A Nonparametric Prior for Simultaneous Covariance Estimation

J. T. Gaskins* and M. J. Daniels†

Abstract:

In the modeling of longitudinal data from several groups, appropriate handling of the dependence structure is of central importance. In many cases, one assumes that the covariance (or correlation) structure is the same for all groups. However, this assumption, if it fails to hold, can have an adverse effect on inference for mean effects. Conversely, if one specifies each of the covariance matrices without regard to the other groups, this can lead to a loss of efficiency if there is information to be gained across groups. It is desirable to develop methods to simultaneously estimate 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 given a huge number of possibilities (e.g., 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 the parameters of their modified Cholesky decomposition (Pourahmadi, 1999). We establish some theoretic properties of these priors, examine their effectiveness via a simulation study, and illustrate the priors using data from a longitudinal clinical trial.

*jgaskins@stat.ufl.edu; Department of Statistics,University of Florida, Gainesville, FL 32611

†mdaniels@stat.ufl.edu; Department of Statistics,University of Florida, Gainesville, FL 32611

1. Introduction

When working with longitudinal data, specifying the model for the dependence structure is a major consideration. Often the data is composed of several groups, such as 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 of mean effects, even leading to bias if data are incomplete (Daniels and Hogan, 2008). 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 (a subset of) the groups, and appropriate equality of parameters within a group. We accomplish this task by developing nonparametric priors for the set of covariance matrices.

We will begin by establishing the necessary notation. 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 the Y_{mi} is

$$Y_{mi} \,|\, \Phi_m, \Gamma_m \sim \text{i.i.d. } \mathrm{N}_p\big(0, \Sigma(\Phi_m, \Gamma_m)\big),\ i = 1, \ldots, n_m;\ m = 1, \ldots, M,$$

with the covariance matrix $\Sigma_m = \Sigma(\Phi_m, \Gamma_m)$ parameterized by the generalized autoregressive parameters (GARPs), Φ_m, with innovation variances (IVs), Γ_m, as described by Pourahmadi (1999 and 2000). We also refer to this as the modified Cholesky parameterization, since the parameters

are derived by performing a Cholesky decomposition on Σ_m. That is,

$$\begin{aligned}\Sigma(\Phi_m, \Gamma_m)^{-1} &= T(\Phi_m)\, D(\Gamma_m)\, T(\Phi_m)' \\ &= \begin{bmatrix} 1 & -\phi_{m1} & -\phi_{m2} & \cdots \\ & 1 & -\phi_{m3} & \cdots \\ & & 1 & \cdots \\ & & & \ddots \end{bmatrix} \begin{bmatrix} \frac{1}{\gamma_{m1}} & & & \\ & \frac{1}{\gamma_{m2}} & & \\ & & \ddots & \\ & & & \frac{1}{\gamma_{mp}} \end{bmatrix} \begin{bmatrix} 1 & & & \\ -\phi_{m1} & 1 & & \\ -\phi_{m2} & -\phi_{m3} & 1 & \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.\end{aligned}$$

The $T(\Phi_m)$ matrix is upper-triangular with ones on its diagonal. Note that there are p parameters for each $\Gamma_m = (\gamma_{m1}, \ldots, \gamma_{mp})$ and $J = p(p-1)/2$ parameters associated with each $\Phi_m = (\phi_{m1}, \ldots, \phi_{mJ})$. The natural interpretability of the GARPs 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 $\Sigma = \{\Sigma_1, \ldots, \Sigma_M\}$ 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 (2002) or correlation (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 and Rayner, 1987). Pourahmadi et al. (2007) developed estimation and testing procedures for equality among the GARPs and subsets of the GARPs. Daniels (2006) considered a Bayesian perspective by introducing priors for the GARPs and IVs, 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. Other techniques have been proposed that model the covariance matrix as a function of a continuous covariate. These include models that perform regressions on certain model components (Chiu et al., 1996; Daniels, 2006; Fox and Dunson, 2011; Hoff and Niu, 2011), as well as those that treat the covariance matrices as realization of a stochastic volatility process (Philiov and Glickman, 2006a,

2006b; Lopes et al., 2011).

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. In the longitudinal data setting we wish to find more covariance structure than just common zeros across all groups. We want to consider models that allow subsets of the model parameters to be equal across (a subset of the) groups at non-zero values; the Guo et al. estimators do not accommodate this goal. We additionally note that it is not clear how one could easily adapt their penalty term into a Bayesian prior on the set covariance matrices for our setting.

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 and Pourahmadi, 2002). Our goal is to develop a prior for the set of GARPs and IVs 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 GARPs of a common lag. Another consideration for prior development is to encourage sparsity of the elements of $T(\Phi_m)$, that is, containing few non-zero elements. Because each GARP 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 (MSBP) introduced by Dunson, Xue, and Carin (2008).

The layout of the paper is as follows. We first review the MSBP and specify some of its key properties that will be influential in the development of our priors. We propose a class of priors on the set of covariance matrices by specifying a prior on the set of GARPs and a prior for the set of IVs. Section 3 considers the priors for the GARPs and IVs. We explore several resulting properties of the GARP and IV parameters in Section 4. We remark on some of the bias and

efficiency issues regarding the mean function under covariance estimation in Section 5, and then describe the computational issues involved in setting up a Markov chain Monte Carlo (MCMC) scheme to perform inference using these priors in Section 6. In Section 7 the risk performance of the Bayes estimators resulting from the grouping priors are compared with other methods. In Section 8 we apply our priors to a longitudinal dataset obtained from a study of depression patients. We conclude with some discussion about further areas of exploration.

2. Review of the MSBP

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. The MSBP 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 correlations among the F_{mj} measures that in turn produce desirable properties on the model parameters θ_m. The measure F_{mj} has the following form

$$F_{mj} = \sum_{h=1}^{H} \pi_{mjh} \delta_{\xi_{jh}}, \quad \xi_{jh} \sim \text{i.i.d. } F_{0j},$$

where $\boldsymbol{\Xi} = \{\xi_{jh}\}$ is a $p \times H$ matrix of random elements and δ_x represents a point mass at x. The rows of $\boldsymbol{\Xi}$ ($j = 1, \ldots, p$) correspond to each of the model parameters, which have a nonatomic base distribution F_{0j}. The H columns are referred to as the clusters. We sometimes refer to the elements of $\boldsymbol{\Xi}$ 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

$$\pi_{mjh} = V_{mjh} \prod_{l<h} (1 - V_{mjl}), \quad V_{mjh} = U_{mh} X_{jh}, \quad U_{mh} \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad X_{jh} \overset{iid}{\sim} \text{Beta}(1, \beta).$$

We see that the V_{mjh} is the product of U_{mh}, which controls the likelihood that parameters for group m come from cluster h, and X_{jh}, controlling the likelihood that parameter j is drawn from cluster h. Because U_{mh}'s are shared across parameters and the X_{jh} across groups, they induce the dependence among the probability measures of $\mathcal{F}$. We 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 F_{mj} is a valid distribution.

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

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

So we may choose the number of clusters H such that this expected approximation error (1) is arbitrarily small, so that the MSBP$_H$ truncation is a good approximation to MSBP.

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 the stick-breaking parameters α and β.

We finally recall the model behavior for some special cases of the stick-breaking parameters. In the following, we assume $H = \infty$. As $\alpha \to 0$, Beta(1,α) $\to \delta_1$, and so for all m, $V_{mjh} = X_{jh}$ almost surely (a.s.). Thus, for all m, $F_{mj} = \sum_{h=1}^{\infty} X_{jh} \prod_{l<h}(1 - X_{jl})\delta_{\xi_{jh}} = F_j$, which implies $F_j \sim DP(\beta F_{0j})$ and that F_{mj} are equal across all groups. This is equivalent to placing

independent Dirichlet process priors on each of the parameters with no sharing of information across parameters. Hence, α controls that amount of information shared across parameters, with small values of α indicating little borrowing of information. In the case where $\alpha \to 0$ and $\beta \to \infty$, $\theta_{mj} \sim F_{0j}$, that is, the prior for the parameters collapses to the base distribution. If both $\alpha, \beta \to 0$, we make no use of the groupings and treat the data as a single population with global parameters $\theta_j = \theta_{mj}$ for all m.

3. Grouping Priors for GARPs and IVs

We now propose priors to use for simultaneous covariance estimation based on the MSBP. These priors are referred to as grouping priors because of the way the MSBP induces grouping among the values of the various parameters. To this end, we independently place a MSBP-type prior on the set of GARPs $\Phi = \{\Phi_1, \ldots, \Phi_M\}$ and another MSBP-type prior on the set of IVs $\Gamma = \{\Gamma_1, \ldots, \Gamma_M\}$. The prior on Σ is induced by the mapping defined by $\Sigma_m = \Sigma(\Phi_m, \Gamma_m)$, which forms the covariance matrix corresponding to GARPs Φ_m and IVs Γ_m. Because the GARPs and IVs are orthogonal parameters (Pourahmadi, 2007), it is sensible to choose priors with Φ independent of Γ. In this section we discuss three priors for Φ and a pair of priors on Γ.

3.1. Sparsity Grouping Prior for Φ

The first prior proposal for the GARPs, referred to as the sparsity grouping prior, is defined as follows.

$$\begin{aligned}
\phi_{mj} &\sim F_{mj}(\cdot) = \sum_{h=1}^{H_\phi} \pi_{mjh}\, \delta_{\xi_{jh}}(\cdot), \;\; m = 1, \ldots, M, \; j = 1, \ldots, J; && (2)\\
\xi_{jh} &\sim \epsilon_{q(j)}\delta_0 + (1 - \epsilon_{q(j)})\mathrm{N}(0, \sigma^2), \;\; j = 1, \ldots, J, \; h = 1, \ldots, H_\phi; && (3)\\
\pi_{mjh} &= U_{mh}X_{jh}\prod_{l<h}(1 - U_{ml}X_{jl}) \;\; \text{for all } m, j, h;\\
U_{mh} &\sim \text{Beta}(1, \alpha_\phi), \; h = 1, \ldots, H_\phi - 1, \;\; U_{mH\phi} \sim \delta_1, \;\; m = 1, \ldots, M;\\
X_{jh} &\sim \text{Beta}(1, \beta_\phi), \; h = 1, \ldots, H_\phi - 1, \;\; X_{jH\phi} \sim \delta_1, \;\; j = 1, \ldots, J.
\end{aligned}$$

This prior assumes each ϕ_{mj} is drawn from the random probability measure F_{mj} in (2). The distribution (3) of ξ_{jh}, the candidate values for the jth GARP, is a mixture of a mean-zero normal and a distribution degenerate at zero. Here, $q(\cdot) : \{1, \ldots, J\} \mapsto \{1, \ldots, p-1\}$ denotes the function that gives the lag value associated with the GARPs. Note there are $(p-1)$ of the ϵ_q, each of which represents the probability that ξ_{jh} will be zero for the lag-q GARPs. The presence of the zero point mass promotes sparsity in $T(\Phi_m)$, and because a zero GARP parameter 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 the MSBP. 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 GARP stick-breaking parameters from the IV parameters to be defined later. Note α_ϕ is the defining parameter for the U_{mh}'s, which control the likelihood that the group m GARPs come from cluster h. While the β_ϕ is the parameter associated with X_{jh}, influencing the probability the GARP j will be drawn from cluster h. The previously described special cases as α_ϕ and β_ϕ converge to zeros and infinity

are still applicable. In the following section, we will derive further properties of the distribution of the GARP parameters that will be functions of these two stick-breaking parameters.

This sparsity grouping prior for Φ is defined similarly to the MSBP, with the key difference being that Dunson et al. (2008) specify that the base distribution of the ξ_{jh}'s be nonatomic. This is not the case with our prior since we use a distribution for the ξ_{jh}'s that contain a point mass at zero. This does not lead to a problem, but it does alter some of the theoretic properties as discussed in Section 4.

3.2. Non-sparse Grouping Prior for Φ

We now consider the non-sparse grouping prior, which is a slight alteration of the sparsity grouping prior obtained by dropping the zero point mass. The non-sparse grouping prior for Φ_m is defined by replacing equation (3) in the sparsity prior by

$$\xi_{jh} \sim \mathrm{N}(0, \sigma^2).$$

Having removed the zero point mass from the ξ-level, ϕ_{mj} is non-zero a.s. Thus, we no longer allow for conditional independence relationships or gain sparsity in the $T(\Phi)$ matrix, but by fixing the mean at zero for the distribution of ξ, we still encourage shrinkage toward independence. This distribution for Φ follows exactly the MSBP framework of Dunson et al. (2008).

3.3. Lag-block Sparsity Grouping Prior for Φ

Having developed a pair of priors for Φ which follow closely to the MSBP, we consider a variant that differs more significantly from the MSBP framework and is better suited for covariance matrices in an ordered data setting (e.g., longitudinal data).

As will be shown in the Section 4.1, the distributions F_{mj} and $F_{mj'}$ of the parameters ϕ_{mj} and $\phi_{mj'}$ (different GARPs for a common group) are uncorrelated for the sparsity and non-sparse

grouping priors. This is because the only information shared across parameters is whether the parameter values for different groups are drawn from the same cluster. No information relating to the value of the parameter is borrowed. This may be viewed unfavorably in the context of covariance estimation, as one might expect that different GARPs that correspond to a common lag value to be similar, or even equal. For example, it may be reasonable to consider that the regression effect of Y_{t-1} onto Y_t be the same for different values of t. By deviating from the MSBP, we develop a new prior for Φ which we refer to as the lag-block sparsity grouping prior. This will induce a correlation structure between F_{mj} and $F_{mj'}$ when GARP j and j' correspond to the same lag value, $q(j) = q(j')$.

The lag-block prior consists of replacing equations (2) and (3) from the sparsity prior with

$$\begin{aligned} \phi_{mj} &\sim F_{mj}(\cdot) = \sum_{h=1}^{H_\phi} \pi_{mjh}\, \delta_{\xi_{q(j)h}}(\cdot), \;\; m = 1, \ldots, M, \; j = 1, \ldots, J; \\ \xi_{qh} &\sim \epsilon_q \delta_0 + (1 - \epsilon_q)\mathrm{N}(0, \sigma^2), \;\; q = 1, \ldots, p-1, \; h = 1, \ldots, H_\phi. \end{aligned}$$

Note that for this prior each cluster h no longer has a parameter associated with each GARP, but only a ξ-term for each of the $p - 1$ lags. This means that all of the lag-q GARP parameters are drawn from the same set of H_ϕ candidate values. In Section 4.2, we show that this yields a non-zero correlation between F_{mj} and $F_{mj'}$ and a non-zero probability that ϕ_{mj} and $\phi_{mj'}$ are equal, for j and j' of the same lag, unlike the previous priors.

3.4. InvGamma Grouping Prior for Γ

We now consider two priors for the innovation variances Γ, the first of which follows from the MSBP. We define the InvGamma grouping prior to be

$$\begin{aligned}
\gamma_{mj} &\sim G_{mj}(\cdot) = \sum_{h=1}^{H_\gamma} \tau_{mjh}\, \delta_{\eta_{jh}}(\cdot), \;\; m = 1, \ldots, M, \; j = 1, \ldots, p; && (4)\\
\eta_{jh} &\sim \text{InvGamma}(\lambda_1, \lambda_2), \;\; j = 1, \ldots, p, \; h = 1, \ldots, H_\gamma; && (5)\\
\tau_{mjh} &= W_{mh} Z_{jh} \prod_{l<h} (1 - W_{ml} Z_{jl}) \;\; \text{for all } m, j, h;\\
W_{mh} &\sim \text{Beta}(1, \alpha_\gamma), \; h = 1, \ldots, H_\gamma - 1, \;\; W_{mH_\gamma} \sim \delta_1, \;\; m = 1, \ldots, M;\\
Z_{jh} &\sim \text{Beta}(1, \beta_\gamma), \; h = 1, \ldots, H_\gamma - 1, \;\; Z_{jH_\gamma} \sim \delta_1, \;\; j = 1, \ldots, p.
\end{aligned}$$

We draw the IV γ_{mj} from the stick-breaking measure G_{mj}, where the candidate atoms are drawn from an inverse Gamma distribution (InvGamma). We choose the InvGamma distribution for η to exploit conjugacy for the IVs, as will be described in the web appendix. We use the parameterization of InvGamma where the λ_2 parameter defines the rate of the distribution. 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.

3.5. Correlated-logNormal Grouping Prior for Γ

Recall that lag-block grouping prior moves beyond the MSBP framework to more closely align with our intuition of the behavior of the GARPs. In a similar spirit we now expand beyond the previously defined prior for Γ to develop a more sophisticated prior which allows us to exploit some reasonable assumptions about the innovation variances. One sometimes considers the IVs 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.

To form the correlated-logNormal prior for Γ we replace line (5) in the InvGamma grouping prior specification with the following

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

The development in this prior is that the candidate IVs η_{jh} are drawn in a correlated fashion and marginally follow a logNormal distribution, providing the name of this prior, the correlated-logNormal grouping prior. We introduce the intermediate variable ω_h which is a p-dimensional normally distributed random vector with mean vector $\psi \mathbf{1}_p$ 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 auto-regressive function of order 1. The (i, j) component of $R(\rho)$ is $\rho^{|i-j|}$. We let η_{jh} be the exponentiated version of ω_{jh} so that marginally η_{jh} has a logNormal distribution.

In the previous InvGamma grouping prior, the distribution of η relied on two hyperparameters, λ_1 and λ_2. The correlated-logNormal prior relies on three hyperparameters to define the distribution of η: ψ, Ω, and ρ. While one might be tempted to use an unstructured mean and/or covariance matrix for ω_h, this does not appear to be feasible due to lack of information one has to estimate these hyperparameters at this level of the model. Hence, we choose the common mean and AR(1) covariance, which still leads to improved estimation as will be shown in the risk simulations of Section 7.

Note that in the special case where $\rho = 0$, that the components of the ω_h vector are independent. Consequently, the IV candidates η_{jh} are i.i.d. from the logNormal(ψ, Ω) distribution, and so this special case follows the MSBP framework. Comparing the two MSBP priors, the InvGamma grouping prior and the correlated-logNormal grouping prior with $\rho = 0$, we recommend using the InvGamma because of the conjugacy that is obtained. The benefit of using the normal (equivalently, logNormal) distribution for ω (η) is that inducing the correlation inside a cluster is straightforward, a gain that outweighs the loss of conjugacy. See Section 6 and Web Appendix 2 for more details

on the conditional distributions and computational issues involved in using this prior.

4. Theoretical 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 MSBP is formally defined as the limit as the number of clusters approaches infinity, and the finite number of clusters case, while necessary for implementation, is viewed as an approximation. Our grouping priors follow in the same way. While defined using finite H_ϕ and H_γ for computation, we view them as approximations to the distributions defined in the limiting cases when $H_\phi, H_\gamma \to \infty$. The following properties 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 these properties are provided in the Web Appendix 1.

4.1. Properties for the GARPs for the Sparsity and Non-Sparse Grouping Priors

First, we consider the behavior of the GARPs from the sparsity grouping prior. For the following calculations, we assume that the ϵ_q's 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,σ^2) distribution. Define $\Psi(\cdot) = \epsilon\delta_0(\cdot) + (1-\epsilon)\Phi(\cdot)$, the probability measure for the mixture distribution of the ξ_{jh}'s.

1. For all sets A in the Borel field of the real line $\mathcal{B}(\mathcal{R})$, E$(F_{mj}(A)) = \Psi(A)$. This property shows that it is appropriate to refer to the δ_0-normal mixture Ψ as the base distribution for the GARPs. This unbiasedness property also holds for the finite H_ϕ case, as can be seen in the derivation.

2. For $A \in \mathcal{B}(\mathcal{R})$, $\text{Var}\left(F_{mj}(A)\right) = \frac{2}{(2+\alpha)(2+\beta)-2}\Psi(A)\left(1-\Psi(A)\right)$.

   We see that α and β control the extent to which the random measure F_{mj} differs from the base distribution. As either α or β approach infinity, the distribution of ϕ_{mj} collapses to the parametric base Ψ. Small values of α and β allow for a more flexible prior.

3. For $A \in \mathcal{B}(\mathcal{R})$ and $m \neq m'$, $\text{Corr}\left(F_{mj}(A), F_{m'j}(A)\right) = \frac{\alpha+\alpha\beta/2+\beta+1}{2\alpha+\alpha\beta+\beta+1}$.

   This quantity is the correlation between the distributions of the GARP j for two groups m and m'. Because this does not depend on the choice of Borel set A, it may be used as a simple univariate measure of the degree to which information is shared across groups. Simple algebra shows that $1/2 \leq \text{Corr}\left(F_{mj}(A), F_{m'j}(A)\right) \leq 1$. In particular, $\text{Corr}\left(F_{mj}(A), F_{m'j}(A)\right) \to 1/2$ as either α or β approach infinity. Additionally, $\text{Corr}\left(F_{mj}(A), F_{m'j}(A)\right) \to 1$ as $\alpha \to 0$.

4. Having considered the correlation for the distributions of the GARP j parameters for two different groups, we derive the probability of equality for these two parameters. For $m \neq m'$, $Pr(\phi_{mj} = \phi_{m'j}) = \epsilon^2 + \frac{1-\epsilon^2}{(1+\alpha)(2+\beta)-1}$.

   The presence of the zero point mass causes these clustering properties to differ from those derived in Dunson et al. (2008). Note that as either α or β approach infinity, this probability approaches ϵ^2, which is the probability that both ϕ_{mj} and $\phi_{m'j}$ are equal to zero if drawn from the parametric base distribution Ψ. This probability of equality is increasing in ϵ, which is intuitive since larger values of ϵ indicate that both terms are more likely to be zero whether or not they come from the same cluster. Additionally, this probability increases as either α and β decreases coinciding with the increase in $\text{Corr}\left(F_{mj}(A), F_{m'j}(A)\right)$.

5. In Property 3., the correlation between the distributions of two parameters representing the same GARP but different groups was considered. We now study the relationship between the distributions for two different GARPs of the same group. For $A \in \mathcal{B}(\mathcal{R})$ and $j \neq j'$, $\text{Corr}(F_{mj}(A), F_{mj'}(A)) = 0$, that is, F_{mj} and $F_{m'j}$ are uncorrelated.

Additionally, for $m \neq m'$, Corr$(F_{mj}(A), F_{m'j'}(A)) = 0$, that is, the distributions of two different GARPs of different groups are also uncorrelated.

6. Property 4. considered the probability of equality among two GARP parameters with common j and differing group m. The probability of equality for two different GARPs j and j' of the same group m relies only on the probability of setting each to zero independently. For $j \neq j'$, $Pr(\phi_{mj} = \phi_{mj'}) = \epsilon_{q(j)}\epsilon_{q(j')} = Pr(\phi_{mj} = 0)\, Pr(\phi_{mj'} = 0)$.

   This implies that there is no sharing strength in the estimation of different GARPs of the same group, even if the GARPs share a common lag value $(q(j) = q(j'))$. This, in conjunction with the preceding property, further motivates the development of the lag-block grouping prior.

   If we consider two different groups m, m', we again obtain $Pr(\phi_{mj} = \phi_{m'j'}) = \epsilon_{q(j)}\epsilon_{q(j')}$.

We additionally point out that the non-sparse grouping prior is equivalent to the sparsity grouping prior if we take $\epsilon_q = 0$ for all q. Hence, the respective properties for this prior are obtained by substituting $\epsilon = 0$ (and consequently, $\Psi(\cdot) = \Phi(\cdot)$) in the above properties. Because the non-sparse prior is a MSBP prior with a non-atomic base distribution, the properties may also be taken from Propositions 1, 2, and 4 in Dunson et al. (2008).

4.2. Properties for the GARPs for the Lag-block Sparsity Grouping Prior

We also explore some of the theoretical properties of the lag-block prior. Properties 1.–4. are the same as for the sparsity grouping prior so we do not list them again. The critical deviation from the sparsity grouping properties is the change for Properties 5.-8.

5. For $A \in \mathcal{B}(\mathcal{R})$ and $j \neq j'$ with $q(j) = q(j')$, Corr $(F_{mj}(A), F_{mj'}(A)) = \frac{\beta+\alpha\beta/2+\alpha+1}{2\beta+\alpha\beta+\alpha+1}$.
   For $j \neq j'$ with $q(j) \neq q(j')$, Corr$(F_{mj}(A), F_{mj'}(A)) = 0$.

6. For $j \neq j'$ with $q = q(j) = q(j')$, $Pr(\phi_{mj} = \phi_{mj'}) = \epsilon_q^2 + \frac{1-\epsilon_q^2}{(2+\alpha)(1+\beta)-1}$.
   For $j \neq j'$ with $q(j) \neq q(j')$, $Pr(\phi_{mj} = \phi_{mj'}) = \epsilon_q \epsilon_{q'}$.

Properties 5. and 6. consider the correlation and matching probabilities for different GARPs j and j' for a common group m. Under the lag-block prior we have imposed a correlation structure on the distribution functions of the GARP parameters of a common lag. As mentioned during the introduction of this prior, we are now borrowing strength in the estimation of GARP parameters from the same lag. The correlation is guaranteed to lie in the interval $[1/2, 1]$ and does not depend on the chosen Borel set A, only on the parameters α_ϕ and β_ϕ. The correlation is the same as that for Property 3. with the role of α_ϕ and β_ϕ in reverse. Likewise, the probability of matching across GARPs of common lag in Property 6. is also equivalent to the probability of matching across group for common GARP in Property 4. with α_ϕ and β_ϕ exchanged.

We compare the correlations between the distributions of the same group to distributions of the same GARP j. Let $m \neq m'$, and $j \neq j'$ with $q(j) = q(j')$, and assume $\alpha_\phi > \beta_\phi$. Then,

$$\begin{aligned} \text{Corr}\left(F_{mj}(A), F_{m'j}(A)\right) &< \text{Corr}\left(F_{mj}(A), F_{mj'}(A)\right), \\ Pr(\phi_{mj} = \phi_{m'j}) &< Pr(\phi_{mj} = \phi_{mj'}). \end{aligned}$$

If α_ϕ is the larger then β_ϕ, then there is more similarity in the distributions of common group and differing GARP than for the distributions of differing group but common GARP. The probability that ϕ_{mj} agrees with other GARPs of its group is greater than that of matching other groups' GARPs for the same j. If $\alpha_\phi < \beta_\phi$, then the inequalities reverse.

It had previously been unnecessary to consider correlations or matching simultaneously across group and GARP for the previous priors, because there was no influence across GARP. As seen in the previous two properties, that is no longer the case for the lag-block prior, so it becomes beneficial to look at the behavior of these quantities here.

7. For $A \in \mathcal{B}(\mathcal{R})$, $m \neq m'$, and $j \neq j'$ with $q = q(j) = q(j')$,

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

   If j and j' do not correspond to a common lag (i.e., $q(j) \neq q(j')$), then $F_{mj}(A)$ and $F_{m'j'}(A)$ are uncorrelated.

8. For $m \neq m'$ and $j \neq j'$ with $q = q(j) = q(j')$, $Pr(\phi_{mj} = \phi_{m'j'}) = \epsilon_q^2 + \frac{1-\epsilon_q^2}{2(1+\alpha)(1+\beta)-1}$.
   If $q(j) \neq q(j')$, then $Pr(\phi_{mj} = \phi_{m'j'}) = \epsilon_q \epsilon_{q'}$.

It is instructive to compare the correlations and matching probabilities across both group and GARP to the respective quantities when only group or only GARP differ. Simple algebra gives that for all Borel sets A, $m \neq m'$, and $j \neq j'$ with $q = q(j) = q(j')$,

$$\begin{aligned}
\text{Corr}\left(F_{mj}(A), F_{m'j'}(A)\right) &< \text{Corr}\left(F_{mj}(A), F_{mj'}(A)\right), \\
\text{Corr}\left(F_{mj}(A), F_{m'j'}(A)\right) &< \text{Corr}\left(F_{mj}(A), F_{m'j}(A)\right), \\
Pr(\phi_{mj} = \phi_{m'j'}) &< Pr(\phi_{mj} = \phi_{mj'}), \\
Pr(\phi_{mj} = \phi_{m'j'}) &< Pr(\phi_{mj} = \phi_{m'j}).
\end{aligned}$$

That is, the correlations of the distribution functions and the probability of matching across both groups and GARP parameters both are strictly smaller than the correlation and matching probability across just one. We noted earlier that Corr $(F_{mj}(A), F_{m'j}(A))$ and Corr $(F_{mj}(A), F_{mj'}(A))$ are each contained in $[1/2, 1]$. Comparatively, the lower bound for Corr $(F_{mj}(A), F_{m'j'}(A))$ is $1/4$. The lag-block specification induces a positive correlation between each pair of distributions of a common lag.

We conclude the discussion of the GARP 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 MSBP (and hence, the sparsity and non-sparse prior). As $\alpha \to 0$, it remains true that $F_{mj} = F_j$ for all m and that $F_j \sim DP(\beta F_{0j})$. 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 both α and β approach zero, the data is pooled such that a common parameter value is assigned for all groups, i.e. $\phi_{mj} = \phi_j$. Due to the lag-block structure, we will now have a common parameter across all groups and all GARPs of a common lag, $\phi_{mj} = \phi_q$ for all m and j with $q = q(j)$.

4.3. Innovation Variance Properties

We now explore the behavior of the IVs and their distributions G_{mj}. Let $\mathcal{R}_+$ denote the positive real line, and $\text{I}\Gamma(\cdot)$ denote the probability function of the InvGamma(λ_1, λ_2) distribution, with fixed values for the hyperparameters. Because the InvGamma grouping prior is a MSBP prior, the relevant properties follow immediately from Dunson et al. (2008). We list them here for ease of comparison with the properties obtained from the correlated-logNormal prior.

1. For $A \in \mathcal{B}(\mathcal{R}_+)$, $\text{E}(G_{mj}(A)) = \text{I}\Gamma(A)$.
2. For $A \in \mathcal{B}(\mathcal{R}_+)$, $\text{Var}(G_{mj}(A)) = \frac{2}{(2+\alpha)(2+\beta)-2}\,\text{I}\Gamma(A)\,(1 - \text{I}\Gamma(A))$.
3. For $A \in \mathcal{B}(\mathcal{R}_+)$ and $m \neq m'$, $\text{Corr}(G_{mj}(A), G_{m'j}(A)) = \frac{\alpha+\alpha\beta/2+\beta+1}{2\alpha+\alpha\beta+\beta+1}$.
4. For $m \neq m'$, $Pr(\gamma_{mj} = \gamma_{m'j}) = \frac{1}{(1+\alpha)(2+\beta)-1}$.
5. For $A \in \mathcal{B}(\mathcal{R}_+)$, $j \neq j'$ and $1 \leq m, m' \leq M$ (possibly equal),
   $\text{Corr}\,(G_{mj}(A), G_{m'j'}(A)) = 0$.
6. For $j \neq j'$ and $1 \leq m, m' \leq M$, $Pr(\gamma_{mj} = \gamma_{m'j'}) = 0$.

We additionally consider the properties when we specify the correlated-logNormal prior for the IVs. Let $\log A$ be the set $\{\log x : x \in A\}$ for any $A \in \mathcal{B}(\mathcal{R}_+)$ and $\Phi(\cdot)$ the probability function for the N(ψ, Ω) distribution, assuming the hyperparameters ψ, Ω are fixed.

1. For $A \in \mathcal{B}(\mathcal{R}_+)$, $\text{E}(G_{mj}(A)) = \Phi(\log A)$.

2. For $A \in \mathcal{B}(\mathcal{R}_+)$, $\text{Var}(G_{mj}(A)) = \frac{2}{(2+\alpha)(2+\beta)-2}\Phi(\log A)\,(1-\Phi(\log A))$.

   Properties 1. and 2. vary from the MSBP-InvGamma prior due to changing the distribution of η from InvGamma to (marginally) logNormal.

3. For $A \in \mathcal{B}(\mathcal{R}_+)$ and $m \neq m'$, $\text{Corr}(G_{mj}(A), G_{m'j}(A)) = \frac{\alpha+\alpha\beta/2+\beta+1}{2\alpha+\alpha\beta+\beta+1}$.

4. For $m \neq m'$, $Pr(\gamma_{mj} = \gamma_{m'j}) = \frac{1}{(1+\alpha)(2+\beta)-1}$.

   The correlation between distributions and the probability of matching across groups for a common time point has the same structure for both of these IV priors.

5. For $A \in \mathcal{B}(\mathcal{R}_+)$ and $j \neq j'$,

$$\text{Corr}\left(G_{mj}(A), G_{mj'}(A)\right) = \frac{\beta + \alpha\beta/2 + \alpha + 1}{2\beta + \alpha\beta + \alpha + 1}\,\text{Corr}\left(I\{\omega_{j1} \in \log A\},\, I\{\omega_{j'1} \in \log A\}\right).$$

6. For $A \in \mathcal{B}(\mathcal{R}_+)$, $j \neq j'$, and $m \neq m'$,

$$\text{Corr}\left(G_{mj}(A), G_{m'j'}(A)\right) = \frac{\alpha\beta/2 + \alpha + \beta + 1}{2\alpha\beta + 2\alpha + 2\beta + 1}\,\text{Corr}\left(I\{\omega_{j1} \in \log A\},\, I\{\omega_{j'1} \in \log A\}\right).$$

   It is no longer the case that the correlation of these distributions is independent of 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 $(\omega_{j1}, \omega_{j'1}) \sim \text{N}_2\left(\psi\mathbf{1}_2, \Omega R^*(\rho)\right)$, where $[R^*(\rho)]_{(1,1)} = [R^*(\rho)]_{(2,2)} = 1$ and $[R^*(\rho)]_{(1,2)} = [R^*(\rho)]_{(2,1)} = \rho^{|j-j'|}$. The higher correlations for neighboring IVs (for $\rho > 0$) implies a smoothing of the IVs as a function of j. We observe that the leading term gives the same correlation structure as the common-lag GARPs when using the lag-block prior.

7. For $j \neq j'$ and $1 \leq m, m' \leq M$, $Pr(\gamma_{mj} = \gamma_{m'j'}) = 0$. With the correlated-logNormal grouping prior there is no matching of IVs across time points. This is a consequence of the fact that two points drawn from a correlated normal distribution ($|\rho| < 1$) will be equal with probability zero. However, as is apparent from the previous two correlation properties, they can be arbitrary close depending on ρ, ψ, Ω and the choices of A.

5. Effect of Covariance Estimation on the Mean Function

In this section, we briefly discuss the importance of covariance estimation on the mean structure. In the complete data case the mean and covariance parameters are orthogonal in the sense of Cox and Reid (1987) and the estimates of the mean parameters will be consistent under mis-specification of the covariance structure. However, if there is missingness in the data, there is no longer orthogonality, even at the true value of the covariance matrix (Little and Rubin, 2002). Hence, for the posterior distribution of the mean parameters to be consistent, the dependence structure must be correctly specified. So, even in the missing at random case (MAR), where missingness depends only on the observed values not the unobserved data, it is no longer appropriate to treat the covariance structure as a nuisance parameter. In this case biased mean estimates can result if we do not use the correct model for the dependence. For further discussion of the mean-covariance issues that arise in incomplete data modeling, see Daniels and Hogan (2008, Section 6.2).

Although the mean and dependence are asymptotically independent in the complete data case, efficiency gains may be possible for small or moderately sized, fully-observed samples. Through four simulation examples with relatively small sample sizes, Cripps et al. (2005) demonstrated improvements in estimating regression coefficients, fitted values, and the predictive density for the Wong et al. (2003) covariance selection prior over a more dispersed covariance prior choice.

We examine the impact of covariance estimation on the means in the simulations (Section 7) and the data example (Section 8).

6. Computational Considerations

6.1. Selecting the Number of Clusters

Recall that equation (1) provided us with the expected approximation error which we employ to choose the number of clusters necessary for the MSBP truncation. This formula continues to hold for each of the proposed grouping priors, since the stick-breaking weights are always formed using the MSBP framework. Hence, if the values of α, β (for either the GARP or IV parameters) are assumed known, then we choose the number of clusters H such that $[1 - (1+\alpha)^{-1}(1+\beta)^{-1}]^{H-1}$ 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 Gamma(1,1) priors for α and β. To choose the value of H when using a prior on for the stick-breaking parameters, we run the MCMC chain for approximately 10% of its total length and use the posterior means to test whether (1) is below our threshold.

Dunson et al. (2008) point out that for α, β values less than 1, $H = 20$ clusters leads to an expected approximation error less than 0.01. They recommend using stick-breaking parameter values less than 1, but it has been our experience through various simulation studies and data analyses that this need not strictly be the case. In many examples using the Gamma(1,1) priors, we received posterior means for the stick-breaking parameters between 1 and 2. While these situations do require more clusters ($20 \leq H \leq 40$) to give a small approximation error, they continue to give improved covariance estimation versus competing priors.

6.2. Using the Grouping Priors in MCMC Analysis

One of the nice properties of the MSBP prior is that by introducing appropriate latent variables an MCMC algorithm can be devised that employs well-known conjugate distributions that are easy to

sample from (Dunson et al., 2008). Because a normal prior for the GARPs and an InvGamma prior for the IVs provide conjugacy, the sampling for ξ and η are from normal (or a zero-normal mixture) and InvGamma distributions, respectively, for the sparsity, non-sparse, lag-block, and InvGamma grouping priors. The conjugacy for the IV candidates is lost for the correlated-logNormal prior, but we can sample η efficiently by incorporating a slice sampling step (Neal, 2003) or another sampling technique. We review the algorithm of Dunson et al. (2008) and further discuss the implementation of a posterior sampling scheme for our grouping priors and other computational challenges in Web Appendix 2.

7. Risk Simulations

We now examine the operating characteristics of the proposed grouping priors via risk simulations. For the purposes of comparison with the proposed grouping priors, we introduce some additional naive priors based on the modified Cholesky parameterization of the covariance matrix. We consider two priors (NB1, NB2) for the GARPs and one prior (NB) for the IVs, defined as follows.

$$\begin{aligned} \text{Naive Bayes 1 (NB1):} \quad & \phi_{mj} \sim \epsilon_{q(j)}\delta_0 + (1-\epsilon_{q(j)})\text{N}(0,\sigma^2) \\ \text{Naive Bayes 2 (NB2):} \quad & \phi_{mj} \sim \text{N}(0,\sigma^2) \\ \text{Naive Bayes (NB):} \quad & \gamma_{mj} \sim \text{InvGamma}(\lambda_1,\lambda_2) \end{aligned}$$

These priors are simple choices for the GARPs and IVs so that conjugacy is maintained, leading to relatively simple MCMC algorithms.

The NB1 prior is the model that follows as α_ϕ approaches zero with β_ϕ going to infinity for the sparsity grouping prior. This clearly corresponds to independently (given the hyperparameters) drawing from the parametric model of ϕ. While not allowing any kind of grouping in the GARP parameters, this model will induce a level of sparsity into the $T(\Phi_m)$ matrices, as does the sparsity grouping prior.

The naive Bayes 2 prior removes the point mass portion from NB1 and, when used in conjunction with NB for the IVs, is a simplified version of the prior suggested by Daniels and Pourahmadi (2002). NB2 is the limiting case of the non-sparse prior when α_ϕ converges to zero and β_ϕ approaches infinity. Since NB2 does not have the zero point mass for the GARPs, this prior will not induce sparsity in the $T(\Phi_m)$ matrices. However, since the prior mean for the ϕ_{mj} is fixed at zero, there is shrinking of the conditional dependencies toward zero.

We note that the NB prior for the innovation variances gives the model corresponding to the InvGamma grouping prior when $\alpha_\gamma \to 0$ and $\beta_\gamma \to \infty$.

For each of the simulations, we generate 50 datasets and run an MCMC chain 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: for ϵ_q, independent Unif(0,1); for α_ϕ, β_ϕ, α_γ, β_γ, λ_1, and λ_2, independent Gamma(1,1); $\sigma^2 \sim$ InvGamma(0.1, 0.1). For the correlated-logNormal prior, we use $\Omega \sim$ InvGamma(0.1,0.1) and $\psi \sim$ N(0, $c^2\Omega$), $c^2 = 1000$, and we fix the value of ρ to be 0.75 (for explanation, see Web Appendix 2.5).

We measure the performance of our proposed priors by estimating the risk associated with the Bayes estimators under two common loss functions (Yang and Berger, 1994),

$$\begin{aligned} L_1(\Sigma_m, \hat{\Sigma}_{m1}) &= tr(\Sigma_m^{-1}\hat{\Sigma}_{m1}) - \log|\Sigma_m^{-1}\hat{\Sigma}_{m1}| - p \\ L_2(\Sigma_m, \hat{\Sigma}_{m2}) &= tr\left\{(\Sigma_m^{-1}\hat{\Sigma}_{m2} - I)^2\right\}. \end{aligned}$$

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 weighted average of the losses from the individual covariance matrices, with weights proportional to the group's sample size.

Additionally, to represent two of the more common methods of dealing with this situation, we run the MCMC 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 on this matrix. The group-specific prior places independent flat priors on each of the M groups. The resulting conditional distributions are inverse-Wishart, making this distribution simple to incorporate in the MCMC algorithm.

7.1. Risk Simulation 1

We first consider $M = 5$ (groups) and $p = 4$ (four-dimensional) normally distributed mean-zero random variables. The five covariance matrices are defined by the following specification of the GARP and IV parameters:

$$
\begin{array}{llll}
\Phi_1 = (\ 0.7, & 0.2, & 0.7, & 0, \quad 0.2, \quad 0.7), \\
\Phi_2 = (\ 0.7, & 0, & 0.3, & 0, \quad 0, \quad 0.7), \\
\Phi_3 = (\ 0.3, & 0, & 0.3, & 0, \quad 0, \quad 0.3), \\
\Phi_4 = (\ 0.7, & 0.2, & 0.7, & 0.1, \quad 0.2, \quad 0.7), \\
\Phi_5 = (\ 0.7, & 0, & 0.7, & 0, \quad 0, \quad 0.3),
\end{array}
\qquad
\begin{array}{llll}
\Gamma_1 = (\ 1, & 1, & 1, & 1), \\
\Gamma_2 = (\ 2, & 2, & 2, & 2), \\
\Gamma_3 = (\ 2, & 2, & 1, & 1), \\
\Gamma_4 = (\ 5, & 5, & 5, & 5), \\
\Gamma_5 = (\ 1, & 1, & 2, & 2).
\end{array}
$$

We use the sample sizes of $n_1 = \ldots = n_4 = 30, n_5 = 15$. Note that for this specification many of the parameters across groups are equal and that many of the higher lag GARP terms are zero. Additionally with the smaller sample size for the fifth group, the grouping priors should improve estimation of Σ_5 by sharing information across similar groups. Thus, this should be an ideal situation for our priors. Using the technique suggested in Section 6.1, we specify $H_\phi = H_\gamma = 40$ for the grouping priors.

Risk estimates are given in Table 1. The prior composed of the lag-block structure on the GARPs and the correlated-logNormal specification for the IVs has the best risk estimates of the collection. Comparing the lag-block/InvGamma and sparsity/correlated-logNormal priors to the sparsity/InvGamma grouping prior, the modification on either the GARPs or the IVs produces improved risk performance. The lag-block/correlated-logNormal produces risk estimates that are 15% and 12% lower than the NB1/NB naive prior. It is natural to compare the NB1/NB prior to the sparsity/InvGamma because the first is a limiting case of the latter. Likewise, we compare

NB2/NB and grouping/InvGamma. For both loss functions, the sparsity/InvGamma beats NB1/NB and grouping/InvGamma beats NB2/NB, indicating the borrowing of information across groups induced by the grouping priors improves the estimation. We also see that the sparsity prior performs better than the grouping prior, but this is to be expected since we know that there are GARP parameters that are equal to zero. Comparatively, the estimators from the flat priors perform very poorly; the risks for the grouping priors are 37–47% smaller than the group-specific estimator for L_1 and 30–39% for L_2.

Additional risk simulations were tested with the mean fixed to zero, some of which are included in Web Appendix 3. The grouping prior continued to perform very well 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. Throughout the lag-block/correlated-logNormal prior performed the best with the other grouping priors performing as well or better than the relevant naive choices. The choice of the flat priors continued to perform poorly compared to the grouping (and naive) choices.

7.2. Risk Simulation 2

We now study our estimators in the presence of a more realistic longitudinal scenario. We incorporate a non-zero mean, and the simulated data will suffer from ignorable missingness due to a dropout process. There are $M = 8$ groups with $n_m = 50$ measurements of dimension $p = 6$. Letting D_i denote the time $t = 2, \ldots, p+1$ of dropout for subject i ($D_i = p+1$ indicates a subject who completes the study), the dropout is induced according to the model

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

This missing data mechanism yields data that follow the MAR assumption. The mean, GARP, IV, and dropout parameters for the simulation are listed in Table 2.

This choice of Φ and Γ do not provide any equalities across groups as was present in the previous example. However with the small sample sizes, it will generally still be advantageous to share information across the eight groups. Also, note a moderate amount of sparsity in the GARP parameters, which is typical for ordered data. Table 3 gives the probability of Y_{it} will be unobserved by t and m due to the dropout process (6). We see that groups 3 and 8 experience a large amount of attrition over the study which will have adverse effects on the mean estimation. As noted in Section 5, the improved modeling of the dependence structure provided by our grouping priors should yield improvements in the estimation of the mean function.

For the MCMC chain, additional steps are needed to sample the mean vectors and the missing data. We assume a flat prior on the group-specific mean vectors μ_m; as a result, its full conditional distribution will be multivariate normal. We use data augmentation to sample the missing data values from the (normal) distributions conditional on the observed data.

The estimated risk associated with estimating the covariance matrices for each of the two loss functions is shown in Table 4. With the increased values of p and M, all of the grouping priors beat the naive priors. The ability to borrow strength across groups improves the estimation such that even the non-sparse grouping prior, which does not allow the correct independence relationships, beats the NB1 prior, which correctly incorporates the potential independence. The lag-block/correlated-logNormal prior continues to beat the remainder of the grouping priors, with a risk improvement of 30 and 25% over the NB1/NB prior and 52 and 41% over the group-specific flat prior. From these and other simulation studies, 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. This is particularly important since the number of possible models increases as p and M increase.

In addition to improvement in the estimation of the covariance matrix, we additionally examined the ability to recover the mean structure for each group. Having placed a flat prior on the

group-specific mean vectors μ_m, we measure the accuracy of the means using the loss function $L(\hat{\mu}_m, \mu_m) = (\hat{\mu}_m - \mu_m)'\Sigma_m^{-1}(\hat{\mu}_m - \mu_m)$, which is standardized by the true covariance matrix Σ_m; we use the weighted average of these losses across groups for an overall loss. The final column of Table 4 displays the performance.

There is clear improvement in the mean estimation under the grouping priors. The lag-block/correlated-logNormal prior produces a risk 14% smaller than the NB1/NB 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. By considering flexible, 'parsimonious' priors on the dependence, we see a meaningful improvement in the estimation of mean trajectory.

8. Data Example

We now 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 further analyzed by Pourahmadi and Daniels (2002), consists of weekly measures of depression over a sixteen week study period. Depression scores were measured weekly using the Hamilton Rating Scale for Depression (HRSD). As noted in previous analyses of the dataset, the severity of the depression symptoms at baseline influences the rate of the improvement of HRSD scores. There are two treatments under consideration in the study, a psychotherapy-only treatment versus a treatment regimen which includes 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 HRSD scores (baseline through sixteen weeks) is assumed to be normally distributed with a quadratic mean function and covariance matrix specific to treatment-severity group.

Approximately 30% of the possible measurements from the study are missing. We assume that

the missingness was MAR and incorporate a data augmentation step to sample the missing values given the observed data and the current parameter values, as in the previous risk simulation. The modeling of the mean function is accomplished by assuming that the expected depression score is a group-specific, quadratic function in time.

To compare the fits induced by the various covariance priors used, we use the deviance information criteria (DIC) (Spiegelhalter et al., 2002). Let

$$Dev = Dev(\bar{\mu}, \overline{\Sigma^{-1}} \,|\, y_{obs}) = -2 \,\text{loglik}(\bar{\mu}, \overline{\Sigma^{-1}} \,|\, y_{obs})$$

be the deviance, or twice the negative observed data log-likelihood evaluated at the posterior means of the parameters. Here $\bar{\mu}$ and $\overline{\Sigma^{-1}}$ represents the posterior expectations of μ and Σ^{-1}. Wang and Daniels (2011) recommended DIC based on the observed data likelihood for missing data settings, such as this. The complexity of each of the models is measured by p_D, which can be viewed as the effective number of parameters. The p_D is defined by

$$p_D = \overline{Dev(\mu, \Sigma^{-1} \,|\, y_{obs})} - Dev(\bar{\mu}, \overline{\Sigma^{-1}} \,|\, y_{obs}),$$

where the over-bar notation denotes the posterior mean of the deviances in the first term. Note that we use the precision (inverse of the covariance) matrix in the p_D calculation to obtain a more numerically stable estimate (Spiegelhalter et al., 2002). We form the DIC as our model comparison measure by DIC$= Dev + 2p_D$. Smaller values of DIC indicate a good combination of model fit and simplicity.

We perform an analysis using each of the pairs of grouping covariance priors: sparsity/InvGamma, non-sparse/InvGamma, lag-block/InvGamma, sparsity/correlated-logNormal, and lag-block/correlated-logNormal. 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 Section 6, appropriate values for H_ϕ and H_γ were chosen by running the first 10% of the chain. The number of clusters necessary to approximate the limiting distribution to within 0.01 ranged

from twenty to thirty. For the correlated-logNormal prior, we run an MCMC chain for three fixed values for ρ of 0.5, 0.75, and 0.9 (see Web Appendix 2.5); otherwise, we used the same hyper-priors as in Section 7. The analysis was also performed with the NB1/NB and NB2/NB prior combinations, as well as the common-Σ and group-specific flat priors.

Table 5 contains the model deviance, the effective number of parameters p_D, and the model selection criteria DIC when using various prior choices for Σ. We see that the lag-block/correlated-logNormal with $\rho = 0.9$ prior gives the best fit to the dataset. For the correlated-logNormal prior the largest of the three correlations $\rho = 0.9$ slightly beats the other two when we combine them with either the lag-block or sparsity grouping priors, but the (relatively) small difference in DIC provides further evidence of the robustness of the prior to the choice of ρ. Comparing the model complexities between the NB2/NB prior to the non-sparse/InvGamma prior, as well as the NB1/NB to the sparsity/InvGamma, there is a reduction of 34 and 6 parameters, respectively. Using the grouping priors over their naive Bayes counterparts produces a more structured model. The structure added by allowing conditional independence is even more evident, since the sparsity/InvGamma p_D contains about 54 fewer parameters than the non-sparse/InvGamma. There is also a dramatic improvement in model fit comparing analyses with the lag-block prior to the corresponding analysis with the sparsity prior. This is mainly due to decreased model complexity (i.e. fewer free parameters). We conclude the prior comparison by noting that 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 covariance priors effect the mean estimation. We show the treatment effect (difference in mean value between baseline and week 16) and 95% credible intervals for the first two groups in Table 5. In group $m = 2$ we see that there are clear differences for this effect across the different priors on Φ. We see a treatment effect of around 9.5 points for the four

lag-block analyses, while the four grouping priors with the sparse GARP structure show an effect around 8.7. We see major deviations for the two flat priors, 10.2 for the common-Σ flat and 6.9 for the group-specific flat prior. For group 1 we do not see 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 Section 5. The effects for groups 3 and 4 (not shown) also indicate little difference in treatment effect, as in group 1. We note that the differences do not rise to the level of statistical significance but they are large enough to be of practical importance.

We provide some additional details from the analysis using the lag-block and correlated-logNormal ($\rho = 0.9$) grouping priors. This was the prior that produced the best model fit (according to the DIC criterion). We obtained the following posterior means (and 95% credible intervals) for the stick-breaking parameters: $\hat{\alpha}_\phi = 0.673(0.34, 1.30), \hat{\beta}_\phi = 0.584(0.49, 0.71), \hat{\alpha}_\gamma = 0.454(0.27, 0.78), \hat{\beta}_\phi = 0.431(0.32, 0.60)$. We see that the intervals for these parameters are much smaller than for the Gamma(1,1) prior, indicating substantial learning about the parameter values. The posterior mean for ϵ_1 is 0.109, the mean of ϵ_2 is 0.215, with the remaining means ranging from 0.31 to 0.78. This agrees with our intuition that the first few lags will usually be non-zero with higher lags more likely to be zero.

Figures 1–3 show the grouping nature of the proposed priors. Figure 1 shows the posterior probabilities of $Pr(\gamma_{mj} = \gamma_{m'j})$ for each m, m', j combination. Larger boxes indicate higher grouping probabilities, with the boxes on the $y = x$ diagonal having area one. We see that there is substantial matching for the groups 1 and 2 (the two low initial severity groups), as well as for groups 3 and 4 (the two high initial severity groups), with less matching across the two pairs.

Figures 2 and 3 give the posterior probabilities of matching for the lag-1 and lag-4 GARPs, respectively. We show only the first eight of each due to space limitations. Because we use the lag-

block sparsity grouping prior for Φ, there is a positive probability of equality across GARPs of a common lag. This is indicated in the figure by the presence of large matching probabilities for $j \neq j'$, (e.g. $j = 1$ and $j' = 3$). We note the pairwise probability of equality is very high for all combinations of the $j = 1$ GARPs and the group 2, lag-1 GARPs (and group 3, lag-1 GARPs to a lesser extent). One would be unlikely to learn of this relationship or to consider a model with equality across all (or a large subset) of the parameters ($\phi_{11}, \phi_{21}, \phi_{31}, \phi_{41}, \phi_{23}, \phi_{26}, \phi_{2,10}, \phi_{2,15}, \phi_{2,21}, \phi_{2,28}, \phi_{2,36}$) using other techniques. The black boxes that overlay the $y = x$ diagonal are proportional to the posterior of $Pr(\phi_{mj} = 0)$. We see that the lag-1 GARPs are rarely set to zero.

Considering Figure 3 one notes that the lag-4 GARPs are quite often equal to zero, except for the first group. There are larger matching probabilities for the lag-4 parameters, much of which is due to matching with both parameters set to zero. The matching relationships are not always due to equality with zero, as can be seen from the large probabilities of matching across the group 1 GARPs, as well as in group 4 for GARPs 7, 12, 18, and 25. Recall that the lag-block is the only grouping priors that is able to exploit equality across GARPs within lag, a property that greatly improves the model fit for the depression data.

9. Discussion

In this article we have proposed a new class of priors for sets of covariance matrices. These grouping priors seek to promote sparsity in the Cholesky decomposition of the inverse covariance matrix, as well as all for equality relationships in the parameters across different groups and across GARPs (of the same lag). The theoretical properties explored in Section 4 show that the priors seek to promote sensible structures in Σ, while allowing the data to inform the extent to which the posterior agrees with the structures. The risk simulations of Section 7 and the depression data example of Section 8 clearly show that the proposed priors lead to improved covariance estimation as measured by the estimated risk (for the mean parameters under ignorable missingness as well)

and model fit, respectively. In particular, the lag-block grouping prior for the GARPs and the correlated-logNormal grouping prior on the IVs perform much better than the naive Bayes and flat prior competitors, as well as the other grouping priors. Because each of the grouping priors performs well, one can choose to use grouping priors other than lag-block/corr-logNormal based on the context of the data being used for inference.

In addition to the grouping priors previously defined, there are a number of other natural extensions and possible variations of our grouping priors that one could use. For instance, one could allow for differing values of σ^2, the GARP parameter variance, that depend on the lag value of the associated GARP. This might be beneficial in a situation where p is large and one believes that the GARPs after the first few lags vary more tightly around zero. Additionally, we can remove the sparsity from the lag-block grouping prior by deleting the point mass at zero from the distribution of the ξ's. As another choice, instead of specifying the Φ and Γ as separate blocks with different values of the stick-breaking parameters α and β, one could draw both the GARP and IV terms with the same values of α and β. Instead of specifying that the candidate GARPs ξ are zero according to the probability ϵ, another extension is to modify the grouping prior by introducing a zero-th cluster where $\xi_{j0} = 0$ for all j. The selection of ϕ_{mj} would then follow by $Pr(\phi_{mj} = 0) = Pr(\phi_{mj} = \xi_{j0}) = \epsilon_{q(j)}$ and for $h = 1, \ldots, H$, $Pr(\phi_{mj} = \xi_{jh}) = (1 - \epsilon_{q(j)})\pi_{mjh}$. The properties derived in Section 5 are easily obtained and compared to those obtained in the sparsity grouping case. While these or others may be more natural in certain contexts, we believe that those discussed here are the most applicable priors for general longitudinal data.

We briefly point out some of the potential drawbacks of our grouping priors. The main weakness is the computational time necessary to run the MCMC sampling scheme. We note that the computational time increases as any of M, p, H_ϕ, or H_γ increase. This is due in part to the large number of latent variable used in the MCMC algorithm. While computation time will also increase as the group sample sizes n_m increase, this is comparatively minor, as it does not lead to any ad-

ditional latent variables. While the time to run one chain with a grouping prior may be slower than other methods, model selection on the covariance will generally not be necessary, as our prior encompasses an extraordinarily large collection of models for the GARPs/IVs. For instance, there are $2^{p(p+1)/2}$ different models for Σ when we only consider those that allow each GARP/IV to be either constant across all groups or differ across all groups. When we increase the model space to contain all combinations where each GARP/IV is be constant across all possible subsets of the groups, we have $B_M^{p(p+1)/2}$ models to consider, where B_M is the Mth Bell number (Spivey, 2008) or the sum of the first M Stirling numbers of the second kind. 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 model in a single MCMC analysis and accounting for uncertainty appropriately. It is our belief that running an MCMC chain with one of these grouping prior is a necessary alternative to the unreasonable time and energy required to fit and compare the extremely large class of models.

When modeling multivariate data whose components do not have a natural ordering, it is still possible to use a modified form of these grouping priors. It is inappropriate to make use of the (time) lags between measurements when we are outside of the situation of ordered data, so it would be inadvisable to apply the lag-block prior to non-longitudinal data. Similarly, because we no longer view the IVs as values of a function of time, we do not recommend the correlated-logNormal model for Γ. The sparsity and non-sparse grouping priors with the InvGamma prior on Γ remain reasonable choices for prior specification due to their flexibility, clustering properties, and conjugacy for computations. We finally note that for the sparsity grouping prior we would no longer vary ϵ by lag but use a common value for all ξ.

Throughout this article, we have treated the different groups as exchangeable. Future areas of work include exploring methods to incorporate an ordering of the groups to guide the extent of the clustering. This could be useful in a situation where the groups are defined by increasing strengths

of a treatment or by the dropout time in a pattern mixture model (Little, 1994), and we wish to make use of this ordering to inform the grouping. Additionally, adaptations of the grouping priors could be employed for binary data models with mean and dependence parameters, such as those in the Ising model. Finally, the development of faster algorithms for MCMC sampling would further increase the attractiveness of the grouping priors.

We point out that that the ideas behind the lag-block and correlated-logNormal grouping priors provide a general recipe for extending the MSBP to form nonparametric priors in situations outside of simultaneous covariance estimation. For the lag-block prior, rather than using i.i.d. atoms in the stick-breaking measures for each of the parameters, we required that all of the common-lag GARPs be drawn from the same set of atoms. In the correlated-logNormal prior, the IV candidates η_{jh} were formed so they were correlated across time point j for a common cluster h. In both cases, we formed the stick-breaking weights of these atoms in the same way as in Dunson et al. (2008). However, by choosing the candidate atoms dependently, either through strict equality across sets of the parameters or through specifying correlations in each cluster, we have adapted the MSBP to better fit our specific inference problem. For a different estimation situation, one can devise a way to choose the candidate atoms dependently that is appropriate for that particular situation and form a nonparametric grouping prior by coupling these atoms with stick-breaking weights formed according to the MSBP. If the dependence structure among atoms is reasonable, then this prior should yield more efficient estimation than the MSBP.

Acknowledgments

This work was partially supported by NIH grant CA85295.

References

Boik, R. J. (2002). Spectral models for covariance matrices. *Biometrika*, 89(1):159–182.

Boik, R. J. (2003). Principal component models for correlation matrices. *Biometrika*, 90(3):679–701.

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

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, 49:1–18.

Cripps, E., Carter, C. K., and Kohn, R. (2005). Variable selection and covariance selection in multivariate regression models. In *Handbook of Statistics*, volume 25, pages 519–552.

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. *Journal of Multivariate Analysis*, 97:1185–1207.

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

Daniels, M. J. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89:553–566.

Dunson, D. B., Xue, Y., and Carin, L. (2008). The matrix stick-breaking process: Flexible Bayes meta-analysis. *Journal of the American Statistical Association*, 103(481):317–327.

Fox, E. and Dunson, D. B. (2011). Bayesian nonparametric covariance regression. *Journal of the American Statistical Association*. Submitted.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Hoff, P. and Niu, X. (2011). A covariance regression model. To appear in *Statistica Sinica*.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81:471–483.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Lopes, H. F., McCulloch, R. E., and Tsay, R. S. (2011). Cholesky stochastic volatility. *Submitted*.

Manly, B. F. J. and Rayner, J. C. W. (1987). The comparison of sample covariance matrices using likelihood ratio tests. *Biometrika*, 74:841–847.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767.

Philipov, A. and Glickman, M. E. (2006a). Factor multivariate stochastic volatility via Wishart processes. *Econometric Reviews*, 25(2-3):311–334.

Philipov, A. and Glickman, M. E. (2006b). Multivariate stochastic volatility via Wishart processes. *Journal of Business & Economic Statistics*, 24(3):313–328.

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

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87(2):425–435.

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

Pourahmadi, M. and Daniels, M. J. (2002). Dynamic conditionally linear mixed models for longitudinal data. *Biometrics*, 58(1):225–231.

Pourahmadi, M., Daniels, M. J., and Park, T. (2007). Simultaneous modelling of the cholesky decomposition of several covariance matrices. *Journal of Multivariate Analysis*, 98(3):568 – 587.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64(4):583–639.

Spivey, M. Z. (2008). A generalized recurrence for bell numbers. *Journal of Integer Sequences*, 11.

Thase, M., Greenhouse, J., Frank, E., Reynolds, C., Pilkonis, P., Hurley, K., Grochocinski, V., and Kupfer, D. (1997). Treatment of major depression with psychotherapy or psychotherapy-pharmacotherapy combinations. *Archives of General Psychiatry*, 54:1009–1015.

Wang, C. and Daniels, M. J. (2011). A note on MAR, identifying restrictions, model comparison, and sensitivity analysis in pattern mixture models with and without covariates for incomplete data. *Biometrics*, 67(3):810–818.

Wong, F., Carter, C. K., and Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, 90(4):809–830.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 22:1195–1211.

Priors		Estimated Risk	
GARP	IV	Loss Fcn 1	Loss Fcn 2
Lag-block	Corr-logNormal	0.247	0.429
Lag-block	InvGamma	0.257	0.448
Sparsity	Corr-logNormal	0.270	0.462
Sparsity	InvGamma	0.281	0.480
NB1	NB	0.291	0.488
Non-sparse	InvGamma	0.292	0.493
NB2	NB	0.322	0.530
Group-specific flat		0.463	0.700
Common-Σ flat		1.560	6.623

Table 1: Risk Estimates for Simulation 1

$$
\begin{aligned}
\mu_1 &= (\ 0,\ 1.9,\ 5.2,\ 9.9,\ 16.0,\ 23.5\ )\\
\mu_2 &= (\ 0,\ 1.8,\ 4.8,\ 9.0,\ 14.4,\ 21.0\ )\\
\mu_3 &= (\ 0,\ 1.8,\ 5.6,\ 11.4,\ 19.2,\ 29.0\ )\\
\mu_4 &= (\ 0,\ 2.0,\ 5.0,\ 9.0,\ 14.0,\ 20.0\ )\\
\mu_5 &= (\ 0,\ 2.0,\ 5.2,\ 9.6,\ 15.2,\ 22.0\ )\\
\mu_6 &= (\ 0,\ 3.0,\ 6.0,\ 9.0,\ 12.0,\ 15.0\ )\\
\mu_7 &= (\ 0,\ 1.8,\ 4.8,\ 9.0,\ 14.4,\ 21.0\ )\\
\mu_8 &= (\ 0,\ 2.8,\ 7.2,\ 13.2,\ 20.8,\ 30.0\ )
\end{aligned}
$$

$$
\begin{aligned}
\Phi_1 &= (\ 0.7,\ 0.2,\ 0.7,\ 0,\ 0.2,\ 0.7,\ 0,\ 0,\ 0.2,\ 0.7,\ 0,\ 0,\ 0,\ 0.2,\ 0.7\ )\\
\Phi_2 &= (\ 0.6,\ 0.1,\ 0.6,\ 0.1,\ 0.1,\ 0.6,\ -0.1,\ 0.1,\ 0.1,\ 0.6,\ -0.1,\ -0.1,\ 0.1,\ 0.1,\ 0.6\ )\\
\Phi_3 &= (\ 0.4,\ 0.3,\ 0.4,\ -0.2,\ 0.3,\ 0.4,\ 0,\ -0.2,\ 0.3,\ 0.4,\ -0.2,\ 0,\ -0.2,\ 0.3,\ 0.4\ )\\
\Phi_4 &= (\ 0.3,\ 0,\ 0.3,\ -0.1,\ 0,\ 0.3,\ 0,\ -0.1,\ 0,\ 0.3,\ 0,\ 0,\ -0.1,\ 0,\ 0.3\ )\\
\Phi_5 &= (\ 1,\ -0.5,\ 1,\ 0.2,\ -0.5,\ 1,\ 0,\ 0.2,\ -0.5,\ 1,\ 0,\ 0,\ 0.2,\ -0.5,\ 1\ )\\
\Phi_6 &= (\ 0.8,\ -0.4,\ 0.8,\ 0.3,\ -0.4,\ 0.8,\ 0,\ 0.3,\ -0.4,\ 0.8,\ 0,\ 0,\ 0.3,\ -0.4,\ 0.8\ )\\
\Phi_7 &= (\ 0.9,\ -0.2,\ 1,\ -0.2,\ -0.2,\ 1,\ -0.2,\ -0.2,\ -0.2,\ 1,\ -0.2,\ -0.2,\ -0.2,\ -0.2,\ 1\ )\\
\Phi_8 &= (\ -0.9,\ 0.1,\ -0.9,\ 0,\ 0.1,\ -1,\ 0.2,\ 0,\ 0.1,\ -0.8,\ -0.2,\ 0.2,\ 0,\ 0.1,\ -0.8\ )
\end{aligned}
$$

$$
\begin{aligned}
\Gamma_1 &= (\ 1,\ 1,\ 1,\ 1,\ 1,\ 1\ )\\
\Gamma_2 &= (\ 1.5,\ 1.5,\ 1.5,\ 1.5,\ 1.5,\ 1.5\ )\\
\Gamma_3 &= (\ 3.4,\ 3.1,\ 2.8,\ 2.5,\ 2.2,\ 1.8\ )\\
\Gamma_4 &= (\ 3,\ 3,\ 2,\ 2,\ 2,\ 1\ )\\
\Gamma_5 &= (\ 3.5,\ 3.2,\ 2.9,\ 3.5,\ 3.2,\ 2.9\ )\\
\Gamma_6 &= (\ 5,\ 3.7,\ 3,\ 3,\ 2,\ 2\ )\\
\Gamma_7 &= (\ 2,\ 1.8,\ 1.6,\ 1.4,\ 1.2,\ 1\ )\\
\Gamma_8 &= (\ 3.3,\ 3,\ 2.7,\ 2.4,\ 2.2,\ 1.9\ )
\end{aligned}
$$

$$\zeta_0 = (-2.5,\ -3.5,\ -9,\ -13,\ -20)$$
$$\zeta_1 = (\ 0.4,\ 0.5,\ 0.8,\ 1.0,\ 1.2)$$
$$\zeta_2 = (\ 0,\ 0.2,\ -2,\ 0,\ 0,\ 0,\ 0.1,\ -4)$$

Table 2: Parameter Values for Simulation 2

m	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$
1	0.080	0.156	0.167	0.239	0.498
2	0.100	0.188	0.199	0.241	0.356
3	0.014	0.029	0.034	0.112	0.670
4	0.090	0.180	0.190	0.225	0.303
5	0.093	0.199	0.224	0.323	0.496
6	0.100	0.251	0.282	0.333	0.348
7	0.094	0.183	0.197	0.251	0.374
8	0.002	0.007	0.014	0.172	0.726

Table 3: Probability Y_{it} is missing by group m for Simulation 2

Priors		Estimated Risk		
GARP	IV	Loss Fcn 1	Loss Fcn 2	Mean Loss
Lag-block	Corr-logNormal	0.425	0.742	0.175
Lag-block	InvGamma	0.437	0.759	0.174
Sparsity	Corr-logNormal	0.553	0.915	0.196
Sparsity	InvGamma	0.565	0.934	0.196
Non-sparse	InvGamma	0.551	0.912	0.200
NB1	NB	0.605	0.987	0.203
NB2	NB	0.630	1.010	0.210
Group-specific flat*		0.892	1.255	0.248
Common-Σ flat		8.105	84.339	0.925

Table 4: Risk Estimates for Simulation 2. (The group-specific flat prior is only over 49 datasets because the MCMC chain failed to converge for one of the datasets.)

Covariance Prior		Model Fit			Treatment Effect	
GARP Prior	IV Prior	Dev	p_D	DIC	Group 1	Group 2
Lag-block	Corr-logN ($\rho = 0.90$)	39,006	342	39,690	9.23 (7.03, 11.48)	9.51 (6.85, 12.19)
Lag-block	InvGamma	38,999	350	39,698	9.22 (6.98, 11.45)	9.39 (6.73, 12.13)
Lag-block	Corr-logN ($\rho = 0.75$)	39,006	347	39,700	9.22 (6.99, 11.42)	9.56 (6.85, 12.27)
Lag-block	Corr-logN ($\rho = 0.50$)	39,003	349	39,700	9.24 (7.00, 11.53)	9.41 (6.74, 12.14)
Sparsity	Corr-logN ($\rho = 0.90$)	38,887	464	39,816	9.25 (7.02, 11.49)	8.78 (6.33, 11.24)
Sparsity	Corr-logN ($\rho = 0.75$)	38,887	466	39,819	9.23 (6.96, 11.42)	8.82 (6.39, 11.27)
Sparsity	Corr-logN ($\rho = 0.50$)	38,883	472	39,827	9.25 (7.01, 11.51)	8.68 (6.23, 11.20)
Sparsity	InvGamma	38,884	475	39,834	9.25 (7.02, 11.53)	8.64 (6.16, 11.15)
NB1	NB	38,875	481	39,837	9.25 (7.11, 11.46)	8.56 (6.16, 10.99)
Non-sparse	InvGamma	38,818	529	39,876	9.29 (7.01, 11.59)	8.81 (6.21, 11.52)
NB2	NB	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)

Table 5: Model fit statistics and treatment effects for the first two groups for the depression data using each of the priors.

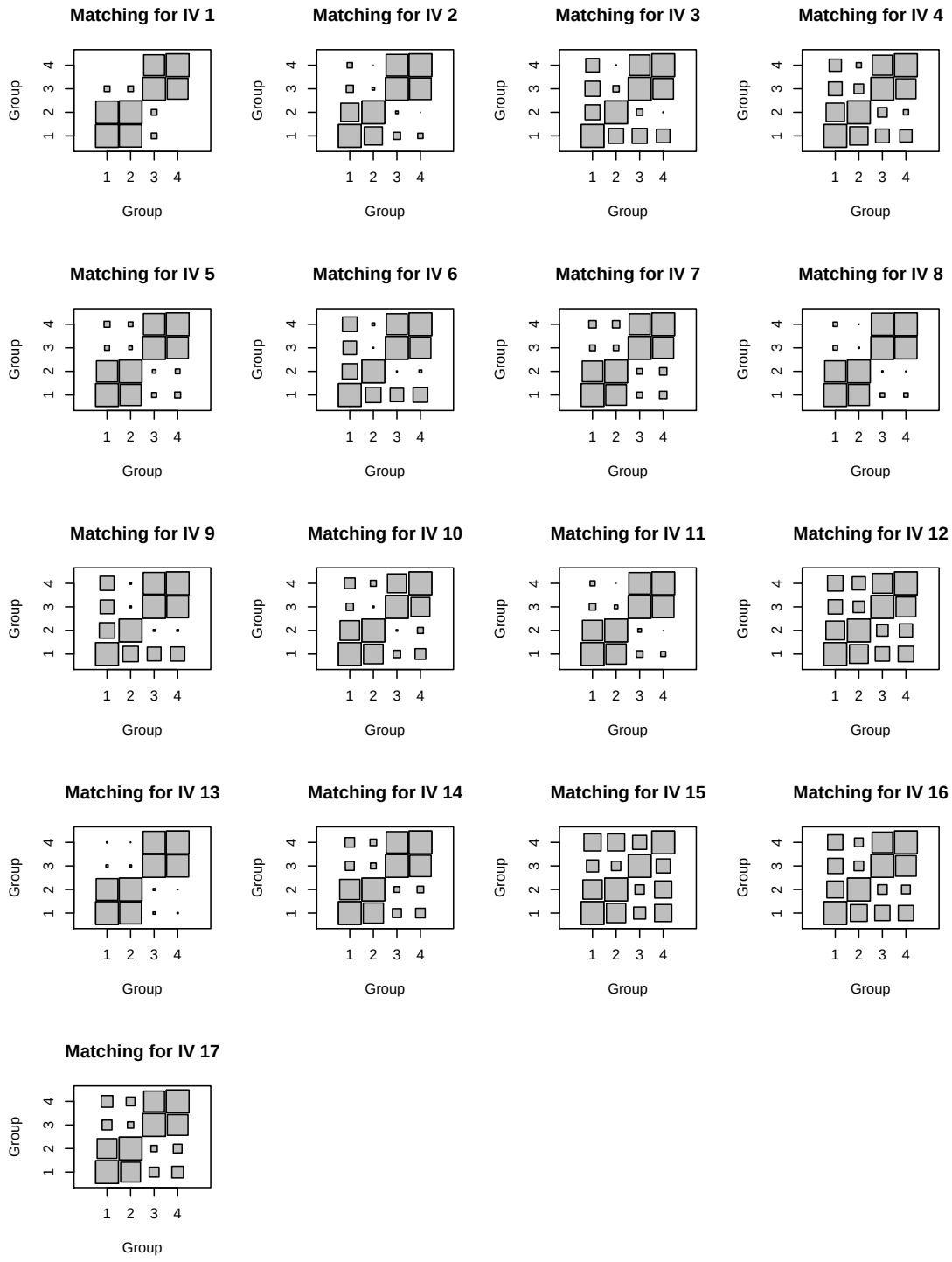

Figure 1: The posterior probabilities of matching for the innovation variances. The size of the boxes are proportional to $Pr(\gamma_{mj} = \gamma_{m'j} \mid y_{obs})$, with the boxes on the diagonal having area one for comparison.

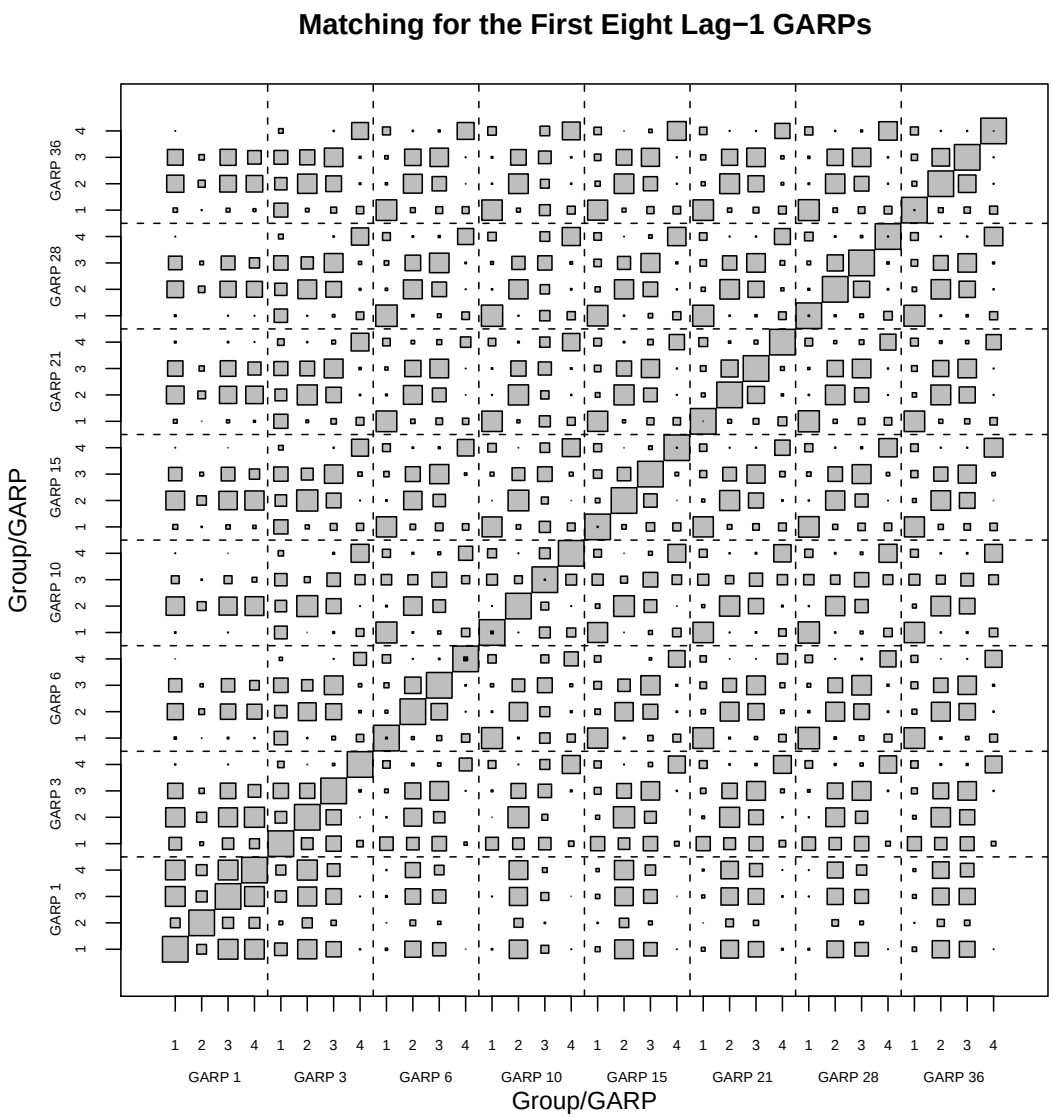

Figure 2: The posterior probabilities of matching for the first eight lag-1 GARPs. The size of the grey boxes are proportional to $Pr(\phi_{mj} = \phi_{m'j'} \mid y_{obs})$, with the boxes on the diagonal having area one for comparison. The black boxes overlaying the diagonal are proportional to the posterior of $Pr(\phi_{mj} = 0)$.

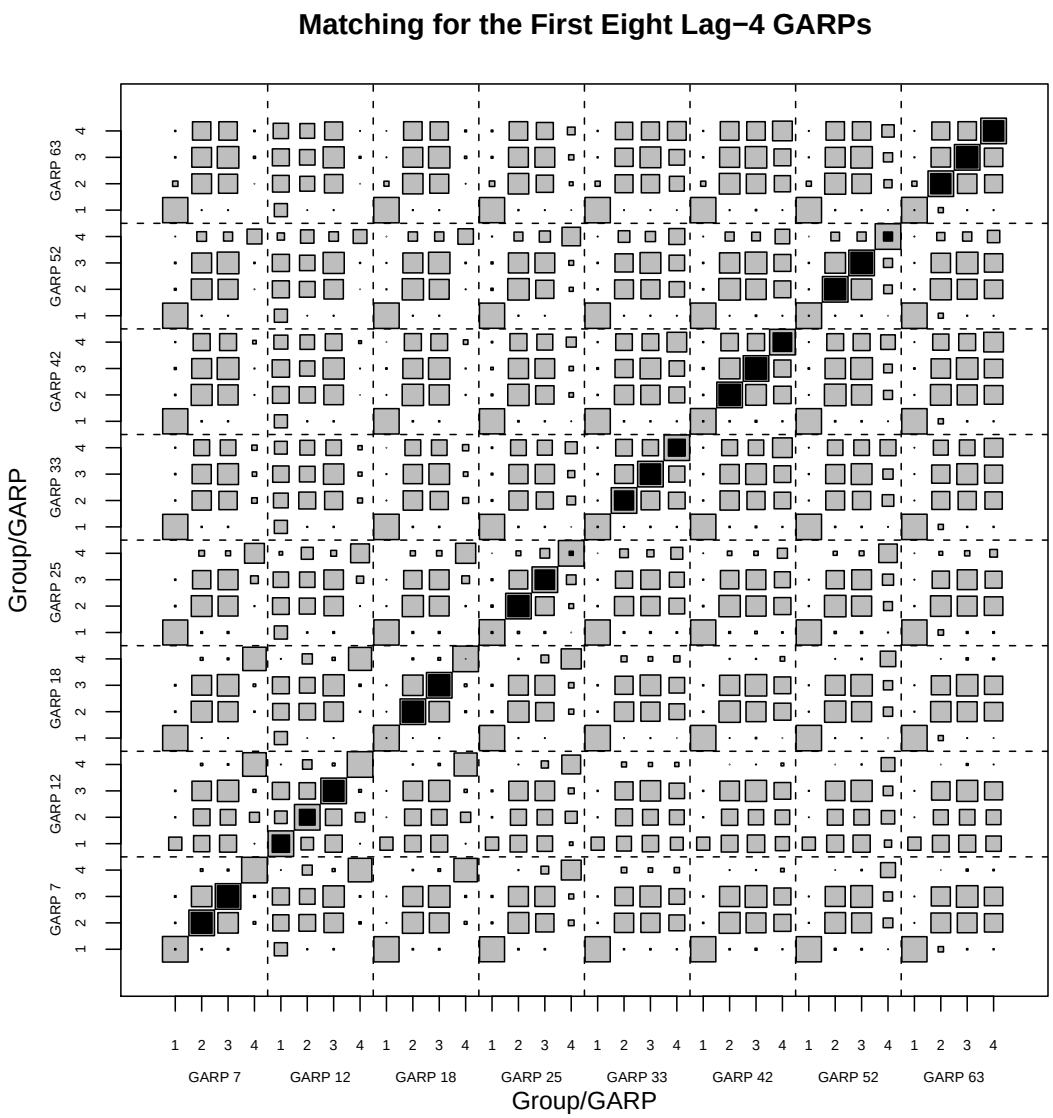

Figure 3: The posterior probabilities of matching for the first eight lag-4 GARPs. The size of the grey boxes are proportional to $Pr(\phi_{mj} = \phi_{m'j'} \mid y_{obs})$, with the boxes on the diagonal having area one for comparison. The black boxes overlaying the diagonal are proportional to the posterior of $Pr(\phi_{mj} = 0)$.