Bayesian semiparametric modelling of covariance matrices for multivariate longitudinal data

Georgios Papageorgiou
Department of Economics, Mathematics and Statistics
Birkbeck, University of London
g.papageorgiou@bbk.ac.uk

Abstract

The article develops marginal models for multivariate longitudinal responses. Overall, the model consists of five regression submodels, one for the mean and four for the covariance matrix, with the latter resulting by considering various matrix decompositions. The decompositions that we employ are intuitive, easy to understand, and they do not rely on any assumptions such as the presence of an ordering among the multivariate responses. The regression submodels are semiparametric, with unknown functions represented by basis function expansions. We use spike-slap priors for the regression coefficients to achieve variable selection and function regularization, and to obtain parameter estimates that account for model uncertainty. An efficient Markov chain Monte Carlo algorithm for posterior sampling is developed. The simulation study presented investigates the the gains that one may have when considering multivariate longitudinal analyses instead of univariate ones, and whether these gains can counteract the negative effects of missing data. We apply the methods on a highly unbalanced longitudinal dataset with four responses observed over a period of 20 years.

Keywords: Cholesky decomposition; Clustering; Model averaging; Semiparametric regression; Variable selection

1 Introduction

There is a variety of systems that evolve over time and are too intricate to be appropriately characterized by a single outcome variable. Consider for example 'cognition', the topic of the data example presented later in this article. According to the American Psychological Association's Dictionary of Psychology, cognition is a term that refers to 'all forms of knowing and awareness, such as perceiving, conceiving, remembering, reasoning, judging, imagining, and problem solving'. Cognition is a complex, not directly observable, system that evolves over time. Measuring cognitive function relies on administering multiple psychometric tests, and investigating how it evolves and declines in the elderly requires observations over multiple years. Analysing datasets generated by such investigations requires flexible regression models for multivariate longitudinal responses. The main goal of this article is to develop, within a Bayesian framework, marginal models for multivariate longitudinal continuous responses, non-parametrically linking the mean vectors

and covariance matrices to the covariates. The current literature on modelling covariance matrices for multivariate longitudinal data relies on decompositions of the covariance matrices that are either difficult to interpret or that make assumptions that are not tenable. Here, we avoid all such decompositions and we only utilize those that are easy to understand and justify.

In order to set the notation, let $\boldsymbol{Y}_j$ denote a vector of p responses observed at time point t_j, $j = 1, \ldots, n$, $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top)^\top$ the np-dimensional vector of all responses, and $\boldsymbol{\Sigma}$ the covariance matrix of $\boldsymbol{Y}$. Matrix $\boldsymbol{\Sigma}$ must satisfy the positive definiteness (pd) condition, and that creates a major challenge when modelling its elements as functions of covariates. To ensure that this condition is satisfied, it is necessary to factorized $\boldsymbol{\Sigma}$ into a product of matrices. The decomposition[1]

$$\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^\top = \boldsymbol{D} \tag{1}$$

has been utilized for modelling multivariate longitudinal data by Xu and Mackenzie,[2] Kim and Zimmerman,[3] Kohli et al,[4] and Lee et al.[5] Matrices $\boldsymbol{L}$ and $\boldsymbol{D}$ have the following structures

$$\boldsymbol{L} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ -\boldsymbol{\Phi}_{21} & \boldsymbol{I} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{\Phi}_{n1} & -\boldsymbol{\Phi}_{n2} & \ldots & \boldsymbol{I} \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_2 & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{D}_n \end{bmatrix}, \tag{2}$$

where submatrices $\boldsymbol{\Phi}_{jk}$ are unconstrained, $j = 2, \ldots, n, k < j$, while submatrices $\boldsymbol{D}_j$ are pd, $j = 1, \ldots, n$, and all submatrices in $\boldsymbol{L}$ and $\boldsymbol{D}$ are of dimension $p \times p$. The following interpretation that has appeared in, among others, Pourahmadi[6] and Daniels and Pourahmadi[7] for the univariate case, and Xu and Mackenzie[2] and Kim and Zimmerman[3] for the multivariate case, provides justification for the lack of constraints on the elements of $\boldsymbol{\Phi}_{jk}$ and the pd constraint on the elements of $\boldsymbol{D}_j$. For the sake of simplicity we assume that vectors $\boldsymbol{Y}_j$ have zero mean. Let $\hat{\boldsymbol{Y}}_j$ denote the linear least squares predictor of $\boldsymbol{Y}_j$ based on its predecessors $\boldsymbol{Y}_{j-1}, \ldots, \boldsymbol{Y}_1$, and let $\boldsymbol{\epsilon}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ denote the prediction error. It can be shown that the predictor $\hat{\boldsymbol{Y}}_j$ is expressed in terms of the negatives of the submatrices of $\boldsymbol{L}$ as

$$\hat{\boldsymbol{Y}}_j = \sum_{k=1}^{j-1} \boldsymbol{\Phi}_{jk}\boldsymbol{Y}_k. \tag{3}$$

Further, the prediction error variance is given by the corresponding block of $\boldsymbol{D}$ i.e. $\text{var}(\boldsymbol{\epsilon}_j) = \boldsymbol{D}_j$. To see how factorization (1) is satisfied, let $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_n^\top)^\top$. We have that $\text{var}(\boldsymbol{\epsilon}) = \text{var}(\boldsymbol{L}\boldsymbol{Y}) = \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^\top$, and since consecutive prediction errors are uncorrelated, it follows that $\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^\top = \boldsymbol{D}$. The sets of matrices $\{\boldsymbol{\Phi}_{jk} : j = 2, \ldots, n, k < j\}$ and $\{\boldsymbol{D}_j : j = 1, \ldots, n\}$ have been termed the generalized autoregressive matrices and innovation covariance matrices of $\boldsymbol{\Sigma}$.[2]

A simple modification of (1) can provide an alternative interpretation. Since $\boldsymbol{L}^{-1}$ has the same block triangular structure as $\boldsymbol{L}$, the decomposition

$$\boldsymbol{L}^{-1}\boldsymbol{\Sigma}(\boldsymbol{L}^{-1})^\top = \boldsymbol{D}^*, \tag{4}$$

also holds, where $\boldsymbol{D}^*$ has the same structure as $\boldsymbol{D}$. This decomposition, for univariate responses, has been studied by Zhang and Leng,[8] and for multivariate ones by Feng et al.[9] With this decomposition, the submatrices of $\boldsymbol{L}$ that are below the main diagonal can be interpreted as moving average coefficient matrices in a time series context, hence, again, they are unconstrained.

In this article, we adopt the decomposition in (1) that leads to the interpretation in (3). Since matrices $\boldsymbol{\Phi}_{jk}$ are unconstrained, their elements can be modelled utilizing semiparametric regression models. These are described in Section 2.

The decompositions in (1) and (4) have been unanimous among researchers modelling multivariate longitudinal data. By contrast, there have been several proposals for modelling the innovation covariance matrices $\{\boldsymbol{D}_j : j = 1, \ldots, n\}$. These are reviewed in the next paragraph. The difficulty in modelling these matrices arises from having to satisfy the pd condition and from the absence of any natural ordering in the elements of the errors $\boldsymbol{\epsilon}_j$.

Based on the spectral decomposition and the matrix logarithmic transformation,[10] Xu and Mackenzie[2] model matrices $\boldsymbol{D}_j, j = 1, \ldots, n$, in terms of the explanatory variable 'time', t. Further, based on the modified Cholesky decomposition[6;11], i.e. the decomposition in (1) applied for $p = 1$, Kim and Zimmerman[3] model the innovation covariance matrices as functions of t. Lastly, Kohli et al[4] avoid decompositions and instead model each $\boldsymbol{D}_j^{-1}$ using linear covariance models,[12] that is, $\boldsymbol{D}_j^{-1} = \alpha_1 \boldsymbol{M}_1 + \cdots + \alpha_j \boldsymbol{M}_j$, where $\boldsymbol{M}_k$ are known pd matrices and α_k are unknown weights, $k = 1, \ldots, j$. The spectral decomposition and the modified Cholesky decomposition outside the context of longitudinal studies, lack simple statistical interpretation. Further, the linear covariance model does not express the matrices in terms of covariates and its implementation in practice can be difficult. For instance, the best performing model in the study of Kohli et al[4] does not guarantee that the estimated covariance matrix is pd.

A decomposition, however, that is statistically simple and intuitive, was proposed by Barnard et al.[13] It separates the matrices $\boldsymbol{D}_j$ into diagonal matrices of innovation variances $\boldsymbol{S}_j$ and innovation correlation matrices $\boldsymbol{R}_j$, $\boldsymbol{D}_j = \boldsymbol{S}_j^{1/2} \boldsymbol{R}_j \boldsymbol{S}_j^{1/2}$. With this decomposition it is easy to model the innovation variances in terms of covariates as the only constrained they must satisfy is the positiveness. This was the approach taken by Lee et al[5] and the approach that we take in this article too.

As was remarked by Pourahmadi,[14] of the available decompositions, the separation of variances and correlations is the least effective in satisfying the pd condition. Clearly, this condition must now be satisfied by the innovation correlation matrices $\{\boldsymbol{R}_j : j = 1, \ldots, n\}$. To overcome this condition, and model the elements of a correlation matrix $\boldsymbol{R}$ of longitudinal observations in terms of covariates, Zhang et al[15] utilize the Cholesky decomposition $\boldsymbol{R} = \boldsymbol{T}\boldsymbol{T}^\top$, parametrizing the non-zero elements of $\boldsymbol{T}$ using hyperspherical coordinates. In the applications presented by Zhang et al,[15] the elements of $\boldsymbol{R}$ are modelled in terms of covariate 'lag'. Lee et al[5] follow the same approach as Zhang et al,[15] but merely reparametrise the innovation correlation matrices without modelling their elements in terms of covariates. In the data analysis they presented, this resulted in a common, time invariant, innovation correlation matrix, even though they stressed the importance of correctly specifying the correlation structure.

Here, we take a different, more intuitive approach, by specifying a normal prior on the Fisher's z transformation of the nonredundant elements of the matrices $\boldsymbol{R}_j = \{r_{jkl}\}$,

$$\log\{(1 + r_{jkl})/(1 - r_{jkl})\}/2 \sim N(\mu_{cj}, \sigma_{cj}^2) I[\boldsymbol{R}_j \in \mathcal{C}], \tag{5}$$

where $\mathcal{C}$ denotes the space of correlation matrices and $I[.]$ the indicator function that maintains positive definiteness. In addition, the indicator function truncates the range of the correlations and induces dependence among them.[16] Due to this truncation, parameters μ_{cj} and σ_{cj}^2 are no longer interpreted as the mean and variance of the distribution. Here, we model both parameters as unknown functions of time t, $\mu_{cj} = \eta_0 + f_\mu(t_j)$ and $\log \sigma_{cj}^2 = \omega_0 + f_\sigma(t_j)$.

The model in (5) can be restrictive because it only allows for a single function μ_{cj} that is common to all correlations. However, it is conceivable that correlations evolve differently over time. Failing to specify a prior that allows for the needed flexibility can have a negative impact on the estimated correlations, especially in small samples.[16] Here, this flexibility is reached by considering mixtures of normal distributions $\log\{(1 + r_{jkl})/(1 - r_{jkl})\}/2 \sim \sum_h \pi_h N(\mu_{cjh}, \sigma_{cj}^2) I[\boldsymbol{R}_j \in \mathcal{C}]$. We refer to this model as the 'grouped correlations model'.[17] We also consider a 'grouped variables model' that clusters the variables instead of the correlations and it is more structured than the grouped correlations model.

The overall model consists of 5 regression submodels. These are the models for the mean of the response vector, the elements of the generalized autoregressive matrices, the innovation variances, and the 2 parameters of distribution of the innovation correlation matrices. In the approach presented here, each regression model involves nonparametrically modelled effects, represented utilizing basis function expansions. We enable flexible estimation by utilizing several basis functions and we implement variable selection and function regularization by utilizing spike-slab priors (see e.g. George and McCulloch[18]). The work presented in this article builds upon the work of Chan et al[19] and Papageorgiou[20] who describe methods for univariate response regression with nonparametric models for the mean and variance, and the work of Papageorgiou and Marshall[21] who presented methods for multivariate response regression with nonparametric models for the means, the variances and the correlation matrix.

We develop an efficient stochastic search variable selection algorithm by using Zellner's g-prior[22] that allows integrating out the regression coefficients in both the mean function of the responses $E(\boldsymbol{Y})$ and the parameter μ_{cj} (or μ_{cjh}) of the innovation correlation matrices. In addition, the Markov chain Monte Carlo (MCMC) algorithm generates the variable selection indicators in blocks[19;20] and selects the MCMC tuning parameters adaptively.[23]

The remainder of this article is arranged as follows. Section 2 describes the model in detail and Section 3 describes the main elements of the MCMC algorithm. In Section 4 we present results from a simulation study that investigates the gains that one may have, in terms of reduced posterior mean squared error (MSE), when fitting multivariate longitudinal models instead of univariate ones. Additionally, we examine the effects that missing data have on the posterior MSE and whether adding more than one response to the model can counteract these effects. Section 5 applies the methods on a highly unbalanced dataset on cognitive function and depressive symptomatology, with 4 responses observed over of period of 20 years. Section 6 concludes the paper with a brief discussion. All the methods described in this article are freely available in the R package BNSP.[24]

2 Multivariate longitudinal response model

Let $\boldsymbol{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijp})^\top$ denote the vector of p responses observed on individual $i, i = 1, \ldots, n$, at time point t_{ij}, $j = 1, \ldots, n_i$. Here, we allow the observational time points t_{ij} to be unequally spaced. We let T denote the ordered set of all unique observational times, $T = \{t_1, \ldots, t_M\}$, and we denote its cardinality by M. Further, let $\boldsymbol{u}_{ij}$ denote the covariate vector that is observed along with $\boldsymbol{Y}_{ij}$ and that may include time, other time-dependent covariates and time-independent ones. In addition, let $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}^\top, \ldots, \boldsymbol{Y}_{in_i}^\top)^\top$ denote the ith response vector. With $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i)$ and $\boldsymbol{\Sigma}_i = \mathrm{cov}(\boldsymbol{Y}_i)$, the overall model takes the form

$$\boldsymbol{Y}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, 2, \ldots, n.$$

In the following subsections we detail how the means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ are modelled semiparametrically in terms of covariates.

2.1 Mean model

The means $\mu_{ijk} = E(Y_{ijk}), k = 1, \ldots, p$, are modelled utilizing semiparametric regression methods

$$\mu_{ijk} = \beta_{k0} + \sum_{l=1}^{K_1} u_{ijl}\beta_{kl} + \sum_{l=K_1+1}^{K} f_{\mu,k,l}(u_{ijl}), \tag{6}$$

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

Unknown functions are represented utilizing basis function expansions

$$f_{\mu,k,l}(u_{ijl}) = \sum_{s=1}^{q_{\mu l}} \beta_{kls}\kappa_{\mu ls}(u_{ijl}) = \boldsymbol{x}_{ijl}^{\top}\boldsymbol{\beta}_{kl}, \tag{7}$$

where $q_{\mu l}$ is fixed and it represents the maximum number of basis functions that can be used in modelling $f_{\mu,k,l}(.)$. Further, $\boldsymbol{x}_{ijl} = (\kappa_{\mu l1}(u_{ijl}), \ldots, \kappa_{\mu l q_{\mu l}}(u_{ijl}))^{\top}$ is the vector of basis functions and $\boldsymbol{\beta}_{kl} = (\beta_{kl1}, \ldots, \beta_{klq_{\mu l}})^{\top}$ is the vector of the corresponding coefficients. Here, the basis functions of choice are the radial basis functions, given by $\kappa_1(u) = u, \kappa_2(u) = |u - \xi_1|^2 \log(|u - \xi_1|^2), \ldots, \kappa_{q-1}(u) = |u - \xi_{q-1}|^2 \log(|u - \xi_{q-1}|^2)$, where $\xi_1, \ldots, \xi_{q-1}$ are fixed knots.

Now, model (6) can be linearised as

$$\mu_{ijk} = \beta_{k0} + \sum_{l=1}^{K_1} u_{ijl}\beta_{kl} + \sum_{l=K_1+1}^{K} \boldsymbol{x}_{ijl}^{\top}\boldsymbol{\beta}_{kl} = \beta_{k0} + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}_k, = (\boldsymbol{x}_{ij}^{*})^{\top}\boldsymbol{\beta}_k^{*}, \tag{8}$$

where $\boldsymbol{x}_{ij} = (u_{ij1}, \ldots, u_{ijK_1}, \boldsymbol{x}_{ij,K_1+1}^{\top}, \ldots, \boldsymbol{x}_{ijK}^{\top})^{\top}$ and $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kK_1}, \boldsymbol{\beta}_{k,K_1+1}^{\top}, \ldots, \boldsymbol{\beta}_{kK}^{\top})^{\top}$. Further, $\boldsymbol{x}_{ij}^{*} = (1, \boldsymbol{x}_{ij}^{\top})^{\top}$ and $\boldsymbol{\beta}_k^{*} = (\beta_{k0}, \boldsymbol{\beta}_k^{\top})^{\top}$.

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

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

where $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{p0})^{\top}$, $\boldsymbol{X}_{ij} = \boldsymbol{I}_p \otimes \boldsymbol{x}_{ij}^{\top}$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\top}, \ldots, \boldsymbol{\beta}_p^{\top})^{\top}$. In the final expression, $\boldsymbol{X}_{ij}^{*}$ and $\boldsymbol{\beta}^{*}$ are as $\boldsymbol{X}_{ij}$ and $\boldsymbol{\beta}$ but with $\boldsymbol{x}_{ij}$ and $\boldsymbol{\beta}_k$ replaced by $\boldsymbol{x}_{ij}^{*}$ and $\boldsymbol{\beta}_k^{*}, k = 1, \ldots, p$.

Recalling that $\boldsymbol{Y}_i$ denotes the ith response vector, we may write the mean model as $E(\boldsymbol{Y}_i) = \boldsymbol{X}_i^{*}\boldsymbol{\beta}^{*}$, where $\boldsymbol{X}_i^{*} = [(\boldsymbol{X}_{i1}^{*})^{\top}, (\boldsymbol{X}_{i2}^{*})^{\top}, \ldots, (\boldsymbol{X}_{in_i}^{*})^{\top}]^{\top}$. Alternatively, the model can be written in the usual form $E(\boldsymbol{Y}) = \boldsymbol{X}^{*}\boldsymbol{\beta}^{*}$, where $\boldsymbol{Y} = (\boldsymbol{Y}_1^{\top}, \ldots, \boldsymbol{Y}_n^{\top})^{\top}$ and $\boldsymbol{X}^{*} = [(\boldsymbol{X}_1^{*})^{\top}, \ldots, (\boldsymbol{X}_n^{*})^{\top}]^{\top}$.

We note that $\boldsymbol{x}_{ij}$ is a maximal vector covariates and it is common to the p responses. By introducing binary indicators for variable selection, we allow each response to have its own set of covariates. Vector $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kK_1}, \boldsymbol{\gamma}_{k,K_1+1}^{\top}, \ldots, \boldsymbol{\gamma}_{kK}^{\top})^{\top}$ has the same length and structure are vector $\boldsymbol{\beta}_k$, introduced just below (8), and its elements are binary indicators that select which regressors enter the mean model of the kth response. Given $\boldsymbol{\gamma}_k$, model (8) is written as

$$\mu_{ijk} = \beta_{k0} + \boldsymbol{x}_{ijk}^{\top}\boldsymbol{\beta}_{\gamma_k} = (\boldsymbol{x}_{ijk}^{*})^{\top}\boldsymbol{\beta}_{\gamma_k}^{*},$$

where $\boldsymbol{\beta}_{\gamma_k}$ consists of all non-zero elements of $\boldsymbol{\beta}_k$ and $\boldsymbol{x}_{ijk}$ of the corresponding elements of $\boldsymbol{x}_{ij}$.

Letting $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\top}, \ldots, \boldsymbol{\gamma}_p^{\top})^{\top}$ denote the vector of binary indicators and $\boldsymbol{\beta}_{\gamma}^{*} = ((\boldsymbol{\beta}_{\gamma_1}^{*})^{\top}, \ldots, (\boldsymbol{\beta}_{\gamma_k}^{*})^{\top})^{\top}$ the vector of non-zero regression coefficients for the p responses, we can write the regression model for the mean of $\boldsymbol{Y}_i$ as $E(\boldsymbol{Y}_i) = \boldsymbol{X}_{\gamma i}^{*}\boldsymbol{\beta}_{\gamma}^{*}$. Similarly, we can write $E(\boldsymbol{Y}) = \boldsymbol{X}_{\gamma}^{*}\boldsymbol{\beta}_{\gamma}^{*}$.

2.2 Covariance model

Let $\boldsymbol{\Sigma}_i$ denote the covariance matrix of $\boldsymbol{Y}_i$. As was discussed in the introduction, to model the elements of $\boldsymbol{\Sigma}_i$ in terms of covariates, we begin by considering the factorization $\boldsymbol{L}_i\boldsymbol{\Sigma}_i\boldsymbol{L}_i^{\top} = \boldsymbol{D}_i$, where matrices $\boldsymbol{L}_i$ and $\boldsymbol{D}_i$ have the form shown in (2). The next two subsections describe semiparametric models for the generalized autoregressive matrices $\{\boldsymbol{\Phi}_{ijk} : j = 2, \ldots, n_i, k < j\}$ and the innovation covariance matrices $\{\boldsymbol{D}_{ij} : j = 1, \ldots, n_i\}$.

2.3 Generalized autoregressive matrices

For ϕ_{ijklm}, the (l,m) element of $\boldsymbol{\Phi}_{ijk}$, $l,m=1,\ldots,p$, we consider the following semiparametric model

$$\phi_{ijklm}=\psi_{lm0}+\sum_{b=1}^{B_1}v_{ijkb}\psi_{lmb}+\sum_{b=B_1+1}^{B}f_{\phi,l,m,b}(v_{ijkb})=\psi_{lm0}+\boldsymbol{z}_{ijk}^{\top}\boldsymbol{\psi}_{lm}=(\boldsymbol{z}_{ijk}^{*})^{\top}\boldsymbol{\psi}_{lm}^{*}, \tag{9}$$

where the developments in (9) follow along the same lines as those for the mean model, detailed in (6) - (8). Functions $f_{\phi,l,m,b}(v_{ijkb})=\boldsymbol{z}_{ijkb}^{\top}\boldsymbol{\psi}_{lmb}$ are linearised utilizing a fixed number $q_{\phi b}$ of basis functions, $b=B_1+1,\ldots,B$. Vector $\boldsymbol{z}_{ijk}^{*}$ denotes the maximal set covariates that is common to the p^2 autoregressive coefficients. Further, vector $\boldsymbol{\psi}_{lm}^{*}=(\psi_{lm0},\psi_{lm1},\ldots,\psi_{lmB_1},\boldsymbol{\psi}_{lm,B_1+1}^{\top},\ldots,\boldsymbol{\psi}_{lmB}^{\top})^{\top}$ consists of the regression coefficients, grouped by the effect they model. Vector $\boldsymbol{\xi}_{lm}$ is the corresponding vector of binary indicators that selects which coefficients enter the model of the (l,m) autoregressive coefficient. We note that here the intercepts are subject to selection. Given $\boldsymbol{\xi}_{lk}$, model (9) is expressed as

$$\phi_{ijklm}=(\boldsymbol{z}_{ijklm}^{*})^{\top}\boldsymbol{\psi}_{\xi_{lm}}^{*}, \tag{10}$$

where $\boldsymbol{\psi}_{\xi_{lm}}^{*}$ consists of all non-zero elements of $\boldsymbol{\psi}_{lm}^{*}$ and $\boldsymbol{z}_{ijklm}^{*}$ of the corresponding elements of $\boldsymbol{z}_{ijk}^{*}$.

Let $\boldsymbol{\psi}^{*}=((\boldsymbol{\psi}_{11}^{*})^{\top},\ldots,(\boldsymbol{\psi}_{1p}^{*})^{\top},(\boldsymbol{\psi}_{21}^{*})^{\top},\ldots,(\boldsymbol{\psi}_{pp}^{*})^{\top})^{\top}$ be the vector of all regression coefficients, and $\boldsymbol{\xi}=(\boldsymbol{\xi}_{11}^{\top},\ldots,\boldsymbol{\xi}_{1p}^{\top},\boldsymbol{\xi}_{21}^{\top},\ldots,\boldsymbol{\xi}_{pp}^{\top})^{\top}$ be the vector of all binary indicators. Further, let $\boldsymbol{\psi}_{\xi}^{*}$ be the vector of all non-zero coefficients. Now, from (3) and the definition of the prediction error, we have

$$\boldsymbol{Y}_{ij}=\sum_{k=1}^{j-1}\boldsymbol{\Phi}_{ijk}\boldsymbol{Y}_{ik}+\boldsymbol{\epsilon}_{ij}=\sum_{k=1}^{j-1}[\boldsymbol{I}_p\otimes\boldsymbol{Y}_{ik}^{\top}\otimes(\boldsymbol{z}_{ijk}^{*})^{\top}]_{\xi}\boldsymbol{\psi}_{\xi}^{*}+\boldsymbol{\epsilon}_{ij}=\boldsymbol{V}_{ij\xi}\boldsymbol{\psi}_{\xi}^{*}+\boldsymbol{\epsilon}_{ij}, \tag{11}$$

that represents a dynamic linear model, since the design matrix $\boldsymbol{V}_{ij\xi}=\sum_{k=1}^{j-1}[\boldsymbol{I}_p\otimes\boldsymbol{Y}_{ik}^{\top}\otimes(\boldsymbol{z}_{ijk}^{*})^{\top}]_{\xi}$ involves the predecessors of $\boldsymbol{Y}_{ij}$. Note that, by using ξ as subscript in a matrix $[.]_{\xi}$, we mean the matrix with the columns that correspond to the zero elements of $\boldsymbol{\xi}$ removed. When the mean of $\boldsymbol{Y}_{ij}$ is not zero, we replace $\boldsymbol{Y}_{ij}$ in (11) by its centred version $\boldsymbol{Y}_{ij}-\boldsymbol{X}_{ij}^{*}\boldsymbol{\beta}^{*}$. This leads to

$$\boldsymbol{Y}_{ij}-\boldsymbol{X}_{ij}^{*}\boldsymbol{\beta}^{*}=\sum_{k=1}^{j-1}\boldsymbol{\Phi}_{ijk}(\boldsymbol{Y}_{ik}-\boldsymbol{X}_{ik}^{*}\boldsymbol{\beta}^{*})+\boldsymbol{\epsilon}_{ij}, \tag{12}$$

which can be written in the more familiar form of a multivariate mixed model

$$\boldsymbol{Y}_{ij}=\boldsymbol{X}_{ij}^{*}\boldsymbol{\beta}^{*}+\boldsymbol{V}_{ij\xi\beta}\boldsymbol{\psi}_{\xi}^{*}+\boldsymbol{\epsilon}_{ij}, \tag{13}$$

where the design matrix $\boldsymbol{V}_{ij\xi\beta}$ depends on $\boldsymbol{\beta}^{*}$.

2.4 Innovation covariance matrices

For modelling the innovation covariance matrices $\boldsymbol{D}_{ij}, i=1,\ldots,n, j=1,\ldots,n_i$, we begin by employing the separation strategy of Barnard et al,[13] by which $\boldsymbol{D}_{ij}$ is decomposed into a diagonal matrix of variances $\boldsymbol{S}_{ij}=\mathrm{diag}(\sigma_{ij1}^2,\ldots,\sigma_{ijp}^2)$ and a correlation matrix $\boldsymbol{R}_{ij}$,

$$\boldsymbol{D}_{ij}=\boldsymbol{S}_{ij}^{1/2}\boldsymbol{R}_{ij}\boldsymbol{S}_{ij}^{1/2}. \tag{14}$$

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

2.4.1 Diagonal innovation variance matrices

It is easy to model matrix $\boldsymbol{S}_{ij}$ in terms of covariates as the only requirement on its diagonal elements is that they are nonnegative. It is clear that the following semiparametric model satisfies this requirement

$$\log \sigma^2_{ijk} = \alpha_{k0} + \sum_{l=1}^{L_1} w_{ijl}\alpha_{kl} + \sum_{l=L_1+1}^{L} f_{\sigma,k,l}(w_{ijl}) = \alpha_{k0} + \boldsymbol{w}_{ij}^\top \boldsymbol{\alpha}_k. \tag{15}$$

Similar models for innovation variances have appeared in Leng et al[25] and Lin and Pan.[26]

Consider now vectors of indicator variables for selecting the elements of $\boldsymbol{w}_{ij}$ that enter the kth variance regression model. In line with the indicator variables for the mean and autoregressive models, these are denoted by $\boldsymbol{\delta}_k = (\delta_{k1}, \ldots, \delta_{kL_1}, \boldsymbol{\delta}^\top_{k,L_1+1}, \ldots, \boldsymbol{\delta}^\top_{kL})^\top$. Given $\boldsymbol{\delta}_k$, model (15) can be expressed as

$$\log \sigma^2_{ijk} = \alpha_{k0} + \boldsymbol{w}_{ijk}^\top \boldsymbol{\alpha}_{\delta_k},$$

or equivalently

$$\sigma^2_{ijk} = \exp(\alpha_{k0}) \exp(\boldsymbol{w}_{ijk}^\top \boldsymbol{\alpha}_{\delta_k}) = \sigma^2_k \exp(\boldsymbol{w}_{ijk}^\top \boldsymbol{\alpha}_{\delta_k}). \tag{16}$$

2.4.2 Time-varying innovation correlation matrices

The approach we take for modelling the correlation matrices $\boldsymbol{R}_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_i$, extends the work of Liechty et al[17] and Papageorgiou and Marshall[21] who proposed models for correlation matrices in the context of cross-sectional studies. In the context of longitudinal studies, we propose to model these matrices utilizing time as the only covariate. Hence, we use symbols $\boldsymbol{R}_t = \{r_{tkl}\}, t \in T$, to denote these matrices. Below we describe three priors, termed 'common correlations', 'grouped correlations' and 'grouped variables' priors.

2.4.3 Common correlations

The common correlations model is defined as follows

$$f(\boldsymbol{R}_t|\mu_{ct}, \sigma^2_{ct}) = \pi(\mu_{ct}, \sigma^2_{ct}) \prod_{k<l} \left\{\exp\{-(g(r_{tkl}) - \mu_{ct})^2/2\sigma^2_{ct}\} J[g(r_{tkl}) \to r_{tkl}]\right\} I[\boldsymbol{R}_t \in \mathcal{C}], \tag{17}$$

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

$$\pi^{-1}(\mu_{ct}, \sigma^2_{ct}) = \int_{\boldsymbol{R}_t \in \mathcal{C}} \prod_{k<l} \left\{\exp\{-(g(r_{tkl}) - \mu_{ct})^2/2\sigma^2_{ct}\} J[g(r_{tkl}) \to r_{tkl}]\right\} dr_{tkl}.$$

A typical choice for $g(r)$ is the Fisher's z transformation $g(r) = \log([1 + r]/[1 - r])/2$ that leads to $J[g(r) \to r] = (1 - r)^{-1}(1 + r)^{-1}$.

We model the parameters of this distribution using nonparametric regression, with time t as the only independent variable

$$\mu_{ct} = \eta_0 + f_\mu(t) = \eta_0 + \boldsymbol{z}_{\mu t}^\top \boldsymbol{\eta}, \tag{18}$$

$$\log \sigma^2_{ct} = \omega_0 + f_\sigma(t) = \omega_0 + \boldsymbol{z}_{\sigma t}^\top \boldsymbol{\omega}. \tag{19}$$

Equivalently, we can write (19) as

$$\sigma_{ct}^2 = \exp(\omega_0)\exp(\boldsymbol{z}_{\sigma t}^\top \boldsymbol{\omega}) = \sigma^2 \exp(\boldsymbol{z}_{\sigma t}^\top \boldsymbol{\omega}).$$

If the indicator function were not present in (17), then the model described in (17) - (19) could be thought of as a semi-parametric regression model for the mean and the variance of normally distributed responses (see e.g. Chan et al[19] and Papageorgiou[20]). Let vector $\boldsymbol{r}_t$ consist of the non-redundant elements of $\boldsymbol{R}_t, t \in T$, and let $\boldsymbol{r} = (\boldsymbol{r}_{t_1}^\top, \ldots, \boldsymbol{r}_{t_M}^\top)^\top$. Subject to the pd constraint, we can express the common correlations model as

$$g(\boldsymbol{r}) = \boldsymbol{Z}^* \boldsymbol{\eta}^* + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{D}^2(\boldsymbol{\omega})), \tag{20}$$

where $\boldsymbol{\eta}^* = (\eta_0, \boldsymbol{\eta}^\top)^\top$, $\boldsymbol{Z}^*$ is a design matrix with rows equal to $\boldsymbol{z}_{\mu t}^* = (1, \boldsymbol{z}_{\mu t}^\top)^\top$ and $\boldsymbol{D}(\boldsymbol{\omega})$ is a diagonal matrix with elements equal to $\exp(\boldsymbol{z}_{\sigma t}^\top \boldsymbol{\omega}/2), t \in T$, where $\boldsymbol{z}_{\mu t}^*$ and $\exp(\boldsymbol{z}_{\sigma t}^\top \boldsymbol{\omega}/2)$ appear in the corresponding matrices $d = p(p-1)/2$ times, the number of unique elements of $\boldsymbol{R}_t$.

Introducing now vectors of indicators $\boldsymbol{\nu}$ and $\boldsymbol{\varphi}$ for selecting the elements of $\boldsymbol{z}_{\mu t}$ and $\boldsymbol{z}_{\sigma t}$ that enter the models for μ_{ct} and σ_{ct}^2, respectively, model (20) can be written as

$$g(\boldsymbol{r}) = \boldsymbol{Z}_\nu^* \boldsymbol{\eta}_\nu^* + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{D}^2(\boldsymbol{\omega}_\varphi)), \tag{21}$$

where $\boldsymbol{\eta}_\nu^*$ and $\boldsymbol{\omega}_\varphi$ consist of the non-zero elements of $\boldsymbol{\eta}^*$ and $\boldsymbol{\omega}$, and $\boldsymbol{Z}_\nu^*$ consists of the columns of $\boldsymbol{Z}^*$ that correspond to the non-zero elements of $\boldsymbol{\eta}^*$.

2.4.4 Grouped correlations

The common correlations model is restrictive as it implies that all correlations evolve in the same way over time. It is important, however, from a theoretical standpoint to estimate the correlation structure correctly. Hence, here we generalize the model to include multiple surfaces by considering the following prior

$$\begin{aligned} &f(\boldsymbol{R}_t | \boldsymbol{\mu}_{ct}, \sigma_{ct}^2, \boldsymbol{\lambda}) = \pi(\boldsymbol{\mu}_{ct}, \sigma_{ct}^2, \boldsymbol{\lambda}) \\ &\times \prod_{k<l} \left\{ \sum_{h=1}^{H} I[\lambda_{kl} = h] \exp[-(g(r_{tkl}) - \mu_{cth})^2 / 2\sigma_{ct}^2] \right\} J[g(r_{tkl}) \to r_{tkl}] I[\boldsymbol{R}_t \in \mathcal{C}], \end{aligned} \tag{22}$$

where H denotes the number of surfaces in the model, λ_{kl} is the surface assignment indicator and

$$\mu_{cth} = \eta_{h0} + f_{\mu h}(t) = \eta_{h0} + \boldsymbol{z}_{\mu th}^\top \boldsymbol{\eta}_{\nu_h} = (\boldsymbol{z}_{\mu th}^*)^\top \boldsymbol{\eta}_{\nu_h}^*.$$

In the above, $\boldsymbol{\eta}_h$ are surface specific regression coefficients and $\boldsymbol{\nu}_h$ are surface specific variable selection indicators, $h = 1, \ldots, H$.

Let vector $\boldsymbol{r}_{th}$ consist of the non-redundant elements of $\boldsymbol{R}_t, t \in T$, assigned to surface h. Further, let $\boldsymbol{r}_h = (\boldsymbol{r}_{t_1 h}^\top, \ldots, \boldsymbol{r}_{t_M h}^\top)^\top$. Subject to the pd constraint, we can express model (22) as

$$g(\boldsymbol{r}_h) = \boldsymbol{Z}_{\nu_h}^* \boldsymbol{\eta}_{\nu_h}^* + \boldsymbol{\epsilon}_h, h = 1, \ldots, H, \tag{23}$$

where $\boldsymbol{\epsilon}_h \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{D}_h^2(\boldsymbol{\omega}_\varphi))$.

2.4.5 Grouped variables

Here we describe another clustering model that, instead of correlations, clusters the variables. We have the following prior

$$f(\boldsymbol{R}_t|\boldsymbol{\mu}_{ct},\sigma^2_{ct},\boldsymbol{\lambda}) = \pi(\boldsymbol{\mu}_{ct},\sigma^2_{ct},\boldsymbol{\lambda})$$
$$\times\prod_{k<l}\left\{\sum_{h_1,h_2=1}^{G} I[\lambda_k=h_1]I[\lambda_l=h_2]\exp[-(g(r_{tkl})-\mu_{cth_1h_2})^2/2\sigma^2_{ct}]\right\} J[g(r_{tkl})\to r_{tkl}]I[\boldsymbol{R}_t\in\mathcal{C}], \quad (24)$$

where

$$\mu_{cth_1h_2} = \eta_{h_1h_20} + f_{\mu h_1h_2}(t) = \eta_{h_1h_20} + \boldsymbol{z}^\top_{\mu th_1h_2}\boldsymbol{\eta}_{\nu_{h_1h_2}} = (\boldsymbol{z}^*_{\mu th_1h_2})^\top\boldsymbol{\eta}^*_{\nu_{h_1h_2}},$$

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

2.5 Prior specification

In this section we specify the priors for the model parameters. We begin by describing the priors of the parameters of the mean model. For $\boldsymbol{\beta}^*_\gamma$ we specify a g-prior[22],

$$\boldsymbol{\beta}^*_\gamma|c_\beta,\boldsymbol{\gamma},\boldsymbol{\Sigma}\sim N(\boldsymbol{0},c_\beta(\tilde{\boldsymbol{X}}^\top_\gamma\tilde{\boldsymbol{X}}_\gamma)^{-1}), \quad (25)$$

where $\tilde{\boldsymbol{X}}_\gamma = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}^*_\gamma$, and $\boldsymbol{\Sigma}$ is the covariance matrix of the $\boldsymbol{Y}$, that is $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_i, i=1,\ldots,n)$. Further, the prior for c_β is taken to be an inverse Gamma, $c_\beta\sim\text{IG}(a_\beta,b_\beta)$.

Consider now the vectors $\boldsymbol{\gamma}_k = (\gamma_{k1},\ldots,\gamma_{kK_1},\boldsymbol{\gamma}^\top_{k,K_1+1},\ldots,\boldsymbol{\gamma}^\top_{kK})^\top, k=1,\ldots,p$, of variable selection indicators for the mean functions. We specify independent binomial priors for each of their K subvectors,

$$P(\boldsymbol{\gamma}_{kl}|\pi_{\mu kl}) = \pi^{N(\gamma_{kl})}_{\mu kl}(1-\pi_{\mu kl})^{q_{\mu l}-N(\gamma_{kl})}, l=1,\ldots,K, \quad (26)$$

where, for parametric effects, $N(\gamma_{kl})=\gamma_{kl}$ and $q_{\mu l}=1$, $l=1,\ldots,K_1$, while for nonparametric effects, $N(\gamma_{kl})=\sum_{s=1}^{q_{\mu l}}\gamma_{kls}$ and $q_{\mu l}$ was defined in (7), $l=K_1+1,\ldots,K$. Independent Beta priors are specified for $\pi_{\mu kl}\sim\text{Beta}(c_{\mu kl},d_{\mu kl})$, $k=1,\ldots,p, l=1,\ldots,K$.

Moving on to the priors of the parameter of the autoregressive coefficients, a normal prior is specified for the non-zero coefficients $\boldsymbol{\psi}^*_{\xi lm}, l,m=1,\ldots,p$, of the model in (10)

$$\boldsymbol{\psi}^*_{\xi lm}\sim N(\boldsymbol{0},c^2_{\psi lm}\boldsymbol{I}).$$

For the scale parameter $c^2_{\psi lm}$, we consider inverse Gamma and half-normal priors, $c^2_{\psi lm}\sim\text{IG}(a_{\psi lm},b_{\psi lm})$ and $c_{\psi lm}\sim N(0,\phi^2_{\psi lm})I[c_{\psi lm}>0]$. Further, the specification of the priors for the vectors of variable selection indicators $\boldsymbol{\xi}_{lm} = (\xi_{lm0},\xi_{lm1},\ldots,\xi_{lmB_1},\boldsymbol{\xi}^\top_{lm,B_1+1},\ldots,\boldsymbol{\xi}^\top_{lmB})^\top, l,m=1,\ldots,p$, follows the same pattern as for $\boldsymbol{\gamma}_k$ indicators. That is, independent binomial priors are specified for each of the $1+B$ subvectors of $\boldsymbol{\xi}_{lm}$,

$$P(\boldsymbol{\xi}_{lmb}|\pi_{\phi lmb}) = \pi^{N(\xi_{lmb})}_{\phi lmb}(1-\pi_{\phi lmb})^{q_{\phi b}-N(\xi_{lmb})}, b=0,\ldots,B,$$

where $N(\xi_{lmb})=\xi_{lmb}$ and $q_{\phi b}=1$ for $b=0,\ldots,B_1$, and $N(\xi_{lmb})=\sum_{s=1}^{q_{\phi b}}\xi_{lmbs}$ and $q_{\phi b}$ was defined after (9) for $b=B_1+1,\ldots,B$. Independent Beta priors are specified for $\pi_{\phi lmb}\sim\text{Beta}(c_{\phi lmb},d_{\phi lmb}), l,m=1,\ldots,p$.

Next, we describe the priors for the innovation variance parameters. First, for $\boldsymbol{\alpha}_{\delta_k}$ we specify independent normal priors

$$\boldsymbol{\alpha}_{\delta_k}|c_{\alpha k}, \boldsymbol{\delta}_k \sim N(\mathbf{0}, c_{\alpha k}^2 \boldsymbol{I}), k = 1, \ldots, p.$$

Further, we consider inverse Gamma and half-normal priors for $c_{\alpha k}$, $c_{\alpha k}^2 \sim \text{IG}(a_{\alpha k}, b_{\alpha k})$ and $c_{\alpha k} \sim N(0, \phi_{\alpha k}^2) I[c_{\alpha k} > 0], k = 1, \ldots, p$. For the scale parameter σ_k^2 in (16), $k = 1, \ldots, p$, we also consider inverse Gamma and half-normal priors, $\sigma_k^2 \sim \text{IG}(a_{\sigma k}, b_{\sigma k})$ and $\sigma_k \sim N(0, \phi_{\sigma k}^2) I[\sigma_k > 0]$. Continuing with the priors for the vectors of indicators $\boldsymbol{\delta}_k = (\delta_{k1}, \ldots, \delta_{kL_1}, \boldsymbol{\delta}_{k,L_1+1}^\top, \ldots, \boldsymbol{\delta}_{kL}^\top)^\top$, $k = 1, \ldots, p$, we specify independent binomial priors for each of their L subvectors,

$$P(\boldsymbol{\delta}_{kl}|\pi_{\sigma kl}) = \pi_{\sigma kl}^{N(\delta_{kl})} (1 - \pi_{\sigma kl})^{q_{\sigma l} - N(\delta_{kl})}, l = 1, \ldots, L,$$

where $N(\delta_{kl})$ and $q_{\sigma l}$ are analogous to $N(\gamma_{kl})$ and $q_{\mu l}$ (and $N(\xi_{lmb})$ and $q_{\phi b}$). We specify independent Beta priors for $\pi_{\sigma kl} \sim \text{Beta}(c_{\sigma kl}, d_{\sigma kl}), k = 1, \ldots, p, l = 1, \ldots, L$.

Lastly, we describe the priors placed on the parameters of the models of the innovation correlation matrices. Starting with the 'common correlations model' in (21), let $\tilde{\boldsymbol{Z}} = \boldsymbol{D}(\boldsymbol{\omega})^{-1} \boldsymbol{Z}^*$. The prior for the non-zero part of $\boldsymbol{\eta}^*$ is the following g-prior

$$\boldsymbol{\eta}_\nu^*|c_\eta, \sigma^2, \boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{\varphi} \sim N(\mathbf{0}, c_\eta \sigma^2 (\tilde{\boldsymbol{Z}}_\nu^\top \tilde{\boldsymbol{Z}}_\nu)^{-1}). \tag{27}$$

Further, the prior for c_η is specified as $c_\eta \sim \text{IG}(a_\eta, b_\eta)$. Furthermore, we specify a binomial prior for vector $\boldsymbol{\nu}$

$$P(\boldsymbol{\nu}|\pi_\nu) = \pi_\nu^{N(\nu)} (1 - \pi_\nu)^{q_{R\mu} - N(\nu)},$$

where $q_{R\mu}$ denotes the number of basis functions used in linearising $f_\mu(t)$ in (18), and $N(\nu) = \sum_{j=1}^{q_{R\mu}} \nu_j$. A Beta prior is specified for $\pi_\nu \sim \text{Beta}(c_\nu, d_\nu)$.

Further, the prior for $\boldsymbol{\omega}_\varphi$ is specified as

$$\boldsymbol{\omega}_\varphi|c_\omega, \boldsymbol{\varphi} \sim N(\mathbf{0}, c_\omega^2 \boldsymbol{I}).$$

For c_ω we consider inverse Gamma and half-normal priors, $c_\omega^2 \sim \text{IG}(a_\omega, b_\omega)$ and $c_\omega \sim N(0, \phi_{c_\omega}^2) I[c_\omega > 0]$. Continuing with the prior for vector $\boldsymbol{\varphi}$, this is taken to be

$$P(\boldsymbol{\varphi}|\pi_\varphi) = \pi_\varphi^{N(\varphi)} (1 - \pi_\varphi)^{q_{R\sigma} - N(\varphi)},$$

where $q_{R\sigma}$ denotes the number of basis functions used in linearising $f_\sigma(t)$ in (19) and $N(\varphi) = \sum_{j=1}^{q_{R\sigma}} \varphi_j$. We specify a Beta prior for $\pi_\varphi \sim \text{Beta}(c_\varphi, d_\varphi)$. Further, we specify inverse Gamma and half-normal priors for σ^2, $\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma)$ and $\sigma \sim N(\sigma; 0, \phi_\sigma^2) I[\sigma > 0]$.

Next we describe the priors for the cluster specific parameters $(\boldsymbol{\nu}_h, \boldsymbol{\eta}_h)$ and prior weights $w_h, h = 1, \ldots, H$, of the grouped correlations model. Firstly, conditionally on $\boldsymbol{\nu}_h$, the non-zero elements of $\boldsymbol{\eta}_h$ are specified to have the following prior

$$\begin{aligned} \boldsymbol{\eta}_{\nu_h}|c_\eta, \sigma^2, \boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{\varphi} &\sim N(\mathbf{0}, c_\eta \sigma^2 (\tilde{\boldsymbol{Z}}_{\nu_h}^\top \tilde{\boldsymbol{Z}}_{\nu_h})^{-1}), \quad \text{if the cluster is non-empty,} \\ \boldsymbol{\eta}_{\nu_h}|c_\eta, \sigma^2, \boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{\varphi} &\sim N(\mathbf{0}, c_\eta \sigma^2 \boldsymbol{I}), \quad \text{if the cluster is empty.} \end{aligned} \tag{28}$$

Secondly, for vectors $\boldsymbol{\nu}_h$ we specify binomial priors

$$P(\boldsymbol{\nu}_h|\pi_{\nu_h}) = \pi_{\nu_h}^{N(\nu_h)} (1 - \pi_{\nu_h})^{q_{R\mu} - N(\nu_h)},$$

with the prior on π_{ν_h} being $\pi_{\nu_h} \sim \text{Beta}(c_\nu, d_\nu), h = 1, \ldots, H$.

Lastly, the prior weights w_h are constructed utilizing the so called stick-breaking process.[27;28] Let $v_h, h = 1, \ldots, H-1$, be independent draws from a $\text{Beta}(1, \alpha^*)$ distribution. We have: $w_1 = v_1$, for $2 \leq l < H$, $w_l = v_l \prod_{h=1}^{l-1}(1 - v_h)$, and $w_H = \prod_{h=1}^{H-1}(1 - v_h)$. 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 $a_{\alpha*}/b_{\alpha*}$.

In the numerical illustrations that we present in this article, we use the following priors. For c_β we specify $\text{IG}(1/2, np/2)$, as a p-variate analogue to the prior of Liang et al[29]. For $\pi_{\mu kl}, \pi_{\phi lmb}, \pi_{\sigma kl}, \pi_\nu$, and π_φ we specify $\text{Beta}(1, 1)$ priors. The priors on $c_{\alpha k}, k = 1, \ldots, p$, and c_ω are specified to be $\text{IG}(1.1, 1.1)$. Further, for $c_{\psi lm}, l, m = 1, \ldots, p$, $\sigma_k, k = 1, \ldots, p$, and σ we define $\text{HN}(2)$ priors. Furthermore, for c_η, we specify $\text{IG}(1/2, Md/2)$, again, as an analogue to the prior of Liang et al.[29] Lastly, the prior on the DP concentration parameter is specified as $\alpha^* \sim \text{Gamma}(5, 2)$.

3 Posterior Sampling

Posterior sampling is carried out by expressing the likelihood function in different ways and using the expression that is more computationally convenient for each step of the MCMC algorithm. Further, where possible, we integrate out parameters to improve mixing, and where beneficial, we augment the model with additional parameters to make sampling easier. Here we only present the main concepts and a detailed description of all MCMC steps is available in the supplementary material.

We first consider the likelihood that is obtained from the normality assumption, $\boldsymbol{Y}_i \sim N(\boldsymbol{X}^*_{\gamma i}\boldsymbol{\beta}^*_\gamma, \boldsymbol{\Sigma}_i), i = 1, \ldots, n$. Due to the factorization in (1), the quadratic form $Q = \sum_{i=1}^n (\boldsymbol{Y}_i - \boldsymbol{X}^*_{\gamma i}\boldsymbol{\beta}^*_\gamma)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}^*_{\gamma i}\boldsymbol{\beta}^*_\gamma)$ of the likelihood may be written as

$$Q = \sum_{i=1}^{n} (\boldsymbol{L}_i \boldsymbol{r}_i)^\top \boldsymbol{D}_i^{-1} (\boldsymbol{L}_i \boldsymbol{r}_i) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} (\boldsymbol{r}_{ij} - \hat{\boldsymbol{r}}_{ij})^\top \boldsymbol{D}_{ij}^{-1} (\boldsymbol{r}_{ij} - \hat{\boldsymbol{r}}_{ij}), \tag{29}$$

where $\boldsymbol{r}_i = \boldsymbol{Y}_i - \boldsymbol{X}^*_i \boldsymbol{\beta}^*$.

Further, recalling (12) and the definition of prediction error, and the separation of variances and correlations in (14), Q may be written as

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \boldsymbol{\epsilon}_{ij}^\top [\boldsymbol{S}_{ij}^{1/2} \boldsymbol{R}_{t_j} \boldsymbol{S}_{ij}^{1/2}]^{-1} \boldsymbol{\epsilon}_{ij} = \sum_{t \in T} \sum_{i \in O_t} \check{\boldsymbol{\epsilon}}_{it}^\top \boldsymbol{R}_t^{-1} \check{\boldsymbol{\epsilon}}_{it}, \tag{30}$$

where $\check{\boldsymbol{\epsilon}}_{ij} = \boldsymbol{S}_{ij}^{-1/2} \boldsymbol{\epsilon}_{ij}$. In the first equality, matrix $\boldsymbol{R}$ is subscripted by t_j because here we model its elements in terms of covariate 'time' t only. In the second equality, T denotes the set of unique observational time points and O_t the set of sampling units observed at time point t. We denote the cardinality of O_t by n_t.

In addition, recalling (11) and its centred versions in (12) and (13), we can write (29) as

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n_i} (\boldsymbol{r}_{ij} - \boldsymbol{V}_{ij\xi\beta} \boldsymbol{\psi}^*_\xi)^\top \boldsymbol{D}_{ij}^{-1} (\boldsymbol{r}_{ij} - \boldsymbol{V}_{ij\xi\beta} \boldsymbol{\psi}^*_\xi), \tag{31}$$

where $\boldsymbol{V}_{ij\xi\beta} = \sum_{k=1}^{j-1} [\boldsymbol{I}_p \otimes \boldsymbol{r}_{ik}^\top \otimes (\boldsymbol{z}^*_{ijk})^\top]_\xi$. We take $\boldsymbol{V}_{ij\xi\beta}$ to be a matrix of zeros when $j = 1$.

To improve mixing of the MCMC algorithm, we can integrate out vector $\boldsymbol{\beta}^*_\gamma$ from the likelihood of $\boldsymbol{Y}$

$$f(\boldsymbol{Y}) = (2\pi)^{-\frac{Np}{2}} \big(\prod_{i=1}^{n} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \big) (c_\beta + 1)^{-\frac{N(\gamma)+p}{2}} \exp(-S/2), \tag{32}$$

where

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

$\tilde{\boldsymbol{Y}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Y}$, $N = \sum_{i=1}^{n} n_i$, $\tilde{\boldsymbol{X}}_\gamma = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}_\gamma^*$ and $N(\gamma)+p$ is the total number of columns in $\boldsymbol{X}_\gamma^*$.

To write function S in a computationally convenient way, first note that $\tilde{\boldsymbol{Y}}^{\top}\tilde{\boldsymbol{Y}} = \sum_{i=1}^{n} \boldsymbol{Y}_i^{\top}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Y}_i$, which, due to (1), may be written as $\sum_{i=1}^{n}(\boldsymbol{L}_i\boldsymbol{Y}_i)^{\top}\boldsymbol{D}_i^{-1}\boldsymbol{L}_i\boldsymbol{Y}_i$. Further, due to the separation of variances and correlations in (14), we may further express the latter as $\sum_{i=1}^{n}(\boldsymbol{S}_i^{-1/2}\boldsymbol{L}_i\boldsymbol{Y}_i)^{\top}\boldsymbol{R}_i^{-1}\boldsymbol{S}_i^{-1/2}\boldsymbol{L}_i\boldsymbol{Y}_i = \sum_{i=1}^{n}\check{\boldsymbol{Y}}_i^{\top}\boldsymbol{R}_i^{-1}\check{\boldsymbol{Y}}_i$, where $\check{\boldsymbol{Y}}_i = \boldsymbol{S}_i^{-1/2}\boldsymbol{L}_i\boldsymbol{Y}_i$. Furthermore, we write the last expression as $\sum_{i=1}^{n}\sum_{j=1}^{n_i}\check{\boldsymbol{y}}_{ij}^{\top}\boldsymbol{R}_{t_j}^{-1}\check{\boldsymbol{y}}_{ij}$. Likewise, we may write that $\tilde{\boldsymbol{X}}_\gamma^{\top}\tilde{\boldsymbol{X}}_\gamma = \sum_{i=1}^{n}(\boldsymbol{L}_i\boldsymbol{X}_{\gamma i}^*)^{\top}\boldsymbol{D}_i^{-1}\boldsymbol{L}_i\boldsymbol{X}_{\gamma i}^* = \sum_{i=1}^{n}(\boldsymbol{S}_i^{-1/2}\boldsymbol{L}_i\boldsymbol{X}_{\gamma i}^*)^{\top}\boldsymbol{R}_i^{-1}\boldsymbol{S}_i^{-1/2}\boldsymbol{L}_i\boldsymbol{X}_{\gamma i}^* = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(\check{\boldsymbol{X}}_{\gamma ij}^*)^{\top}\boldsymbol{R}_{t_j}^{-1}\check{\boldsymbol{X}}_{\gamma ij}^*$. In addition, $\tilde{\boldsymbol{X}}_\gamma^{\top}\tilde{\boldsymbol{Y}} = \sum_{i=1}^{n}(\boldsymbol{L}_i\boldsymbol{X}_{\gamma i}^*)^{\top}\boldsymbol{D}_i^{-1}\boldsymbol{L}_i\boldsymbol{Y}_i = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(\check{\boldsymbol{X}}_{\gamma ij}^*)^{\top}\boldsymbol{R}_{t_j}^{-1}\check{\boldsymbol{y}}_{ij}$. It follows that

$$S = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\check{\boldsymbol{y}}_{ij}^{\top}\boldsymbol{R}_{t_j}^{-1}\check{\boldsymbol{y}}_{ij} -$$
$$c_\beta(1+c_\beta)^{-1}[\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\check{\boldsymbol{y}}_{ij})^{\top}\boldsymbol{R}_{t_j}^{-1}\check{\boldsymbol{X}}_{\gamma ij}^*][\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\check{\boldsymbol{X}}_{\gamma ij}^*)^{\top}\boldsymbol{R}_{t_j}^{-1}\check{\boldsymbol{X}}_{\gamma ij}^*]^{-1}[\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\check{\boldsymbol{X}}_{\gamma ij}^*)^{\top}\boldsymbol{R}_{t_j}^{-1}\check{\boldsymbol{y}}_{ij}].$$

The most computationally expensive step of the MCMC algorithm is the one that samples from the posterior of the parameters of the correlation matrices. This is because acceptance probabilities of the Metropolis-Hastings step involve the ratio of the normalizing constants of the density in (17) (or the ones in (22) and (24)). Calculating this ratio is very computationally demanding, but it can be avoided by introducing the 'shadow prior'.[17] The basic idea is to introduce latent variables θ_{tkl} between the correlations r_{tkl} and the means μ_{ct}. This modifies prior (17) as follows

$$f(\boldsymbol{R}_t|\boldsymbol{\theta}_t,\tau^2) = \pi(\boldsymbol{\theta}_t,\tau^2)\prod_{k<l}\exp[-(g(r_{tkl})-\theta_{tkl})^2/2\tau^2]J[g(r_{tkl})\to r_{tkl}]I[\boldsymbol{R}_t\in\mathcal{C}], \tag{33}$$

where

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

and τ is a fixed, small constant. Further, variables θ_{tkl} are independently distributed with time specific means and variances, $\theta_{tkl}\sim N(\mu_{ct},\sigma_{ct}^2)$. Hence, the semiparametric model (21) now applies on the unconstrained $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} = \boldsymbol{Z}_\nu^*\boldsymbol{\eta}_\nu^* + \boldsymbol{\epsilon}, \boldsymbol{\epsilon}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{D}^2(\boldsymbol{\omega}_\varphi)). \tag{34}$$

Sampling from the posterior of $\boldsymbol{\theta}_t = \{\theta_{tkl}\}$ still involves the ratio of the normalising constants of the density in (33), but that, as was argued by Liechty et al,[17] for small τ, can reasonably be approximated by one. In all numerical examples presented here, we take $\tau = 0.01$.

Considering now the 'grouped correlations' model, with the introduction of the shadow prior, model (22) becomes the same as the one in (33). The difference is that here variables θ_{tkl} are independent with conditional distribution $\theta_{tkl}|\lambda_{kl} = h \sim N(\mu_{cth},\sigma_{ct}^2)$. Further, letting $\boldsymbol{\theta}_h$ be analogous to $\boldsymbol{r}_h$, defined above (23), it is easy to see that semiparametric model (23) now applies on the unconstrained $\boldsymbol{\theta}_h$

$$\boldsymbol{\theta}_h = \boldsymbol{Z}_{\nu_h}^*\boldsymbol{\eta}_{\nu_h}^* + \boldsymbol{\epsilon}_h, h = 1,\ldots,H. \tag{35}$$

Additionally, letting $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H)^\top$, model (35) may be written in vectorized form in precisely the same way as (34).

To sample from the full conditional of $\boldsymbol{R}_t, t \in T$, recall (30) and (33). It follows that

$$f(\boldsymbol{R}_t|\ldots) \propto |\boldsymbol{R}_t|^{-\frac{n_t}{2}} \text{etr}(-\boldsymbol{R}_t^{-1}\boldsymbol{S}_t/2) \prod_{k<l} \exp[-(g(r_{tkl}) - \theta_{tkl})^2/2\tau^2] J[g(r_{tkl}) \to r_{tkl}] I[\boldsymbol{R}_t \in \mathcal{C}], \tag{36}$$

where n_t denotes the total number of observations at time point t, $\text{etr}(.) = \exp(\text{tr}(.))$, and $\boldsymbol{S}_t = \sum_{i \in O_t} \check{\boldsymbol{\epsilon}}_{it} \check{\boldsymbol{\epsilon}}_{it}^\top$.

We obtain a proposal density and sample from (36) by utilizing the method of Zhang et al[30] and Liu and Daniels.[31] We start by substituting $\boldsymbol{R}_t$ for symmetric pd and otherwise unconstrained matrices $\boldsymbol{E}_t$. These are assumed to have an inverse Wishart prior $\boldsymbol{E}_t \sim \text{IW}(\zeta_t, \boldsymbol{\Psi}_t), t \in T$, with mean equal to the realization of $\boldsymbol{E}_t$ from the previous iteration of the sampler. Given the inverse Wishart prior on $\boldsymbol{E}_t$, we obtain the following easy to sample from inverse Wishart posterior

$$g(\boldsymbol{E}_t|\ldots) \propto |\boldsymbol{\Psi}_t|^{\frac{\zeta_t}{2}} |\boldsymbol{E}_t|^{-\frac{n_t+\zeta_t+p+1}{2}} \text{etr}\{-\boldsymbol{E}_t^{-1}(\boldsymbol{S}_t + \boldsymbol{\Psi}_t)/2\}. \tag{37}$$

We decompose $\boldsymbol{E}_t = \boldsymbol{B}_t^{1/2} \boldsymbol{R}_t \boldsymbol{B}_t^{1/2}$ into a diagonal matrix of variances $\boldsymbol{B}_t = \text{Diag}(b_{t1}^2, \ldots, b_{tp}^2)$, and a correlation matrix $\boldsymbol{R}_t$. The Jacobian associated with this transformation is $J(\boldsymbol{E}_t \to \boldsymbol{B}_t, \boldsymbol{R}_t) = \prod_{k=1}^p (b_{tk})^{p-1} = |\boldsymbol{B}_t|^{(p-1)/2}$. It follows that the joint density for $(\boldsymbol{B}_t, \boldsymbol{R}_t)$ is

$$h(\boldsymbol{B}_t, \boldsymbol{R}_t|\ldots) \propto |\boldsymbol{\Psi}_t|^{\frac{\zeta_t}{2}} |\boldsymbol{B}_t|^{(p-1)/2} |\boldsymbol{E}_t|^{-\frac{n_t+\zeta_t+p+1}{2}} \text{etr}\{-\boldsymbol{E}_t^{-1}(\boldsymbol{S}_t + \boldsymbol{\Psi}_t)/2\}. \tag{38}$$

Sampling from (38) at iteration $u+1$ of the sampler proceeds by sampling $\boldsymbol{E}_t^{(u+1)}$ from (37) and decomposing $\boldsymbol{E}_t^{(u+1)}$ into $(\boldsymbol{B}_t^{(u+1)}, \boldsymbol{R}_t^{(u+1)})$. Further, the pair $(\boldsymbol{B}_t^{(u+1)}, \boldsymbol{R}_t^{(u+1)})$ is accepted as a sample from (36) with probability

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

where in $h(,|)$, $\boldsymbol{\Psi}_t = (\zeta_t - p - 1)\boldsymbol{E}_t^{(u)}$, and ζ_t is a tuning parameter.

To improve mixing of the MCMC algorithm over the parameters of the models of the correlation matrices, we can integrate out vector $\boldsymbol{\eta}$ from the likelihood of $\boldsymbol{\theta}$. The marginal of $\boldsymbol{\theta}$, computed from (34) and (27), is

$$f(\boldsymbol{\theta}|\boldsymbol{\omega}, c_\eta, \boldsymbol{\nu}, \boldsymbol{\varphi}, \sigma^2) \propto |\sigma^2 D^2(\boldsymbol{\omega}_\varphi)|^{-\frac{1}{2}} (c_\eta + 1)^{-\frac{N(\nu)+1}{2}} \exp\{-S^*/2\sigma^2\}, \tag{39}$$

where $S^* = S^*(\boldsymbol{\theta}, \boldsymbol{\omega}, c_\eta, \boldsymbol{\nu}, \boldsymbol{\varphi}) = \tilde{\boldsymbol{\theta}}^\top \tilde{\boldsymbol{\theta}} - \frac{c_\eta}{1+c_\eta} \tilde{\boldsymbol{\theta}}^\top \tilde{\boldsymbol{Z}}_\nu (\tilde{\boldsymbol{Z}}_\nu^\top \tilde{\boldsymbol{Z}}_\nu)^{-1} \tilde{\boldsymbol{Z}}_\nu^\top \tilde{\boldsymbol{\theta}}$, and $\tilde{\boldsymbol{\theta}} = \boldsymbol{D}^{-1}(\boldsymbol{\omega}_\varphi)\boldsymbol{\theta}$.

Lastly, under the grouped correlations model, the marginal likelihood of $\boldsymbol{\theta}$ is given by

$$f(\boldsymbol{\theta}|\boldsymbol{\omega}, c_\eta, \boldsymbol{\nu}, \boldsymbol{\varphi}, \sigma^2) \propto |\sigma^2 D^2(\boldsymbol{\omega}_\varphi)|^{-\frac{1}{2}} (c_\eta + 1)^{-\frac{N(\nu)+H}{2}} \exp\{-S^*/2\sigma^2\},$$

where $N(\nu) = \sum_h N(\nu_h)$, and $S^* = S^*(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{c}_\eta, \boldsymbol{\nu}, \boldsymbol{\varphi}) = \sum_h S_h^* = \sum_h \tilde{\boldsymbol{\theta}}_h^\top \left\{\boldsymbol{I} - \frac{c_\eta}{1+c_\eta} \tilde{\boldsymbol{Z}}_{\nu_h} (\tilde{\boldsymbol{Z}}_{\nu_h}^\top \tilde{\boldsymbol{Z}}_{\nu_h})^{-1} \tilde{\boldsymbol{Z}}_{\nu_h}^\top\right\} \tilde{\boldsymbol{\theta}}_h$.

4 Simulation Study

The simulation study investigates the gains that one may have, in terms of reduced posterior mean squared error (MSE), when fitting multivariate longitudinal models instead of univariate ones. Additionally, the

study examines the effects that missing data have on the posterior MSE and whether adding more than one response to the model can counteract these effects.

We consider a data-generating mechanism that consists of 3 responses, Y_{1t}, Y_{2t}, Y_{3t}, observed over $M = 6$ time points, $t = -0.5, -0.3, \ldots, 0.5$. The 3 responses are generated from a multivariate normal distribution with mean $\boldsymbol{\mu}_t = (\beta_{10} + \beta_{11}t, 0, 0)^\top$, that is, the mean of the first response is a linear function of t, while the other 2 responses have constant zero mean. Further, the 18×18 covariance matrix is taken to be $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\rho_1, \rho_2) = \boldsymbol{\Sigma}_1(\rho_1) \otimes \boldsymbol{\Sigma}_{23}(\rho_2)$, where $\boldsymbol{\Sigma}_1(\rho_1) = \{\rho_1^{|k-l|}\}$ is the marginal covariance matrix of $\boldsymbol{Y}_1$ and $\boldsymbol{\Sigma}_{23}(\rho_2)$ has diagonal elements equal to 1 and all off diagonal elements equal to ρ_2.

The choice of autoregressive covariance structure $\boldsymbol{\Sigma}_1$ is a standard one. It implies that correlations between observations on Y_1 at different time points decrease quickly with increasing lag. The value of ρ_1 is not of particular interest here, and hence it is fixed at $\rho_1 = 0.5$, implying that correlations get halved every time the lag increases by 0.2. The structure of $\boldsymbol{\Sigma}$ indicates that Y_1 is equally correlated with Y_2 and Y_3 at all time points. This correlation is equal to ρ_2 at lag 0 and it gets halved when the lag increases by 0.2.

The chosen mean and covariance structures are fairly simple, and this suffices for the purposes of the current study. The main interest here is on the quality of the estimate of the mean function of the first response. We examine how this quality, as measured by the posterior MSE, or its components bias and variance, depends on the dimension of the response in the fitted model, the value of the correlation coefficient ρ_2, the sample size n, and the mechanism and percentage of missing data. The effect of the dimension of the response is evaluated by fitting 1-, 2-, and 3-dimensional response models. The effect of the correlation coefficient is examined by letting ρ_2 take values in the set $\{0.2, 0.4, 0.6, 0.8\}$. The effect of the sample size is assessed by letting n take values in the set $\{100, 300\}$. Lastly, the effect of the missing data is investigated by considering full datasets (0% probability of missingness), and missing completely at random (MCAR) and missing at random (MAR) mechanisms. In the MCAR mechanism observations are missing with probability 25% and 50%, while in the MAR mechanism observations are missing with probability given by

$$\text{logit}P(\text{missingness at time } t) = m_0 + m_1 Y_{1,t-1}.$$

We consider 2 values for the pair (m_0, m_1): firstly $(-1.53, -0.30)$ (denoted by MAR_1) and secondly $(0.49, 0.34)$ (denoted by MAR_2). For both MCAR and MAR we require that each subject is observed at least twice: at 'baseline' ($t = -0.5$) and at one occasion thereafter. The MAR_1 mechanism is associated with the following probabilities of missingness at the 6 times points: $0\%, 10\%, 13\%, 17\%, 20\%, 25\%$, while the MAR_2 with: $0\%, 30\%, 35\%, 40\%, 45\%, 50\%$.

Considering a 3-dimensional response model as an example, the mean functions are modelled using $\mu_{tk} = \beta_{k0} + \beta_{k1}t$, $k = 1, 2, 3$. This specification is correct for the first response and wrong for the other two responses. Further, the autoregressive coefficients and innovation variances are modelled using $\phi_{jklm} = \phi_{lm0} + f_{\phi,l,m}(t_j - t_k)$ and $\log \sigma^2_{tk} = \alpha_{k0} + f_{\sigma,k}(t)$, where $f_{\phi,l,m}(t_j - t_k)$ and $f_{\sigma,k}(t)$ include 5 and 6 knots, respectively, allowing for the possibility of non-linear effects. Lastly, we fit the common correlations model (17) with constant mean, $\mu_{ct} = \eta_0$, and constant variance, $\log \sigma^2_{ct} = \omega_0$. These choices for μ_{ct} and $\log \sigma^2_{ct}$ were arrived at after experimentation with more complex models that showed the more complex models to not be necessary.

The regression coefficients are taken to be $\beta_{10} = 0$ and $\beta_{11} = 2.95$, where the value of β_{11} was so chosen to achieve a signal-to-noise ratio (SNR) equal to 1. The SNR is defined as $\text{SNR} = (\text{SST} - \text{SSE})/\text{SSE}$, where SST the total sum of squares $\text{SST} = \sum_{i,j}(y_{ij1} - \bar{y}_1)^2$ and SSE the error sum of squares $\text{SSE} = \sum_{i,j}(y_{ij1} - \hat{y}_{ij1})^2$. Predictions $\hat{y}_{ij1}$ were obtained by fitting the true data-generating model.

For all models we obtain 3×10^4 posterior samples of which we discard the first 10^4 as burn in, and of the remaining 2×10^4 we keep one in two. Based on the 10^4 retained samples, we obtain samples for the

parameter of main interest, $\mu_{t1} = E(Y_{t1}|t) = \beta_{10} + \beta_{11}t$, by replacing the regression coefficients β_{10} and β_{11} by the corresponding sampled values, $\mu_{t1}^{(s)} = \beta_{10}^{(s)} + \beta_{11}^{(s)}t, s = 1, \ldots, 10^4$. We compare fitted models in terms of their posterior bias and variance in estimating μ_{t1}. Using subscript p to denote quantities obtained from models with fitted response dimension $p = 1, 2, 3$, the posterior bias and variance are computed as $B_p(t) = (\bar{\mu}_{t1,p} - \mu_{t1})^2$ and $V_p(t) = \sum_s (\mu_{t1,p}^{(s)} - \bar{\mu}_{t1,p})^2$, where $\bar{\mu}_{t1,p}$ is the mean of $\mu_{t1,p}^{(s)}, s = 1, \ldots, 10^4$. Further, because $B_p(t)$ and $V_p(t)$ are computed for $t = -0.5, -0.3, \ldots, 0.5$, results are summarised by computing sums: $B_p = \sum_t B_p(t)$ and $V_p = \sum_t V_p(t)$. Results presented are based on 30 replicate datasets.

Table 1 compares model performance by reporting the ratio $100B_p/B_1$, that we refer to as relative bias, while Table 2 compares model performance by reporting the ratio $100V_p/V_1$, that we refer to as relative variance, $p = 2, 3$. We can see from Table 1 that relative bias decreases as the correlation ρ_2 between the responses increases, for all n, p, and missingness mechanisms and probabilities. Comparing the left-hand side of the Table, that refers to 2-dimensional response models, to the right-hand side, that refers to 3-dimensional models, we see that the comparison depends on the missingness mechanisms and probabilities. For the 0%, MCAR and MAR_1 mechanisms, 2- and 3-dimensional models fair almost equally, but for the MAR_2 mechanism, the relative bias of the 3-dimensional model is usually lower than that of the 2-dimensional model. Table 2 displays the results on relative variances. We can observe that as correlations between the responses increase, the relative variance decreases, for all n, p, and missingness mechanisms and probabilities. The relative variance for 3-dimensional response models is lower than that of 2-dimensional models, for all missingness mechanisms and probabilities. These results are very much in line with what one would expect given the results of Zellner.[32]

We now study the effects of missing data on the posterior bias and variance. To do so, we introduce a second subscript: we let $B_{p,m}$ and $V_{p,m}$ denote the bias and variance in estimating μ_{t1} based on a p-dimensional response model where the observations can be either MCAR or MAR. In the former case, subscript m denotes the probability of missingness, $m = 0, 0.25, 0.50$, and in the latter case the MAR mechanism, $m = \text{MAR}_1$, MAR_2. Table 3 presents results on the ratios $100B_{p,m}/B_{1,0}, p = 1, 2, 3$, that is, it compares the bias of univariate response models that are based on the full dataset with the performance of univariate and multivariate response models that are based on subsets of the full dataset. First, looking at univariate response models ($p = 1$) and the MCAR 25% mechanism, bias increases by 8.4% and 8.7% for $n = 100$ and $n = 300$, respectively. These two percentages are derived by averaging the 4 relevant numbers in the Table. Adding a second response to the model is enough to counteract the increase in bias due to missing 25% of the data. We can see that almost all the 8 relevant entries in the Table are less than 100%, and as low as 75.19%. Adding a third response slightly improves the model performance relative to the model with 2 responses. Further, when the probability of missingness increases to 50% in the MCAR mechanism, the bias in the univariate response models increases, on average, by 29% and 37.9% for the 2 sample sizes (again computed by averaging the 4 relevant numbers in the Table). Adding a second response improves performance, but it cancels out the increase in the bias only when the correlation ρ_2 is as high as 0.80. Adding a third response does not further improve model performance. Moving on to the MAR_1 mechanism, the bias in the univariate response models increases, on average, by 17.3% and 30.4% for the 2 sample sizes. We can see that the addition of the second response reduces bias and that the addition of the third one offers a slight further improvement. Lastly, due to the MAR_2 mechanism, the average bias increase in the univariate response models is 62.3% and 38% for the 2 sample sizes. Clearly, the addition of the second response in the models reduces bias. The addition of the third response offers further improvements for all but one sample size by correlation combination.

Table 4 presents results on the ratios $100V_{p,m}/V_{1,0}, p = 1, 2, 3$. Concerning univariate response models, $p = 1$, when the mechanism is MCAR with missingness probability $m = 0.25$, the variance increases by a modest 4.6% and 1.1% for $n = 100$ and $n = 300$, respectively. Adding a second response to the

model counterbalances the effect of missing completely at random 25% of the data. It can be seen in the Table that, with only 1 exception, all entries are below 100%, and as low as 79.71%. The addition of a third response further improves model performance. Further, when the MCAR mechanism is associated with missingness probability of $m = 0.50$, the variance increases, on average, by 13.8% and 12.2% for the 2 sample sizes. Adding a second response completely cancels out the increase in the variance when $\rho_2 = 0.80$ and it almost cancels it out when $\rho_2 = 0.6$, for both sample sizes. A third response further improves performances and it neutralizes the increase in the variances when ρ_2 is 0.60 or higher. The MRA_1 mechanism increases the variance of univariate response models by 6.1% and 5% for $n = 100$ and $n = 300$, respectively, while the MRA_2 by 20.1% and 17.9% for the two sample sizes, respectively. We can see that for MAR_1, the addition of the second response completely cancels out the increase in the variance when ρ_2 is 0.4 or higher. The addition of the third response offers further reductions in the variance. For the MAR_2 mechanism we can see that the addition of the second and third response results in important reduction is the variances, for both sample sizes and for correlations ρ_2 that are 0.4 or higher.

5 Application

The dataset that we analyse here is a random subsample of $n = 500$ subjects from the Paquid prospective cohort study[33] and it is available in the R package `lcmm`.[34] The Paquid study aimed at investigating cerebral and functional aging in individuals aged 65 years and older, and it involved individuals living in south-western France. Longitudinal observations over a maximum period of 20 years were collected on 3 cognitive tests and depressive symptomatology. The cognitive tests evaluated the global mental status using the Mini Mental State Examination (y_1), verbal fluency using the Isaacs Set Test (y_2), and visual memory using the Benton's Visual Retention Test (y_3). Depressive symptomatology was measured by the Center for Epidemiological Study Depression scale (y_4). We note that for the first 3 responses higher scores indicate better cognitive function performance, while for the fourth response, higher scores indicate more symptoms. The available covariates are a binary gender indicator, $x_1 = 1$ for males and $x_1 = 0$ for females, a binary indicator of educational level, $x_2 = 1$ for individuals who graduated from primary school and $x_2 = 0$ otherwise, age at entry in the cohort x_3, and time of observation t.

Because the distribution of y_1 is very asymmetric, we analyse a normalized version of it.[35] Certain power transformations are needed on the other 3 responses in order for the normality assumption to appear plausible. Furthermore, to avoid computational problems, we scale the time of observation, $t = 0, 1, 2, \ldots, 20$, by dividing by 20, hence, $t = 0, 0.05, 0.1, \ldots, 1$. Further, we centre and scale the age at entry x_3 by subtracting the minimum age and dividing by 15. Lastly, we note that the dataset is highly unbalanced, with individuals observed at no more than 9 time points of the possible 21. The number of individuals observed at each time point decreases quickly: there are about 300 observations at $t = 0.1$, about 200 at $t = 0.2$, about 150 between $t = 0.3$ and $t = 0.4$, and thereafter the number of observations stays well below 100, with only 18 individuals observed at the last time point.

We model the mean using

$$\mu_{ijk} = \beta_{k0} + \beta_{k1}x_{i1} + \beta_{k2}x_{i2} + f_{\mu,k,3}(x_{3i}) + f_{\mu,k,4}(t_{ij}), k = 1, \ldots, 4, \tag{40}$$

where $f_{\mu,k,3}(x_{3i})$ and $f_{\mu,k,4}(t_{ij})$ are modelled using up to 10 and 8 knots, equivalently 11 and 9 basis functions, respectively. These knots are chosen as the unique quantiles of x_3 and t that correspond to the 10 equally spaced probabilities between 0 and 1. Although for x_3 there are 10 unique quantiles, for t there are only 8, hence the difference in the number of knots for modelling these 2 functions. Thus, for modelling the means there are in total 92 parameters, of which 88 are subject to selection.

Further, the 84×84 covariance matrix includes 3570 unique elements. These are modelled using 4 regression models that we describe next. First, the autoregressive coefficients are modelled using

$$\phi_{ijklm} = \psi_{lm0} + f_{\phi,l,m}(t_{ij} - t_{ik}), l, m = 1, \ldots, 4, \tag{41}$$

where $f_{\phi,l,m}(t_{ij} - t_{ik})$ is modelled using up to 10 knots. These are selected as the unique quantiles of lag, $t_{ij} - t_{ik}$, that correspond to the 10 equally spaced probabilities between 0 and 1. Thus, there are up to 192 parameters for modelling the autoregressive coefficients and all of these are subject to selection.

The innovation variances are modelled as

$$\log \sigma^2_{ijk} = \alpha_{k0} + f_{\sigma,k,1}(x_{3i}) + f_{\sigma,k,2}(t_{ij}), k = 1, \ldots, 4, \tag{42}$$

where $f_{\sigma,k,1}(x_{3i})$ and $f_{\sigma,k,2}(t_{ij})$ are modelled using up to 10 and 8 knots, just as was done for the mean functions. Hence, there are up to 84 parameters for modelling the innovation variances, of which 4 are always in the model, while the other 80 are subject to selection.

Lastly, we describe the model for the correlation matrices. Since the 4 responses can be naturally clustered into the group of 3 variables that measure cognitive function and the group of 1 variable that measures depressive symptomatology, a grouped variables model is a priori justifiable. Despite that, we prefer the less structured grouped correlations model, and let the data decide if a grouped variables model is preferable. The parameters of the correlation matrices are modelled using $\mu_{cth} = \eta_{h0} + f_{\mu h}(t_{ij})$ and $\log \sigma^2_{ct} = \omega_0 + f_\sigma(t_{ij})$. As the 4×4 correlation matrices have 6 unique elements, we allow up to $H = 6$ different trajectories for the means $\mu_{cth}, h = 1, \ldots, 6$. Functions $f_{\mu h}(t_{ij})$ and $f_\sigma(t_{ij})$ are modelled using up to 10 knots. These are selected as the unique quantiles of t that correspond to the 10 equally spaced probabilities between 0 and 1. Although for modelling innovation variances there are 8 unique quantiles, here there are 10, as there is a single correlation matrix observed at each time point. Hence, there are up to 72 parameters for modelling μ_{cth}, of which 66 are subject to selection. For modelling $\log \sigma^2_{ct}$ there are 12 parameters, of which 11 are subject to selection. This brings the number of parameters for modelling the covariance matrix to 360, 10 times smaller than the number of unique elements in the covariance matrix.

Results presented below are based on 25×10^3 posterior samples that are obtained from 10^5 iterations of the MCMC sampler, with the first 5×10^4 discarded as burn-in, and of the remaining keeping every second sample. Convergence of the MCMC algorithm was diagnosed by running two parallel chains with different initializations and computing the Gelman-Rubin diagnostic, as implemented in the R package coda.[36] It was observed that the 'potential scale reduction factor' had upper limits close one. Further, results from the two chains where indistinguishable.

Concerning the mean regression models, results in the form of posterior means and 80% credible intervals are presented in the Figure of the supplementary material. Covariate x_1 (gender) has little to no effect on the means of three cognitive function responses but it has an important effect on the mean of depressive symptomatology, indicating that males have, on average, less symptoms than females. The corresponding posterior probabilities of inclusion of x_1 in the mean models of the 4 responses are 3.3%, 23.4%, 33.2% and 100%, respectively. Further, covariate x_2 (education) has important effects on the means of the responses relating to cognitive function, with posterior probability of inclusion 100%, and with those who graduated from primary school having higher mean cognitive function scores. In addition, x_2 appears to have no effect on mean depressive symptomatology, with posterior probability of inclusion 2.9%. Covariates x_3 (age) and t (time) have negative linear (or almost linear) associations with the means of the 3 responses that relate to cognitive function. Furthermore, x_3 has a curved relationship with mean depressive symptomatology that is characterized by high uncertainty. Lastly, t has positive linear association with the mean of fourth response, indicating that the average depressive symptomatology increases with time. Visually, the covariates have simple relationships with the mean responses. This can also be confirmed by the small number of regression

coefficients that are selected to fit the mean functions: of the 88, on average, 16.5 or 18.8% were selected during MCMC sampling.

Results on the autoregressive coefficient regression models are shown in Figure 1. Recalling the discussion around (3), the fitted models presented in the diagonal of Figure 1 correspond to the coefficients for predicting responses based on passed observations on the same response. These 4 curves have similar shapes: they start from around 0.35 to 0.45 and they decrease towards 0. Based on univariate data analyses, Pan and MacKenzie[37] and Papageorgiou[38] reported similar curves. The remaining plots include a variety of fitted curves: complex and clearly important (such as ϕ_{jk12} and ϕ_{jk13}), complex but not very important (such as ϕ_{jk14} and ϕ_{jk32}) and flat at 0 (such as ϕ_{jk21}, ϕ_{jk31} and ϕ_{jk41}). For fitting these curves, of the 192 regression coefficients, on average, 62.7 or 32.7% were selected during MCMC sampling.

Results relating to innovation standard deviations are displayed in Figure 2. The first row shows almost flat fitted curves $\exp(f_{\sigma,1,1}(x_3)/2)$ and $\exp(f_{\sigma,4,1}(x_3)/2)$, and more complex fitted curves $\exp(f_{\sigma,2,1}(x_3)/2)$ and $\exp(f_{\sigma,3,1}(x_3)/2)$. The four fitted curves $\exp(f_{\sigma,k,2}(t)/2)$, shown in the second row of Figure 2, have similar shapes: they decrease as t increases and they are characterized by increasing uncertainty as t increases. For fitting these curves, of the 80 regression coefficients, on average, 32.8 or 41% were selected during MCMC sampling.

Figure 3 shows the fitted models of the parameters of the innovation correlation matrices. The first plot shows the clustering structure that the MCMC sampler visited most often: for 81.9% of the MCMC samples, two clusters were formed, the first being $\{(1,2),(1,3),(2,3)\}$ (denoted by the darker colour on the plot) while the second being $\{(1,4),(2,4),(3,4)\}$ (denoted by the lighter colour). While this clustering was obtained by a 'grouped correlations' model, it is also compatible with the 'grouped variables' model, the first cluster consisting of variables $1,2,3$ (the results on the cognitive tests) and the second consisting of variable 4 (the test result on depressive symptomatology). The other 2 plots in the Figure show the fitted curves for $\mu_{cth}, h = 1,2$, and σ_{ct}. The fitted curve μ_{t1} starts at around 0.25, decreases below 0.2 and then increases to just below 0.35. The fitted curve μ_{t2} is much simpler: it is constant at 0.1 for all time points. For fitting these curves, of the 66 regression coefficients, on average, 24.5 or 37.1% were selected during MCMC sampling. The fitted smooth curve for σ_{ct} is shown in the third plot of Figure 3. For fitting this curve, of the 11 regression coefficients, on average, 3.5 or 31.8% were selected during MCMC sampling.

Hence, overall, for fitting the 3570 unique elements of the 84×84 covariance matrix, the model selects, on average, 123.5 of the available 360 parameters, per MCMC iteration.

We conclude this section by providing an interpretation of the results in terms of the elements of the covariance matrix $\boldsymbol{\Sigma}$ of the 4 responses. To do so, we construct two covariance matrices based on the sampled parameter values, both for time points $t = 0, 0.15, 0.30, 0.45, 0.60, 0.75, 0.90$ and for x_3 equal to the lower ($x_3 = 0.18$) and upper ($x_3 = 0.73$) quartiles of age. With these choices we can examine the effects of time t and age x_3 on the elements of $\boldsymbol{\Sigma}$. For every iteration of the MCMC sampler (that is not discarded), we first obtain the innovation variances from (42). Based on these and on the sampled correlation matrices $\boldsymbol{R}$, using (14), we reconstruct the innovation covariance matrices $\boldsymbol{D}$. Further, we obtain the autoregressive coefficients from (41), and based on these we construct the generalized autoregressive matrices $\boldsymbol{\Phi}$, which, in turn, give us the matrix $\boldsymbol{L}$ in (2). Lastly we rearrange (1) to obtain $\boldsymbol{\Sigma} = \boldsymbol{L}^{-1}\boldsymbol{D}(\boldsymbol{L}^{\top})^{-1}$. This process results in 25×10^3 realizations of the 2 covariance matrices that correspond to $x_3 = 0.18$ and $x_3 = 0.73$. Results are shown in Figures 4, 5 and 6.

First, Figure 4 shows the posterior means and 80% credible intervals for the standard deviations of Y_1 and Y_2 over time. The standard deviations of Y_3 and Y_4 have similar shapes as the ones for Y_2, and hence they are omitted. We can see that in all 4 plots the standard deviations first decrease and then increase with time, t. The first two plots display the standard deviation of Y_1 for $x_3 = 0.18$ and $x_3 = 0.73$, respectively. We can see that age has no effect on the standard deviation of Y_1. However, it does have an effect on the standard deviation of Y_2 as the curve in the fourth plot, when compared with that in the third

plot, is shifted downwards. Further, Figure 5 shows how the correlations among responses evolve over time. We can see that correlations do not change by much over time and that there is high uncertainty about the posterior means. Lastly, Figure 6 displays autocorrelations at lag 0.15 over time. These autocorrelations can be seen to be increasing over time.

6 Discussion

The article describes a Bayesian framework for the analysis of multivariate longitudinal Gaussian responses, with nonparametric models for the means, the elements of the generalized autoregressive matrices, the elements of the innovation variance matrices, and the 2 parameters of the innovation correlation matrices. The use of spike-slab priors is important here as it allows for automatic variable selection and function regularization, simultaneously in the 5 submodels. Further, it obviates the need for model selection based on information criteria, see e.g. Pan and Mackenzie,[39] and it allows for the uncertainty in the model choice to propagate in the model parameter estimates. In addition, the automatic variable selection allows each regression equation to have its own set of covariates. For example, looking back at (40), it may appear that the responses have the same mean model. However, the binary indicators for variable selection allow each response to have its own mean model. The same is true for the autoregressive coefficients and innovation variances.

The major assumption of the framework is that the errors have a multivariate Gaussian distribution. It is worthwhile to relax this assumption, and to that end, in future work, we will be exploring the use of the multivariate skew-normal distribution Azzalini.[40]

References

1. Hamilton JD. *Time Series Analysis*. Princeton, NJ: Princeton Univ. Press 1994.

2. Xu J, Mackenzie G. Modelling covariance structure in bivariate marginal models for longitudinal data. *Biometrika.* 2012;99:649-662.

3. Kim C, Zimmerman DL. Unconstrained models for the covariance structure of multivariate longitudinal data. *J Multivar Anal.* 2012;107:104-118.

4. Kohli P, Garcia TP, Pourahmadi M. Modeling the Cholesky factors of covariance matrices of multivariate longitudinal data. *J Multivar Anal.* 2016;145:87-100.

5. Lee K, Cho H, Kwak M-S, Jang EJ. Estimation of covariance matrix of multivariate longitudinal data using modified Choleksky and hypersphere decompositions. *Biometrics.* 2020;76:75-86.

6. Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika.* 1999;86:677-690.

7. Daniels MJ, Pourahmadi M. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika.* 2002;89:553-566.

8. Zhang W, Leng C. A moving average Cholesky factor model in covariance modelling for longitudinal data. *Biometrika.* 2012;99:141-150.

9. Feng S, Lian H, Xue L. A new nested Cholesky decomposition and estimation for the covariance matrix of bivariate longitudinal data. *Comput Stat Data Anal.* 2016;102:98-109.

10. Chiu TYM, Leonard T, Tsui K-W. The matrix-logarithmic covariance model. *J Am Stat Assoc.* 1996;91:198-210.

11. Pourahmadi M. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika.* 2000;87:425-435.

12. Anderson TW. Asymptotically efficient estimation of covariance matrices with linear structure. *Ann Stat.* 1973;1:135-141.

13. Barnard J, McCulloch R, Meng X-L. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Stat Sin.* 2000;10:1281-1311.

14. Pourahmadi M. Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance-correlation parameters. *Biometrika.* 2007;94:1006-1013.

15. Zhang W, Leng C, Tang CY. A joint modelling approach for longitudinal studies. *J Royal Stat Soc Ser B (Stat Methodol).* 2015;77:219-238.

16. Daniels MJ, Kass RE. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J Am Stat Assoc.* 1999;94:1254-1263.

17. Liechty JC, Liechty MW, Müller P. Bayesian correlation estimation. *Biometrika.* 2004;91:1-14.

18. George EI, McCulloch RE. Approaches for Bayesian variable selection. *Stat Sin.* 1997;7:339-373.

19. Chan D, Kohn R, Nott D, Kirby C. Locally adaptive semiparametric estimation of the mean and variance functions in regression models. *J Comput Graph Stat.* 2006;15:915-936.

20. Papageorgiou Georgios. BNSP: an R package for fitting Bayesian semiparametric regression models and variable selection. *R J.* 2018;10:526-548.

21. Papageorgiou G, Marshall BC. Bayesian semiparametric analysis of multivariate continuous responses, with variable selection. *J Comput Graph Stat.* 2020;29:896-909.

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

23. Roberts GO, Rosenthal JS. Optimal scaling for various Metropolis-Hastings algorithms. *Stat Sci.* 2001;16:351-367.

24. Papageorgiou G. *BNSP: Bayesian non- and semi-parametric model fitting.* 2020. R package version 2.1.5.

25. Leng C, Zhang W, Pan J. Semiparametric mean–covariance regression analysis for longitudinal data. *J Am Stat Assoc.* 2010;105:181-193.

26. Lin H, Pan J. Nonparametric estimation of mean and covariance structures for longitudinal data. *Can J Stat.* 2013;41:557-574.

27. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Ann Stat.* 1973;1:209-230.

28. Sethuraman J. A constructive definition of Dirichlet priors. *Stat Sin.* 1994;4:639-650.

29. Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g priors for Bayesian variable selection. *J Am Stat Assoc.* 2008;103:410-423.

30. Zhang X, Boscardin JW, Belin TR. Sampling Correlation Matrices in Bayesian Models With Correlated Latent Variables *J Comput Graph Stat.* 2006;15:880–896.

31. Liu X, Daniels MJ. A New Algorithm for Simulating a Correlation Matrix Based on Parameter Expansion and Reparameterization *J Comput Graph Stat.* 2006;15:897-914.

32. Zellner A. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J Am Stat Assoc.* 1962;57:348-368.

33. Letenneur L, Commenges D, Dartigues JF, Barberger-Gateau P. Incidence of dementia and Alzheimer's disease in elderly community residents of south-western France. *Int J Epidemiol.* 1994;23:1256-1261.

34. Proust-Lima C, Philipps V, Diakite A, Liquet B. *lcmm: extended mixed models using latent classes and latent processes.* 2020. R package version: 1.9.2.

35. Philipps V, Amieva H, Andrieu S, et al. Normalized mini-mental state examination for assessing cognitive change in population-based brain aging studies. *Neuroepidemiology.* 2014;43:15-25.

36. Plummer Martyn, Best Nicky, Cowles Kate, Vines Karen. CODA: Convergence Diagnosis and Output Analysis for MCMC *R News.* 2006;6:7–11.

37. Pan J, MacKenzie G. Regression models for covariance structures in longitudinal studies. *Stat Model.* 2006;6:43-57.

38. Papageorgiou G. Restricted maximum likelihood estimation of joint mean-covariance models. *Can J Stat.* 2012;40:225-242.

39. Pan J, Mackenzie G. On modelling mean-covariance structures in longitudinal studies. *Biometrika.* 2003;90:239-244.

40. Azzalini A, Dalla Valle A. The multivariate skew-normal distribution. *Biometrika.* 1996;83:715-726.

Table 1: Simulation study results: the entries of the table are the relative biases $100B_p/B_1, p = 2, 3$. Rows refer to the sample size $n = 100, 300$, columns to the correlation coefficient $\rho_2 = 0.2, 0.4, 0.6, 0.8$, and the 5 parts of the table to the mechanism and probabilities of missingness. Results are based on 30 replicate datasets per sample size by correlation combination.

Missingness 0%

$p = 2$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	93.79	85.81	76.42	67.07	100	96.37	87.48	76.00	66.49
300	95.79	89.59	81.43	71.70	300	96.79	90.03	82.50	73.78

MCAR 25%

$p = 2$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	95.69	88.12	77.99	67.12	100	95.04	84.01	74.29	64.48
300	93.27	86.67	79.45	70.32	300	92.43	85.44	78.51	70.58

MCAR 50%

$p = 2$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	96.39	91.55	84.72	75.56	100	100.80	94.8	87.24	79.93
300	94.53	87.70	78.89	68.02	300	94.71	86.6	78.04	66.83

MAR_1

$p = 2$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	93.95	87.91	79.57	67.94	100	91.04	83.41	73.95	65.39
300	97.18	93.49	86.46	78.17	300	97.20	93.64	85.53	79.78

MAR_2

$p = 2$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	99.99	87.57	85.42	73.08	100	92.60	85.61	80.02	73.48
300	100.87	91.59	83.28	71.36	300	102.86	87.43	77.64	70.27

Table 2: Simulation study results: the entries of the table are the relative variances $100V_p/V_1, p = 2, 3$. Rows refer to the sample size $n = 100, 300$, columns to the correlation coefficient $\rho_2 = 0.2, 0.4, 0.6, 0.8$, and the 5 parts of the table to the mechanism and probabilities of missingness. Results are based on 30 replicate datasets per sample size by correlation combination.

Missingness 0%									
$p = 2$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	99.06	95.74	90.36	81.64	100	97.90	93.10	85.67	75.57
300	98.77	95.59	89.47	79.73	300	97.42	92.19	84.91	74.13
MCAR 25%									
$p = 2$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	98.58	95.13	90.04	81.38	100	97.58	92.97	85.48	75.09
300	98.87	94.96	88.76	78.84	300	98.27	92.07	83.95	72.72
MCAR 50%									
$p = 2$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	98.35	95.13	90.22	80.74	100	97.40	92.56	85.41	75.22
300	98.94	95.25	89.37	80.18	300	98.13	93.21	85.07	73.65
MAR_1									
$p = 2$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	98.61	95.10	89.43	81.34	100	97.35	92.47	84.78	74.29
300	99.11	94.68	87.68	77.82	300	97.61	91.77	83.42	72.69
MAR_2									
$p = 2$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	97.53	94.86	88.89	78.23	100	96.67	91.96	83.74	71.62
300	98.58	94.49	87.21	77.63	300	97.36	92.39	83.18	72.33

Table 3: Simulation study results: the entries of the table are the relative biases $100B_{p,m}/B_{1,0}, p = 1, 2, 3$. Rows refer to the sample size $n = 100, 300$, columns to the correlation coefficient $\rho_2 = 0.2, 0.4, 0.6, 0.8$, and the 4 parts of the table to the MCAR and MAR mechanisms of missingness. Results are based on 30 replicate datasets per sample size by correlation combination.

MCAR 25%					MCAR 50%				
$p = 1$	0.2	0.4	0.6	0.8	$p = 1$	0.2	0.4	0.6	0.8
100	105.33	107.67	108.65	112.03	100	131.80	130.82	127.64	125.61
300	110.74	109.26	107.88	107.11	300	141.21	139.64	137.11	133.82
$p = 2$	0.2	0.4	0.6	0.8	$p = 2$	0.2	0.4	0.6	0.8
100	100.79	94.87	84.74	75.19	100	127.04	119.77	108.13	94.91
300	103.28	94.70	85.72	75.32	300	133.49	122.47	108.17	91.02
$p = 3$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	100.11	90.45	80.72	72.24	100	132.86	124.01	111.35	100.40
300	102.36	93.35	84.70	75.59	300	133.73	120.93	107.00	89.42
MAR_1					MAR_2				
$p = 1$	0.2	0.4	0.6	0.8	$p = 1$	0.2	0.4	0.6	0.8
100	119.64	116.94	114.38	118.34	100	146.04	165.84	175.16	162.06
300	128.10	128.65	135.69	129.04	300	138.22	144.98	147.67	121.19
$p = 2$	0.2	0.4	0.6	0.8	$p = 2$	0.2	0.4	0.6	0.8
100	112.40	102.80	91.01	80.39	100	146.03	145.23	149.62	118.44
300	124.48	120.28	117.32	100.87	300	139.42	132.79	122.97	106.12
$p = 3$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	108.92	97.53	84.59	77.38	100	135.24	141.97	140.15	119.09
300	124.51	120.47	116.06	102.94	300	142.17	126.76	114.65	105.16

Table 4: Simulation study results: the entries of the table are the relative variances $100V_{p,m}/V_{1,0}, p = 1, 2, 3$. Rows refer to the sample size $n = 100, 300$, columns to the correlation coefficient $\rho_2 = 0.2, 0.4, 0.6, 0.8$, and the 4 parts of the table to the MCAR and MAR mechanisms of missingness. Results are based on 30 replicate datasets per sample size by correlation combination.

MCAR 25%					MCAR 50%				
$p = 1$	0.2	0.4	0.6	0.8	$p = 1$	0.2	0.4	0.6	0.8
100	104.35	104.54	104.41	104.98	100	113.97	114.17	113.45	113.60
300	100.79	101.19	101.36	101.11	300	111.57	112.18	112.77	112.45
$p = 2$	0.2	0.4	0.6	0.8	$p = 2$	0.2	0.4	0.6	0.8
100	102.87	99.45	94.00	85.43	100	112.09	108.60	102.36	91.71
300	99.65	96.09	89.97	79.71	300	110.39	106.85	100.79	90.16
$p = 3$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	101.82	97.19	89.25	78.83	100	111.01	105.68	96.90	85.44
300	99.05	93.17	85.10	73.53	300	109.48	104.57	95.93	82.81
MAR_1					MAR_2				
$p = 1$	0.2	0.4	0.6	0.8	$p = 1$	0.2	0.4	0.6	0.8
100	105.90	106.10	106.28	106.10	100	119.78	118.29	121.32	121.06
300	104.32	104.86	105.91	104.93	300	116.84	118.22	118.95	117.50
$p = 2$	0.2	0.4	0.6	0.8	$p = 2$	0.2	0.4	0.6	0.8
100	104.43	100.90	95.05	86.30	100	116.82	112.22	107.83	94.70
300	103.39	99.28	92.85	81.65	300	115.18	111.71	103.73	91.22
$p = 3$	0.2	0.4	0.6	0.8	$p = 3$	0.2	0.4	0.6	0.8
100	103.09	98.12	90.11	78.83	100	115.79	108.78	101.60	86.70
300	101.83	96.23	88.35	76.27	300	113.76	109.23	98.94	84.99

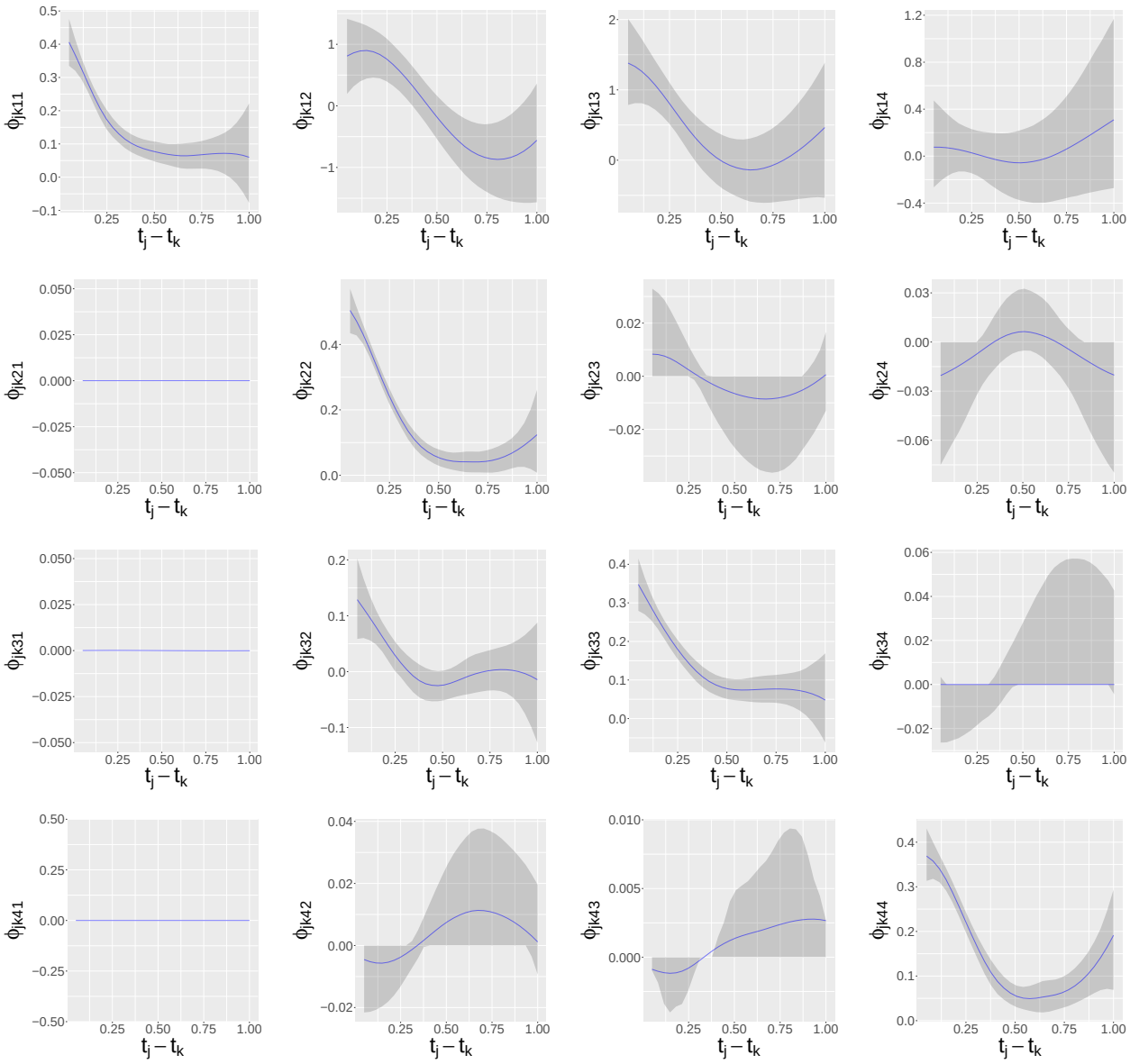

Figure 1: Application results: autoregressive coefficient regression models. Posterior means and 80% credible intervals.

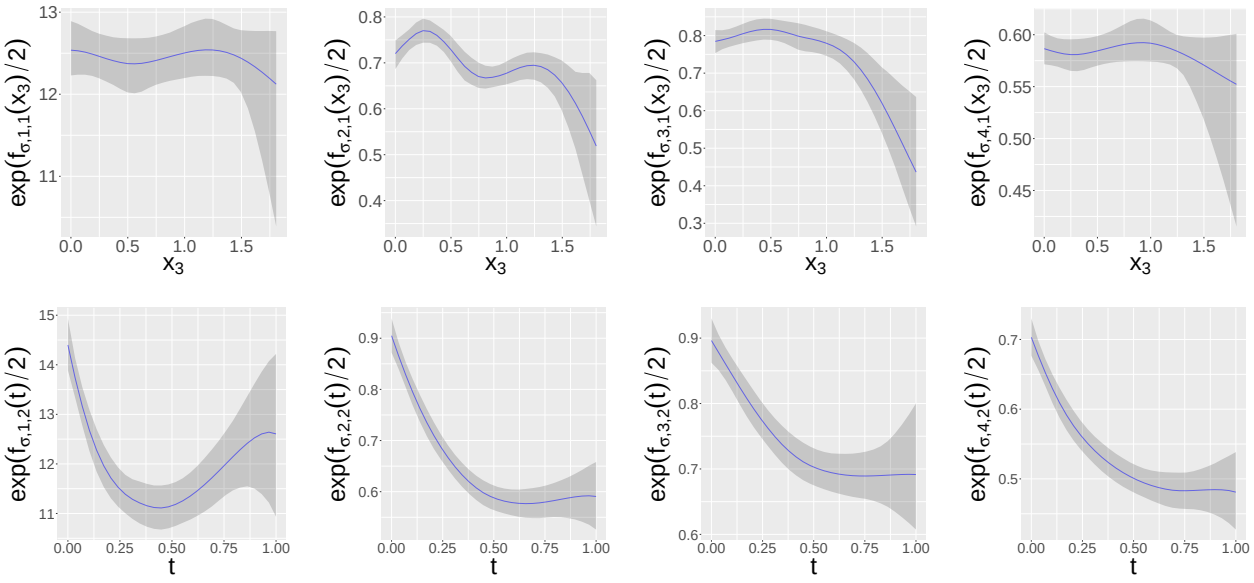

Figure 2: Application results: innovation standard deviation regression models. Posterior means and 80% credible intervals. Rows refer to the covariate effects and columns to the innovation standard deviations.

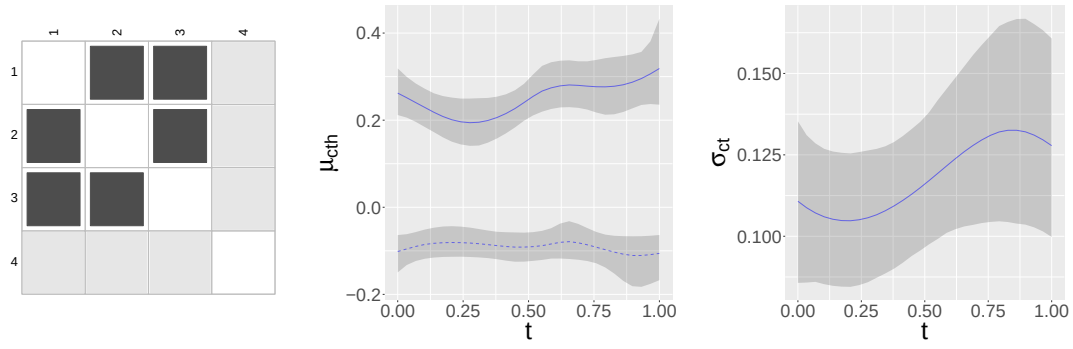

Figure 3: Application results: clustering structure of the elements of the 4×4 correlation matrices, posterior means and 80% credible intervals for $\mu_{cth}, h = 1, 2$, and posterior mean and 80% credible interval for σ_{ct}.

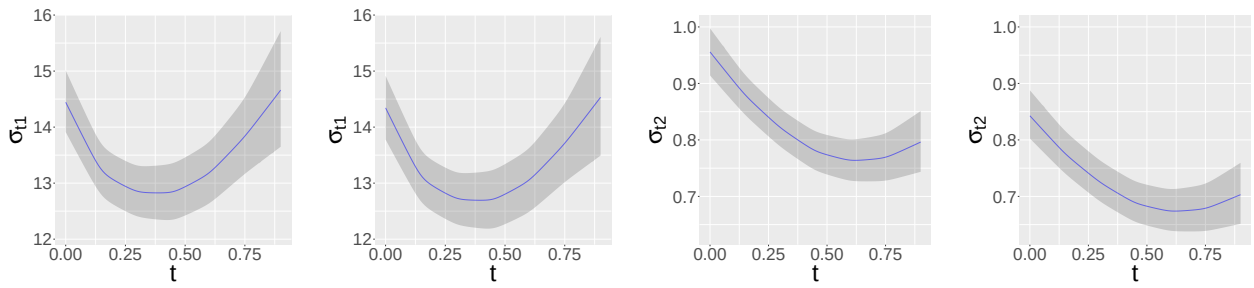

Figure 4: Application results: standard deviations of Y_1 and Y_2 over time, for ages $x_3 = 0.18$ and $x_3 = 0.73$, respectively. Posterior means and 80% credible intervals.

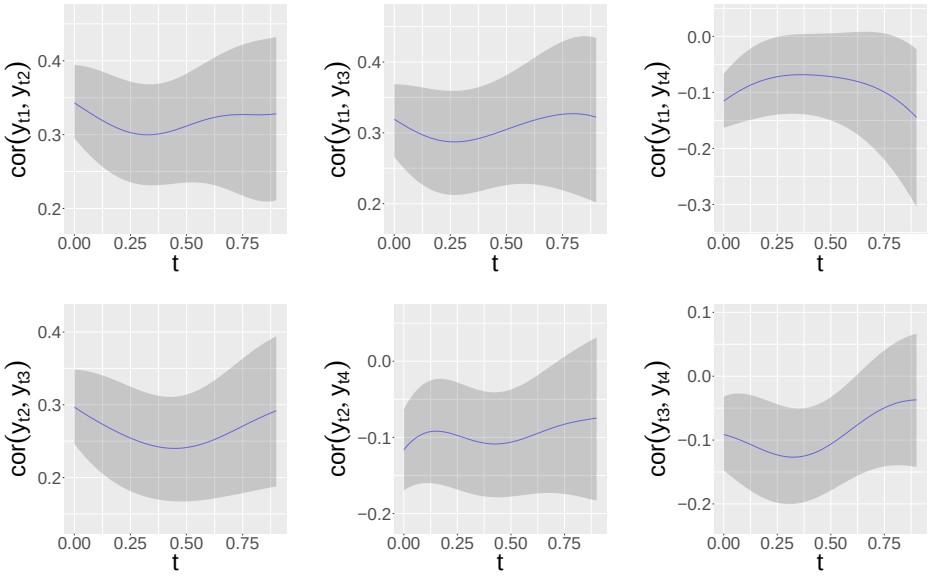

Figure 5: Application results: correlations over time. Posterior means and 80% credible intervals.

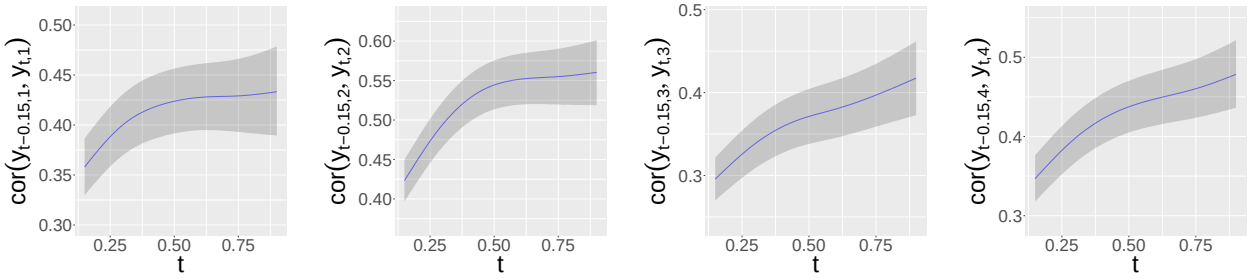

Figure 6: Application results: autocorrelations at lag 0.15 over time. Posterior means and 80% credible intervals.

Supplementary material of: 'Bayesian semiparametric modelling of covariance matrices for multivariate longitudinal data'

Georgios Papageorgiou
Department of Economics, Mathematics and Statistics
Birkbeck, University of London
g.papageorgiou@bbk.ac.uk

1 MCMC algorithm

Here we provide all details of the MCMC sampler of the three correlation models. We note that for all steps that involve tuning parameters, the values of these are chosen adaptively[1] to achieve an acceptance probability of $20\% - 25\%$.[2]

1.1 Common correlations model

Starting from the common correlations model, the algorithm proceeds as follows:

1. The elements of $\boldsymbol{\gamma}_{kl}, k = 1, \ldots, p, l = 1, \ldots, K$, are updated in random order and in blocks of random size.[3] Let $\boldsymbol{\gamma}_{Bkl}$ be a block of elements of $\boldsymbol{\gamma}_{kl}$. The proposed value for $\boldsymbol{\gamma}_{Bkl}$ is obtained from its prior with the remaining elements of $\boldsymbol{\gamma}_{kl}$, denoted by $\boldsymbol{\gamma}_{Ckl}$, kept at their current value. The proposal pmf is obtained from the binomial prior in (26) with $\pi_{\mu kl}$ integrated out

$$p(\boldsymbol{\gamma}_{Bkl}|\boldsymbol{\gamma}_{Ckl}) = \frac{p(\boldsymbol{\gamma}_{kl})}{p(\boldsymbol{\gamma}_{Ckl})} = \frac{\text{Beta}(c_{\mu kl} + N(\boldsymbol{\gamma}_{kl}), d_{\mu kl} + q_{\mu l} - N(\boldsymbol{\gamma}_{kl}))}{\text{Beta}(c_{\mu kl} + N(\boldsymbol{\gamma}_{Ckl}), d_{\mu kl} + q_{\mu l} - L(\boldsymbol{\gamma}_{Bkl}) - N(\boldsymbol{\gamma}_{Ckl}))},$$

   where $L(\boldsymbol{\gamma}_{Bkl})$ denotes the length of $\boldsymbol{\gamma}_{Bkl}$ i.e. the size of the block. For this proposal pmf, the acceptance probability of the Metropolis-Hastings move reduces to the ratio of the likelihoods in (32)

$$\min\left\{1, (c_\beta + 1)^{\{N(\gamma^C) - N(\gamma^P)\}/2} \exp\{(S^C - S^P)/2\}\right\},$$

   where superscripts P and C denote proposed and currents values respectively.

2. Parameter c_β is updated from the marginal likelihood (32) and the $\text{IG}(a_\beta, b_\beta)$ prior

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

   To sample from the above, we utilize a normal approximation. Let $\ell(c_\beta) = \log\{f(c_\beta|\ldots)\}$. 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 parameter. With superscripts P and C denoting proposed and currents values, the acceptance probability is the minimum between one and

$$\frac{f(c_\beta^P|\dots)}{f(c_\beta^C|\dots)}\frac{N(c_\beta^C;\hat{c}_\beta,-g^2/\ell''(\hat{c}_\beta))}{N(c_\beta^P;\hat{c}_\beta,-g^2/\ell''(\hat{c}_\beta))}.$$

3. Pairs $(\boldsymbol{\delta}_{kl},\boldsymbol{\alpha}_{kl}), k=1,\dots,p, l=1,\dots,L$, are updated simultaneously. Similarly to the updating of $\boldsymbol{\gamma}_{kl}$, the elements of $\boldsymbol{\delta}_{kl}$ are updated in random order and in blocks of random size. Let $\boldsymbol{\delta}_{Bkl}$ denote a block. Blocks $\boldsymbol{\delta}_{Bkl}$ and the whole vector $\boldsymbol{\alpha}_{kl}$ are generated simultaneously. As was mentioned by Chan et al,[3] generating the whole vector $\boldsymbol{\alpha}_{kl}$, instead of subvector $\boldsymbol{\alpha}_{Bkl}$, is necessary in order to make $\boldsymbol{\alpha}_{kl}$ consistent with the proposed value of $\boldsymbol{\delta}_{kl}$.

   Generating the proposed value for $\boldsymbol{\delta}_{Bkl}$ is done in a similar way as was done for $\boldsymbol{\gamma}_{Bkl}$. Let $\boldsymbol{\delta}_{kl}^P$ denote the proposed value of $\boldsymbol{\delta}_{kl}$. Next, we describe how the proposed vale for $\boldsymbol{\alpha}_{\delta_{kl}^P}$ is obtained. To avoid clutter, proposed values $\boldsymbol{\alpha}_{\delta_{kl}^P}^P$ will be denoted by the simpler $\boldsymbol{\alpha}_{kl}^P$. The development that follows is in the spirit of Chan et al[3] who built on the work of Gamerman.[4]

   Let $\hat{\boldsymbol{\beta}}_\gamma^C=\{c_\beta/(1+c_\beta)\}(\tilde{\boldsymbol{X}}_\gamma^\top\tilde{\boldsymbol{X}}_\gamma)^{-1}\tilde{\boldsymbol{X}}_\gamma^\top\tilde{\boldsymbol{y}}$ denote the current value of the posterior mean of $\boldsymbol{\beta}_\gamma^*$. Define the current squared residuals

$$e_{ijk}^C=(y_{ijk}-(\boldsymbol{x}_{ijk}^*)^\top\hat{\boldsymbol{\beta}}_{\gamma_k}^C)^2.$$

   These have an approximate $\sigma_{ijk}^2\chi_1^2$ distribution, where $\sigma_{ijk}^2=\sigma_k^2\exp(\boldsymbol{w}_{ijk}^\top\boldsymbol{\alpha}_{\delta_k})$. The latter defines a Gamma generalized linear model (GLM) for the squared residuals with mean σ_{ijk}^2, which, utilizing a log-link, can be thought of as Gamma GLM with an offset term: $\log(\sigma_{ijk}^2)=\log(\sigma_k^2)+\boldsymbol{w}_{ijk}^\top\boldsymbol{\alpha}_{\delta_k}$. Given $\boldsymbol{\delta}_{kl}^P$, the proposal density for $\boldsymbol{\alpha}_{kl}^P$ is derived utilizing the one step iteratively reweighted least squares algorithm. This proceeds as follows. First define the transformed observations

$$d_{ijk}^C(\boldsymbol{\alpha}_k^C)=\log(\sigma_k^2)+\boldsymbol{w}_{ij}^\top\boldsymbol{\alpha}_k^C+\frac{e_{ijk}^C-(\sigma_{ijk}^2)^C}{(\sigma_{ijk}^2)^C},$$

   where superscript C denotes current values.

   Further, let $\boldsymbol{\sigma}_k^2=(\sigma_{11k}^2,\dots,\sigma_{nn_nk}^2)^\top$. Now, the model (16) for $\boldsymbol{\sigma}_k^2$ can be expressed as

$$\boldsymbol{\sigma}_k^2=\sigma_k^2\exp(\boldsymbol{W}_{\delta_k}\boldsymbol{\alpha}_{\delta_k}),$$

   where $\boldsymbol{W}_{\delta_k}=[\boldsymbol{w}_{11k},\dots,\boldsymbol{w}_{nn_nk}]^\top$.

   Next we define

$$\boldsymbol{\Delta}(\boldsymbol{\delta}_{kl}^P)=(c_{\alpha k}^{-1}\boldsymbol{I}+\boldsymbol{W}_{\delta_{kl}^P}^\top\boldsymbol{W}_{\delta_{kl}^P})^{-1}\text{ and }\hat{\boldsymbol{\alpha}}(\boldsymbol{\delta}_{kl}^P,\boldsymbol{\alpha}_k^C)=\boldsymbol{\Delta}_{\delta_{kl}^P}\boldsymbol{W}_{\delta_{kl}^P}^\top\boldsymbol{d}_k^C,$$

   where $\boldsymbol{W}_{\delta_{kl}}$ is a submatrix of $\boldsymbol{W}_{\delta_k}$ that considers only the columns that pertain to the lth effect, and $\boldsymbol{d}_k^C$ denotes the vector of d_{ijk}^C. The proposed value $\boldsymbol{\alpha}_{kl}^P$ is obtained from the multivariate normal distribution $N(\boldsymbol{\alpha}_{kl}^P;\hat{\boldsymbol{\alpha}}(\boldsymbol{\delta}_{kl}^P,\boldsymbol{\alpha}_k^C),h_{kl}\boldsymbol{\Delta}(\boldsymbol{\delta}_{kl}^P))$, where h_{kl} is a tuning parameter.

   Let $N(\boldsymbol{\alpha}_{kl}^C;\hat{\boldsymbol{\alpha}}(\boldsymbol{\delta}_{kl}^C,\boldsymbol{\alpha}_k^P),h_{kl}\boldsymbol{\Delta}(\boldsymbol{\delta}_{kl}^C))$ denote the proposal density for taking a step in the reverse direction, from model $\boldsymbol{\delta}_{kl}^P$ to $\boldsymbol{\delta}_{kl}^C$. Then the acceptance probability of the pair $(\boldsymbol{\delta}_{kl}^P,\boldsymbol{\alpha}_{kl}^P)$ is

$$\min\left\{1,\frac{|\boldsymbol{\Sigma}^P|^{-\frac{1}{2}}\exp\{-S^P/2\}}{|\boldsymbol{\Sigma}^C|^{-\frac{1}{2}}\exp\{-S^C/2\}}\frac{(2\pi c_{\alpha k})^{-\frac{N(\delta_{kl}^P)}{2}}\exp\{-\frac{1}{2c_{\alpha k}}(\boldsymbol{\alpha}_{kl}^P)^\top\boldsymbol{\alpha}_{kl}^P\}}{(2\pi c_{\alpha k})^{-\frac{N(\delta_{kl}^C)}{2}}\exp\{-\frac{1}{2c_{\alpha k}}(\boldsymbol{\alpha}_{kl}^C)^\top\boldsymbol{\alpha}_{kl}^C\}}\frac{N(\boldsymbol{\alpha}_{kl}^C;\hat{\boldsymbol{\alpha}}_{\delta_{kl}^C},h_{kl}\boldsymbol{\Delta}_{\delta_{kl}^C})}{N(\boldsymbol{\alpha}_{kl}^P;\hat{\boldsymbol{\alpha}}_{\delta_{kl}^P},h_{kl}\boldsymbol{\Delta}_{\delta_{kl}^P})}\right\},$$

   where the determinants, for centred variables, are equal to one, otherwise, the ratio of the determinants may be computed as $\prod_{i=1}^n\prod_{j=1}^{n_i}\{(\sigma_{ijk}^2)^C/(\sigma_{ijk}^2)^P\}^{1/2}$.

4. The full conditional of $\sigma_k^2, k = 1, \ldots, p$, is given by

$$f(\sigma_k^2|\ldots) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp(-S/2)\xi(\sigma_k^2),$$

where $\xi(\sigma_k^2)$ denotes either the IG or half-normal prior. We follow a random walk algorithm obtaining proposed values $(\sigma_k^2)^{(P)} \sim N((\sigma_k^2)^{(C)}, v_k^2)$, where v_k^2 is a tuning parameter. Proposed values are accepted with probability $f((\sigma_k^2)^{(P)}|\ldots)/f((\sigma_k^2)^{(C)}|\ldots)$, which reduces to

$$\{(\sigma_k^2)^C/(\sigma_k^2)^P\}^{N/2} \exp\{(S^C - S^P)/2\})\xi((\sigma_k^2)^{(P)})/\xi((\sigma_k^2)^{(C)}).$$

5. Concerning parameter $c_{\alpha k}, k = 1, \ldots, p$, the full conditional that corresponds to the IG$(a_{\alpha k}, b_{\alpha k})$ prior is another inverse Gamma density IG$(a_{\alpha k} + N(\delta_k)/2, b_{\alpha k} + \boldsymbol{\alpha}_{\delta_k}^\top \boldsymbol{\alpha}_{\delta_k}/2)$.

   The full conditional that corresponds to the half-normal prior $\sqrt{c_{\alpha k}} \sim N(0, \phi_{\alpha k}^2) I[\sqrt{c_{\alpha k}} > 0]$ is

$$f(c_{\alpha k}|\ldots) \propto c_{\alpha k}^{-N(\delta_k)/2} \exp\{-\boldsymbol{\alpha}_{\delta_k}^\top \boldsymbol{\alpha}_{\delta_k}/2c_{\alpha k}\} \exp\{-c_{\alpha k}/2\phi_{\alpha k}^2\} I[\sqrt{c_{\alpha k}} > 0].$$

   We obtain proposed values from $c_{\alpha k}^{(P)} \sim N(c_{\alpha k}^{(C)}, v_k^2)$, where $c_{\alpha k}^{(C)}$ denotes the current value and v_k^2 denotes a tuning parameter. Proposed values are accepted with probability $f(c_{\alpha k}^{(P)}|\ldots)/f(c_{\alpha k}^{(C)}|\ldots)$.

6. Samples from the posterior of $\boldsymbol{\beta}_\gamma^*$ are generated from

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

   which is obtained using the first version of Q, just above (29), and prior (25).

7. Pairs $(\boldsymbol{\xi}_{lmb}, \boldsymbol{\psi}_{lmb}), l, m = 1, \ldots, p, b = 0, \ldots, B$, are updated simultaneously. Let $\boldsymbol{\xi}_{Blmb}$ denote a randomly chosen block of $\boldsymbol{\xi}_{lmb}$. The block $\boldsymbol{\xi}_{Blmb}$ and the whole of $\boldsymbol{\psi}_{lmb}$ are updated simultaneously. Proposed values for $\boldsymbol{\xi}_{Blmb}$ are obtained from its prior, in a similar way as was done for $\boldsymbol{\gamma}_{Bkl}$. Now, to obtain a proposal for $\boldsymbol{\psi}_{lmb}$, note that the full conditional of $\boldsymbol{\psi}_\xi^*$, as obtained from its normal prior and the likelihood that corresponds to (31), is

$$\boldsymbol{\psi}_\xi^*|\cdots \sim N\left[\boldsymbol{A}\left(\sum_{i=1}^n \sum_{j=1}^{n_i} \boldsymbol{V}_{ij\xi\beta}^\top \boldsymbol{D}_{ij}^{-1} \boldsymbol{r}_{ij}\right), \boldsymbol{A} \equiv \left(\boldsymbol{D}_{c\psi}^{-1} + \sum_{i=1}^n \sum_{j=1}^{n_i} \boldsymbol{V}_{ij\xi\beta}^\top \boldsymbol{D}_{ij}^{-1} \boldsymbol{V}_{ij\xi\beta}\right)^{-1}\right],$$

   where $\boldsymbol{D}_{c\psi}$ is a diagonal matrix with diagonal elements equal to $c_{\psi lm}$, each having multiplicity equal to the length of $\boldsymbol{\psi}_{\xi lm}^*$. We use as proposal the conditional of $\boldsymbol{\psi}_{lmb}$ given all other elements of $\boldsymbol{\psi}_\xi^*$, denoted by $N(\boldsymbol{\psi}_{lmb}|\boldsymbol{\psi}_{Clmb})$. Then, the acceptance probability of the pair $(\boldsymbol{\xi}_{lmb}^P, \boldsymbol{\psi}_{lmb}^P)$ is

$$\min\left\{1, \frac{\exp(-S^P/2)}{\exp(-S^C/2)} \frac{N(\boldsymbol{\psi}_{lmb}^P|\boldsymbol{0}, c_{\psi lm}\boldsymbol{I})}{N(\boldsymbol{\psi}_{lmb}^C|\boldsymbol{0}, c_{\psi lm}\boldsymbol{I})} \frac{N(\boldsymbol{\psi}_{lmb}^C|\boldsymbol{\psi}_{Clmb})}{N(\boldsymbol{\psi}_{lmb}^P|\boldsymbol{\psi}_{Clmb})}\right\}.$$

8. For parameter c_ψ the full conditional that corresponds to the IG(a_ψ, b_ψ) prior is another inverse Gamma density, IG$(a_\psi + N(\xi)/2, b_\psi + (\boldsymbol{\psi}_\xi^*)^\top \boldsymbol{\psi}_\xi^*/2)$.

   The full conditional that corresponds to the half-normal prior $\sqrt{c_\psi} \sim N(0, \phi_\psi^2) I[\sqrt{c_\psi} > 0]$ is

$$f(c_\psi|\ldots) \propto c_\psi^{-N(\xi)/2} \exp[-(\boldsymbol{\psi}_\xi^*)^\top \boldsymbol{\psi}_\xi^*/(2c_\psi)] \exp(-c_\psi/2\phi_\psi^2) I(\sqrt{c_\psi} > 0).$$

   We obtain proposed values from $c_\psi^{(P)} \sim N(c_\psi^{(C)}, s_\psi^2)$, where $c_\psi^{(C)}$ is the current value and s_ψ^2 is a tuning parameter. Proposed values are accepted with probability $f(c_\psi^{(P)}|\ldots)/f(c_\psi^{(C)}|\ldots)$.

9. Sampling from the full conditional of $\boldsymbol{R}_t, t \in T$, is described in detail in the main body of the paper.

10. To sample from the full conditional of $\boldsymbol{\theta}_t, t \in T$, first write $f(\boldsymbol{r}|\boldsymbol{\theta}, \tau^2) = \prod_{t\in T} \pi(\boldsymbol{\theta}_t, \tau^2) N(g(\boldsymbol{r}); \boldsymbol{\theta}, \tau^2 \boldsymbol{I})$ for the likelihood of (33). Further, the prior for $\boldsymbol{\theta}$, as derived from (39), is $\boldsymbol{\theta} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\theta)$ where $\boldsymbol{\Sigma}_\theta^{-1} = \sigma^{-2}\boldsymbol{D}^{-1}(\boldsymbol{\omega}_\varphi)\left\{\boldsymbol{I} - \frac{c_\eta}{1+c_\eta}\tilde{\boldsymbol{Z}}_\nu(\tilde{\boldsymbol{Z}}_\nu^\top\tilde{\boldsymbol{Z}}_\nu)^{-1}\tilde{\boldsymbol{Z}}_\nu^\top\right\}\boldsymbol{D}^{-1}(\boldsymbol{\omega}_\varphi)$. Hence, it is easy to show that the posterior is

$$f(\boldsymbol{\theta}|\dots) = \left\{\prod_{t\in T}\pi(\boldsymbol{\theta}_t, \tau^2)\right\} N\left(\boldsymbol{\theta}; \tau^{-2}\boldsymbol{A}g(\boldsymbol{r}), \boldsymbol{A} \equiv (\tau^{-2}\boldsymbol{I} + \boldsymbol{\Sigma}_\theta^{-1})^{-1}\right). \tag{43}$$

At iteration $u+1$ we sample $\boldsymbol{\theta}^{(u+1)}$ utilizing as proposal the normal distribution that appears on the right of (43), ignoring the normalizing constants. The proposed $\boldsymbol{\theta}^{(u+1)}$ is accepted with probability

$$\min\left\{1, \frac{\prod_{t\in T}\pi(\boldsymbol{\theta}_t^{(u+1)}, \tau^2)}{\prod_{t\in T}\pi(\boldsymbol{\theta}_t^{(u)}, \tau^2)}\right\}.$$

This, as was argued by Liechty et al,[5;6] for a small value of τ^2, can reasonably be assumed to be unity.

11. The elements of $\boldsymbol{\nu}$ are updated in random order and in blocks of random size. Let $\boldsymbol{\nu}_B$ be such a block. The proposed value for $\boldsymbol{\nu}_B$ is obtained from its prior with the remaining elements of $\boldsymbol{\nu}$, denoted by $\boldsymbol{\nu}_{B^C}$, kept at their current value, in a similar way as was done for $\boldsymbol{\gamma}_{Bkl}$. The acceptance probability reduces to the ratio of the likelihoods in (39)

$$\min\left\{1, \frac{(c_\eta+1)^{-\frac{N(\nu^P)+1}{2}}\exp\{-(S^*)^P/2\sigma^2\}}{(c_\eta+1)^{-\frac{N(\nu^C)+1}{2}}\exp\{-(S^*)^C/2\sigma^2\}}\right\},$$

where superscripts P and C denote proposed and currents values respectively.

12. Vectors $\boldsymbol{\omega}$ and $\boldsymbol{\varphi}$ are updated following a similar approach as that for step 3. The elements of $\boldsymbol{\varphi}$ are updated in random order and in blocks of random size. Let $\boldsymbol{\varphi}_B$ denote a block. Blocks $\boldsymbol{\varphi}_B$ and the whole vector $\boldsymbol{\omega}$ are generated simultaneously. Generating the proposed value for $\boldsymbol{\varphi}_B$ is done in a similar way as was done for $\boldsymbol{\gamma}_{Bkl}$. Let $\boldsymbol{\varphi}^P$ denote the proposed value of $\boldsymbol{\varphi}$. Next, we describe how the proposed vale for $\boldsymbol{\omega}_{\varphi^P}$ is obtained. Let $\hat{\boldsymbol{\eta}}_\nu^C = \{c_\eta/(1+c_\eta)\}(\tilde{\boldsymbol{Z}}_\nu^\top\tilde{\boldsymbol{Z}}_\nu)^{-1}\tilde{\boldsymbol{Z}}_\nu^\top\tilde{\boldsymbol{\theta}}$ denote the current value of the posterior mean of $\boldsymbol{\eta}_\nu$. Define the current squared residuals

$$e_{tkl}^C = (\theta_{tkl} - (\boldsymbol{z}_{t\nu}^*)^\top\hat{\boldsymbol{\eta}}_\nu^C)^2,$$

$t \in T, k = 1, \dots, p-1, l > k$. These will have an approximate $\sigma_t^2\chi_1^2$ distribution, where $\sigma_t^2 = \sigma^2\exp(\boldsymbol{z}_t^\top\boldsymbol{\omega})$. The latter defines a Gamma generalized linear model (GLM) for the squared residuals with mean $E(\sigma_t^2\chi_1^2) = \sigma_t^2 = \sigma^2\exp(\boldsymbol{z}_t^\top\boldsymbol{\omega})$, which, utilizing a log-link, can be thought of as Gamma GLM with an offset term: $\log(\sigma_t^2) = \log(\sigma^2) + \boldsymbol{z}_t^\top\boldsymbol{\omega}$. Given the proposed value of $\boldsymbol{\varphi}$, denoted by $\boldsymbol{\varphi}^P$, the proposal density for $\boldsymbol{\omega}_{\varphi^P}$ is derived utilizing the one step iteratively reweighted least squares algorithm. This proceeds as follows. First define the transformed observations

$$d_{tkl}^C(\boldsymbol{\omega}^C) = \log(\sigma^2) + \boldsymbol{z}_t^\top\boldsymbol{\omega}^C + \frac{e_{tkl}^C - (\sigma_t^2)^C}{(\sigma_t^2)^C},$$

where superscript C denotes current values. Further, let $\boldsymbol{d}^C$ denote the vector of d^C_{tkl}.

Next we define

$$\boldsymbol{\Delta}(\boldsymbol{\varphi}^P) = (c_\omega^{-1}\boldsymbol{I} + \boldsymbol{Z}^\top_{\varphi^P}\boldsymbol{Z}_{\varphi^P})^{-1} \text{ and} \hat{\boldsymbol{\omega}}(\boldsymbol{\varphi}^P, \boldsymbol{\omega}^C) = \boldsymbol{\Delta}_{\varphi^P}\boldsymbol{Z}^\top_{\varphi^P}\boldsymbol{d}^C,$$

where $\boldsymbol{Z}$ is the design matrix $\boldsymbol{Z}^*$ without the intercept column. The proposed value $\boldsymbol{\omega}^P_{\varphi^P}$ is obtained from a multivariate normal distribution with mean $\hat{\boldsymbol{\omega}}(\boldsymbol{\varphi}^P, \boldsymbol{\omega}^C)$ and covariance $h\boldsymbol{\Delta}(\boldsymbol{\omega}^P)$, denoted as $N(\boldsymbol{\omega}^P_{\varphi^P};\hat{\boldsymbol{\omega}}(\boldsymbol{\varphi}^P, \boldsymbol{\omega}^C), h\boldsymbol{\Delta}(\boldsymbol{\varphi}^P))$, where h is a tuning parameter.

Let $N(\boldsymbol{\omega}^C_{\varphi^C};\hat{\boldsymbol{\omega}}(\boldsymbol{\varphi}^C, \boldsymbol{\omega}^P), h\boldsymbol{\Delta}(\boldsymbol{\varphi}^C))$ denote the proposal density for taking a step in the reverse direction, from model $\boldsymbol{\varphi}^P$ to $\boldsymbol{\varphi}^C$. Then the acceptance probability of the pair $(\boldsymbol{\varphi}^P, \boldsymbol{\omega}^P_{\varphi^P})$ is

$$\min\left\{1, \frac{|D^2(\boldsymbol{\omega}^P_{\varphi^P})|^{-\frac{1}{2}}\exp\{-S^P/2\sigma^2\}}{|D^2(\boldsymbol{\omega}^C_{\varphi^C})|^{-\frac{1}{2}}\exp\{-S^C/2\sigma^2\}} \frac{(2\pi c_\omega)^{-\frac{N(\varphi^P)}{2}}\exp\{-\frac{1}{2c_\omega}(\boldsymbol{\omega}^P_{\varphi^P})^\top\boldsymbol{\omega}^P_{\varphi^P}\}}{(2\pi c_\omega)^{-\frac{N(\varphi^C)}{2}}\exp\{-\frac{1}{2c_\omega}(\boldsymbol{\omega}^C_{\varphi^C})^\top\boldsymbol{\omega}^C_{\varphi^C}\}} \frac{N(\boldsymbol{\omega}^C_{\varphi^C};\hat{\boldsymbol{\alpha}}_{\varphi^C}, h\boldsymbol{\Delta}_{\varphi^C})}{N(\boldsymbol{\omega}^P_{\varphi^P};\hat{\boldsymbol{\alpha}}_{\varphi^P}, h\boldsymbol{\Delta}_{\varphi^P})}\right\}.$$

13. We update σ^2 utilizing the marginal (39) and the two prior specifications. The full conditional that corresponds to the IG(a_σ, b_σ) prior is

$$f(\sigma^2|\dots) \propto (\sigma^2)^{-\frac{Md}{2}-a_\sigma-1}\exp\{-(S^*/2 + b_\sigma)/\sigma^2\},$$

where M denotes the number of unique observational times and $d = p(p-1)/2$. The above is recognized as IG$(Md/2 + a_\sigma, S^*/2 + b_\sigma)$.

The full conditional that corresponds to the half-normal prior $\sigma \sim N(\sigma; 0, \phi^2_\sigma)I[\sigma > 0]$ is

$$f(\sigma^2|\dots) \propto (\sigma^2)^{-\frac{Md}{2}}\exp\{-S^*/(2\sigma^2)\}\exp\{-\sigma^2/(2\phi^2_\sigma)\}I[\sigma > 0].$$

Proposed values are obtained from $\sigma^2_p \sim N(\sigma^2_c, f_1^2)$, where σ^2_c denotes the current value and f_1^2 is a tuning parameter. Proposed values are accepted with probability $f(\sigma^2_p|\dots)/f(\sigma^2_c|\dots)$.

14. Parameter c_η is updated from the marginal (39) and the IG(a_η, b_η) prior

$$f(c_\eta|\dots) \propto (c_\eta + 1)^{-\frac{N(\nu)+1}{2}}\exp\{-S^*/2\sigma^2\}(c_\eta)^{-a_\eta-1}\exp\{-b_\eta/c_\eta\}.$$

To sample from the above, we utilize a normal approximation to it. Let $\ell(c_\eta) = \log\{f(c_\eta|\dots)\}$. We utilize a normal proposal density $N(\hat{c}_\eta, -g^2/\ell''(\hat{c}_\eta))$ where $\hat{c}_\eta$ is the mode of $\ell(c_\eta)$, found using a Newton-Raphson algorithm, $\ell''(\hat{c}_\eta)$ is the second derivative of $\ell(c_\eta)$ evaluated at the mode, and g^2 is a tuning parameter. At iteration $u+1$ the acceptance probability is the minimum between one and

$$\frac{f(c_\eta^{(u+1)}|\dots)}{f(c_\eta^{(u)}|\dots)}\frac{N(c_\eta^{(u)};\hat{c}_\eta, -g^2/\ell''(\hat{c}_\eta))}{N(c_\eta^{(u+1)};\hat{c}_\eta, -g^2/\ell''(\hat{c}_\eta))}.$$

15. Concerning parameter c_ω, the full conditional that corresponds to the IG(a_ω, b_ω) prior is another inverse Gamma density IG$(a_\omega + N(\varphi)/2, b_\omega + \boldsymbol{\omega}^\top_\varphi\boldsymbol{\omega}_\varphi/2)$.

The full conditional that corresponds to the half-normal prior $\sqrt{c_\omega} \sim N(\sqrt{c_\omega}; 0, \phi^2_{c_\omega})I[\sqrt{c_\omega} > 0]$ is

$$f(c_\omega|\dots) \propto c_\omega^{-N(\varphi)/2}\exp\{-\boldsymbol{\omega}^\top_\varphi\boldsymbol{\omega}_\varphi/(2c_\omega)\}\exp\{-c_\omega/(2\phi^2_{c_\omega})\}I[\sqrt{c_\omega} > 0].$$

We obtain proposed values $c^P_\omega \sim N(c^C_\omega, f_2^2)$, where c^C_ω denotes the current value and f_2^2 is a tuning parameter. Proposed values are accepted with probability $f(c^P_\omega|\dots)/f(c^C_\omega|\dots)$.

16. The sampler utilizes the marginal in (39) to improve mixing. However, if samples are required from the posterior of $\boldsymbol{\eta}$, they can be generated from

$$\boldsymbol{\eta}_\nu|\cdots \sim N(\frac{c_\eta}{1+c_\eta}(\tilde{\boldsymbol{Z}}_\nu^\top\tilde{\boldsymbol{Z}}_\nu)^{-1}\tilde{\boldsymbol{Z}}_\nu^\top\tilde{\boldsymbol{\theta}}, \frac{\sigma^2 c_\eta}{1+c_\eta}(\tilde{\boldsymbol{Z}}_\nu^\top\tilde{\boldsymbol{Z}}_\nu)^{-1}),$$

where $\boldsymbol{\eta}_\nu$ is the non-zero part of $\boldsymbol{\eta}$.

1.2 Grouped correlations model

The algorithm for the 'grouped correlations' model requires 3 additional steps (the first 3 below) and a modification to a step from the algorithm of the 'common correlations' model (the last one below). We describe these next.

1. Update $v_h \sim \text{Beta}(d_h+1, d-\sum_{l=1}^h d_h+\alpha^*), h=1,\ldots,H-1$, where $d=p(p-1)/2$ is the number of correlations, and d_h is the number of correlations allocated in the hth cluster. Given v_h, update the weights $w_h, h=1,\ldots,H$.

2. To calculate the cluster assignment probabilities first define $\boldsymbol{\theta}_{kl}=(\theta_{1kl},\ldots,\theta_{Mkl})^\top$ to be the profile of the (k,l) correlation. Based on (35), we have the following posterior probability

$$P(\lambda_{kl}=h|\ldots) \propto w_h N(\boldsymbol{\theta}_{kl}; \boldsymbol{Z}^*_{kl\nu_h}\boldsymbol{\eta}^*_{\nu_h}, \sigma^2\boldsymbol{D}^2_{kl}(\boldsymbol{\omega}_\varphi)).$$

   Here matrix $\boldsymbol{D}^2_{kl}$ is diagonal of dimension M with elements given by $\exp\{\boldsymbol{z}^\top_{\varphi t}\boldsymbol{\omega}_\varphi\}, t\in T$, and matrix $\boldsymbol{Z}^*_{kl\nu_h}$ has M rows given by $(\boldsymbol{z}^*_{th})^\top, t\in T$, i.e. it is the subset of $\boldsymbol{Z}^*_{\nu_h}$ that corresponds to $\boldsymbol{\theta}_{kl}$. Vector $\boldsymbol{\eta}^*_{\nu_h}$ is imputed from its posterior, or if the cluster is empty from prior (28).

3. We update concentration parameter α^* using the method described by.[7] With the $\alpha^*\sim\text{Gamma}(a_{\alpha*}, b_{\alpha*})$ prior, the posterior can be expressed as a mixture of two Gamma distributions

$$\alpha^*|\eta,k \sim \pi_\eta\text{Gamma}(a_{\alpha*}+k, b_{\alpha*}-\log(\eta)) + (1-\pi_\eta)\text{Gamma}(a_{\alpha*}+k-1, b_{\alpha*}-\log(\eta)), \quad (44)$$

   where k is the number of non-empty clusters, $\pi_\eta=(a_{\alpha*}+k-1)/\{a_{\alpha*}+k-1+n(b_{\alpha*}-\log(\eta))\}$ and

$$\eta|\alpha^*,k \sim \text{Beta}(\alpha^*+1, d). \quad (45)$$

   Hence, the algorithm proceeds as follows: with α^* and k fixed at their current values, we sample η from (45). Then, based on the same k and the newly sampled value of η, we sample a new α^* from (44).

4. A modification to step 12 of the previous algorithm is needed. Here, we calculate the squared residuals based on $\hat{\boldsymbol{\eta}}^C_{\nu_h}=\{c_\eta/(1+c_\eta)\}(\tilde{\boldsymbol{Z}}^\top_{\nu_h}\tilde{\boldsymbol{Z}}_{\nu_h})^{-1}\tilde{\boldsymbol{Z}}^\top_{\nu_h}\tilde{\boldsymbol{\theta}}_h$. Furthermore, function S^* is needed for this step.

1.3 Grouped variables model

The algorithm for the 'grouped variables' model differs from that of the 'grouped correlations' model in the calculation of the cluster assignment probabilities. Here, they are computed as follows. Recall that $\boldsymbol{\theta}_{kl}=(\theta_{1kl},\ldots,\theta_{Mkl})^\top$ and let w_h be the prior probability that a variable is assigned to cluster h. We have the following posterior probability

$$P(\lambda_k=h|\ldots) \propto w_h\prod_{l\neq k} N(\boldsymbol{\theta}_{kl}; \boldsymbol{Z}^*_{kl\nu_{h\lambda_l}}\boldsymbol{\eta}^*_{\nu_{h\lambda_l}}, \sigma^2\boldsymbol{D}^2_{kl}(\boldsymbol{\omega}_\varphi)).$$

2 Application

Figure 7 displays the effects of the 4 covariates on the means of the 4 responses.

References

1. Roberts GO, Rosenthal JS. Examples of adaptive MCMC. *J Comput Graph Stat.* 2009;18:349–367.

2. Roberts GO, Rosenthal JS. Optimal scaling for various Metropolis-Hastings algorithms. *Stat Sci.* 2001;16:351-367.

3. Chan D, Kohn R, Nott D, Kirby C. Locally adaptive semiparametric estimation of the mean and variance functions in regression models. *J Comput Graph Stat.* 2006;15:915-936.

4. Gamerman D. Sampling from the posterior distribution in generalized linear mixed models. *Stat Comput.* 1997;7:57–68.

5. Liechty JC, Liechty MW, Müller P. Bayesian correlation estimation. *Biometrika.* 2004;91:1-14.

6. Liechty MW, Liechty JC, Müller P. The shadow prior. *J Comput Graph Stat.* 2009;18:368-383.

7. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *J Am Stat Assoc.* 1995;90:577–588.

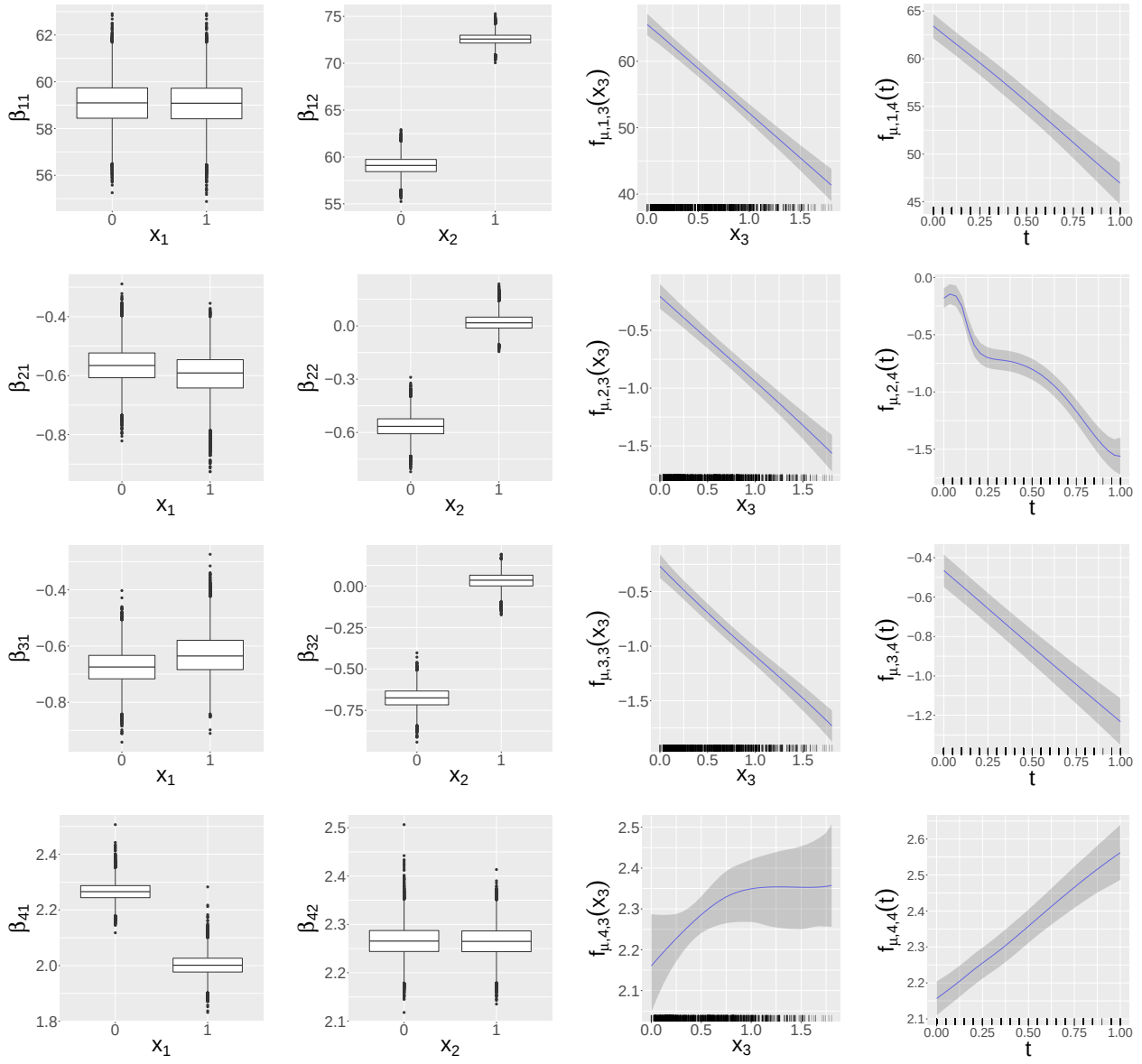

Figure 7: Application results: mean regression models. Posterior means and 80% credible intervals. Rows refer to the response means and columns to the covariate effects.