Received: 3 July 2019 | Accepted: 23 March 2022

DOI: 10.1111/rssc.12570

ORIGINAL ARTICLE

Bayesian nonparametric modelling of multiple graphs with an application to ethnic metabolic differences

Marco Molinari[1] | **Andrea Cremaschi**[2] | **Maria De Iorio**[1,2,3,4] | **Nishi Chaturvedi**[5] | **Alun D. Hughes**[5] | **Therese Tillin**[5]

[1]Department of Statistical Science, UCL, London, UK

[2]Singapore Institute of Clinical Sciences, Agency for Science, Technology and Research, Singapore, Singapore

[3]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[4]Yale-NUS College, Singapore, Singapore

[5]Department of Population Science & Experimental Medicine, Institute of Cardiovascular Science, UCL, London, UK

Correspondence
Marco Molinari, Department of Statistical Science, UCL, London WC1E 6BT, UK.
Email: molinari_marco@runbox.com

Abstract

We propose a novel approach to the estimation of multiple Gaussian graphical models (GGMs) to analyse patterns of association among a set of metabolites, under different conditions. Our motivating application is the SABRE (Southall And Brent REvisited) study, a triethnic cohort study conducted in the United Kingdom. Through joint modelling of pattern of association corresponding to different ethnic groups, we are able to identify potential ethnic differences in metabolite levels and associations, with the aim of gaining a better understanding of different risk of cardiometabolic disorders across ethnicities. 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. Inference is performed via Markov chain Monte Carlo.

KEYWORDS

biomarkers, Dirichlet process, Gaussian graphical models, MCMC, metabolomics

1 INTRODUCTION

Diabetes 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 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 United Kingdom with prevalence of diabetes in South Asians estimated to be 2–4 times higher than that of the general population (Sproston & Mindell, 2006). People of African-Caribbean origin in United Kingdom, although fewer, 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 & 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 the underlying mechanistic pathways are poorly understood, although are likely to involve a complex interplay between environmental, behavioural, metabolic, genetic and epigenetic influences.

The SABRE (Southall And Brent REvisited) population-based cohort was initiated in the late 1980s in north-west London with the aim of studying ethnic differences in cardiovascular disease and diabetes. Since then, it has accumulated a wide range of phenotypic and disease outcome data. The study includes people of European, South Asian and African-Caribbean descent, aged 40–69 years at baseline. Recently, metabolic analyses have been performed on over 3000 stored blood samples from the baseline and 20-year follow-up studies. Metabolomics is the large-scale study of metabolites, within cells, biofluids, tissues or organisms. Collectively, these metabolites and their interactions within a biological system are known as the *metabolome*. Measurements of over 200 metabolites or ratios of metabolites, obtained through nuclear magnetic resonance spectroscopy (Soininen et al., 2015) are available for more than 3000 stored baseline serum samples. Lipoproteins are classified according to their density (very-low-density, low-density, intermediate-density and high-density lipoproteins). Each lipoprotein subclass can be further characterised by its lipid composition (i.e. triglycerides, phospholipids, cholesterol esters and free cholesterol) and its particle size. The full list of metabolites included in the analysis is reported in Table S1 in Supplementary Material. We exclude from the analysis individuals with known diabetes at the time of the first visit. This is motivated by the fact that people with known diabetes were already receiving treatment that may alter their metabolite levels. Furthermore, we include in the analysis: (a) clinical markers, such as the homeostasis model assessment as an index of insulin resistance (Matthews et al., 1985)—an important risk factor for the development of diabetes; (b) three important enzymes (alanine aminotransferase, aspartate aminotransferase and gamma-glutamyl transferase) of which the first two are clinical biomarkers indicators of liver health while latter is used as a diagnostic marker for liver disease; and (c) anthropometric variables measuring the body fat distribution such as the waist-to-hip ratio (WHR). The full list of covariates included is given in Table S2 in Supplementary Material.

In this work, we focus on the SABRE study fasting baseline metabolic and phenotypic data set, with a view to identifying and elucidating potential mechanistic pathways to insulin resistance (and hence risk of developing type 2 diabetes), and to explore ethnic differences in these pathways. The statistical analysis poses several challenges: intersubject variability, the large number of variables under investigation and the high correlation between metabolite levels. There is a wealth of proposals in the statistical literature on how to tackle these problems. We employ the seemingly unrelated regression (SUR) model introduced by Zellner (1971). The SUR model can be seen as a generalisation of the linear regression model, where multiple regression equations, each one having its own dependent variable and possibly a specific set of regression covariates, are linked together by specifying a joint distribution on the error terms, which can exhibit a correlation structure. The SUR model offers a flexible modelling tool, but at the price of a large number of parameters to be estimated. To regularise posterior inference we adopt a sparse SUR approach, assuming a local–global shrinkage prior for the regression coefficients, that is the horseshoe prior (Carvalho et al., 2010), and we model the associations 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 constraints 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 its underlying structure. However, the convergence of the posterior distribution can be slow due to the single edge update and the intractable normalising 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. The problem of estimating sparse matrices of regression coefficients and precision matrices jointly has been tackled before in the frequentist (Cai et al., 2013; Rothman et al., 2010), as well as Bayesian (Bhadra & Mallick, 2013; Deshpande et al., 2019) framework. The former modelling approach is usually based on ℓ_1 penalisation, while the latter exploits the specification of spike-and-slab prior distributions. In both frameworks, the estimation of the precision matrix is the most demanding part, as the parameter space is restricted to the cone of positive-definite matrices, often requiring computationally intensive algorithms. Here we use the stochastic search structure learning (SSSL) algorithm of Wang (2015) to specify the precision matrix prior distribution. The SSSL uses the best aspects of the G-Wishart and Bayesian graphical Lasso priors, enabling explicit structure learning while maintaining good scalability.

We specify a generalised Dirichlet process prior (GDP, Hjort, 2000), an extension of the well-known Dirichlet process (DP, Antoniak, 1974; Ferguson, 1973, 1974), on the edge inclusion probabilities, allowing for clustering of the edges and through the calibration of the GDP base measure we ensure the desired degree of sparsity in the graph. The GDP is a probability model defined on the space of probability distributions. Similarly to the DP (Sethuraman, 1994), the GDP can be defined through a constructive definition of the process. If a random probability measure P is distributed according to a GDP, with concentration parameter α, mean parameter μ and base measure P_0, then

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

where $\theta_1, \theta_2, \ldots$ are *iid* realisations from P_0 and δ_{θ_k} is the Dirac measure that assigns unit mass probability in correspondence of the location θ_k. The weights ψ_k are generated according to the *stick-breaking* construction:

$$\begin{aligned} \psi_k &= \phi_k \prod_{j=1}^{k-1} (1 - \phi_j), \qquad k = 2, 3, \ldots \\ \psi_1 &= \phi_1 \end{aligned} \tag{1}$$

with $\varphi_k \overset{iid}{\sim} \text{Beta}(\alpha_k \mu_k, \alpha_k(1-\mu_k))$, where μ_k represents the expected value of the beta random variable and α_k is a concentration parameter, for $k = 1, 2, \ldots$ Here we use the more parsimonious parametrisation given by Hjort (2000), where $\alpha_k = \alpha$ and $\mu_k = \mu$, for $k = 1, 2, \ldots$ By construction we have $0 \leq \psi_k \leq 1$ and $\sum_{k=1}^{\infty} \psi_k = 1$. We extend the GDP prior to multiple GGMs, enabling borrowing information between graphs under different biological conditions. More in detail, we assume a SUR model for the metabolites for each ethnic group. We model the error precision matrix in each group conditionally on an ethnic-specific graph, and we propose a joint nonparametric prior for the graphs. This strategy allows us to highlight common patterns and structural differences. In this context, each graph is characterised by the same set of nodes (representing the dependent variables of the SUR model), connected by a set of group-specific edges. Thanks to the clustering property of the GDP prior, we allow edges from different graphs to share the same edge probability and consequently to inform each other.

The paper is organised as follows. Section 2 introduces the sparse SUR model, the nonparametric prior on the edge inclusion probabilities and its extension to handle multiple GGMs. Section 3 illustrates the performance of the model on simulated data sets. Section 4 presents the analysis of the SABRE data, with a discussions of the clinical relevance of the findings. Finally, Section 5 concludes the work with a discussion of the main results and future directions.

2 | 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.

2.1 | Sparse SUR model

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

$$\boldsymbol{y}_l = X_l \boldsymbol{\beta}_l + \boldsymbol{u}_l, \quad l = 1, \ldots, M \tag{2}$$

where the X_l is a $n \times p_l$ response-specific matrix of explanatory variables, $\boldsymbol{\beta}_l = \left(\beta_{l1}, \ldots, \beta_{lp_l}\right)$ is a p_l-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, $\text{N}(\boldsymbol{0}, I_n)$, where I_n is the identity matrix of dimension $n \times n$.

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

$$\begin{aligned} \boldsymbol{y}_{1:M} &= X\boldsymbol{\beta} + \boldsymbol{u} \\ \boldsymbol{u} &\sim \text{N}(\boldsymbol{0}, \Omega \otimes I_n) \end{aligned}$$

by concatenating the responses in a unique column vector $\boldsymbol{y}_{1:M}$ 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 a 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 $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_{lj}$, which corresponds to the regression coefficient associated to the j-th covariate in the l-th equation.

2.2 | Background on graphical models

We give a brief introduction to graphical models, following Lauritzen (1996). Let $G = (V, E)$, with $V = \{1, 2, \ldots, M\}$ the vertex set and $E \subset \{(i, j) \in V \times V : i < j\}$ the edge set, be an undirected graph whose vertices are associated with a M-dimensional vector of variables $\boldsymbol{y} = (y_1, \ldots, y_M)$ following a multivariate normal distribution, $\mathrm{N}(\mathbf{0}, \Omega)$. Note that in this section, for simplicity of notation, we use $\boldsymbol{y}$ to denote the vector of variable corresponding to each node in the graph. The graph G can be represented by a set of $r = M(M-1)/2$ binary variables $Z = (z_{ij})_{i<j}$, where $z_{ij} = 1 \iff e_{ij} \in E$, with e_{ij} denoting the edge between node i and j in the graph G, for $i, j \in \{1, \ldots, M\}$. Thus, $r = M(M-1)/2$ is the total number of possible edges in the graph G. There is a direct correspondence between the elements of the precision matrix Ω and the edges in the graph G. A missing edge in E implies $\omega_{ij} = 0$ (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 Ω is constrained to belong to the cone PD_G, that is the set of positive definite matrices with entries equal to zero for all $e_{ij} \notin E$.

2.3 | Prior specification

We adopt the horseshoe prior of Carvalho et al. (2010) to impose regularisation on the regression coefficients $\boldsymbol{\beta}$. The horseshoe prior has the desirable property of being characterised by a singularity at zero to strongly shrink small or negligible coefficients, while leaving important coefficients unaffected thanks to its heavy tails. The horseshoe prior is specified as follows

$$\begin{aligned} \beta_{lj} | \lambda_{lj}, \tau_l &\sim \mathrm{N}\left(0, \lambda_{lj}^2 \tau_l^2\right) \\ \lambda_{lj} &\sim \mathrm{C}^+(0, 1) \\ \tau_l &\sim \mathrm{C}^+(0, 1) \end{aligned} \tag{3}$$

with $j = 1, \ldots, p_l$ and $l = 1, \ldots, M$. C^+ denotes the standard half-Cauchy distribution, λ_{lj}^2 is the local shrinkage parameter, specific for the coefficient β_{lj}, while τ_l^2 represents the overall shrinkage level for equation l. 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. Carvalho et al. (2010) compares the performance of the variable selection based on (3) with that of a spike and slab prior (George & McCulloch, 1993), showing that the posterior selection given by the horseshoe is consistent with that of the spike and slab. See, for instance, Piironen and Vehtari (2017) for a discussion on how to set the horseshoe prior hyperparameters. To perform posterior inference, we adopt the conjugate sampler proposed by Makalic and Schmidt (2016), which allows a fast Gibbs sampling, avoiding working directly with the half-Cauchy distribution. Makalic and Schmidt (2016) exploit the following relationship. Let κ and ρ be random variables such that

$$\kappa^2|\rho \sim \mathrm{IG}(1/2, 1/a) \text{ and } \rho \sim \mathrm{IG}(1/2, 1/A^2) \tag{4}$$

then $\kappa \sim \mathrm{C}^+(0,\ A)$, where IG() is the inverse-gamma distribution. Exploiting the scale mixture representation in Equation (4) we can express (3) as

$$\begin{aligned} \beta_{lj}|\lambda_{lj}, \tau_l &\sim \mathrm{N}\left(0, \lambda_{lj}^2\tau_l^2\right) \\ \lambda_{lj}^2|v_{lj} &\sim \mathrm{IG}\left(1/2, 1/v_{lj}\right) \\ \tau_l^2|\xi_l &\sim \mathrm{IG}\left(1/2, 1/\xi_l\right) \\ v_{lj}, \xi_l &\sim \mathrm{IG}(1/2, 1). \end{aligned} \tag{5}$$

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

$$p(\Omega) = C(\theta)^{-1}\prod_{i<j}\left\{(1-\pi)\mathrm{N}\left(\omega_{ij}|0, v_0^2\right) + \pi\mathrm{N}\left(\omega_{ij}|0, v_1^2\right)\right\}\prod_i \exp\left(\omega_{ii}|\frac{\eta}{2}\right) 1_{\{\Omega\in PD_G\}} \tag{6}$$

where $\mathrm{Exp}(\omega|\eta)$ represents the exponential density with expectation $1/\eta$ and $1_{\{\cdot\}}$ 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 η, fixing its value to 1 as done in Wang (2015). The prior on π is discussed later. The first product in Equation (6) involving the off-diagonal elements of Ω, is a mixture of two normal distributions. The second product multiplies M exponential densities for the diagonal elements of Ω. Now, recalling the connection between the graph G and its binary representation through the adjacency matrix $Z = (z_{ij})_{i<j}$, (6) can be rewritten as

$$p(\Omega|Z, \theta) = C(Z, v_0, v_1, \eta)^{-1}\prod_{i<j}\mathrm{N}\left(\omega_{ij}|0, v_{z_{ij}}^2\right)\prod_i \exp\left(\omega_{ii}|\frac{\eta}{2}\right) \tag{7}$$

$$p(Z|\theta) = C(\theta)^{-1}C(Z, v_0, v_1, \eta)\prod_{i<j}\left\{\pi_{ij}^{z_{ij}}(1-\pi_{ij})^{1-z_{ij}}\right\} \tag{8}$$

where $v_{z_{ij}}^2 = v_1^2$ if $z_{ij} = 1$ and $v_{z_{ij}}^2 = v_0^2$ if $z_{ij} = 0$, and $C(\theta)$ and $C(Z, v_0, v_1, \eta)$ are normalising constant for the respective densities. The joint distribution $p(\Omega, Z|\theta)$ admits (6) as a marginal distribution for Ω. In the representation in Equations (7) and (8), small values of v_0 give high probability to the event $z_{ij} = 0$, so that the distribution of ω_{ij} is concentrated around 0, implying that the correspondent edge will have a close-to-zero probability to be included in the graph G. For an appropriately chosen large value of v_1, the event $z_{ij} = 1$ implies that the distribution of ω_{ij} is the diffuse component $\mathrm{N}(0, v_1^2)$ and so ω_{ij} can be estimated to be substantially different from zero (Wang, 2015). See also Malsiner-Walli and Wagner (2018) for a comparison of spike and slab priors.

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 ω_{ij} is small enough to be replaced by zero. Wang (2015) discuss 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 η is easier, as with standardised data, a choice of $\eta = 1$ assigns probability to the entire region of plausible values for the inverse variances ω_{ii}. There is a wealth of literature regarding the choice of the prior distribution for π_{ij}, the edge inclusion probability. See, for example Carvalho and Scott (2009) and Tan et al. (2017) 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 π_{ij} as follows

$$\begin{aligned} \{\pi_{ij}\}_{i<j}|P &\overset{iid}{\sim} P \\ P|\alpha,\mu,P_0 &\sim \mathrm{GDP}(\alpha,\mu,P_0) \\ P_0|a_\pi,b_\pi &= \mathrm{Beta}(a_\pi,b_\pi) \\ \alpha|\alpha_a,\alpha_b &\sim \mathrm{Gamma}(\alpha_a,\alpha_b) \\ \mu &\sim \mathrm{Beta}(a_\mu,b_\mu). \end{aligned} \tag{9}$$

The choice of a nonparametric prior allows for flexible modelling of the edge inclusion probabilities. Moreover, we can tune the hyperprior parameters characterising the base measure P_0 to achieve the desired level of sparsity. The parameters α and μ control the clustering structure of the GDP (note that posterior clustering depends also on the choice of the base measure). The choice of the hyperparameters depends on the particular application. The model for e_{ij} is then given by:

$$\begin{aligned} e_{ij}|\pi_{ij} &\overset{ind}{\sim} \mathrm{Ber}(\pi_{ij}), \quad i<j \\ \{\pi_{ij}\}_{i<j}|P &\overset{iid}{\sim} P \\ P|\alpha,\mu,P_0 &\sim \mathrm{GDP}(\alpha,\mu,P_0). \end{aligned} \tag{10}$$

The above equations defines a GDP Mixture model (GDPM, Lo, 1984; Barcella et al., 2017) for $\{e_{ij}\}_{i<j}$. Recalling the discrete nature of the GDP. we can rewrite (10) as

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

where the π_k denote the (unique) locations of the GDP prior.

2.4 | 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 constrain 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 graph structures supported by a GDP prior. We follow the framework of Tan et al. (2017) 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 Supplementary Material):

1. Conditionally on π_{ij}, the probability that a node i is connected to a node j is π_{ij}.
2. The degree of a node i is distributed as a mixture of Binomial distributions, with mixing weights given by the GDP

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

where, once again for ease of notation, we have substituted the index (ij) with k. We have that $\mathrm{E}\left[D_i|P\right] = \sum_{k=1}^{\infty} \psi_k (M-1)\pi_k \quad Var\left[D_i|P\right] = (M-1)\sum_{k=1}^{\infty} \psi_k \pi_k \left[(1-\pi_k) + (M-1)\pi_k\right] - \left[\sum_{k=1}^{\infty} \psi_k (M-1)\pi_k\right]^2$.
3. Marginalising over the random measure, we obtain (see Supplementary Material):

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

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 hyperparameters 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 in Equation (9), we perform a sensitivity analysis for different values of α and μ and different parametrisation of the base distribution P_0. Figures S1 and S2 in Supplementary Material present the resulting degree distribution for different combinations of hyperparameters. It is evident that our prior choice is able to accommodate different shapes. However, simulations show that, by appropriate choice of hyperparameters, we can obtain an exponential decay in the tails of the functions, but not a power law decay.

2.5 | 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 subsamples, 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, Saegusa and Shojaie (2016) estimate multiple graphs specifying a global penalisation and using optimisation techniques, while in a Bayesian framework Peterson et al. (2015) propose a joint model for multiple GGMs employing a Markov random field prior, encouraging sharing of common edges. The Markov random field prior is also used by Lin et al. (2017) for multiple graphs presenting both spatial and temporal dependence. Also relevant are the works of Tan et al. (2017), which propose a multiplicative prior to capture common and group-specific structures, and of Bilgrau et al. (2020), which presents a penalisation approach to estimate multiple precision matrices, allowing for the incorporation of prior information. We propose to model multiple graphs through an extension of the GDP prior, that is 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 g-th group 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 is characterised by a specific sample size n_g and its own graph G_g, for $g = 1, \dots, R$. 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 Equations (7) and (8) are generalised to handle multiple precision matrices Ω_g, and therefore multiple adjacency matrices Z_g as follows:

$$p(\Omega_g|Z_g, \theta_g) = C(Z_g, \nu_0, \nu_1, \eta_g)^{-1}\prod_{i<j}\mathrm{N}\left(\omega_{g,ij}|0, \nu^2_{z_{g,ij}}\right)\prod_i \exp\left(\omega_{g,ii}|\frac{\eta_g}{2}\right) \tag{11}$$

$$p(Z_g|\theta_g) = C(\theta_g)^{-1}C(Z_g, \nu_0, \nu_1, \eta_g)\prod_{i<j}\left\{\pi_{g,ij}^{z_{g,ij}}(1-\pi_{g,ij})^{1-z_{g,ij}}\right\}. \tag{12}$$

The hyperparameters ν_0^2 and ν_1^2 remain unchanged and are common to all groups. We can see that, conditional on the inclusion probabilities $\boldsymbol{\pi}_{g,ij}$, η_g, ν_0^2 and ν_1^2, Equations (11) and (12) are independent across groups. The prior in Equation (9) on $\pi_{g,ij}$ can be extended in the 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 ψ_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 DGDP, which is defined as follows. Let

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

be the random measure associated to the g-th group, for $g = 1, \dots, R$. The locations are *iid* draws from a common base measure P_0, as before. The weights still admits the stick-breaking representation:

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

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

Each φ_{kg} has a beta distribution, $\mathrm{Beta}(\alpha\mu_g, \alpha(1-\mu_g))$, but now μ_g is group specific. (Barcella et al., 2017) propose to introduce dependence across the $\{\mu_g\}$ employing a beta regression framework and letting the μ_g depend on a categorical covariates denoting group. Using the DGDP, the model in Equation (9) can then be extended to the multiple graphs as follows, for $g = 1, \dots, R$:

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

where $\boldsymbol{x}_g$ 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 parametrisation. 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 & 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 Supplementary Material.

3 | SIMULATION RESULTS

We perform a simulation study to investigate the efficacy of the proposed model. We simulate data from multivariate normal distributions, focussing on the estimation of the precision matrix and the multiple graphs. We present here two simulation scenarios, while further simulated examples can be found in Section 4 of Supplementary Material, together with details on the required computational times.

To assess the performance of the proposed model, we consider 20 replicas for each scenario described and we compare the resulting estimates with existing methods. In particular, we consider: (a) the ANOVA-DDP Ishwaran and James (2001) and Barcella et al. (2017); (b) a parametric version of the proposed SSSL model, in which the edge-inclusion probabilities are beta distributed and group specific; (c) the Bayesian structure learning model of Mohammadi et al. (2015), based on a birth–death MCMC in a standard conjugate model specification involving a multivariate normal likelihood and a G-Wishart prior distribution for the precision matrix, available through the `R` package *BDgraph*; (d) the graphical Lasso (Friedman et al., 2008) as implemented in the `R` package *glasso*; (e) the graphical group Lasso (Danaher et al., 2014), which imposes an additional regularisation between multiple precision matrices to enforce a similar structure, as implemented the `R` package *JGL*. We use different metrics for the comparisons: (a) the receiver operating characteristic-area under the curve (ROC-AUC), which is a normalised measure of the area under the ROC curve created by plotting the true positive rate against the false positive rate at various thresholds; (b) the mean square error (MSE), evaluated as the mean squared difference between the true precision matrices and the estimated ones; (c) the MSE restricted to the set of non-zero entries present in the simulated graph. Notice that some of the models implemented for comparison purposes do not allow for direct estimation of the multiple graph structures, namely the graphical Lasso and the graphical group Lasso. Therefore, these two methods are only compared in terms of MSE in the following analyses. Throughout all simulations, we run the Bayesian algorithms for 15000 iterations (of which 10,000 are discarded for the burn-in period) and we use the following parameter specifications:

1. **DGDP** $K = 10$, $\nu_0 = 0.01$, $h = 100$, $\alpha_a = 0.1$, $\alpha_b = 1$, $a_\pi = 0.01$, $b_\pi = 0.01$, $\boldsymbol{\zeta}_\mu = \mathbf{0}$, $\boldsymbol{\zeta}_\mu = \mathrm{I}_G$
2. **ANOVA-DDP** $K = 10$, $\nu_0 = 0.01$, $h = 100$, $\alpha_a = 0.1$, $\alpha_b = 1$, $\boldsymbol{\eta}_\mu = \mathbf{0}$, $\boldsymbol{\eta}_\Sigma = \mathrm{I}_G$
3. **Parametric** $\nu_0 = 0.01$, $h = 100$, $a_\pi = 1$, $b_\pi = 1$
4. **BDgraph** Default parametrisation, G-Wishart diagonal base measure = I_M
5. **Graph-LASSO** $\rho = 0.15$
6. **Group-LASSO** $\lambda_1 = 0.15$, $\lambda_2 = 0.2$

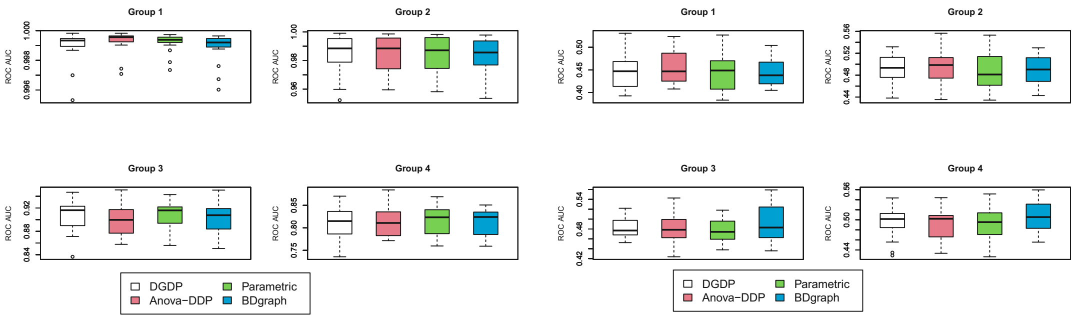
3.1 | Scenarios with 20 nodes and 4 groups

We first generate four multiple graphs following the guidelines of Peterson et al. (2015). We construct four precision matrices $\Omega_1, \Omega_2, \Omega_3$ and Ω_4 corresponding to graphs G_1, G_2, G_3 and G_4, of $M = 20$ nodes (for a total number of possible edges of $r = M(M-1)/2 \times 4 = 760$). We first define the precision matrix Ω_1 and then we derive the others as a perturbation of the first. Ω_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 $i = 1, \ldots, 8$, while the rest of the elements are set to zero. This defines an AR structure for the element of Ω_1. The total number of non-zero off-diagonal elements is 37. To construct Ω_2, we remove ten edges at random from Ω_1, setting the corresponding entries to zero. Then, we randomly add ten edges that are not present in Ω_1, giving a value of 0.5 to the new precision coefficients. The procedure is repeated similarly for Ω_3 and Ω_4, avoiding the replacement of edges that were previously modified. 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 `R` package `Matrix`. The precision matrices $\Omega_2, \Omega_3, \Omega_4$ 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 60, 50, 50 and 40 for group 1, 2, 3 and 4, respectively.

The second simulation scenario is similar to the first, but Ω_1 is now a unit diagonal matrix and we add 60 non-zero off-diagonal elements, chosen randomly from the r possible edges, and fix the corresponding entries of Ω_1 to 0.5. Ω_2, Ω_3 and Ω_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 ROC-AUC distributions for 10 simulated data sets for the first and second scenarios are displayed in Figure 1. The distributions are concentrated between 0.7 and 1 for all groups in the scenario with the AR structures, denoting the ability of all models to recover the true graphical structure. In the second scenario, the performance is very similar across models, but none of them seems to be able to effectively estimate the edges in the graph.

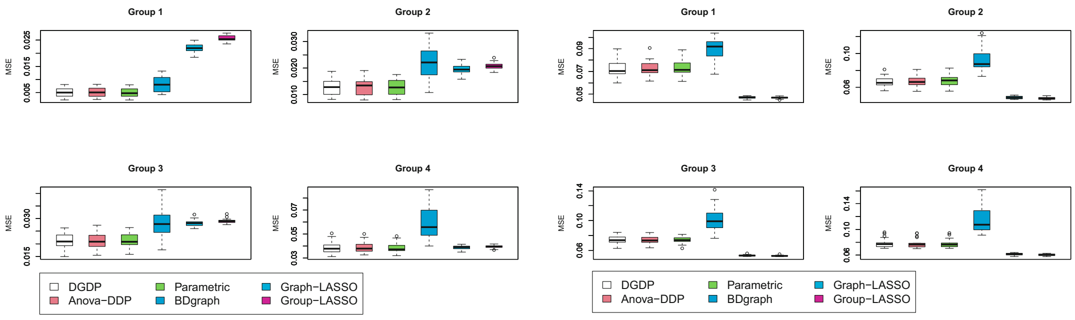
The boxplots of the MSE distributions are displayed in Figures 2 (full precision matrix) and 3 (considering only the non-zero entries of the precision matrix) for all models considered. The error metric takes low values in general, with the Lasso-based algorithms achieving the lowest values of MSE in the random structure scenario. The DGDP model shows an almost identical performance when compared with the ANOVA-DDP and the parametric models in both scenarios and all groups. The BDgraph yields higher MSE values in most groups and scenarios. In general, while the comparison between the models yields similar results in Figures 2 and 3,



(a) AR structure

(b) Random structure

FIGURE 1 Scenarios with 20 nodes and 4 groups: receiver operating characteristic-area under the curve (ROC-AUC) boxplots comparing dependent generalised Dirichlet process, ANOVA-DDP, parametric and BDgraph models over the two simulated scenarios. The ROC-AUC distributions are evaluated over 20 replicas for each scenario.



(a) AR structure

(b) Random structure

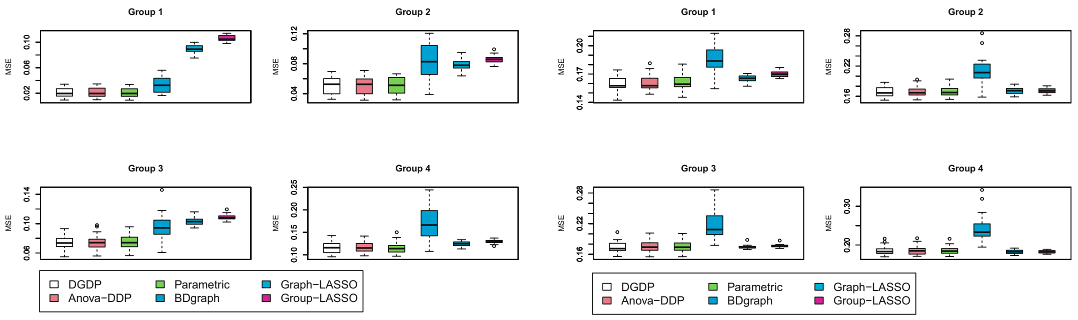
FIGURE 2 Scenarios with 20 nodes and 4 groups: mean square error (MSE) boxplots comparing dependent generalised Dirichlet process, ANOVA-DDP, parametric and BDgraph models plus graph-Lasso and group-Lasso, over the two simulated settings. The MSE distributions are evaluated over 20 replicas for each scenario. The MSE is computed considering all the entries of the precision matrices.

overall the MSE values are higher when only the non-zero entries of the precision matrix are considered.

Section 4.1 of Supplementary Material shows the results relative to an additional simulation based on the one just described, but characterised by $M = 200$ nodes and two groups with $n_1 = n_2 = 100$ in the first scenario, while $n_1 = 100$ and $n_2 = 50$ in the second one. The comparison with alternative models, based on the computation of ROC-AUC and MSE values for 10 replicas, shows comparable performance of the proposed model with the ANOVA-DDP and parametric models, and improvements with respect to the BDgraph and Lasso-based methods.

3.2 | Scenario with 91 nodes and 3 groups

We investigate the model performance on three imbalanced groups as in the SABRE study. We construct three graphs, with $M = 91$ nodes each, which is equal to the number of metabolites used



(a) AR structure

(b) Random structure

FIGURE 3 Scenarios with 20 nodes and 4 groups: mean square error (MSE) boxplots comparing dependent generalised Dirichlet process, ANOVA-DDP, parametric and BDgraph models plus graph-Lasso and group-Lasso, over the two simulated settings. The MSE distributions are evaluated over 20 replicas for each scenario. The MSE is computed considering only the non-zero entries of the precision matrices.

in the application presented in Section 4. The precision matrices of each group are set equal to the empirical precision matrix among all metabolites of each ethnic group in the SABRE study, setting to zero those elements smaller than 0.1 in absolute value. The resulting graphs have 683, 706 and 2259 non-zero edges, respectively. The new empirical precision matrices are not positive definite; therefore, we use as before the R package *nearPD* (Higham, 2002). We simulate observations for each group using the same sample sizes of the three groups in the SABRE study, that is 1103, 978 and 119.

In Figure 4, we show the ROC-AUC for each group. We can see that the ROC-AUC values are concentrated around 0.9 for the DGDP model on the first and second group, and are higher than 0.6 for the third group, which is clearly the most difficult to accurately estimate because of the much smaller sample size. In general, the DGDP model outperforms the alternative models considered.

In Figure 5 we report the boxplots of the MSE, where we can notice a good performance of the proposed model when compared to BDgraph and Lasso-based models. The performance in terms of MSE compared to the ANOVA-DDP and the parametric models is slightly worse for the DGDP, but comparable. Once again, the MSE values are in general higher when only the non-zero entries of the precision matrix are considered.

We also examine the posterior distribution of the partition induced by the proposed model, in comparison with the one obtained from the Anova-DDP model. In Figure 6, we present a summary of the posterior distribution of the number of clusters $K^{\star}$ within each group. The distribution of the number of clusters is concentrated on lower values for the DGDP model, indicating that the proposed model favours coarser partitions. Considering the partition estimate obtained by minimising the Binder loss function (Binder, 1978; Lau & Green, 2007), the medians and ranges of the number of clusters across replicas for the three groups are $1\,(1,2)$, $1\,(1,3)$ and $1\,(1,2)$ for the DGDP. For the ANOVA-DDP, they are equal to $6\,(1,9)$, $6\,(1,10)$ and $5\,(1,10)$, respectively. To better highlight the features of the partitions implied by the two models, in Table 1, we report summary statistics for the following features of the partitions: the number of singleton clusters (posterior median and range) and the size of the largest cluster (posterior median and range). The results are averaged over the 20 replicas. The Table shows that the DGDP provides a more parsimonious representation of the data in terms of number of singleton clusters and size of the largest cluster.

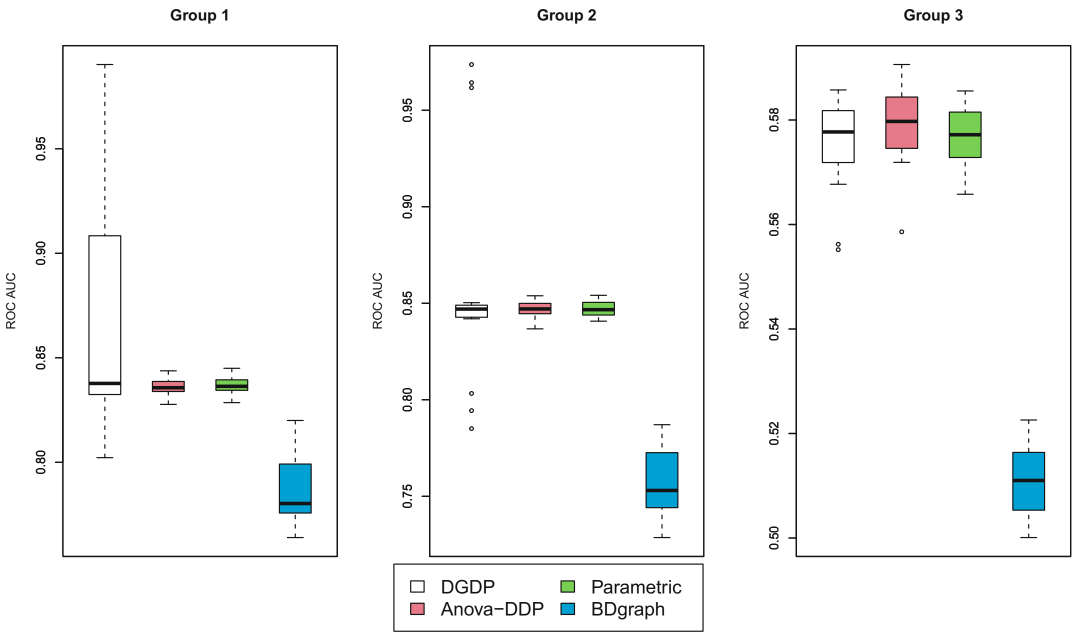


FIGURE 4 Scenario with 91 nodes and 3 groups: receiver operating characteristic-area under the curve boxplots obtained over 20 replicas of the proposed scenario, comparing the dependent generalised Dirichlet process, the ANOVA-DDP, the parametric and the BDgraph models.

Section 4.2 of Supplementary Material shows the results for an additional simulation similar to the one described in this section, with the inclusion of a non-zero mean term. The comparison shows a general difficulty in estimating the graph structure in the smallest group, especially when the graph structure in the third group is characterised by a large number of edges, as in our case.

4 | SABRE RESULTS

In this Section, we fit the proposed model for multiple GGMs to the SABRE metabolic data set. The data set described in Section 1 has a total of 2200 observations, stratified in three ethnicities, 1103 Europeans, 978 South Asians and 119 African-Caribbean. The number of nodes (i.e. the number of equations in the SUR model) is $M = 91$, a list of which can be found in Table S1 in Supplementary Material. The number of metabolites is reduced here from 200 to 91 because we focus on the absolute concentrations of the metabolites and we exclude the ratios between the concentrations. As predictors in the regression term of the mean, we include the covariates listed in Table S2, consisting of measures of body-fat distributions, liver health and other risk factors, such as smoking habits, sex and age (the total number of the covariates is $p = 18$ with an additional 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 Equation (11) 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

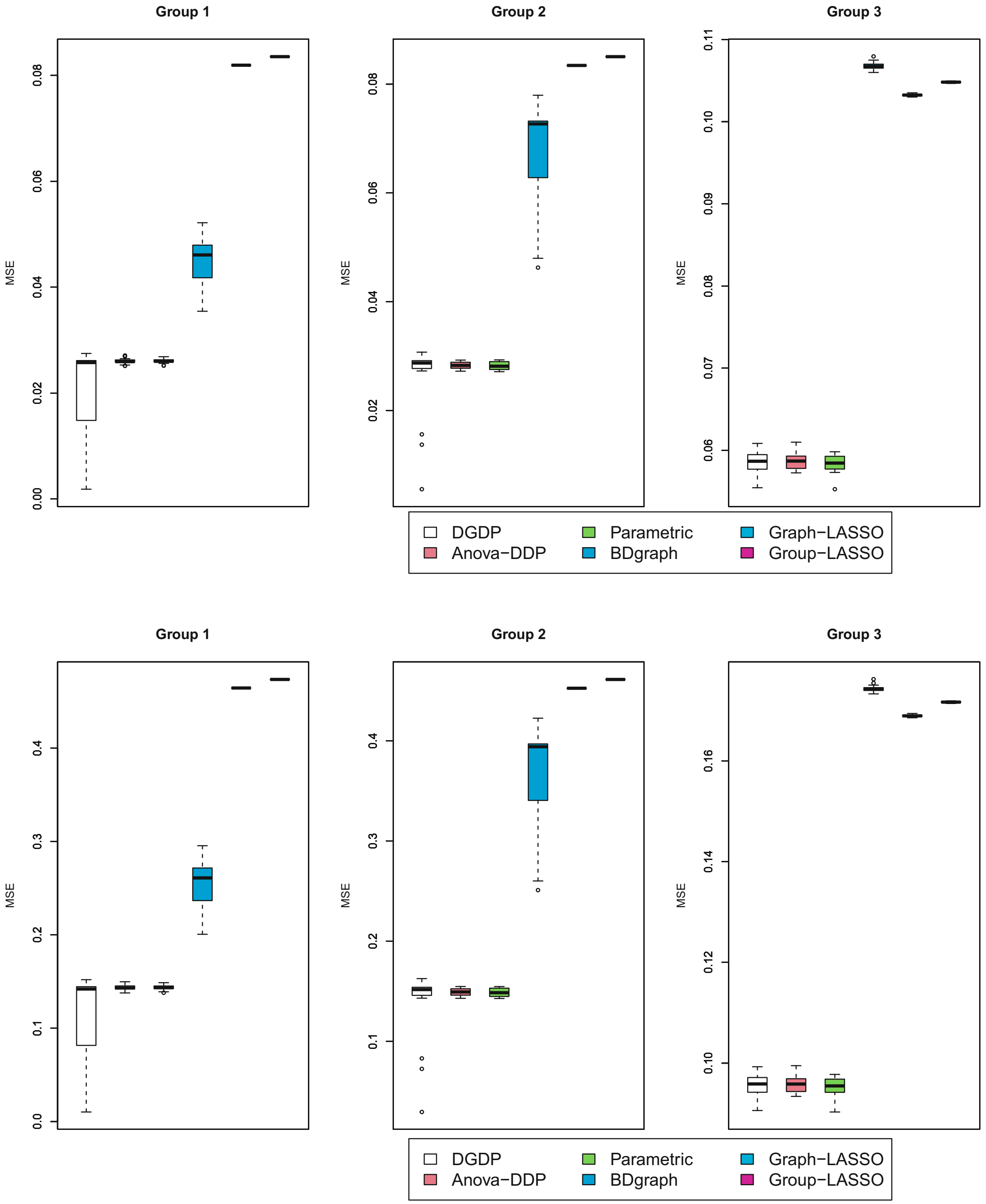


FIGURE 5 Scenario with 91 nodes and 3 groups: mean square error (MSE) boxplots obtained over 20 replicas of the proposed scenario, comparing the dependent generalised Dirichlet process, the ANOVA-DDP, The parametric, the BDgraph, the graph-Lasso and the group Lasso models. The top row refers to MSE computed considering all the entries of the precision matrices, while the bottom row only considers the non-zero entries of the precision matrices.

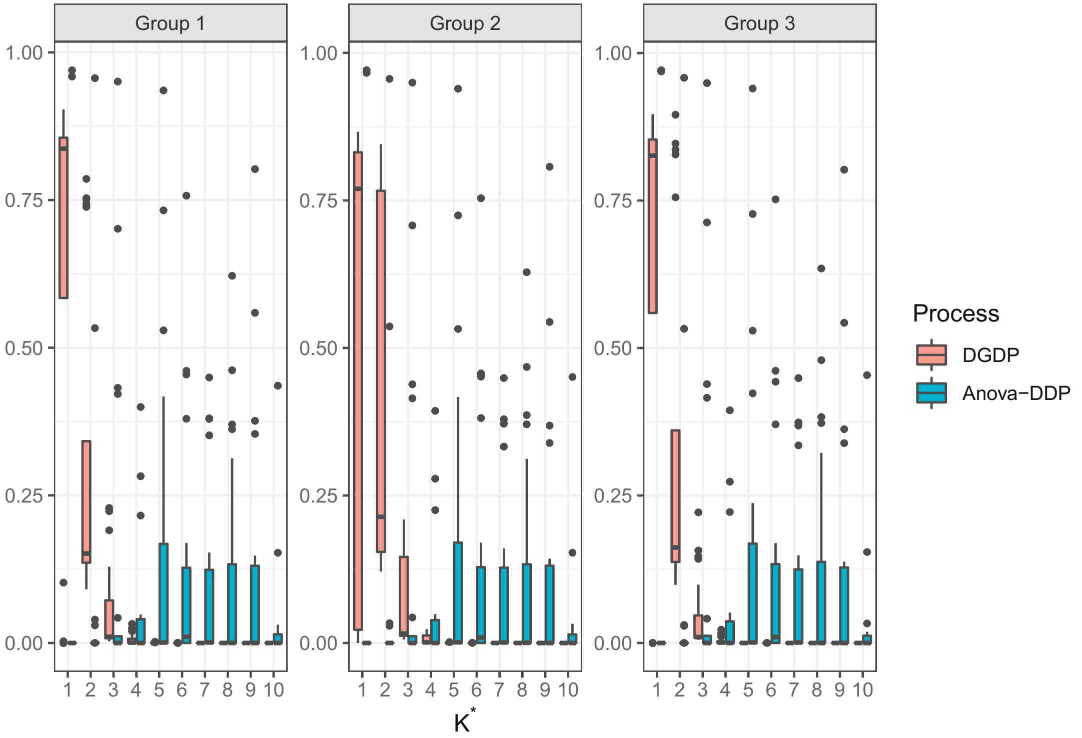


FIGURE 6 Scenario with 91 nodes and 3 groups: posterior distribution of the number of clusters for the dependent generalised Dirichlet process and ANOVA-DDP models, for each group. The x-axis reports the number of clusters. The boxplots are obtained from 20 replicas and represent the distribution across replicas of the probability of the number of clusters to be equal to a particular value.

Beta($a_\pi = 0.5, b_\pi = 0.5$). The concentration parameter α 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 Equation (5). We run the MCMC for 12,000 iterations, comprising a burn-in period of 2000 iterations and thinning every 5 iterations. In addition to the multiple graphs, using the output of the MCMC algorithm, we also estimate the differential networks (De la 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, as it was also observed in the simulation scenarios with analogous sample sizes of Section 3.2 and 4.2 in Supplementary Material.

In Figure 7 are shown the posterior means of the regression coefficients β_{lj}, for equation $l = 1, \ldots, M$ and covariate $j = 1, \ldots, p$. The only covariates that do not show association with any metabolite are waist-to-hip ratio and diastolic blood pressure. It is worth noting that WHR has a strong positive correlation with some of the other measures of adiposity, such as sagittal

TABLE 1 Scenario with 91 nodes and 3 groups: summary statistics of the number of singleton clusters and size of the largest cluster (posterior median and range) under the dependent generalised Dirichlet process (DGDP) or ANOVA-DDP models. The estimates are obtained by averaging over the 20 replicas

Summary	Group	DGDP	ANOVA-DDP
Number of singletons	1	0 (0, 2.45)	0.3 (0, 2.95)
	2	0 (0, 2.15)	0.3 (0, 3.10)
	3	0 (0, 2.40)	0.3 (0, 2.90)
Size of the largest cluster	1	3993.05 (3882.8, 4017.2)	2839.4 (2723, 2959.3)
	2	3954.95 (3743.85, 4011.3)	2837.5 (2722.55, 2956.05)
	3	3885.07 (3719.2, 3959.95)	2838.2 (2731.6, 2958.15)

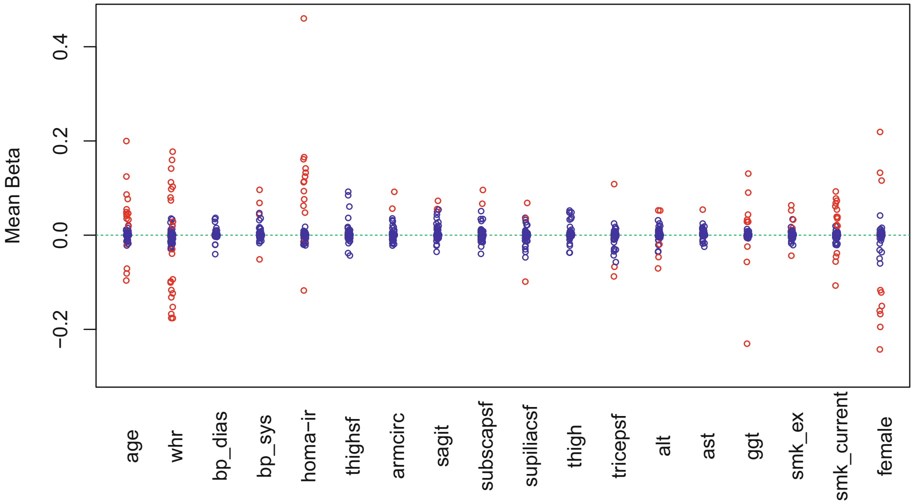


FIGURE 7 Southall And Brent REvisited study: Each dot represents the mean of the posterior distribution of a coefficient β_{lj}. Red dots denote coefficients whose 95% credible interval does not contain the zero.

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.

We summarise the attributes of the three group-specific networks inferred by the proposed model (i.e., Europeans, South Asians and African-Caribbean) in Table 2, while their plots can be found in Figures S11, S12 and S13 of Supplementary Material. Table 2 shows the main features of the individual networks, namely: the number of inferred edges, the number of connected and isolated nodes; the average edge and node betweenness (i.e. the number of shortest paths passing through an edges or node); the average node degree (i.e. the number of connections departing a node) and the average clustering coefficient (also called transitivity, i.e. the probability that the adjacent nodes of a node are connected). All these measures are computed using the R package igraph. The individual networks are characterised by a high number of edges, especially in the European and South Asian groups, connecting all nodes in the networks. In particular, the first two groups are very similar in all the features reported in Table 2, while the African-Caribbean group is characterised by less connections and lower betweenness and clustering coefficient. This indicates different network organisations and metabolite interactions

TABLE 2 Southall And Brent REvisited (SABRE) study: properties of the inferred networks for each ethnic group

	Europeans	South Asians	Africans-Caribbean
Number_edges	2803	2834	632
Number_connected_nodes	91	91	91
Number_isolated_nodes	0	0	0
Edges_betweenness	1.922	1.890	14.372
Nodes_betweenness	0.004	0.003	0.014
Node_degree	61.604	62.286	13.890
Clustering coeff.	0.846	0.838	0.423

within the three ethnic groups, especially when comparing the European and South Asian groups with the African-Caribbean group.

In Figure 8, we show the differential network between Europeans and South Asians, where an edge between two nodes is added to the differential graph if the posterior mean of the absolute difference between the adjacency matrices in the corresponding position is higher than 0.5. This quantity is computed by averaging over the posterior MCMC chain. 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 the underlying metabolic processes. For example, the amino acid *Histidine* features many connections with other amino acids and a subset of lipoprotein subfractions, with these edges only present in the South Asian group. Other central nodes in the differential network are *acetoacetate*, *acetate*, *pyruvate* and *lactate*.

We also show the differential networks between Europeans or South Asians and African-Caribbean in Figures S14 and S15 in Section 5 of Supplementary Material. These differential networks have more edges compared to the one between Europeans and South Asians, due to the fact that the network for the African-Caribbean ethnic group is much sparser. Therefore, in order to highlight the few connections unique to the Africans Caribbean, we limit the inclusion of differential edges in the Europeans and South Asian groups to those with an inclusion probability higher than 0.8.

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 of Figure 8. 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 conditions identified by this pathway analysis, *pyruvate dehydrogenase deficiency (E3)*, *pyruvate carboxylase deficiency*, *diabetes mellitus (MODY) non-insulin-dependent* and *chronic progressive external ophthalmoplegia (CPEO)/Kearns–Sayre syndrome (KSS)*. Pyruvate dehydrogenase and pyruvate carboxylase

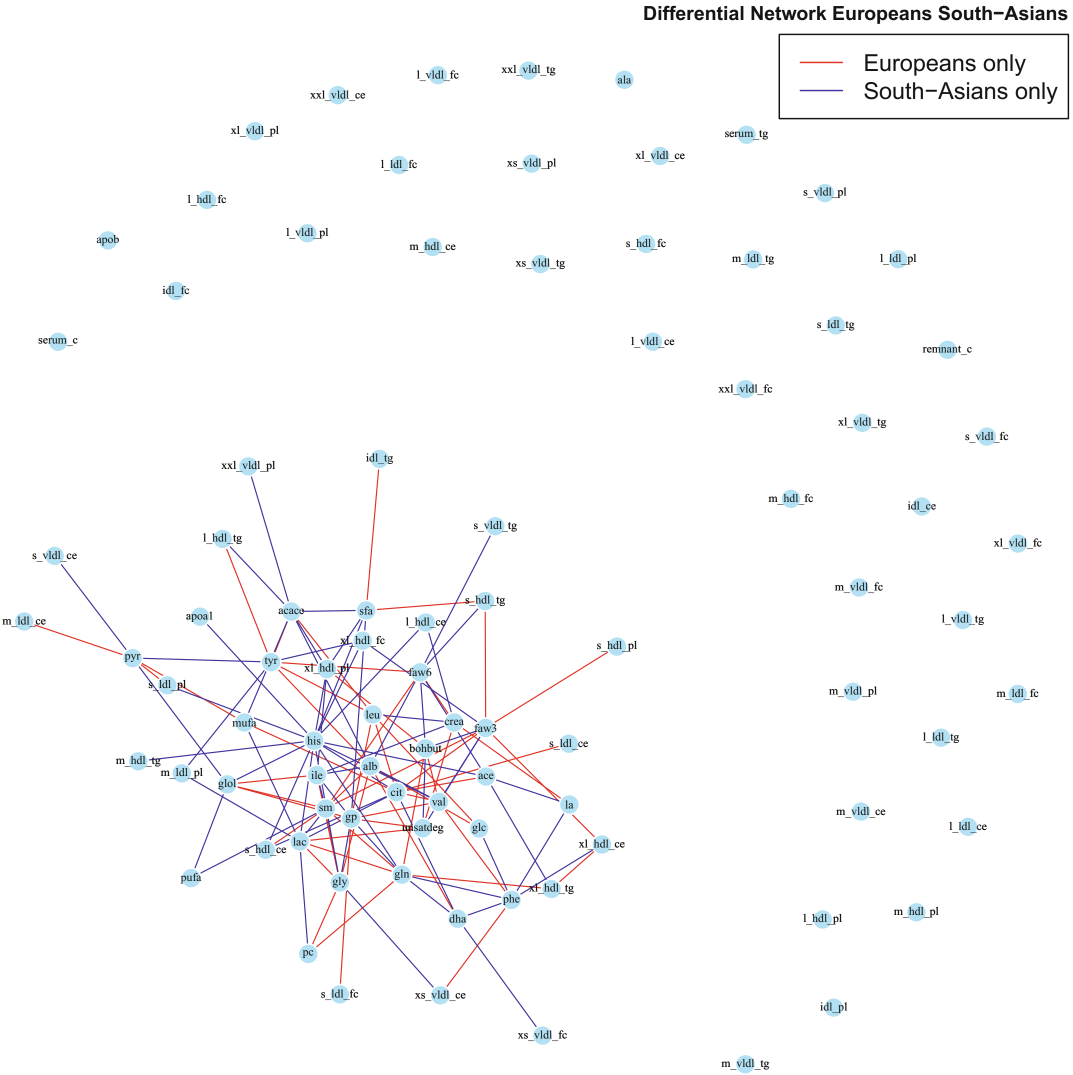


FIGURE 8 Southall And Brent REvisited study: differential network between Europeans and South Asians groups. The red edges represent the connections between metabolites for the European ethnicity whose probability of inclusion in the differential network is higher than 0.5, while blue edges represent the connections between metabolites for the South Asian ethnicity whose probability of inclusion in the differential network is higher than 0.5.

deficiency are the most common disorders in pyruvate metabolism. Pyruvate dehydrogenase (PDH) is an enzyme complex made of three catalytic subunits, pyruvate dehydrogenase (E1), dihydrolipoamide acyltransferase (E2) and dihydrolipoamide dehydrogenase (E3), and two cofactors, thiamine pyrophosphate and lipoic acid. The enzyme complex converts pyruvate, after it enters the mitochondria, 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 E1 subunit, while mutations in E2 and E3 are less often the cause for PDH Complex Deficiency. The enzyme defect causes more pyruvate to be metabolised to lactate and leads to lactic acidosis (Bissonnette & 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, PDH Complex is important in regulating the glucose metabolism with PDH deficiency 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 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 may be further metabolised in two different means. In non-gluconeogenic tissues, like muscles and brain, pyruvate is decarboxylated to form acetyl-CoA catalysed by the pyruvate dehydrogenase complex. In gluconeogenic tissues, where pyruvate carboxylase is highly abundant, 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. MODY is an autosomal dominant monogenic disorder of pancreatic beta cells that usually manifests itself before the age of 30 and accounts for 1%–3% of diabetes in this age group (Misra & 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, 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 KSS, 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 retinopathy, optic atrophy, and rarely, cataracts. Cardiac manifestations include cardiac conduction block and cardiomyopathy. Cerebral manifestations include epilepsy, cerebellar ataxia and dementia. The peripheral nervous system can also be affected, typically with axonal sensory neuropathy. Endocrine involvement includes diabetes mellitus, hypothyroidism, hypoparathyroidism and hypogonadism. Sensorineural hearing loss and gastrointestinal involvement are also possible (Vorgerd & Deschauer, 2011). In interpreting the pathway over-representation analysis, we note that all the highlighted conditions are associated with defective mitochondrial function and altered pyruvate metabolism. We therefore speculate that alterations in metabolic flexibility and mitochondrial aerobic metabolism (Smith et al., 2018) may be a fruitful area for further study. Consistent with this suggestion, a previous small experimental study reported that overweight South Asian men had impaired metabolic flexibility compared with matched European counterparts (Bakker et al., 2015) and we have previously observed poorer oxidative capacity in skeletal muscle independent of diabetes in South Asians compared with Europeans (Jones et al., 2020).

To evaluate the sensitivity of the model to the prior choice of ν_0, playing a major role in determining the level of sparsity of the graphs, we repeat the analysis for $\nu_0 \in \{0.01, 0.1\}$. A comparison of the estimates given $\nu_0 = 0.01$ and $\nu_0 = 0.1$ shows differences in the number of edges included in each of the three graphs and consequently in the differential networks. However, the main characteristics of the individual graphs and the differential network are maintained. For example, the differential network between Europeans and South Asians estimated with $\nu_0 = 0.1$, reported in Figure S16 in Supplementary Material, highlights similar connectivity patterns to the one estimated with $\nu_0 = 0.01$, for example the amino acid Histidine presents connections with other amino acids and lipids subfractions, with these edges being only present in the South Asian group, as before. Furthermore, varying the prior values of π_a, π_b between 0.01 and 0.5 does not lead to substantial changes in posterior inference.

5 | CONCLUSIONS

This paper proposes the use of a GDP prior on the edge inclusion probabilities of a GGM together with the SSSL prior on the precision matrix. The model 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 data set. 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. Our findings are interesting because they can lead to formulation of new hypothesis, for example metabolic pathways associated with diabetes and cardiovascular disorders, and guide further experimentation.

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 & 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). The authors 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.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

SUPPLEMENTARY MATERIAL

The reader is referred to the online Supplementary Material for additional information regarding this manuscript, where one can find: proofs and a priori simulations regarding the degree distribution; details on the MCMC algorithm; additional tables reporting data features; additional simulation studies; the differential networks not reported in the main text.

The C++ code is made available at the following GitLab link: https://gitlab.com/molinari_marco/bayesiannonparametricggms

ORCID

Marco Molinari https://orcid.org/0000-0002-3374-9099

REFERENCES

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

Bakker, L.E., Guigas, B., van Schinkel, L.D., van der Zon, G.C., Streefland, T.C., van Klinken, J.B. et al. (2015) Middle-aged overweight south asian men exhibit a different metabolic adaptation to short-term energy restriction compared with europeans. *Diabetologia*, 58, 165–177.

Barcella, W., De Iorio, M., Favaro, S. & Rosner, G.L. (2017) Dependent generalized Dirichlet process priors for the analysis of acute lymphoblastic leukemia. *Biostatistics*, 19, 342–348.

Bhadra, A. & Mallick, B.K. (2013) Joint high-dimensional bayesian variable and covariance selection with an application to eqtl analysis. *Biometrics*, 69, 447–457.

Bilgrau, A.E., Peeters, C.F., Eriksen, P.S., Bøgsted, M. & van Wieringen, W.N. (2020) Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. *Journal of Machine Learning Research*, 21, 1–52.

Binder, D.A. (1978) Bayesian cluster analysis. *Biometrika*, 65, 31–38.

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

Cai, T.T., Li, H., Liu, W. & Xie, J. (2013) Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100, 139–156.

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

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

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G. et al. (2018) Metaboanalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, gky310. Available from: http://dx.doi.org/10.1093/nar/gky310.

Danaher, P., Wang, P. & Witten, D.M. (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 76, 373.

De Iorio, M., Müller, P., Rosner, G. & MacEachern, S. (2004) An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99, 205–215.

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

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

Deshpande, S.K., Rocková, V. & George, E.I. (2019) Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics*, 28, 921–931.

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

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

Ferguson, T.S. (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics*, 615–629.

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

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

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

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

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

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

Jones, S., Tillin, T., Williams, S., Eastwood, S.V., Hughes, A.D. & Chaturvedi, N. (2020) Type 2 diabetes does not account for ethnic differences in exercise capacity or skeletal muscle function in older adults. *Diabetologia*, 63, 624–635.

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

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

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

Lee, I.-K. (2014) The role of pyruvate dehydrogenase kinase in diabetes and obesity. *Diabetes & Metabolism Journal*, 38, 181–186.

Lenkoski, A. & Dobra, A. (2011) Computational aspects related to inference in gaussian graphical models with the g-wishart prior. *Journal of Computational and Graphical Statistics*, 20, 140–157.

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

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

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

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

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

Misra, S. & Owen, K.R. (2018) Genetics of monogenic diabetes: present clinical challenges. *Current Diabetes Reports*, 18, 141. Available from: https://doi.org/10.1007/s11892-018-1111-4

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

Peterson, C., Stingo, F.C. & Vannucci, M. (2015) Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110, 159–174.

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

Rothman, A.J., Levina, E. & Zhu, J. (2010) Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19, 947–962.

Saegusa, T. & Shojaie, A. (2016) Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics*, 10, 1341.

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

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

Smith, R.L., Soeters, M.R., Wüst, R.C. & Houtkooper, R.H. (2018) Metabolic flexibility as an adaptation to energy resources and requirements in health and disease. *Endocrine Reviews*, 39, 489–517.

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

Soininen, P., Kangas, A.J., Würtz, P., Suna, T. & Ala-Korpela, M. (2015) Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circulation: Cardiovascular Genetics*, 8, 192–206.

Sproston, K. & Mindell, J. (2006) Health survey for England 2004. the health of minority ethnic groups.

Tan, L.S., Jasra, A., De Iorio, M. & Ebbels, T.M. (2017) Bayesian inference for multiple gaussian graphical models with application to metabolic association networks. *The Annals of Applied Statistics*, 11, 2222–2251.

Valcárcel, B., Würtz, P., al Basatena, N.-K.S., Tukiainen, T., Kangas, A.J., Soininen, P. et al. (2011) A differential network approach to exploring differences between biological states: an application to prediabetes. *PLoS ONE*, 6, e24702.

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

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

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

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

Zellner, A. (1971) An introduction to bayesian inference in econometrics. *Technical Report.*

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Molinari, M., Cremaschi, A., De Iorio, M., Chaturvedi, N., Hughes, A.D. & Tillin, T. (2022) Bayesian nonparametric modelling of multiple graphs with an application to ethnic metabolic differences. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1–24. Available from: https://doi.org/10.1111/rssc.12570