Bayesian semiparametric analysis of multivariate continuous responses, with variable selection

Georgios Papageorgiou and Benjamin C. Marshall

Department of Economics, Mathematics and Statistics

Birkbeck, University of London, UK

g.papageorgiou@bbk.ac.uk

February 12, 2020

Abstract

This paper presents an approach to Bayesian semiparametric inference for Gaussian multivariate response regression. We are motivated by various small and medium dimensional problems from the physical and social sciences. The statistical challenges revolve around dealing with the unknown mean and variance functions and in particular, the correlation matrix. To tackle these problems, we have developed priors over the smooth functions and a Markov chain Monte Carlo algorithm for inference and model selection. Specifically: Dirichlet process mixtures of Gaussian distributions is used as the basis for a cluster-inducing prior over the elements of the correlation matrix. The smooth, multidimensional means and variances are represented using radial basis function expansions. The complexity of the model, in terms of variable selection and smoothness, is then controlled by spike-slab priors. A simulation study is presented, demonstrating performance as the response dimension increases. Finally, the model is fit to a number of real world datasets. An R package, scripts for replicating synthetic and real data examples, and a detailed description of the MCMC sampler are available in the supplemental materials online.

Keywords: Clustering; Covariance matrix models; Model averaging; Multivariate response regression; Seemingly unrelated regression models; Semiparametric regression

1 Introduction

Many systems are too complex to be adequately described by a single response variable. For example, in medical investigations, understanding how the body reacts to certain drugs requires multiple blood tests. Similarly, in the social sciences, multiple exams are needed in order to build a complete picture of a student's academic ability. Scientific investigations into these systems therefore produce multiple outcome variables. Typically the outcomes are correlated and ignoring this correlation can result in loss of optimality. Multivariate response models are needed for the analysis of data arising from these and many other experimental setups. Our main goal here is to develop Bayesian multivariate response models for continuous responses with nonparametric models for the mean vectors and covariance matrices, assuming a multivariate Gaussian likelihood.

Modelling unconstrained means nonparametrically, as general functions of the covariates, is straight forward and by now fairly standard. In the work that we present here, nonparametric effects are represented as linear combinations of radial basis functions. Generally, our approach is to utilize a large number of basis functions because this enables flexible estimation of true effects that are locally adaptive. Potential over-fitting is mitigated by utilizing spike-slab priors for variable selection and regularization (see e.g. O'Hara and Sillanpää (2009) for a review on variable selection methods).

Modelling covariance matrices nonparametrically is not as straight forward as modelling the means, due the positive definiteness constraint that complicates matters. To overcome this constraint and model the elements of the covariance matrix in terms of regressors, a first, necessary step is to decompose the covariance matrix $\mathbf{\Sigma}$ into a product of matrices. Such decompositions include the spectral and Cholesky, and variations of the latter. Pinheiro and Bates (1996) review the spectral and Cholesky decompositions with several different parametrisations. Based on the spectral decomposition and the matrix logarithmic transformation, Chiu et al. (1996) model the structure of a covariance matrix in terms of explanatory variables. Pourahmadi (1999) and Chen and Dunson (2003) describe two modifications of the Cholesky decomposition that result in statistically meaningful, unconstrained reparametrisation of the covariance

matrix, provided that there is a natural ordering in the responses (Pourahmadi, 2007), as it happens in longitudinal studies, where the time of observation provides this natural ordering.

The spectral and the modified Cholesky decompositions, outside the context of longitudinal studies, lack simple statistical interpretation, making it difficult for practitioners to incorporate prior beliefs into the model. A decomposition, however, that is statistically simple and intuitive, comes from the separation strategy of Barnard et al. (2000), according to which $\boldsymbol{\Sigma}$ is separated into a diagonal matrix of variances $\boldsymbol{S}$ and a correlation matrix $\boldsymbol{R}$, $\boldsymbol{\Sigma} = \boldsymbol{S}^{1/2}\boldsymbol{R}\boldsymbol{S}^{1/2}$. This decomposition makes it easy to model the variances in terms of covariates as the only constrained on them is the positiveness. Here we use a log-link and linear predictors that are constructed in the same way as for the mean parameters.

Chan et al. (2006) describe several reasons why allowing the variances to be general functions of the covariates is meaningful. Firstly, prediction intervals obtained from heteroscedastic regression models can be more realistic than those obtained by assuming constant error variance, or as Müller and Mitra (2013) put it, it can result in more honest representation of uncertainties. Secondly, it allows the practitioner to examine and understand which covariates drive the variances, and in the multivariate response case, examine if the same or different subsets of covariates are associated with the variances of the responses. Thirdly, modelling the variances in terms of covariates results in more efficient estimation of the mean functions. Lastly, it produces more accurate standard errors for the estimates of unknown parameters.

Our approach for variable selection and model averaging can be thought of as a generalization of the approach of George and McCulloch (1993) who describe methods for univariate linear regression and the approach of Chan et al. (2006) and Papageorgiou (2018) who focus on methods for flexible mean and variance modelling for a single response. The current paper is a generalization of the work of Chan et al. (2006) and Papageorgiou (2018) from univariate to multivariate responses. Whereas in the univariate case one has to fit a single smooth mean and a single smooth variance function, in the multivariate case, multiple such functions have to be fit. However, the representation of these functions, and their prior distributions, are constructed in the same way as in the univariate case. The most important challenge that one has to face when dealing with multivariate regression is modelling the correlation matrix and sampling from its posterior. In this paper we discuss three intuitive correlation matrix priors and strategies for posterior sampling. In addition, we develop an efficient stochastic search variable selection algorithm by using

Zellner's g-prior (Zellner, 1986) that allows integrating out the regression coefficients in the mean function. Further, in our Markov chain Monte Carlo (MCMC) algorithm, we generate the variable selection indicators in blocks (Chan et al., 2006; Papageorgiou, 2018) and choose the MCMC tuning parameters adaptively (Roberts and Rosenthal, 2001).

Of course, the separation of the variances from the correlations alone does not solve the problem of positive definiteness, as the constraint has now been transferred from the covariance matrix $\boldsymbol{\Sigma}$ to the correlation matrix $\boldsymbol{R} = \{r_{kl}\}$. Here, we place a normal prior on the Fisher's z transformation of the nonredundant elements of $\boldsymbol{R}$, $\log\{(1+r_{kl})/(1-r_{kl})\}/2 \sim N(\mu_R, \sigma_R^2) I[\boldsymbol{R} \in \mathcal{C}]$, where $\mathcal{C}$ denotes the space of correlation matrices and $I[.]$ denotes the indicator function that restricts the range of the correlations and induces dependence among them (Daniels and Kass, 1999). We rely on the 'shadow prior' of Liechty et al. (2004) to maintain positive definiteness. The model is intuitive and easy to interpret, allowing practitioners to represent their substantive prior knowledge.

However, the normal model for the correlations is quite restrictive, and this can have a negative impact on the estimated correlations, especially in small samples (Daniels and Kass, 1999). Here, to achieve a nonparametric model for the correlation matrix, we consider mixtures of normal distributions $\log\{(1+r_{kl})/(1-r_{kl})\}/2 \sim \sum_h \pi_h N(\mu_{R,h}, \sigma_R^2) I[\boldsymbol{R} \in \mathcal{C}]$ for the transformed r_{kl}. This is in the spirit of the 'grouped correlations model' of Liechty et al. (2004) who also propose a 'grouped variables model'. The latter clusters the variables instead of the correlations and it is more structured than the nonparametric grouped correlations model. Here, we consider both the grouped correlations and variables models.

In what follows, we work with generic Dirichlet process (Ferguson, 1973) mixtures of normal distributions for the correlations, utilizing the stick breaking construction (Sethuraman, 1994). However, one of the attractive features of the grouped correlations and variables models is that they allow the researcher to represent prior information and beliefs about the strength of correlations among variables and the general structure of the correlation matrix. See Liechty et al. (2004) and Tsay and Pourahmadi (2017) for examples on structured correlation matrices.

Our work is related to two further strands of the literature. The first one is known as 'seemingly unrelated regressions' (SUR) and it originates from the work of Zellner (1962). The second one is known as 'generalized additive models for location, scale and shape' (GAMLSS) and it originates from the work

of Rigby and Stasinopoulos (2005).

Concerning SUR, Zellner (1962) showed how efficiency gains can be achieved by simultaneous estimation of linear regression equations, accommodating potentially correlated error terms. This gain in efficiency, measured in terms of reduction in the variance of the estimates of regression coefficients, can be substantial when the correlations among the error terms are high and covariates in different regression equations are not highly correlated. As the methodology presented in this paper is a Bayesian semi-parametric version of Zellner's model, similar gains are to be expected from our approach too, and these are investigated in a simulation study presented in Section 4.

GAMLSS, and the Bayesian analogue termed as BAMLSS (Umlauf et al., 2018), provides a general framework for the analysis of data in a very wide class of univariate distributions, utilizing flexible models for the parameters of the response distribution. The popularity of these methods stems from the fact that for most realistic problems, the assumption that the parameters are linearly dependent on the covariates, or even constant (as in homoscedastic regression), is not tenable. Applying this level of regression flexibility to multivariate response models is currently an active area of research. Smith and Kohn (2000) implemented the multivariate normal regression model with smooth additive terms in the mean function and with homoscedastic errors. Klein et al. (2015) present applications of the GAMLSS framework to bivariate regression with normal and t-distributed errors, and on Dirichlet regression. Klein and Kneib (2016) used copulas in bivariate response models, relating the parameters of the marginals and those of the dependence structure to additive predictors. Here, we focus attention to models with Gaussian errors and we develop a fully multivariate model with nonparametric models for the means, the variances and the correlation matrix, with automatic variable selection based on spike-slab priors.

The remainder of this paper is arranged as follows. Section 2 develops the proposed model in more detail. Posterior sampling is discussed in Section 3. Section 4 presents results from a simulation study that examines the efficiency gains one may have when fitting multivariate models instead of univariate ones. We also look into the performance of the method in automatically choosing the appropriate level of function complexity. Lastly, the simulation study reports the run-times needed to fit the described models. In Section 5 we present two applications of the model. In the first application, the objective is to understand how the human body responds to a drug overdose. This is a common setting, where there are multiple

outcomes each depending on a number of covariates. The statistical task is to understand how responses and covariates relate to each other. The second application is taken from the social sciences. Statistically, the problem can be seen as graphical modelling, where the conditional independence properties of the inverse covariance matrix are combined with flexible regression modelling. The paper concludes with a brief discussion. All the methods we used in this paper are freely available in the R package BNSP (Papageorgiou, 2019).

2 Multivariate response model

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})^\top$ denote a p-dimensional response vector and $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ denote covariate vectors, observed on the ith sampling unit, $i = 1, \ldots, n$. Below we detail how the mean and covariance matrix of $\boldsymbol{y}_i$ are modelled in terms of covariates. Here $\boldsymbol{x}_i$ denotes the vector of covariates for the mean and $\boldsymbol{z}_i$ that for the covariance. Hence, we do not assume the covariates for the mean are necessarily the same as those for the covariance. Typically, $\boldsymbol{z}_i$ is a subset of $\boldsymbol{x}_i$.

The model for the mean of the jth response, $j = 1, \ldots, p$, is expressed as

$$E(Y_{ij}) = \mu_{ij} = \mu(\boldsymbol{x}_i, \boldsymbol{\beta}_j^*) = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j = (\boldsymbol{x}_i^*)^\top \boldsymbol{\beta}_j^*, \tag{1}$$

where $\boldsymbol{x}_i^* = (1, \boldsymbol{x}_i^\top)^\top$ and $\boldsymbol{\beta}_j^* = (\beta_{0j}, \boldsymbol{\beta}_j^\top)^\top$. As we detail below, the linear predictor may include parametric and nonparametric terms. Even though it may appear from (1) that all regression equations have the same set of predictors, the introduction of binary indicators for variable selection will allow each response to have its own set of covariates.

The implied model for the mean of vector $\boldsymbol{Y}_i$ is

$$E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i = \boldsymbol{\mu}(\boldsymbol{X}_i, \boldsymbol{\beta}^*) = \boldsymbol{\beta}_0 + \boldsymbol{X}_i \boldsymbol{\beta},$$

where

$$\boldsymbol{\beta}_0 = \begin{pmatrix} \beta_{01} \\ \beta_{02} \\ \vdots \\ \beta_{0p} \end{pmatrix}, \boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{x}_i^\top & \mathbf{0}^\top & \dots & \mathbf{0}^\top \\ \mathbf{0}^\top & \boldsymbol{x}_i^\top & \dots & \mathbf{0}^\top \\ \vdots & & & \\ \mathbf{0}^\top & \mathbf{0}^\top & \dots & \boldsymbol{x}_i^\top \end{bmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_p \end{pmatrix}.$$

The mean vector can also be written as $\boldsymbol{\mu}_i = \boldsymbol{X}_i^* \boldsymbol{\beta}^*$, where $\boldsymbol{X}_i^*$ and $\boldsymbol{\beta}^*$ have the same structure as $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ above, but with $\boldsymbol{x}_i$ and $\boldsymbol{\beta}_j$ replaced by $\boldsymbol{x}_i^*$ and $\boldsymbol{\beta}_j^*$, $j = 1, \dots, p$.

We let $\boldsymbol{\Sigma}_i$ denote the covariance matrix of the ith response vector

$$\text{cov}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}(\boldsymbol{R}, \boldsymbol{z}_i, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2). \tag{2}$$

We factorize $\boldsymbol{\Sigma}_i = \boldsymbol{S}_i^{1/2} \boldsymbol{R} \boldsymbol{S}_i^{1/2}$ into a matrix of correlations $\boldsymbol{R}$ and a diagonal matrix of variances $\boldsymbol{S}_i = \text{diag}(\sigma_{i1}^2, \dots, \sigma_{ip}^2)$. The variances $\sigma_{ij}^2, j = 1, \dots, p$, are modelled in terms of covariates $\boldsymbol{z}_i$ using $\sigma_{ij}^2 = \sigma_j^2 \exp(\boldsymbol{z}_i^\top \boldsymbol{\alpha}_j)$, where $\boldsymbol{\alpha}_j$ is a vector of regression coefficients and σ_j^2 is a multiplicative variance term. Let $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_p^2)^\top$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \dots, \boldsymbol{\alpha}_p^\top)^\top$. Clearly, $\boldsymbol{\Sigma}_i$ depends on $\boldsymbol{R}, \boldsymbol{z}_i, \boldsymbol{\alpha}$, and $\boldsymbol{\sigma}^2$, and this is emphasised by the notation in (2).

The model specification is completed by assuming a normal distribution for the response vector

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i^* \boldsymbol{\beta}^*, \boldsymbol{\Sigma}_i), i = 1, 2, \dots, n. \tag{3}$$

Alternatively, the model can be written in the usual form

$$\boldsymbol{Y} \sim N(\boldsymbol{X}^* \boldsymbol{\beta}^*, \boldsymbol{\Sigma}), \tag{4}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \dots, \boldsymbol{Y}_n^\top)^\top$, $\boldsymbol{X}^* = [(\boldsymbol{X}_1^*)^\top, \dots, (\boldsymbol{X}_n^*)^\top]^\top$, and $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_i, i = 1, \dots, n)$.

In the following subsections we detail how the mean and covariance functions are modelled nonparametrically.

2.1 Mean model

The mean function $\mu_{ij} = \mu(\boldsymbol{x}_i, \boldsymbol{\beta}_j^*)$ takes the following general form

$$\mu_{ij} = \beta_{0j} + \sum_{k=1}^{K_1} u_{ik}\beta_{jk} + \sum_{k=K_1+1}^{K} f_{\mu,j,k}(u_{ik}), \tag{5}$$

where $u_{ik}, k = 1, \ldots, K_1$, denotes the regressors with parametrically modelled effects and $u_{ik}, k = K_1 + 1, \ldots, K$, denotes the regressors with effects that are modelled as smooth functions. Further, K denotes the total number of regressors that enter the p mean models.

When the assumption of the linearity of the effects of a covariate on the mean function is unrealistic or suspect, it can be relaxed by the use of smooth functions $f_{\mu,j,k}(.)$, as these can capture non-linear effects. They are represented using

$$f_{\mu,j,k}(u_{ik}) = \sum_{l=1}^{q_{\mu k}} \beta_{jkl}\phi_{\mu kl}(u_{ik}) = \boldsymbol{x}_{ik}^\top \boldsymbol{\beta}_{jk}, \tag{6}$$

where $\boldsymbol{x}_{ik} = (\phi_{\mu k1}(u_{ik}), \phi_{\mu k2}(u_{ik}), \ldots, \phi_{\mu k q_{\mu k}}(u_{ik}))^\top$ and $\boldsymbol{\beta}_{jk} = (\beta_{jk1}, \beta_{jk2}, \ldots, \beta_{jk q_{\mu k}})^\top$ are the vectors of basis functions and regression coefficients. In the current paper, the basis functions of choice are the radial basis functions, given by $\boldsymbol{x}_{ik} = \left(u_{ik}, |u_{ik} - \xi_{k1}|^2 \log\left(|u_{ik} - \xi_{k1}|^2\right), \ldots, |u_{ik} - \xi_{k q_{\mu k}-1}|^2 \log\left(|u_{ik} - \xi_{k q_{\mu k}-1}|^2\right)\right)^\top$, where $\xi_{k1}, \ldots, \xi_{k q_{\mu k}-1}$ are the knots.

Now, model (5) can be linearised and expressed as model (1)

$$\mu_{ij} = \beta_{0j} + \sum_{k=1}^{K_1} u_{ik}\beta_{jk} + \sum_{k=K_1+1}^{K} \boldsymbol{x}_{ik}^\top \boldsymbol{\beta}_{jk} = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \tag{7}$$

where $\boldsymbol{x}_i = (u_{i1}, \ldots, u_{iK_1}, \boldsymbol{x}_{iK_1+1}^\top, \ldots, \boldsymbol{x}_{iK}^\top)^\top$ and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK_1}, \boldsymbol{\beta}_{jK_1+1}^\top, \ldots, \boldsymbol{\beta}_{jK}^\top)^\top$.

Our general approach for representing smooth functions is to utilize a large number of basis functions. With this approach, under-fitting may be avoided. Chan et al. (2006) use the same strategy to capture covariate effects that are locally adaptive, that is, effects that vary rapidly in some parts of the covariate space and slowly in some other parts. We deal with potential over-fitting by allowing positive prior probability that the regression coefficients are exactly zero. This is achieved by the introduction of binary variables that allow coefficients to drop out of the model. These, for parametric effects, are denoted as $\gamma_{jk} = I[\beta_{jk} \neq 0], k = 1, \ldots, K_1$, and for nonparametric effects as $\gamma_{jkl} = I[\beta_{jkl} \neq 0], k = K_1 + 1, \ldots, K, l =$

$1, \ldots, q_{\mu k}$. Binary indicators are grouped in the same way as the regression coefficients $\boldsymbol{\beta}_j$ after (7), $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jK_1}, \boldsymbol{\gamma}_{jK_1+1}^\top, \ldots, \boldsymbol{\gamma}_{jK}^\top)^\top$.

Given $\boldsymbol{\gamma}_j$, model (7) is expressed as

$$\mu_{ij} = \beta_{0j} + \boldsymbol{x}_{\gamma_j i}^\top \boldsymbol{\beta}_{\gamma_j j},$$

where $\boldsymbol{\beta}_{\gamma_j j}$ consists of all non-zero elements of $\boldsymbol{\beta}_j$ and $\boldsymbol{x}_{\gamma_j i}$ of the corresponding elements of $\boldsymbol{x}_i$. Likewise, letting $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_p^\top)^\top$, the mean model implied by (3) and (4) may be expressed as $E(\boldsymbol{Y}_i) = \boldsymbol{X}_{\gamma i}^* \boldsymbol{\beta}_\gamma^*$ and $E(\boldsymbol{Y}) = \boldsymbol{X}_\gamma^* \boldsymbol{\beta}_\gamma^*$.

2.2 Covariance model

A first step in modelling the covariance matrices $\boldsymbol{\Sigma}_i$ in terms of covariates is to employ the separation strategy of Barnard et al. (2000), according to which $\boldsymbol{\Sigma}_i$ is expressed as a diagonal matrix of variances, $\boldsymbol{S}_i = \text{diag}(\sigma_{i1}^2, \ldots, \sigma_{ip}^2)$, and a correlation matrix $\boldsymbol{R}$,

$$\boldsymbol{\Sigma}_i = \boldsymbol{S}_i^{1/2} \boldsymbol{R} \boldsymbol{S}_i^{1/2}. \tag{8}$$

The next subsections consider models for the diagonal elements of $\boldsymbol{S}_i$ and for the correlation matrix $\boldsymbol{R}$.

2.2.1 Diagonal variance matrices

Modelling the diagonal matrices $\boldsymbol{S}_i$ in terms of covariates is straight forward as the only requirement on these elements is that they are nonnegative. Hence, an additive model with a log-link may be utilised

$$\log \sigma_{ij}^2 = \alpha_{0j} + \sum_{k=1}^{Q_1} v_{ik} \alpha_{jk} + \sum_{k=Q_1+1}^{Q} f_{\sigma,j,k}(v_{ik}), \tag{9}$$

where $v_{ik}, k = 1, \ldots, Q_1$, and $v_{ik}, k = Q_1 + 1, \ldots, Q$, denote covariates with parametric and nonparametric effects on the log-variance, respectively. Further, Q denotes the total number of effects that enter the p variance models. Additionally, $f_{\sigma,j,k}(.)$ are smooth functions of covariates, represented as linear combinations of $q_{\sigma k}$ radial basis functions and regression coefficients. By analogy to (6), we write $f_{\sigma,j,k}(v_{ik}) = \boldsymbol{z}_{ik}^\top \boldsymbol{\alpha}_{jk}$.

Hence, by analogy to (7), model (9) may be written as

$$\log \sigma_{ij}^2 = \alpha_{0j} + \boldsymbol{z}_i^\top \boldsymbol{\alpha}_j, \tag{10}$$

where $\boldsymbol{z}_i = (v_{i1}, \ldots, v_{iQ_1}, \boldsymbol{z}_{iQ_1+1}^\top, \ldots, \boldsymbol{z}_{iQ}^\top)^\top$ and $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jQ_1}, \boldsymbol{\alpha}_{jQ_1+1}^\top, \ldots, \boldsymbol{\alpha}_{jQ}^\top)^\top$.

Consider now vectors of indicator variables for selecting the elements of $\boldsymbol{z}_i$ that enter the jth variance regression model. In line with the indicator variables for the mean model, these are denoted by $\boldsymbol{\delta}_j = (\delta_{j1}, \ldots, \delta_{jQ_1}, \boldsymbol{\delta}_{jQ_1+1}^\top, \ldots, \boldsymbol{\delta}_{jQ}^\top)^\top$. Given $\boldsymbol{\delta}_j$, model (10) becomes

$$\log \sigma_{ij}^2 = \alpha_{0j} + \boldsymbol{z}_{\delta_j i}^\top \boldsymbol{\alpha}_{\delta_j j},$$

or equivalently

$$\sigma_{ij}^2 = \exp(\alpha_{0j}) \exp(\boldsymbol{z}_{\delta_j i}^\top \boldsymbol{\alpha}_{\delta_j j}) = \sigma_j^2 \exp(\boldsymbol{z}_{\delta_j i}^\top \boldsymbol{\alpha}_{\delta_j j}).$$

Let $\boldsymbol{\sigma}_j^2 = (\sigma_{1j}^2, \ldots, \sigma_{nj}^2)^\top$. Then, the model for $\boldsymbol{\sigma}_j^2$ can be expressed as

$$\boldsymbol{\sigma}_j^2 = \sigma_j^2 \exp(\boldsymbol{Z}_{\delta_j} \boldsymbol{\alpha}_{\delta_j j}), \tag{11}$$

where the design matrix $\boldsymbol{Z}_{\delta_j} = [\boldsymbol{z}_{\delta_j 1}, \ldots, \boldsymbol{z}_{\delta_j n}]^\top$ consists of n rows, with the ith row containing the elements of $\boldsymbol{z}_i$ that corresponds to the non-zero elements of $\boldsymbol{\delta}_j$.

2.2.2 Common correlations model

Turning our attention to the correlation matrix $\boldsymbol{R}$, the first prior model we consider, termed the 'common correlations model', takes the following form

$$f(\boldsymbol{R}|\mu_R, \sigma_R^2) = \nu(\mu_R, \sigma_R^2) \prod_{k<l} \exp\{-[g(r_{kl}) - \mu_R]^2 / 2\sigma_R^2\} J[g(r_{kl}) \to r_{kl}] I[\boldsymbol{R} \in \mathcal{C}]. \tag{12}$$

Here $\mathcal{C}$ denotes the space of correlation matrices, $I[.]$ is the indicator function that ensures the correlation matrix is positive definite and $\nu(.,.)$ is the normalizing constant

$$\nu^{-1}(\mu_R, \sigma_R^2) = \int_{\boldsymbol{R}\in\mathcal{C}} \prod_{k<l} \exp\{-[g(r_{kl}) - \mu_R]^2/2\sigma_R^2\} J[g(r_{kl}) \to r_{kl}] dr_{kl}.$$

Function $g(r)$ may be taken to be the Fisher's z transformation $g(r) = \log([1+r]/[1-r])/2$, considered within Bayesian hierarchical modelling by Daniels and Kass (1999). With this choice, $J[g(r) \to r] = (1-r)^{-1}(1+r)^{-1}$. Another choice is the identity function $g(r) = r$ that simplifies the model formulation.

Making the simplifying model choice of $g(r) = r$ and ignoring the normalizing constant, (12) reduces to

$$f(\boldsymbol{R}|\mu_R, \sigma_R^2) \propto \prod_{k<l} \exp\{-(r_{kl} - \mu_R)^2/2\sigma_R^2\} I[\boldsymbol{R} \in \mathcal{C}], \tag{13}$$

where the product is over the nonredundant, upper triangular, elements of $\boldsymbol{R}$ and the kernel is that of a normal density with mean μ_R and variance σ_R^2. Although it may appear that $\{r_{kl} : k < l\}$ are independent, this is not the case as the indicator function restricts the range of the correlations and induces dependence among them. The 'common correlations model' is intuitive and easy to interpret, however it can can be quite restrictive since all correlations are tied to a common mean μ_R and a common variance σ_R^2. For this reason, we consider two models that are more flexible, the 'grouped correlations' and 'grouped variables' models.

2.2.3 Grouped correlations model

The 'grouped correlations model' includes a clustering on the elements of $\boldsymbol{R}$, and it takes the form

$$\begin{aligned} &f(\boldsymbol{R}|\boldsymbol{\mu}_R, \sigma_R^2, \boldsymbol{\lambda}) = \nu(\boldsymbol{\mu}_R, \sigma_R^2, \boldsymbol{\lambda}) \\ &\times \prod_{k<l} \left\{ \sum_{h=1}^{H} I[\lambda_{kl} = h] \exp\{-[g(r_{kl}) - \mu_{R,h}]^2/2\sigma_R^2\} \right\} J[g(r_{kl}) \to r_{kl}] I[\boldsymbol{R} \in \mathcal{C}], \end{aligned} \tag{14}$$

where H denotes the number of correlation groups and $\mu_{R,h}$ denotes the mean of the hth group, $h = 1, \ldots, H$.

Consider, for example, the case depicted in Figure 1: a correlation matrix of a 5-dimensional response,

where the 10 nonredundant correlations are partitioned into $H = 3$ groups, namely, $A = \{r_{12}, r_{13}, r_{23}\}$, $B = \{r_{14}, r_{15}, r_{24}, r_{25}, r_{34}, r_{35}\}$, and the singleton group $C = \{r_{45}\}$, where each group has its own mean. Making the same simplifying choices as those that gave rise to (13), prior (14) for the current scenario can be written as

$$f(\boldsymbol{R}|\boldsymbol{\mu}_R, \sigma_R^2, \boldsymbol{\lambda}) \propto \prod_{r_{kl} \in A} \exp\{-(r_{kl} - \mu_{R,A})^2/2\sigma_R^2\}$$
$$\times \prod_{r_{kl} \in B} \exp\{-(r_{kl} - \mu_{R,B})^2/2\sigma_R^2\} \prod_{r_{kl} \in C} \exp\{-(r_{kl} - \mu_{R,C})^2/2\sigma_R^2\} I[\boldsymbol{R} \in \mathcal{C}]$$

2.2.4 Grouped variables model

The 'grouped variables model' is another clustering model that clusters the variables instead of the correlations. The prior takes the form

$$f(\boldsymbol{R}|\boldsymbol{\mu}_R, \sigma_R^2, \boldsymbol{\lambda}) = \nu(\boldsymbol{\mu}_R, \sigma_R^2, \boldsymbol{\lambda})$$
$$\times \prod_{k<l} \left\{ \sum_{h_1,h_2=1}^{G} I[\lambda_k = h_1] I[\lambda_l = h_2] \exp\{-[g(r_{kl}) - \mu_{R,h_1,h_2}]^2/2\sigma_R^2\} \right\} J[g(r_{kl}) \to r_{kl}] I[\boldsymbol{R} \in \mathcal{C}],$$

where G is the number of groups in which the variables are partitioned, creating $H = G(G+1)/2$ clusters for the correlations.

A clustering on the variables is more structured than a clustering on the correlations. In other words, a clustering on the variables implies a clustering on the correlations. The converse, however, is not necessarily true. Revisiting Figure 1, we see that the $p = 5$ responses are grouped into two clusters, the first group consisting of variables $\{1, 2, 3\}$, and the second one of variables $\{4, 5\}$. These two groups create three groups of correlations, two of which describe the correlations within each group and one that describes the correlation between the two groups.

2.3 Prior specification

Let $\tilde{\boldsymbol{X}} = \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{X}^*$. The prior for $\boldsymbol{\beta}_\gamma^*$ is specified as (Zellner, 1986)

$$\boldsymbol{\beta}_\gamma^* | c_\beta, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{R} \sim N(\boldsymbol{0}, c_\beta (\tilde{\boldsymbol{X}}_\gamma^\top \tilde{\boldsymbol{X}}_\gamma)^{-1}). \tag{15}$$

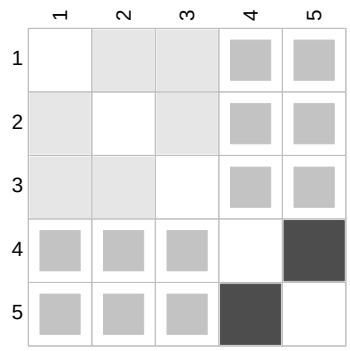

Figure 1: A 5×5 correlation matrix with two groups of variables, $\{1, 2, 3\}$ and $\{4, 5\}$, and three groups of correlations, denoted by different colours.

Further, the prior for c_β is specified as inverse Gamma, $c_\beta \sim \text{IG}(a_\beta, b_\beta)$.

For the vector $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jK_1}, \boldsymbol{\gamma}_{jK_1+1}^\top, \ldots, \boldsymbol{\gamma}_{jK}^\top)^\top, j = 1, \ldots, p$, of indicator variables, we specify independent binomial priors for each of its K subvectors,

$$P(\boldsymbol{\gamma}_{jk}|\pi_{\mu jk}) = \pi_{\mu jk}^{N(\gamma_{jk})}(1 - \pi_{\mu jk})^{q_{\mu k} - N(\gamma_{jk})}, k = 1, \ldots, K,$$

where $N(\gamma_{jk}) = \gamma_{jk}$ for parametric effects, $k = 1, \ldots, K_1$, and $N(\gamma_{jk}) = \sum_{l=1}^{q_{\mu k}} \gamma_{jkl}$ for nonparametric effects, $k = K_1 + 1, \ldots, K$. We work with Beta priors for $\pi_{\mu jk}$, $\pi_{\mu jk} \sim \text{Beta}(c_{\mu jk}, d_{\mu jk})$, $j = 1, \ldots, p, k = 1, \ldots, K$, although sparsity inducing, zero-inflated Beta priors, are also an attractive option.

Continuing with the priors on the covariance parameters, we specify independent normal priors for $\boldsymbol{\alpha}_{\delta_j j}$

$$\boldsymbol{\alpha}_{\delta_j j}|c_{\alpha j}, \boldsymbol{\delta}_j \sim N(\mathbf{0}, c_{\alpha j}\boldsymbol{I}), j = 1, \ldots, p.$$

Further, the priors we consider for $c_{\alpha j}$, are the half-normal, $\sqrt{c_{\alpha j}} \sim \text{HN}(\phi^2_{c_{\alpha j}}) \equiv N(0, \phi^2_{c_{\alpha j}})I[\sqrt{c_{\alpha j}} > 0]$, and the inverse Gamma, $c_{\alpha j} \sim \text{IG}(a_{\alpha j}, b_{\alpha j})$, $j = 1, \ldots, p$.

For the Q subvectors of $\boldsymbol{\delta}_j = (\delta_{j1}, \ldots, \delta_{jQ_1}, \boldsymbol{\delta}_{jQ_1+1}^\top, \ldots, \boldsymbol{\delta}_{jQ}^\top)^\top, j = 1, \ldots, p$, we specify independent binomial priors

$$P(\boldsymbol{\delta}_{jk}|\pi_{\sigma jk}) = \pi_{\sigma jk}^{N(\delta_{jk})}(1 - \pi_{\sigma jk})^{q_{\sigma k} - N(\delta_{jk})}, k = 1, \ldots, Q,$$

where $N(\delta_{jk}) = \delta_{jk}$ for parametric effects, $k = 1, \ldots, Q_1$, and $N(\delta_{jk}) = \sum_{k=1}^{q_{\sigma k}} \delta_{jkl}$ for nonparametric effects, $k = Q_1 + 1, \ldots, Q$. We specify independent Beta priors for $\pi_{\sigma jk}$, $\pi_{\sigma jk} \sim \text{Beta}(c_{\sigma jk}, d_{\sigma jk})$, $j = 1, \ldots, p, k = 1, \ldots, Q$.

For $\sigma_j^2, j = 1, \ldots, p$, we consider inverse Gamma and half-normal priors, denoted as $\sigma_j^2 \sim \text{IG}(a_{\sigma j}, b_{\sigma j})$ and $\sigma_j \sim \text{HN}(\phi_{\sigma j}^2) \equiv N(0, \phi_{\sigma j}^2) I[\sigma_j > 0]$.

Lastly, we describe the priors on the parameters of the correlation models. Starting with the 'common correlations model' in (12), we place the following priors on its parameters

$$\mu_R \sim N(0, \varphi_R^2) \text{ and } \sigma_R \sim \text{HN}(\phi_R^2) \equiv N(\sigma_R; 0, \phi_R^2) I[\sigma_R > 0].$$

We take the 'grouped correlations model' to be arising from the 'common correlations model', by treating the prior on μ_R as another unknown model parameter. In symbols, $\mu_R \sim P$, where P is an unknown distribution. Here, we place a Dirichlet process (DP) prior on P (Ferguson, 1973). Due to the almost sure discreteness of the DP, the prior P admits the following representation

$$P(.) = \sum_{h=1}^{\infty} w_h \delta_{\mu_{R,h}}(.),$$

where $\delta_x(.)$ is an indicator function, $\delta_x(y) = I[x = y]$. The prior weights w_h are constructed utilising the so called stick-breaking process (Sethuraman, 1994). Let $v_h, h = 1, 2, \ldots,$ be independent draws from a $\text{Beta}(1, \alpha^*)$ distribution. We have, $w_1 = v_1$, for $l \geq 2$, $w_l = v_l \prod_{h=1}^{l-1}(1 - v_h)$. We take the concentration parameter α^* to be unknown and we assign to it a gamma prior $\alpha^* \sim \text{Gamma}(a_{\alpha*}, b_{\alpha*})$ with mean $a_{\alpha*}/b_{\alpha*}$. Further, $\mu_{R,h}$ are generated from the so called base distribution, here taken to be $N(0, \varphi_R^2)$.

The 'grouped correlations model' in (14) is obtained by first writing

$$\int_{\mu_R} f(\boldsymbol{R}|\mu_R, \sigma_R^2) dP(\mu_R) = \sum_{h=1}^{\infty} w_h f(\boldsymbol{R}|\mu_{R,h}, \sigma_R^2) =$$
$$\nu(\boldsymbol{\mu}_R, \sigma_R^2, \boldsymbol{w}) \sum_{h=1}^{\infty} w_h \prod_{k<l} \exp\{-[g(r_{kl}) - \mu_{R,h}]^2 / 2\sigma_R^2\} J[g(r_{kl}) \to r_{kl}] I[\boldsymbol{R} \in \mathcal{C}],$$

where $\boldsymbol{\mu}_R$ and $\boldsymbol{w}$ denote the vectors of group means and the stick-breaking weights, respectively. In practice, we truncate $P()$ to include H components. In this case, the prior weights are constructed as

before, except for the Hth one that is now constructed as $w_H = \prod_{h=1}^{H-1}(1 - v_h)$. Further, we introduce allocation variables λ_{kl} to indicate the component in which r_{kl} is assigned to, $k = 1, \ldots, p, k < l$. The stick-breaking weights provide the prior on the allocation variables: $P(\lambda_{kl} = h) = w_h, h = 1, \ldots, H$. With these observations, it is clear how model (14) follows.

The development on the 'grouped variables model' is very similar, with the clustering now performed on the variables rather than the correlations.

In the simulation study and applications that we present in Sections 4 and 5, we use the following priors, unless otherwise stated within the relevant sections. For c_β, we specify IG$(1/2, np/2)$, as a p-variate analogue of the prior of Liang et al. (2008). For all inclusion probabilities, $\pi_{\mu jk}$ and $\pi_{\sigma jk}$, we define Beta$(1, 1)$, i.e. uniform, priors. The prior on all $c_{\alpha j}$ is specified to be IG$(1.1, 1.1)$. Further, for all σ_j, we define the prior to be HN(2). In addition, we specify $\mu_R \sim N(0, 1)$ and $\sigma_R \sim$ HN(1). Lastly, the DP base distribution is taken to be the standard normal while the concentration is taken to have a $\alpha^* \sim$ Gamma$(5, 2)$ prior.

3 Posterior Sampling

To carry out posterior sampling we consider two likelihood functions and use the one that is more computationally convenient for each step of the MCMC algorithm.

We first consider the full likelihood i.e. the one that involves all model parameters. The contribution of $\boldsymbol{Y}_i, i = 1, \ldots, n$, using decomposition (8), may be expressed as

$$
\begin{aligned}
&f(\boldsymbol{Y}_i|\boldsymbol{\beta}^*, \boldsymbol{\gamma}, c_\beta, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\sigma}^2, \boldsymbol{R}) \propto |\boldsymbol{\Sigma}_i(\boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\sigma}^2)|^{-\frac{1}{2}} \exp\{-(\boldsymbol{Y}_i - \boldsymbol{X}_i^*\boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i^*\boldsymbol{\beta}^*)/2\} \\
&\propto |\boldsymbol{S}_i^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{S}_i^{\frac{1}{2}}|^{-\frac{1}{2}} \exp\{-(\boldsymbol{S}_i^{-\frac{1}{2}}\boldsymbol{r}_i)^\top \boldsymbol{R}^{-1}(\boldsymbol{S}_i^{-\frac{1}{2}}\boldsymbol{r}_i)/2\} \propto |\boldsymbol{S}_i|^{-\frac{1}{2}}|\boldsymbol{R}|^{-\frac{1}{2}} \exp\{-\text{tr}(\boldsymbol{R}^{-1}\tilde{\boldsymbol{S}}_i)/2\},
\end{aligned}
$$

where $\boldsymbol{r}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i^*\boldsymbol{\beta}^*$ and $\tilde{\boldsymbol{S}}_i = (\boldsymbol{S}_i^{-1/2}\boldsymbol{r}_i)(\boldsymbol{S}_i^{-1/2}\boldsymbol{r}_i)^\top$. Hence, the likelihood function, based on all observations, is

$$
f(\boldsymbol{Y}|\boldsymbol{\beta}^*, \boldsymbol{\gamma}, c_\beta, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\sigma}^2, \boldsymbol{R}) \propto \prod_{i=1}^{n} |\boldsymbol{S}_i|^{-\frac{1}{2}}|\boldsymbol{R}|^{-\frac{n}{2}} \exp\{-\text{tr}(\boldsymbol{R}^{-1}\tilde{\boldsymbol{S}}_i)/2\} \tag{16}
$$

where $\tilde{\boldsymbol{S}} = \sum_{i=1}^{n} \tilde{\boldsymbol{S}}_i$.

To improve mixing of the MCMC algorithm, we can integrate out vector $\boldsymbol{\beta}^*$ from the likelihood (16), to obtain

$$f(\boldsymbol{Y}|\boldsymbol{\gamma}, c_\beta, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\sigma}^2, \boldsymbol{R}) = (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}(\boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\sigma}^2)|^{-\frac{1}{2}} (c_\beta + 1)^{-\frac{N(\gamma)+p}{2}} \exp(-S/2), \tag{17}$$

where

$$S = S(\boldsymbol{Y}, \boldsymbol{\gamma}, c_\beta, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\sigma}^2, \boldsymbol{R}) = \tilde{\boldsymbol{Y}}^\top \big(I - \frac{c_\beta}{1+c_\beta} \tilde{\boldsymbol{X}}_\gamma (\tilde{\boldsymbol{X}}_\gamma^\top \tilde{\boldsymbol{X}}_\gamma)^{-1} \tilde{\boldsymbol{X}}_\gamma^\top \big) \tilde{\boldsymbol{Y}},$$

with $\tilde{\boldsymbol{Y}} = \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{Y}$ and $N(\gamma) + p$ the total number of columns in $\tilde{\boldsymbol{X}}_\gamma$. We compute S using the following, more convenient, expression

$$S = \mathrm{tr}\big(\boldsymbol{R}^{-1} \sum_{i=1}^{n} \check{\boldsymbol{y}}_i \check{\boldsymbol{y}}_i^\top\big) - c_\beta (1+c_\beta)^{-1} \big(\sum_{i=1}^{n} \check{\boldsymbol{y}}_i^\top \boldsymbol{R}^{-1} \check{\boldsymbol{X}}_{\gamma i}\big) \big(\sum_{i=1}^{n} \check{\boldsymbol{X}}_{\gamma i}^\top \boldsymbol{R}^{-1} \check{\boldsymbol{X}}_{\gamma i}\big)^{-1} \big(\sum_{i=1}^{n} \check{\boldsymbol{X}}_{\gamma i}^\top \boldsymbol{R}^{-1} \check{\boldsymbol{y}}_i\big)$$

where $\check{\boldsymbol{X}}_{\gamma i} = \boldsymbol{S}_i^{-1/2} \boldsymbol{X}_{\gamma i}^*$ and $\check{\boldsymbol{y}}_i = \boldsymbol{S}_i^{-1/2} \boldsymbol{Y}_i$.

Sampling from the posterior of the parameters of the correlation matrices poses the greatest challenge. Consider, for instance, sampling from the posterior of parameter μ_R of the 'common correlations model', given in (12), using the Metropolis-Hastings algorithm. Letting μ_R^C and μ_R^P denote current and proposed values, the acceptance probability will involve the ratio of the normalising constants $\nu(\mu_R^P, \sigma_R^2)/\nu(\mu_R^C, \sigma_R^2)$, which can be very computationally demanding to calculate. Posterior sampling, however, may be simplified by utilising the 'shadow prior' (Liechty et al., 2004). The basic idea is to introduce latent variables θ_{kl} between the correlations r_{kl} and the mean μ_R, by which prior (12) becomes

$$f(\boldsymbol{R}|\boldsymbol{\theta}, \tau^2) = \nu(\boldsymbol{\theta}, \tau^2) \prod_{k<l} \exp\{-[g(r_{kl}) - \theta_{kl}]^2 / 2\tau^2\} J[g(r_{kl}) \to r_{kl}] I[\boldsymbol{R} \in \mathcal{C}], \tag{18}$$

where

$$\nu^{-1}(\boldsymbol{\theta}, \tau^2) = \int_{\boldsymbol{R} \in \mathcal{C}} \prod_{k<l} \exp\{-[g(r_{kl}) - \theta_{kl}]^2 / 2\tau^2\} J[g(r_{kl}) \to r_{kl}] dr_{kl}.$$

Further, variables θ_{kl} are assumed to be independently distributed as

$$\theta_{kl} \sim N(\mu_R, \sigma_R^2), l = 1, \ldots, p, k < l, \tag{19}$$

and τ is taken to be a small constant. Sampling from the posterior of $\boldsymbol{\theta} = \{\theta_{kl}\}$ still involves the ratio of the normalising constants, $\nu(\boldsymbol{\theta}^P, \tau^2)/\nu(\boldsymbol{\theta}^C, \tau^2)$, but that, as was argued by Liechty et al. (2004), for small τ, can reasonably be approximated by one. In addition, now sampling for the posterior of μ_R given $\boldsymbol{\theta}$ is straight forward. Hence, the computational burden is greatly alleviated.

We now provide details on the step of the MCMC algorithm that updates $\boldsymbol{R}$. This step uses the prior in (18) and the likelihood in (16). Hence, the posterior of $\boldsymbol{R}$ is

$$f(\boldsymbol{R}|\ldots) \propto |\boldsymbol{R}|^{-\frac{n}{2}} \exp\{-\mathrm{tr}(\boldsymbol{R}^{-1}\tilde{\boldsymbol{S}})/2\} \prod_{k<l} \exp\{-[g(r_{kl}) - \theta_{kl}]^2/2\tau^2\} J[g(r_{kl}) \to r_{kl}] I[\boldsymbol{R} \in \mathcal{C}]. \tag{20}$$

To obtain a proposal density and sample from (20) we utilize the method of Zhang et al. (2006) and Liu and Daniels (2006). We start by considering a symmetric, positive definite and otherwise unconstrained matrix $\boldsymbol{E}$ in place of $\boldsymbol{R}$, assumed to have an inverse Wishart prior $\boldsymbol{E} \sim \mathrm{IW}(\zeta, \boldsymbol{\Psi})$, with mean equal to the realization of $\boldsymbol{E}$ from the previous iteration of the sampler. Given the inverse Wishart prior on $\boldsymbol{E}$, we obtain the following, easy to sample from, inverse Wishart posterior

$$g(\boldsymbol{E}|\ldots) \propto |\boldsymbol{\Psi}|^{\frac{\zeta}{2}} |\boldsymbol{E}|^{-\frac{n+\zeta+p+1}{2}} \exp\{-\mathrm{tr}[\boldsymbol{E}^{-1}(\tilde{\boldsymbol{S}} + \boldsymbol{\Psi})]/2\}. \tag{21}$$

We decompose $\boldsymbol{E} = \boldsymbol{D}^{1/2}\boldsymbol{R}\boldsymbol{D}^{1/2}$ into a diagonal matrix of variances $\boldsymbol{D} = \mathrm{diag}(d_1^2, \ldots, d_p^2)$, and a correlation matrix $\boldsymbol{R}$. The Jacobian associated with this transformation is $J(\boldsymbol{E} \to \boldsymbol{D}, \boldsymbol{R}) = \prod_{k=1}^{p} (d_k)^{p-1} = |\boldsymbol{D}|^{(p-1)/2}$. It follows that the joint density for $(\boldsymbol{D}, \boldsymbol{R})$ is

$$h(\boldsymbol{D}, \boldsymbol{R}|\ldots) \propto |\boldsymbol{\Psi}|^{\frac{\zeta}{2}} |\boldsymbol{D}|^{(p-1)/2} |\boldsymbol{E}|^{-\frac{n+\zeta+p+1}{2}} \exp\{-\mathrm{tr}[\boldsymbol{E}^{-1}(\boldsymbol{S} + \boldsymbol{\Psi})]/2\}. \tag{22}$$

Sampling from (22) at iteration $u+1$ proceeds by sampling $\boldsymbol{E}^{(u+1)}$ from (21) and decomposing $\boldsymbol{E}^{(u+1)}$

into $(\boldsymbol{D}^{(u+1)}, \boldsymbol{R}^{(u+1)})$. Further, the pair $(\boldsymbol{D}^{(u+1)}, \boldsymbol{R}^{(u+1)})$ is accepted as a sample from (20) with probability

$$\alpha = \min\left\{1, \frac{f(\boldsymbol{R}^{(u+1)}|\dots)h(\boldsymbol{D}^{(u)}, \boldsymbol{R}^{(u)}|\dots)}{f(\boldsymbol{R}^{(u)}|\dots)h(\boldsymbol{D}^{(u+1)}, \boldsymbol{R}^{(u+1)}|\dots)}\right\},$$

where, in $h(,|)$, $\boldsymbol{\Psi} = (\zeta - p - 1)\boldsymbol{E}^{(u)}$. We treat ζ as a tuning parameter and we automatically adjust it's value (Roberts and Rosenthal, 2009) so as to obtain an acceptance probability of $20\% - 25\%$ (Roberts and Rosenthal, 2001). Further details on the MCMC steps are provided in the Appendix and the supplemental materials online.

4 Simulation study

The first purpose in this simulation study is to quantify, in a simple scenario, the gains that one may have, in terms of reduced bias and variance, when estimating a posterior mean function by fitting the multivariate model of the highest available response dimension instead of a lower dimensional model. The second one is to report the run-times needed to fit models of increasing response dimension. To achieve these goals, it suffices to consider data-generating mechanisms with simple mean and variance functions. Simulation studies that illustrate the performance of the univariate version of the current model in capturing complex mean and variance functions have been presented by Chan et al. (2006) and Papageorgiou (2018), and hence will not be revisited here. Additionally, we evaluate the model's ability to select important variables with its spike-slab priors, and whether the variable selection ability depends on the dimension of the response. Lastly, we compare the performance of the models presented here with the performance of other models for sparse multivariate regression that have appeared in the literature.

The data-generating mechanism that we consider consists of ten orthogonal covariates, $x_1, \dots, x_{10}$, each generated from a uniform distribution in the $(-0.5, 0.5)$ interval, and ten responses, $Y_1, \dots, Y_{10}$, that are generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = (\beta_{01} + \beta_{11}x_1, 0, \dots, 0)^\top$ and covariance $\boldsymbol{\Sigma}(\rho)$. The first element of the mean is a linear function of x_1, while all other elements are zero. The covariance matrix is taken to have diagonal elements equal to one and all off diagonal elements equal to ρ.

The main interest here is on the quality of the estimate of the mean function of the first dimension. We examine how this quality, as measured by the posterior bias and variance, depends on the dimension

of the response, the value of the correlation coefficient ρ, and the chosen linear predictor. The effect of the dimension of the response is evaluated by fitting one-, two-, four-, six-, and ten-dimensional response models to the dataset that includes the ten responses. The effect of the correlation coefficient is examined by letting ρ take values in the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The effect of the choice of the linear predictor is evaluated by fitting mean models that include only the relevant covariate, x_1, models that include the first three covariates, x_1, x_2, x_3, and models that include all available covariates, $x_1, \ldots, x_{10}$.

For example, a four-dimensional response model considers $Y_1, \ldots, Y_4$ and ignores $Y_5, \ldots, Y_{10}$. The mean functions of the responses are modelled using one of the following three options

$$\mu_j = \beta_{0j} + \beta_{j1}x_1, \mu_j = \beta_{0j} + \sum_{k=1}^{3} \beta_{jk}x_k, \mu_j = \beta_{0j} + \sum_{k=1}^{10} \beta_{jk}x_k, \tag{23}$$

$j = 1, 2, 3, 4$, where the first specification is correct for the first response and wrong for the other three responses, while the second and third are wrong for all responses. Further, we fit models with constant variance functions σ_j^2, $j = 1, 2, 3, 4$, and the common correlations model given in (12). Both the variance and correlation model specifications are the correct ones.

The regression coefficients are taken to be $\beta_{01} = 0$ and $\beta_{11} = 3.47$. The chosen value of β_{11} achieves a signal-to-noise ratio (SNR) equal to one, where SNR is defined as SNR $= (\text{SST} - \text{SSE})/\text{SSE}$, with SST the total sum of squares SST$=\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)^2$ and SSE the error sum of squares SSE$=\sum_{i=1}^{n}(y_{i1} - \hat{y}_{i1})^2$. In addition, two values for the sample size are considered, $n = 50, 150$.

For all models we run the MCMC sampler for $40,000$ sweeps, discarding the first $20,000$ as burn in, and of the remaining $20,000$ keeping one in two. This results in $10,000$ samples for $\mu_{i1} = E(Y_{i1}|x_{i1}) = \beta_{01} + \beta_{11}x_{i1}$, obtained by replacing the regression coefficients in the chosen linear predictor by the corresponding sampled values. We recall that the choices of the linear predictor are given in (23). Our final estimate of $\mu_{i1}, i = 1, \ldots, n$, is taken to be the median of the sampled values, which we denote by $\hat{\mu}_{i1d}$, where subscript d denotes the dimension of the response, $d = 1, 2, 4, 6, 10$. We quantify uncertainty about these estimates by forming 90% credible intervals $(\hat{\mu}_{q_1,i1d}, \hat{\mu}_{q_2,i1d})$ where the end-points of these intervals are the 5% and 95% quantiles of the sampled values.

We compare the models in terms of their bias and variance in estimating μ_{i1}. As we estimate μ_{i1} for a range of x_1 values, we summarize the bias by computing the sum of squared deviations of the estimates

from the targets, $B(d) = \sum_{i=1}^{n}(\mu_{1i} - \hat{\mu}_{i1d})^2$. Further, the variance of the estimates is summarized by computing the sum of the squared lengths of the credible intervals, $V(d) = \sum_{i=1}^{n}(\hat{\mu}_{q_1,i1d} - \hat{\mu}_{q_2,i1d})^2$. To obtain representative results and independent of the generated dataset, we repeat the above process on 40 replicate datasets for each sample size n by correlation ρ combination.

Results for the first choice of the mean model, $\mu_j = \beta_{0j} + \beta_{j1}x_1$, are presented in Tables 1 and 2. Table 1 compares models by reporting the ratio $B(d)/B(1)$ (as a percentage), that we refer to as the relative bias, while Table 2 compares models by reporting the ratio $V(d)/V(1)$, that we refer to as the relative variance, $d = 2, 4, 6, 10$. In Table 1 we see a clear decreasing trend of the relative bias as the correlation between the responses increases. Although the gains are low when the correlation between the responses is low, we observe a rapid decrease in the relative bias as the correlation increases, for all sample sizes n and for all d. We also observe that for $d = 4$, relative bias is lower than for $d = 2$, especially for $n = 50$ and for correlations higher than 0.1. However, relative bias for $d = 6$ and $d = 10$ is very similar to that for $d = 4$. Similar patterns are observed for the relative variances in Table 2. There is a clear decreasing trend as the correlation increases, for all sample sizes and all dimensions. This decrease is more pronounced for high correlations between the responses, as one would expect given the results of Zellner (1962). Results for the second and third mean models, as displayed in (23), are available in the supplement. Generally, the patterns of relative bias and variance are the same as those seen above, however, the gains are generally more pronounced.

It is always useful to compare new methods, such as the one presented here, with methods that have appeared in the literature. Here we make comparisons, in terms of posterior bias, with the method for multivariate regression of Rothman et al. (2010), that has been implemented in the R package MRCE (Rothman, 2017). Rothman et al. (2010) present a method for sparse multivariate regression with covariance estimation (henceforth abbreviated as MRCE) that estimates regression coefficients by maximizing a multivariate normal likelihood with lasso penalties for the regression coefficients and the elements of the precision matrix. To each of the simulated datasets we fit MRCE models of response dimension $d = 2, 4, 6, 10$ and compute the total bias, $B_M(d) = \sum_{i=1}^{n}(\mu_{1i} - \tilde{\mu}_{i1d})^2$, where $\tilde{\mu}_{i1d}$ is either $\tilde{\mu}_{i1d} = \tilde{\beta}_{01d} + \sum_{k=1}^{3}\tilde{\beta}_{1kd}x_{ik}$ or $\tilde{\mu}_{i1d} = \tilde{\beta}_{01d} + \sum_{k=1}^{10}\tilde{\beta}_{1kd}x_{ik}$, depending on the mean model choice, and $\tilde{\beta}_{01d}, \ldots, \tilde{\beta}_{1,3d}, \ldots, \tilde{\beta}_{1,10d}$ are the MRCE coefficient estimates for the second and third mean models. We note that comparisons are based

Table 1: Simulation study results: the entries of the table are the relative biases $B(d)/B(1), d = 2, 4, 6, 10$, expressed as percentages. Rows refer to the sample size $n = 50, 150$, and columns to the correlation between the responses, $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$. Results are based on the first mean model, and 40 replicate datasets per sample size by correlation combination.

$d = 2$	0.1	0.3	0.5	0.7	0.9	$d = 4$	0.1	0.3	0.5	0.7	0.9
50	94.41	89.19	82.81	69.72	53.22	50	93.83	81.73	69.97	59.49	50.17
150	99.42	96.71	93.41	88.14	79.25	150	99.35	97.36	92.99	86.91	78.25
$d = 6$	0.1	0.3	0.5	0.7	0.9	$d = 10$	0.1	0.3	0.5	0.7	0.9
50	97.82	83.28	69.96	58.47	49.67	50	103.06	81.67	70.86	58.08	49.73
150	99.41	97.00	92.01	85.33	77.17	150	97.49	96.12	91.39	85.30	77.36

Table 2: Simulation study results: the entries of the table are the relative variances $V(d)/V(1), d = 2, 4, 6, 10$, expressed as percentages. Rows refer to the sample size $n = 50, 150$, and columns to the correlation between the responses, $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$. Results are based on the first mean model, and 40 replicate datasets per sample size by correlation combination.

$d = 2$	0.1	0.3	0.5	0.7	0.9	$d = 4$	0.1	0.3	0.5	0.7	0.9
50	101.06	98.32	92.09	82.99	68.22	50	100.55	94.00	85.39	74.03	60.29
150	102.15	97.77	90.27	78.80	62.66	150	100.62	93.61	83.91	72.03	58.36
$d = 6$	0.1	0.3	0.5	0.7	0.9	$d = 10$	0.1	0.3	0.5	0.7	0.9
50	97.93	91.59	82.38	75.41	65.03	50	95.99	88.28	81.78	73.69	65.30
150	100.34	92.14	82.34	70.71	58.14	150	98.32	89.23	79.15	68.92	59.04

on the second and third the mean models, but not the first one, as both approaches have a mechanism for inducing sparseness. Results, in the form of ratios $B(d)/B_M(d), d = 2, 4, 6, 10$, expressed as percentages, for the second mean model, the two sample sizes and the five correlation coefficients are presented in Table 3. We see that all entries are well below 100%, with the minimum at 42.29% and the maximum at 72.28%. Results for the third mean model are available in the supplement, and they are generally more pronounced than those of Table 3, ranging from 27.44% to 61.39%. A major advantage, however, of the MRCE approach is that the resulting algorithm is less computationally intensive than the MCMC sampler and thus model fitting is typically very fast.

To evaluate the variable selection performance of the proposed model, and to check its possible dependence on the response dimension, the correlation coefficient, the sample size, and the mean model choice, we compute the posterior probabilities that at least one of the irrelevant regressors, x_2, x_3, or $x_2, \ldots, x_{10}$, is included in the mean model of the first response. Results, for the second mean model choice, expressed as percentages, are displayed in Table 4. We note that this evaluation depends on the choice of the prior

Table 3: Simulation study results: the entries of the table are the relative biases $B(d)/B_M(d), d = 2, 4, 6, 10$, expressed as percentages. Rows refer to the sample size $n = 50, 150$, and columns to the correlation between the responses, $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$. Results are based on the second mean model, and 40 replicate datasets per sample size by correlation combination.

$d = 2$	0.1	0.3	0.5	0.7	0.9	$d = 4$	0.1	0.3	0.5	0.7	0.9
50	66.22	65.19	58.12	53.17	49.68	50	61.74	55.34	47.99	50.15	53.66
150	43.24	45.74	50.00	52.91	54.16	150	42.29	49.70	53.00	53.50	61.98
$d = 6$	0.1	0.3	0.5	0.7	0.9	$d = 10$	0.1	0.3	0.5	0.7	0.9
50	53.46	51.28	55.13	57.24	60.15	50	45.62	45.44	46.41	50.38	59.00
150	42.47	47.91	49.78	51.67	62.42	150	42.96	49.09	46.92	53.02	72.28

Table 4: Simulation study results: the entries of the table are the posterior probabilities, expressed as percentages, that at least one of x_2, x_3 is included in the mean model of the first response. Rows refer to the dimension of the fitted model $d = 2, 4, 6, 10$, columns to the correlation coefficient $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$, and the two parts of the table to the two sample sizes $n = 50, 150$. Results are based on 40 replicate datasets per sample size by correlation combination.

	$n = 50$					$n = 150$				
	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
2	7.91	7.79	7.43	6.89	5.44	6.53	6.04	5.59	4.62	3.63
4	8.33	8.14	7.72	7.13	5.45	6.62	6.32	5.96	5.48	4.27
6	8.48	8.38	7.90	7.36	5.93	6.53	6.27	6.13	5.73	4.62
10	8.03	7.87	7.46	7.22	5.78	6.65	6.40	6.21	5.80	4.79

for the inclusion probabilities, described in Section 2.3. The current evaluation is based on a Beta$(1, 3)$ prior, but the inclusion probabilities can be made smaller by changing the parameters of the Beta prior in a way that decreases the prior probability of inclusion. We further note that the relevant predictor, x_1, was almost always included in the model, regardless of the choice of the Beta prior. For this reason, we do not provide results on the inclusion probabilities of this regressor. When fitting one-dimensional models, the irrelevant regressors were included 8.26% of the time when $n = 50$, and 5.17% of the time when $n = 150$. From this observation and from Table 4, it is clear that the probabilities of inclusion decrease as the sample size increases. From Table 4, it is also clear that, for fixed dimension d, the probabilities decrease as the correlation coefficient increases. There is no clear pattern between inclusion probabilities and the dimension d of the fitted model. Results for the third mean model choice are available in the supplement. They follow the same pattern as the results of Table 4, with the probabilities being, as expected, a bit higher.

Table 5: Simulation study results: the entries of the table are the run-times, measured in seconds, required to obtain 40,000 posterior samples. Rows refer to the sample size $n = 50, 150$, columns to the dimension of the fitted model $d = 1, 2, 4, 6, 10$, and the three parts of the table to the three mean model specifications. Results are based on 40 replicate datasets per sample size by correlation combination.

	$\mu_j = \beta_{0j} + \beta_{j1}x_1$					$\mu_j = \beta_{0j} + \sum_{k=1}^{3} \beta_{jk}x_k$				
	1	2	4	6	10	1	2	4	6	10
50	1.64	10.42	25.74	59.77	253.38	3.07	15.29	44.60	112.34	526.81
150	2.96	25.49	63.02	148.93	639.28	5.36	36.64	106.06	270.72	1270.04

	$\mu_j = \beta_{0j} + \sum_{k=1}^{10} \beta_{jk}x_k$				
	1	2	4	6	10
50	9.50	37.09	147.99	466.96	2683.43
150	16.29	82.03	308.95	908.60	4890.23

We conclude this section by reporting run-times of the models. We note that the MCMC sampler has been implemented in the C programming language and that the current simulations were run on an Intel Core i7 3.40GHz processor. Run-times are reported in Table 5. These range from 1.64 seconds for a univariate, simple linear regression model to about 81.5 minutes, or 4890 seconds, for a ten-dimensional response model. For both sample sizes, and all mean models, increasing the number of responses increases the run-time in a manner that is consistent with a cubic polynomial. Further, for all response dimensions, increasing the sample size increases the run-time linearly. This last point is not obvious from Table 5 as there are only two sample sizes, however, this was observed in other simulation studies not reported here.

5 Applications

This section describes two applications of the multivariate response model. The first application investigates how the human cardiovascular system responds to a particular kind of drug overdose. Due to the complexity of the cardiovascular system, a multivariate response measurement has been taken, thus the scientific objectives demand flexible regression models within a multivariate framework. The second application shows how the multivariate model can be used to semi-parametrically condition on additional information when fitting graphical models. We elaborate on a particularly nice example of this type of modelling described in Whittaker (2009, p.1). The data used in the first application comes from Johnson and Wichern (2014). Data for the second application comes from Whittaker (2009) who in turn cites

Mardia et al. (1979) as the original source.

5.1 Multiple response regression

The cardiovascular system of $n = 17$ patients who had overdosed on amitriptyline (used to treat headaches and depression) was measured by taking a blood pressure reading (bp, y_1) and also by recording each patients' PRQRS wave - as produced by an electrocardiogram. The PRQRS wave was broken down into two parts; the PR part (pr, y_2) and the QRS part (qrs, y_3). Hence, in this example, the number of responses is $p = 3$. Covariates include the size of the overdose that was measured in terms of the amount of the drug taken (amt), total blood plasma level (tot) and the amount of amitriptyline found inside the plasma (ami). The objective of this analysis is to obtain graphical and numerical summaries of the effects of the drug overdose, along with a quantification of the uncertainty around those summaries.

To avoid numerical instability as a result of the variables being measured on different scales, we work with centred and scaled versions of the responses. In addition, a new covariate defined as ratio = ami/tot is introduced, and the explanatory variables are taken to be centred and scaled versions of $\log(\text{amt}), \log(\text{tot}), \log(\text{ratio})$, henceforth simply refer to as amt, tot, and ratio. The specific form of the model is

$$\boldsymbol{Y}_i \sim N(\boldsymbol{\mu}(\boldsymbol{x}_i, \boldsymbol{\beta}^*), \boldsymbol{\Sigma}(\boldsymbol{R}, \boldsymbol{\sigma}^2)), i = 1, 2, \ldots, 17,$$

with the means $\boldsymbol{\mu}(\boldsymbol{x}_i, \boldsymbol{\beta}^*) = (\mu(\boldsymbol{x}_i, \boldsymbol{\beta}_1^*), \mu(\boldsymbol{x}_i, \boldsymbol{\beta}_2^*), \mu(\boldsymbol{x}_i, \boldsymbol{\beta}_3^*))^\top$ given the following shared representation

$$\mu(\boldsymbol{x}_i, \boldsymbol{\beta}_j^*) = \beta_{0j} + f_{\mu,j,1}(u_{i1}) + f_{\mu,j,2}(u_{i2}) + f_{\mu,j,3}(u_{i3}), j = 1, 2, 3,$$

where u_1, u_2 and u_3 denote the three explanatory variables. The functions $f_{\mu,j,k}$ are represented as

$$f_{\mu,j,k}(u_{ik}) = \sum_{l=1}^{6} \beta_{jkl} \phi_{\mu kl}(u_{ik}), j = 1, 2, 3, k = 1, 2, 3.$$

The same number of knots, 5, or equivalently 6 basis functions, was chosen for all three semi-parametric terms. For each function $f_{\mu,j,k}$, the same $\pi_{\mu jk} = 0.5$ prior probability for the inclusion of $\phi_{\mu kl}(.)$, $j, k = 1, 2, 3, l = 1, \ldots, 6$, was used. These decisions were motivated by not having any reason to want to build

in differing levels of functional complexity across the responses, nor across the explanatory variables.

Initial plots suggest little to no change in the variances of the response variables, although it is doubtful whether the eye or a model would be able to detect this with $n = 17$. For this reason $\boldsymbol{S}$ was taken to consist of constant terms

$$\boldsymbol{S} = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2).$$

The grouped variables prior was placed on $\boldsymbol{R}$, with the upper limit G on the number of clusters set to 3. This choice was guided by the the fact that responses pr (y_2) and qrs (y_3) are both measurements of the same biological feature (the PRQRS curve) and it would make sense for them to be similarly related to bp (y_3). By choosing G to be equal to the number of responses, we allow for the possibility that such a grouping is not supported by the data.

The MCMC sampler was run for a total $400,000$ iterations discarding the first $200,000$ as burn in and thereafter retaining every second sample. Results are displayed in Figure 2. The first row displays the fitted curves for input amt and the three responses, bp, pr, and qrs. There is some evidence of nonlinear relationships, with the corresponding 90% credible intervals being very wide, reflecting a high level of uncertainty due to the high variance in the responses and the small sample size. Figure 2, row two, plots the fitted function for covariate tot and the three responses. Again, we observe some evidence of nonlinear relationships, with very wide 90% credible intervals. Lastly, the third row plots the fitted functions for covariate ratio. These plots highlight the way in which the credible intervals adapt to data sparsity. Where there is less data, the 90% credible interval is much wider.

The posterior summaries of the correlations in $\boldsymbol{R}$ are given in Table 6. Displayed are the posterior means, standard deviations, 90% point-wise credible intervals and probabilities of being allocated in the same cluster. The credible intervals are wide, reflecting the high degree of uncertainty in the values of the residual correlations. The posterior over the clustering structure places pr with qrs 56% of the time, and places bp with pr and qrs 50% and 52% of the time, respectively.

5.2 Graphical Models

The multivariate normal allows for conditional independence results to be inferred from the structure found in the precision (inverse covariance) matrix. In particular, suppose vectors $\boldsymbol{X}_a, \boldsymbol{X}_b, \boldsymbol{X}_c$ are jointly

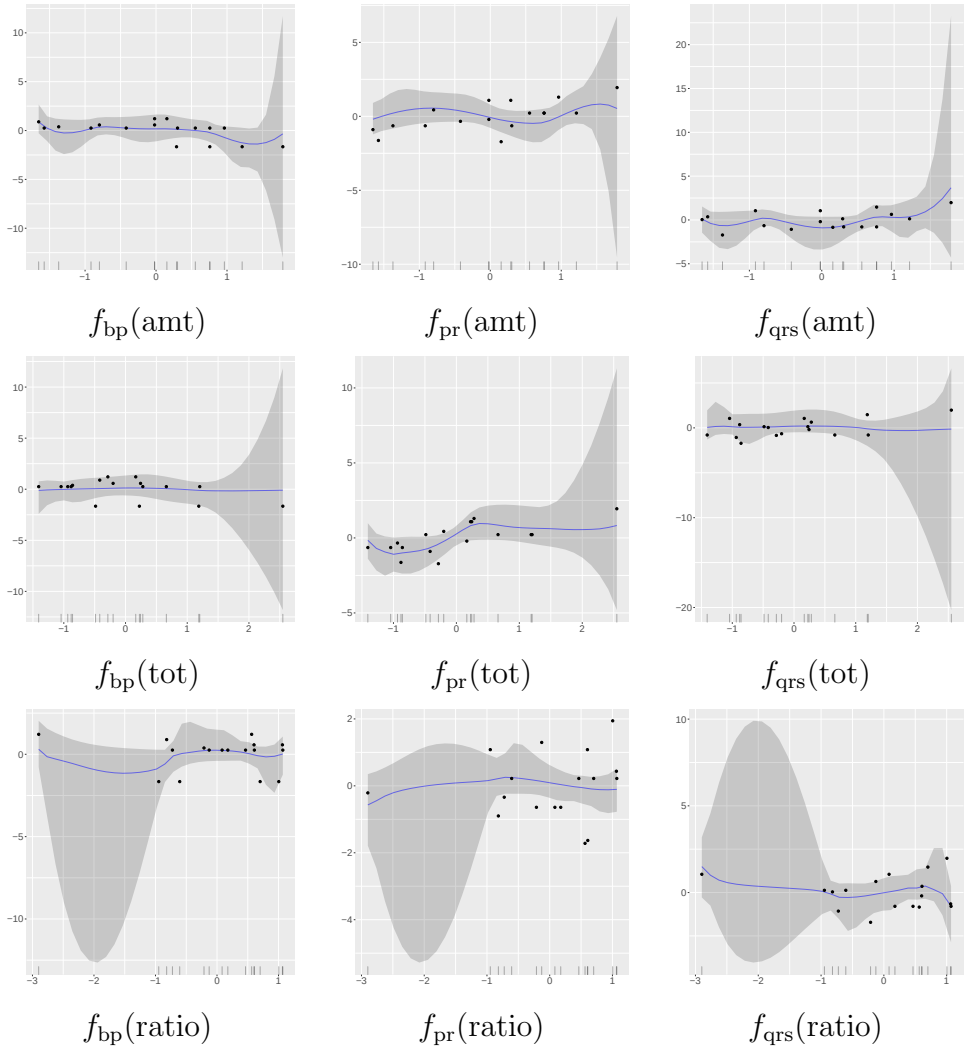

Figure 2: Results on multiple response regression: posterior means and 90% credible intervals over the nonlinear functions that enter the mean models. Rows correspond to the three covariates (amt, tot, ratio) and columns to the three responses (bp,pr,qrs).

	(pr, qrs)	(pr, bp)	(qrs, bp)
mean	0.18	-0.31	-0.61
sd	0.37	0.37	0.36
5%	-0.42	-0.87	-0.95
95%	0.77	0.26	0.17
cluster	0.56	0.50	0.52

Table 6: Results on multiple response regression: posterior correlation summary. Rows correspond to the posterior mean, standard deviation, 5% and 95% quantiles, and the probabilities of being allocated in the same cluster. Columns correspond to response variable pairs.

normal. It follows that $\boldsymbol{X}_a$ is independent of $\boldsymbol{X}_b$ given $\boldsymbol{X}_c$, if and only if, the (two identical) blocks of precision parameters relating $\boldsymbol{X}_a$ with $\boldsymbol{X}_b$ are all zero. This relation between conditional independence and the precision matrix is proven by considering how the multivariate normal density factorises when the precision matrix contains blocks of zeros.

Whittaker (2009) presents an application of this technique. The data consist of scores on $p = 5$ tests given to $n = 88$ school children. The tests are Mechanics (M), Statistics (S), Vectors (V), Analysis (An) and Algebra (Al). Matrices (a), (b) and (c) in Table 7 contain the empirical correlation matrix, scaled negative precision matrix and the suggested independence structure. The independence structure was arrived at by setting to zero all precision terms smaller in absolute value than $\alpha = 0.1$. The same inference would be made for $0.08 < \alpha < 0.23$. The interpretation of this structure is that test results on M and V are independent of results on An and S given results on Al.

Putting aside worries about how to choose a threshold value α in some principled way, we might also wish to explicitly condition on additional information about the school children. If the variables describing this additional information are not normally distributed, then they cannot be added directly into the graphical model. The model presented in this paper allows a solution to this problem. We demonstrate this methodology by explicitly conditioning on Al and repeating the above analysis on the reduced 4×4 correlation matrix describing the associations between the remaining test results. The analysis described previously suggests we ought to find that there is near zero precision term between the pairs (M, An), (M, S), (V, An) and (V, S).

The model we fit takes the form of

$$\boldsymbol{Y}_i \sim N(\boldsymbol{\mu}(\boldsymbol{x}_i, \boldsymbol{\beta}^*), \boldsymbol{\Sigma}_i), i = 1, 2, \ldots, 88.$$

Here $\boldsymbol{Y}_i \in \mathbb{R}^4$ is a vector containing the scores (M, V, An, S) for the ith child. The mean vector, $\boldsymbol{\mu}(\boldsymbol{x}_i, \boldsymbol{\beta}^*)$, is a function of the single explanatory variable Al:

$$\mu(\boldsymbol{x}_i, \boldsymbol{\beta}_j^*) = \beta_{0j} + f_{\mu,j}(u_i), j = 1, 2, 3, 4,$$

where $u_i = \text{Al}_i$ is the Algebra test score. In this example, there is sufficient data to warrant allowing the

variances to vary smoothly with Al. We chose a structure that mirrors the mean model:

$$\log \sigma_{ij}^2 = \alpha_{0j} + f_{\sigma,j}(u_i), j = 1, 2, 3, 4.$$

To complete the specification, a prior needs to be placed over $\boldsymbol{R} \in \mathbb{R}^{4\times4}$. In light of the objectives of this analysis, and motivated by the results obtained previously, we apply the grouped variables prior, with $G = 4$, expecting to find two groups: (M, V) and (An, S).

The MCMC sampler was run for 400,000 iterations, discarding 100,000 as burn in and thereafter retaining every second sample. Figure 3 presents the estimated functions and 90% credible intervals. There is evidence of non-linear dependency of the means on Al. The credible intervals are much tighter in this example, reflecting the larger sample size. The credible intervals can also be seen to adapt to the amount of available data.

The posterior probabilities that the elements of the precision matrix exceed the threshold $\alpha = 0.1$ are displayed in Table 7, matrix (d). These are estimated by inverting and scaling every sampled correlation matrix $\boldsymbol{R}$, and counting the number of times its elements exceed α. The results do conform to a large extent to what was expected. The precision term relating V and M is almost certainly larger than α, with posterior probability essentially one. Likewise, the term relating An with S is greater in magnitude than α with probability 0.98. On the other hand, the terms relating the pairs (An, S) and (M, V) all have posterior probabilities of exceeding α far below one. Interestingly, there is still a 0.61 chance that An and V are dependent, even after conditioning on Al, thus displaying the utility of being able to check the assumptions behind a graphical model, by explicitly conditioning - in a semiparametric way - on part of the response vector.

6 Discussion

The article describes a framework for the analysis of multivariate normal responses, with nonparametric models for the means, the variances and the correlation matrix. By utilizing spike-slab priors, the described framework allows covariates that enter the mean and variance functions to automatically drop out of the model. This automatic variable selection can be of great importance when one has to deal with high

	M	V	Al	An	S
M	1.00				
V	0.55	1.00			
Al	0.55	0.61	1.00		
An	0.41	0.49	0.71	1.00	
S	0.39	0.44	0.66	0.61	1.00

(a) Sample correlation matrix

	M	V	Al	An	S
M					
V	0.33				
Al	0.23	0.28			
An	0.00	0.08	0.43		
S	0.03	0.02	0.36	0.25	

(b) Negative scaled precision matrix

	M	V	Al	An	S
M	1				
V	1	1			
Al	1	1	1		
An	0	0	1	1	
S	0	0	1	1	1

(c) Independence structure

	M	V	An	S
M	1.00			
V	1.00	1.00		
An	0.35	0.61	1.00	
S	0.22	0.21	0.98	1.00

(d) Independence structure conditioning on Al

Table 7: Results on the graphical modelling application: matrices (a), (b) and (c) are based directly on the analysis given in Whittaker (2009). The matrix in (a) is a covariance-correlation matrix with variances on the diagonal, and covariances and correlations on the lower and upper triangles. Matrices (b) and (c) show the scaled negative precision matrix and the suggested independence structure. Matrix (d) contains the posterior probabilities that the elements of the scaled negative precision matrix are greater than $\alpha = 0.1$ in absolute value, after conditioning on Al.

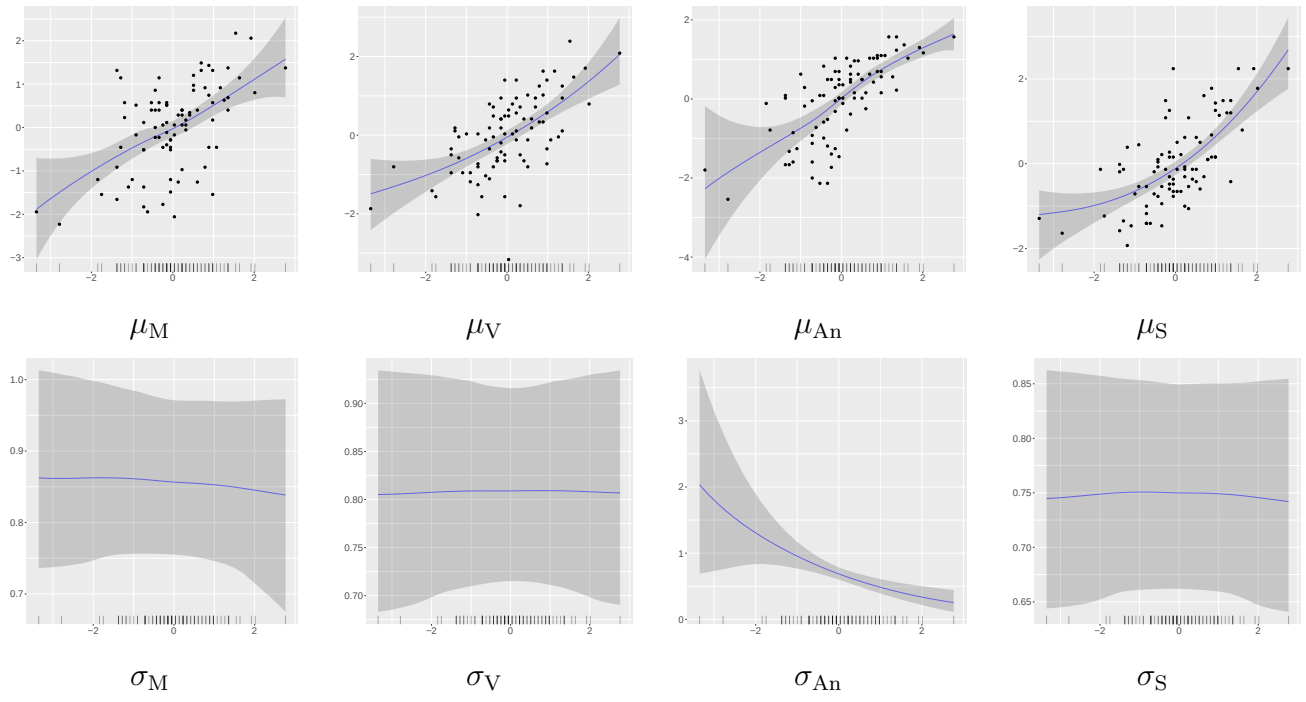

Figure 3: Results on the graphical modelling application: posterior means and 90% credible intervals of the mean (first row) and standard deviation (second row) functions of the four response variables (M, V, An, S) plotted in the four columns.

dimensional datasets.

Our framework builds on the intuitive separation strategy that factorizes the covariance matrix into a diagonal matrix of variances and a correlation matrix. We have described parametric and nonparametric models for the correlation matrix, based on normal and DP mixtures of normals for the (transformed) elements of the correlation matrix. Even though we emphasised DP mixtures in the applications we presented, this certainly is not the only choice. In fact, since the models are intuitive and easy to understand, it is easy for practitioners to incorporate prior knowledge about the correlation structure into the model. In a simulation study we illustrated the efficiency gains that one may have when fitting a multivariate models. Hence, the method can be useful in practice, since multiple responses naturally arise in many applications.

Scheipl et al. (2012) present a different flavour of spike-slab priors for function selection in univariate structured additive regression models. Their model can include varying coefficient terms, smooth interactions between covariates, spatial effects and cluster-specific random effects. Allowing for such diverse effects within a multivariate setting is certainly worth pursuing as it would increase the practical utility of the methods presented here.

7 Appendix: MCMC algorithm

At the first step of our sampler, we update the elements of $\boldsymbol{\gamma}_{jk}, j = 1, \ldots, p, k = 1, \ldots, K,$. This is done as suggested by Chan et al. (2006), hence details are omitted, but are available in the supplement.

At the second step, pairs $(\boldsymbol{\delta}_{jk}, \boldsymbol{\alpha}_{jk}), j = 1, \ldots, p, k = 1, \ldots, Q$, are updated simultaneously. Again, this is done as in Chan et al. (2006), who built on the work of Gamerman (1997), but with the introduction of a free parameter that we select adaptively (Roberts and Rosenthal, 2009) in order to achieve an acceptance probability of $20\% - 25\%$ (Roberts and Rosenthal, 2001).

The full conditional of $\sigma_j^2, j = 1, \ldots, p$, is given by

$$f(\sigma_j^2 | \ldots) \propto |\boldsymbol{\Sigma}(\boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\sigma}^2)|^{-\frac{1}{2}} \exp(-S/2)\xi(\sigma_j^2),$$

where $\xi(\sigma_j^2)$ denotes either the IG or half-normal prior. To sample from the above, we follow a random

walk algorithm.

The full conditional for parameter c_β is obtained from the marginal (17) and the IG(a_β, b_β) prior

$$f(c_\beta|\dots) \propto (c_\beta+1)^{-\frac{N(\gamma)+p}{2}} \exp(-S/2)(c_\beta)^{-a_\beta-1}\exp(-b_\beta/c_\beta).$$

To sample from the above, we utilize a normal approximation. Let $\ell(c_\beta) = \log\{f(c_\beta|\dots)\}$. We utilize a normal proposal density $N(\hat{c}_\beta, -g^2/\ell''(\hat{c}_\beta))$ where $\hat{c}_\beta$ is the mode of $\ell(c_\beta)$, found using a Newton-Raphson algorithm, $\ell''(\hat{c}_\beta)$ is the second derivative of $\ell(c_\beta)$ evaluated at the mode, and g^2 is a tuning variance parameter that we choose adaptively

Concerning parameter $c_{\alpha j}, j = 1, \dots, p$, the full conditional corresponding to the IG$(a_{\alpha j}, b_{\alpha j})$ prior is another inverse Gamma density, IG$(a_{\alpha j} + N(\delta_j)/2, b_{\alpha j} + \boldsymbol{\alpha}_{\delta_j j}^\top \boldsymbol{\alpha}_{\delta_j j}/2)$.

Further, using likelihood (16) and prior (15), we find the posterior $\boldsymbol{\beta}_\gamma^*$ to be

$$\boldsymbol{\beta}_\gamma^*|\cdots \sim N(\frac{c_\beta}{1+c_\beta}(\tilde{\boldsymbol{X}}_\gamma^\top\tilde{\boldsymbol{X}}_\gamma)^{-1}\tilde{\boldsymbol{X}}_\gamma^\top\tilde{\boldsymbol{Y}}, \frac{c_\beta}{1+c_\beta}(\tilde{\boldsymbol{X}}_\gamma^\top\tilde{\boldsymbol{X}}_\gamma)^{-1}).$$

The next step of the algorithm updates $\boldsymbol{R}$. This step has been described in the main body of the paper.

Further, to sample from the full conditional of $\boldsymbol{\theta}$, write $f(\boldsymbol{r}|\boldsymbol{\theta}, \tau^2) = \nu(\boldsymbol{\theta}, \tau^2)N(g(\boldsymbol{r}); \boldsymbol{\theta}, \tau^2\boldsymbol{I})$ for the likelihood in (18). Further, the prior for $\boldsymbol{\theta}$ is given in (19), $\boldsymbol{\theta} \sim N(\mu_R\mathbf{1}, \sigma_R^2\boldsymbol{I})$. Hence, it is easy to show that the posterior is

$$f(\boldsymbol{\theta}|\dots) = \nu(\boldsymbol{\theta}, \tau^2)N\left(\boldsymbol{\theta}; \boldsymbol{A}(\tau^{-2}g(\boldsymbol{r}) + \sigma_R^{-2}\mu_R\mathbf{1}), \boldsymbol{A} \equiv (\tau^{-2}+\sigma_R^{-2})^{-1}\boldsymbol{I}\right). \tag{24}$$

At iteration $u+1$, we sample $\boldsymbol{\theta}^{(u+1)}$ utilizing as proposal the normal distribution that appears on the right hand side of (24). The proposed $\boldsymbol{\theta}^{(u+1)}$ is accepted with probability

$$\min\left\{1, \frac{\nu(\boldsymbol{\theta}^{(u+1)}, \tau^2)}{\nu(\boldsymbol{\theta}^{(u)}, \tau^2)}\right\},$$

which, for a small value of τ^2 can reasonably be assumed to be unity (Liechty et al., 2004; Yu et al., 2014; Liechty et al., 2009).

We update μ_R from $\mu_R \sim N((d/\sigma_R^2 + 1/\varphi_r^2)^{-1}(d/\sigma_R^2)\bar{\theta}, (d/\sigma_R^2 + 1/\varphi_r^2)^{-1})$, where $\bar{\theta}$ is the mean of the

elements of vector $\boldsymbol{\theta}$.

Lastly, we update σ_R^2 utilizing the following full conditional

$$f(\sigma_R^2|\dots) \propto (\sigma_R^2)^{-\frac{d}{2}} \exp\{-\sum_{i=1}^{d}(\theta_i - \mu_R)^2/(2\sigma_R^2)\} \exp\{-\sigma_R^2/(2\phi_R^2)\} I[\sigma_R > 0].$$

Proposed values are obtained from $(\sigma_R^2)^{(p)} \sim N((\sigma_R^2)^{(c)}, f_1^2)$ where $(\sigma_R^2)^{(c)}$ denotes the current value and f_1^2 denotes a tuning parameter.

8 Supplementary Materials

Supplement : Additional tables with simulation results and a detailed MCMC sampler. (.pdf file)

R package BNSP : An R package that implements the MCMC algorithm and various functions for processing the posterior samples. The package is also available on CRAN. (.tar.gz)

Examples : Folder containing R scripts for replicating simulations and data examples.

References

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1311.

Chan, D., Kohn, R., Nott, D., and Kirby, C. (2006). Locally adaptive semiparametric estimation of the mean and variance functions in regression models. *Journal of Computational and Graphical Statistics*, 15(4):915–936.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769.

Chiu, T. Y. M., Leonard, T., and Tsui, K.-W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198–210.

Daniels, M. J. and Kass, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94(448):1254–1263.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7(1):57–68.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Johnson, R. and Wichern, D. (2014). *Applied Multivariate Statistical Analysis*. Pearson, Essex. ISBN 1292024941.

Klein, N. and Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Statistics and Computing*, 26(4):841–860.

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2015). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(4):569–591.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Liechty, J. C., Liechty, M. W., and Müller, P. (2004). Bayesian correlation estimation. *Biometrika*, 91(1):1–14.

Liechty, M. W., Liechty, J. C., and Müller, P. (2009). The shadow prior. *Journal of Computational and Graphical Statistics*, 18(2):368–383.

Liu, X. and Daniels, M. J. (2006). A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterization. *Journal of Computational and Graphical Statistics*, 15(4):897–914.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Probability and mathematical statistics. Academic Press, London.

Müller, P. and Mitra, R. (2013). Bayesian nonparametric inference - why and how. *Bayesian Analysis*, 8(2):269–302.

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

Papageorgiou, G. (2018). BNSP: an R package for fitting Bayesian semiparametric regression models and variable selection. *The R Journal*, 10(2):526–548.

Papageorgiou, G. (2019). *BNSP: Bayesian Non- And Semi-Parametric Model Fitting*. R package version 2.1.1.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika*, 94(4):1006–1013.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Rothman, A. J. (2017). *MRCE: Multivariate Regression with Covariance Estimation*. R package version 2.1.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532.

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

Smith, M. and Kohn, R. (2000). Nonparametric seemingly unrelated regression. *Journal of Econometrics*, 98(2):257–281.

Tsay, R. S. and Pourahmadi, M. (2017). Modelling structured correlation matrices. *Biometrika*, 104(1):237–242.

Umlauf, N., Klein, N., and Zeileis, A. (2018). BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27(3):612–627.

Whittaker, J. (2009). *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing.

Yu, Philip, L. H., Li, W. K., and Ng, F. C. (2014). Formulating hypothetical scenarios in correlation stress testing via a Bayesian framework. *The North American Journal of Economics and Finance*, 27:17–33.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. and Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. Elsevier Science Publishers.

Zhang, X., Boscardin, J. W., and Belin, T. R. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, 15(4):880–896.