

Bayesian dynamic network modelling: an application to metabolic associations in cardiovascular diseases

Marco Molinari, Andrea Cremaschi, Maria De Iorio, Nishi Chaturvedi, Alun Hughes & Therese Tillin

To cite this article: Marco Molinari, Andrea Cremaschi, Maria De Iorio, Nishi Chaturvedi, Alun Hughes & Therese Tillin (2024) Bayesian dynamic network modelling: an application to metabolic associations in cardiovascular diseases, Journal of Applied Statistics, 51:1, 114-138, DOI: [10.1080/02664763.2022.2116746](https://doi.org/10.1080/02664763.2022.2116746)

To link to this article: <https://doi.org/10.1080/02664763.2022.2116746>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 02 Sep 2022.



[Submit your article to this journal](#)



Article views: 2073



[View related articles](#)



[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)

Bayesian dynamic network modelling: an application to metabolic associations in cardiovascular diseases

Marco Molinari^a, Andrea Cremaschi^b, Maria De Iorio ^{a,b,c}, Nishi Chaturvedi^d, Alun Hughes^d and Therese Tillin^d

^aDepartment of Statistical Science, University College, London, London, UK; ^bSingapore Institute for Clinical Sciences, A*STAR, Singapore; ^cYong Loo Lin School of Medicine, National University of Singapore, Singapore; ^dDepartment of Population Science and Experimental Medicine, University College London, London, UK

ABSTRACT

We propose a novel approach to the estimation of multiple Graphical Models to analyse temporal patterns of association among a set of metabolites over different groups of patients. Our motivating application is the Southall And Brent REvisited (SABRE) study, a tri-ethnic cohort study conducted in the UK. 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 a nodewise regression approach to infer the structure of the graphs, borrowing information across time as well as across ethnicities. The response variables of interest are metabolite levels measured at two time points and for two ethnic groups, Europeans and South-Asians. We use nodewise regression to estimate the high-dimensional precision matrices of the metabolites, imposing sparsity on the regression coefficients through the dynamic horseshoe prior, thus favouring sparser graphs. We provide the code to fit the proposed model using the software `Stan`, which performs posterior inference using Hamiltonian Monte Carlo sampling, as well as a detailed description of a block Gibbs sampling scheme.

ARTICLE HISTORY

Received 5 October 2020
Accepted 14 August 2022


KEYWORDS

Dynamic shrinkage priors;
Gibbs sampling; graphical
models; metabolomics;
nodewise regression

1. Introduction

The Southall And Brent REvisited (SABRE) [68] population based cohort study was initiated in the late 1980's in north-west London with the aim of studying ethnic differences in cardiovascular diseases and diabetes. The study includes individuals of European, South-Asian and African-Caribbean descent, aged 40–69 years at baseline. Recently, metabolites measurements, obtained using nuclear magnetic resonance spectroscopy [61], have been collected on over 3000 stored blood samples at baseline and 20-years follow-up. All participants gave written informed consent. Approval for the baseline study was obtained from Ealing, Hounslow and Spelthorne, Parkside, and University College London research ethics

CONTACT Maria De Iorio  m.deiorio@ucl.ac.uk

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/02664763.2022.2116746>.

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

committees, and at follow-up from St. Mary's Hospital Research Ethics Committee (reference 07/H0712/109). In this work, we investigate the complex metabolic pathways involved in cardiac metabolism, combining metabolite concentration data with more traditional clinical makers. Metabolites are small molecules that participate in metabolic reactions and are involved in biochemical pathways associated with metabolism in health and disease [19]. Indeed, there is increasing evidence that most cardiovascular diseases, such as diabetes mellitus, involve disturbances in cardiac metabolism and that, on the other hand, heart diseases can affect the metabolism, hence initiating a vicious cycle [49]. In the last two decades, metabolomics (i.e. the large-scale study of metabolites within cells, bio-fluids, tissues or organisms) has emerged as a powerful tool for defining changes in metabolism that occur across a range of cardiovascular disease states. Findings from metabolomic 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.

Here, we focus on type 2 diabetes, a main clinical outcome of interest in the SABRE study, as it poses an enormous individual and societal burden, with high risk of major complications and diminished quality and length of life. Hence, it is imperative to understand causal mechanisms in order to identify subjects 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 estimated to be 2–4 times higher than that of the general population [63]. Research to date suggests that insulin resistance and differences in body fat distribution explain some of the ethnic differences in diabetes risk, but the underlying mechanistic pathways are poorly understood, although they are likely to involve a complex interplay between environmental, behavioural, metabolic, genetic and epigenetic influences. Therefore, it is essential not to limit the analysis to differences in metabolite levels across ethnicities over time, but to also consider patterns of variations in metabolic associations to gain better insight in molecular mechanisms of disease pathogenesis and to formulate novel data-driven scientific hypotheses.

In the SABRE study measurements are available on over 200 metabolite concentrations, covering a wide range of chemical classes, at baseline and 20-years follow up. 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 [50]. In the nodewise regression approach, 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/metabolites, and each variable is regressed on all the others. It is a local method, because it infers the neighbourhood 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. The element j, l in the precision matrix is estimated to be non-zero (implying an edge in the graph between the two variables) if either the estimated regression coefficient of variable j on l , or the estimated coefficient of variable l on j , is non-zero. Alternative approaches to graphical models estimation are available in the literature. Graphical Lasso is a popular global method in both frequentist and Bayesian domain, based on a penalised maximum likelihood estimator [24] or on a double exponential prior [74], respectively. Graphical Lasso has the advantage of

ensuring a positive definite estimate of Ω , but requires a greater computational effort and it is less flexible in estimating the individual scaling levels (i.e. the diagonal elements of Ω) compared to the nodewise approach [36]. 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 precision matrix is selected. For example, [42] propose an algorithm to perform posterior inference on graph space assuming a 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. A more efficient algorithm, based on a birth-death MCMC, has been proposed by [51].

In several applications, including the one presented in this work, information regarding the grouping of the subjects is available, e.g. different ethnic groups. Therefore, it is of interest to estimate a group-specific graph, by specifying a joint model for multiple graphs able to identify common structure as well as group-specific connections. This is particularly relevant in our application as it allows to highlight biological mechanism which differ between ethnicity and potentially lead to different disease development. Examples of models for multiple graphs are found in the work of [60], which specify a global penalisation and use optimisation techniques, and of [54], which in a Bayesian framework propose a joint model involving a Markov random field prior to encourage sharing of information among edges. A Markov random field prior modelling both spatial and temporal dependence is also used by [44]. Additional examples are the approaches of [65], where shared structures are modelled via a multiplicative prior, and of [8], where multiple precision matrices are estimated via a penalisation approach which allows the inclusion of information a-priori.

The main methodological contribution of this work is to extend nodewise regression to (i) achieve sparsity in the resulting precision matrix (ii) accommodate multiple groups (iii) account for different time points. This will allow not only to better understand group-specific and global dependence structure, but also how they evolve across time. To this end, we exploit recent contribution on shrinkage prior and dynamic stochastic volatility models. In a nodewise regression framework, inducing sparsity in the graph is equivalent to estimating some regression coefficients equal to zero. A wealth of proposals is available in the literature to impose sparsity on regression coefficients (see, for instance, [53] for a review on shrinkage priors). We opt for the Horseshoe prior, which belongs to the class of continuous global-local shrinkage priors (see, for example, [2,6,12]). The Horseshoe prior 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. Linking the hyper-parameters of the Horseshoe priors across groups and employing the dynamic extension of the Horseshoe prior proposed by [40], we are able to accurately estimate dynamic evolving complex precision matrices, across multiple groups of observations of different sample sizes at different time points. A similar approach is taken by [44], where the neighbourhood selection of each node is based on the nodewise regression method of [50], but it is different from our approach in two ways: (i) a Spike and Slab prior is imposed on the regression coefficients to induce sparsity on edge selection, increasing computational cost; (ii) information is shared across multiple groups and time points through a Markov Random Field structure, which also involves neighbourhood selection. Our choice of a Horseshoe prior, which is continuous and does

not require hyper-parameter tuning (in its original formulation), significantly speeds up computations, allowing the use of different inference approaches, such as Markov chain Monte Carlo and Hamiltonian Monte Carlo (HMC). Furthermore, information across multiple groups is shared by imposing a hierarchical structure on the hyper-parameters of the horseshoe prior, while time dependence is incorporated through an autoregressive model specified on the regression coefficients as proposed by [40]. Both extensions are computationally efficient, being also suitable for parallel computations on multicore machines. Moreover, the approach proposed by [40] scales linearly in the number of time points. A limitation of the neighbourhood selection based on nodewise regression is the lack of a direct posterior estimate of the joint precision matrix. As detailed in [44], the Bayesian version of nodewise regression allows to accurately estimate the 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 to approximate the precision matrix and as a measure of the strength and direction of each specific connection in the graph.

The paper is organised as follows. In Section 2 we introduce nodewise regression, its extensions and discuss prior specification. In Section 3 we show the performance of the proposed model in simulations. Section 4 illustrates the application to the SABRE study. Section 5 concludes the paper summarising main results and contributions.

2. Methodology

In this Section, we introduce nodewise regression and its use for graph estimation. We first provide a detailed description of the basic model for a single graph. Next, we explain our strategy to extend nodewise regression to multiple graphs and to the more general dynamic multiple graphs framework. Hence, we explain how to perform posterior inference on such models.

2.1. Nodewise regression for graphical models

We first introduce basic concepts from graph theory. For more details, we refer to [41]. Let $G = (V, E)$ be an undirected graph, with vertex set $V = \{1, \dots, M\}$ and edge set $E \subset \{(j, l) \in V \times V : j < l\}$. The vertices of the graph are associated with a M -dimensional vector of variables $\mathbf{y} = (y_1, \dots, y_M)$ assumed to follow a multivariate Normal distribution

$$\mathbf{y} \sim N_M(\boldsymbol{\mu}, \Omega) \quad (1)$$

where $\boldsymbol{\mu}$ is the mean vector and $\Omega = [\omega_{jl}]_{j,l=1}^M$ is the $M \times M$ precision matrix. In the context of Gaussian Graphical Models (GGM) [17], it is standard practice to specify a prior distribution on G and, then, conditional on G , a prior for Ω as there is a direct correspondence between the elements of the precision matrix Ω and the edges in the graph G . An edge is present between nodes V_j and V_l , that is $(j, l) \in E$, if and only if $\omega_{jl} \neq 0$ [41,76]. If $\omega_{jl} = 0$ (absence of an edge), then y_j and y_l are conditionally independent given the remaining variables \mathbf{y}_{-jl} , where \mathbf{y}_{-jl} denotes the random vector \mathbf{y} excluding the elements j and l . In this context the main object of inference is the graph G and usually Ω is treated as a latent variable. Posterior inference is often complicated by the complexity of graph space.

In this paper we opt for a different approach as it is more scalable to higher dimensions. To estimate the graph G we use nodewise regression [50], a technique that exploits the relationship between the partial correlation coefficients and the regression coefficients of a linear regression, without directly specifying a probability model for G . Recall that for the standard linear regression

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

the regression coefficients β_{jl} can be expressed in terms of partial correlations ω_{jl} , i.e. $\beta_{jl} = \frac{-\omega_{jl}}{\omega_{ll}}$ and analogously $\beta_{lj} = \frac{-\omega_{lj}}{\omega_{jj}}$. 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. Here, σ_l^2 denotes the error variance. In our context, consider the partition where the random variable y_l is the l -th coordinate of \mathbf{y} and \mathbf{y}_{-l} corresponds to the remaining coordinates. The conditional distribution of y_l given \mathbf{y}_{-l} is

$$y_l \mid \mathbf{y}_{-l}, \boldsymbol{\mu}, \Omega \sim \mathcal{N} \left(\mu_l - \sum_{j \neq l} \frac{\omega_{jl}}{\omega_{ll}} (y_j - \mu_j), \frac{1}{\omega_{ll}} \right) \quad (3)$$

where $\omega_{jl}/\omega_{ll} = \beta_{jl}$, for $j \neq l$. In many application, as in Section 4, μ_l captures covariate effects on the response y_l , i.e. $\mu_l = \mathbf{Z}_l \boldsymbol{\eta}_l$, where \mathbf{Z}_l is the covariate vector and $\boldsymbol{\eta}_l$ is the vector of regression coefficients. In this case, the nodewise regression model is defined on the responses centred on the covariate term. For ease of explanation, in the following sections we assume, without loss of generality, that $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M) = \mathbf{0}$.

This framework allows us to express the problem of graphical model selection as M independent linear regression problems, since a regression coefficient estimated equal to zero implies a zero in the corresponding element of the precision matrix and, implicitly, a lack of an edge in the underlying graph. For this reason, we are interested in identifying regression coefficients equal to zero. In the Bayesian framework, a wealth of sparse Bayesian regression techniques is available. When performing variable selection, a popular choice is to impose shrinkage priors on the regression coefficients. Typical examples include the class of two components discrete mixture priors, known as spike and slab [26], and the class of continuous shrinkage priors, of which examples are the Horseshoe and the Horseshoe+ prior see [6] 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 high number of parameters, due to the large state space. In contrast, continuous priors are easier to implement and are usually more computationally efficient, although the probability for the coefficient to be exactly zero is null and further thresholding is required. We opt for the Horseshoe prior [12], which is a scale mixture prior defined as

$$\beta_j \mid \lambda_j, \tau \sim \mathcal{N} \left(0, \lambda_j^2 \tau^2 \right)$$

$$\lambda_j, \tau \stackrel{ind}{\sim} \mathcal{C}^+(0, 1)$$

where $C^+(0, 1)$ denotes the standard half-Cauchy distribution, with probability density function

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

The Horseshoe belongs to the family of global-local shrinkage priors, as the global (i.e. common to all β_j) parameter τ shrinks all the coefficients towards zero, while the thick half-Cauchy tails for the local scales λ_j allow some regression coefficients β_j to escape the shrinkage and counter-balance the effect of τ . Moreover, the Horseshoe prior is characterised by a singularity at zero to strongly shrink small or negligible coefficients, while leaving the important ones unaffected thanks to its heavy tails (given by the half-Cauchy distribution).

We now describe the proposed model for single graph estimation. Let \mathbf{Y} be a $n \times M$ matrix of observations, where n is the sample size and M is the number of random variables/nodes. Each column $\mathbf{y}_l = (y_{1l}, y_{2l}, \dots, y_{nl})^T$, for $l = 1, \dots, M$, contains the measurements of the l -th variable. The regression model for the l -th column can be written as

$$\begin{aligned} \mathbf{y}_l \mid \boldsymbol{\beta}_l, \sigma_l^2 &\sim N_n(\mathbf{X}\boldsymbol{\beta}_l, \sigma_l^2 I_n) \\ \beta_{jl} \mid \lambda_{jl}, \tau_l &\sim N\left(0, \lambda_{jl}^2 \tau_l^2\right) \\ \sigma_l^2 \mid a_\sigma, b_\sigma &\sim \text{Inverse} - \text{Gamma}(a_\sigma, b_\sigma) \\ \lambda_{jl} &\sim C^+(0, 1) \\ \tau_l &\sim C^+(0, 1) \end{aligned} \tag{4}$$

where \mathbf{X} is the matrix of explanatory variables given by \mathbf{Y}_{-l} (i.e. \mathbf{Y} excluding the l -th column) and $\boldsymbol{\beta}_l = (\beta_{1l}, \beta_{2l}, \dots, \beta_{pl})$, with $p = M - 1$, is a vector of regression coefficients for the l -th regression. A pseudo inclusion probability parameter κ_{jl} is defined by [12] as

$$\kappa_{jl} = \frac{1}{1 + \text{Var}(\beta_{jl} \mid \lambda_{jl}, \tau_l)} = \frac{1}{1 + \lambda_{jl}^2 \tau_l^2} \tag{5}$$

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 and leading to the inclusion of the variable/edge in the graph. As previously shown [12], variable selection based on (5) (with a threshold level of 0.5) performs similarly to explicit variable selection based on spike and slab prior. Throughout this work, we follow this criterion to decide which edges to include in the graph.

To improve computational efficiency, we use the representation of the standard half-Cauchy distribution employed by [25,55]. The standard half-Cauchy distribution can be expressed as the product of a standard half-Normal distribution times the square root of an Inverse-Gamma distribution. Let $z \sim N^+(0, 1)$ and $y \sim \text{Inverse} - \text{Gamma}(1/2, 1/2)$ and define $x = z\sqrt{y}$, then $x \sim C^+(0, 1)$. $N^+(0, 1)$ denotes the standard half-Normal distribution, which is defined as the absolute value of a Normal distribution [43]. This re-parametrisation can help to avoid divergent transitions in an HMC algorithm (a problem commonly encountered with funnel shaped distributions). As suggested by [55], we

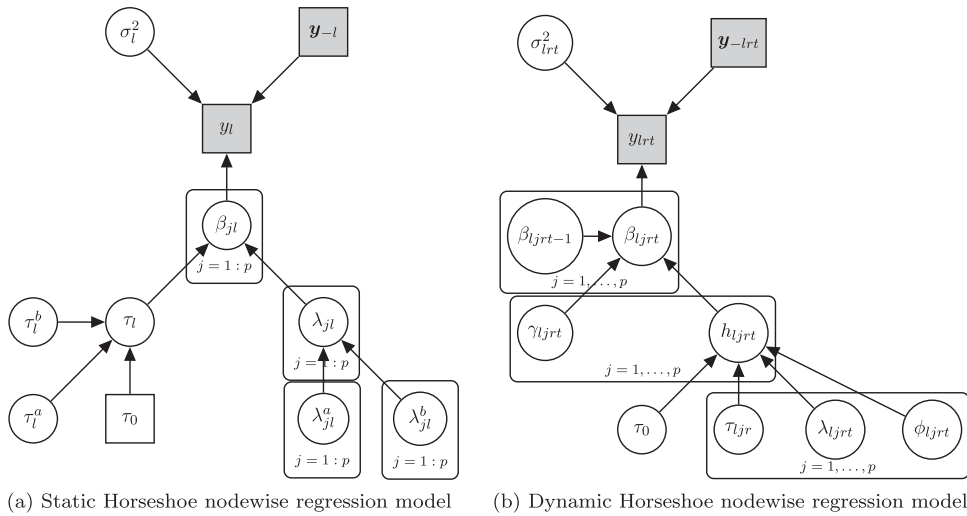


Figure 1. Summary of the nodewise regression models, highlighting the relationship between the observations and the parameters of the model. Circles represent random variables, while squares correspond to observations or fixed hyper-parameters. (a) Static Horseshoe model for each observed variable y_l , with $l = 1, \dots, M$, when only a group of observations is under investigation at one time point. (b) Dynamic Horseshoe model for each observed variable y_{lrt} at time t and within group r , with $l = 1, \dots, M$.

also allow for a tunable global scale parameter τ_l , which can help achieving the desired level of sparsity. Thus the prior distribution on the regression parameters becomes

$$\begin{aligned}
 \beta_{jl} \mid \lambda_{jl}, \tau_l &\sim N(0, \lambda_{jl}^2 \tau_l^2) \\
 \lambda_{jl} &= \lambda_{jl}^a \sqrt{\lambda_{jl}^b} \\
 \tau_l &= \tau_l^a \sqrt{\tau_l^b} \tau_0 \\
 \lambda_{jl}^a &\sim N^+(0, 1) \\
 \lambda_{jl}^b &\sim \text{Inverse - Gamma}(1/2, 1/2) \\
 \tau_l^a &\sim N^+(0, 1) \\
 \tau_l^b &\sim \text{Inverse - Gamma}(1/2, 1/2)
 \end{aligned} \tag{6}$$

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 [55]. A schematic representation of model (6) is shown in Figure 1(a).

2.1.1. Extension to multiple groups

We now extend the Horseshoe prior to allow borrowing information across multiple groups of observations. The groups are usually defined by the problem under investigation, for example they might correspond to different biological conditions, disease status or spatial regions. Estimating a single graphical model would lead to an implicit assumption of homogeneity of the underlying graphs across groups, with a consequent loss of information

about their heterogeneity and a 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. Several approaches for joint inference of multiple GGMs have been recently developed (see, for example, [14,15,30,52,54,65,67,75]). We opt for nodewise regression for computational convenience and ease of extension to more complex setups.

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 $r = 1, \dots, R$. We introduce dependence across multiple graphs through the group specific global shrinkage parameter τ_{lr} of the Horseshoe prior, extending easily model (6) to this setup. The model is now defined as follows

$$\begin{aligned}
 y_{lr} \mid \boldsymbol{\beta}_{lr}, \sigma_{lr}^2 &\sim N_p(\mathbf{X}_r \boldsymbol{\beta}_{lr}, \sigma_{lr}^2 I_{n_r}) \\
 \boldsymbol{\beta}_{jlr} \mid \lambda_{jlr}, \tau_{lr} &\sim N\left(0, \lambda_{jlr}^2 \tau_{lr}^2\right) \\
 \lambda_{jlr} &= \lambda_{jlr}^a \sqrt{\lambda_{jlr}^b} \\
 \tau_{lr} &= \tau_l^a \sqrt{\tau_l^b} \tau_{0r} \\
 \sigma_{lr}^2 \mid a_\sigma, b_\sigma &\stackrel{\text{ind}}{\sim} \text{Inverse - Gamma}(a_\sigma, b_\sigma) \\
 \lambda_{jlr}^a &\sim N^+(0, 1) \\
 \lambda_{jlr}^b &\sim \text{Inverse - Gamma}(1/2, 1/2) \\
 \tau_l^a &\sim N^+(0, 1) \\
 \tau_l^b &\sim \text{Inverse - Gamma}(1/2, 1/2)
 \end{aligned} \tag{7}$$

where \mathbf{X}_r is a $n_r \times p$ matrix corresponding to Y_{-lr} (i.e. Y_r excluding the l -th column) and $\boldsymbol{\beta}_{lr} = (\beta_{1lr}, \beta_{2lr}, \dots, \beta_{plr})$ is a vector of regression coefficients specific to equation l and group r . We exploit the structure of the Horseshoe prior, retaining group-coefficient specific local shrinkage parameters λ_{jlr} , while we link the parameters τ_{lr} . The rationale behind this strategy is the following. The global shrinkage parameter τ_{lr} pulls all the coefficients globally towards zero, while the thick half-Cauchy tails for the local variances λ_{jlr}^2 allow the important coefficients to escape the global shrinkage independently in each group [12]. In practice, we expect the graphs to have group specific connection patterns, allowed by the group-coefficient specific local shrinkage parameters λ_{jlr} . We also expect groups to share some common structures and to be sparse. For this reason, we link the global shrinkage parameters τ_{lr} , allowing borrowing information across groups about the global level of sparsity of the graphs.

2.1.2. 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 graphs from multiple groups. Some recent proposals for inference of dynamic Graphical Models can be found, for instance, in [1,21,28,31,33,44].

In this work, we consider the evolution over time of the patterns of metabolic associations for two ethnic groups. Our goal is to estimate sparse multiple-graphs evolving over time. Here, we provide an extension of the nodewise regression which enables estimation

of time dependent graphs for multiple groups, enabling borrowing of information across time and groups. One of the main goals is to gain a better understanding how the differences in associations among groups evolve over time. To this end, we extend the model in (7) by imposing a time structure over the shrinkage scale parameters of the Horseshoe prior following the approach proposed by [40]. Time dependence is introduced by specifying a stochastic volatility model on the log-variance of each regression coefficient, as well as an autoregressive term in the distribution of β_{jrt} . Let $t = 1, \dots, T$ be the time index and let \mathbf{Y}_{rt} be a $n_{rt} \times M$ matrix containing the observations belonging to group r at time t . In what follows, we omit the equation subscript l for ease of notation. Define h_{jrt} as

$$h_{jrt} = \log \left(\tau_{jr}^2 \tau_0^2 / \sqrt{pn_{rt}} \right) + \phi_{jr} \left(h_{jrt-1} - \log \left(\tau_{jr}^2 \tau_0^2 / \sqrt{pn_{rt}} \right) \right) + \log \left(\lambda_{jrt}^2 \right) \quad (8)$$

where $p = M - 1$, ϕ_{jr} is an autoregressive coefficient specific to the j -th covariate and r -th group, τ_0 is a global shrinkage parameter common to all regression coefficients and shared by all groups, τ_{jr} is a group-coefficient specific shrinkage parameter and λ_{jrt}^2 is a time-group-coefficient specific local shrinkage parameter. Note that τ_0 allows sharing information across groups, while the time dependence is captured by the parameter ϕ_{jr} . Finally, the time-dependent multiple-groups nodewise regression model becomes

$$\begin{aligned} \mathbf{y}_{rt} \mid \boldsymbol{\beta}_{rt}, \sigma_{rt}^2 &\sim N_p \left(\mathbf{X}_{rt} \boldsymbol{\beta}_{rt}, \sigma_{rt}^2 \mathbf{I}_{n_{rt}} \right) \\ \sigma_{rt}^2 \mid a_\sigma, b_\sigma &\sim \text{Inverse - Gamma} (a_\sigma, b_\sigma) \\ \beta_{jrt} &= \beta_{jrt-1} + \gamma_{jrt} \exp(h_{jrt}/2) \\ \gamma_{jrt} &\stackrel{\text{ind}}{\sim} N(0, 1) \\ h_{jrt} &= \log \left(\tau_{jr}^2 \tau_0^2 / \sqrt{pn_{rt}} \right) \\ &\quad + \phi_{jr} \left(h_{jrt-1} - \log \left(\tau_{jr}^2 \tau_0^2 / \sqrt{pn_{rt}} \right) \right) \\ &\quad + \log \left(\lambda_{jrt}^2 \right) \\ (\phi_{jr} + 1)/2 \mid \phi_a, \phi_b &\sim \text{Beta}(\phi_a, \phi_b) \\ \tau_0 &\sim C^+(0, 1) \\ \tau_{jr} &\sim C^+(0, 1) \lambda_{jrt} \qquad \qquad \qquad \sim C^+(0, 1) \end{aligned} \quad (9)$$

where $\boldsymbol{\beta}_{rt} = (\beta_{1rt}, \dots, \beta_{prt})$ is a time and group specific vector of regression parameters and \mathbf{X}_{rt} is a $n_{rt} \times p$ matrix corresponding to \mathbf{Y}_{-lrt} (i.e. \mathbf{Y}_{rt} excluding the l -th column). A schematic representation of model (9) is shown in Figure 1(b).

The distribution of the logarithm of the square of a half-Cauchy random variable is a Z-distribution [40]. In particular, if $\zeta = \log(\lambda^2)$, where $\lambda \sim C^+(0, 1)$, then ζ has probability density function

$$g(\zeta) = \pi^{-1} \exp(\zeta) [1 + \exp(\zeta)]^{-1}, \quad \zeta \in \mathbb{R} \quad (10)$$

The Z-distribution can be represented as a mean-variance mixture of Gaussian distributions [4] and, thanks to the Polya-Gamma expansion proposed by [40], we can develop a multiple groups hierarchy similar to the one in model (7). To this end, we define $\zeta_{jrt} =$

$\log(\lambda_{jrt}^2)$, $\mu_0 = \log(\tau_0^2)$ and $\mu_{jr} = \log(\tau_{jr}^2 \tau_0^2)$ and we re-write the prior distributions for the parameters on a log scale as

$$\begin{aligned}
 \zeta_{jrt} \mid \xi_{\zeta_{jrt}} &\sim N\left(0, \xi_{\zeta_{jrt}}^{-1}\right) \\
 \mu_{jr} \mid \mu_0, \xi_{\mu_{jr}} &\sim N\left(\mu_0, \xi_{\mu_{jr}}^{-1}\right) \\
 \mu_0 \mid \xi_{\mu_0} &\sim N\left(0, \xi_{\mu_0}^{-1}\right) \\
 \xi_{\zeta_{jrt}} &\sim \text{Polya} - \text{Gamma}(1, 0) \\
 \xi_{\mu_{jr}} &\sim \text{Polya} - \text{Gamma}(1, 0) \\
 \xi_{\mu_0} &\sim \text{Polya} - \text{Gamma}(1, 0)
 \end{aligned} \tag{11}$$

This modelling strategy allows us to propagate the shrinkage profile of each regression coefficient over time, allowing fast structural changes or slowly adjusting processes. The theoretical properties of the dynamic Horseshoe prior are discussed by [40], who also show its good performance when compared to alternative priors. Moreover, the Polya-Gamma expansion in (11) leads to efficient computations as it allows to design a fast block-Gibbs sampler (although, for the application in this work, posterior inference is performed through Hamiltonian Monte Carlo methods as implemented in the software `Stan`).

2.2. Posterior inference

For datasets of moderate size, e.g. $M \leq 30$ and sample size $n \leq 1000$, posterior inference for the nodewise regression model with Horseshoe prior (static and dynamic) can be performed efficiently using Bayesian softwares like `Stan` [11] or `JAGS` (<http://mcmc-jags.sourceforge.net/>). The main computational bottleneck for our model is sampling from the posterior distribution of the regression coefficients, which can be expensive for large values of p with complexity equal to $O(p^3)$ in general settings [7,59]. [37] propose an approximation scheme for the horseshoe posterior, which exploits the structural sparsity of the posterior to reduce the cost per step and that can scale to hundreds of thousands of predictors. The per step computational cost of the approximate algorithm is of the order $(\kappa \vee p)N$, where κ is the number of included variables at every step.

In Section 1 of Supplemental Material, we provide sample code to implement the proposed approach in `Stan`, which implements Hamiltonian Monte Carlo (HMC, [9], chapter 5). When M is large (> 100), implementation in a low level language is advisable. We also develop a block Gibbs sampling, extending the algorithm provided by [40] to allow for multiple groups of different sample sizes across time. Details of the MCMC algorithm are provided in Section 1 of Supplemental Material. The computational efficiency of the two algorithms is assessed through a simulation study. We consider $N = 100$ observations split equally between two groups and different values for the number of nodes M and of time points T . We record the computational times (in seconds) per iteration for both the `stan` implementation and the block Gibbs sampler MCMC algorithm (implemented in `R`). The computer on which we have run the simulations is a Linux machine with Intel Core i7

1.8 GHZ. As it is shown in Table 1 of Supplemental Section 1, both algorithms are computationally efficient when $T = 3$, with the Gibbs sampler outperforming `Stan` for larger number of time points. This is due to the ability of the Gibbs sampler to perform joint updates without looping over the number of time points. Recall also that the computational burden of the dynamic Horseshoe approach scales linearly with the number of time points [40]. The results can be further improved by implementing a parallel programming strategy, thanks to the conditional independence of the different regressions.

Finally, posterior inference on the implied graph can be performed by applying either the *OR* rule or the *AND* rule. In practice, at every iteration of the MCMC algorithm, we obtain a draw for the graph, by including an edge between nodes j and l if and only if the draws $\beta_{jl} \neq 0$ or $\beta_{lj} \neq 0$, in the case of the *OR* rule. When using the *AND* rule, we include an edge between nodes l and j if and only if the draws $\beta_{jl} \neq 0$ and $\beta_{lj} \neq 0$. Moreover, given β_{jl} and σ_l^2 , we obtain a sampled value for Ω by setting the diagonal elements equal to $1/\sigma_l^2$ and off-diagonal elements equal to $-\beta_{jl}/\sigma_l^2$.

3. Simulation study

We investigate the performance of the proposed models on synthetic datasets. We evaluate the ability of each model to estimate the precision matrix Ω through the Mean Absolute Error (MAE), calculated between the true matrix and its posterior mean estimate. We also investigate the ability of the 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 for the edge inclusion probability. We also include results relative to the false positive rates (FPR) and false negative rates (FNR) when comparing different network estimates. For all the simulation scenarios presented in this Section, we show the results obtained by applying the *AND* rule. Using the *OR* rule yield comparable results.

In the first simulation, we compare the estimate of the multiple groups model in (7) with that of the R package `BDgraph`, which implements a birth-death MCMC algorithm for Bayesian structure learning in graphical models, and with the Graphical Group Lasso [15]. We construct three precision matrices Ω_1 , Ω_2 and Ω_3 , corresponding to graphs G_1 , G_2 and G_3 , of $M = 20$ nodes. Following [54], we first define the precision matrix Ω_1 and then derive the others as a perturbation of the first. We set the main diagonal elements of Ω_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. Ω_2 is derived from Ω_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. Ω_3 is derived from Ω_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` ([32], from the R package `Matrix`). The precision matrices Ω_2 and Ω_3 constructed with this procedure are a perturbation of Ω_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 the second simulation scenario we compare the multiple groups dynamic node-wise model in (9) with (i) the static multiple groups nodewise model in (7) (where we assume the three time periods to be independent); (ii) the `BDgraph` package and (iii) the Graphical Group LASSO. We consider $M = 20$ nodes, two groups and we construct two matrices Ω_{1t_1} and Ω_{2t_1} , one for each group at time 1. First we set the main diagonal elements of Ω_{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 Ω_{2t_1} is constructed removing two edges from Ω_{1t_1} at random and adding three 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 two 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 3 in Section 2 of Supplemental Material, we show the networks generated with such procedure from which we simulate the datasets. 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 third 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.

3.1. Simulation results

In Figure 1 in Supplemental Section 2 we report the results of the first simulation scenario by displaying the boxplots of the MAE and AUC, calculated over twenty replicates obtained with the multiple groups nodewise model (7) and with the `BDgraph` package. The node-wise model works better in terms of MAE, for which a value closer to 0 denotes an estimate of Ω close to the truth, and in terms of AUC, for which a value close to 1 denotes a better recovery of the true graph. The values of the FPR and FNR for the two models, shown in Figure 2 in Section 2 of Supplemental Material, are comparable in all groups, and in particular equal to zero for groups G_2 and G_3 .

In Figures 4 and 5 in Supplemental Section 2 we show the results for the second simulation scenario. We compare the results obtained using the dynamic and static multiple groups nodewise models in (7) and (9), the `BDgraph` package and the Graphical Group LASSO method. We notice a competitive recovery offered by the proposed model of the true precision matrices (Figure 4, MAE, top two panels). The graphical structure (Figure 4, AUC, bottom two panel) is recovered well by the proposed method in two of the time points for the first group, and in the second time point for the second group. In this simulation setting, no method consistently outperforms the others. This is not surprising given the sample sizes. In Figure 5 we display False Positive and False Negative Rates for the four models. As expected, given the use of a shrinkage prior, the proposed models perform better in terms of FPR and, in general, similarly to `BDgraph`. We note that Graphical Group LASSO induces less sparsity in graph reconstruction.

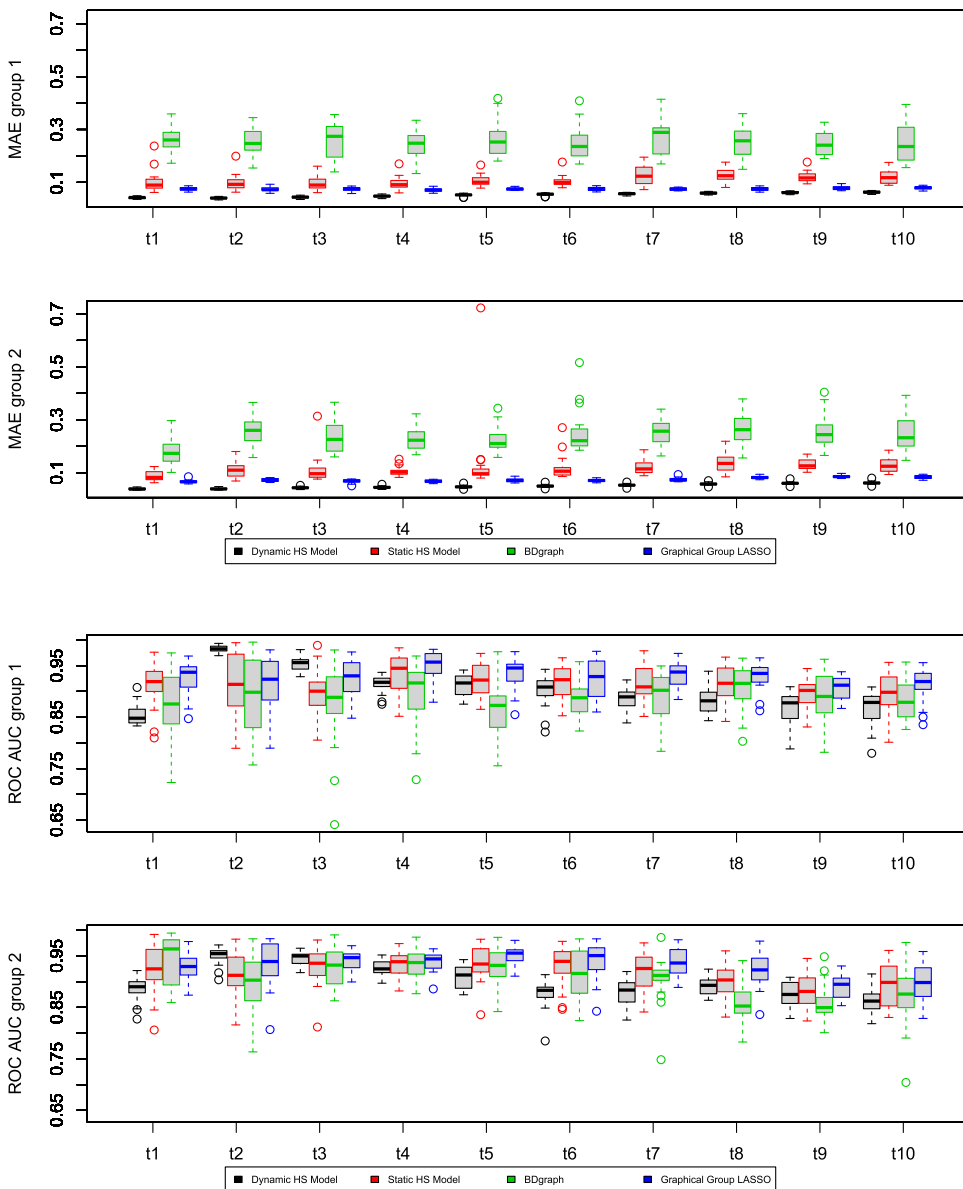


Figure 2. Third simulation scenario. MAE and AUC comparison between the results obtained using the dynamic model in (9) (black), the static model in (7) (red), *BDgraph* (green) and Graphical Group LASSO (blue). Each row refers to one of the groups and shows the boxplots for each of the ten time points. The results are obtained over twenty replicates and by adopting the *AND* rule. In terms of MAE, the dynamic model yields the lowest values, indicating overall a good performance in estimating the precision matrix. On the other hand, *BDgraph* shows the worst performance in terms of MAE. The methods are comparable in terms of AUC.

The results of the comparison between the dynamic and static nodewise models obtained from the third simulation scenario are shown in Figures 2 and 3, where we report the boxplots of the MAE, AUC, FPR and FNR calculated over twenty replicates.

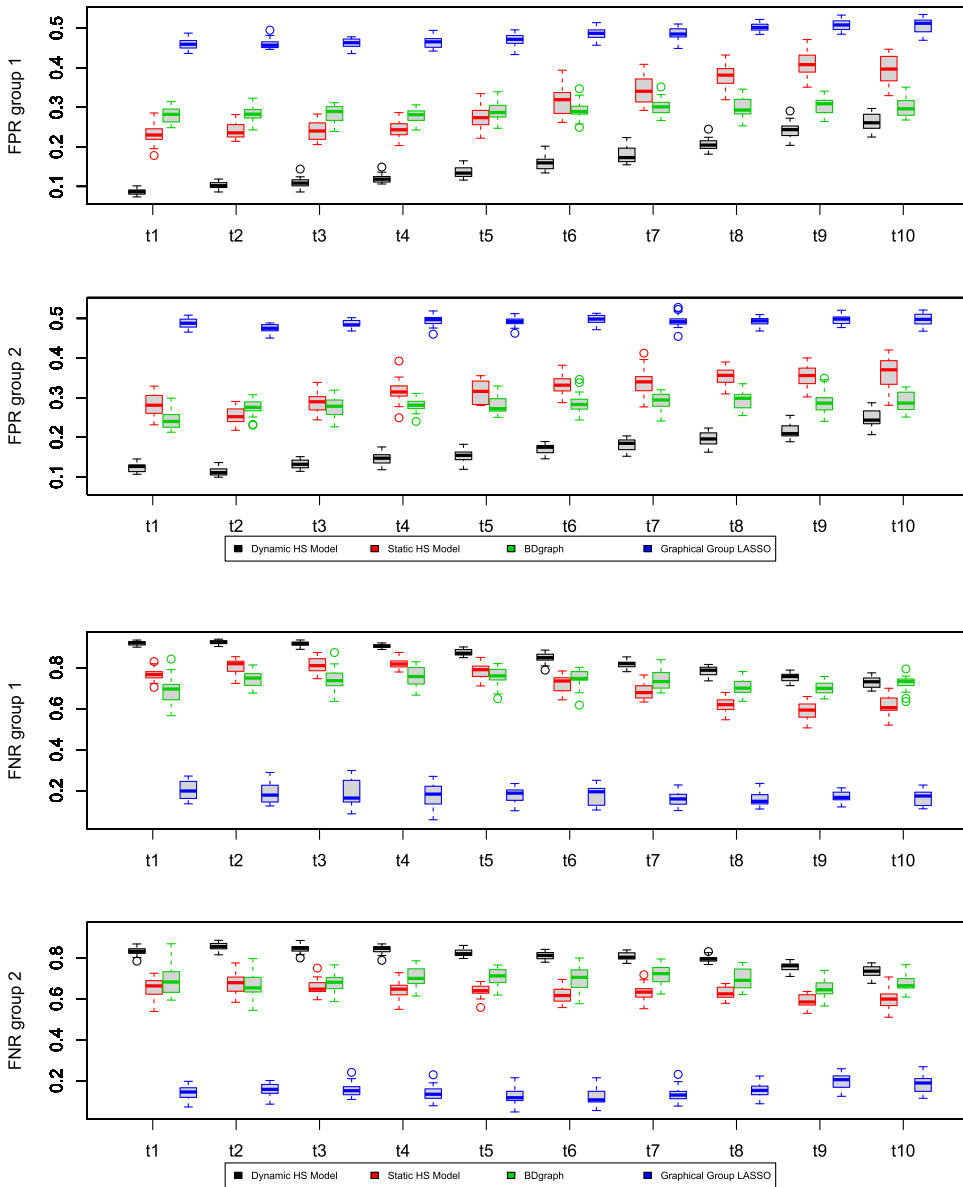


Figure 3. Third simulation scenario. False positive rates (FPR) and false negative rates (FNR) for the dynamic model in (9) (black), the static model in (7) (red), *BDgraph* (green) and Graphical Group LASSO (blue). Each row refers to one of the groups and shows the boxplots for each of the ten time points. The results are obtained over twenty replicates and by adopting the *AND* rule. In both groups, the dynamic and Graphical Group LASSO models differ the most in terms of FPR and FNR. On the other hand, the results for the static model and *BDgraph* are comparable at the first time points, and diverge slightly at later time points.

The dynamic model has the lowest MAE values in both groups and at all time points, while *BDgraph* shows the worst performance in terms of MAE. In terms of graph recovery, as measured by AUC, the methods are comparable. In terms of FPR, the dynamic Horseshoe model performs most shrinkage, although it suffers the highest FNR.

4. Application

In this Section, we apply the nodewise regression model for dynamic multiple graphs to the SABRE metabolic data. The dataset has a total of 1246 observations at baseline (T_1) and 875 at follow-up time (T_2). In this application, groups are defined by ethnicity. Individuals are stratified into 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 . Measurements of different classes of metabolites are available. Lipoproteins, which constitute the majority of the metabolomic dataset, are classified according to their density (very-low-density lipoprotein (VLDL), low-density lipoprotein (LDL), intermediate-density lipoprotein (IDL) and high-density lipoprotein (HDL)). Each lipoprotein subclass can be further characterised by its lipid composition (i.e. triglycerides, phospholipids, free cholesterol and cholesteryl esters) and its particle size. The full list of metabolites included in the analysis is reported in Table 2 in Section 2 of Supplemental Material. We include in the analysis clinical markers, such as the homeostasis model assessment (Homa IR) as an index of insulin resistance [48] - an important risk factor for the development of type 2 diabetes -, waist to hip ratio (WHR) as a measure of body fat distribution, control variables such as gender, age, smoking habits, physical activity and alcohol consumption. We control for previous disease status, in particular we include indicators of coronary heart disease (CHD), stroke and diabetes mellitus. We also include indicators of drug treatments for blood pressure, diabetes and blood lipids. The full list of covariates included is given in Table 3 in Section 2 of Supplemental Material.

The nodes of the graph correspond to the different metabolites and there are a total of $M = 88$ nodes (i.e. the number of equations in the nodewise regression). When analysing these 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 μ_l of the multivariate Normal distribution in (3) through a linear predictor and assume that $\mu_{lrt} = \mathbf{Z}\eta_{lrt}$, where \mathbf{Z} is a matrix of predictors common to all equations and η_{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 η_{lrt} as follows

$$\begin{aligned}\eta_{klrt} &= \eta_{klrt-1} + \gamma_{klrt}^z \\ \gamma_{klrt}^z &\sim \mathbf{N}(0, s_0)\end{aligned}$$

for $k = 1, \dots, p_z$, where s_0 is the prior variance, here specified to induce a flat Normal distribution. Posterior inference is performed by updating jointly the regression coefficients β_{lrt} and η_{lrt} . We also include an intercept term so that the total number of covariates is $p_z = 20$. We run the MCMC for 10,000 iterations, including a burn-in period of 2000 iterations and thinning every 4 iterations. We perform posterior analysis applying both the *AND* and *OR* rule. In Figures 6 to 9 in Supplemental Section 2, we report the posterior expectation of κ for each group and time point. Furthermore, in Figures 10 to 17 we display the inferred individual networks for each ethnic group at both baseline and follow-up times. These networks are characterised by a high number of edges, in particular, we can notice a very highly connected group of lipoproteins in all graphs. In addition to the individual networks we also estimate the differential networks [16,65,71] derived from the pairwise comparison between the graphs corresponding to the two ethnicities for each time point and the pairwise comparison between T_1 and T_2 for each ethnicity. Following [65], 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 shedding light on the main differences between ethnicities and the evolution of the metabolic associations over time. Differential networks obtained with the *AND* are shown in Figures 4 to 7, while those obtained with *OR* are reported in Figures 18 to 21. In a differential network, an edge between two nodes is added if the posterior probability of an edge being in one graph but not in the other is higher than 0.5. In this Section, we only discuss the results obtained with the *AND* rule since it implies a more stringent selection criteria.

In Figures 4 and 5 we show the differential networks between T_1 and T_2 for Europeans and South-Asians, respectively. 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 is shared by the respective ethnicity both at T_1 and T_2 . Moreover, for both ethnic groups, the majority of the edges in the differential networks derives from edges present at baseline, but not at follow-up. The connected metabolites in the differential networks for both ethnicities (Figures 4 and 5) belong predominantly to the groups of very low density lipoproteins and high density lipoproteins, with the addition of an edge between leucine and isoleucine. These are, together with valine, essential amino acids which account for 35–40% of the dietary indispensable amino acids in body protein and 14% of the total amino acids in skeletal muscle [46]. Their association with a number of disorders has often been reported in the literature, including insulin resistance, type-2 diabetes [29] and cardiovascular diseases [69].

In Figures 6 and 7 we report the differential networks between Europeans and South-Asians, respectively at time T_1 and T_2 . These networks present edges connecting amino-acids and lipoproteins sub-fractions, highlighting potential differences in underlying metabolic processes. In the last column of Table 2 in Supplemental Section 2 we report the metabolites differentially connected at both time points. 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 IMPaLA [Integrated Molecular Pathway Level Analysis][38]. 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 four top-ranked pathways at time T_2 : *Cholesterol metabolism*, *Fat digestion and absorption metabolism*, *Lipid and atherosclerosis metabolism* and *Vitamin digestion and absorption*. Note that the differential network at time T_1 contains only one extra metabolites as compared to T_2 , leading to substantially the same pathways.

Cholesterol metabolism in humans is a complex process which involves multiple metabolic pathways subject to many points of regulation under genetic and metabolic control [27]. In brief, cholesterol is either supplied from diet or synthesised by many cells of the body, with the liver being one of the major sites of cholesterol synthesis. A diet characterised by high intakes of cholesterol itself, or of saturated fats and excessive calories, may increase the level of cholesterol in the blood. The response to changes in dietary cholesterol is heterogeneous in humans and many studies show that cholesterol consumption should

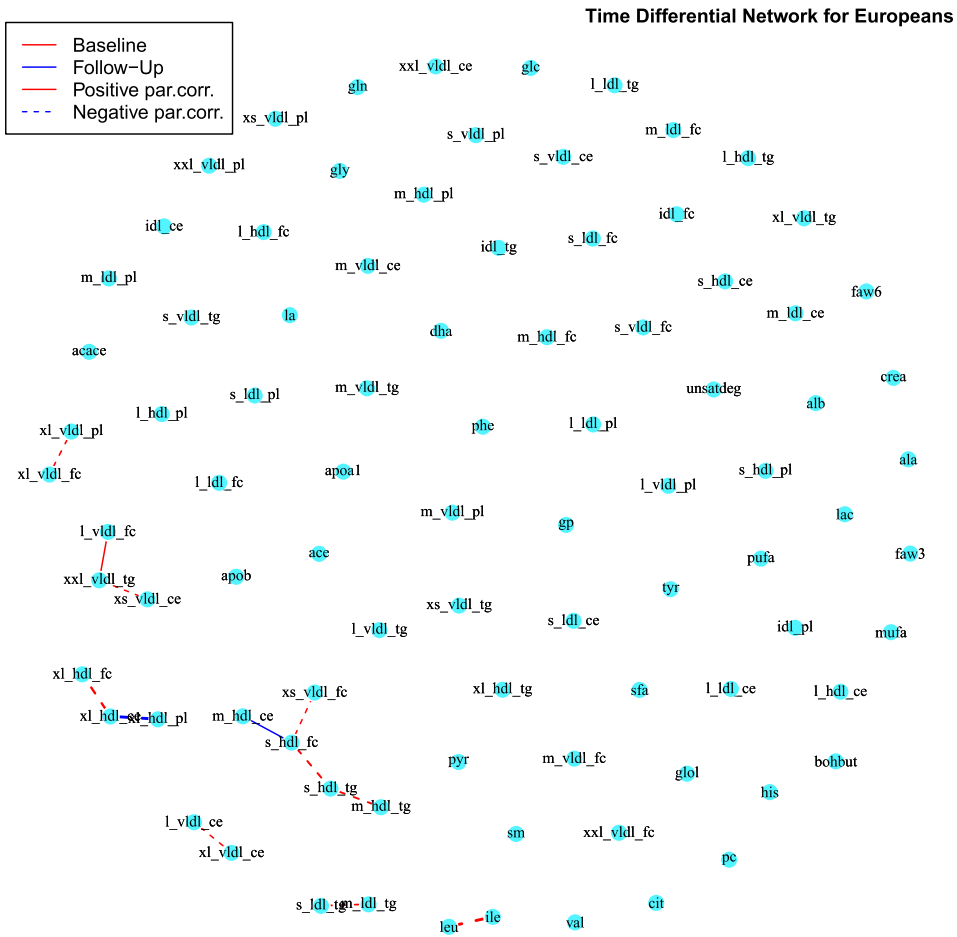


Figure 4. European Differential network between baseline and follow-up obtained with the *AND* rule. Red lines correspond to edges only present in the baseline network, while blue lines to those only present in the follow-up network. Continuous lines represent differential positive partial correlations, while dashed lines indicate negative ones. Note that the majority of the differential edges comes from those only present in the baseline network.

be restricted by diabetics and others at risk for cardiovascular disease [18]. Moreover, levels of cholesterol in blood are also influenced by environmental factors such as dietary fatty acids and metabolic perturbations such as diabetes, obesity and genetic factors [3]. Lipids and lipoproteins crucially contribute to atherosclerosis, the pathological basis of cardiovascular diseases and influence inflammatory processes as well as function of leukocytes, vascular and cardiac cells, thereby impacting on vessels and heart. Clinical studies have found a clear causal relationship between hypercholesterolaemia and atherosclerotic disease [see, for instance,][56]. Furthermore, beyond LDL cholesterol, there is evidence that also other lipid mediators contribute to cardiovascular risk [62]. Consistently with our analysis, ethnic differences in dyslipidemia patterns, lipid metabolism, total and HDL cholesterol have been reported in the literature [20,23,70].

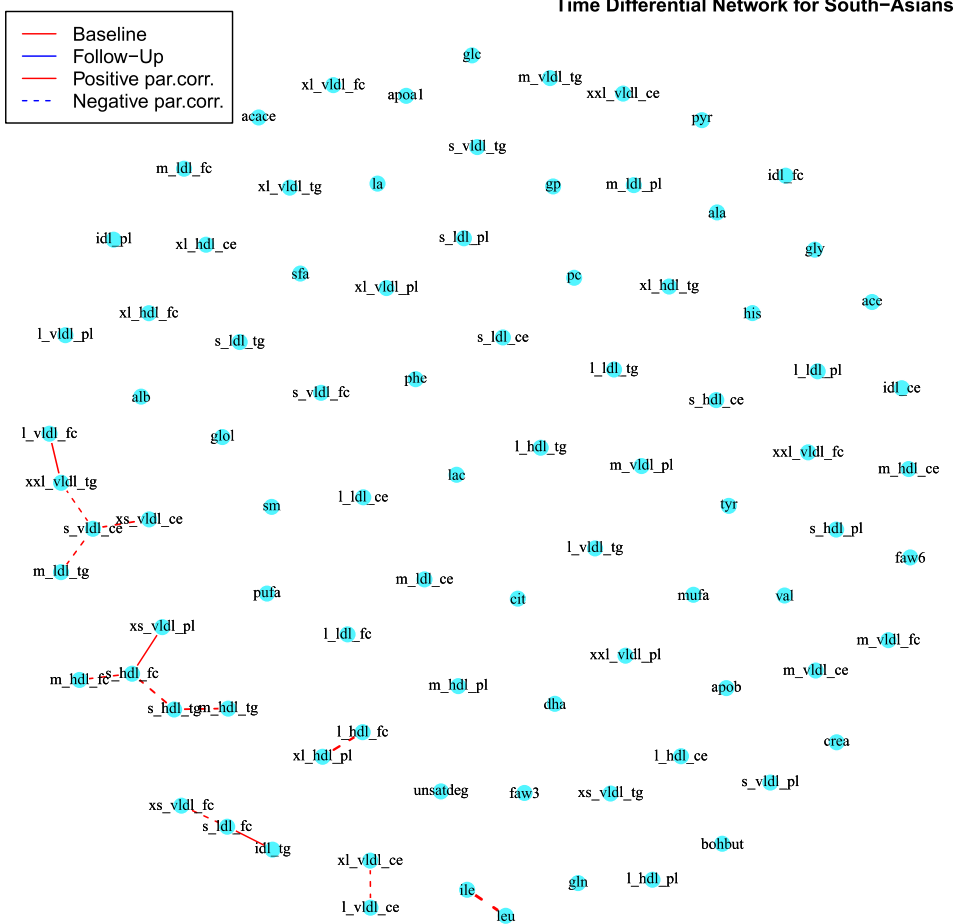


Figure 5. South-Asians Differential network between baseline and the follow-up obtained with the *AND* rule. Red lines correspond to edges only present in the baseline network, while blue lines to those only present in the follow-up network. Continuous lines represent differential positive partial correlations, while dashed lines indicate negative ones. Note that the only difference between the two networks comes from edges only present in the baseline network.

Lipids have physical, chemical, and physiological characteristics that make them important factors in human nutrition. They form a group of compounds of varied chemical nature, with the common property of being soluble in organic solvents but insoluble in water. This property affects their digestion, absorption and transport in the blood and metabolism at cellular level [57]. Dietary fats consist of a wide range of polar and nonpolar lipids [10]. Triacylglycerol is the dominant fat in the diet, contributing 90–95% of the total energy derived from dietary fat. Dietary fats also include phospholipids, sterols (e.g. cholesterol), and many other lipids (e.g. fat-soluble vitamins). Lipid digestion, absorption, and metabolism have been linked to lipid-associated disorders, including dyslipidemias and cardiovascular diseases. Efficient absorption of dietary fats is important as fat can be used as a source of energy to support various cellular functions or stored until it is needed to support intracellular processes [35]. Most circulating and stored fatty acids are derived from

Ethnicity Differential Network for the Follow-Up

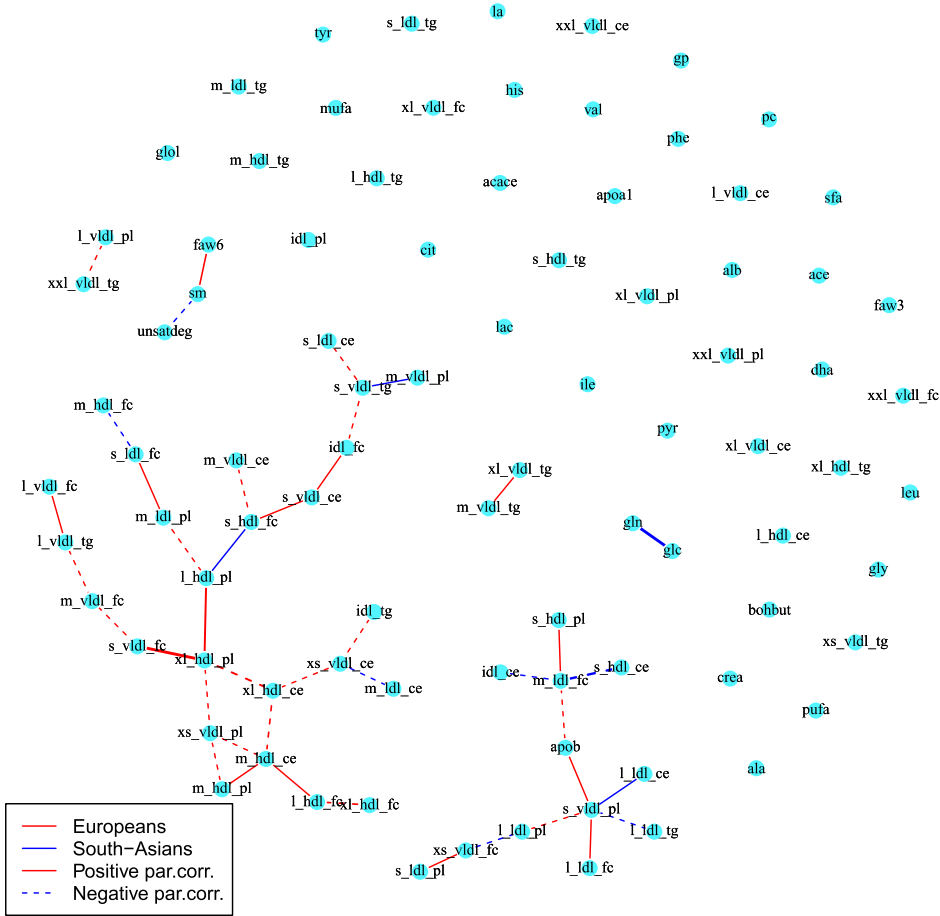


Figure 7. Differential network between Europeans and South-Asians at follow-up obtained with the AND rule. Red lines correspond to edges only present in the European network, while blue lines to those only present in the South-Asian network. Continuous lines represent differential positive partial correlations, while dashed lines indicate negative ones.

the walls of arterial blood vessels [39]. Elevated levels of low density lipoprotein (LDL) cholesterol constitute a major risk factor for genesis of atherosclerosis as LDL can accumulate within the blood vessel wall and undergo modification by oxidation. It is generally accepted that atherosclerotic lesions are initiated via an enhancement of LDL uptake by monocytes and macrophages [13,45]. Ethnic differences have been reported in the risk of atherosclerosis. For example, Lipoprotein(a) (LP(a)) is a well-known risk factor for atherosclerosis and Lp(a)-associated risk may vary by ethnicity [e.g.][22]. Lp(a) has been related to greater risk of carotid plaque and its progression in whites and Lp(a)-associated risk of plaque outcomes vary significantly in white and black individuals [64]. Moreover, there are distinct patterns of lipid profiles associated with ethnicity regardless of the glucose levels [77] and in the biological relationships underlying the serum lipids-disease association [5].

The fourth identified pathway is *Vitamin digestion and absorption*. Vitamins are organic substances which are essential for health and well-being as they catalyse numerous biochemical reactions. Vitamins must be obtained from exogenous sources via intestinal absorption since humans and other mammals cannot synthesise these compounds [39]. The importance of vitamins in health has been widely reported [47,66].

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 Figures 22, 23, 24 and 25 we plot the posterior mean of the regression coefficients η_{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, particularly 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 stronger effect (see the 95% credible region) on the metabolites. High blood triglycerides and low HDL are among the risk factors that determine the metabolic syndrome [58], 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.

5. Discussion and conclusion

This paper extends nodewise regression technique to infer dynamic evolving multiple graphs. The model allows us to analyse multiple groups of different sample sizes observed at multiple time points, borrowing information across time and groups. We impose regularisation on the regression coefficients and the model allows for 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 algorithm 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, hence 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 as well as a frequentist method. 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. In most scenarios, the proposed methods outperforms the competitors, still maintaining a computational advantage. 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 follow-up. Our approach enables us to detect an interpretable set of unique association patterns which can aid mechanistic understanding of between-group

and between-times differences in the development of insulin resistance, diabetes and cardiovascular diseases and has the potential to help generating new scientific hypotheses. In doing this, we correct for potential confounders and clinical events that could affect metabolites levels.

Acknowledgments

We are 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.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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 [grant numbers 067100, 37055891 and 086676/7/08/Z], the British Heart Foundation [grant numbers PG/06/145, PG/08/103/26133, PG/12/29/29497 and CS/13/1/30327] and Diabetes UK [grant number 13/0004774]. The SABRE study team also acknowledges the support of the National Institute of Health Research Clinical Research Network (NIHR CRN). This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 2 [grant number MOE2019-T2-2-100].

ORCID

Maria De Iorio  <http://orcid.org/0000-0003-3109-0478>

References

- [1] A. Abbruzzo, C. Di Serio, and E. Wit, *Dynamic Gaussian graphical models for modelling genomic networks*, in *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, Springer, 2013, pp. 3–12.
- [2] A. Armagan, D.B. Dunson, and J. Lee, *Generalized double pareto shrinkage*, *Stat. Sin.* 23 (2013), pp. 119.
- [3] D. Arnold and P. Kwiterovich, *Cholesterol-absorption, function, and metabolism*, in *Encyclopedia of Food Sciences and Nutrition*, B. Caballero, ed., 2nd ed., Academic Press, Oxford, 2003, pp. 1226–1237. Available at <https://www.sciencedirect.com/science/article/pii/B012227055X00225X>.
- [4] O. Barndorff-Nielsen, J. Kent, and M. Sørensen, *Normal variance-mean mixtures and z distributions*, *International Statistical Review/Revue Internationale De Statistique.* 50 (1982), pp. 145–159.
- [5] A.R. Bentley and C.N. Rotimi, *Interethnic differences in serum lipids and implications for cardiometabolic disease risk in African ancestry populations*, *Glob. Heart.* 12 (2017), pp. 141–150.
- [6] A. Bhadra, J. Datta, N.G. Polson, and B. Willard, *The horseshoe+ estimator of ultra-sparse signals*, *Bayesian Analysis.* 12 (2017), pp. 1105–1131.
- [7] A. Bhattacharya, A. Chakraborty, and B.K. Mallick, *Fast sampling with Gaussian scale mixture priors in high-dimensional regression*, *Biometrika.* 103 (2016), pp. 985–991.

- [8] A.E. Bilgrau, C.F. Peeters, P.S. Eriksen, M. Bøgsted, and W.N. van Wieringen, *Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes.*, *J. Mach. Learn. Res.* 21 (2020), pp. 1–52.
- [9] S. Brooks, A. Gelman, G. Jones, and X.L. Meng, *Handbook of Markov Chain Monte Carlo*, CRC press, 2011.
- [10] M.C. Carey, *Small dm, bliss cm. Lipid digestion and absorption*, *Annu. Rev. Physiol.* 45 (1983), pp. 651–677.
- [11] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, *Stan: A probabilistic programming language*, *J. Stat. Softw.* 76 (2017), pp. 1–32. Available at <https://www.jstatsoft.org/v076/i01>.
- [12] C.M. Carvalho, N.G. Polson, and J.G. Scott, *The horseshoe estimator for sparse signals*, *Biometrika* 97 (2010), pp. 465–480.
- [13] P.C. Choy, Y.L. Siow, D. Mymin, and K. O, *Lipids and atherosclerosis*, *Biochem. Cell Biol.* 82 (2004), pp. 212–224.
- [14] H. Chun, X. Zhang, and H. Zhao, *Gene regulation network inference with joint sparse Gaussian graphical models*, *J. Comput. Graph. Stat.* 24 (2015), pp. 954–974.
- [15] P. Danaher, P. Wang, and D.M. Witten, *The joint graphical lasso for inverse covariance estimation across multiple classes*, *J. R. Stat. Soc. Series. B. Stat. Methodol.* 76 (2014), pp. 373–397.
- [16] A. De la Fuente, *From ‘differential expression’ to ‘differential networking’ – identification of dysfunctional regulatory networks in diseases*, *Trends. Genet.* 26 (2010), pp. 326–333.
- [17] A.P. Dempster, *Covariance selection*, *Biometrics* 28 (1972), pp. 157–175.
- [18] R.G. Elkin, *Cholesterol in chicken eggs: Still a dietary concern for some*, in *Egg Innovations and Strategies for Improvements*, Elsevier, 2017, pp. 189–198.
- [19] S. Ellul, M. Wake, S.A. Clifford, K. Lange, P. Würtz, M. Juonala, T. Dwyer, J.B. Carlin, D.P. Burgner, and R. Saffery, *Metabolomics: Population epidemiology and concordance in Australian children aged 11–12 years and their parents*, *BMJ. Open.* 9 (2019), pp. 106–117.
- [20] B. Enkhmaa, E. Anuurad, and L. Berglund, *Lipoprotein (a): Impact by ethnicity and environmental and medical conditions*, *J. Lipid. Res.* 57 (2016), pp. 1111–1125.
- [21] S. Epskamp, L.J. Waldorp, R. Möttus, and D. Borsboom, *The Gaussian graphical model in cross-sectional and time-series data*, *Multivariate. Behav. Res.* 53 (2018), pp. 453–480.
- [22] N.I. Forbang, M.H. Criqui, M.A. Allison, J.H. Ix, B.T. Steffen, M. Cushman, and M.Y. Tsai, *Sex and ethnic differences in the associations between lipoprotein (a) and peripheral arterial disease in the multi-ethnic study of atherosclerosis*, *J. Vasc. Surg.* 63 (2016), pp. 453–458.
- [23] A.T. Frank, B. Zhao, P.O. Jose, K.M. Azar, S.P. Fortmann, and L.P. Palaniappan, *Racial/ethnic differences in dyslipidemia patterns*, *Circulation* 129 (2014), pp. 570–579.
- [24] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, *Biostatistics* 9 (2008), pp. 432–441.
- [25] A. Gelman, *Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)*, *Bayesian Analysis* 1 (2006), pp. 515–534.
- [26] E. George and R. McCulloch, *Variable selection via Gibbs sampling*, *J. Am. Stat. Assoc.* 88 (1993), pp. 881–889.
- [27] J.L. Goldstein and M.S. Brown, *A century of cholesterol and coronaries: From plaques to genes to statins*, *Cell* 161 (2015), pp. 161–172.
- [28] K. Greenewald, S. Park, S. Zhou, and A. Giessing, *Time-dependent spatially varying graphical models, with application to brain fMRI data analysis*, *Adv. Neural Inform. Process. Syst.* 30 (2017), pp. 5832–5840.
- [29] M. Guasch-Ferré, A. Hruby, E. Toledo, C.B. Clish, M.A. Martínez-González, J. Salas-Salvadó, and F.B. Hu, *Metabolomics in prediabetes and diabetes: A systematic review and meta-analysis*, *Diabetes. Care.* 39 (2016), pp. 833–846.
- [30] J. Guo, E. Levina, G. Michailidis, and J. Zhu, *Joint estimation of multiple graphical models*, *Biometrika* 98 (2011), pp. 1–15.
- [31] P.L. Harrington Jr and A.O. Hero III, *Spatio-temporal graphical model selection*, preprint (2010). Available at arXiv:1004.2304.

- [32] N.J. Higham, *Computing the nearest correlation matrix—a problem from finance*, IMA J. Numer. Anal. 22 (2002), pp. 329–343.
- [33] F. Huang and S. Chen, *Learning dynamic conditional Gaussian graphical models*, IEEE. Trans. Knowl. Data. Eng. 30 (2017), pp. 703–716.
- [34] D.Y. Hui, *Intestinal phospholipid and lysophospholipid metabolism in cardiometabolic disease*, Curr. Opin. Lipidol. 27 (2016), pp. 507–512.
- [35] J. Iqbal and M.M. Hussain, *Intestinal lipid absorption*, Am. J. Physiol.-Endocrinol. Metab. 296 (2009), pp. E1183–E1194.
- [36] J. Janková and S. van de Geer, *Inference in high-dimensional graphical models*, preprint (2018). Available at arXiv:1801.08512.
- [37] J. Johndrow, P. Orenstein, and A. Bhattacharya, *Scalable approximate MCMC algorithms for the horseshoe prior*, J. Mach. Learn. Res. 21 (2020), pp. 00.
- [38] A. Kamburov, R. Cavill, T.M. Ebbels, R. Herwig, and H.C. Keun, *Integrated pathway-level analysis of transcriptomics and metabolomics data with impala*, Bioinformatics 27 (2011), pp. 2917–2918.
- [39] M. Kanehisa and S. Goto, *Kegg: Kyoto encyclopedia of genes and genomes*, Nucleic. Acids. Res. 28 (2000), pp. 27–30.
- [40] D.R. Kowal, D.S. Matteson, and D. Ruppert, *Dynamic shrinkage processes*, preprint (2017). Available at arXiv:1707.00763.
- [41] S. Lauritzen, *Graphical Models*, Oxford Statistical Science Series, Clarendon Press, 1996. Available at <https://books.google.co.uk/books?id=mGQWkx4guhAC>.
- [42] A. Lenkoski and A. Dobra, *Computational aspects related to inference in Gaussian graphical models with the g -wishart prior*, J. Comput. Graph. Stat. 20 (2011), pp. 140–157.
- [43] F. Leone, L. Nelson, and R. Nottingham, *The folded normal distribution*, Technometrics 3 (1961), pp. 543–550.
- [44] Z. Lin, T. Wang, C. Yang, and H. Zhao, *On joint estimation of Gaussian graphical models for spatial and temporal data*, Biometrics 73 (2017), pp. 769–779.
- [45] M.F. Linton, P.G. Yancey, S.S. Davies, W.G. Jerome, E.F. Linton, W.L. Song, A.C. Doran, and K.C. Vickers, *The role of lipids and lipoproteins in atherosclerosis*, (2019). Endotext [Internet].
- [46] I. Manoli and C. Venditti, *Disorders of branched chain amino acid metabolism*, Transl. Sci. Rare Dis. 1 (2016), pp. 91–110.
- [47] M.A. Maqbool, M. Aslam, W. Akbar, and Z. Iqbal, *Biological importance of vitamins for human health: A review*, J. Agric. Basic Sci. 2 (2017), pp. 50–58.
- [48] D.R. Matthews, J.P. Hosker, A.S. Rudenski, B.A. Naylor, D.F. Treacher, and R.C. Turner, *Homeostasis model assessment: Insulin resistance and β -cell function from fasting plasma glucose and insulin concentration in man*, Diabetologia 28 (1985), pp. 412–419.
- [49] R.W. McGarrah, S.B. Crown, G.F. Zhang, S.H. Shah, and C.B. Newgard, *Cardiovascular metabolomics*, Circ. Res. 122 (2018), pp. 1238–1258.
- [50] N. Meinshausen and P. Bühlmann, *High-dimensional graphs and variable selection with the lasso*, Ann. Stat. 34 (2006), pp. 1436–1462.
- [51] A. Mohammadi and E.C. Wit, *Bayesian structure learning in sparse Gaussian graphical models*, Bayesian Analysis 10 (2015), pp. 109–138.
- [52] K. Mohan, P. London, M. Fazel, D. Witten, and S.I. Lee, *Node-based learning of multiple Gaussian graphical models*, J. Mach. Learn. Res. 15 (2014), pp. 445–488.
- [53] R.B. O’Hara and M.J. Sillanpää, *A review of bayesian variable selection methods: What, how and which*, Bayesian Anal. 4 (2009), pp. 85–117.
- [54] C. Peterson, F.C. Stingo, and M. Vannucci, *Bayesian inference of multiple Gaussian graphical models*, J. Am. Stat. Assoc. 110 (2015), pp. 159–174.
- [55] J. Piironen and A. Vehtari, *Sparsity information and regularization in the horseshoe and other shrinkage priors*, Electron. J. Stat. 11 (2017), pp. 5018–5051.
- [56] D.J. Rader, J. Cohen, and H.H. Hobbs, *Monogenic hypercholesterolemia: New insights in pathogenesis and treatment*, J. Clin. Invest. 111 (2003), pp. 1795–1803.
- [57] M. Ramirez, L. Amate, and A. Gil, *Absorption and distribution of dietary fatty acids from different sources*, Early. Hum. Dev. 65 (2001), pp. S95–S101.

- [58] C.K. Roberts, A.L. Hevener, and R.J. Barnard, *Metabolic syndrome and insulin resistance: Underlying causes and modification by exercise training*, *Compr. Physiol.* 3 (2013), pp. 1–58.
- [59] H. Rue, *Fast sampling of Gaussian Markov random fields*, *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* 63 (2001), pp. 325–338.
- [60] T. Saegusa and A. Shojaie, *Joint estimation of precision matrices in heterogeneous populations*, *Electron. J. Stat.* 10 (2016), pp. 1341.
- [61] P. Soininen, A.J. Kangas, P. Würtz, T. Suna, and M. Ala-Korpela, *Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics*, *Circ.: Cardio-vasc. Genet.* 8 (2015), pp. 192–206.
- [62] J. Soppert, M. Lehrke, N. Marx, J. Jankowski, and H. Noels, *Lipoproteins and lipids in cardiovascular disease: From mechanistic insights to therapeutic targeting*, *Adv. Drug. Deliv. Rev.* 159 (2020), pp. 4–33.
- [63] K. Sproston and J. Mindell, *Health survey for England 2004. The health of minority ethnic groups*, 2006.
- [64] B.T. Steffen, G. Thanassoulis, D. Duprez, J.H. Stein, A.B. Karger, M.C. Tattersall, J.D. Kaufman, W. Guan, and M.Y. Tsai, *Race-based differences in lipoprotein (a)-associated risk of carotid atherosclerosis: The multi-ethnic study of atherosclerosis*, *Arterioscler. Thromb. Vasc. Biol.* 39 (2019), pp. 523–529.
- [65] L.S. Tan, A. Jasra, M. De Iorio, and T.M. Ebbels, *Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks*, *Ann. Appl. Stat.* 11 (2017), pp. 2222–2251.
- [66] T.G. Taylor, *The Importance of Vitamins to Human Health*, Springer, 1979.
- [67] D. Telesca, P. Müller, G. Parmigiani, and R.S. Freedman, *Modeling dependent gene expression*, *Ann. Stat.* 6 (2012), pp. 542.
- [68] T. Tillin, N.G. Forouhi, P.M. McKeigue, and N. Chaturvedi, *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*, *Int. J. Epidemiol.* 41 (2012), pp. 33–42.
- [69] D.K. Tobias, P.R. Lawler, P.H. Harada, O.V. Demler, P.M. Ridker, J.E. Manson, S. Cheng, and S. Mora, *Circulating branched-chain amino acids and incident cardiovascular disease in a prospective cohort of us women*, *Circ.: Genom. Precis. Med.* 11 (2018), pp. e002157.
- [70] J.K. Ujcic-Voortman, G. Bos, C.A. Baan, D.G. Uitenbroek, A.P. Verhoeff, and J.C. Seidell, *Ethnic differences in total and HDL cholesterol among Turkish, Moroccan and Dutch ethnic groups living in Amsterdam, the Netherlands*, *BMC. Public. Health.* 10 (2010), pp. 1–8.
- [71] B. Valcárcel, P. Würtz, N.K.S. al Basatena, T. Tukiainen, A.J. Kangas, P. Soininen, M.R. Järvelin, M. Ala-Korpela, T.M. Ebbels, M. de Iorio, and P. Csermely, *A differential network approach to exploring differences between biological states: An application to prediabetes*, *PLoS. ONE.* 6 (2011), pp. e24702.
- [72] F. Visioli, M.A. Crawford, S. Cunnane, P. Rise, and C. Galli, *Lipid transport, dietary fats, and endogenous lipid synthesis: Hypotheses on saturation and competition processes*, *Nutr. Health.* 18 (2006), pp. 127–132.
- [73] F. Visioli and A. Poli, *Fatty acids and cardiovascular risk. Evidence, lack of evidence, and diligence*, *Nutrients* 12 (2020), pp. 3782.
- [74] H. Wang, *Bayesian graphical lasso models and efficient posterior computation*, *Bayesian Analysis.* 7 (2012), pp. 867–886.
- [75] R. Warnick, M. Guindani, E. Erhardt, E. Allen, V. Calhoun, and M. Vannucci, *A Bayesian approach for estimating dynamic functional network connectivity in fMRI data*, *J. Am. Stat. Assoc.* 113 (2018), pp. 134–151.
- [76] N. Wermuth, *Analogies between multiplicative models in contingency tables and covariance selection*, *Biometrics* 32 (1976), pp. 95–108.
- [77] L. Zhang, Q. Qiao, J. Tuomilehto, E.D. Janus, T.H. Lam, A. Ramachandran, V. Mohan, C.D.A. Stehouwer, Y. Dong, T. Nakagami, A. Onat, and S. Söderberg, *Distinct ethnic differences in lipid profiles across glucose categories*, *J. Clin. Endocrinol. Metab.* 95 (2010), pp. 1793–1801.