the metabolic profile of each individual. By adopting a Bayesian nonparametric approach, we are able to obtain data-driven clustering of the observations, highlighting the presence of subpopulations in the SABRE data, with a multi-ethnic composition, characterised by different levels of HOMA IR, which can cause different risk of development of type 2 diabetes. Moreover, we regularise the estimation of the regression by performing variable selection on the covariates with a Spike and Slab prior, pointing out a group of metabolites and anthropometric covariates that predict HOMA IR. The proposed model gives us promising results, for example, identifying clusters, characterised by high levels of tyrosine, alanine, ALT and subcutaneous adiposity. The modelling approach has the potential to highlight areas for further research, such as the investigation of metabolic specific profiles that can lead to a greater risk of onset of type 2 diabetes and replications in bigger samples with a formal pathway analysis.

The second part of the thesis, which includes Chapters 5 and 6, investigates the analysis of the metabolic interactions, their differences across ethnicities and across time. In Chapter 5 we build our second contribution with a Bayesian nonparametric model to infer multiple graphs in a Sparse SUR framework, where we control for confouders of interest, while estimating the underlying graphs. The nonparametric prior that we adopt over the edges inclusion probabilities allows great flexibility and imply borrowing of information across multiple groups (in our case corresponding to ethnicities). The Sparse SUR model is completed with the adoption of a continuous shrinkage prior, i.e. the Horseshoe prior, on the coefficients associated with the regression covariates, which ensures good computational scalability and a successful regularisation. In order to highlight the key differences between graphs we estimate the pairwise differential networks and we analyse them through a dedicated software. We conduct a pathway enrichment analysis, in order to put our findings in a more general metabolic context and understand if the highlighted connections are indicator of diseases of interest.

Finally, in our third contribution (Chapter 6), we build a more general model through which we can estimate multiple dynamic graphs, allowing borrowing of information across times and groups. We develop a Bayesian model, merging together the technique of Nodewise Regression (Meinshausen and Bühlmann, 2006) and the dynamic shrinkage process of Kowal et al. (2017), extending them to allow multiple groups of different sample sizes to be analysed. Nodewise Regression together with the dynamic Horseshoe prior ensures great scalability, both over the number of time points and the number of nodes in the graph. The resulting technique can be used to infer dynamic graphs in high-dimensional datasets, while adjusting for confunders of interest. As

for the static model, we put our findings in a more general metabolic context, analysing the differential networks between times and between ethnicities with the online software *MetaboAnalyst*.
The SSSL algorithm adopted in Chapter 5 has the advantage of ensuring a positive definite estimate of the precision matrices, conditionally on the graphs, while retaining a good scalability for medium sized models (in terms of number of nodes). A limitation of this approach is the necessity to specify a priori a threshold for the discrete mixture of Normal distributions in (5.5). We test the goodness of our choice through a sensitivity analysis, repeating the estimation of the model for different values of the threshold, confirming that the specified value of $\nu_0 = 0.01$ gives a good compromise between complexity and ease of interpretation. On the other hand, the Nodewise Regression model in Chapter 6 is based solely on continuous priors, therefore it has a great scalability to graphs of higher dimension and it does not require the choice of a threshold during inference. The inclusion or exclusion of an edge in the graph is based on a functional of the Horseshoe prior that possesses some desirable properties (Carvalho et al., 2010). Moreover, the proposed model based on Nodewise Regression with Horseshoe prior can be more easily extended to more complex scenarios.

In conclusion, this thesis originates from a collaboration with the SABRE research team and focuses on a thorough analysis of the metabolomics SABRE dataset, with an interest in understanding the implications of alterations in the human metabolism on the development of Insulin Resistance, and hence development of type 2 diabetes. Moreover, given the multi-ethnic nature of SABRE, we want to understand how the onset of diabetes and cardio-vascular diseases differs across ethnicities and the differences in the underlying metabolic processes. The findings that we expose will be used as a starting point for further hypothesis generation and in-depth analysis by the epidemiologists and clinicians working in SABRE. The availability of `R` routines will be beneficial for the work of the team and for that of other researchers.

## 7.2 Open research questions

Some open research questions are left that we think would constitute relevant extensions or improvements of the proposed models.

- Investigate the performance of alternative nonparametric priors on edge inclusion probabilities in the multiple GGMs model in Chapter 5. A relatively newer type of Bayesian nonparametric prior is the class of normalised completely random measures (Kingman, 1993). It would be interesting to asses the effect of choosing this different prior on the estimation of multiple graphs.

- Extend the multiple GGMs model to allow the automatic detection of multiple graphs on the basis of a continuous covariate. In our contribution we define groups of observation according to a categorical variable of interest (ethnicity in our case), which naturally defines sub-samples in the data. The automatic identification of sub-groups of observation would allow the discovery of differences in the patterns of associations of metabolites that correspond to particular values of a variable of interest.

- The dynamic model in Chapter 6 allows the estimation of multiple graphs over multiple time points efficiently, thanks to the scalability of the Horseshoe shrinkage prior. The extension to non-Gaussian and mixed data would provide an even more general modelling strategy that would allow the joint analysis of continuous and discrete variables. An example of graphical model for mixed data, for a single graph, is given by Dobra and Lenkoski (2011). Haslbeck and Waldorp (2015) provide a model to infer dynamic mixed graphical models (for a single group) in a time-series framework.

There are other real data applications that would be interesting to study. First of all the application of predictive Bayesian models to outcomes of interest in the SABRE study, such as diabetes, stroke and coronary heart disease. The implications of alterations in metabolic processes on the development of cardiovascular diseases is an ongoing area of active research, therefore the detailed study of these variables would be beneficial to the metabolomics research.

The SABRE study encompasses a rich collection of variables. In this thesis we concentrated on metabolomics, but other *'omics* data are available, of which genomics is a major one. The inclusion in our analysis of genomic data could also lead to employ causal inference techniques, in fact genomic data can be used to asses a causal effect through the technique of Mendelian Randomisation (Davey Smith and Ebrahim, 2003).

Finally, the SABRE team is collecting at present date data from the second follow-up. The second follow-up would be extremely informative and useful for the analyse of the evolution pattern of the metabolic interactions for a longer time period and to eventually perform a formal survival analysis.

# Bibliography

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. In: *The Annals of Statistics*, **2**, 1152–1174.

Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. In: *Biometrika*, **92**,2, 317–335.

Barcella, W., De Iorio, M., Favaro, S. and Rosner, G. L. (2017). Dependent generalized Dirichlet process priors for the analysis of acute lymphoblastic leukemia. In: *Biostatistics.*

Barndorff-Nielsen, O., Kent, J. and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. In: *International Statistical Review/Revue Internationale de Statistique*, 145–159.

Bhadra, A., Datta, J., Polson, N. G., Willard, B. et al. (2017). The horseshoe+ estimator of ultra-sparse signals. In: *Bayesian Analysis*, **12**,4, 1105–1131.

Billio, M., Casarin, R. and Rossini, L. (2017). Bayesian nonparametric sparse seemingly unrelated regression model (SUR). In: *Available at SSRN 2832728.*

Binder, D. A. (1978). Bayesian Cluster Analysis. In: *Biometrika*, **65**, 31–38.

Bissonnette, B. and Bissonnette, B. (2006). *Syndromes: rapid recognition and perioperative implications.* McGraw-Hill New York, NY.

Bonnet, F., Ducluzeau, P.-H., Gastaldelli, A., Laville, M., Anderwald, C. H., Konrad, T., Mari, A., Balkau, B., Group, R. S. et al. (2011). Liver enzymes are associated with hepatic insulin resistance, insulin secretion, and glucagon concentration in healthy men and women. In: *Diabetes*, **60**,6, 1660–1667.

Brooks, S., Gelman, A., Jones, G. and Meng, X.-L. (2011). *Handbook of markov chain monte carlo.* CRC press.

Brown, P. J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**,3, 627–641.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017). Stan: A Probabilistic Programming Language. In: *Journal of Statistical Software, Articles*, **76**,1, 1–32. ISSN: 1548-7660. DOI: `10.18637/jss.v076.i01`. URL: `https://www.jstatsoft.org/v076/i01`.

Carvalho, C. M. and Scott, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. In: *Biometrika*, **96**,3, 497–512.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. In: *Biometrika*, **97**,2, 465–480.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. In: *The American Statistician*, **46**,3, 167–174.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D. S. and Xia, J. (2018). MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. In: *Nucleic Acids Research*, gky310. DOI: `10.1093/nar/gky310`. eprint: `/oup/backfile/content_public/journal/nar/pap/10.1093_nar_gky310/1/gky310.pdf`. URL: `http://dx.doi.org/10.1093/nar/gky310`.

Davey Smith, G. and Ebrahim, S. (Feb. 2003). ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease?*. In: *International Journal of Epidemiology*, **32**,1, 1–22. ISSN: 0300-5771. DOI: `10.1093/ije/dyg070`. eprint: `http://oup.prod.sis.lan/ije/article-pdf/32/1/1/11213726/320001\_dyg070.pdf`. URL: `https://doi.org/10.1093/ije/dyg070`.

De Iorio, M., Müller, P., Rosner, G. and MacEachern, S. (2004). An ANOVA Model for Dependent Random Measures. In: *Journal of the American Statistical Association*, **99**, 205–215.

Dedkova, E. N. and Blatter, L. A. (2014). Role of $\beta$-hydroxybutyrate, its polymer poly-$\beta$-hydroxybutyrate and inorganic polyphosphate in mammalian health and disease. In: *Frontiers in physiology*, **5**, 260.

Dempster, A. P. (1972). Covariance selection. In: *Biometrics*, 157–175.

Dobra, A., Lenkoski, A. et al. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. In: *The Annals of Applied Statistics*, **5**,2A, 969–993.

Ehtisham, S, Hattersley, A., Dunger, D. and Barrett, T. (2004). First UK survey of paediatric type 2 diabetes and MODY. In: *Archives of Disease in Childhood*, **89**,6, 526–529.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. In: *The Annals of Statistics*, **1**, 209–230.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. In: *Biostatistics*, **9**,3, 432–441.

Fuente, A. De la (2010). From 'differential expression'to 'differential networking'–identification of dysfunctional regulatory networks in diseases. In: *Trends in genetics*, **26**,7, 326–333.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. In: *Journal of the American statistical association*, **85**,410, 398–409.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). In: *Bayesian analysis*, **1**,3, 515–534.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In: *IEEE Transactions on pattern analysis and machine intelligence* 6, 721–741.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. In: *Journal of the American Statistical Association*, **88**, 881–889.

— (1997). Approaches for Bayesian variable selection. In: *Statistica Sinica*, **7**, 339–373.

Gilman, S. (2011). *Neurobiology of disease.* Elsevier.

Giudici, P. and Green, P. (1999). Decomposable graphical Gaussian model determination. In: *Biometrika*, **86**,4, 785–801.

Haslbeck, J. M. and Waldorp, L. J. (2015). mgm: Structure Estimation for time-varying Mixed Graphical Models in high-dimensional Data. In: *J Stat Softw.*

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. In: *Biometrika*, **57**,1, 97–109.

Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. In: *IMA journal of Numerical Analysis*, **22**,3, 329–343.

Hjort, N. L. (2000). Bayesian analysis for a generalised Dirichlet process prior. In: *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420.*

Hoff, P. D. (2009). *A first course in Bayesian statistical methods.* Springer Science & Business Media.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. In: *Journal of the American Statistical Association*, **96**,453, 161–173.

Janková, J. and Geer, S. van de (2018). Inference in high-dimensional graphical models. In: *arXiv preprint arXiv:1801.08512.*

Kastner, G. and Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. In: *Computational Statistics & Data Analysis*, **76**, 408–423.

Kingman, J. (1993). Poisson Processes. In: *Oxford University Press.*

Kowal, D. R., Matteson, D. S. and Ruppert, D. (2017). Dynamic shrinkage processes. In: *arXiv preprint arXiv:1707.00763.*

Lao-On, U., Attwood, P. V. and Jitrapakdee, S. (2018). Roles of pyruvate carboxylase in human diseases: from diabetes to cancers and infection. In: *Journal of Molecular Medicine*, **96**,3-4, 237–247.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. In: *Journal of Computational and Graphical Statistics*, **16**,3, 526–558.

Lauritzen, S. (1996). *Graphical Models.* Oxford Statistical Science Series. Clarendon Press. ISBN: 9780191591228. URL: `https://books.google.co.uk/books?id=mGQWkx4guhAC`.

Lee, I.-K. (2014). The role of pyruvate dehydrogenase kinase in diabetes and obesity. In: *Diabetes & metabolism journal*, **38**,3, 181–186.

Lehninger, A. L., Nelson, D. L., Cox, M. M., Cox, M. M. et al. (2005). *Lehninger principles of biochemistry.* Macmillan.

Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. In: *Journal of Computational and Graphical Statistics*, **20**,1, 140–157.

Leone, F., Nelson, L. and Nottingham, R. (1961). The folded normal distribution. In: *Technometrics*, **3**,4, 543–550.

Lin, Z., Wang, T., Yang, C. and Zhao, H. (2017). On joint estimation of Gaussian graphical models for spatial and temporal data. In: *Biometrics*, **73**,3, 769–779.

Lo, A. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. In: *The Annals of Statistics*, **12**, 351–357.

MacEachern, S. N. (1999). Dependent nonparametric processes. *ASA proceedings of the section on Bayesian statistical science.* Alexandria, Virginia. Virginia: American Statistical Association; 1999, 50–55.

Madigan, D., York, J. and Allard, D. (1995). Bayesian graphical models for discrete data. In: *International Statistical Review/Revue Internationale de Statistique*, 215–232.

Makalic, E. and Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. In: *IEEE Signal Processing Letters*, **23**,1, 179–182.

Malsiner-Walli, G. and Wagner, H. (2018). Comparing spike and slab priors for Bayesian variable selection. In: *arXiv preprint arXiv:1812.07259.*

Matthews, D. R., Hosker, J. P., Rudenski, A. S., Naylor, B. A., Treacher, D. F. and Turner, R. C. (1985). Homeostasis model assessment: insulin resistance and $\beta$-cell function from fasting plasma glucose and insulin concentration in man. In: *Diabetologia*, **28**,7, 412–419.

McGarrah, R. W., Crown, S. B., Zhang, G.-F., Shah, S. H. and Newgard, C. B. (2018). Cardiovascular metabolomics. In: *Circulation research*, **122**,9, 1238–1258.

Meinshausen, N., Bühlmann, P. et al. (2006). High-dimensional graphs and variable selection with the lasso. In: *The annals of statistics*, **34**,3, 1436–1462.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. In: *The journal of chemical physics*, **21**,6, 1087–1092.

Misra, S. and Owen, K. R. (2018). Genetics of Monogenic Diabetes: Present Clinical Challenges. In: *Current Diabetes Reports*, **18**,12, 141. ISSN: 1539-0829. DOI: 10.1007/s11892-018-1111-4. URL: https://doi.org/10.1007/s11892-018-1111-4.

Mohammadi, A., Wit, E. C. et al. (2015). Bayesian structure learning in sparse Gaussian graphical models. In: *Bayesian Analysis*, **10**,1, 109–138.

O'Hara, R. B., Sillanpää, M. J. et al. (2009). A review of Bayesian variable selection methods: what, how and which. In: *Bayesian analysis*, **4**,1, 85–117.

Omori, Y., Chib, S., Shephard, N. and Nakajima, J. (2007). Stochastic volatility with leverage: Fast and efficient likelihood inference. In: *Journal of Econometrics*, **140**,2, 425–449.

Peterson, C., Stingo, F. C. and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. In: *Journal of the American Statistical Association*, **110**,509, 159–174.

Piironen, J., Vehtari, A. et al. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. In: *Electronic Journal of Statistics*, **11**,2, 5018–5051.

Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. In: *Journal of the American statistical Association*, **108**,504, 1339–1349.

Psarakis, S. and Panaretoes, J (1990). The folded t distribution. In: *Communications in Statistics-Theory and Methods*, **19**,7, 2717–2734.

Robert, C. and Casella, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. In: *Statistical Science*, 102–115.

Roberts, C. K., Hevener, A. L. and Barnard, R. J. (2013). Metabolic syndrome and insulin resistance: underlying causes and modification by exercise training. In: *Comprehensive Physiology*, **3**,1, 1–58.

Roverato, A. (2002). Hyper Inverse Wishart Distribution for Non-decomposable Graphs and its Application to Bayesian Inference for Gaussian Graphical Models. In: *Scandinavian Journal of Statistics*, **29**,3, 391–411.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**,2, 325–338.

Schervish, M. J. and Carlin, B. P. (1992). On the convergence of successive substitution sampling. In: *Journal of Computational and Graphical statistics*, **1**,2, 111–127.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. In: *Statistica Sinica*, **4**, 639–650.

Shoffner, J. M., Lott, M. T., Lezza, A. M., Seibel, P., Ballinger, S. W. and Wallace, D. C. (1990). Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRNALys mutation. In: *Cell*, **61**,6, 931–937.

Smits, B. W., Fermont, J., Delnooz, C. C., Kalkman, J. S., Bleijenberg, G. and Engelen, B. G. van (2011). Disease impact in chronic progressive external ophthalmoplegia: more than meets the eye. In: *Neuromuscular Disorders*, **21**,4, 272–278.

Sproston, K. and Mindell, J. (2006). Health Survey for England 2004. The health of minority ethnic groups. In:

Stafstrom, C. E. (2003). Hyperglycemia lowers seizure threshold. In: *Epilepsy currents*, **3**,4, 148–149.

Tam, C. S., Xie, W., Johnson, W. D., Cefalu, W. T., Redman, L. M. and Ravussin, E. (2012). Defining insulin resistance from hyperinsulinemic-euglycemic clamps. In: *Diabetes care*, **35**,7, 1605–1610.

Tan, L. S., Jasra, A., De Iorio, M. and Ebbels, T. (2016). Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. In: *arXiv preprint arXiv:1603.06358*.

Tillin, T., Forouhi, N. G., McKeigue, P. M. and Chaturvedi, N. (2010). Southall And Brent REvisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. In: *International journal of epidemiology*, **41**,1, 33–42.

Tillin, T., Hughes, A. D., Godsland, I. F., Whincup, P., Forouhi, N. G., Welsh, P., Sattar, N., McKeigue, P. M. and Chaturvedi, N. (2012). Insulin resistance and truncal obesity as important determinants of the greater incidence of diabetes in Indian Asians and African Caribbeans compared with Europeans: the Southall And Brent REvisited (SABRE) cohort. In: *Diabetes care*, DC_120544.

Udler, M. S., Kim, J., Grotthuss, M. von, Bonas-Guarch, S., Cole, J. B., Chiou, J., Boehnke, M., Laakso, M., Atzmon, G., Glaser, B. et al. (2018). Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. In: *PLoS medicine*, **15**,9, e1002654.

Valcárcel, B., Würtz, P., Basatena, N.-K. S. al, Tukiainen, T., Kangas, A. J., Soininen, P., Järvelin, M.-R., Ala-Korpela, M., Ebbels, T. M. and De Iorio, M. (2011). A differential network approach to exploring differences between biological states: an application to prediabetes. In: *PLoS One*, **6**,9, e24702.

Vorgerd, M. and Deschauer, M. (2011). "Treatment and management of hereditary metabolic myopathies". In: *Neuromuscular Disorders: Treatment and Management*. Elsevier, 409–429.

Wang, H. (2010). Sparse seemingly unrelated regression modelling: Applications in finance and econometrics. In: *Computational Statistics & Data Analysis*, **54**,11, 2866–2877.

Wang, H. et al. (2012). Bayesian graphical lasso models and efficient posterior computation. In: *Bayesian Analysis*, **7**,4, 867–886.

— (2015). Scaling it up: Stochastic search structure learning in graphical models. In: *Bayesian Analysis*, **10**,2, 351–377.

Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. In: *Biometrics*, 95–108.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. Tech. rep.

# Appendix A

# Supplementary Material Bayesian Nonparametric Modelling of Insulin Resistance. MCMC Details

In this appendix we describe the steps of the MCMC algorithm to estimate the nonparametric random intercept/error model presented in 4. We implement a Gibbs sampling, which requires a Metropolis update for some parameters.

1. The update of $\boldsymbol{\beta}$, the vector of regression coefficients, is the standard conjugate update from a Normal model, but conditional on the Spike and Slab selection. Note that in our application the observation $y_i$ are also indexed by the ethnicity indicator $g$. For ease of notation we drop the subscript $g$ as we assume the regression coefficients to be the same across ethnicities and we simply assume to have a total of $n$ observations. We introduce the latent variable indicator vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{p-1})$, where the element $\omega_j$ is equal to 1 if the $j_{th}$ covariate is included in the model and 0 otherwise. If $\omega_j = 0$ then the corresponding $\beta_j$ is equal to zero. Let $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ denote the sub-vector of $\boldsymbol{\beta}$ including elements for which the corresponding $\omega_j$ is equal to 1 (slab component of the model) and let $X_{\boldsymbol{\omega}}$ be the design matrix consisting only of those columns of $X$ corresponding to non-zero effects. Then the conditional distribution of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$

$$\begin{aligned} p(\boldsymbol{\beta}_{\boldsymbol{\omega}} \mid rest) &\propto \prod_{i=1}^{n} \mathrm{N}\left(y_i \mid \beta_{0i} + \boldsymbol{x}_{\boldsymbol{\omega} i}\boldsymbol{\beta}_{\boldsymbol{\omega}}, \tau_i^2\right) \times \prod_{j:\omega_j=1} \mathrm{N}\left(\beta_j \mid \mu_\beta, \tau_\beta^2\right) \\ &= \mathrm{N}\left(\boldsymbol{\beta}_{\boldsymbol{\omega}} \mid \tilde{\boldsymbol{\mu}}_\beta, \tilde{C}_\beta\right) \end{aligned}$$

where

$$\begin{aligned}
\tilde{C}_\beta &= \tau_\beta^2 I + X_{\boldsymbol{\omega}}' V X_{\boldsymbol{\omega}} \\
V &= \begin{bmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \ddots & 0 \\ \vdots & 0 & \ddots & \cdots \\ 0 & \cdots & 0 & \tau_n^2 \end{bmatrix} \\
\tilde{\boldsymbol{\mu}}_\beta &= \tilde{C}_\beta^{-1}\left(\tau_\beta^2 \boldsymbol{\mu}_\beta + X_{\boldsymbol{\omega}}' V \boldsymbol{y}\right)
\end{aligned}$$

and $\boldsymbol{y} = (y_1, \dots, y_n)$. Here $\boldsymbol{\mu}_\beta$ is the vector of appropriate dimension whose elements are all equal to $\mu_\beta$.

2. The update of $\boldsymbol{\omega}$ is performed evaluating the model marginal likelihood individually for each covariate (with the intercept $\beta_{0i}$ always included) as

$$p(\omega_j = 1 \mid \boldsymbol{\omega}_{\setminus \boldsymbol{j}}, rest) = \left[1 + \frac{1-\pi}{\pi}\frac{p(\boldsymbol{y} \mid \omega_j = 0, \boldsymbol{\omega}_{\setminus \boldsymbol{j}}, \tau_1^2, \dots, \tau_n^2)}{p(\boldsymbol{y} \mid \omega_j = 1, \boldsymbol{\omega}_{\setminus \boldsymbol{j}}, \tau_1^2, \dots, \tau_n^2)}\right]^{-1}$$

where $\boldsymbol{\omega}_{\setminus \boldsymbol{j}}$ denotes the vector $\boldsymbol{\omega}$ excluding $\omega_j$. $p(\boldsymbol{y} \mid \omega_j = 1, \boldsymbol{\omega}_{\setminus \boldsymbol{j}}, \tau_1^2, \dots, \tau_n^2)$ represents the marginal likelihood of the model obtained marginalising with respect to $\boldsymbol{\beta}$:

$$\begin{aligned}
& p(\boldsymbol{y} \mid \omega_j = 1, \boldsymbol{\omega}_{\setminus \boldsymbol{j}}, \tau_1^2, \dots, \tau_n^2) = \\
& = \int_{\boldsymbol{\beta}} p(\boldsymbol{y}, \boldsymbol{\beta} | \tau_1^2, \dots, \tau_n^2, \omega_j = 1, \boldsymbol{\omega}_{\setminus \boldsymbol{j}}) \; d\boldsymbol{\beta} \\
& = \int_{\boldsymbol{\beta}} p(\boldsymbol{y} | \boldsymbol{\beta}, \tau_1^2, \dots, \tau_n^2, \omega_j = 1, \boldsymbol{\omega}_{\setminus \boldsymbol{j}}) p(\boldsymbol{\beta}) \; d\boldsymbol{\beta} \\
& = -\frac{1}{2} n \log(2\pi) + \frac{1}{2}\log\left(|\tau_\beta^2 I|\right) - \frac{1}{2}\log\left(|\tilde{C}_\beta|\right) \\
& = -\frac{1}{2}\left(\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\mu}}_\beta' \tilde{C}_\beta \tilde{\boldsymbol{\mu}}_\beta\right)
\end{aligned}$$

where $|A|$ is the determinant of the matrix $A$ and $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_n)'$, where $\tilde{y}_i = (y_i - \beta_{0i})\tau_i$.

3. The update of $\pi$ is a straightforward conjugate update from a Beta-Bernoulli model

$$p(\pi \mid rest) = \text{Beta}\left(\pi_a + \sum \omega_j, \pi_b + (p-1) - \sum \omega_j\right)$$

4. To update the DGDP we adopt a truncated stick-breaking approach, i.e. we approximate the infinite mixture with a finite mixture with $L$ components where $L$ is large. A discussion on the truncation level can be found in Ishwaran and James (2001) and Barcella et al. (2017). We perform the following steps in order to update the parameters of the DGDP.

   (a) *Resampling the cluster allocation vector, given the rest.* Conditionally on the remaining parameters in the model, the allocation vectors, $\boldsymbol{s}_g$, are independent. Note that we have an allocation vector for each ethnicity $g$. Let $s_{ig}$ be the cluster indicator for observation $i$ in group $g$, with $s_{ig} \in 1, \ldots, L$, for $i = 1, \ldots, n$. We draw $s_{ig}$ from

   $$p(s_{ig} = k \mid rest) \propto \psi_{kg} \mathrm{N}(y_{ig} \mid \beta_{0i} + \sum_{j:\omega_j=1} \beta_j x_{ij}, \tau_i^2)$$

   for $k = 1, \ldots, L$.

   (b) *Resampling the mixture weights, $\psi_{kg}$, given the rest.* Conditionally on $g$ and the remaining parameters in the model, the mixture weights for each group are independent. This is a straightforward update due to the conjugacy between the Generalised Dirichlet distribution on $\psi_{1g}, \ldots, \psi_{Lg}$ and the Multinomial distribution on $\boldsymbol{s}$:

   $$\phi_{kg} \mid rest \sim \mathrm{Beta}\left(\mu_g \upsilon + \sum_{i=1}^{n_g} \mathrm{I}(s_{ig} = k), (1 - \mu_g)\upsilon + \sum_{i=1}^{n_g} \mathrm{I}(s_{ig} > k)\right)$$

   where $n_g$ is the number of observations in group $g$ and $\mathrm{I}(\cdot)$ represents the indicator function, assuming value 1 if the inner condition is satisfied and 0 otherwise. Then the weights $\psi_{kg}$ can be obtained using a stick-breaking procedure.

   (c) *Resampling $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_G)$ given the rest.* Conditionally on the weights $\psi_{kg}$ and $\upsilon$ we update $\boldsymbol{\mu}$ with a Metropolis-Hastings step, using a Multivariate Normal proposal. In this case the data corresponds to the set of sticks $\phi_{kg}$ and the likelihood is given by the product of Beta distributions. See Barcella et al. (2017) for details.

   (d) *Resampling $\upsilon$ given the rest.* Conditionally on the weights $\psi_{kg}$ and $\boldsymbol{\mu}$ we update $\upsilon$ with a Metropolis-Hastings step, with a Gamma proposal and data given by the set sticks $\phi_{kg}$ for $g = 1, \ldots, G$.

(e) *Resampling of the locations* $\theta_k = (\beta_{0k}, \tau_k^2)$ *given the rest.* The locations of the DGDP are a priori *iid* realisations of the base measure $G_0 = \mathrm{N}\,(m_0, \kappa_0^2) \times \mathrm{Gamma}(\tau_a, \tau_b)$. Given the clustering structure defined by the allocation vector $\boldsymbol{s}$, the update of $\theta_k$ is performed separately for each cluster and it is a straightforward conjugate update:

$$p(\beta_{0k}, \tau_k^2 \mid rest) \propto G_0(\beta_{0k}, \tau_k^2) \prod_{i,g:s_{ig}=k} \mathrm{N}(y_{ig} \mid \beta_{0k} + \sum_{j:\omega_j=1} \beta_j x_{ig}, \tau_k^2)$$

for $k = 1, \ldots, L$

# Appendix B

# Supplementary Material Bayesian Nonparametric Gaussian Graphical Models

In this Appendix we provide the proofs of the degree distribution generated by a Generalised Dirichlet Process prior adopted over the edges inclusion parameters of a GGM. We also provide the details of the MCMC algorithm to estimate the Sparse SUR model with multiple GGMs presented in Chapter 5.

## B.1 Degree Distribution properties

The degree, $D_i$, of a node $i$ is the number of connections that involve node $i$, so $D_i = \sum_{j \neq i} e_{ij}$, where $e_{ij} \in \{0, 1\}$, with $e_{ij} = 1$ denoting the presence of an edge connecting nodes $i$ and $j$ and 0 otherwise. The degree $D_i$ is then bounded between 0 and $M-1$, the total number of nodes minus one. Conditionally on $\pi_{ij}$, the probability that a node $i$ is connected to a chosen node $j$ is $\pi_{ij}$. The degree of a node $i$ is distributed as a mixture of Binomial distributions, with mixing weights and locations defined by the GDP prior:

$$D_i \mid P \sim \sum_{k=1}^{\infty} \psi_k \text{Binomial}(M-1, \pi_k)$$

where $\pi_k$ refers to a unique location in the GDP prior. The conditional first and second moments are:

$$\begin{aligned}
\mathrm{E}\left[D_i \mid P\right] = \mathrm{E}\left[\sum_{k=1}^{\infty} \psi_k \text{Binomial}(M-1, \pi_k)\right] &= \sum_{k=1}^{\infty} \psi_k \mathrm{E}\left[\text{Binomial}(M-1, \pi_k)\right] \\
&= \sum_{k=1}^{\infty} \psi_k \times (M-1)\pi_k
\end{aligned}$$

and

$$\mathrm{E}\left[D_i^2 \mid P\right] = \sum_{i=1}^{\infty} \psi_k (M-1)\pi_k(1-\pi_k+(M-1)\pi_k)$$

Marginalising over the probabilities $\pi_k$s, we obtain

$$\begin{aligned}
\mathrm{E}\left[D_i \mid \psi\right] &= \frac{a_\pi}{a_\pi+b_\pi}\sum_{k=1}^{\infty} \psi_k \times (M-1) \\
\mathrm{E}\left[D_i^2 \mid \psi\right] &= \sum_{k=1}^{\infty} \psi_k (M-1)\left[\frac{a_\pi}{a_\pi+b_\pi} + (M-2)\frac{a_\pi(a_\pi+1)}{(a_\pi+b_\pi)(a_\pi+b_\pi+1)}\right]
\end{aligned}$$

Marginalising over $\psi_k$ gives

$$\begin{aligned}
\mathrm{E}\left[D_i\right] &= \frac{a_\pi}{a_\pi+b_\pi}\mu\sum_{k=1}^{\infty}(1-\mu)^{k-1} \times (M-1) \\
&= \frac{a_\pi}{a_\pi+b_\pi}(M-1)\mu\frac{1}{1-(1-\mu)} = (M-1)\frac{a_\pi}{a_\pi+b_\pi} \\
\mathrm{E}\left[D_i^2\right] &= (M-1)\mu\sum_{k=1}^{\infty}(1-\mu)^{k-1}\left\{\frac{a_\pi}{a_\pi+b_\pi} + (M-2)\frac{(a_\pi+1)a_\pi}{(1+a_\pi+b_\pi)(a_\pi+b_\pi)}\right\} \\
&= (M-1)\left\{\frac{a_\pi}{a_\pi+b_\pi} + (M-2)\frac{(a_\pi+1)a_\pi}{(1+a_\pi+b_\pi)(a_\pi+b_\pi)}\right\}
\end{aligned}$$

Given the first two moments, the variance is easily derived.

## B.2 Posterior Inference details

Here we provide the details of the MCMC algorithm for the SUR model in the case multiple GGMs. We use a block Gibbs Sampling with Metropolis steps. Define $\boldsymbol{y}_g$ to be the sub-vector of $\boldsymbol{y}$, of dimension $Mn_g$, including only the observations belonging to group $g$. Similarly, define $X_g$ to be a block diagonal sub-matrix of $X$, of dimension $Mn_g \times Q$, including only the observation belonging to group $g$.

1. *Resampling $\boldsymbol{\beta}$, $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ given all the rest.* This step requires only conjugate

updates.

$$
\begin{aligned}
\boldsymbol{\beta} \mid rest &\sim \mathrm{N}\left(\boldsymbol{b}, A^{-1}\right) \\
A &= \sum_{g=1}^{R} \left\{ X_g^T \left(\Omega_g \otimes I_{n_g}\right) X_g \right\} + \Lambda_*^{-1} \\
\boldsymbol{b} &= A^{-1} \sum_{g=1}^{R} \left\{ X_g^T \left(\Omega_g \otimes I_{n_g}\right) \boldsymbol{y}_g \right\} \\
\Lambda_* &= \begin{bmatrix} \tau_1^2 \Lambda_1 & 0 & \cdots & 0 \\ 0 & \tau_2^2 \Lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tau_M^2 \Lambda_M \end{bmatrix}
\end{aligned}
$$

where $\Lambda_l$, for $l = 1, \ldots, M$, is a square matrix with diagonal equal to $(\lambda_{l1}^2, \ldots, \lambda_{lp_l}^2)$ and zeros elsewhere, $p_l$ is the number of covariates in regression $l$ and $\otimes$ denotes the Kronecker product. The conditional posterior distributions for the local and global shrinkage parameters are

$$
\begin{aligned}
\lambda_{lj}^2 \mid rest &\sim \text{Inv-Gamma}\left(1, \frac{1}{\nu_{lj}} + \frac{\beta_{lj}^2}{2\tau_l^2}\right) \\
\tau_l^2 \mid rest &\sim \text{Inv-Gamma}\left(\frac{p_l + 1}{2}, \frac{1}{\xi_l} + \frac{1}{2} \sum_{j=1}^{p} \frac{\beta_{lj}^2}{\lambda_{lj}^2}\right)
\end{aligned}
$$

with $l = 1, \ldots, M$ and $j = 1, \ldots, p_l$. The hyper-parameters are updated as

$$
\begin{aligned}
\nu_{lj} \mid rest &\sim \text{Inv-Gamma}\left(1, 1 + \frac{1}{\lambda_{lj}^2}\right) \\
\xi_l \mid rest &\sim \text{Inv-Gamma}\left(1, 1 + \frac{1}{\tau_l^2}\right)
\end{aligned}
$$

with $l = 1, \ldots, M$ and $j = 1, \ldots, p_l$.

2. *Resampling $\Omega_g$ given all the rest.* $\Omega_g$ is updated column-wise following the algorithm described in Wang (2015). Let $\tilde{Y}_g = [\tilde{\boldsymbol{y}}_1 \ldots \tilde{\boldsymbol{y}}_M]$, a $n_g \times M$ matrix, where each column is defined as $\tilde{\boldsymbol{y}}_l = \boldsymbol{y}_l - X_l \boldsymbol{\beta}_l$ and $V_g = (v_{z_{ij}}^2)$, is a $M \times M$ symmetric matrix with zeros on the diagonal. For simplicity of explanation, consider the last column and partition $\Omega_g$, $S_g = \tilde{Y}_g^T \tilde{Y}_g$ and $V_g$ as follows (index $g$ omitted for ease of notation)

$$
\Omega = \begin{bmatrix} \Omega_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{12}^T & \omega_{22} \end{bmatrix}, \; S = \begin{bmatrix} S_{11} & \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}^T & s_{22} \end{bmatrix}, \; V = \begin{bmatrix} V_{11} & \boldsymbol{v}_{12} \\ \boldsymbol{v}_{12}^T & v_{22} \end{bmatrix}
$$

Then, through the change of variable $\omega_{22} \to u = \omega_{22} - \boldsymbol{\omega}_{12}^T \Omega_{11}^{-1} \boldsymbol{\omega}_{12}$, we have the following full conditionals

$$\boldsymbol{\omega}_{12} \mid rest \sim \mathrm{N}\left(-C\boldsymbol{s}_{12}, C\right)$$
$$u \mid rest \sim \mathrm{Gamma}\left(\frac{n_g}{2} + 1, \frac{s_{22} + \eta}{2}\right)$$

where $C = \left\{(s_{22} + \eta)\Omega_{11}^{-1} + \mathrm{diag}(\boldsymbol{v}_{12})^{-1}\right\}^{-1}$. The steps are repeated for each column, leading to a fast block update.

3. *Resampling $G_g$ given all the rest.* The graph for group $g$ is independent from all the other graphs given the edge inclusion probabilities $\pi_{g,ij}$. We update the graph through its adjacency matrix $Z_g$. The $z_{g,ij}$s are independent Bernoulli random variables with probability of success

$$p(z_{g,ij} = 1 \mid \Omega_g, rest) = \frac{\pi_{g,ij}\mathrm{N}(\omega_{g,ij} \mid 0, v_1^2)}{\pi_{g,ij}\mathrm{N}(\omega_{g,ij} \mid 0, v_1^2) + (1 - \pi_{g,ij})\mathrm{N}(\omega_{g,ij} \mid 0, v_0^2)}$$

To update the parameters in the DGDP we follow the algorithm proposed Barcella et al. (2017), which is based on a truncation to $K$ components of the infinite mixture.

1. *Resampling the cluster allocation vector, given the rest.* Let $c_{g,ij}$ be the cluster indicator for edge $e_{ij}$ in group $g$, with $c_{g,ij} \in \{1, \dots, K\}$, for $i, j = 1, \dots, M, i < j$. Draw $c_{g,ij}$ from

$$p(c_{g,ij} = k \mid rest) \propto \psi_{kg}\mathrm{Ber}(e_{g,ij} \mid \pi_k), \text{ for } k = 1, \dots, K$$

2. *Resampling the mixture weights, $\psi_{kg}$, given the rest.* Conditionally on the group $g$ and the remaining parameters in the model, the mixture weights for each group are independent. This is a straightforward update due to the conjugacy of the Generalised Dirichlet distribution on $\psi_{1g}, \dots, \psi_{Kg}$ with the Multinomial distribution on $\boldsymbol{c}_g$:

$$\phi_{kg} \mid rest \sim \mathrm{Beta}\left(\alpha\mu_g + \sum_{i<j} 1(c_{g,ij} = k),\ \alpha(1 - \mu_g) + \sum_{i<j} 1(c_{g,ij} > k)\right)$$

where $1(\cdot)$ denotes the indicator function, assuming value 1 if the inner condition is satisfied and 0 otherwise. The weights $\psi_{kg}$ can be obtained using the stick-breaking procedure.

3. *Resampling $\alpha$ and $\boldsymbol{\zeta}$, given the rest.* We use two Metropolis steps for $\alpha$ and $\boldsymbol{\zeta}$. Draw a new value $\alpha^*$ by sampling $\log \alpha^*$ from a Normal proposal centred around the current value and accept $\alpha^*$ with probability $\min\{a_\alpha, 1\}$, where the ratio $a_\alpha$ is given by (taking into account the Jacobian $J()$ of the transformation $\alpha \to \log \alpha$)

$$a_\alpha = \frac{p(\boldsymbol{\phi}|\alpha^*, \boldsymbol{\mu}^s)p(\alpha^*)J(\alpha^*)}{p(\boldsymbol{\phi}|\alpha^s, \boldsymbol{\mu}^s)p(\alpha^s)J(\alpha^s)}$$

where $\boldsymbol{\mu}^s = (\mu_1^s, \ldots, \mu_R^s)$, $\mu_g^s = \boldsymbol{x}_g \boldsymbol{\zeta}^s$ and the super-index $s$ refers to the current value of $\boldsymbol{\mu}$ and $\boldsymbol{\zeta}$.

Draw a new value $\boldsymbol{\zeta}^*$ from a multivariate Normal proposal centred on the current values and accept with probability $\min\{a_\zeta, 1\}$, where the ratio $a_\zeta$ is given by

$$a_\zeta = \frac{p(\boldsymbol{\phi}|\alpha^s, \boldsymbol{\mu}^*)p(\boldsymbol{\zeta}^*)}{p(\boldsymbol{\phi}|\alpha^s, \boldsymbol{\mu}^s)p(\boldsymbol{\zeta}^s)}$$

where $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_R^*)$ and $\mu_g^* = \text{logit}(\boldsymbol{x}_g \boldsymbol{\zeta}^*)$ and $\boldsymbol{\mu}^s$.

4. *Resampling $P_0$ given the rest.* This step is a conjugate update from the Beta-Binomial model. For each component draw $\pi_k$ from

$$\begin{aligned}\pi_k \mid rest &\sim \text{Beta}\left(a_\pi^*, b_\pi^*\right)\\ a_\pi^* &= a_\pi + \sum_{g=1}^{R} \sum_{i<j} 1(c_{g,ij} = k \wedge e_{g,ij} = 1)\\ b_\pi^* &= b_\pi + \sum_{g=1}^{R} \sum_{i<j} 1(c_{g,ij} = k \wedge e_{g,ij} = 0)\end{aligned}$$

where $\wedge$ is the *and* boolean condition.

# Appendix C

# Supplementary Material Bayesian Dynamic Multiple Graphical Models

In this appendix we provide the Gibbs sampling algorithm details and the Stan code to perform inference on the Dynamic Multiple Graphical Models via Nodewise Regression presented in Chapter 6.

## C.1 Gibbs sampling algorithm

Here we provide details of the Gibbs sampling. Our starting point is the algorithm described in Kowal et al. (2017) and we extend it to allow for multiple groups of different sample sizes. The sampler consists of two main components: a stochastic volatility sampling algorithm (Kastner and Frühwirth-Schnatter, 2014) augmented with a Polya-Gamma sampler (Polson et al., 2013), and a Cholesky Factor Algorithm (Rue, 2001) to sample the regression coefficients in the dynamic linear model. We provide details of the update steps that differ from the original paper, for the others refer to Kowal et al. (2017).

1. *Resampling the log-variances $h_{jrt}$ given the rest.* Omori et al. (2007) propose a method to sample directly from the full-conditional distribution of $\boldsymbol{h_{jr}} = (h_{jr1}, \dots, h_{jrT})$, working on the log-scale, where the ensuing log-chi-square distribution of the likelihood $\log(\gamma^2_{jrt})$, from

$$\log\left((\beta_{jrt} - \beta_{jrt-1})^2\right) = \log(\gamma^2_{jrt}) + h_{jrt}$$

is approximated via a known normal mixture approximation of 10 components. To sample the new log-variances we use the all-without-a-loop (AWOL) sampler of Kastner and Frühwirth-Schnatter (2014), which allows to sample $\boldsymbol{h_{jr}}$ jointly without the need of a sequential algorithm. Conditional on all the other parameters the log-variances are independent across groups. Resampling proceeds as in Kowal et al. (2017).

2. *Resampling the mixture components indicator $s_{jrt}$ given the rest.* The discrete mixture probabilities are straightforward to update. The prior mixture probabilities are the pre-specified mixing proportions given by Omori et al. (2007) and the likelihood is $\log(\gamma_{jrt}^2 + c) \sim \mathrm{N}(h_{jrt} + m_{s_{jrt}}, v_{s_{jrt}})$, where $c$ is a small offset to avoid numerical underflows and $m$ and $v$ are the pre-specified mean and variance, respectively, of the 10 mixture components. Resampling proceeds as in Kowal et al. (2017), independently for each group.

3. *Resampling the mean parameters $\mu_{jr}$ given the rest.* The update of $\mu_{jr}$ is done independently for each group. We can re-write the log-variance equation in (6.8) as an ordinary linear regression, where the parameters $\mu_{jr}$ play the role of regression coefficients, as follows:

$$
\begin{aligned}
h_{jrt} &= \mu_{jr} + \phi_{jr}(h_{jrt-1} - \mu_{jr}) + \eta_{jrt} \\
\tilde{h}_{jrt} &= \mu_{jr}(1 - \phi_{jr}) + \eta_{jrt}
\end{aligned}
$$

where $\tilde{h}_{jrt} = h_{jrt} - \phi_{jr} h_{jrt-1}$ and $\tilde{h}_{jrt} \sim \mathrm{N}\left(\mu_{jr}(1 - \phi_{jr}), \xi_{\eta_{jrt}}^{-1}\right)$. Finally, we can write

$$\tilde{\tilde{h}}_{jrt} \sim \mathrm{N}\left(\mu_{jr} z_{jrt}, 1\right)$$

where $z_{jrt} = (1 - \phi_{jr})\sqrt{\xi_{\eta_{jrt}}}$ are the regression covariates, $\mu_{jr}$ is the regression coefficient and $\tilde{\tilde{h}}_{jrt} = \tilde{h}_{jrt}\sqrt{\xi_{\eta_{jrt}}}$, for $t = 2, \cdots, T$, are the response variables. The posterior follows from the standard update of a Normal prior with Normal likelihood. For $j = 1, \ldots, p$ and $r = 1, \ldots, R$ we sample from

$$
\begin{aligned}
\mu_{jr} \mid rest &\sim \mathrm{N}\left(l_\mu / q_\mu, 1/q_\mu\right) \\
q_\mu &= \xi_{\mu_{jr}} + \xi_{\eta_{jr1}} + \sum_{t=2}^{T} z_{jrt}^2 \\
l_\mu &= \xi_{\mu_{jr}} \mu_0 + h_{jr1} \xi_{\eta_{jr1}} + \sum_{t=2}^{T} \tilde{\tilde{h}}_{jrt} z_{jrt}
\end{aligned}
$$

4. *Resampling the mean parameter $\mu_0$ given the rest.* The posterior distribution of $\mu_0$ follows from the standard update of a Normal prior and Normal likelihood represented by $\mu_{jr}$. We sample from

$$\begin{aligned}\mu_0 \mid rest &\sim \mathrm{N}(l_0/q_0, 1/q_0)\\ q_0 &= \xi_{\mu_0} + \sum_{r=1}^{R}\sum_{j=1}^{p} \xi_{\mu_{jr}}\\ l_0 &= \sum_{r=1}^{R}\sum_{j=1}^{p} \xi_{\mu_{jr}}\mu_{jr}\end{aligned}$$

5. *Resampling the autoregressive coefficients $\phi_{jr}$ given the rest.* The new value of $\phi_{jr}$ is drawn via slice sampler, independently for each group. The parametrisation $(\phi_{jr}+1)/2 \sim \mathrm{Beta}(a_\phi, b_\phi)$ implies that $|\phi_{jr}| < 1$, which ensures a stationary stochastic process for $\boldsymbol{h_{jr}}$. Resampling is performed as in Kowal et al. (2017).

6. *Resampling the Polya-Gamma mixing parameters $\xi_{\eta_{jrt}}$, $\xi_{\mu_{jr}}$ and $\xi_{\mu_0}$ given the rest.* This step is a conjugate update of a Polya-Gamma prior given a Gaussian likelihood (Polson et al., 2013).

$$\begin{aligned}\xi_{\eta_{jrt}} \mid rest &\sim \text{Polya-Gamma}(1, \eta_{jrt}), \qquad \forall j, r, t\\ \xi_{\mu_{jr}} \mid rest &\sim \text{Polya-Gamma}(1, \mu_{jr} - \mu_0), \qquad \forall j, r\\ \xi_{\mu_0} \mid rest &\sim \text{Polya-Gamma}\left(1, \mu_0\right)\end{aligned}$$

7. *Resampling the regression coefficients $\boldsymbol{\beta}_{rt}$ given the rest.* Conditional on the rest, the regression coefficients of each group are independent. We perform a joint update of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_t^T, \ldots, \boldsymbol{\beta}_T^T)$, where $\boldsymbol{\beta}_t = (\beta_{1t}, \ldots, \beta_{pt})$, exploiting the block-diagonal structure of the posterior precision matrix of $\boldsymbol{\beta}$. The posterior distribution is (omitting the group subscript for ease of notation)

$$\begin{aligned}\boldsymbol{\beta} \mid rest &\sim \mathrm{N}\left(Q_\beta^{-1}\boldsymbol{m}_\beta, Q_\beta^{-1}\right)\\ Q_\beta &= A_\sigma + A_h\end{aligned}$$

where $A_\sigma$ is a $Tp \times Tp$ block-diagonal matrix defined as follows:

$$A_\sigma = \begin{bmatrix} X_1^T X_1/\sigma_1^2 & 0 & \cdots & \cdots & 0\\ \vdots & \ddots & \vdots & \vdots & \vdots\\ 0 & \cdots & X_t^T X_t/\sigma_t^2 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \cdots & X_T^T X_T/\sigma_T^2 \end{bmatrix}$$

and $A_h$ is a $Tp \times Tp$ matrix defined as follows:

$$A_h = \left(D^T \otimes I_p\right) \Sigma_h^{-1} \left(D \otimes I_p\right)$$

where $\otimes$ denotes the Kronecker product, $D$ is a $T \times T$ tri-diagonal matrix with diagonal entries equal to 1 and first off-diagonal elements equal to $-1$, $I_p$ is the $p \times p$ identity matrix and $\Sigma_h^{-1}$ is a $Tp \times Tp$ diagonal matrix with diagonal entries equal to

$$(\exp(h_{11}/2), \ldots, \exp(h_{p1}/2), \ldots, \exp(h_{1T}/2), \ldots, \exp(h_{pT}/2))$$

The posterior mean $\boldsymbol{m}_\beta$ is a $Tp$-dimensional vector defined as:

$$\boldsymbol{m}_\beta = \left(X_1^T \boldsymbol{y}_1/\sigma_1^2, \ldots, X_t^T \boldsymbol{y}_t/\sigma_t^2, \ldots, X_T^T \boldsymbol{y}_T/\sigma_T^2\right)$$

8. *Resampling the observation error variances $\sigma_{rt}^2$ given the rest.* This step is a conjugate update of an Inverse-Gamma prior given a Gaussian likelihood. For each group $r$ and time $t$ we sample:

$$\sigma_{rt}^2 \mid rest \sim \text{Inverse-Gamma}\left(a_\sigma + \frac{n_{rt}}{2}, b_\sigma + \frac{\sum_{i=1}^{n_{rt}}(y_{irt} - \boldsymbol{x}_{irt}\boldsymbol{\beta}_{rt})^2}{2}\right)$$

## C.2 Stan Code

```
data {
int < lower = 1 > NT;            // Time points
int < lower = 1 > n_t[ NT ]; // Number of observations for each time point
int < lower = 1 > Ngr; // Number of groups
int < lower = 1 > n_groups[ Ngr*NT ]; // Dimension of the groups for each time
int < lower = 1 > m;                  // Number of regressors
int < lower = 1, upper = Ngr > G_t[ sum(n_groups) ];    // Vector with groups membership
vector [ sum(n_t) ] y_t;             // Response variable
row_vector[m] X_t[ sum(n_t) ];         // Input matrix (= Y_-j)
// ---------------
real < lower = 0 > scale_global_tau_0[ Ngr ]; // prior scale for the global shrinkage parameter Tau
real < lower = 1 > df_global_0;  // degrees of freedom for Tau
real < lower = 1 > df_global_j;
real < lower = 1 > df_local_t; // degrees of freedom for Lambdas
matrix <lower = 0> [m, Ngr] devs_X_t[ NT ]; // Deviances of X (diag of XtX )
// ------ Params for the regularisation of the Horseshoe
real <lower = 0> slab_scale_c_t;  // Slab of student-t
real <lower = 0> slab_df_c_t;  // df of student-t
// Parameters Beta on Phi_j
real <lower = 0> phi_ab[ 2 ];
real <lower = 0> sigma_prior;
```

```
}
transformed data{
int n_cumsum_gr[ Ngr * NT ]; // Vector of cumulative sum of observations in groups per time
int pos;
pos = 1;
for(tt in 1:NT){
for(gr in 1:Ngr){
if( pos == 1) n_cumsum_gr[ pos ] = 1;
if( pos > 1) n_cumsum_gr[ pos ] = n_cumsum_gr[ pos - 1 ] + n_groups[ pos - 1 ];
pos += 1;
}
}
}
parameters {
vector < lower = 0 > [ Ngr * NT ] sigma_t; // Time varying noise std
// Parameters T
vector[ m * Ngr ] z_t[ NT ]; //Innovations for Beta_t
// Tau_j
vector < lower = 0 > [ m * Ngr ] aux1_global_j;
vector < lower = 0 > [ m * Ngr ] aux2_global_j;
// Tau_0
real < lower = 0 > aux1_global_0;
real < lower = 0 > aux2_global_0;
// Local Lambda_jt
vector < lower = 0 > [ m * Ngr ] aux1_local_t[ NT ];
vector < lower = 0 > [ m * Ngr ] aux2_local_t[ NT ];
vector < lower = 0 > [ Ngr ] aux_c_t[ NT ];
// Autoregressive Parameter phi_j for the dynamic HS
vector < lower = 0 , upper = 1 > [ m * Ngr ] phi_j_pos;
}
transformed parameters {
vector < lower = -1 , upper = 1 > [ m * Ngr ] phi_j;
// Tau_j
vector < lower = 0 > [ m ] tau_j[ Ngr ];
// Tau_0
real < lower = 0 > tau_0[ Ngr ];
// Local Lambda_j,t
vector < lower = 0 > [ m ] lambda_t[ NT * Ngr ];
vector < lower = 0 > [ m ] lambda_tilde_t[ NT * Ngr ]; // truncated local shrinkage parameter
vector < lower = 0 > [ Ngr ] c_t[ NT ]; // c for reg HS
// log_scale variance h_j,t
vector [ m ] log_h[ NT * Ngr ];
vector < lower = 0 > [ m ] exp_h[ NT * Ngr ];
// Beta_t
vector[ m ] beta_t[ NT * Ngr ];
vector [ sum( n_t ) ] f_t;
vector < lower = 0 > [ sum(n_t) ] sigma_long;  // To store the sigma for each obs
// <<<<<<<<<<<<<<<<<<<<<>>>>>>>>>>>>>>>>>>>>>>>>
// <<<<<<<<<<<<<<< Now operations >>>>>>>>>>>>>>
// <<<<<<<<<<<<<<<<<<<<<>>>>>>>>>>>>>>>>>>>>>>>>
// Adjust phi_j
```

```
phi_j = (phi_j_pos * 2) - 1;
// In Time == 1
c_t[ 1 ] = slab_scale_c_t * sqrt( aux_c_t[ 1 ] );
for(gr in 1:Ngr){
// Global Shrinkage parameters
tau_j[ gr ] = aux1_global_j[((gr-1)*m + 1):((gr-1)*m + m)] .*
sqrt( aux2_global_j[((gr-1)*m + 1):((gr-1)*m + m)] );
tau_0[ gr ] = aux1_global_0 * sqrt( aux2_global_0 ) * scale_global_tau_0[ gr ] * sigma_t[ gr ];
// tau_j[ gr ] = aux1_global_j .* sqrt( aux2_global_j[ ((gr-1)*m + 1):((gr-1)*m + m) ] );
lambda_t[ gr ] = aux1_local_t[ 1 ][((gr-1)*m + 1):((gr-1)*m + m)].*
sqrt( aux2_local_t[ 1 ][((gr-1)*m + 1):((gr-1)*m + m)] );
lambda_tilde_t[ gr ] = sqrt( ( c_t[ 1 ][ gr ]^2 *
square( lambda_t[ gr ] ) ) ./ ( c_t[ 1 ][ gr ]^2 +
square( tau_j[ gr ] ) .* square( lambda_t[ gr ] )));
// Begin the creation og log-volatility h
log_h[ gr ] = 2*log( lambda_tilde_t[ gr ] ) + 2*( log( tau_j[ gr ] ) + log(tau_0[ gr ]) );
exp_h[ gr ] = exp( 0.5 * log_h[ gr ] );
}
// For time > 1
for(tt in 2:NT){
c_t[ tt ] = slab_scale_c_t * sqrt( aux_c_t[ tt ] );
for(gr in 1:Ngr ){
lambda_t[ (tt-1)*Ngr + gr ] = aux1_local_t[ tt ][((gr-1)*m + 1):((gr-1)*m + m)] .*
sqrt( aux2_local_t[ tt ][((gr-1)*m + 1):((gr-1)*m + m)] );
lambda_tilde_t[ (tt-1)*Ngr + gr ] = sqrt( ( c_t[ tt ][ gr ]^2 *
square( lambda_t[ (tt-1)*Ngr + gr ] ) ) ./
( c_t[ tt ][ gr ]^2 + square( tau_j[ gr ] ) .* square(lambda_t[ (tt-1)*Ngr + gr ])));
for(jj in 1:m){
log_h[ (tt-1)*Ngr + gr ][jj] = 2*log( lambda_tilde_t[ (tt-1)*Ngr + gr ][ jj ] ) +
2*( log( tau_j[ gr ][jj]) + log( tau_0[ gr ] ) ) +
phi_j[ (gr-1)*m + jj ] * ( log_h[ (tt-2)*Ngr + gr ][jj] -
2*( log( tau_j[ gr ][jj]) + log( tau_0[ gr ] )));
exp_h[ (tt-1)*Ngr + gr ][jj] = exp( 0.5 * log_h[ (tt-1)*Ngr + gr ][jj] );
}
}// gr
}// tt
// Beta_t
for(gr in 1:Ngr){
beta_t[ gr ] = z_t[ 1 ][((gr-1)*m + 1):((gr-1)*m + m)] .* exp_h[ gr ];
}
for(tt in 2:NT){
for(gr in 1:Ngr){
beta_t[ (tt-1)*Ngr + gr ] = beta_t[ (tt-2)*Ngr + gr ] + z_t[ tt ][((gr-1)*m + 1):((gr-1)*m + m)] .*
exp_h[ (tt-1)*Ngr + gr ];
}
}
// Mean function
{
int ii_ind;
ii_ind = 1;
for(tt in 1:NT ){
```

```
for(gr in 1:Ngr){
for(ii in 1:n_groups[ (tt-1)*Ngr + gr ] ){
sigma_long[ ii_ind ] = sigma_t[ (tt-1)*Ngr + gr ];
f_t[ ii_ind ] = X_t[ ii_ind ] * beta_t[ (tt-1)*Ngr + gr ];
ii_ind += 1;
}
}
}
}
}
model {
// Here we use auxiliary variables for tau and lambda
// Locals shared across groups
sigma_t ~ inv_gamma( 0.5, 0.5 );
phi_j_pos ~ beta( phi_ab[1], phi_ab[2] );
// One global tau for each group
aux1_global_0 ~ std_normal();
aux2_global_0 ~ inv_gamma( 0.5 * df_global_0, 0.5 * df_global_0 );
aux1_global_j ~ std_normal();
aux2_global_j ~ inv_gamma( 0.5 * df_global_j, 0.5 * df_global_j );
for(tt in 1:NT){
aux2_local_t[ tt ] ~ inv_gamma( 0.5 * df_local_t, 0.5 * df_local_t );
aux1_local_t[ tt ] ~ std_normal();
aux_c_t[ tt ] ~ inv_gamma( 0.5 * slab_df_c_t, 0.5 * slab_df_c_t );
z_t[ tt ] ~ std_normal();
}
// Likelihood
y_t ~ normal(f_t , sigma_long );
}
generated quantities {
// Vector of k_j, the pseudo inclusion probability
vector [ m ] k_j_t[ NT * Ngr ];
// Vector to hold precision Omega elements for the current equation
vector [ m ] omega_vec_t[ NT * Ngr ];
// Elements of Omega
for(tt in 1:NT){
for(gr in 1:Ngr ){
for(jj in 1:m){
k_j_t[ (tt-1)*Ngr + gr ][jj] = 1/( 1 + 1/square( sigma_t[ gr ] ) * (exp_h[ (tt-1)*Ngr + gr ][jj] *
exp_h[ (tt-1)*Ngr + gr ][jj]) * devs_X_t[ tt ][ jj, gr ] );
omega_vec_t[ (tt-1)*Ngr + gr ][ jj ] = (- beta_t[ (tt-1)*Ngr + gr ][ jj ] ) / square( sigma_t[ gr ] );
}
}
}
}
```

# Appendix D

# Tables

Table D.1: List of metabolites included in the analysis. Each listed lipoprotein is further fractioned according to the content of triglycerides, phospholipids, cholesterol esters and free cholesterol. All the metabolites concentrations here listed are measured in millimoles per litre (mmol/L)

| **Abbreviation** | **Full name** |
|---|---|
| acace | Acetoacetate |
| ace | Acetate |
| ala | Alanine |
| alb | Albumin |
| apoa1 | Apolipoprotein A-I |
| apob | Apolipoprotein B |
| bohbut | 3-hydroxybutyrate |
| cit | Citrate |
| crea | Creatinine |
| dha | 22:6, docosahexaenoic acid |
| faw3 | Omega-3 fatty acids |
| faw6 | Omega-6 fatty acids |
| glc | Glucose |
| gln | Glutamine |
| glol | Glycerol |
| gly | Glycine |
| gp | Glycoprotein acetyls, mainly a1-acid glycoprotein |
| his | Histidine |
| ile | Isoleucine |
| la | 18:2, linoleic acid |
| lac | Lactate |
| leu | Leucine |
| mufa | Monounsaturated fatty acids; 16:1, 18:1 |
| pc | Phosphatidylcholine and other cholines |
| phe | Phenylalanine |
| pufa | Polyunsaturated fatty acids |
| pyr | Pyruvate |
| sfa | Saturated fatty acids |
| sm | Sphingomyelins |
| tyr | Tyrosine |
| unsatdeg | Estimated degree of unsaturation |
| val | Valine |

| **Abbreviation** | **Full name** |
|---|---|
| lipids_s_hdl | Small HDL lipids compounds |
| lipids_m_hdl | Medium HDL lipids compounds |
| lipids_l_hdl | Large HDL lipids compounds |
| lipids_xl_hdl | Extra large HDL lipids compounds |
| lipids_s_ldl | Small LDL lipids compounds |
| lipids_m_ldl | Medium LDL lipids compounds |
| lipids_l_ldl | Large LDL lipids compounds |
| lipids_idl | IDL lipids compounds |
| lipids_xs_vldl | Extra small VLDL lipids compounds |
| lipids_s_vldl | Small VLDL lipids compounds |
| lipids_m_vldl | Medium VLDL lipids compounds |
| lipids_l_vldl | Large VLDL lipids compounds |
| lipids_xl_vldl | Extra large VLDL lipids compounds |
| lipids_xxl_vldl | Extra extra large VLDL lipids compounds |

Table D.2: List of clinical and anthropometric covariates

| **Abbreviation** | **Full name** |
|---|---|
| Age | Age at the first visit (baseline) |
| WHR | Waist to Hip Ratio |
| HOMA IR | Homeostasis Model Assessment Insulin Resistance |
| Thigh skinfold | Thigh skinfold |
| Arm circumf | Arm circumference |
| Sagittal diam | Sagittal diameter |
| Subscap skinfold | Subscapular skinfold |
| Supiliac skinfold | Suprailiac skinfold |
| Thigh circumf | Thigh circumference |
| Triceps skinfold | Triceps skinfold |
| bp_avdias | Blood pressure diastolic |
| bp_avsys | Blood pressure systolic |
| ALT | Alanine aminotransferase |
| AST | Aspartate aminotransferase |
| GGT | Gamma glutamyltransferase |
| Sex female | Dummy variable for female sex |
| Smoke_Ex | Dummy variable ex smoker |
| Smoke_Current | Dummy variable current smoker |

Table D.3: Bayesian Variable Selection from the MC$^3$ method

| | acace | ace | ala | alb | apoa1 | apob | bohbut | cit | crea | dha | faw3 | faw6 | glc | gln | glol | gly | gp | his | ile | la | lac |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | . | . | . | . | . | ✓ | . | ✓ | ✓ | ✓ | ✓ | ✓ | . | . | . | ✓ | ✓ | . | . | ✓ | . |
| WHR | . | ✓ | ✓ | . | . | ✓ | . | . | . | . | ✓ | . | ✓ | . | ✓ | . | ✓ | . | ✓ | . | ✓ |
| HOMA IR | ✓ | ✓ | ✓ | . | ✓ | . | . | . | . | . | . | . | ✓ | ✓ | ✓ | . | ✓ | . | ✓ | . | ✓ |
| thigh_skinfold | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| arm_circumf | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| sagittal_diam | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| subscap_skinfold | . | . | . | . | ✓ | . | . | . | . | . | . | . | . | . | . | . | ✓ | . | . | . | . |
| supiliac_skinfold | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| thigh_circumf | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| triceps_skinfold | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| ALT | . | ✓ | ✓ | . | ✓ | ✓ | . | . | . | ✓ | ✓ | ✓ | . | . | . | ✓ | . | . | . | ✓ | . |
| AST | . | ✓ | . | . | ✓ | ✓ | . | . | . | ✓ | ✓ | ✓ | . | . | ✓ | . | . | . | . | ✓ | . |
| GGT | . | . | . | . | ✓ | . | . | . | . | . | . | . | . | ✓ | . | . | ✓ | . | . | . | . |
| Ethn_Asian | ✓ | ✓ | . | . | . | ✓ | ✓ | . | . | ✓ | ✓ | ✓ | ✓ | ✓ | . | . | ✓ | . | . | ✓ | ✓ |
| Ethn_African | . | . | ✓ | . | . | ✓ | . | . | ✓ | ✓ | . | . | ✓ | . | ✓ | . | ✓ | . | . | . | . |
| Somke_Ex | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Smoke_Current | . | . | . | . | . | . | . | ✓ | . | . | . | ✓ | . | . | . | ✓ | ✓ | . | . | ✓ | ✓ |
| Sex_female | ✓ | ✓ | ✓ | . | ✓ | . | ✓ | . | ✓ | . | . | ✓ | ✓ | ✓ | ✓ | ✓ | . | ✓ | ✓ | ✓ | ✓ |

| | leu | mufa | pc | phe | pufa | pyr | serum_c | serum_tg | sfa | sm | tyr | unsatdeg | val | lip_vldl | lip_ldl | lip_idl | lip_hdl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | . | . | ✓ | . | ✓ | . | ✓ | ✓ | . | ✓ | . | ✓ | . | ✓ | ✓ | ✓ | . |
| WHR | ✓ | ✓ | . | ✓ | . | . | . | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | . | . | . |
| HOMA IR | ✓ | . | . | ✓ | . | ✓ | . | ✓ | ✓ | . | ✓ | ✓ | ✓ | ✓ | . | . | ✓ |
| thigh_skinfold | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| arm_circumf | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| sagittal_diam | . | . | . | . | . | . | . | . | . | . | . | ✓ | . | . | . | . | . |
| subscap_skinfold | . | . | . | . | . | . | . | . | . | . | . | . | ✓ | . | . | . | ✓ |
| supiliac_skinfold | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| thigh_circumf | . | . | . | . | . | . | . | . | . | . | . | ✓ | . | . | . | . | . |
| triceps_skinfold | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| ALT | . | ✓ | ✓ | . | ✓ | . | ✓ | ✓ | ✓ | ✓ | . | ✓ | . | ✓ | ✓ | ✓ | . |
| AST | . | ✓ | ✓ | . | ✓ | . | ✓ | ✓ | . | ✓ | . | ✓ | . | ✓ | ✓ | ✓ | . |
| GGT | . | ✓ | ✓ | . | . | ✓ | . | ✓ | . | . | . | . | . | . | . | . | ✓ |
| Ethn_Asian | ✓ | ✓ | ✓ | . | ✓ | ✓ | ✓ | ✓ | ✓ | . | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | . |
| Ethn_African | . | ✓ | . | . | . | ✓ | . | ✓ | ✓ | ✓ | . | ✓ | . | ✓ | . | . | . |
| Somke_Ex | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Smoke_Current | . | ✓ | . | . | . | . | . | . | . | ✓ | . | ✓ | . | . | . | . | . |
| Sex_female | ✓ | ✓ | ✓ | . | ✓ | . | ✓ | . | ✓ | ✓ | . | ✓ | ✓ | . | ✓ | ✓ | ✓ |

Table D.4: Proportion of Ethnic sub-groups of origin in each cluster. (Max per row in bold).

| **Cluster** | **Africans** | **Gujarati** | **Irish** | **Muslim** | **Native** | **Other** | **Other** | **Punjabi** | **Punjabi** |
|---|---|---|---|---|---|---|---|---|---|
| **Number** | **Caribbean** | **Hindu** | | | **British** | **Europeans** | **South-Asians** | **Hindu** | **Sikh** |
| 1 | .07 | .02 | .07 | .04 | **.53** | .03 | .04 | .04 | .16 |
| 2 | .00 | .00 | .18 | .00 | **.73** | .09 | .00 | .00 | .00 |
| 3 | .00 | .00 | .10 | .00 | **.84** | .06 | .00 | .00 | .00 |
| 4 | .00 | .16 | .00 | .13 | .00 | .00 | .13 | .11 | **.46** |
| 5 | .00 | .00 | .00 | .00 | **.90** | .10 | .00 | .00 | .00 |
| 6 | .12 | .03 | .04 | .08 | **.32** | .01 | .07 | .04 | .29 |
| 7 | .02 | .07 | .01 | .11 | **.33** | .02 | .06 | .05 | **.33** |
| 8 | .00 | .08 | .03 | .10 | .11 | .00 | .12 | .11 | **.43** |
| 9 | .00 | .21 | .00 | **.26** | .00 | .00 | .21 | .16 | .16 |
| 10 | .03 | .11 | .00 | .16 | .19 | .02 | .08 | .03 | **.39** |

Table D.5: Anthropometric covariates, diabetes indicator, CVD indicators and other control variables. Every variable is used at time 1 ($T_1$) and time 2 ($T_2$)

| **Variable name** | **Label** |
|---|---|
| Age | Respondent's age |
| WHR | Waist to Hip Ratio |
| Systolic blood pressure | Systolic blood pressure |
| HOMA IR | Homeostasis model assessment Insulin Resistance |
| Sex female | Dummy variable for female sex (male as reference category) |
| Smoke_Ex | Indicator variable for ex smoker category |
| Smoke_Current | Indicator variable current smoker category |
| Pt_alcohol | Weekly quantity of alcohol consumed |
| Phys_score | Physical activity score |
| Edu_years | Years of education (assumed constant over time) |
| bl_hdl | Concentration of HDL |
| bl_trig | Concentration of blood Triglycerides |
| bl_chol | Concentration of blood Cholesterol |
| CHD | Indicator variable for the presence of Coronary Heart Disease up to $T_1$ and $T_2$ |
| stroke | Indicator variable for the presence of stroke up to $T_1$ and $T_2$ |
| Diabetes | Indicator variable for the presence of Diabetes Mellitus (type 2) up to $T_1$ and $T_2$ |
| dm_treat | Indicator variable for the presence of drug treatment for diabetes up to $T_1$ and $T_2$ |
| bp_treat | Indicator variable for the presence of blood pressure lowering drugs treatment up to $T_1$ and $T_2$ |
| lipids_treat | Indicator variable for the presence of blood lipids lowering drugs treatment up to $T_1$ and $T_2$ |

# Appendix E

# Figures

## E.1 Figures Exploratory Data Analysis

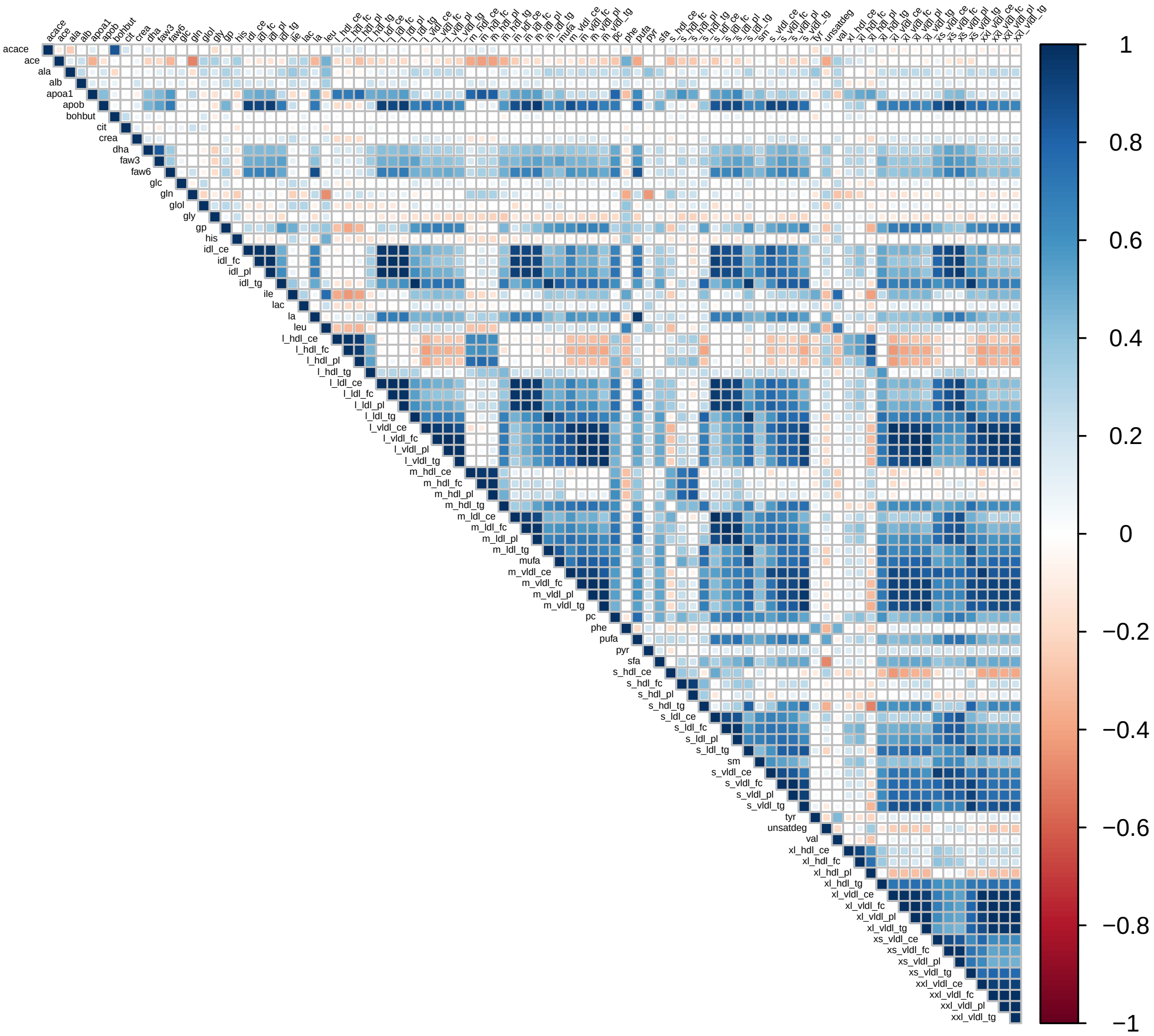


Figure E.1: Correlation Matrix between metabolites in the fasting dataset

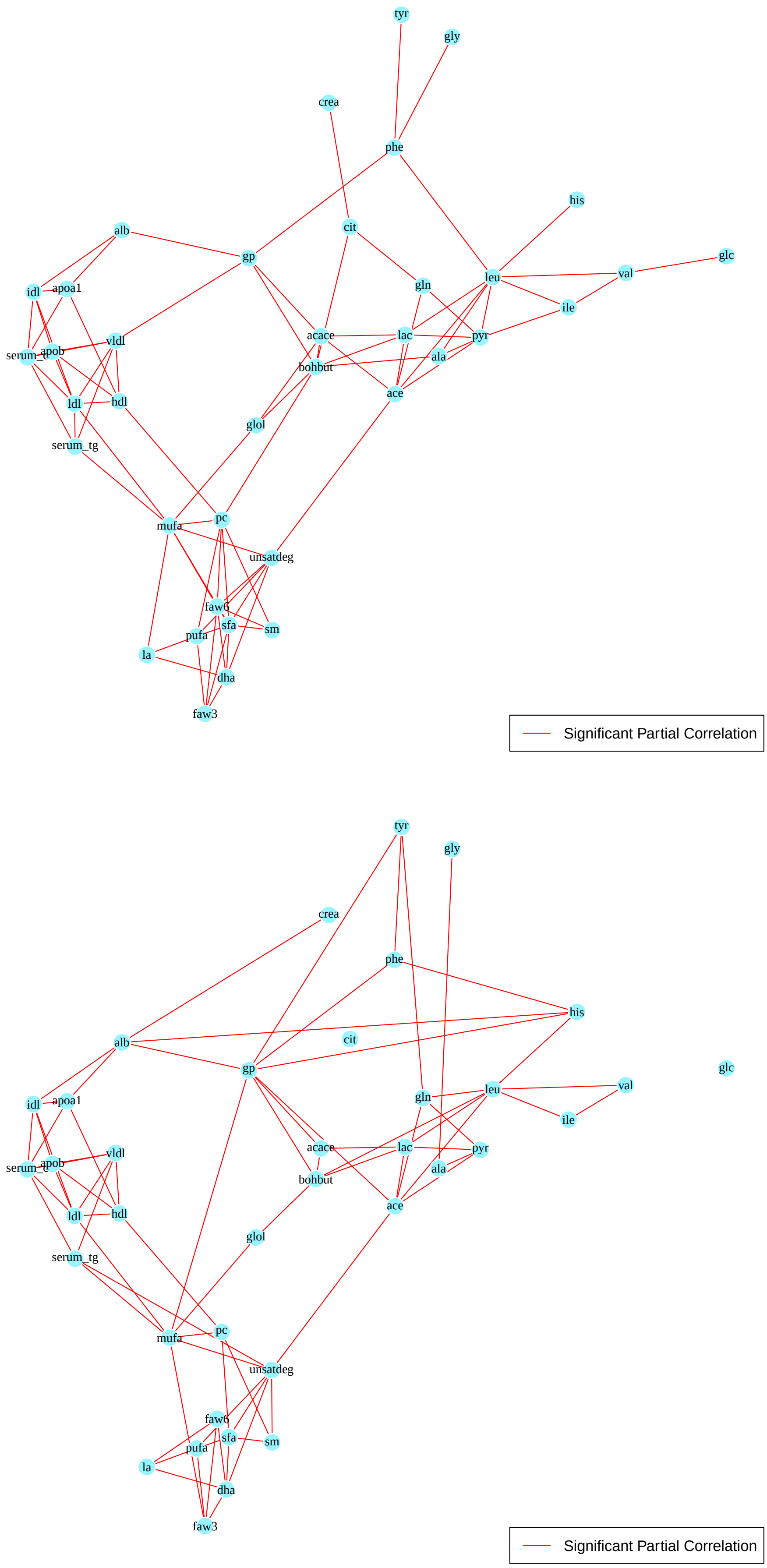


Figure E.2: Individual Networks for the Europeans (top panel) and the South-Asians (bottom panel)

## E.2 Figures Bayesian Nonparametric Modelling of HOMA IR

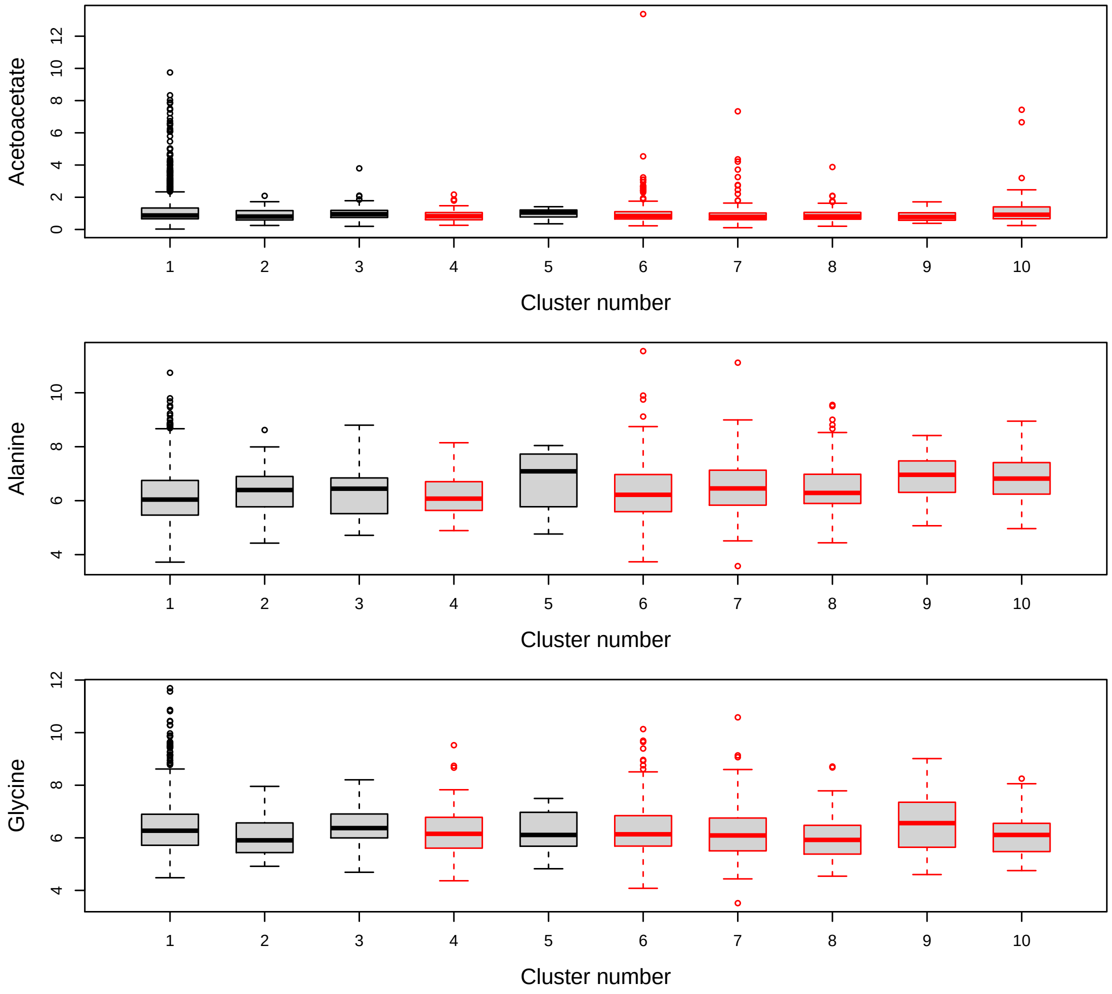


Figure E.3: From top to bottom, boxplots of Acetoacetate, Alanine and Glycine. Black boxplots indicate clusters with a majority of Europeans, while red boxplots indicate clusters with a majority of South-Asians.

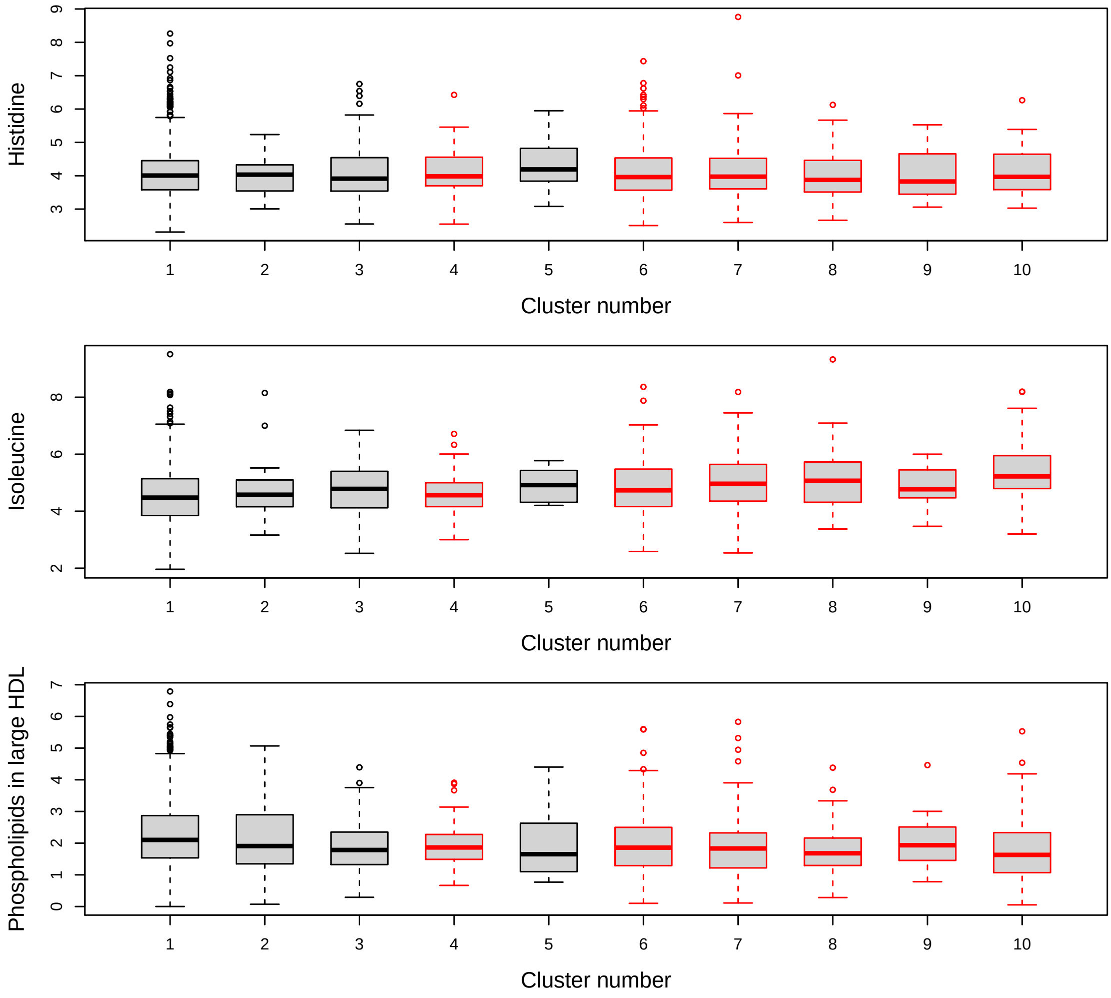


Figure E.4: From top to bottom, boxplots of Histidine, Isoleucine and Phospholipids in large HDL. Black boxplots indicate clusters with a majority of Europeans, while red boxplots indicate clusters with a majority of South-Asians.

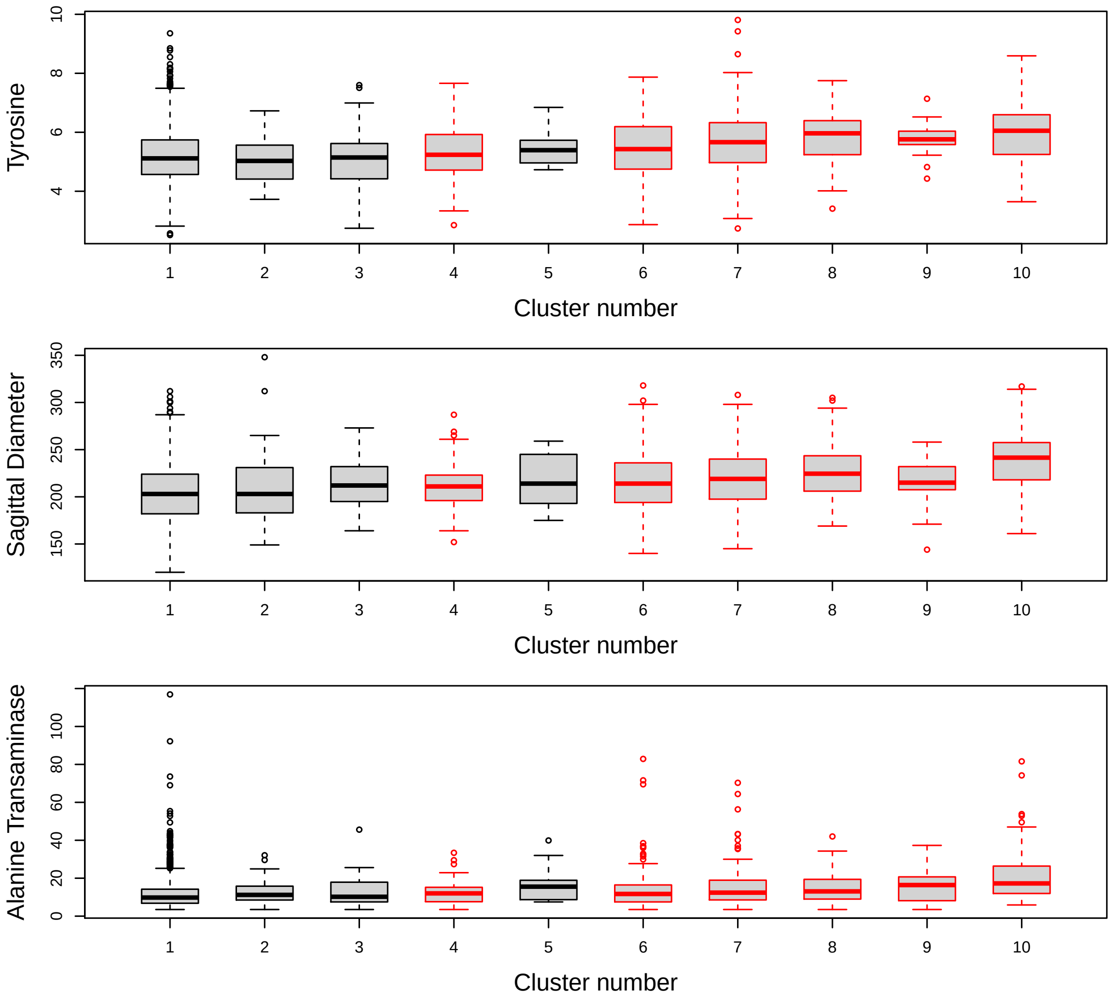
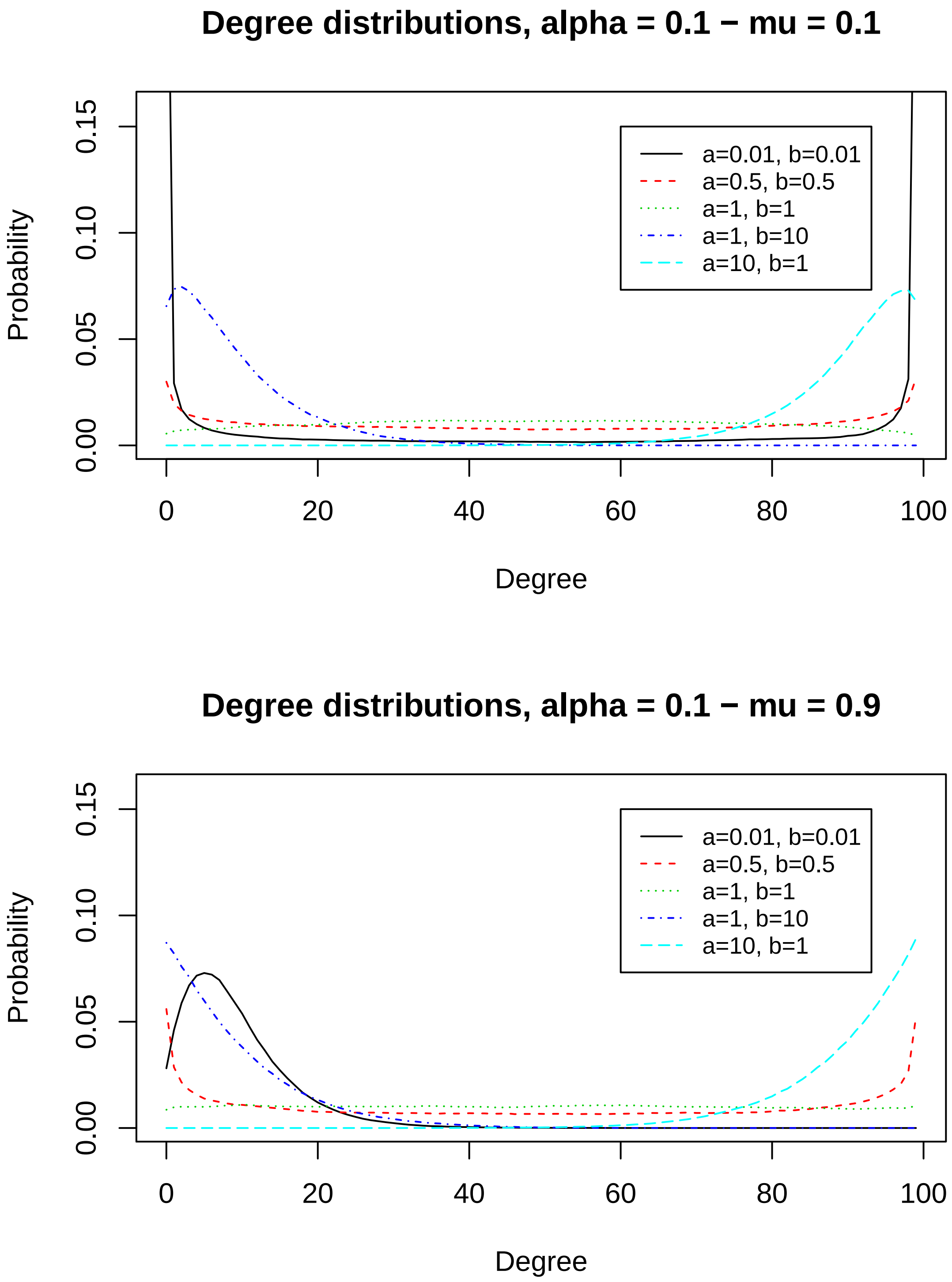


Figure E.5: From top to bottom, boxplots of Tyrosine, sagittal diameter and Alanine Aminotransferase. Black boxplots indicate clusters with a majority of Europeans, while red boxplots indicate clusters with a majority of South-Asians.

## E.3 Figures Bayesian Nonparametric Gaussian Graphical Models


Degree distributions, alpha = 0.1 − mu = 0.1

Degree distributions, alpha = 0.1 − mu = 0.9

Figure E.6: Degree Distribution for $\alpha = 0.1$ and $\mu = 0.1$ (top panel) and $\mu = 0.9$ (bottom panel) and different combinations of the Beta hyper-parameters $a_\pi, b_\pi$.

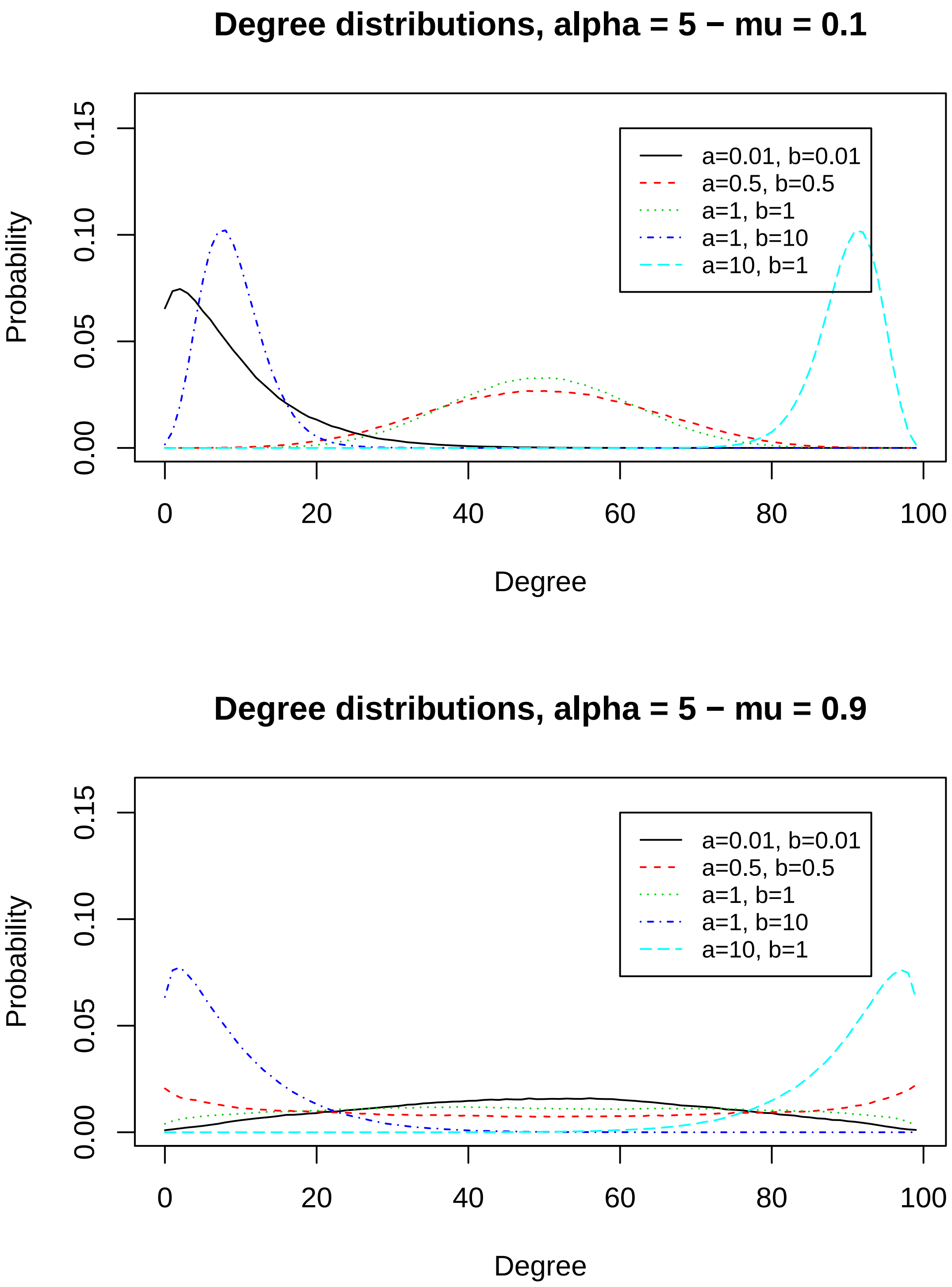


Figure E.7: Degree Distribution for $\alpha = 5$ and $\mu = 0.1$ (top panel) and $\mu = 0.9$ (bottom panel) and different combinations of the Beta hyper-parameters $a_\pi, b_\pi$.

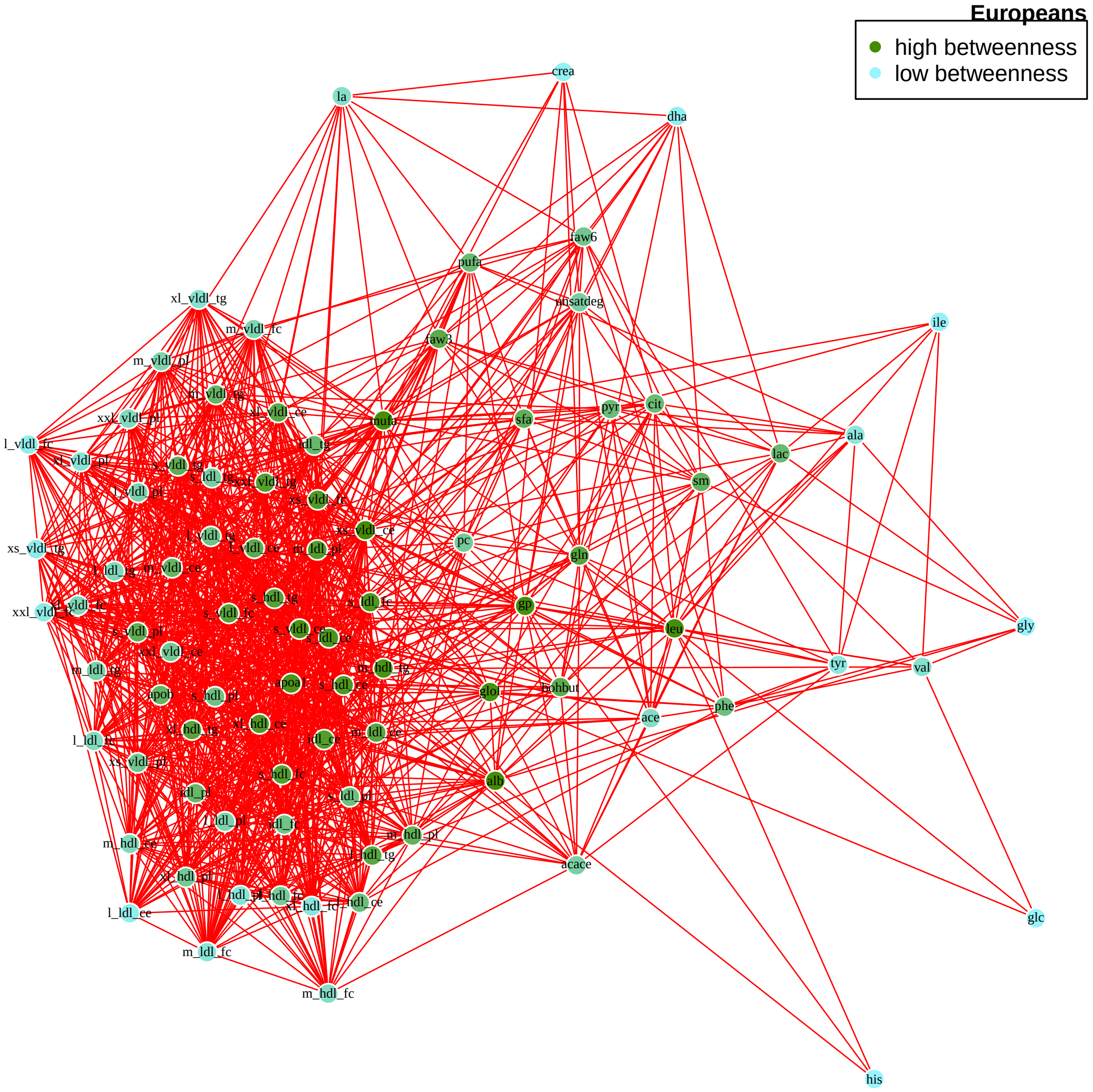


Figure E.8: Graph for the European ethnicity. An edge between two nodes is included in the graph if its posterior probability of inclusion is higher than 0.5.

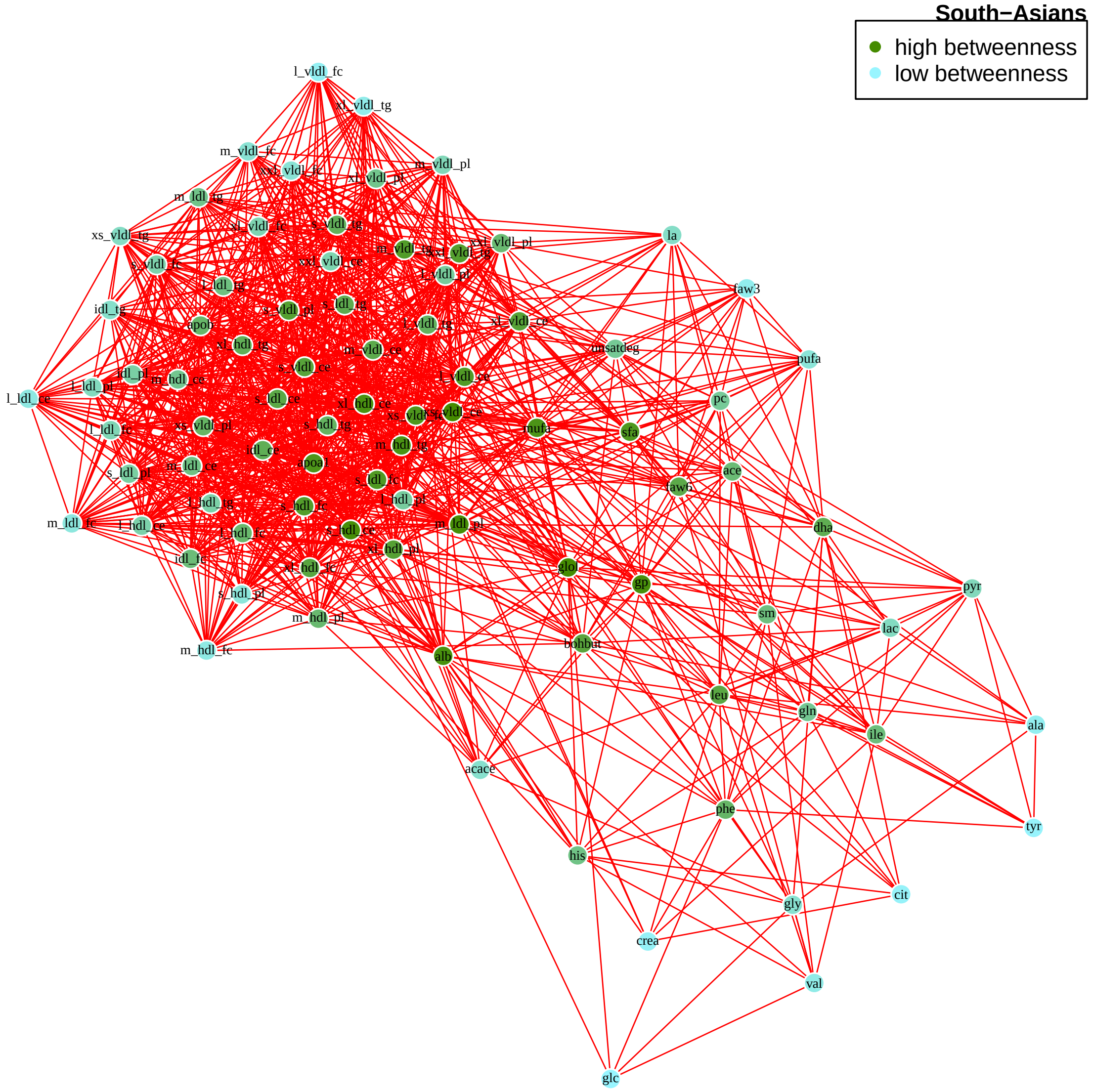


Figure E.9: Graph for the South-Asian ethnicity. An edge between two nodes is included in the graph if its posterior probability of inclusion is higher than 0.5.

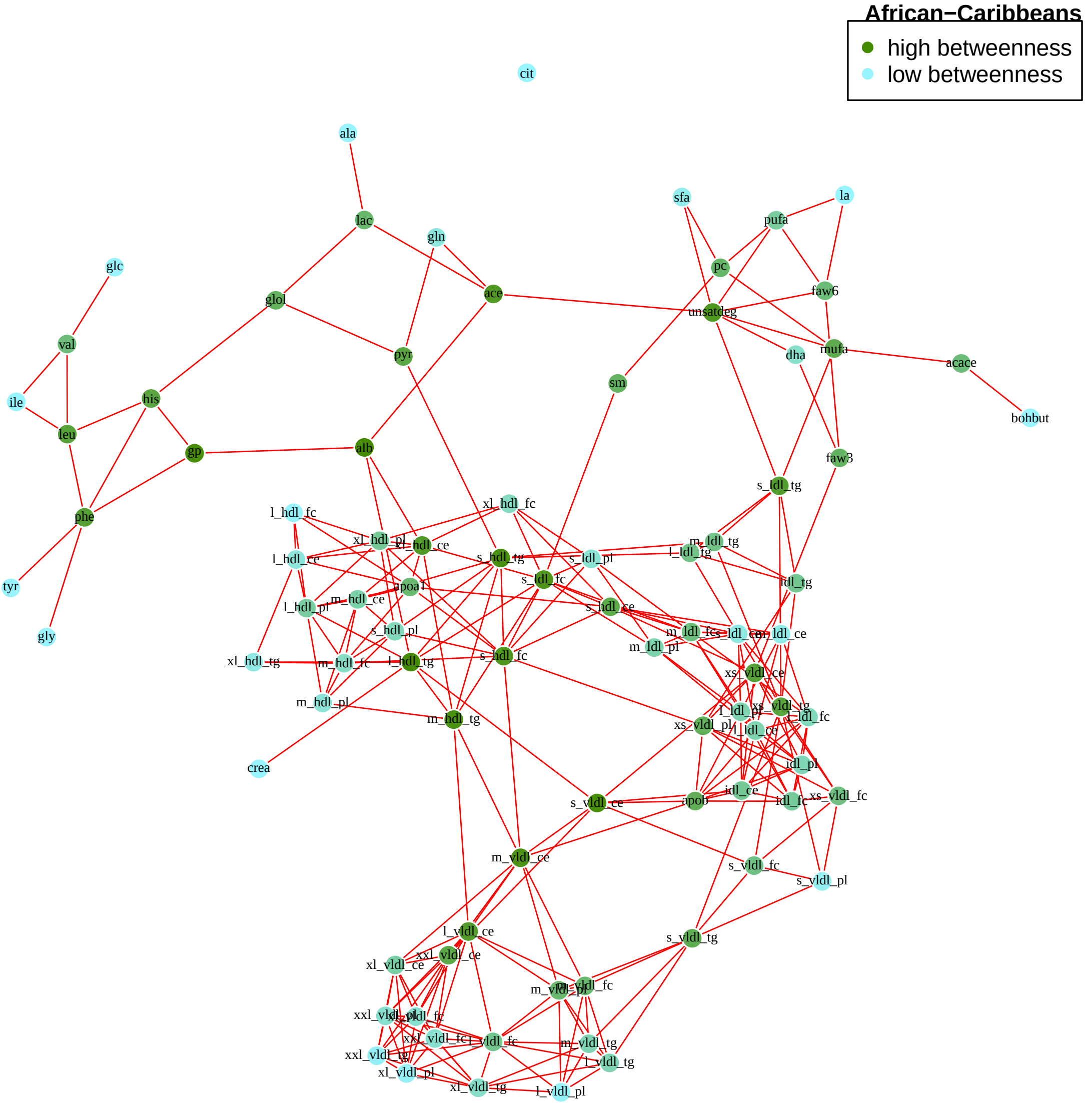


Figure E.10: Graph for the African-Caribbean ethnicity. An edge between two nodes is included in the graph if its posterior probability of inclusion is higher than 0.5.

## E.4 Figures Bayesian Dynamic Multiple Graphical Models

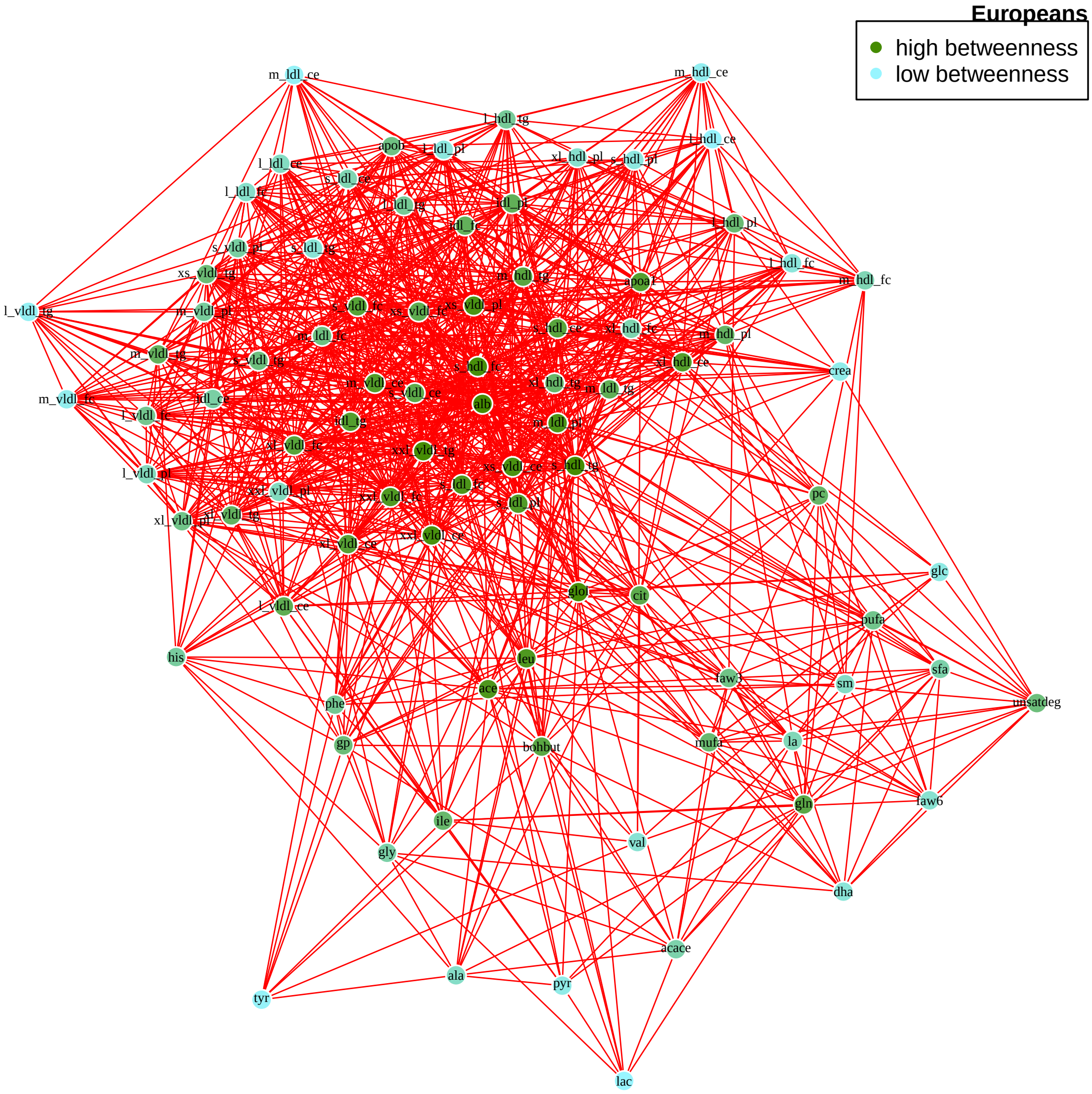


Figure E.11: Individual network for the European ethnicity at baseline. An edge between two nodes is included in the graph if its posterior probability of inclusion is higher than 0.8.

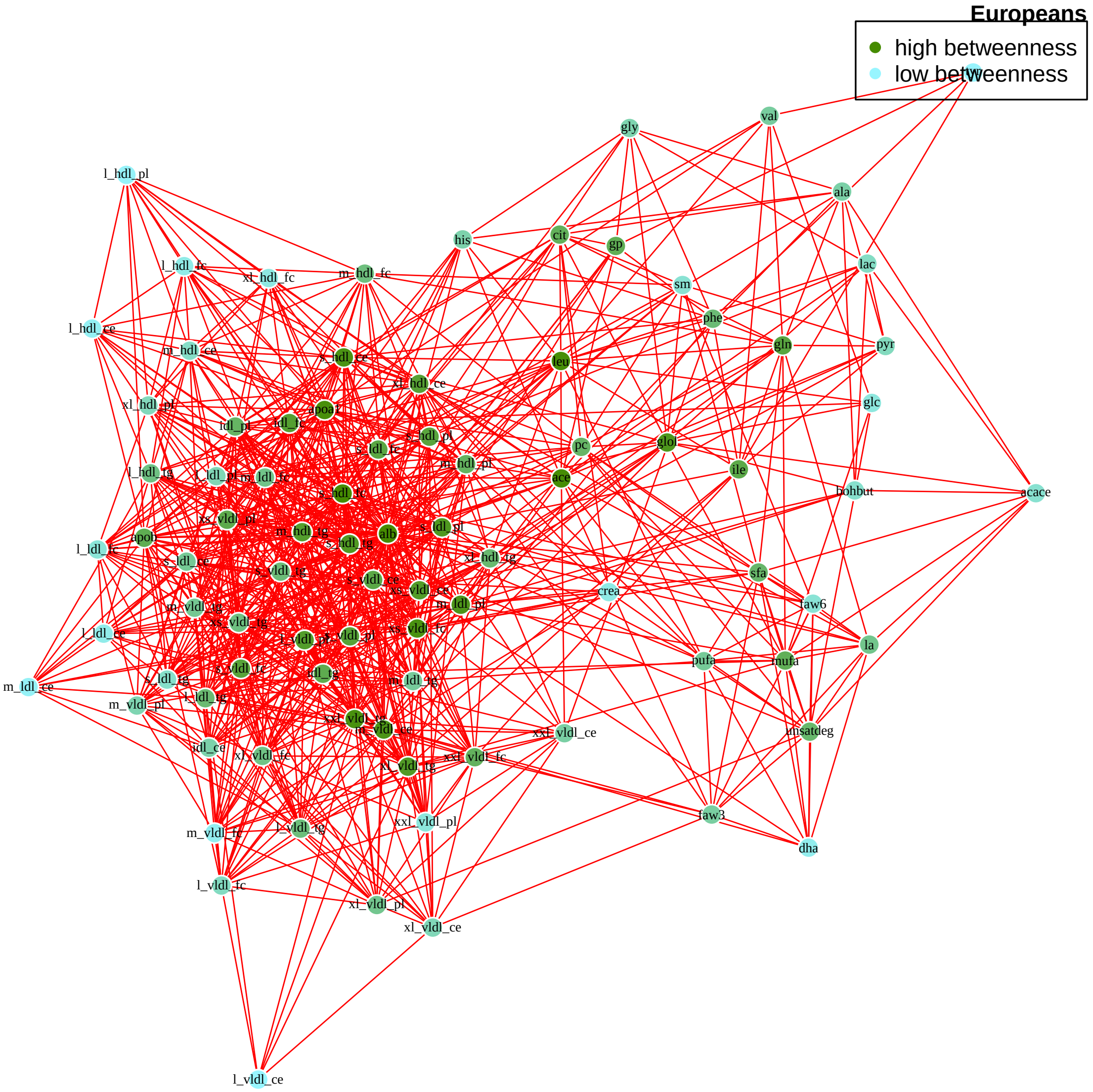


Figure E.12: Individual network for the European ethnicity at follow-up. An edge between two nodes is included in the graph if its posterior probability of inclusion is higher than 0.8

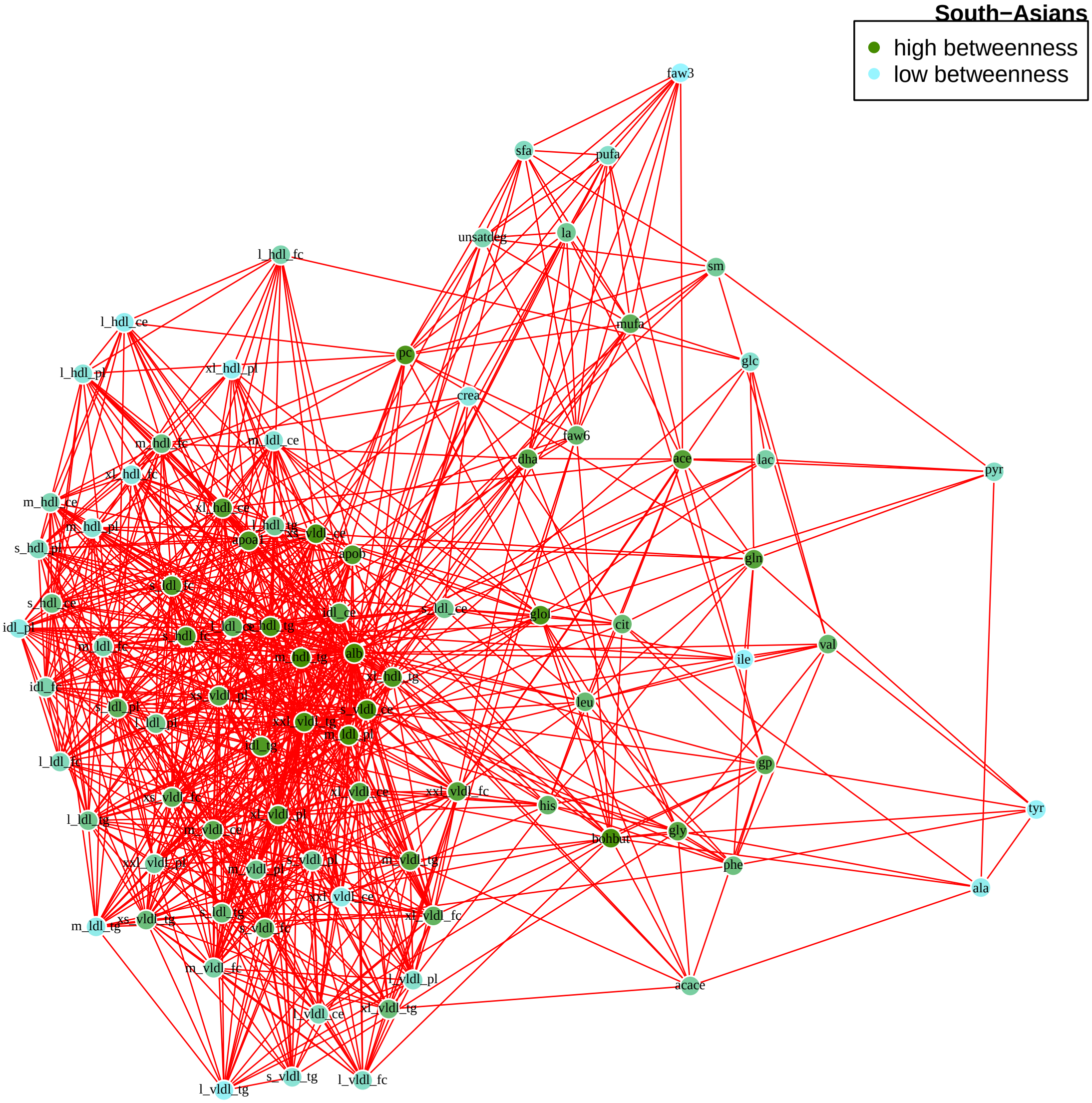


Figure E.13: Individual network for the South-Asian ethnicity at baseline. An edge between two nodes is included in the graph if its posterior probability of inclusion is higher than 0.8

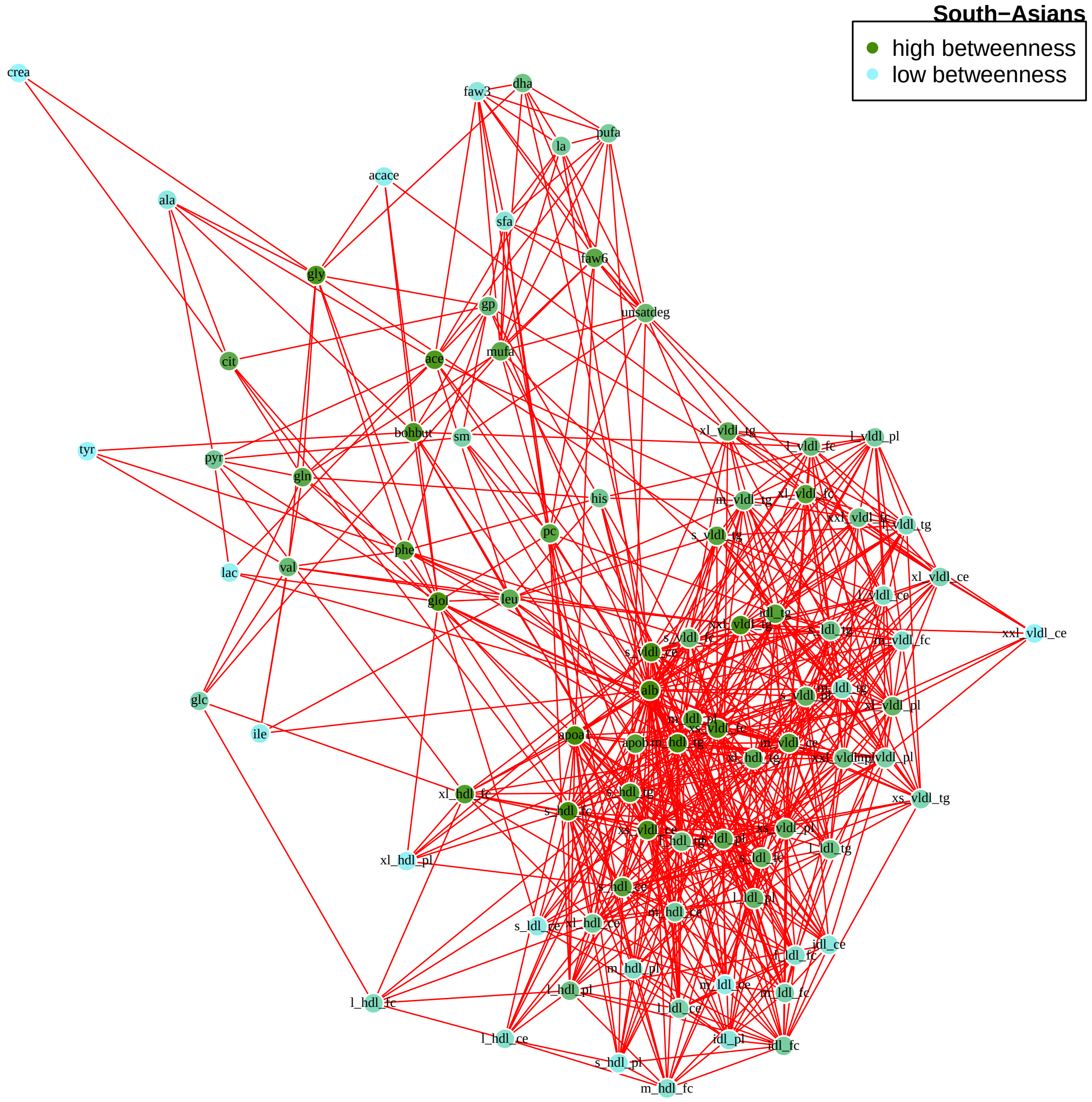


Figure E.14: Individual network for the South-Asian ethnicity at follow-up. An edge between two nodes is included in the graph if its posterior probability of inclusion is higher than 0.8

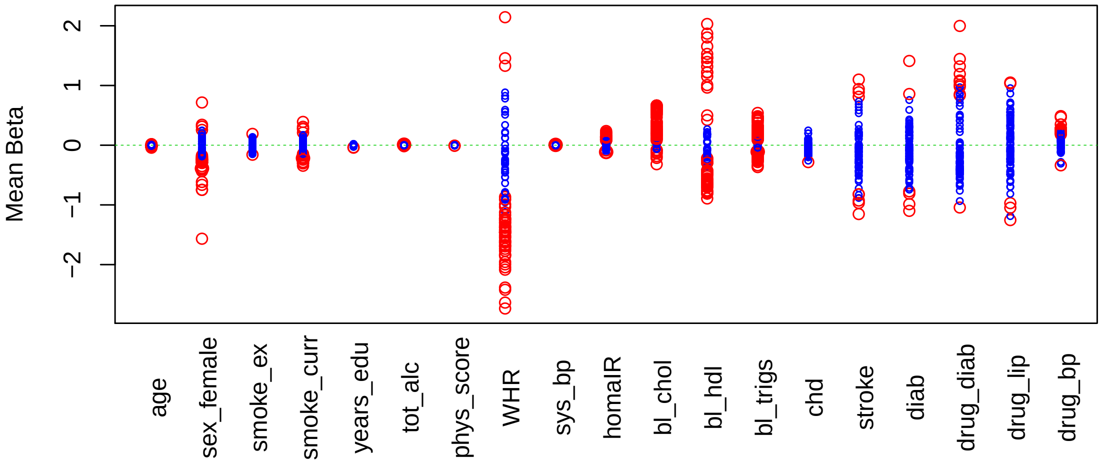


Figure E.15: Posterior means of $\beta_{lj}$ for the Europeans at baseline. Each dot represents the mean of the posterior distribution of a coefficient $\beta_{lj}$, $l = 1 \ldots, M$. Red dots denote coefficients whose 95% credible interval does not contain the zero.

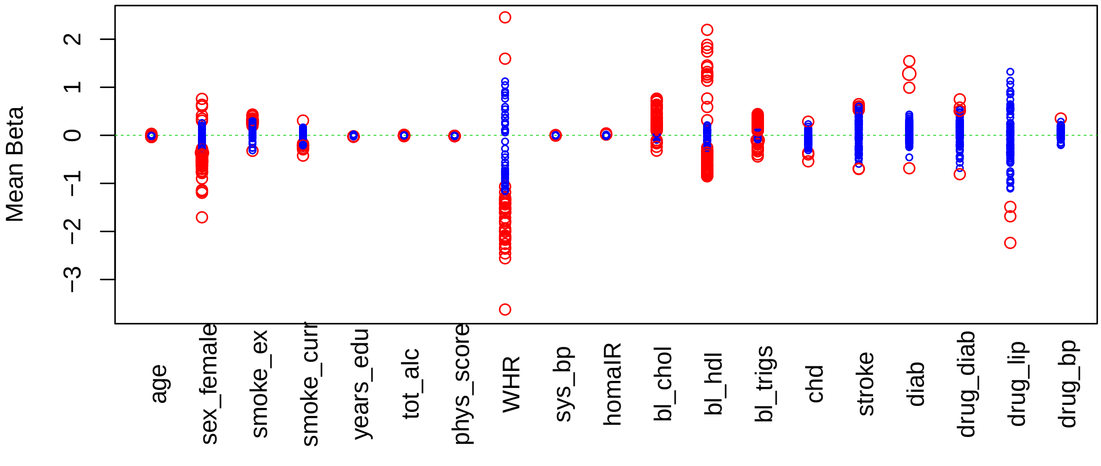


Figure E.16: Posterior means of $\beta_{lj}$ for the South-Asians at baseline. Each dot represents the mean of the posterior distribution of a coefficient $\beta_{lj}$, $l = 1 \ldots, M$. Red dots denote coefficients whose 95% credible interval does not contain the zero.

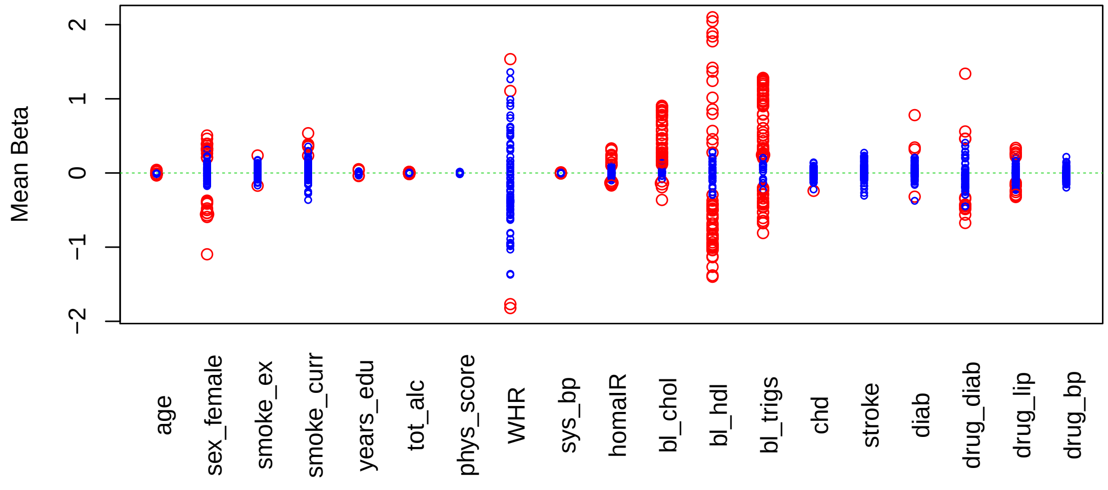


Figure E.17: Posterior means of $\beta_{lj}$ for the Europeans at follow-up. Each dot represents the mean of the posterior distribution of a coefficient $\beta_{lj}$, $l = 1\ldots, M$. Red dots denote coefficients whose 95% credible interval does not contain the zero.

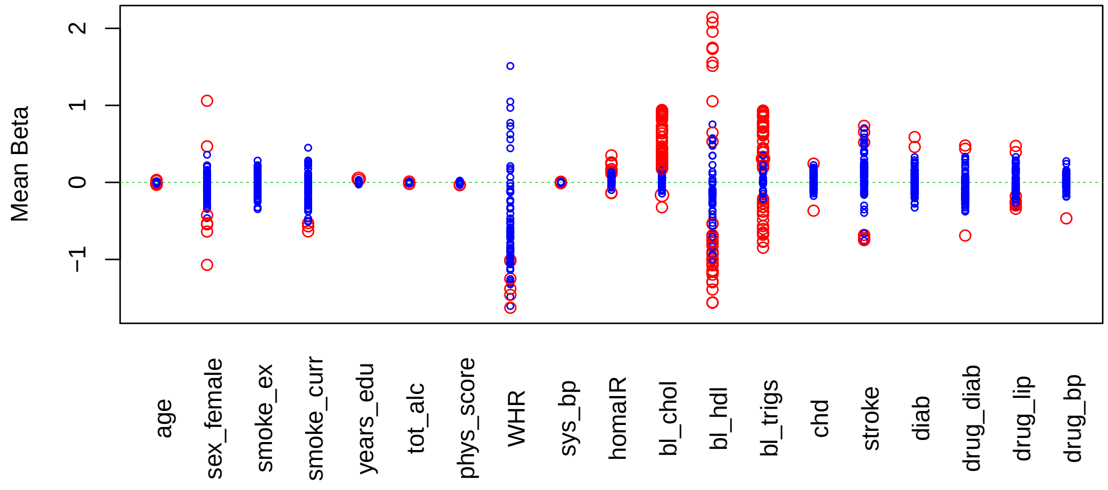


Figure E.18: Posterior means of $\beta_{lj}$ for the South-Asians at follow-up. Each dot represents the mean of the posterior distribution of a coefficient $\beta_{lj}$, $l = 1\ldots, M$. Red dots denote coefficients whose 95% credible interval does not contain the zero.