Biometrika (2013), **100**, 1, *pp.* 125–138
doi: 10.1093/biomet/ass060
Advance Access publication 17 December 2012

A nonparametric prior for simultaneous covariance estimation

By JEREMY T. GASKINS

Department of Statistics, University of Florida, Gainesville, Florida 32611, U.S.A.
jgaskins@stat.ufl.edu

and MICHAEL J. DANIELS

Section of Integrative Biology, Division of Statistics & Scientific Computation, University of Texas, Austin, Texas 78712, U.S.A.
mjdaniels@austin.utexas.edu

Summary

In the modelling of longitudinal data from several groups, appropriate handling of the dependence structure is of central importance. Standard methods include specifying a single covariance matrix for all groups or independently estimating the covariance matrix for each group without regard to the others, but when these model assumptions are incorrect, these techniques can lead to biased mean effects or loss of efficiency, respectively. Thus, it is desirable to develop methods for simultaneously estimating the covariance matrix for each group that will borrow strength across groups in a way that is ultimately informed by the data. In addition, for several groups with covariance matrices of even medium dimension, it is difficult to manually select a single best parametric model among the huge number of possibilities given by incorporating structural zeros and/or commonality of individual parameters across groups. In this paper we develop a family of nonparametric priors using the matrix stick-breaking process of Dunson et al. (2008) that seeks to accomplish this task by parameterizing the covariance matrices in terms of their modified Cholesky decompositions (Pourahmadi, 1999). We establish some theoretical properties of these priors, examine their effectiveness via a simulation study, and illustrate the priors using data from a longitudinal clinical trial.

Some key words: Bayesian nonparametric inference; Cholesky decomposition; Matrix stick-breaking process; Simultaneous covariance estimation; Sparsity.

1. Introduction

When working with longitudinal data, specifying the model for the dependence structure is a major consideration. Often the data are composed of several groups, such as for differing treatments in a clinical trial. In many cases, particularly if one does not have many observations per group, one assumes that the covariance or correlation structure is constant across all groups. However, this assumption, if it fails to hold, can have a dramatic effect on the inference for mean effects, even sometimes leading to bias. Conversely, if one specifies each of the covariance matrices without regard to the other groups, this can lead to a loss of information. Dealing with these competing models for the covariance structure is a concern in many statistical applications, such as classification and model-based clustering. Therefore, it is desirable to develop methods to simultaneously estimate the set of covariance matrices that will borrow information

across groups in a coherent, automated manner allowing for structural zeros, commonality across subsets of the groups, and appropriate equality of parameters within a group.

When the data are fully observed under multivariate normality, the mean and covariance parameters are orthogonal in the sense of Cox & Reid (1987), and the mean parameters will be consistent under misspecification of the covariance structure. However, if there is missingness, as is often the case for longitudinal data, there is no longer orthogonality, even at the true value of the covariance matrix (Little & Rubin, 2002). Hence, for the posterior distribution of the mean parameters to be consistent, the dependence structure must be correctly specified, and it is not appropriate to treat the covariance matrix as a nuisance parameter (Daniels & Hogan 2008, § 6.2). Further, Cripps et al. (2005) demonstrate efficiency gains for the regression parameters for fully observed data by using parsimonious models for the covariance matrix.

Assume that we have M groups of normally distributed longitudinal data with n_m responses of dimension p, Y_{mi} for the mth group. We assume without loss of generality that the mean vector for each group is zero. The distribution of Y_{mi} is

$$Y_{mi} \mid \Sigma_m \sim N_p\left(0, \Sigma_m\right) \quad (i = 1, \ldots, n_m;\ m = 1, \ldots, M),$$

with the covariance matrix $\Sigma_m = \Sigma(\Phi_m, \Gamma_m)$ parameterized by the generalized autoregressive parameters, Φ_m, and innovation variances, Γ_m, as described by Pourahmadi (1999, 2000). For brevity, we sometimes refer to the generalized autoregressive parameters as the autoregressive parameters. We also refer to this as the modified Cholesky parameterization, since the parameters are derived by performing a Cholesky decomposition on Σ_m, $\Sigma(\Phi_m, \Gamma_m)^{-1} = T(\Phi_m)\, D(\Gamma_m)\, T(\Phi_m)^{\mathrm{T}}$. Here, $\Gamma_m = \left(\gamma_{m1}, \ldots, \gamma_{mp}\right)$, and $D(\Gamma_m)$ is a $p \times p$ diagonal matrix with (j, j)-element $(\gamma_{mj})^{-1}$. The $T(\Phi_m)$ matrix is upper-triangular with ones on the main diagonal, and the above-diagonal elements are given by the negatives of Φ_m. The elements of $\Phi_m = (\phi_{m1}, \ldots, \phi_{mJ})$ are indexed by $j = 1, \ldots, J = p(p-1)/2$ corresponding to the location (j_1, j_2) in $T(\Phi_m)$ $(1 < j_1 < j_2 \leqslant p)$ for $-\phi_{mj}$. The natural interpretability of the autoregressive parameters relies on an assumed order of the p components of Y. This is quite natural in the longitudinal data setting where the elements of Y are measurements of the same quantity obtained at p different time-points. This assumed ordering may, however, not be appropriate in other multivariate data settings.

Many authors have developed frequentist estimators of the collection $\Sigma = \left\{\Sigma_1, \ldots, \Sigma_M\right\}$ by inducing commonality among some feature of the Σ_m. Boik proposed models to induce structure by imposing commonality on some or all of the principal components of the covariance (Boik, 2002) or correlation (Boik, 2003) matrix. Others have used the variance-correlation decomposition for estimation by imposing structures such as proportionality of all Σ_m or commonality among the correlation matrices (Manly & Rayner, 1987). Guo et al. (2011) considered an automated approach using the lasso to estimate sparse graphical models by selecting sets of edges common to all groups, as well as group-specific edges. However, their method shares little information about nonzero parameters across the groups. Pourahmadi et al. (2007) developed estimation and testing procedures for equality among subsets of the ϕ_{mj}. In a clustering context, models assuming $T(\Phi_m)$ and/or $D(\Gamma_m)$ to be either constant or distinct across all groups were developed by McNicholas & Murphy (2010). Daniels (2006) considered a Bayesian perspective by introducing priors for the parameters of the Cholesky decomposition, as well as the principal components of the covariance matrices, that induce pooling across groups. Unfortunately, it is computationally challenging to select among all the possible models within these classes. Hoff (2009) also considers a model that shrinks towards a common eigenvector structure, allowing the extent of the pooling to vary across each principal axis. Other methods have been proposed that

model the covariance matrix as a regression function of a continuous covariate (Chiu et al., 1996; Daniels, 2006; Fox & Dunson, 2011; Hoff & Niu, 2012). However, covariance regression models are often plagued by the difficulty of interpreting the regression parameters.

In this article we focus solely on the modified Cholesky parameterization because of the unrestrictedness of the parameters, the interpretability for longitudinal data, and the computational advantages via conjugacy (Daniels & Pourahmadi, 2002). Our goal is to develop a prior for the sets $\Phi = \{\Phi_1, \ldots, \Phi_M\}$ and $\Gamma = \{\Gamma_1, \ldots, \Gamma_M\}$ in such a way that we borrow strength across the M groups. Additionally, we want to share information across Γ_m and Φ_m values, particularly those autoregressive parameters of a common lag. Another consideration for prior development is to encourage sparsity of the elements of $T(\Phi_m)$. Because each ϕ_{mj} represents a conditional dependency, setting ϕ_{mj} to zero establishes a conditional independence relationship between a pair of components of Y. It is necessary to consider priors that allow the data to inform the balance between these two goals: pooling across groups and introducing sparsity. Above all, we seek to accomplish this in an automated, stochastic fashion. To form such a nonparametric prior, we employ the matrix stick-breaking process introduced by Dunson et al. (2008).

2. Matrix stick-breaking process

Before we introduce the proposed priors for Σ, we first review some of the key components of the matrix stick-breaking process (Dunson et al., 2008). The authors consider the case where n_m subjects from group m are drawn from a parametric model that depends on the p-dimensional parameter vector θ_m, as well as possible global parameters or subject-specific covariates. This process induces a prior for the set of θ_m that allows for clustering of parameters by drawing $\theta_{mj} \sim F_{mj}$ for $m = 1, \ldots, M$ and $j = 1, \ldots, p$, where F_{mj} is a random probability measure. They define the matrix $\mathcal{F}$ of random probability measures by $\{F_{mj} : m = 1, \ldots, M, j = 1, \ldots, p\}$, which will have a distribution that induces dependence among the F_{mj} measures that in turn produce desirable properties on the model parameters θ_m. The measure F_{mj} is a discrete distribution $F_{mj} = \sum_{h=1}^{H} \pi_{mjh}\delta_{\xi_{jh}}$, where δ_x represents a point mass at x, $\xi_{jh} \sim F_{0j}$ independently ($h = 1, \ldots, H$), and Ξ is the $p \times H$ matrix with (j, h)-element ξ_{jh}. The rows of Ξ ($j = 1, \ldots, p$) correspond to each of the model parameters, which have a nonatomic base distribution F_{0j}. The H columns are referred to as the clusters. The elements of Ξ are referred to as the parameter candidates because they constitute the set of potential values for the model parameters θ_{mj}.

The dependence among the F_{mj} is controlled by the specification of the stick-breaking weights π_{mjh}. These are defined by $\pi_{mjh} = V_{mjh}\prod_{l<h}(1 - V_{mjl})$, where $V_{mjh} = U_{mh}X_{jh}$, $U_{mh} \sim \text{Be}(1, \alpha)$, and $X_{jh} \sim \text{Be}(1, \beta)$. Here V_{mjh} is the product of U_{mh}, which controls the likelihood that parameters for group m come from cluster h, and X_{jh}, which controls the likelihood that parameter j is comes from cluster h. Because the U_{mh} are shared across parameters and the X_{jh} are shared across groups, they induce the dependence among the probability measures of $\mathcal{F}$. The authors require that $U_{mH} = 1$ for all m and that $X_{jH} = 1$ for all j, so that the stick-breaking weights sum to one, guaranteeing that F_{mj} is a valid distribution.

The matrix stick-breaking process is then defined using the above specification as $H \to \infty$, and the authors refer to the finite H case as the truncation approximation to the matrix stick-breaking process. The adequacy of this approximation is measured using a method similar to that employed by Ishwaran & James (2001). Dunson et al. (2008) show that for a set $\{\pi_{mjh}\}$ drawn from the full process,

$$E\left(\sum_{h=H}^{\infty}\pi_{mjh}\right) = \left\{1 - \frac{1}{(1+\alpha)(1+\beta)}\right\}^{H-1}. \quad (1)$$

The number of clusters H can be chosen such that this expected approximation error (1) is arbitrarily small, so the effect of the approximation is negligible.

Because the probability measures F_{mj} and $F_{m'j}$ for two groups m and m' share the same set of atoms $\{\xi_{j1}, \ldots, \xi_{jH}\}$, there is a positive probability that θ_{mj} will equal $\theta_{m'j}$. This occurs when θ_{mj} and $\theta_{m'j}$ are drawn from the same cluster, that is, if $\theta_{mj} = \theta_{m'j} = \xi_{jh}$ for some h in $1, \ldots, H$. The probability of this occurring is a known function of α and β.

3. COVARIANCE GROUPING PRIORS

3·1. *Lag-block grouping prior for* Φ

We now propose priors to use for simultaneous covariance estimation based on the matrix stick-breaking process. These priors are referred to as grouping priors because they induce grouping among the values of the various parameters. To this end, we independently place priors on Φ and Γ with the prior on Σ induced by the mapping $\Sigma_m = \Sigma(\Phi_m, \Gamma_m)$. Because Φ and Γ are orthogonal parameters (Pourahmadi, 2007), it is sensible to choose independent priors.

The prior for Φ, referred to as the lag-block grouping prior, is defined as follows:

$$\phi_{mj} \sim F_{mj}(\cdot) = \sum_{h=1}^{H_\phi} \pi_{mjh}\, \delta_{\xi_{q(j)h}}(\cdot) \quad (m = 1, \ldots, M;\ j = 1, \ldots, J), \tag{2}$$

$$\xi_{qh} \sim \epsilon_q \delta_0 + (1 - \epsilon_q)\mathrm{N}(0, \sigma^2) \quad (q = 1, \ldots, p-1;\ h = 1, \ldots, H_\phi), \tag{3}$$

$$\pi_{mjh} = U_{mh} X_{jh} \prod_{l<h} (1 - U_{ml} X_{jl}) \quad (\text{for all } m, j, h),$$

$$U_{mh} \sim \mathrm{Be}(1, \alpha_\phi) \quad (h = 1, \ldots, H_\phi - 1), \quad U_{mH_\phi} \sim \delta_1 \quad (m = 1, \ldots, M),$$

$$X_{jh} \sim \mathrm{Be}(1, \beta_\phi) \quad (h = 1, \ldots, H_\phi - 1), \quad X_{jH_\phi} \sim \delta_1 \quad (j = 1, \ldots, J).$$

Under this prior ϕ_{mj} is drawn from the random probability measure F_{mj} in (2). Here, $q(\cdot): \{1, \ldots, J\} \mapsto \{1, \ldots, p-1\}$ denotes the function that gives the lag value associated with the jth autoregressive parameter. Unlike the original specification of Dunson et al. (2008), the point masses in F_{mj} are not drawn for each parameter j but for all parameters j of the same lag $q(j)$. We refer to this as the lag-block prior because all of the lag-q ϕ are drawn from the same set of H_ϕ candidate values ξ_{qh}.

The distribution (3) of the ξ_{qh}, the candidate values for the lag-q generalized autoregressive parameters, is a mixture of a mean zero normal and a distribution degenerate at zero. Mixture priors for elements of a covariance matrix have been frequently used to promote sparsity (e.g., Smith & Kohn, 2002; Chen & Dunson, 2003; Cai & Dunson, 2006; Frühwirth-Schnatter & Tüchler, 2008). There are $(p-1)$ of the ϵ_q, each of which represents the probability that ξ_{qh} will be zero. The presence of the zero point mass promotes sparsity in $T(\Phi_m)$, and because $\phi_{mj} = 0$ represents a conditional independence relationship, the sparsity has a desirable interpretation. Allowing the probability of conditional independence to depend on lag follows from common intuition as one generally expects decreased relevance for higher lag terms. We can specify a beta prior for each of the ϵ_q with parameters that potentially depend on lag.

We form the probabilities $\{\pi_{mjh}\}$ as in Dunson et al. (2008). The α_ϕ and β_ϕ stick-breaking parameters serve the same role as α and β before. We subscript them with ϕ to distinguish the stick-breaking parameters for the prior on Φ from the parameters to be defined for the prior on Γ. The U_{mh} and X_{jh} parameters have the same interpretation as in the matrix stick-breaking

process, but while we share candidates across autoregressive parameters of the same lag, each parameter has its own values X_{jh}.

A key distinction between our prior and the original process of Dunson et al. (2008) is the use of the same set of candidate values for different parameters. This has important consequences for the theoretical properties of our priors. In particular, for ϕ_{mj} and $\phi_{mj'}$ with $j \neq j'$ and $q(j) = q(j')$, i.e., different autoregressive parameters for a common group and common lag, their distributions F_{mj} and $F_{mj'}$ are positively correlated, whereas under the original specification they would be uncorrelated. This implication is quite attractive for longitudinal data as it follows common intuition. For example, it may be reasonable to consider the regression effect of Y_t onto Y_{t-1} to be the same for different values of t. We discuss these properties further in § 4.

3·2. *Correlated-lognormal grouping prior for* Γ

We now define the prior for the innovation variances Γ as follows:

$$\gamma_{mj} \sim G_{mj}(\cdot) = \sum_{h=1}^{H_\gamma} \tau_{mjh}\, \delta_{\eta_{jh}}(\cdot) \quad (m = 1, \ldots, M;\ j = 1, \ldots, p), \tag{4}$$

$$\begin{aligned} \eta_{jh} &= \exp(\omega_{jh}) \quad (j = 1, \ldots, p;\ h = 1, \ldots, H_\gamma), \\ \omega_h &= (\omega_{1h}, \ldots, \omega_{ph})^{\mathrm{T}} \sim \mathrm{N}_p\{\psi 1_p, \Omega R(\rho)\} \quad (h = 1, \ldots, H_\gamma), \end{aligned} \tag{5}$$

$$\begin{aligned} \tau_{mjh} &= W_{mh} Z_{jh} \prod_{l<h} (1 - W_{ml} Z_{jl}) \quad (\text{for all } m, j, h), \\ W_{mh} &\sim \mathrm{Be}(1, \alpha_\gamma) \quad (h = 1, \ldots, H_\gamma - 1), \quad W_{mH_\gamma} \sim \delta_1 \quad (m = 1, \ldots, M), \\ Z_{jh} &\sim \mathrm{Be}(1, \beta_\gamma) \quad (h = 1, \ldots, H_\gamma - 1), \quad Z_{jH_\gamma} \sim \delta_1 \quad (j = 1, \ldots, p). \end{aligned}$$

We draw the innovation variance γ_{mj} from the stick-breaking measure G_{mj}, where the candidate atoms are drawn by exponentiating a multivariate normal variable ω_h. The probability τ_{mjh} of each of the atoms is formed using the stick-breaking method on the product of W and Z. These beta random variables depend on the parameters α_γ and β_γ.

The candidates η_{jh} are drawn in a correlated fashion, unlike the original matrix stick-breaking process, and marginally follow a lognormal distribution, providing the name of this prior. We introduce the intermediate variable ω_h in (5), which is a p-dimensional normally distributed random vector with mean vector $\psi 1_p$ and covariance matrix $\Omega R(\rho)$. Here, ψ and Ω are scalar quantities, $\Omega > 0$, and $R(\rho)$ is the correlation matrix corresponding to an autoregressive function of order 1. The (i, j) component of $R(\rho)$ is $\rho^{|i-j|}$. This choice is motivated by the fact that one sometimes considers the innovation variances as realized values of some unknown smooth function of time. Similar to the lag-block prior we will obtain the atoms η_{jh} for the random measure G_{mj} in a dependent way, while leaving the construction of the probability weights τ_{mjh} unchanged.

In the special case where $\rho = 0$, the components of the ω_h vector are independent. Consequently, the innovation variance candidates η_{jh} are independently distributed according to the log-normal(ψ, Ω) distribution, and this special case follows the original matrix stick-breaking process framework.

In addition to the grouping priors that we have defined here, there are other possibilities to form similar priors on the set $\{\Sigma_1, \ldots, \Sigma_M\}$ using the matrix stick-breaking process framework. We explore some of these in the online Supplementary Material.

4. THEORETICAL PROPERTIES

4·1. *Generalized autoregressive parameter properties*

We now explore some of the theoretical properties of the proposed grouping priors in the case where $H_\phi, H_\gamma \to \infty$. Recall that the matrix stick-breaking process is formally defined to be the limiting distribution as the number of clusters approaches infinity; using a finite number of clusters, while necessary for implementation, is viewed as an approximation. Our grouping priors follow in the same way. The following properties, numbered (6)–(12), are derived for these limiting distributions, and we ensure that the number of clusters is chosen large enough that these properties may be considered to hold approximately. The initial properties mirror Propositions 1, 2, and 4 of Dunson et al. (2008). Partial derivations of (6)–(12) are provided in the Supplementary Material.

First, we consider the behaviour of Φ from the lag-block grouping prior. For the following calculations, we assume that each ϵ_q and all hyperparameters are fixed. Additionally, for ease of notation, we ignore the subscript on $\epsilon_q, \alpha_\phi, \beta_\phi$ when it is clear from context, and let $\Phi(\cdot)$ denote the probability measure for the $N(0, \sigma^2)$ distribution. Define $\Psi(\cdot) = \epsilon\delta_0(\cdot) + (1-\epsilon)\Phi(\cdot)$, the probability measure for the mixture distribution of the ξ_{qh}.

For all sets A in the Borel field of the real line $\mathcal{B}(\mathcal{R})$,

$$E\left\{F_{mj}(A)\right\} = \Psi(A), \quad \text{var}\left\{F_{mj}(A)\right\} = \frac{2}{(2+\alpha)(2+\beta)-2}\Psi(A)\left\{1-\Psi(A)\right\}. \tag{6}$$

This unbiasedness property shows that it is appropriate to refer to mixture Ψ as the base distribution for Φ. The form of the variance shows that α and β control the extent to which the random measure F_{mj} differs from the base distribution. As either α or β approaches infinity, the distribution of ϕ_{mj} collapses to the parametric base Ψ; small α and β give a more flexible prior.

For two different groups $m \neq m'$,

$$\text{corr}\left\{F_{mj}(A), F_{m'j}(A)\right\} = \frac{\alpha + \alpha\beta/2 + \beta + 1}{2\alpha + \alpha\beta + \beta + 1}. \tag{7}$$

Because this correlation between the amounts of mass the distribution functions assign to the set A does not depend on the choice of A, it may be used as a simple univariate measure of the degree to which information is shared across groups. Simple algebra shows that $1/2 \leqslant \text{corr}\left\{F_{mj}(A), F_{m'j}(A)\right\} \leqslant 1$. In particular, $\text{corr}\left\{F_{mj}(A), F_{m'j}(A)\right\}$ approaches $1/2$ as either α or β approaches infinity and approaches 1 as $\alpha \to 0$.

For groups $m \neq m'$, the probability of matching for the jth autoregressive parameter is

$$\text{pr}(\phi_{mj} = \phi_{m'j}) = \epsilon^2 + \frac{1-\epsilon^2}{(1+\alpha)(2+\beta)-1}. \tag{8}$$

The presence of the zero point mass in Ψ causes our properties to differ from those derived in Dunson et al. (2008). As either α or β approach infinity, this probability approaches ϵ^2, the probability that both ϕ_{mj} and $\phi_{m'j}$ are zero if drawn independently from the parametric base distribution Ψ. The right-hand side of (8) is increasing in ϵ, as larger values of ϵ indicate that both terms are more likely to be zero whether or not they come from the same cluster. Additionally, (8) increases when either α or β decreases coinciding with the increase in (7).

Considering two different autoregressive parameters $j \neq j'$ of the same lag $q = q(j) = q(j')$ from the same group m, we have

$$\text{corr}\left\{F_{mj}(A), F_{mj'}(A)\right\} = \frac{\beta + \alpha\beta/2 + \alpha + 1}{2\beta + \alpha\beta + \alpha + 1}, \quad \text{pr}(\phi_{mj} = \phi_{mj'}) = \epsilon_q^2 + \frac{1 - \epsilon_q^2}{(2+\alpha)(1+\beta) - 1}. \tag{9}$$

The grouping prior has imposed a correlation structure on the distribution functions of the ϕ of a common lag, allowing us to borrow strength in the estimation of the dependence parameters from the same lag. This correlation is the same as (7) for the earlier $m \neq m'$ case with the roles of α and β in reverse. Likewise, the probability of matching across parameters of common lag (9) is also equivalent to the probability of matching across group for common parameter (8) with α and β exchanged. This is a key distinction from the process of Dunson et al. (2008) where the correlation and matching probabilities would be zero.

For different groups $m \neq m'$ and different autoregressive parameters $j \neq j'$ of the same lag,

$$\text{corr}\left\{F_{mj}(A), F_{m'j'}(A)\right\} = \frac{\alpha\beta/2 + \alpha + \beta + 1}{2\alpha\beta + 2\alpha + 2\beta + 1},$$
$$\text{pr}(\phi_{mj} = \phi_{m'j'}) = \epsilon_q^2 + \frac{1 - \epsilon_q^2}{2(1+\alpha)(1+\beta) - 1}. \tag{10}$$

Some algebra shows that this correlation is less than both (7) and (9). Likewise, $\text{pr}(\phi_{mj} = \phi_{m'j'})$ is smaller than (8) and (9). That is, the correlations of the distribution functions and the probability of matching across both group and autoregressive parameter are strictly smaller than the correlation and matching probability across just one. If $\alpha > \beta$, then $\text{corr}\left\{F_{mj}(A), F_{m'j}(A)\right\} < \text{corr}\left\{F_{mj}(A), F_{mj'}(A)\right\}$ and $\text{pr}(\phi_{mj} = \phi_{m'j}) < \text{pr}(\phi_{mj} = \phi_{mj'})$. Hence, there is more similarity in the distributions of common group and differing parameter than for the distributions of differing group but common parameter. The ordering reverses for $\alpha < \beta$. We noted previously that $\text{corr}\left\{F_{mj}(A), F_{m'j}(A)\right\}$ and $\text{corr}\left\{F_{mj}(A), F_{mj'}(A)\right\}$ are each contained in $[1/2, 1]$. Comparatively, $\text{corr}\left\{F_{mj}(A), F_{m'j'}(A)\right\}$ is in $[1/4, 1]$, again guaranteeing a strictly positive dependence between all distributions corresponding to the same lag.

For two different generalized autoregressive parameters $j \neq j'$ with different lag, $q(j) \neq q(j')$, $F_{mj}(A)$ and $F_{m'j'}(A)$ are uncorrelated for $1 \leqslant m, m' \leqslant M$. Further, $\text{pr}(\phi_{mj} = \phi_{m'j'}) = \epsilon_q \epsilon_{q'}$, the product of the marginal probabilities that each are set to zero.

We conclude the discussion of the Φ grouping prior by noting that distributions obtained in the special cases $\alpha \to 0$ and $\alpha, \beta \to 0$ are different from those obtained in the original matrix stick-breaking process. As $\alpha \to 0$, it remains true that all F_{mj} converge to $F_j \sim \text{DP}(\beta F_{0j})$, common for all groups. However, these are not independent across j, because the set of F_j of common lag share the same set of atoms ξ_{qh}. Previously as $\alpha, \beta \to 0$, the data are pooled such that a common parameter value is assigned for all groups, i.e., $\phi_{mj} = \phi_j$. Since atoms are shared across all autoregressive parameters of a common lag, the limiting distribution will yield $\phi_{mj} = \phi_q$ for all m and j with $q = q(j)$. As $\alpha \to 0$ and $\beta \to \infty$, $\phi_{mj} \sim \Psi$ as in Dunson et al. (2008).

4·2. *Innovation variance properties*

We now explore the behaviour of the innovation variances and their distributions G_{mj}. Let $\mathcal{R}_+$ denote the positive real line, $\log A$ be the set $\{\log x : x \in A\}$ for any $A \in \mathcal{B}(\mathcal{R}_+)$, and $\Phi(\cdot)$ the probability function for the $\text{N}(\psi, \Omega)$ distribution, assuming the hyperparameters ψ, Ω are fixed. Properties (6)–(8) hold as in the autoregressive parameter case with $\Psi(A)$ replaced $\Phi(\log A)$ and ϵ set to 0.

For innovation variances of the same group and different times $j \neq j'$,

$$\operatorname{corr}\left\{G_{mj}(A), G_{mj'}(A)\right\} = \frac{\beta + \alpha\beta/2 + \alpha + 1}{2\beta + \alpha\beta + \alpha + 1} \operatorname{corr}\left\{\delta_{\omega_{j1}}(\log A),\ \delta_{\omega_{j'1}}(\log A)\right\}, \qquad (11)$$

and for both different groups $m \neq m'$ and different times $j \neq j'$,

$$\operatorname{corr}\left\{G_{mj}(A), G_{m'j'}(A)\right\} = \frac{\alpha\beta/2 + \alpha + \beta + 1}{2\alpha\beta + 2\alpha + 2\beta + 1} \operatorname{corr}\left\{\delta_{\omega_{j1}}(\log A),\ \delta_{\omega_{j'1}}(\log A)\right\}. \qquad (12)$$

The correlation of these distributions now depends on the choice of Borel set A. However, they are the products of a term that depends solely on the stick-breaking parameters α and β and a term that depends only on A and the distribution of $(\omega_{j1}, \omega_{j'1}) \sim N_2\left\{\psi 1_2, \Omega R^*(\rho)\right\}$, where $R^*(\rho)$ is the 2×2 correlation matrix with off-diagonal elements $\rho^{|j-j'|}$. The higher correlations for neighbouring terms imply a smoothing of the variances as a function of j for $\rho > 0$. We observe that the leading terms of (11) and (12) gives the same correlation structure as in (9) and (10), respectively. Additionally, with the choice of $\rho = 0$, the term depending on A is zero, and the distributions are uncorrelated as in Dunson et al. (2008).

For $j \neq j'$ and $1 \leqslant m, m' \leqslant M$, $\operatorname{pr}(\gamma_{mj} = \gamma_{m'j'}) = 0$, that is, there is no matching of the innovation variances across time points. This is a consequence of the fact that two points drawn from a correlated normal distribution with $|\rho| < 1$ will be equal with probability zero.

5. COMPUTATIONAL CONSIDERATIONS

Recall that (1) provided us with the expected approximation error which we employ to choose the number of clusters necessary for the matrix stick-breaking process truncation. This formula continues to hold for the proposed grouping priors, since the stick-breaking weights are formed using the same framework. Hence, if the values of α, β for either the lag-block or correlated-lognormal prior are assumed known, then we choose the number of clusters H such that (1) is less than some threshold, such as 0·01. As we generally do not have any knowledge or prior belief about these stick-breaking parameters, it will often be inappropriate to prespecify values, so we follow the suggestion of Dunson et al. (2008) and specify independent Ga(1, 1) priors for α and β. Analyses using Ga(10, 10) and Ga(0·1, 0·1) indicate little effect of this prior choice on the estimates of Σ. To choose the value of H when using a prior for the stick-breaking parameters, we run a preliminary Markov chain for approximately 10% of the desired chain length and use the posterior means to test whether (1) is below our threshold. If so, we fix this value of H for remaining computation.

One of the nice properties of the matrix stick-breaking process is that introducing appropriate latent variables leads to a computational algorithm that generally samples from well-known conjugate distributions (Dunson et al., 2008). Because a normal prior for the generalized autoregressive parameters provides conjugacy, the sampling for ξ is from a relatively easy-to-sample zero-normal mixture. With the lognormal distribution, conjugacy for the innovation variances is not obtained since inverse gamma is the conjugate distribution for γ, but we can sample η efficiently by incorporating a slice sampling step (Neal, 2003). We appropriately modify the algorithm of Dunson et al. (2008) for posterior sampling from our grouping priors and discuss further computational challenges in the Supplementary Material.

One issue in the sampling algorithm is the behaviour of sampling the correlation ρ. In our experience when considering priors with ρ random, ρ did not seem to be well informed by the data. Hence, we opt to treat ρ as a tuning parameter. We recommend specifying a default value

such as $\rho = 0{\cdot}75$, possibly trying a few other choices and selecting the value based on some model selection criterion. As shown in the depression data study in § 7, the three choices of $\rho = 0{\cdot}5$, $0{\cdot}75$, and $0{\cdot}9$ lead to similar model fits as measured by the deviance. Based on our simulation studies, it appears that the correlated-lognormal prior is fairly robust to the choice of ρ.

6. Risk simulation

We now examine the operating characteristics of the proposed grouping priors via a risk simulation designed to mimic the analysis of a typical longitudinal data scenario. We incorporate a nonzero mean, and the simulated data will suffer from ignorable dropout. There are $M = 8$ groups each with $n_m = 50$ measurements of dimension $p = 6$. Let D_i denote the time $t = 2, \ldots, p+1$ of dropout for subject i, where $D_i = p + 1$ indicates a subject who completes the study. Dropout is induced according to the model

$$\text{logit}\left\{\text{pr}(D_i = t+1 \mid D_i > t, y_{it}, m)\right\} = \zeta_{0t} + \zeta_{1t}\, y_{it} + \zeta_{2m} \quad (t = 1, \ldots, p-1). \tag{6}$$

This missing data mechanism is missing at random because the dropout time depends only on observed values. The mean, covariance, and dropout parameters for the simulation, as well as the probabilities of missingness at time t for each group, are provided in the Supplementary Material.

The choices of Φ and Γ do not have any equalities across groups but some within lag. However, with the small sample sizes, it will generally still be advantageous to share information across the eight groups. Also, there is a moderate amount of sparsity in Φ, as is typical for ordered data. All groups have a mean of zero at time 1, and the mean functions increase at differing rates to the final time $t = 6$. The dropout rates vary across groups with most groups losing 35 to 50% of their subjects by $t = 6$. Groups 3 and 8 experience a larger amount of attrition, about 70%, over the study, which will have adverse effects on the mean estimation; these are also the groups with the largest mean values at $t = 6$.

For the purposes of comparison with the grouping priors, we introduce some additional naive priors based on the modified Cholesky parameterization of the covariance matrix. The comparison priors for the generalized autoregressive parameters are

$$\text{Naive Bayes 1:} \quad \phi_{mj} \sim \epsilon_{q(j)}\delta_0 + (1 - \epsilon_{q(j)})\text{N}(0, \sigma^2),$$
$$\text{Naive Bayes 2:} \quad \phi_{mj} \sim \text{N}(0, \sigma^2).$$

The naive prior 1 independently draws each ϕ_{mj} from the mixture base measure Ψ. This is the simplification of our prior that does not attempt any clustering or dependence across groups or parameters. This prior can be viewed as an extension of the Smith & Kohn (2002) selection prior to a multiple group setting. This prior is also the limiting case as $\alpha_\phi \to 0$ and $\beta_\phi \to \infty$, if the candidates ξ were drawn for all parameters j, instead of blocking within lag $q(j)$. The naive prior 2 can be viewed similarly with all ϵ fixed to zero, so that there is no sparsity. This is a simplification of the prior suggested by Daniels & Pourahmadi (2002).

We pair this with an independent InvGa(λ_1, λ_2) for each γ_{mj}, where λ_2 defines the rate of the inverse gamma distribution. These priors are simple choices for Φ and Γ priors so that conjugacy is maintained, leading to relatively simple sampling algorithms.

Additionally, to represent two of the more common methods of dealing with this situation, we run a Markov chain with a common-Σ flat prior and a group-specific flat prior. The common-Σ prior assumes a common covariance matrix across all groups and uses a flat prior for this matrix. The group-specific prior places independent flat priors on each of the M groups. The resulting conditional distributions are inverse-Wishart, resulting in a simple sampling algorithm.

Table 1. *Estimated risks for each choice of covariance prior from the simulation in § 6. The estimated risk is calculated as the average loss using loss functions* $L_1(\Sigma_m, \hat{\Sigma}_{m1}) = \mathrm{tr}(\Sigma_m^{-1}\hat{\Sigma}_{m1}) - \log|\Sigma_m^{-1}\hat{\Sigma}_{m1}| - p$, $L_2(\Sigma_m, \hat{\Sigma}_{m2}) = \mathrm{tr}\{(\Sigma_m^{-1}\hat{\Sigma}_{m2} - I)^2\}$, *and* $L^{\mu}(\hat{\mu}_m, \mu_m) = (\hat{\mu}_m - \mu_m)^{\mathrm{T}}\Sigma_m^{-1}(\hat{\mu}_m - \mu_m)$

	Estimated risk		
Prior	L_1	L_2	L^{μ}
Covariance grouping prior	0·425	0·742	0·175
Naive Bayes 1	0·605	0·987	0·203
Naive Bayes 2	0·630	1·010	0·210
Group-specific flat*	0·892	1·255	0·248
Common-Σ flat	8·105	84·339	0·925

*The group-specific flat prior is over only 49 datasets because the Markov chain failed to converge for one dataset.

Under this data model, we generate 50 datasets and run our Markov chain Monte Carlo algorithm on each dataset with each prior for 50 000 iterations keeping every tenth iteration, using a burn-in of 10 000. We place the following priors on the hyperparameters when appearing in the prior specification: ϵ_q, independent Unif(0, 1); α_ϕ, β_ϕ, α_γ, β_γ, λ_1, and λ_2, independent Ga(1, 1); $\sigma^2 \sim \mathrm{InvGa}(0{\cdot}1, 0{\cdot}1)$; $\Omega \sim \mathrm{InvGa}(0{\cdot}1, 0{\cdot}1)$; $\psi \sim N(0, c^2\Omega)$, with $c^2 = 1000$. We fix the value of ρ to be 0·75. We assume a flat prior on the group-specific mean vectors μ_m. To handle incomplete data, we use data augmentation to sample the missing data values from normal distributions conditional on the observed data.

We measure the performance of our proposed priors by estimating the risk associated with the Bayes estimators under two common loss functions (Yang & Berger, 1994), $L_1(\Sigma_m, \hat{\Sigma}_{m1}) = \mathrm{tr}(\Sigma_m^{-1}\hat{\Sigma}_{m1}) - \log|\Sigma_m^{-1}\hat{\Sigma}_{m1}| - p$ and $L_2(\Sigma_m, \hat{\Sigma}_{m2}) = \mathrm{tr}\{(\Sigma_m^{-1}\hat{\Sigma}_{m2} - I)^2\}$. Since these losses are defined in terms of a single covariance matrix, we consider the loss for estimating the set of covariance matrices to be the average across groups of the losses from the individual covariance matrices. To study the ability to recover the mean function with differing priors on Σ, we use the posterior mean of μ_m and the loss function $L^{\mu}(\hat{\mu}_m, \mu_m) = (\hat{\mu}_m - \mu_m)^{\mathrm{T}}\Sigma_m^{-1}(\hat{\mu}_m - \mu_m)$, standardizing by the true covariance matrix Σ_m and taking the average across groups.

The estimated risks associated with estimating the covariance matrices for loss functions L_1 and L_2 are shown in Table 1. The grouping prior shows a risk improvement of 30 and 25% over the naive Bayes 1 prior and 52 and 41% over the group-specific flat prior. The success of the grouping prior relative to the naive Bayes priors shows that matching of autoregressive parameters across groups can improve estimation for small-to-moderately sized samples, even if the true parameter values do not exhibit such matching.

The final column of Table 1 displays the risk in mean estimation showing a clear improvement under the grouping prior. The covariance grouping prior produces a risk 14% smaller than the naive Bayes 1 prior and 29% smaller than the group-specific estimator. The risk associated with the common-Σ prior is almost five times that associated with the grouping priors. This corresponds with our observations in the introduction that by considering a more accurate structure on the dependence, we estimate the mean function more efficiently and with less bias.

Additional risk simulations using the grouping prior under simpler data models with fully observed data have been performed, some of which are included in the Supplementary Material. These continued to show that our prior outperforms the naive competitors under many different types of covariance matrix specifications such as situations with no sparsity and dissimilar covariance matrices across groups, and under increasing n_m, M, and p. In particular, we believe that as the number of groups M and the dimension of the covariance matrix p increases, the grouping estimators for Σ will outperform the naive Bayes estimators and the margin by which

they do so increases. The choice of the flat priors continued to perform poorly compared to the grouping and naive choices.

7. Data example

We demonstrate the use of the grouping priors in the fitting of a longitudinal dataset from a depression study. The data, originally presented by Thase et al. (1997) and also analysed by Pourahmadi & Daniels (2002), consist of weekly measures of depression symptoms over a sixteen-week study period. Depression scores were measured weekly using the Hamilton Rating Scale for Depression. As noted in previous analyses of the dataset, the severity of the depression symptoms at baseline influences the rate of the improvement. There are two classes of treatments under consideration in the study, psychotherapy-only versus treatment regimens which include both psychotherapy and pharmacotherapy. We divide the data into $M = 4$ groups for analysis considering each combination of treatment and a binary indicator of the initial severity of depression. The sample sizes for the four groups are 98, 101, 100, and 249. The vector of a patient's seventeen weekly depression scores, measured from baseline to week 16, is assumed to be normally distributed with a quadratic mean function and covariance matrix specific to each treatment-severity combination.

We perform an analysis using the covariance grouping prior, the two naive Bayes priors, and the two flat priors. For each prior specification, the chain ran for a burn-in of 10 000 iterations followed by another 100 000 iterations, of which we retained every tenth value for inference. As described in § 5, appropriate values for H_ϕ and H_γ were fixed at 25 after running a preliminary Markov chain. For the correlated-lognormal prior, we ran a chain for three fixed values for ρ of 0·5, 0·75, and 0·9; otherwise, we used the same hyperpriors as in § 6.

Approximately 30% of the possible measurements from the study are missing. We assume that the missingness was ignorable and incorporated a data augmentation step to sample the missing values given the observed data and the current parameter values, as in the previous risk simulation. A flat prior was used for the regression parameters.

To compare the model fits induced by the various covariance priors used, we use the deviance information criteria (Spiegelhalter et al., 2002). Wang & Daniels (2012) recommend using the observed data likelihood for calculations when there is missing data. The deviance information criteria are calculated by $\text{DIC} = D + 2p_D$, where D is the deviance at the posterior expectation of the parameter values and p_D is a model complexity parameter, often interpreted as the effective number of parameters. Smaller values of DIC indicate a good mix of model fit and simplicity.

Table 2 contains the model deviance, the effective number of parameters p_D, and the model selection criteria DIC for each prior choice for Σ. The covariance grouping priors provide a superior fit compared to the competitors. For the grouping prior the largest of the three correlations $\rho = 0{\cdot}9$ slightly beats the other two, but the relatively small difference in DIC provides evidence of the robustness of the prior to the choice of ρ. Comparing the complexity terms p_D, we see that the grouping prior induces a more structured model than the naive priors, using approximately 135 and 215 fewer parameters. As the decrease in p_D between the two naive priors is attributed to the added sparsity, the decrease from the naive Bayes 1 to the grouping priors is due to the ability to exploit matching of autoregressive parameters across group and lag, as well as matching across groups for the γ. Finally, the grouping priors model the data much more effectively than those methods that assume a flat prior on the covariance matrix, in particular, the treatment-specific prior which has too many parameters to be handled efficiently.

We also consider how the choice of covariance prior affects mean estimation. We show the treatment effect, calculated as the difference in mean value between baseline and week 16, and 95% credible intervals for the first two groups in Table 2. In group $m = 2$ we see clear differences

Table 2. *Model fit statistics and treatment effects for the first two groups for the depression data using each of the priors*

Covariance prior	Model fit			Treatment effect	
	D	p_D	DIC	Group 1	Group 2
Grouping prior ($\rho = 0{\cdot}90$)	39 006	342	39 690	9·23 (7·03 11·48)	9·51 (6·85, 12·19)
Grouping prior ($\rho = 0{\cdot}75$)	39 006	347	39 700	9·22 (6·99, 11·42)	9·56 (6·85, 12·27)
Grouping prior ($\rho = 0{\cdot}50$)	39 003	349	39 700	9·24 (7·00, 11·53)	9·41 (6·74, 12·14)
Naive Bayes 1	38 875	481	39 837	9·25 (7·11, 11·46)	8·56 (6·16, 10·99)
Naive Bayes 2	38 765	563	39 890	9·24 (7·02, 11·49)	7·99 (5·59, 10·46)
Common-Σ flat	39 907	220	40 347	9·44 (6·21, 12·53)	10·17 (7·02, 13·24)
Group-specific flat	39 178	1021	41 219	9·20 (6·44, 12·08)	6·93 (4·22, 9·77)

D, the deviance at the posterior expectation of the parameter values; P_D, a model complexity parameter; DIC, deviance information criteria.

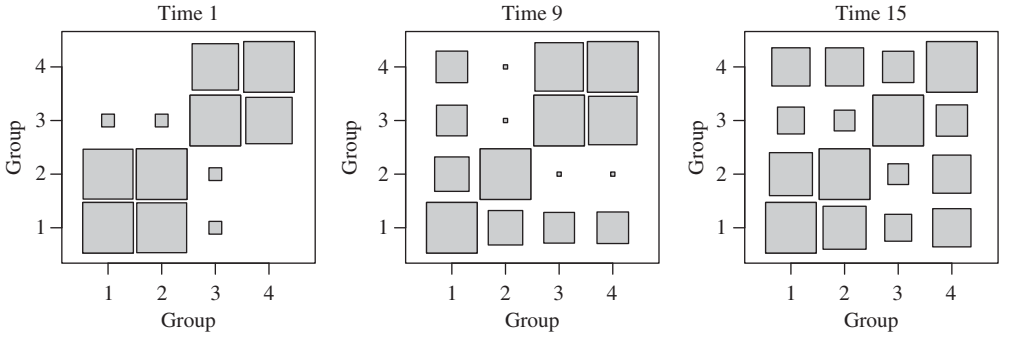

Fig. 1. The posterior probabilities of matching for the innovation variances at times 1, 9, and 15. The sizes of the boxes are proportional to $\text{pr}(\gamma_{mj} = \gamma_{m'j} \mid y_{obs})$.

for this effect across the different priors. The treatment effect under the grouping priors is estimated to be around 9·5 points, but it is 10·2 for the common-Σ flat and 6·9 for the group-specific flat prior. Even between the two naive priors, the estimated treatment effects differ by 0·6. For group 1, as well as groups 3 and 4, we do not observe much difference in the mean effect, except for some deviation with the common-Σ prior, although the confidence interval is more narrow for the grouping priors than the flat versions. These two groups demonstrate the bias and efficiency issues relevant to covariance matrix estimation with missing data as discussed in the introduction. The differences do not rise to the level of statistical significance here, but they are large enough to be of practical importance.

Figures 1 and 2 show the grouping nature of the proposed priors. Figure 1 shows the posterior probabilities of $\text{pr}(\gamma_{mj} = \gamma_{m'j})$ for each m, m' combination at times $j = 1, 9, 15$. These times were chosen as representative of the overall patterns in the data. For $j = 1$ and most of the undisplayed times, there is substantial matching for the groups 1 and 2, the low initial severity group, as well as for groups 3 and 4, the high initial severity groups, with less matching across the pairs. The variances at $j = 9$ and 15 show a stronger propensity to match across all groups.

Figure 2 gives the posterior probabilities of matching for the lag-1 and lag-4 autoregressive parameters. We show only the first four of each due to space limitations. The black boxes that overlay the $y = x$ diagonal are proportional to the posterior of $\text{pr}(\phi_{mj} = 0)$. We see that the lag-1 terms are rarely set to zero, while the lag-4 terms for groups 2 and 3 are frequently zeroed out. Due to the nature of the grouping prior, there is a positive probability of equality across ϕ_{mj} of a common lag. In Fig. 2(a) we note the pairwise probability of equality is very high for all combinations of the first autoregressive parameter of all groups, i.e., the regression of time 2 onto time 1, and the other lag-1, group 2 terms. One would be unlikely to learn of this relationship or to

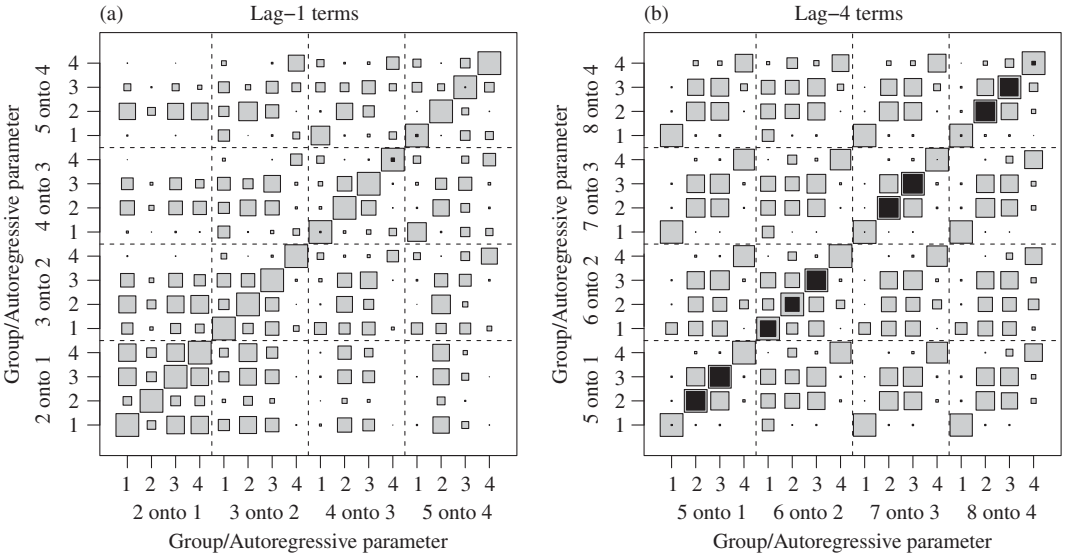

Fig. 2. The posterior probabilities of matching for the generalized autoregressive parameters. Panel (a) contains the matching for the first four lag-1 terms, and (b) displays the first four lag-4 terms. The sizes of the grey boxes are proportional to $\text{pr}(\phi_{mj} = \phi_{m'j'} \mid y_{obs})$. The black boxes overlaying the diagonal are proportional to the posterior of $\text{pr}(\phi_{mj} = 0)$. The axes indicate group m (top line) and autoregressive parameter j of lag-q (bottom line).

consider a model with equality across all, or even a large subset, of these parameters using other approaches. Considering Fig. 2(b), there are larger matching probabilities for the lag-4 parameters, much of which is due to matching with both parameters set to zero. However, the matching is not always due to equality at zero, as can be seen from the large probabilities of matching across the group 1 ϕ. There is similar behaviour for the group 4 generalized autoregressive parameters.

8. Discussion

We have developed a prior on the set of M covariance matrices that simultaneously exploits sparsity and matching of dependence parameters across groups. The model space containing all combinations where each autoregressive parameter/variance is constant across all possible subsets of the groups has $B_M^{p(p+1)/2}$ models, where B_M is the Mth Bell number (Stanley 1997, p. 33). In fact, the grouping priors consider a space that is even larger since we allow matching across autoregressive parameters of a common lag. With this many models we have little hope of finding the most appropriate one. Our grouping priors avoid this problem by stochastically considering the possibility of each of these models in a single analysis and accounting for uncertainty appropriately. It is our belief that running a Markov chain with one of these grouping priors is a necessary alternative to the unreasonable time and energy required to fit and compare this extremely large class of models.

Acknowledgement

This work was partially supported by the National Institutes of Health.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of the properties presented in § 4, additional risk simulations, a detailed description of the sampling algorithm, and additional covariance grouping priors with theoretical properties.

REFERENCES

BOIK, R. J. (2002). Spectral models for covariance matrices. *Biometrika* **89**, 159–82.

BOIK, R. J. (2003). Principal component models for correlation matrices. *Biometrika* **90**, 679–701.

CAI, B. & DUNSON, D. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics* **62**, 446–57.

CHEN, Z. & DUNSON, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762–9.

CHIU, T. Y. M., LEONARD, T. & TSUI, K.-W. (1996). The matrix-logarithmic covariance model. *J. Am. Statist. Assoc.* **91**, 198–210.

COX, D. R. & REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B **49**, 1–18.

CRIPPS, E., CARTER, C. K. & KOHN, R. (2005). Variable selection and covariance selection in multivariate regression models. In *Handbook of Statistics*, vol. 25.

DANIELS, M. J. (2006). Bayesian modelling of several covariance matrices and some results on the propriety of the posterior for linear regression with correlated and/or heterogeneous errors. *J. Mult. Anal.* **97**, 1185–207.

DANIELS, M. J. & HOGAN, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton: Chapman & Hall.

DANIELS, M. J. & POURAHMADI, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **89**, 553–66.

DUNSON, D. B., XUE, Y. & CARIN, L. (2008). The matrix stick-breaking process: Flexible Bayes meta-analysis. *J. Am. Statist. Assoc.* **103**, 317–27.

FOX, E. & DUNSON, D. B. (2011). Bayesian nonparametric covariance regression. *J. Am. Statist. Assoc.* Arxiv preprint arxiv:1101.2017.

FRÜHWIRTH-SCHNATTER, S. & TÜCHLER, R. (2008). Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statist. Comp.* **18**, 1–13.

GUO, J., LEVINA, E., MICHAILIDIS, G. & ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**, 1–15.

HOFF, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *J. R. Statist. Soc.* B **71**, 971–92.

HOFF, P. D. & NIU, X. (2012). A covariance regression model. *Statist. Sinica* **22**, 729–53.

ISHWARAN, H. & JAMES, L. F. (2001). Gibbs sampling methods for stick breaking priors. *J. Am. Statist. Assoc.* **96**, 161–73.

LITTLE, R. J. A. & RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

MANLY, B. F. J. & RAYNER, J. C. W. (1987). The comparison of sample covariance matrices using likelihood ratio tests. *Biometrika* **74**, 841–7.

MCNICHOLAS, P. D. & MURPHY, T. B. (2010). Model-based clustering of longitudinal data. *Can. J. Statist.* **38**, 153–68.

NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31**, 705–67.

POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–90.

POURAHMADI, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425–35.

POURAHMADI, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance correlation parameters. *Biometrika* **94**, 1006–13.

POURAHMADI, M. & DANIELS, M. J. (2002). Dynamic conditionally linear mixed models for longitudinal data. *Biometrics* **58**, 225–31.

POURAHMADI, M., DANIELS, M. J. & PARK, T. (2007). Simultaneous modelling of the Cholesky decomposition of several covariance matrices. *J. Mult. Anal.* **98**, 568–87.

SMITH, M. & KOHN, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Am. Statist. Assoc.* **97**, 1141–53.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B **64**, 583–639.

STANLEY, R. P. (1997). *Enumerate Combinatorics*, vol. 1. New York: Cambridge University Press.

THASE, M., GREENHOUSE, J., FRANK, E., REYNOLDS, C., PILKONIS, P., HURLEY, K., GROCHOCINSKI, V. & KUPFER, D. (1997). Treatment of major depression with psychotherapy or psychotherapy-pharmacotherapy combinations. *Arch. Gen. Psychiat.* **54**, 1009–15.

WANG, C. & DANIELS, M. J. (2012). A note on MAR, identifying restrictions, model comparison, and sensitivity analysis in pattern mixture models with and without covariates for incomplete data. *Biometrics* **68**, 994.

YANG, R. & BERGER, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22**, 1195–211.

[*Received January* 2012. *Revised August* 2012]