Sparsity Inducing Prior Distributions for Correlation Matrices through the Partial Autocorrelations

J. T. Gaskins[*], M. J. Daniels[†], and B. H. Marcus [‡]

Abstract

Modeling a correlation matrix $\mathbf{R}$ can be a difficult statistical task due to both the positive definite and the unit diagonal constraints. Because the number of parameters increases quadratically in the dimension, it is often useful to consider a sparse parameterization. We introduce a pair of prior distributions on the set of correlation matrices for longitudinal data through the partial autocorrelations (PACs), each of which vary independently over [-1,1]. The first prior shrinks each of the PACs toward zero with increasingly aggressive shrinkage in lag. The second prior (a selection prior) is a mixture of a zero point mass and a continuous component for each PAC, allowing for a sparse representation. The structure implied under our priors is readily interpretable because each zero PAC implies a conditional independence relationship in the distribution of the data. Selection priors on the PACs provide a computationally attractive alternative to selection on the elements of $\mathbf{R}$ or $\mathbf{R}^{-1}$ for ordered data. These priors allow for data-dependent shrinkage/selection under an intuitive parameterization in an unconstrained setting. The proposed priors are compared to standard methods through a simulation study and a multivariate probit data example. Supplemental materials for this article (appendix, data, and `R` code) are available online.

Key Words: Bayesian methods, Correlation matrix, Longitudinal data, Multivariate probit, Partial autocorrelation, Selection priors, Shrinkage

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

[†] *mjdaniels@austin.utexas.edu*; Division of Statistics & Scientific Computation, University of Texas, Austin, TX 78712

[‡] *bmarcus@ucsd.edu*; Department of Family and Preventive Medicine, UC San Diego, CA 92093

1. Introduction

Determining the structure of an unknown $J \times J$ covariance matrix $\mathbf{\Sigma}$ is a long standing statistical challenge. A key difficulty in dealing with the covariance matrix is the positive definiteness constraint. This is because the set of values for a particular element σ_{ij} that yield a positive definite $\mathbf{\Sigma}$ depends on the choice of the remaining elements of $\mathbf{\Sigma}$. Additionally, because the number of parameters in $\mathbf{\Sigma}$ is quadratic in the dimension J, methods to find a parsimonious (lower-dimensional) structure can be beneficial.

One of the earliest attempts in this direction is the idea of covariance selection (Dempster, 1972). By setting some of the off-diagonal elements of the concentration matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ to zero, a more parsimonious choice for the covariance matrix of the random vector $\mathbf{Y}$ is achieved. A zero in the (i, j)-th position of $\mathbf{\Omega}$ implies zero correlation (and further, independence under multivariate normality) between Y_i and Y_j, conditional on the remaining components of $\mathbf{Y}$. This property, along with its relation to graphical model theory (e.g., Lauritzen, 1996), has led to the use of covariance selection as a standard part of analysis in multivariate problems (Wong et al., 2003; Yuan and Lin, 2007; Rothman et al., 2008). However, one should be cautious when using such selection methods as not all produce positive definite estimators. For instance, thresholding the sample covariance (concentration) matrix will not generally be positive definite, and adjustments are needed (Bickel and Levina, 2008).

Model specification for $\mathbf{\Sigma}$ may depend on a correlation structure through the so-called separation strategy (Barnard et al., 2000). The separation strategy involves reparameterizing $\mathbf{\Sigma}$ by $\mathbf{\Sigma} = \mathbf{SRS}$, with $\mathbf{S}$ a diagonal matrix containing the marginal standard deviations of $\mathbf{Y}$ and $\mathbf{R}$ the correlation matrix. Let $\mathcal{R}_J$ denote the set of valid correlation matrices, that is, the collection of $J \times J$ positive definite matrices with unit diagonal. Separation can also be performed on the concentration matrix, $\mathbf{\Omega} = \mathbf{TCT}$ so that $\mathbf{T}$ is diagonal and $\mathbf{C} \in \mathcal{R}_J$. The diagonal elements of $\mathbf{T}$ give the partial standard deviations, while the elements c_{ij} of $\mathbf{C}$ are the (full) partial correlations. The covariance selection problem is equivalent to choosing elements of the partial correlation matrix $\mathbf{C}$ to be null. Several authors have constructed priors to estimate $\mathbf{\Sigma}$ by allowing $\mathbf{C}$ to be a sparse matrix (Wong et al., 2003; Carter et al., 2011).

In many cases the full partial correlation matrix may not be convenient to use. In cases where the covariance matrix is fixed to be a correlation matrix such as the multivariate probit case, the elements of the concentration matrix $\mathbf{T}$ and $\mathbf{C}$ are constrained to maintain a unit diagonal for $\mathbf{\Sigma}$

(Pitt et al., 2006). Additionally, interpretation of parameters in the partial correlation matrix can be challenging, particularly for longitudinal settings as the partial correlations are defined conditional on future values. For example, c_{12} gives the correlation between Y_1 and Y_2 conditional on the future measurements $Y_3, \ldots, Y_J$. An additional issue with Bayesian methods that promote sparsity in $\mathbf{C}$ is calculating the volume of the space of correlation matrices with a fixed zero pattern; see Section 4.2 for details.

In addition to the role $\mathbf{R}$ plays in the separation strategy, in some data models the covariance matrix is constrained to be a correlation matrix for identifiability. This is the case for the multivariate probit model (Chib and Greenberg, 1998), Gaussian copula regression (Pitt et al., 2006), certain latent variables models (e.g. Daniels and Normand, 2006), among others. Thus, it is necessary to make use of methods specific for estimating and/or modeling a correlation matrix.

We consider this problem of correlation matrix estimation in a Bayesian context where we are concerned with choices of an appropriate prior distribution $p(\mathbf{R})$ on $\mathcal{R}_J$. Commonly used priors include a uniform prior over $\mathcal{R}_J$ (Barnard et al., 2000) and Jeffrey's prior $p(\mathbf{R}) \propto |\mathbf{R}|^{-(J+1)/2}$. In these cases the sampling steps for $\mathbf{R}$ can sometimes benefit from parameter expansion techniques (Liu, 2001; Zhang et al., 2006; Liu and Daniels, 2006). Liechty et al. (2004) develop a correlation matrix prior by specifying each element ρ_{ij} of $\mathbf{R}$ as an independent normal subject to $\mathbf{R} \in \mathcal{R}_J$. Pitt et al. (2006) extend the covariance selection prior (Wong et al., 2003) to the correlation matrix case by fixing the elements of $\mathbf{T}$ to be constrained by $\mathbf{C}$ so that $\mathbf{T}$ is the diagonal matrix such that $\mathbf{R} = (\mathbf{TCT})^{-1}$ has unit diagonal.

The difficulty of jointly dealing with the positive definite and unit diagonal constraints of a correlation matrix has led some researchers to consider priors for $\mathbf{R}$ based on the partial autocorrelations (PACs) in settings where the data are ordered. PACs suggest a practical alternative by avoiding the complication of the positive definite constraint, while providing easily interpretable parameters (Joe, 2006). Kurowicka and Cooke (2003, 2006) frame the PAC idea in terms of a vine graphical model. Daniels and Pourahmadi (2009) construct a flexible prior on $\mathbf{R}$ through independent shifted Beta priors on the PACs. Wang and Daniels (2013a) construct underlying regressions for the PACs, as well as a triangular prior which shifts the prior weight to a more intuitive choice in the case of longitudinal data. Instead of setting partial correlations from $\mathbf{C}$ to zero to incorporate sparsity, our goal is to encourage parsimony through the PACs. As the PACs are unconstrained, selection does not lead to the computational issues associated with finding the normalizing constant

for a sparse $\mathbf{C}$. We introduce and compare priors for both selection and shrinkage of the PACs that extends previous work on sensible default choices (Daniels and Pourahmadi, 2009).

The layout of this article is as follows. In the next section we will review the relevant details of the partial autocorrelation parameterization. Section 3 proposes a prior for $\mathbf{R}$ induced by shrinkage priors on the PACs. Section 4 introduces the selection prior for the PACs. Simulation results showing the performance of the priors appear in Section 5. In Section 6 the proposed PAC priors are applied to a data set from a smoking cessation clinical trial. Section 7 concludes the article with a brief discussion.

2. Partial autocorrelations

For a general random vector $\mathbf{Y} = (Y_1, \ldots, Y_J)'$ the partial autocorrelation between Y_i and Y_j $(i < j)$ is the correlation between the two given the intervening variables $(Y_{i+1}, \ldots, Y_{j-1})$. We denote this PAC by π_{ij}, and let $\mathbf{\Pi}$ be the upper-triangular matrix with elements π_{ij}. Because the PACs are formed by conditioning on the intermediate components, there is a clear dependence on the ordering of the components of $\mathbf{Y}$. In many applications such as longitudinal data modeling, there is a natural time ordering to the components. With an established ordering of the elements of $\mathbf{Y}$, we refer to the lag between Y_i and Y_j as the time-distance $j - i$ between the two.

We now describe the relationship between $\mathbf{R}$ and $\mathbf{\Pi}$. For the lag-1 components $(j - i = 1)$ $\pi_{ij} = \rho_{ij}$ since there are no components between Y_i and Y_j. The higher lag components are calculated from the formula (Anderson, 1984, Section 2.5),

$$\pi_{ij} = r_1^{-1/2} r_2^{-1/2} \left[\rho_{ij} - \mathbf{r}_1'(i,j) \mathbf{R}_3(i,j)^{-1} \mathbf{r}_2(i,j) \right], \tag{1}$$

where $\mathbf{r}_1'(i,j) = (\rho_{i,i+1}, \ldots, \rho_{i,j-1})$, $\mathbf{r}_2'(i,j) = (\rho_{j,i+1}, \ldots, \rho_{j,j-1})$, and $\mathbf{R}_3(i,j)$ is the sub-correlation matrix of $\mathbf{R}$ corresponding to the variables $(Y_{i+1}, \ldots, Y_{j-1})$. The scalars r_l $(l = 1, 2)$ are $r_l = 1 - \mathbf{r}_l'(i,j) \mathbf{R}_3(i,j)^{-1} \mathbf{r}_l(i,j)$. Equivalent to (1), we may define the partial autocorrelation in terms of the distribution of the (mean zero) variable $\mathbf{Y}$. Let $\widetilde{\mathbf{Y}} = (Y_{i+1}, \ldots, Y_{j-1})'$ be the vector (possibly empty or scalar) of the intermediate responses, and $\mathbf{b}_i' \widetilde{\mathbf{Y}}$ and $\mathbf{b}_j' \widetilde{\mathbf{Y}}$ be the linear least squares predictors of Y_i and Y_j given $\widetilde{\mathbf{Y}}$, respectively. Then, $\pi_{ij} = \text{corr}\{Y_i - \mathbf{b}_i' \widetilde{\mathbf{Y}},\ Y_j - \mathbf{b}_j' \widetilde{\mathbf{Y}}\}$, and it is reasonable to consider π_{ij} to define the correlation between Y_i and Y_j after correcting for $\widetilde{\mathbf{Y}}$.

Examination of formula (1) shows that the operation from $\mathbf{R}$ to $\mathbf{\Pi}$ is invertible. By inverting the previous operations recursively over increasing lag $j - i$, one obtains the correlation matrix

from the PACs by $\rho_{i,i+1} = \pi_{i,i+1}$ and

$$\rho_{ij} = \mathbf{r}_1'(i,j)\mathbf{R}_3(i,j)^{-1}\mathbf{r}_2(i,j) + r_1^{1/2}r_2^{1/2}\pi_{ij}$$

for $j - i > 1$. As the relationship between $\mathbf{R}$ and $\mathbf{\Pi}$ is one-to-one, the Jacobian for the transformation from $\mathbf{R}$ to $\mathbf{\Pi}$ can be computed easily. The determinant of the Jacobian is given by

$$|J(\mathbf{\Pi})| = \prod_{i<j}(1-\pi_{ij}^2)^{-[J-1-(j-i)]/2} \tag{2}$$

(Joe, 2006, Theorem 4). Notationally, we let $\mathbf{R}(\mathbf{\Pi})$ denote correlation matrix corresponding to the PACs $\mathbf{\Pi}$. Similarly, $\mathbf{\Pi}(\mathbf{R})$ represents the set of PACs corresponding to correlation matrix $\mathbf{R}$. When it is clear from context, we continue to use only the matrix $\mathbf{R}$ or $\mathbf{\Pi}$ and not the functional notation.

The key advantage in using PACs is that parameters are unconstrained (Joe, 2006). For the correlation matrix $\mathbf{R}$, the subset of values in $(-1,1)$ that ρ_{ij} can take satisfying the positive definite constraint is determined by the configuration of the other elements of $\mathbf{R}$. For a geometric interpretation of this phenomenon, see Rousseeuw and Molenberghs (1994). For the PACs, each π_{ij} can take any value in $(-1,1)$, regardless of the choice of the remaining π's. This is especially important in the selection context, as setting certain elements of $\mathbf{R}$ (or the partial correlation matrix $\mathbf{C}$) to zero can greatly restrict the sets of values that yield a positive definite matrix for other elements in $\mathbf{R}$ ($\mathbf{C}$).

Define SBeta(α,β) to be the beta distribution shifted to the support $(-1,1)$, i.e., the density proportional to $(1+y)^{\alpha-1}(1-y)^{\beta-1}$ for $y \in (-1,1)$. Daniels and Pourahmadi (2009) use the PACs to form a prior on $\mathbf{R}$ by letting each π_{ij} come from this shifted beta distribution where the two shape parameters depend on the lag $j-i$, with the special case where each $\pi_{ij} \sim$ SBeta$(1,1)$. We call this the flat-PAC (or flat-$\mathbf{\Pi}$) prior since it specifies a uniform distribution for each of the PACs. Wang and Daniels (2013a) advise using a triangular prior with SBeta(2,1) which (weakly) encourages positive values for the PACs.

The result in (2) shows that we can write the flat prior of Barnard et al. (2000) in terms of a prior on the PACs. We call the prior $p_{fR}(\mathbf{R}) \propto I(\mathbf{R} \in \mathcal{R}_J)$ the flat-$\mathbf{R}$ prior since it is uniform over the space $\mathcal{R}_J$. Hence, the flat-$\mathbf{R}$ is equal to $p_{fR}(\mathbf{\Pi}) \propto |J(\mathbf{\Pi})|^{-1}$, which has a contribution from π_{ij} of $(1-\pi_{ij}^2)^{[J-1-(j-i)]/2}$. Note that $p_{fR}(\mathbf{\Pi})$ is the product of independent SBeta(α_{ij},β_{ij}) distributions for each π_{ij}, where $\alpha_{ij} = \beta_{ij} = 1+[J-1-(j-i)]/2$. This provides an unconstrained representation of the flat-$\mathbf{R}$ prior.

In longitudinal/ordered data contexts, we expect the PACs to be negligible for elements that have large lags. We exploit this concept via two types of priors. First, we introduce priors that shrink PACs toward zero with the aggressiveness of the shrinkage depending on the lag. Next, we propose, in the spirit of Wong et al. (2003), a selection prior that will stochastically choose PACs to be set to zero.

3. Partial autocorrelation shrinkage priors

3.1. Specification of the shrinkage prior

Using the PAC framework, we form priors that will shrink the PAC π_{ij} toward zero. It has long been known that shrinkage estimators can produce greatly improved estimation (James and Stein, 1961). As previously noted, $\pi_{ij} = 0$ implies that Y_i and Y_j are uncorrelated given the intervening variables $(Y_{i+1}, \ldots, Y_{j-1})$. In the case where $\mathbf{Y}$ has a multivariate normal distribution, this implies independence between Y_i and Y_j, given $(Y_{i+1}, \ldots, Y_{j-1})$. We anticipate that variables farther apart in time (and conditional on more intermediate variables) are more likely to be uncorrelated, so we will more aggressively shrink π_{ij} for larger values of the lag $j - i$.

We let each $\pi_{ij} \sim \text{SBeta}(\alpha_{ij}, \beta_{ij})$ independently. As we wish to shrink toward zero, we want $\text{E}\{\pi_{ij}\} = 0$, so we fix $\alpha_{ij} = \beta_{ij}$. It is easily shown that

$$\text{Var}\{\pi_{ij}\} = \frac{4\alpha_{ij}\beta_{ij}}{(\alpha_{ij} + \beta_{ij})^2(\alpha_{ij} + \beta_{ij} + 1)},$$

which we denote by ξ_{ij}. We recover the SBeta shape parameters by $\alpha_{ij} = \beta_{ij} = (\xi_{ij}^{-1} - 1)/2$. Hence, the distribution of π_{ij} is determined by its variance ξ_{ij}. Rather than specifying these $J(J-1)/2$ different variances, we parameterize them through

$$\text{Var}\{\pi_{ij}\} = \xi_{ij} = \epsilon_0 |j - i|^{-\gamma}, \tag{3}$$

where $\epsilon_0 \in (0, 1)$ and $\gamma > 0$. Clearly, ξ_{ij} is decreasing in lag so that higher lag terms will generally be closer to zero. We let the positive γ parameter determine the rate that ξ_{ij} decreases in lag.

To fully specify the Bayesian set-up, we must introduce prior distributions on the two parameters, ϵ_0 and γ. To specify these hyperpriors, we use a uniform (or possibly a more general beta) for ϵ_0 and a gamma distribution for γ. We require $\gamma > 0$, so $\xi_{ij} = \epsilon_0 |j - i|^{-\gamma}$ remains an decreasing function of lag. In the simulations and data analysis of Sections 5 and 6, we use $\gamma \sim$ Gamma(5,5),

so that γ has a prior mean of 1 and prior variance of $1/5$. We use a moderately informative prior to keep γ from dominating the role of ϵ_0 in $\xi_{ij} = \epsilon_0 |j - i|^{-\gamma}$. A large value of γ will force all ξ_{ij} of lag greater than one to be approximately zero, regardless of the value of ϵ_0.

3.2. Sampling under the shrinkage prior

The utility of our prior depends on our ability to incorporate it into a Markov chain Monte Carlo (MCMC) scheme. For simplicity we assume that the data consists of $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$, where each $\mathbf{Y}_i$ is a J-dimensional normal vector with mean zero and covariance $\mathbf{R}$, which is a correlation matrix so as to mimic the computations for the multivariate probit case. Let $\mathcal{L}(\mathbf{\Pi}|\mathbf{Y})$ denote the likelihood function for the data, parameterized by the PACs, $\mathbf{\Pi}$.

The MCMC chain we propose involves sequentially updating each of the $J(J-1)/2$ PACs, followed by updating the hyperparameters determining the variance of the SBeta distributions. To sample a particular π_{ij}, we must draw the new value from the distribution proportional to $\mathcal{L}(\pi_{ij}, \mathbf{\Pi}_{(-ij)}|\mathbf{Y})\, p_{ij}(\pi_{ij})$, where $p_{ij}(\pi_{ij})$ is the SBeta(α_{ij}, β_{ij}) density and $\mathbf{\Pi}_{(-ij)}$ represents the set of PACs except π_{ij}. Due to the subtle role of π_{ij} in the likelihood piece, there is no simple conjugate sampling step. In order to sample from $\mathcal{L}(\pi_{ij}, \mathbf{\Pi}_{(-ij)}|\mathbf{Y})\, p_{ij}(\pi_{ij})$, we introduce an auxiliary variable U_{ij} (Damien et al., 1999; Neal, 2003), and note that we can rewrite the conditional distribution as

$$\mathcal{L}(\pi_{ij}, \mathbf{\Pi}_{(-ij)}|\mathbf{Y})\, p_{ij}(\pi_{ij}) = \int_0^\infty I\left\{u_{ij} < \mathcal{L}(\pi_{ij}, \mathbf{\Pi}_{(-ij)}|\mathbf{Y}) p_{ij}(\pi_{ij})\right\}\, \mathrm{d}u_{ij}, \tag{4}$$

suggesting a method to sample π_{ij} in two steps. First, sample U_{ij} uniformly over the interval $[0, \mathcal{L}(\pi_{ij}, \mathbf{\Pi}_{(-ij)}|\mathbf{Y}) p_{ij}(\pi_{ij})]$, using the current value of π_{ij}. We then draw the new π_{ij} from uniformly from the slice set $\mathcal{P} = \left\{\pi : u_{ij} < \mathcal{L}(\pi, \mathbf{\Pi}_{(-ij)}|\mathbf{Y}) p_{ij}(\pi)\right\}$. Because this set lies within the compact set $[-1, 1]$, $\mathcal{P}$ could be calculated numerically to within a prespecified level of accuracy, but this is not generally necessary due to the "stepping out" algorithm of Neal (2003).

The variance parameters, ϵ_0 and γ, are not conjugate so sampling new values in the MCMC chain requires a non-standard step. We also update them using the auxiliary variable technique.

4. Partial autocorrelation selection priors

4.1. Specification of the selection prior

Having developed a prior that shrinks the partial autocorrelations toward zero, we now consider prior distributions that give positive probability to the event that the PAC π_{ij} is equal to zero. Again, this zero implies that Y_i and Y_j are uncorrelated given the intervening variables $(Y_{i+1}, \ldots, Y_{j-1})$ with independence under multivariate normality. The selection priors are formed by independently specifying the prior for each π_{ij} as the mixture distribution,

$$\pi_{ij} \sim \epsilon_{ij} \, \text{SBeta}(\alpha_{ij}, \beta_{ij}) + (1 - \epsilon_{ij}) \, \delta_0, \tag{5}$$

where δ_0 represents a degenerate distribution with point mass at zero. In the shrinkage prior we allowed the shifted Beta parameters α_{ij}, β_{ij} to depend on lag, but here we generally let $\alpha = \alpha_{ij}$ and $\beta = \beta_{ij}$ and incorporate structure through the modeling choices on ϵ_{ij}. While there is flexibility to make any choice of these shifted Beta parameters α, β, we recommend as default choices either a uniform distribution on $[-1, 1]$ through $\alpha = \beta = 1$ (Daniels and Pourahmadi, 2009) or the triangular prior of Wang and Daniels (2013a) by $\alpha = 2, \beta = 1$; alternatively, independent hyperpriors for α, β could be specified.

The value of ϵ_{ij} gives the probability that π_{ij} will be non-zero, i.e. will be drawn from the continuous component in the mixture distribution. Hence, we have the probability that Y_i and Y_j are uncorrelated, given the interceding variables, is $1 - \epsilon_{ij}$. As the values of the ϵ's decrease, the selection prior places more weight on the point-mass δ_0 component of the distribution (5), yielding more sparse choices for $\mathbf{\Pi}$. As with our parameterizations of the variance ξ_{ij} in Section 3.1, we make a structural choice of the form of ϵ_{ij} so that this probability depends on the lag-value. We let

$$\epsilon_{ij} = \epsilon_0 |j - i|^{-\gamma}, \tag{6}$$

similar to our choice of ξ_{ij} in the shrinkage prior.

This choice (6) specifies the continuous component probability to be an polynomial function of the lag. Because ϵ_{ij} is decreasing as the lag $j - i$ increases, $\text{P}(\pi_{ij} = 0)$ increases. Conceptually, this means that we anticipate that variables farther apart in time (and conditional on more intermediate variables) are more likely to be uncorrelated. As with the shrinkage prior, we choose hyperpriors of $\epsilon_0 \sim \text{Unif}(0, 1)$ and $\gamma \sim$ Gamma(5,5).

4.2. Normalizing constant for priors on R

One of the key improvements of our selection prior over other sparse priors for **R** is the simplicity of the normalizing constant, as mentioned in the introduction. Previous covariance priors with a sparse **C** (Wong et al., 2003; Pitt et al., 2006; Carter et al., 2011) place a flat prior on the non-zero components c_{ij} for a given pattern of zeros. However, the needed normalizing constant requires finding the volume of the subspace of $\mathcal{R}_J$ corresponding to the pattern of zeros in **C**. This turns out to be a quite difficult task and provides much of the challenge in the work of the three previously cited papers.

We are able to avoid this issue by specifying our selection prior in terms of the unrestricted PAC parameterization. As the value of any of the π_{ij}'s does not effect the support of the remaining PACs, the volume of $[-1,1]^{J(J-1)/2}$ corresponding to any configuration of $\mathbf{\Pi}$ with J_0 ($\leq J(J-1)/2$) non-zero elements is 2^{J_0}, the volume of a J_0-dimensional hypercube. Because this constant does not depend on which elements are non-zero, we need not explicitly deal with it in the MCMC algorithm to be introduced in the next subsection. Further, we are able the exploit structure in the order of the PACs in selection (i.e. higher lag terms are more likely to be null), whereas in Pitt et al. (2006), the probability that c_{ij} is zero is chosen to minimize the effort required to find the normalizing constant.

An additional benefit of performing selection on the partial autocorrelation as opposed to the partial correlations **C** is that the zero patterns hold under marginalizations of the beginning and/or ending time points. For instance, if we marginalize out the Jth time point, the corresponding matrix of PACs is the original $\mathbf{\Pi}$ after removing the last row and column. However, any zero elements in **C** will not be preserved because $\text{corr}(Y_1, Y_2|Y_3, \ldots, Y_J) = 0$ does not generally imply that $\text{corr}(Y_1, Y_2|Y_3, \ldots, Y_{J-1}) = 0$.

4.3. Sampling under the selection prior

Sampling with the selection prior proceeds similarly to the shrinkage prior scheme with the main difference being the introduction of the point mass in (5). As before we sequentially update each of the PACs, by drawing the new value from the distribution proportional to $\mathcal{L}(\pi_{ij}, \mathbf{\Pi}_{(-ij)}|\mathbf{Y})\, p_{ij}(\pi_{ij})$, where $p_{ij}(\pi_{ij})$ gives the density corresponding the prior distribution in (5) (with respect to the appropriate mixture dominating measure). We cannot use the slice sampling step according to (4)

but must write the distribution as

$$\mathcal{L}(\pi_{ij}, \mathbf{\Pi}_{(-ij)}|\mathbf{Y})\, p_{ij}(\pi_{ij}) = \int_0^\infty I\left\{u_{ij} < \mathcal{L}(\pi_{ij}, \mathbf{\Pi}_{(-ij)}|\mathbf{Y})\right\} p_{ij}(\pi_{ij})\, \mathrm{d}u_{ij}. \tag{7}$$

For the selection prior, we sample U_{ij} uniformly over the interval from zero to $\mathcal{L}(\pi_{ij}, \mathbf{\Pi}_{(-ij)}|\mathbf{Y})$, using the current value of π_{ij}, and then draw π_{ij} from $p_{ij}(\cdot)$, restricted to the slice set $\mathcal{P} = \left\{\pi : u_{ij} < \mathcal{L}(\pi, \mathbf{\Pi}_{(-ij)}|\mathbf{Y})\right\}$.

To sample from $p_{ij}(\cdot)$ restricted to $\mathcal{P}$, let $F(x) = \mathrm{P}(\pi_{ij} \leq x)$ denote the (cumulative) distribution function for the prior (5) of π_{ij}. Note that $F(x)$ is available in closed form when the SBeta distribution is uniform or triangular. We then draw a random variable Z uniformly over the set $F(\mathcal{P}) \subset [0, 1]$, and the updated value of π_{ij} is $F^{-1}(Z) = \inf\{\pi : F(\pi) \geq Z\}$. This is simply a version of the probability integral transform. It is relatively straight-forward to verify that sampling according to (7) instead of (4) using the "stepping out" algorithm of Neal (2003) leaves the stationary distribution invariant.

The similarity between the sampling steps for the shrinkage and selection priors is notable. Consider the situation when the parameter of concern is the vector of regression coefficients for a linear regression model. With a shrinkage prior these regression coefficients may be drawn simultaneously. But when using a selection prior, each coefficient must be sampled one at a time, and each step requires finding the posterior probability it should be set to zero. For linear models the computational effort required for selection is often much greater than under shrinkage.

In the PAC context, this is not the case. We cannot update the PACs in blocks under the shrinkage prior, so there is no computational benefit relative to selection. Because we sample from the probability integral transform restricted to $\mathcal{P}$, there is also no need to compute the posterior probability that the parameter is selected. Hence, the computational effort for the shrinkage and selection is roughly equivalent. Finally, with the exception of the minor step of updating the hyperparameters, the flat-$\mathbf{\Pi}$ and triangular priors also require a similar level of computational time as the selection and shrinkage priors.

To sample the parameters ϵ_0 and γ defining the mixing proportions ϵ_{ij}, we introduce the set of dummy variable $\zeta_{ij} = I(\pi_{ij} \neq 0)$, which have the property that $\mathrm{P}(\zeta_{ij} = 1) = \epsilon_{ij}$. The sampling distributions of ϵ_0 and γ depend on $\mathbf{\Pi}$ only through the set of indicator variables ζ_{ij}. As with the variance parameters of the shrinkage priors, we incorporate a pair of slice sampling steps to update the hyperparameters.

5. Simulations

To better understand the behavior of our proposed priors, we conducted a simulation study to assess the (frequentist) risk of their posterior estimators. We consider four choices A–D for the true covariance matrix in the case of six-dimensional ($J = 6$) data. $\mathbf{R}^A$ will have an autoregressive (AR) structure with $\rho_{ij}^A = 0.7^{|j-i|}$. The corresponding $\mathbf{\Pi}^A$ has values of 0.7 for the lag-1 terms and zero for the others, a sparse parameterization. For the second correlation matrix $\mathbf{R}^B$ we choose the identity matrix so that all of PACs are zero in this case. The $\mathbf{\Pi}^C$ has a structure that decays to zero. For the lag-1 terms $\pi_{i,i+1}^C = 0.7$, and for the remaining terms, $\pi_{ij}^C = 0.4^{j-i-1}$, $j - i > 1$. Neither $\mathbf{\Pi}^C$ nor $\mathbf{R}^C$ have zero elements, but π_{ij}^C decrease quickly in lag $j - i$. Finally, we consider a correlation matrix that comes from a sparse $\mathbf{\Pi}^D$,

$$\mathbf{\Pi}^D = \begin{bmatrix} 1 & .9 & .3 & 0 & 0 & 0 \\ 0.90 & 1 & .8 & .4 & .1 & 0 \\ 0.80 & 0.80 & 1 & .6 & .2 & 0 \\ 0.62 & 0.67 & 0.60 & 1 & .8 & .3 \\ 0.58 & 0.63 & 0.58 & 0.80 & 1 & .7 \\ 0.46 & 0.50 & 0.45 & 0.69 & 0.70 & 1 \end{bmatrix},$$

where the upper-triangular elements correspond to $\mathbf{\Pi}^D$ and the lower-triangular elements depict the marginal correlations from $\mathbf{R}^D$. Note that while $\mathbf{\Pi}^D$ is somewhat sparse, $\mathbf{R}^D$ has only non-zero elements.

For each of these four choices of the true dependence structure and for sample sizes of $N = 20$, 50, and 200, we simulate 50 datasets. For each dataset a posterior sample for $\mathbf{\Pi}$ (and hence, $\mathbf{R}$) is obtained by running an MCMC chain for 5000 iterations, after a burn-in of 1000. We use every tenth iteration for inference, giving a sample of 500 values for each dataset. We consider the performance of both the selection and shrinkage priors on $\mathbf{\Pi}$. For the selection prior, we perform analyses with SBeta$(1, 1)$ (i.e., Unif$(-1, 1)$) and SBeta$(2, 1)$ (triangular prior) for the continuous component of the mixture distributions (5). In both the selection and shrinkage priors, the hyperpriors are $\epsilon_0 \sim$ Unif$(0, 1)$ and $\gamma \sim$ Gamma(5,5). The estimators from the shrinkage and selection priors are compared with the estimators resulting from the flat-$\mathbf{R}$, flat-PAC, and triangular priors. Finally, we consider a naive shrinkage prior where γ is fixed at zero in (3). Here, all PACs are equally shrunk with variance $\xi_{ij} = \epsilon_0$ independent of lag.

We consider two loss functions in comparing the performance of the six prior choices: $L_1(\hat{\mathbf{R}}, \mathbf{R}) =$ tr$(\hat{\mathbf{R}}\,\mathbf{R}^{-1}) - \log|\hat{\mathbf{R}}\,\mathbf{R}^{-1}| - p$ and $L_2(\hat{\mathbf{\Pi}}, \mathbf{\Pi}) = \sum_{i<j}(\hat{\pi}_{ij} - \pi_{ij})^2$. The first loss function is the stan-

dard covariance log-likelihood loss (Yang and Berger, 1994), whose Bayes estimator is $\mathrm{E}\{\mathbf{R}^{-1}\}^{-1}$. Because this quantity generally does not have a unit diagonal, we use $\hat{\mathbf{R}}_1 = \mathbf{S}\,\mathrm{E}\{\mathbf{R}^{-1}\}^{-1}\,\mathbf{S}$, where $\mathbf{S} = [\mathrm{diag}(\mathrm{E}\{\mathbf{R}^{-1}\})]^{1/2}$ is the diagonal matrix that guarantees $\hat{\mathbf{R}}_1$ is a correlation matrix. The Bayes estimator for L_2 is $\hat{\mathbf{R}}_2 = \mathbf{R}\,(\mathrm{E}\{\mathbf{\Pi}\})$, the correlation matrix corresponding to the posterior mean of $\mathbf{\Pi}$.

We estimate the frequentist risk for $\mathbf{R}^k$, $k \in \{A, B, C, D\}$, by averaging the loss over the 50 datasets. Table 1 contains the estimated risk by loss function, prior choice, sample size, and true correlation matrix. When evaluating the risk for loss function l, we are using the estimator $\hat{\mathbf{R}}_l$ for $l = 1, 2$. Figure 1 contains the box plots of the observed losses for L_1 with $\hat{\mathbf{R}}_1$. Plots using loss function 2 look similar and have been excluded for brevity.

It is immediately clear that the shrinkage and selection priors dominate the two flat priors for correlation matrices A and B. These are the matrices that have the most sparsity. From the box plots we see the losses for the middle 50% of datasets for the selection priors fall completely below the middle 50% for the four competitors. For $\mathbf{R}^A$ we see risk reductions between 28 and 61% for the sparse estimators over the estimators from the flat priors with $N = 20$; for $N = 200$ the improvements range from 23 to 64%. In the independence case, the estimators from the shrinkage and selection priors outperform the flat estimators by margins between 83 and 99%. While our focus is mainly on the comparison of the sparse priors to the others, we note that generally the triangular and flat-$\mathbf{\Pi}$ choices are best among the four competitors, with the naive shrinkage prior performing quite well for $\mathbf{R}^B$.

For $\mathbf{\Pi}^C$ all of the seven prior choices perform comparably. From Figure 1 we see that the middle 50% of the losses fall in the same range for each of the sample sizes. For all sample sizes the shrinkage prior is (slightly) favored, and for $N = 20$ the estimated risk for flat-$\mathbf{R}$ is visibly worse than the others. Recall that π_{ij}^C is decreasing in lag but is not equal to zero. In fact, the smallest element $\pi_{16}^C = (0.4)^4 = 0.0256$ which may not be close enough to zero to be effectively zeroed out, explaining why the selection priors are less effective for $\mathbf{\Pi}^C$ than in the other scenarios.

When we consider estimating the sparse correlation matrix $\mathbf{\Pi}^D$, the shrinkage and selection priors outperform the four other priors. From Table 1 we see that for loss function 1 and the $N = 20$ sample size the estimated risk decreases by 45 (25), 45 (24) and 39 (16) percent for the estimates from the shrinkage, selection (2,1), and selection (1,1) priors over the flat-$\mathbf{R}$ (flat-$\mathbf{\Pi}$) priors. This is quite a substantial drop for the small sample size. For the other sample sizes we still

			Risk Estimates by Prior						
R	N	Loss Fcn	Shrinkage	Selection (2,1)	Selection (1,1)	flat-**R**	flat-**Π**	Triangular	Naive Shrink
A	20	1	0.53	0.34	0.36	0.88	0.74	0.69	0.82
A	50	1	0.23	0.13	0.14	0.35	0.32	0.30	0.34
A	200	1	0.057	0.027	0.027	0.076	0.074	0.073	0.075
B	20	1	0.081	0.025	0.021	0.48	0.55	0.53	0.091
B	50	1	0.034	0.010	0.0083	0.23	0.24	0.24	0.039
B	200	1	0.0077	0.0014	0.0012	0.063	0.064	0.064	0.0085
C	20	1	0.69	0.77	0.88	1.02	0.80	0.70	0.68
C	50	1	0.28	0.33	0.36	0.36	0.31	0.29	0.33
C	200	1	0.067	0.081	0.083	0.073	0.070	0.070	0.070
D	20	1	0.61	0.61	0.68	1.11	0.81	0.74	0.88
D	50	1	0.26	0.29	0.30	0.39	0.33	0.32	0.33
D	200	1	0.066	0.068	0.070	0.075	0.073	0.072	0.073
A	20	2	0.25	0.13	0.12	0.52	0.46	0.44	0.47
A	50	2	0.13	0.039	0.042	0.22	0.21	0.20	0.22
A	200	2	0.035	0.0064	0.0066	0.053	0.052	0.052	0.052
B	20	2	0.066	0.015	0.012	0.39	0.45	0.44	0.071
B	50	2	0.031	0.0081	0.0065	0.21	0.22	0.22	0.035
B	200	2	0.0075	0.0013	0.0011	0.062	0.063	0.063	0.0083
C	20	2	0.35	0.42	0.48	0.60	0.47	0.41	0.39
C	50	2	0.15	0.19	0.21	0.21	0.18	0.17	0.19
C	200	2	0.039	0.053	0.055	0.044	0.042	0.042	0.041
D	20	2	0.28	0.28	0.31	0.53	0.44	0.40	0.45
D	50	2	0.14	0.16	0.17	0.22	0.19	0.19	0.20
D	200	2	0.037	0.038	0.040	0.044	0.043	0.042	0.044

Table 1: Risk estimates for simulation study with dimension $J = 6$. Correlation matrices: A autoregressive structure; B independence; C non-zero decaying; D sparse. Loss functions: $L_1(\hat{\mathbf{R}}, \mathbf{R}) = \text{tr}(\hat{\mathbf{R}}\,\mathbf{R}^{-1}) - \log|\hat{\mathbf{R}}\,\mathbf{R}^{-1}| - p$; $L_2(\hat{\mathbf{\Pi}}, \mathbf{\Pi}) = \sum_{i<j}(\hat{\pi}_{ij} - \pi_{ij})^2$.

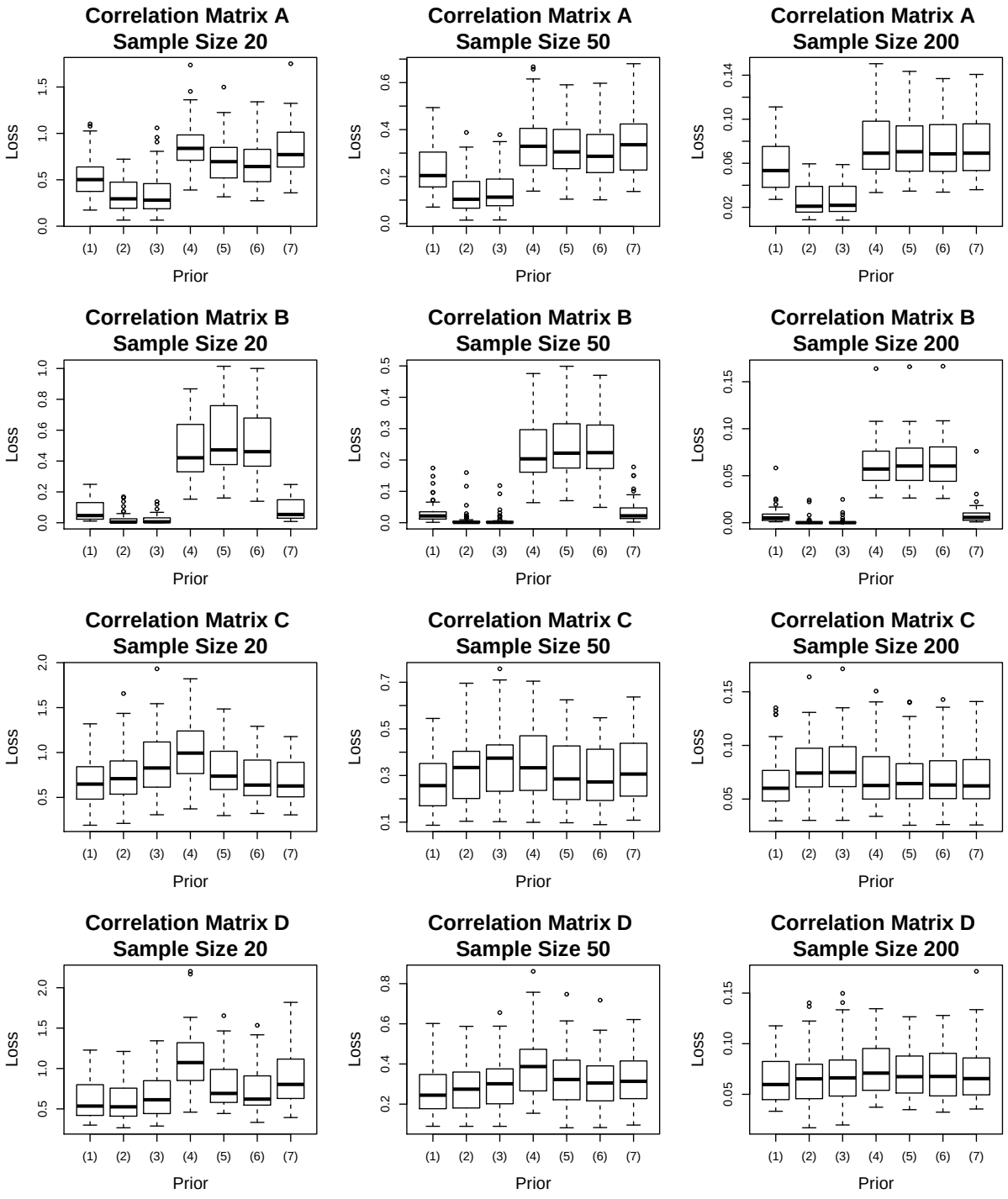

Figure 1: Box plots of the observed loss using $L_1(\hat{\mathbf{R}}_1, \mathbf{R})$ for the $J = 6$ cases. The prior distributions compared are (1) shrinkage, (2) selection (2,1), (3) selection (1,1), (4) flat-$\mathbf{R}$, (5) flat-$\mathbf{\Pi}$, (6) triangular, and (7) naive shrinkage.

$$\mathbf{\Pi}^{D'} = \begin{bmatrix} 1 & .9 & .3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.90 & 1 & .8 & .4 & .1 & 0 & 0 & 0 & 0 & 0 \\ 0.80 & 0.80 & 1 & .6 & .2 & 0 & 0 & 0 & 0 & 0 \\ 0.62 & 0.67 & 0.60 & 1 & .8 & .3 & 0 & 0 & 0 & 0 \\ 0.58 & 0.63 & 0.58 & 0.80 & 1 & .7 & 0 & 0 & 0 & 0 \\ 0.46 & 0.50 & 0.45 & 0.69 & 0.70 & 1 & .8 & .4 & .1 & 0 \\ 0.37 & 0.40 & 0.36 & 0.55 & 0.56 & 0.80 & 1 & .6 & .2 & 0 \\ 0.31 & 0.34 & 0.30 & 0.46 & 0.47 & 0.67 & 0.60 & 1 & .8 & .3 \\ 0.29 & 0.32 & 0.29 & 0.43 & 0.44 & 0.63 & 0.58 & 0.80 & 1 & .7 \\ 0.23 & 0.25 & 0.23 & 0.34 & 0.35 & 0.50 & 0.45 & 0.69 & 0.70 & 1 \end{bmatrix}$$

Table 2: 10×10 PAC matrix $\mathbf{\Pi}^{D'}$ shown above the diagonal and its respective correlation matrix $\mathbf{R}^{D'}$ shown below the diagonal.

observed a clear decrease over the flat priors. For $N = 50$ there is a drop of 32 (20), 26 (14), and 22 (9) percent for the sparse priors over the flat priors, and with $N = 200$ a decrease of 13 (9), 10 (7), and 7 (4) percent.

To investigate how our priors behave as J increases, we repeat the analysis using the non-sparse decaying $\mathbf{R}^C$ and a sparse $\mathbf{R}^{D'}$ with the dimension of the matrix increased to $J = 10$. Again, $\pi^C_{i,i+1} = 0.7$ for the lag-1 terms and $\pi^C_{ij} = 0.4^{j-i-1}$ for all $j - i > 1$, and we expand the previous $\mathbf{R}^D$ to the 10×10 $\mathbf{R}^{D'}$ shown in Table 2. As before the above diagonal elements are from $\mathbf{\Pi}^{D'}$ and the below diagonal elements from the corresponding $\mathbf{R}^{D'}$. $\mathbf{\Pi}^{D'}$ is very sparse, while $\mathbf{R}^{D'}$ has no zero elements. We consider sample sizes of 50 and 200. Risk estimates and box plots for this simulation are displayed in Table 3 and Figure 2.

From both Table 3 and Figure 2 it is clear that estimation of the correlation matrix is improved under the sparse priors. In the simulations of both dimensions we find that the estimators from the triangular selection prior tend to be slightly better than the selection prior with SBeta(1,1). With the sparse correlation matrix $\mathbf{R}^{D'}$ the risk under the sparse priors are about half of the risk of the flat prior under both sample sizes. Recall that $\mathbf{\Pi}^C$ is not sparse but has elements which decay exponentially. Because many of the large lag components are very small, the selection priors provide stability by explicitly zeroing many of these out. For the larger sample size, the flat priors do comparatively better although still worse than the sparse priors.

We have demonstrated that the sparse priors yield improved estimation of the correlation matrix in a variety of data situations. In order to investigate the performance in the standard situation

R	N	Loss Fcn	Risk Estimates by Prior: Shrinkage	Selection (2,1)	Selection (1,1)	flat-**R**	flat-$\mathbf{\Pi}$	Triangular	Naive Shrink
C	50	1	0.61	0.73	0.80	1.29	0.99	0.93	1.09
C	200	1	0.163	0.214	0.219	0.254	0.232	0.228	0.236
D′	50	1	0.47	0.52	0.57	1.22	0.95	0.90	1.00
D′	200	1	0.132	0.129	0.133	0.251	0.231	0.228	0.235
C	50	2	0.37	0.46	0.49	0.86	0.71	0.67	0.74
C	200	2	0.111	0.156	0.160	0.186	0.176	0.172	0.174
D′	50	2	0.26	0.31	0.34	0.76	0.68	0.66	0.68
D′	200	2	0.084	0.079	0.081	0.184	0.177	0.175	0.177

Table 3: Risk estimates for simulation study with dimension $J = 10$. Correlation matrices: C nonzero decaying; D' sparse. Loss functions: $L_1(\hat{\mathbf{R}}, \mathbf{R}) = \text{tr}(\hat{\mathbf{R}}\,\mathbf{R}^{-1}) - \log|\hat{\mathbf{R}}\,\mathbf{R}^{-1}| - p$; $L_2(\hat{\mathbf{\Pi}}, \mathbf{\Pi}) = \sum_{i<j}(\hat{\pi}_{ij} - \pi_{ij})^2$.

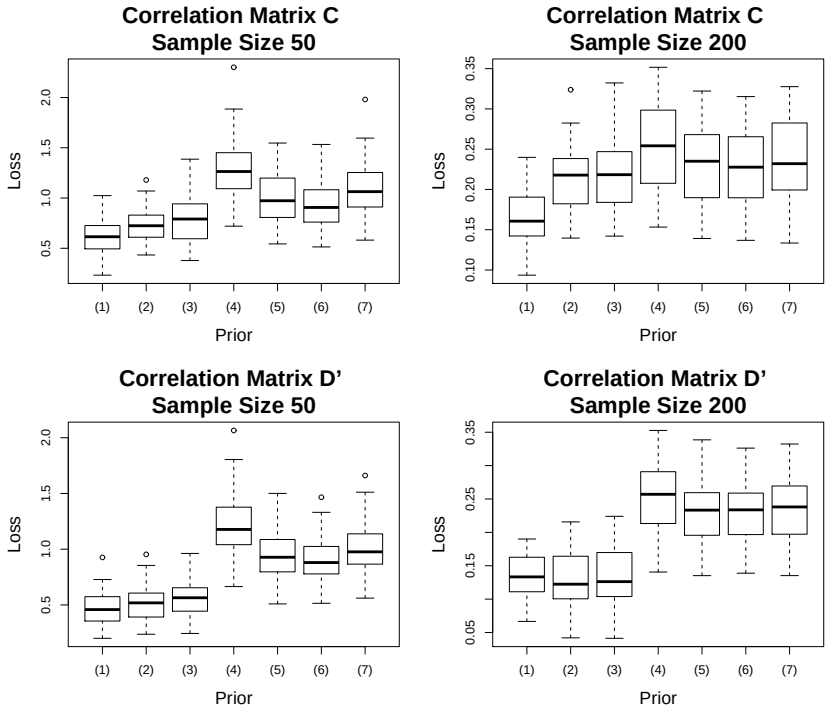

Figure 2: Box plots of the observed loss using $L_1(\hat{\mathbf{R}}_1, \mathbf{R})$ for $J = 10$. The prior distributions compared are (1) shrinkage, (2) selection (2,1), (3) selection (1,1), (4) flat-**R**, (5) flat-$\mathbf{\Pi}$, (6) triangular, and (7) naive shrinkage.

where the true dependence structure is unknown, we apply the sparsity and shrinkage priors to a data set obtained from a smoking cessation clinical trail.

6. Data analysis

The first Commit to Quit (CTQ I) study (Marcus et al., 1999) was a clinical trial designed to encourage women to stop smoking. As weight gain is often a viewed as a factor decreasing the effectiveness of smoking cessation programs, a treatment involving an exercise regimen is utilized to try to increase the quit rate. The control group received an educational intervention of equal time. The study ran for twelve weeks, and patients were encouraged to quit smoking at week 5. As the study required a significant time commitment (three exercise/educational sessions a week), there is substantial missingness due to study dropout. As in previous analyses of this data (Daniels and Hogan, 2008), we assume this missingness is ignorable.

For patient $i = 1, \ldots, N$ ($N = 281$), we denote the vector of quit statuses by $\mathbf{Q}_i = (Q_{i1}, \ldots, Q_{iJ})'$. We only consider the responses after patients are asked to quit, weeks 5 through 12 ($J = 8$). Here $Q_{it} = 1$ indicates a success (not smoking) for patient i at time t ($1 \leq t \leq J$, corresponding to week $t + 4$), $Q_{it} = -1$ for a failure (smoking during the week), and $Q_{it} = 0$ if the observation is missing. Following the usual conventions of the multivariate probit regression model (Chib and Greenberg, 1998), we let $\mathbf{Y}_i$ be the J-dimensional vector of latent variables corresponding to $\mathbf{Q}_i$. Thus, $Q_{it} = 1$ implies that $Y_{it} \geq 0$, and $Q_{it} = -1$ gives $Y_{it} < 0$. When $Q_{it} = 0$, the sign of Y_{it} represents the (unobserved) quit status for the week.

We assume the latent variables follow a multivariate normal distribution $\mathbf{Y}_i \sim \mathrm{N}_J(\boldsymbol{\mu}_i, \mathbf{R})$ for $i = 1, \ldots, N$, where $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$, $\mathbf{X}_i$ is a $J \times q$ matrix of covariates and $\boldsymbol{\beta}$ a q-vector of regression coefficients. As the scale of $\mathbf{Y}$ is unidentified, the covariance matrix of $\mathbf{Y}$ is constrained to be a correlation matrix $\mathbf{R}$. We consider two choices of $\mathbf{X}_i$: 'time-varying' which specifies a different μ_{it} for each time within each treatment group ($q = 2J$) and 'time-constant' which gives the same value of μ_{it} across all times within treatment group ($q = 2$).

With the time-constant and time-varying choices of the mean structure, we consider the following priors for $\mathbf{R}$: shrinkage, selection, flat-$\mathbf{R}$, flat-$\mathbf{\Pi}$, triangular, naive shrinkage, and an autoregressive (AR) prior. The AR prior assumes an AR(1) structure for $\mathbf{R}$, that is, $\rho_{ij} = \rho^{|j-i|}$ and $\pi_{i,i+1} = \rho$ and $\pi_{ij} = 0$ if $|j - i| > 1$. We assume a Unif$(-1, 1)$ distribution for ρ. As in

the risk simulation, we consider the selection prior with both SBeta$(1,1)$ and with SBeta$(2,1)$ for the continuous component. The remaining prior distributions to be specified are $\epsilon_0 \sim$ Unif$(0,1)$, $\gamma \sim$ Gamma$(5,5)$, and the prior on the regression coefficients $\boldsymbol{\beta}$ is flat.

To analyze the data we run an MCMC chain for 12,000 iterations after a burn-in of 3000, retaining every tenth observation. Convergence was assessed through graphical diagnostics and deemed adequate. There are three sets of parameters to sample in the MCMC chain: the regression coefficients, the correlation matrix, and the latent variables. The conditional for $\boldsymbol{\beta}$ given $\mathbf{Y}$ and $\mathbf{R}$ is multivariate normal. Sampling the correlation matrix evolves as discussed in Sections 3.2 and 4.3 using the residuals $\mathbf{Y}_i - \boldsymbol{\mu}_i$. The latent variables $\mathbf{Y}_i$, which are constrained by $\mathbf{Q}_i$, are sampled according to the strategy of Liu et al. (2009, Proposition 1).

To compare the specification based on our prior choices, we make use of the deviance information criterion (DIC; Spiegelhalter et al., 2002). The DIC statistic can be viewed similarly to the Bayesian or Akaike information criterion, but the DIC does not require the user to "count" the number of model parameters. This is key for Bayesian models that utilize shrinkage and/or sparsity priors as it is not clear whether or how one should count a parameter that has been set to or shrunk toward zero. To that end, let

$$\text{Dev} = -2\text{loglik}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{R}}|\mathbf{Q}) = \sum_i -2\text{loglik}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{R}}|\mathbf{Q}_i) \tag{8}$$

be the deviance or twice the negative log-likelihood with the parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{R}}$. Here $\hat{\boldsymbol{\beta}}$ is the posterior mean, and for the correlation estimate $\hat{\mathbf{R}}$, we use the first of the estimators we considered in Section 5, $\hat{\mathbf{R}} = \mathbf{S}\,\text{E}\{\mathbf{R}^{-1}\}^{-1}\,\mathbf{S}$ with $\mathbf{S} = [\text{diag}(\text{E}\{\mathbf{R}^{-1}\})]^{1/2}$. The complexity of the model is measured by the term p_D, sometimes called the effective number of parameters. This p_D is calculated as

$$p_D = \text{E}\{-2\text{loglik}(\boldsymbol{\beta}, \mathbf{R}|\mathbf{Q})\} - \text{Dev}, \tag{9}$$

where the expectation is over the posterior distribution of the parameters $(\boldsymbol{\beta}, \mathbf{R})$. The DIC model comparison statistics is DIC $=$ Dev $+ 2p_D$, the sum of terms measuring model fit and complexity. Smaller values of DIC are preferred.

As Wang and Daniels (2011) point out, the DIC should be calculated using the observed data, which in this case is the quit status responses $\mathbf{Q}_i$ not the latent variables $\mathbf{Y}_i$. Hence the log-likelihood for $\mathbf{Q}_i$ at parameters $(\boldsymbol{\beta}, \mathbf{R})$ is equal to

$$\text{loglik}(\boldsymbol{\beta}, \mathbf{R}|\mathbf{Q}_i) = \log\left(\int_{(-\infty,\infty)^J} I\{Q_{it}y_t \geq 0\,\forall t\}\,\phi(\mathbf{y}|\mathbf{X}_i\boldsymbol{\beta}, \mathbf{R})\,d\mathbf{y}\right), \tag{10}$$

Mean Structure	Correlation Prior	Dev	p_D	DIC
Time-constant	Shrinkage	1031	14	1060
Time-constant	Selection (2,1)	1042	12	1066
Time-constant	Selection (1,1)	1044	12	1068
Time-constant	Triangular	1029	20	1068
Time-constant	flat-$\mathbf{\Pi}$	1029	20	1069
Time-constant	Naive shrinkage	1033	20	1074
Time-constant	AR	1071	3	1078
Time-constant	flat-$\mathbf{R}$	1043	21	1086
Time-varying	Shrinkage	1022	25	1071
Time-varying	Triangular	1017	30	1077
Time-varying	Selection (2,1)	1033	22	1077
Time-varying	Selection (1,1)	1036	22	1080
Time-varying	flat-$\mathbf{\Pi}$	1019	30	1080
Time-varying	Naive shrinkage	1023	31	1085
Time-varying	AR	1068	13	1093
Time-varying	flat-$\mathbf{R}$	1034	31	1097

Table 4: Model comparison statistics for the CTQ data.

where $\phi(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the J-dimensional multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The integral in (10) is not tractable but can be estimated using importance sampling (Robert and Casella, 2004, Section 3.3). See the appendix in the online supplementary materials for details about estimating the DIC. The model fit (Dev), complexity (p_D), and comparison (DIC) statistics are in Table 4; DIC statistics were estimated with a standard error of approximately 0.5.

We see that the models that use a mean structure that depends only on treatment and not time t tend to have lower DIC values. The time-varying models are penalized in the p_D term for having to estimate the additional 14 regression coefficients. Of the correlation priors the flat-$\mathbf{R}$ and AR priors perform much worse than the shrinkage, selection, triangular, and flat-PAC priors with the same mean structure. Additionally, the selection prior that uses the triangular form for SBeta ($\alpha = 2, \beta = 1$) tend to have a smaller DIC than the SBeta(1,1) priors. From Table 4 we determine the prior choice that best balances model fit with parsimony is clearly the model with time-constant mean structure and the shrinkage prior on the correlation matrix prior.

Using this best fitting model, the posterior mean of $\boldsymbol{\beta}$ is $(-0.504, -0.295)$ implying that the marginal probability (95% credible interval) of not smoking during a given study week is $\Phi(-0.504) = 0.307\,(0.24, 0.37)$ for the control group and $\Phi(-0.295) = 0.384\,(0.32, 0.45)$ for

the exercise group, where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. The test of the hypothesis that the control treatment is as effective as the exercise treatment (i.e., $H_0 : \beta_1 \geq \beta_2$) has a posterior probability of 0.06, providing some evidence to the claim that exercise improves cessation results.

We now examine in more detail the effect the shrinkage prior has on modeling the correlation matrix. The posterior means (credible interval) of the shrinkage parameters are $\hat{\epsilon}_0 = 0.406$ $(0.25, 0.60)$ and $\hat{\gamma} = 2.44$ $(1.6, 3.4)$. With a value of γ greater than 1, the variance of π_{ij} is decaying to zero fairly rapidly. The posterior mean of $\mathbf{\Pi}$ is

$$\hat{\mathbf{\Pi}} = \begin{bmatrix} 1.00 & 0.70 & 0.12 & 0.02 & 0.05 & 0.00 & 0.00 & -0.01 \\ 0.71 & 1.00 & 0.83 & 0.16 & 0.09 & 0.02 & 0.01 & 0.00 \\ 0.64 & 0.84 & 1.00 & 0.81 & 0.12 & 0.10 & 0.06 & 0.02 \\ 0.56 & 0.74 & 0.82 & 1.00 & 0.78 & 0.24 & 0.09 & 0.03 \\ 0.51 & 0.64 & 0.69 & 0.79 & 1.00 & 0.81 & 0.37 & 0.04 \\ 0.48 & 0.61 & 0.66 & 0.74 & 0.83 & 1.00 & 0.88 & 0.21 \\ 0.48 & 0.61 & 0.67 & 0.74 & 0.83 & 0.89 & 1.00 & 0.78 \\ 0.40 & 0.52 & 0.57 & 0.63 & 0.70 & 0.77 & 0.80 & 1.00 \end{bmatrix},$$

with the lower diagonal values giving the elements of $\hat{\mathbf{R}}$. We see that the PACs are far from zero in only the first two lags and the remaining π's are close to zero. This is because these partial autocorrelations have been shrunk almost to zero in most iterations.

7. Discussion

In this paper we have introduced two new priors for correlation matrices, a shrinkage prior and a selection prior. These priors choose a sparse parameterization of the correlation matrix through the set of PACs. In the selection context, by stochastically selecting the elements of $\mathbf{\Pi}$ to zero out, our model finds interpretable independence relationships for normal data and avoids the need for complex model selection of the dependence structure. A key improvement of the selection prior over existing methods for sparse correlation matrices is that our approach avoids the complex normalizing constants seen in previous work. Additionally, in settings with time-ordered data, the partial autocorrelations are more interpretable than the full partial correlations, as they do not involve conditioning on future values.

While the examples we have considered here involve situations where the covariance matrix was constrained (as in the data example) or known (as in the simulations) to be a correlation

matrix, the extension to arbitrary $\mathbf{\Sigma}$ is simple. Returning to the separation strategy $\mathbf{\Sigma} = \mathbf{SRS}$ (Barnard et al., 2000), a prior for $\mathbf{\Sigma}$ can be formed by placing independent priors on $\mathbf{S}$ and $\mathbf{R}$, i.e. $p(\mathbf{\Sigma}) = p(\mathbf{R})p(\mathbf{S})$. Using one of the proposed priors for $p(\mathbf{R})$, sensible choices of $p(\mathbf{S})$ include an independent inverse gamma for each of the σ_{jj} or a flat prior on $\{\mathbf{S} = \text{diag}(\sigma_{11}, \ldots, \sigma_{JJ}) : \sigma_{jj} > 0\}$. This leads to a prior on $\mathbf{\Sigma}$ with sparse PACs.

The simulations and data we have considered here deal with $\mathbf{Y}$ of low or moderate dimension. We provide a few comments regarding the scalability of our approach for data with larger J. As we believe that PACs of larger lag play a progressively smaller role in describing the (temporal) dependence, it may be reasonable to specify a maximum allowable lag for non-zero PACs. That is, we choose some k such that $\pi_{ij} = 0$ for all $j - i > k$ and sample π_{ij} $(j - i \leq k)$ from either our shrinkage or selection prior. Banding the $\mathbf{\Pi}$ matrix is related to the idea of banding the covariance matrix (Bickel and Levina, 2008), concentration matrix (Rothman et al., 2008), or the Cholesky decomposition of $\mathbf{\Sigma}^{-1}$ (Rothman et al., 2010). Banding $\mathbf{\Pi}$ has also been studied by Wang and Daniels (2013b). In addition to reducing the number of parameters that must be sampled, other matrix computations will be faster by using properties of banded matrices.

Related to this, modifications to the shrinkage prior may be needed for larger dimension J. Recall that the variance of π_{ij} is $\xi_{ij} = \epsilon_0|j - i|^{-\gamma}$. For large lags, this can be very close to zero leading to numerical instability; recall the parameters of the SBeta distribution are inversely related to ξ_{ij} through $\alpha_{ij} = \beta_{ij} = (\xi_{ij}^{-1} - 1)/2$. Replacing (3) with $\xi_{ij} = \epsilon_0 \min\{|j - i|, k\}^{-\gamma}$ or $\xi_{ij} = \epsilon_0 + \epsilon_1|j - i|^{-\gamma}$ to bound the variances away from zero or banding $\mathbf{\Pi}$ after the first k lags provide two possibilities to avoid such numerical issues.

Further, we have parametrized the variance component and the selection probability in similar ways in our two sparse priors. The quantity is of the form $\epsilon_0|j - i|^{-\gamma}$ for both ξ_{ij} in (3) and ϵ_{ij} in (6), but other parameterizations are possible. We have considered some simulations (not included) allowing the variance/selection probability to be unique for lag, i.e. $\epsilon_{ij} = \epsilon_{|j-i|}$. A prior needs to be specified for each of these $J - 1$ ϵ's, ideally decreasing in lag. Alternatively, one could use $\epsilon_0/|j - i|$, which can be viewed as a special case where the prior on γ is degenerate at 1. In our experience results were not very sensitive to the choice of the parameterization, and posterior estimates of $\mathbf{\Pi}$ and $\mathbf{R}$ were similar.

In addition, we have focused our discussion on the correlation estimation problem in the context of analysis with multivariate normal data. We note that these priors are additionally applicable in

the context of estimating a constrained scale matrix for the multivariate Student t-distribution. Consider the random variable $\mathbf{Y} \sim t_J(\boldsymbol{\mu}, \mathbf{R}, \nu)$. That is, $\mathbf{Y}$ follows a J-dimensional t-distribution with location (mean) vector $\boldsymbol{\mu}$, scale matrix $\mathbf{R}$ (constrained to be a correlation matrix), and ν degrees of freedom (either fixed or random). Using the gamma-mixture-of-normals technique (Albert and Chib, 1993), we rewrite the distribution of $\mathbf{Y}$ to be $\mathbf{Y}|\tau \sim \mathrm{N}_J(\boldsymbol{\mu}, \tau^{-1}\mathbf{R})$ and $\tau \sim \mathrm{Gamma}(\nu/2, \nu/2)$. Sampling for $\mathbf{R}$ as part of an MCMC chain follows as in Sections 3.2 and 4.3 using $\mathbf{Y}^\star = \sqrt{\tau}(\mathbf{Y} - \boldsymbol{\mu})$ as the data. However, one should note that a zero PAC π_{ij} implies that Y_i and Y_j are uncorrelated given $Y_{i+1}, \ldots, Y_{j-1}$, but this is not equivalent to conditional independence as in the normal case.

Acknowledgments

This research was partially supported by NIH CA-85295.

Supplementary Materials

Supplemental Archive: The package contains an Appendix providing details about the DIC calculation, a simulated dataset similar to the CTQ I data, `R` code for the CTQ data analysis, and a file `README.txt` containing describing the code. (PACS_supplement.zip)

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis, 2nd Edition.* Wiley.

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

Bickel, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Carter, C. K., Wong, F., and Kohn, R. (2011). Constructing priors based on model size for non-decoposable Gaussian graphical models: A simulation based approach. *Journal of Multivariate Analysis*, 102:871–883.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61:331–344.

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 Normand, S.-L. (2006). Longitudinal profiling of health care units based on mixed multivariate patient outcomes. *Biostatistics*, 7:1–15.

Daniels, M. J. and Pourahmadi, M. (2009). Modeling covariance matrices via partial autocorrelations. *Journal of Multivariate Analysis*, 100(10):2352 – 2363.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28:157–75.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 311–319. Univesity of California Press.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97:2177–2189.

Kurowicka, D. and Cooke, R. (2003). A parameterization of positive definite matrices in terms of partial correlation vines. *Linear Algebra and its Applications*, 372:225–251.

Kurowicka, D. and Cooke, R. (2006). Completion problem with partial correlation vines. *Linear Algebra and its Applications*, 418:188–200.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press.

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

Liu, C. (2001). Comment on "The art of data augmentation" by D. A. van Dyk and X.-L. Meng. *Journal of Computational and Graphical Statistics*, 10(1):75–81.

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

Liu, X., Daniels, M. J., and Marcus, B. (2009). Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. *Journal of the American Statistical Association*, 104(486):429–438.

Marcus, B., Albrecht, A., King, T., Parisi, A., Pinto, B., Roberts, M., Niaura, R., and Abrams, D. (1999). The efficacy of exercise as an aid for smoking cessation in women: A randomized controlled trial. *Archives of Internal Medicine*, 159:1229–1234.

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

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93:537–554.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, second edition.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rothman, A. J., Levina, E., and Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, 97(3):539–550.

Rousseeuw, P. J. and Molenberghs, G. (1994). The shape of correlation matrices. *The American Statistician*, 48(4):276–279.

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.

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 (with correction). *Biometrics*, 67(3):810–818.

Wang, Y. and Daniels, M. J. (2013a). Bayesian modeling of the dependence in longitudinal data via partial autocorrelations and marginal variances. *Journal of Multivariate Analysis*, 116:130–140.

Wang, Y. and Daniels, M. J. (2013b). Estimating large correlation matrices by banding the partial autocorrelation matrix. Technical report, University of Florida, Gainesville, Florida.

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.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

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