Published in final edited form as:

J Comput Graph Stat. 2016 January 2; 25(1): 167–186. doi:10.1080/10618600.2015.1028549.

Covariance Partition Priors: A Bayesian Approach to Simultaneous Covariance Estimation for Longitudinal Data

J. T. Gaskins* and M. J. Daniels†

J. T. Gaskins: jeremy.gaskins@louisville.edu; M. J. Daniels: mjdaniels@austin.utexas.edu

*Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202

†Department of Integrative Biology, Division of Statistics and Scientific Computation, University of Texas, Austin, TX 78712

Abstract

The estimation of the covariance matrix is a key concern in the analysis of longitudinal data. When data consists of multiple groups, it is often assumed the covariance matrices are either equal across groups or are completely distinct. We seek methodology to allow borrowing of strength across potentially similar groups to improve estimation. To that end, we introduce a covariance partition prior which proposes a partition of the groups at each measurement time. Groups in the same set of the partition share dependence parameters for the distribution of the current measurement given the preceding ones, and the sequence of partitions is modeled as a Markov chain to encourage similar structure at nearby measurement times. This approach additionally encourages a lower-dimensional structure of the covariance matrices by shrinking the parameters of the Cholesky decomposition toward zero. We demonstrate the performance of our model through two simulation studies and the analysis of data from a depression study. This article includes Supplementary Material available online.

Keywords

Cholesky parametrization; clustering; Markov chains; shrinkage; sparsity

1. Introduction

When modeling longitudinal data, estimation of the covariance matrices is of prime importance. In many cases the data may consist of multiple groups defined by differences in treatments and/or baseline covariates. An important consideration for the analyst is whether, and how, the dependence structure varies across these groups. Often, one performs inference under the assumption that the covariance structures are equal across all groups, but if this assumption is incorrect, the resulting inferences may be invalid, even for the mean parameters when the data is subject to missingness (Daniels and Hogan, 2008). Conversely,

Supplementary Materials

Supplemental Archive: The package contains an appendix providing further details about the sampling algorithm and the simulations studies of Section 4, as well as additional simulation results. Also included are a simulated data set similar to the depression data, R code for the depression data analysis, R code for computing $a_Q(\mathcal{P})$ and A_Q for M groups, and a file README.txt containing a description of the code. (CPP_supplement.zip)

modeling the dependence structures independently without regard to the other groups can lead to inefficiency if the group sample sizes are small or the dimension is large. Our goal is to develop methodology that allows the sharing of information across groups to improve estimation efficiency.

Consider M groups containing n_m observations, $\mathbf{Y}_{mi}$ ($i = 1, \dots, n_m$; $m = 1, \dots, M$), of T -dimensional, multivariate normal data. Without loss of generality, we assume $\mathbf{Y}_{mi}$ has mean zero; otherwise, we let $\mathbf{Y}_{mi}$ represent the residual after centering. Each group m has its own covariance matrix Σ_m , and we let $\Sigma = \{\Sigma_1, \dots, \Sigma_M\}$ denote the collection of covariance matrices. We parametrize each Σ_m through the modified Cholesky decomposition

(Pourahmadi, 1999, 2000), so $\Sigma_m^{-1} = \mathbf{T}_m \mathbf{G}_m \mathbf{T}_m^\top$ with $\mathbf{G}_m$ a diagonal matrix with positive entries and $\mathbf{T}_m$ an upper-triangular matrix with unit diagonal. The parameters of this Cholesky parametrization are interpreted by considering the sequential factorization of the distribution of $\mathbf{Y}_{mi}$

$$f(Y_{mi1}, \dots, Y_{miT}) = f(Y_{mi1})f(Y_{mi2}|Y_{mi1}) \cdots f(Y_{miT}|Y_{mi1}, \dots, Y_{mi,T-1}). \quad (1)$$

Under multivariate normality, each of the T sequential distributions, $f(Y_{miT}|Y_{mi1}, \dots,$

$Y_{mi,t-1})$, is a normal distribution with mean $\sum_{j=1}^{t-1} \phi_{mjt} Y_{mij}$ and variance γ_{mt} . Let $\boldsymbol{\phi}_{mt} = (\phi_{m1t}, \dots, \phi_{m,t-1,t})^\top$ be the vector of regression coefficients from the sequential regression for the t th response. From the Cholesky decomposition the unconstrained elements of the t th column of $\mathbf{T}_m$ are $-\phi_{mt}$ and the (t, t) -element of $\mathbf{G}_m$ is $1/\gamma_{mt}$. Hence, these sequential distributions fully and uniquely determine Σ_m . Further, Σ_m will be positive definite as long as $\gamma_{mt} > 0$ for all t , and the normal and inverse gamma distributions provide conditionally conjugate priors for $\boldsymbol{\phi}_{mt}$ and γ_{mt} (Daniels and Pourahmadi, 2002). We refer to the ϕ_{mjt} s as the generalized autoregressive parameters (GARPs) and the γ_{mt} s as the innovation variances (IVs).

There has been a substantial amount of research on the simultaneous covariance estimation problem, but relatively little focus on the longitudinal data scenario. Methods have been suggested by imposing commonalities on a particular feature of the Σ_m s: equality across subsets of the principle components of Σ (Boik, 2002) or its correlation matrix (Boik, 2003); equality across the correlation matrices or the volumes of Σ (Manly and Rayner, 1987); equality across all $\mathbf{T}_m$ and/or all $\mathbf{G}_m$ (McNicholas and Murphy, 2010); equality among arbitrary subsets of the parameters of $\mathbf{T}_m$ and $\mathbf{G}_m$ (Pourahmadi et al., 2007). The later is in the spirit of our method but requires the subsets be provided by the user. This presents a significant computational challenge as there are too many possible configurations to find the best without an automated approach. Daniels (2006) and Hoff (2009) propose shrinkage priors on the Cholesky terms and the eigenvectors, respectively. Guo et al. (2011) and Danaher et al. (2013) propose penalized likelihood methods that induce a sparsity structure common across groups. Other authors have modeled the covariance matrices through regressions of a continuous covariate (e.g., Chiu et al., 1996; Daniels, 2006; Fox and Dunson, 2011; Hoff and Niu, 2012), but the regression parameters in these models often lack interpretation.

Gaskins and Daniels (2013) proposed a nonparametric prior for the Cholesky terms based on the matrix stick-breaking process (Dunson et al., 2008). This methodology considers an enormous model space that includes allowing equality of individual φ_{mjt} across groups m . This leads to equality constraints of order T^2 , which may be unmanageable for data with moderate-to-large dimension. In this article we simplify the set of models under consideration while still permitting a rich set of sparse models with common structures across groups.

In Section 2 we introduce our approach based on a covariance partition prior. To share information across groups, this prior specifies partitions of the groups such that those groups contained in the same set of the partition will have common values for some subset of the dependence parameters. We define a partition corresponding to each of the sequential distributions in (1) and consider the sequence of partitions to behave as a Markov chain. Section 3 contains details of the computational algorithm needed to generate a posterior sample. Performance of the covariance partition prior is studied in Sections 4 and 5 through a simulation study and the analysis of data from a depression study. A brief discussion in Section 6 concludes the article.

2. Covariance partition prior

2.1. Prior on the sequence of partitions

First, it is necessary to introduce notation. Let $\mathcal{M} = \{1, \dots, M\}$ denote the collection of all groups. We are interested in partitioning the groups into sets that share a similar dependence structure. Let $\mathcal{P}$ denote a partition of $\mathcal{M}$, and Ω be the collection of all possible $\mathcal{P}$. The cardinality of Ω is B_M , the M th Bell number (Stanley, 1997, p. 33). B_M is equal to the sum of the Stirling numbers of the second kind. The partition $\mathcal{P}$ can be written as the collection of its sets $\mathcal{P} = \{S_1, \dots, S_d\}$, where each S_i is non-empty, d is the degree (the number of distinct sets) of $\mathcal{P}$, $S_i \cap S_j = \emptyset$ for all $i \neq j$, and $\cup_{i=1}^d S_i = \mathcal{M}$. We also consider two special partitions $\mathcal{P}_{\text{pool}}$ and $\mathcal{P}_{\text{ind}}$. $\mathcal{P}_{\text{ind}} = \{\{1\}, \dots, \{M\}\}$ represents the partition where each group is in a singleton set, corresponding to the case where no information is shared across groups, and $\mathcal{P}_{\text{pool}} = \{\mathcal{M}\}$ pools the data into a single group.

The goal is to partition $\mathcal{M}$, so that groups in a common set of $\mathcal{P}$ share common values of a subset of the dependence parameters. To that end, we introduce T partitions, $\mathcal{P}_1, \dots, \mathcal{P}_T$, one for each measurement time t . Groups in a common set of the t th partition $\mathcal{P}_t$ have the same GARPs and IV corresponding to the sequential distribution $f(Y_{mit} | Y_{mi1}, \dots, Y_{mi,t-1})$. By imposing clustering on these sequential distributions, our method considers matching across groups for the T sequential distributions instead of the $T(T-1)/2$ Cholesky parameters, as when matching is considered for each GARP and IV individually (Gaskins and Daniels, 2013). To employ such an approach, it is necessary to define a prior distribution for this sequence of partitions $\{\mathcal{P}_1, \dots, \mathcal{P}_T\}$.

There is a considerable literature regarding the clustering of data through a single partition $\mathcal{P} = \{S_1, \dots, S_d\}$. The most common methods revolve around the product partition model (Hartigan, 1990) where the likelihood of the partition $\mathcal{P}$ has the form $\text{pr}(\mathcal{P}) = c_0 \prod_{i=1}^d \pi(S_i)$.

This model provides conjugacy and the desirable interpretation that the probability associated with set S_i does not depend on any elements of $\mathcal{M}$ outside of S_i . These models have been most effectively used when the collection $\mathcal{M}$ to partition is quite large (e.g. Crowley, 1997; Booth et al., 2008) and frequently use the partition structure implied by the clustering of a Dirichlet process $\pi(S_i) = \alpha(n_i!)$, where α is the concentration parameter and n_i the degree of set S_i .

However, our focus is on a joint prior $\pi(\wp_1, \dots, \wp_T)$ for an ordered sequence of T partitions, and we desire a distribution that will encourage similar structures across t . If groups m_1 and m_2 are in the same set in $\wp_t$ and therefore, share the same GARPs and IV for response t , it should be more likely that they are in the same set of $\wp_{t+1}$. To that end, we consider the sequence of partitions $\{\wp_1, \dots, \wp_T\}$ to be a Markov process. The distribution of $\wp_t$ given all previous partitions depends only on the most recent partition, that is, $\text{pr}(\wp_t | \wp_1, \dots, \wp_{t-1}) = \text{pr}(\wp_t | \wp_{t-1})$ for all t . Hence, given the adjacent partitions $\wp_{t-1}$ and $\wp_{t+1}$, $\wp_t$ is independent of the remaining $\wp_j$ ($|t-j| > 1$).

To define the transition probability $\text{pr}(\wp_t | \wp_{t-1})$, we employ a commonly used metric on Ω given by $d(\wp_1, \wp_2) = 2|\wp_1 \cap \wp_2| - |\wp_1| - |\wp_2|$, where $|\wp|$ gives the degree of partition $\wp$ and the intersection partition is $\wp_1 \cap \wp_2 = \{S : S \cap \wp_1 = S \cap \wp_2 \neq \emptyset; S = S_1 \cap S_2 \text{ for some sets } S_1 \in \wp_1 \text{ and } S_2 \in \wp_2\}$ (Day, 1981). Note that two groups m_1, m_2 are in the same set of the partition $\wp_1 \cap \wp_2$ if and only if they were in the same set in both $\wp_1$ and $\wp_2$. This distance may be interpreted as the minimum number of merges and splits of the sets of $\wp_1$ needed to obtain $\wp_2$. For instance, the distance between the partitions $\{\{1, 2, 3\}, \{4\}\}$ and $\{\{1, 2\}, \{3, 4\}\}$ is 2, as $\{1, 2, 3\}$ must be split into $\{1, 2\}$ and $\{3\}$ and the sets $\{3\}$ and $\{4\}$ merged. Other metrics on Ω can be found in Arabie and Boorman (1973) or Denoeud and Guénoche (2006) and could alternatively be used with our approach. Based on the metric $d(\cdot, \cdot)$, we define the closeness between the two partitions by $c_q(\wp_1, \wp_2) = [1 + \{d(\wp_1, \wp_2)\}^q]^{-1}$, where q is a non-negative parameter determining the smoothness of Markov process. Note for all positive, finite q , $c_q(\wp_1, \wp_2) = 1$ if and only if $\wp_1 = \wp_2$, $c_q(\wp_1, \wp_2) \in (0, 1)$ for $\wp_1 \neq \wp_2$, and $c_q(\cdot, \cdot)$ is decreasing in $d(\cdot, \cdot)$ at a rate depending on q .

For the partition $\wp$ define $a_q(\wp) = B_M^{-1} \sum_{\wp' \in \Omega} c_q(\wp, \wp')$ to be the attractiveness of $\wp$, given by the average closeness of $\wp$ to $\wp'$ over all $\wp' \in \Omega$. The transition probability from $\wp_{t-1}$ to $\wp_t$ is proportional to the closeness of the partitions and is given by

$$\text{pr}(\wp_t | \wp_{t-1}) = \frac{c_q(\wp_{t-1}, \wp_t)}{a_q(\wp_{t-1}) B_M}. \quad (2)$$

It is easy to verify that this choice of transition function yields a stationary distribution that assigns the probability of $\wp$ to be proportional to its attractiveness $a_q(\wp)$. In particular, $\text{pr}(\wp_t | \wp_{t-1}) > 0$ for all $t, \wp_t$, and $\wp_{t-1}$, so the resulting Markov chain is irreducible, aperiodic, time-reversible, and has stationary probabilities given by

$$\text{pr}(\mathcal{P}_t) = \frac{a_q(\mathcal{P}_t)}{A_q B_M}, \quad (\mathcal{P}_t \in \Omega, t=1, \dots, T), \quad (3)$$

where $A_q = B_M^{-1} \sum_{\mathcal{P}} a_q(\mathcal{P})$ is the average attractiveness. Hence, we choose the distribution for the initial partition $\mathcal{P}_1$ to be the set of stationary probabilities (3) and make all subsequent moves according to (2). Because we are starting the Markov chain at its stationary distribution, the marginal probability for partition $\mathcal{P}_t$ at each time t is given by its stationary probability (3). Combining (2) and (3), the prior distribution of the entire partition process is then

$$\pi(\mathcal{P}_1, \dots, \mathcal{P}_T) = \text{pr}(\mathcal{P}_1) \prod_{t=2}^T \text{pr}(\mathcal{P}_t | \mathcal{P}_{t-1}) = B_M^{-T} \frac{a_q(\mathcal{P}_1)}{A_q} \prod_{t=2}^T \frac{c_q(\mathcal{P}_{t-1}, \mathcal{P}_t)}{a_q(\mathcal{P}_{t-1})}. \quad (4)$$

Our partition prior clearly depends on the value of the smoothness parameter q . When $q = 0$, $c_0(\mathcal{P}_1, \mathcal{P}_2) = 1/2$ for all $\mathcal{P}_1, \mathcal{P}_2$ under the usual convention $0^0 = 1$. Hence, $a_q(\mathcal{P}) = 1/2$ for all $\mathcal{P}$, and $\text{pr}(\mathcal{P}_t | \mathcal{P}_{t-1}) = \text{pr}(\mathcal{P}_t) = 1/B_M$. Thus, the previous partition provides no information about the new partition, and all partitions are equally likely. We call this special case the independent-uniforms prior, because each partition is independent and follows a uniform distribution over Ω .

Conversely, as q gets larger, $c_q(\mathcal{P}_1, \mathcal{P}_2) \approx 0$ if $d(\mathcal{P}_1, \mathcal{P}_2) > 1$. Hence, moves from $\mathcal{P}_{t-1}$ to $\mathcal{P}_t$ that require more than one merge or split are highly unlikely. For large q this places a restrictive structure on the partition process that conflicts with the prior's intended purpose. Hence, we choose an upper bound Q for the parameter q so that $\text{pr}(\mathcal{P}_t | \mathcal{P}_{t-1})$ will be bounded away from zero for all $q \in [0, Q]$ and all $t, \mathcal{P}_{t-1}, \mathcal{P}_t$. Using only values of $q < Q$ means that our prior distribution will have large support on $\{\mathcal{P}_1, \dots, \mathcal{P}_T\}$ and will not be solely concentrated on the subspace of Ω^T with adjacent partitions differing by a distance of at most 1. In the simulation and data examples of Sections 4 and 5, we consider $M = 8$ groups. In this case, we choose $Q = 10$ and note that $\sum_{\mathcal{P}} |\text{pr}(\mathcal{P} | q = 10) - \text{pr}(\mathcal{P} | q = \infty)| \approx .001$, that is, there is little difference in terms of total variation distance between the stationary distributions for q greater than 10.

It is now apparent that the parameter q determines the smoothness of the partition process. Small values of q correspond to a highly variable process where successive partitions may differ greatly. Large values of q yield a sequence where adjacent partitions are unlikely to differ by more than one merge or split. It is then relevant to consider the effect of q on the marginal probabilities (3). For each q the highest probability partition is $\mathcal{P}_{\text{pool}} = \{\mathcal{M}\}$. For small q all marginal probabilities are similar ($\text{pr}(\mathcal{P} | q) \approx B_M^{-1}$), but as q increases, the disparity across $\mathcal{P}$ increases. For instance, when $q = 5$, $\max_{\mathcal{P}, \mathcal{P}'} \{\text{pr}(\mathcal{P}) / \text{pr}(\mathcal{P}')\} = 11.4$, but as the support Ω is large, this difference is not so great as to leave any partitions with a negligible marginal probability. This disparity does not increase substantially after $q = 5$ as each $c_q(\mathcal{P}, \mathcal{P}')$ is approximately either 1, $1/2$, or 0 for $q > 5$.

As it is not clear a priori what would be an optimum value of q , we place a prior on the smoothness. Because a Metropolis-Hastings (MH) sampling step will require the values of A_q and $a_q(\mathcal{P}_1), \dots, a_q(\mathcal{P}_T)$ for each q , we specify a finite support to minimize the computational complexity. To that end, we choose the sequence $\mathcal{Q}$ from 0 to $Q = 10$ with steps of 0.025 ($|\mathcal{Q}| = 401$) to be the support of q for the $M = 8$ case. The prior for q is a discrete uniform over the set $\mathcal{Q}$.

2.2. Prior on the Cholesky parameters

Recall that given the partition $\mathcal{P}_t$ groups in a common set share common values for the dependence parameters associated with the sequential distribution $k(Y_{mi1}|Y_{mi1}, \dots, Y_{mi,t-1})$. The parameters of this distribution are $(\boldsymbol{\varphi}_{mt}, \gamma_{mt})$, where $\boldsymbol{\varphi}_{mt}$ is empty when $t = 1$. To that end, for each set $\mathcal{S}_{it} \in \mathcal{P}_t = \{\mathcal{S}_{1t}, \dots, \mathcal{S}_{d_t t}\}$, we associate the parameters $(\tilde{\boldsymbol{\varphi}}_{it}, \tilde{\gamma}_{it})$, so that $(\boldsymbol{\varphi}_{mt}, \gamma_{mt}) = (\tilde{\boldsymbol{\varphi}}_{it}, \tilde{\gamma}_{it})$ for all $m \in \mathcal{S}_{it}$ and specify the prior in terms of $(\tilde{\boldsymbol{\varphi}}_{it}, \tilde{\gamma}_{it})$.

While the main goal of our methodology is to borrow strength in estimation through the matching across groups induced by our partitions, we additionally develop our prior to induce sparsity in the $\mathbf{T}_m$ matrices. Under multivariate normality $\varphi_{mjt} = 0$ implies Y_{mij} and Y_{mit} are independent given $\{Y_{mik} : k < t, k \neq j\}$. A variety of priors have previously been used that introduce a binary latent variable to determine whether a particular GARP is non-zero, most notably in Smith and Kohn (2002). As we will see in Section 3, our sampling scheme requires marginalization over the GARPs and IVs, which would require summing over all 2^{t-1} ways to select the non-zero parameters. This will become problematic for even moderate t , so it is more natural for us to consider a shrinkage prior to gain sparsity in $\mathbf{T}_m$.

We propose the following prior for $(\tilde{\boldsymbol{\varphi}}_{it}, \tilde{\gamma}_{it})$, the Cholesky parameters for the i th set of the t th partition, conditional on $\mathcal{P}_t$:

$$\delta_{ijt} | \mathcal{P}_t \sim \text{Exp} \left(\frac{1}{2} \xi_0^2 (t-j)^2 \right) \quad (j=1, \dots, t-1), \quad (5)$$

$$\tilde{\gamma}_{it} | \mathcal{P}_t \sim \text{InvGamma}(\lambda_1, \lambda_2), \quad (6)$$

$$\tilde{\boldsymbol{\varphi}}_{it} | \tilde{\gamma}_{it}, \boldsymbol{\Delta}_{it}, \mathcal{P}_t \sim \mathbf{N}_{t-1}(\mathbf{0}_{t-1}, \tilde{\gamma}_{it} \boldsymbol{\Delta}_{it}), \quad (7)$$

where $\boldsymbol{\Delta}_{it}$ is the diagonal matrix with elements $\delta_{11t}, \dots, \delta_{i,t-1,t}$ and the inverse Gamma distribution has kernel $x^{-\lambda_1-1} e^{-\lambda_2/x}$. In (5) we draw the shrinkage factor δ_{ijt} which rescales the variance of $\tilde{\varphi}_{ijt}$. Note that as the lag $t-j$ increases, the shrinkage factor δ_{ijt} tends to be smaller, giving more aggressive shrinkage of $\tilde{\varphi}_{ijt}$ due to its smaller a priori variance $\gamma_{it} \varphi_{ijt}$. For longitudinal data one generally expects only the more recent measurements to be relevant in the distribution of Y_t . This common intuition is captured in the choice (5) by allowing the rate parameter of this distribution to increase in lag. The parameter ξ_0^2 controls the overall level of shrinkage. For conditional conjugacy we use a normal prior for $\tilde{\boldsymbol{\varphi}}_{it}$ and the inverse gamma distribution for the innovation variances. When $t = 1$, there are no GARPs and no shrinkage factors, so only the prior in (6) is needed.

Lines (5) and (7) are in the spirit of the Bayesian Lasso (Park and Casella, 2008). Marginally over δ_{ijt} and conditionally on γ_{it} , the standardized regression coefficients $\tilde{\phi}_{ijt}/\sqrt{\tilde{\gamma}_{it}}$ have independent Laplace (double exponential) priors with location and rate parameters of 0 and $\xi_0(t-j)$, respectively. For a single group with no partitioning, our prior can be considered a Bayesian analog of the penalized likelihood approach of Huang et al. (2006) (also, Yuan and Lin, 2007). The posterior mode from our model would correspond to the penalized

likelihood solution using the weighted penalty $\xi_0 \sum_{t=2}^T \sum_{j=1}^{t-1} w_{jt} |\tilde{\phi}_{jt}|$ for the GARPs with weights $w_{jt} = (t-j)/\sqrt{\tilde{\gamma}_{jt}}$, so that higher lag terms are more strongly penalized. In addition to the Bayesian Lasso, there are numerous alternative shrinkage priors. We discuss some of these further in Section 6 and note some of the challenges they may pose in computations.

For a fully Bayesian analysis we need priors for the innovation variance hyperparameters λ_1 , λ_2 , and the global shrinkage rate ξ_0^2 . In the absence of relevant prior knowledge about these terms, we specify independent Gamma(1,1) priors. Simulations have demonstrated that these choices are relatively uninformative in terms of their effect on the estimated GARPs and IVs.

3. Sampling algorithm

Inference using the covariance partition priors requires a posterior sample from a Markov chain Monte Carlo (MCMC) algorithm. In this section we describe a Gibbs sampler used to obtain such a sample. The main challenge in devising the algorithm is dealing with the partition structure. First, we describe an sampler that updates the partitions one at a time and then extend to a sampling scheme that jointly updates a small neighborhood of partitions within the partition process.

Let $H = (q, \xi_0^2, \lambda_1, \lambda_2)$ denote the set of model hyperparameters, $C_t = \{(\tilde{\phi}_{it}, \tilde{\gamma}_{it})\}_{i=1}^{d_t}$ the set of Cholesky parameters for the t th sequential distributions of all d_t sets (equivalently, all M groups), and $\Delta_t = \{(\delta_{i1t}, \dots, \delta_{i,t-1,t})\}_{i=1}^{d_t}$ the set of shrinkage factors in the t th sequential distributions. Further, $C_{(-t)}$ gives the set of Cholesky parameters $C_1, \dots, C_T$ excluding C_t and $C_{j:k} = \{C_t\}_{t=j}^k$ is the sets of the Cholesky parameters for the j th through the k th sequential distributions ($j < k$). $\mathcal{P}_{(-t)}$, $\mathcal{P}_{j:k}$, $\mathcal{P}_{(-t)}$, and $\mathcal{P}_{j:k}$ are defined similarly. Finally, $\mathcal{Y}$ denotes the data.

3.1. Updating a single partition

We first discuss moves that update the parameters associated with the t th sequential distribution: $\mathcal{P}_t$, C_t , and Δ_t . We can factor this conditional distribution as

$$[\mathcal{P}_t, C_t, \Delta_t | \mathcal{P}_{(-t)}, C_{(-t)}, \Delta_{(-t)}, H, \mathcal{Y}] = [\mathcal{P}_t, \Delta_t | \mathcal{P}_{(-t)}, C_{(-t)}, \Delta_{(-t)}, H, \mathcal{Y}] [C_t | \mathcal{P}_{1:T}, C_{(-t)}, \Delta_{2:T}, H, \mathcal{Y}].$$

Due to conditional independencies, these terms can be further simplified to $[\mathcal{P}_t | \mathcal{P}_{t-1}, \mathcal{P}_{t+1}, H, \mathcal{Y}]$ and $[C_t | \mathcal{P}_t, H, \mathcal{Y}]$.

For the $[C_t | \varphi_{t-1}, H, \mathcal{Y}]$ step we can express this distribution as

$\prod_{i=1}^{d_t} [(\tilde{\phi}_{it}, \tilde{\gamma}_{it}) | S_{it}, \Delta_{it}, H, \mathcal{Y}]$ and sample each $(\tilde{\phi}_{it}, \tilde{\gamma}_{it})$ ($i = 1, \dots, d_t$) exactly from its conditional distribution. The sampling distributions are

$$\tilde{\gamma}_{it} | \Delta_{it}, S_{it}, H, \mathcal{Y} \sim \text{InvGamma} \left(\lambda_1 + \frac{n}{2}, \lambda_2 + \frac{1}{2} \mathbf{y}^\top \left\{ \mathbf{I}_n - \mathbf{X}(\Delta_{it}^{-1} + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right\} \mathbf{y} \right), \quad (8)$$

$$\tilde{\phi}_{it} | \tilde{\gamma}_{it}, \Delta_{it}, S_{it}, H, \mathcal{Y} \sim \mathbf{N}_{t-1} \left((\Delta_{it}^{-1} + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \tilde{\gamma}_{it} (\Delta_{it}^{-1} + \mathbf{X}^\top \mathbf{X})^{-1} \right), \quad (9)$$

where $n = \sum_{m \in S_{it}} n_m$, $\mathbf{y}$ is the n -vector containing Y_{mit} ($i = 1, \dots, n_m; m \in S_{it}$), and $\mathbf{X}$ is the $n \times (t-1)$ matrix with row $(Y_{m1}, \dots, Y_{m,t-1})^\top$. If $t = 1$, $\mathbf{X}$ is empty, and the InvGamma scale parameter in (8) is $\lambda_2 + \frac{1}{2} \mathbf{y}^\top \mathbf{y}$. As there are no GARPs in the $t = 1$ distribution, we ignore (9).

For the $[\varphi_t | \varphi_{t-1}, \varphi_{t+1}, H, \mathcal{Y}]$ step, let φ_t denote the current value of the partition, and we develop a MH step that attempts to update φ_t with a candidate partition $\mathcal{P}_t^*$. The most common ways to propose $\mathcal{P}_t^*$ is to use either a biased random walk (BRW; Crowley, 1997) or a split-merge move (SM; Richardson and Green, 1997; Jain and Neal, 2004). The BRW is constructed by considering a graph structure on Ω where φ and φ' are connected if they differ only in the location of a single group; the candidate φ^* is drawn uniformly from the partitions that have an edge with φ . The SM move either splits a set of φ_t into two or merges two sets into one. Following the advice of Booth et al. (2008), we utilize a mixture of these two moves. With probability π_{BRW} we will perform a BRW step and otherwise, perform a SM step. Since each of the steps is invariant for $[\varphi_{t-1} | \varphi_{t-1}, \varphi_{t+1}, H, \mathcal{Y}]$ (and consequently, for the full posterior $[\varphi_{1:T}, C_{1:T-2:T}, H | \mathcal{Y}]$), the mixture of the steps is as well (Tierney, 1994). Further, the mixture of these two steps is better able to explore the posterior than the samplers that use only the BRW step or only the SM move. We provide additional details of the BRW and SM steps in the Supplementary Materials, as well as the formulas for the candidates probabilities, $\text{pr}(\mathcal{P}_t \rightarrow \mathcal{P}_t^* | \text{BRW})$ and $\text{pr}(\mathcal{P}_t \rightarrow \mathcal{P}_t^* | \text{SM})$.

In addition to proposing a new partition $\mathcal{P}_t^*$, we also must propose candidate values for the shrinkage factors Δ_t^* . One common method would be to use a reversible jump move (Green, 1995), but it turns out that an independent proposal from the candidate density $g(\cdot | \mathcal{P}_t, \mathcal{P}_t^*)$ is more effective for our model. This candidate density $g(\cdot | \mathcal{P}_t, \mathcal{P}_t^*)$ corresponds to conjugately sampling from the distribution of δ_{ijt} given ϕ_j and γ_j , where ϕ_j and γ_j come from a pooled estimator for the GARPs and IV. See the Supplementary Materials for the exact specification of this density.

Computing the probability of accepting the move from (φ_{t-1}) to $(\mathcal{P}_t^*, \Delta_t^*)$ requires the likelihood implied by $[\varphi_{t-1} | \varphi_{t-1}, \varphi_{t+1}, H, \mathcal{Y}]$ and $[\mathcal{P}_t^*, \Delta_t^* | \varphi_{t-1}, \varphi_{t+1}, H, \mathcal{Y}]$, which is given by

$$[\mathcal{P}_t, \Delta_t | \mathcal{P}_{t-1}, \mathcal{P}_{t+1}, H, \mathcal{Y}] \propto \text{pr}(\mathcal{P}_t | \mathcal{P}_{t-1}) \text{pr}(\mathcal{P}_{t+1} | \mathcal{P}_t) \prod_{i=1}^{d_t} \mathcal{L}(Y_{mi}, m \in S_{it} | \Delta_{it}) \pi(\Delta_{it}),$$

where $\pi(\Delta_{it})$ is the prior distribution of the shrinkage factors (5) and $\mathcal{L}(Y_{mi}, m \in S_{it} | \Delta_{it})$ is the likelihood of the data from groups in set S_{it} conditional on the shrinkage after marginalizing out the GARPs and IVs, i.e.,

$$\mathcal{L}(Y_{mi}, m \in S_{it} | \Delta_{it}) = \int_{\mathbb{R}^{t-1}} \int_{\mathbb{R}_+} \left\{ \prod_{m \in S_{it}} \prod_{i=1}^{n_m} f(Y_{miT} | Y_{mi1}, \dots, Y_{mi,t-1}; \phi, \gamma) \right\} \pi(\phi | \Delta_{it}, \gamma) \pi(\gamma) d\phi d\gamma.$$

This integral can be computed in closed form and is proportional to a multivariate t -density; the exact form is contained in the online Supplementary Materials. It is important to note that the integral $\int \mathcal{L}(Y_{mi}, m \in S_{it} | \Delta_{it}) \pi(\Delta_{it}) d\Delta_{it}$ is not available in closed form, which is why we update Δ_{it} jointly with ϕ_t . After combining all the pieces in the usual way (see Supplementary Materials), we have the Metropolis-Hasting acceptance probability for the move from (ϕ_t, Δ_t) to (ϕ_t^*, Δ_t^*) .

3.2. Updating a neighborhood of partitions

The algorithm just described updates a single partition at a time. However, due to the Markovian structure of the partition process, it may be more effective to also include moves allowing us to occasionally update the partitions in a block. For instance, if ϕ_t is equal to or close to its neighbors, ϕ_{t-1} and ϕ_{t+1} , it may be difficult to accept a new value ϕ_t^* for larger values of q . But by attempting to update all three jointly, we may better traverse the high probability regions of the posterior.

To that end, we devise a neighborhood updating step to jointly update the t th through the $(t+k)$ th partitions. First, draw k from $\{1, \dots, T-1\}$ according to $\text{pr}(k) \propto 2^{-k}$, favoring smaller values of k . Conditionally on k , draw T uniformly from $\{1, \dots, T-k\}$. The neighborhood sampling step seeks to jointly update the $k+1$ partitions $\phi_t, \dots, \phi_{t+k}$ and the associated shrinkage factors $\Delta_t, \dots, \Delta_{t+k}$. Using similar conditioning arguments from Section 3.1, we have

$$[\mathcal{P}_{t:(t+k)}, C_{t:(t+k)}, \Delta_{t:(t+k)} | \mathcal{P}_{(-t:(t+k))}, C_{(-t:(t+k))}, \Delta_{(-t:(t+k))}, H, \mathcal{Y}] = [\mathcal{P}_{t:(t+k)}, \Delta_{t:(t+k)} | \mathcal{P}_{t-1}, \mathcal{P}_{t+k+1}, H, \mathcal{Y}] \prod_{j=t}^{t+k} [C_j | \mathcal{P}_j, \Delta_j, H, \mathcal{Y}].$$

The $[C_j | \mathcal{P}_j, \Delta_j, H, \mathcal{Y}]$ steps are as in Section 3.1, and we develop a sampling scheme for the first distribution on the right-hand side.

The neighborhood update consists of two parts: a joint move and an independence move. With probability π_j , we attempt a joint update and uniformly select j in $\{t, \dots, t+k\}$. Using

a BRW move, we propose an update $\mathcal{P}_j^*$ to $\mathcal{P}_j$, and then let $\mathcal{P}_j^*$ also be the candidate for all other partitions in the neighborhood. So $\mathcal{P}_i^* = \mathcal{P}_j^*$ for all $i = t, \dots, t+k$. The probability of proposing such a candidate neighborhood is

$(k+1)^{-1} I(\mathcal{P}_t^* = \dots = \mathcal{P}_{t+k}^*) \sum_{i=t}^{t+k} \text{pr}(\mathcal{P}_i \rightarrow \mathcal{P}_i^* | \text{BRW})$, where $\text{pr}(\mathcal{P}_i \rightarrow \mathcal{P}_i^* | \text{BRW})$ is the probability of moving from $\mathcal{P}_i$ to $\mathcal{P}_i^*$ through the biased random walk. Since $\mathcal{P}_j^*$ is chosen from a biased random walk from $\mathcal{P}_j$, it will frequently be the case that $\mathcal{P}_i$ cannot be reached through a single step of the biased random walk from $\mathcal{P}_i^* = \mathcal{P}_j^*$ ($i \neq j$). Hence, many of these summands may be zero. Further, candidates from the joint move are concentrated on the subspace of Ω^{k+1} where all partitions in the sequence are equal, so the reverse moves will have probability zero unless $\mathcal{P}_t = \dots = \mathcal{P}_{t+k}$.

With probability $1 - \pi_j$, the neighborhood update will propose the new partitions through an independence move. For each $j = t, \dots, t+k$, we independently draw the candidate $\mathcal{P}_j^*$ uniformly from Ω . Clearly, the probability of proposing this transition is $B_M^{-(k+1)}$. In the single partition sampler we calculated the proposal probabilities conditionally on whether we performed a BRW or a SM merge, but here we marginalize over our choice of a joint or independence move. Hence,

$$\text{pr}(\mathcal{P}_{t:(t+k)} \rightarrow \mathcal{P}_{t:(t+k)}^*) = (1 - \pi_j) B_M^{-(k+1)} + \frac{\pi_j}{k+1} I(\mathcal{P}_t^* = \dots = \mathcal{P}_{t+k}^*) \sum_{j=t}^{t+k} \text{pr}(\mathcal{P}_j \rightarrow \mathcal{P}_j^* | \text{BRW}).$$

By considering the probability marginalized over the move type, we always have a non-null probability associated with each transition since both the forward and backward moves are bounded from below by $(1 - \pi_j) B_M^{-(k+1)}$. With the candidate partitions $\{\mathcal{P}_t^*, \dots, \mathcal{P}_{t+k}^*\}$, we propose the candidate shrinkage factors $\Delta_t^*, \dots, \Delta_{t+k}^*$ as in the single partition updater and calculate the MH acceptance probability (Supplementary Materials).

While our main interest is in the independence step as a way to obtain non-null reverse probabilities, accepting an independence step introduces a regeneration in the $(\mathcal{P}_{t:(t+k)})$ -marginal Markov chain (Mykland et al., 1995). Because these independence moves are accepted with low frequency (generally, less than 1%), it would not be appropriate to base inference on the tours of the Markov chain, but allowing these independent moves further suggests that our sampling algorithm will better explore the support of the partition process. While we do not explore this here, it is also possible to add an independence step into the single partition updater to encourage regeneration there. We finally note that as M becomes large, it is usually infeasible to sample independently from Ω ; however, we are generally interested in cases where M is of a size where independence sampling will be straightforward.

3.3. Full sampling algorithm

We have discussed two methods to update the partitions $\mathcal{P}_t$ along with the shrinkage factors ρ_t . Because the neighborhood step tends to be accepted infrequently, our algorithm

incorporates the moves from both Sections 3.1 and 3.2 by performing the neighborhood move K times and then updating each $\wp_t$ through a single partition move. As k is a truncated Geometric(1/2) random variable, its expected value is $2[1 - (T-1)/(2^T-2)] \approx 2$. Hence, each neighborhood step attempts to update $E(k) + 1 \approx 3$ partitions on average, and so K is chosen to be approximately $T/3$ so that moves are proposed for $\wp_t$ under the neighborhood update with similar frequency to the single partition move. Next, the hyperparameters ξ_0^2 and λ_2 are updated conjugately, λ_1 through slice sampling (Neal, 2003), and q through a MH step. The appropriate sampling densities may be found in the Supplementary Materials.

Because we sample the shrinkage factors jointly with partitions, we only get new values of the δ_s when we accept the MH move for the partitions. Hence, our shrinkage factors (and consequently, the GARPs) may be slow mixing. We choose to add an additional step to our sampler to update the shrinkage factors from their sampling distributions conditional on the GARPs and IVs. This involves sampling $1/\delta_{ijt}$ from $\text{InvGaussian}(\tilde{\mu}_{ijt}, \xi_0^2(t-j)^2)$, where $\tilde{\mu}_{ijt} = (\gamma_{it})^{1/2} \xi_0(t-j)/\phi_{ijt}$ ($t = 2, \dots, T; i = 1, \dots, d; j = 1, \dots, t-1$) (Park and Casella, 2008). Empirical results indicate that the entire chain mixes faster by including this step.

In summary, our fixed scan Gibbs sampler iterates between the following steps:

1. Perform K sweeps of the neighborhood sampler with $\pi_J = 1/2$, and after each sweep update the Cholesky parameters $C_{t:(t+K)}$.
2. For each $t = 1, \dots, T$, perform a single partition update move on $(\wp_t, \rho)$ with $\pi_{\text{BRW}} = 1/2$ and update the Cholesky parameters C_t .
3. Conjugately update the full set of shrinkage factors $\delta_{2:T}$ given $C_{2:T}$.
4. Update the hyperparameters $q, \xi_0^2, \lambda_1, \lambda_2$.
5. Conditional on $\Sigma_1, \dots, \Sigma_M$, update other all model parameters such as regression parameters and missing data.

4. Simulation studies

We examine the performance of our priors using two simulation studies. For the first we simulate data consisting of $M = 8$ groups with sample sizes of $n_1 = \dots = n_5 = 60$, $n_6 = \dots = n_8 = 30$. Because there is less data for the final three groups, we expect leveraging information across groups will improve estimation, especially for the smaller groups. Observations are measured at $T = 10$ time points. The true partitions chosen are

$\mathcal{P}_1^{\text{true}} = \{\mathcal{M}\}$, $\mathcal{P}_{2:5}^{\text{true}} = \{\{1, 2, 3, 4\}, \{5, 6, 7, 8\}\}$, $\mathcal{P}_{6:8}^{\text{true}} = \{\{1, 2, 3, 4\}, \{5, 6\}, \{7, 8\}\}$, and $\mathcal{P}_{9:10}^{\text{true}} = \{\{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 8\}\}$. This choice of $\mathcal{P}_{1:T}^{\text{true}}$ is clearly consistent with our assumption that the partitions change smoothly in time. Given the partitions, the true values of $\mathbf{T}_m$ and $\mathbf{G}_m$ are specified to represent a longitudinal data scenario with serial correlation. The true covariance matrices have a sparse structure with between one and three non-zero GARPs. Further details are provided in the Supplementary Materials.

Under this data distribution we generate 100 data sets and run an MCMC chain for each using the algorithm of Section 3.3. After a burn-in of 3000 iterations, the chain runs for

12,000 iterations, and we retain every fifth for posterior inference. The sampler performs $K = 5$ sweeps of the neighborhood move during each iteration. To measure the performance of the resulting estimators, we use the loss functions

$$\begin{aligned} L_{\text{all}}(\Sigma, \hat{\Sigma}) &= L_{\text{all}}(\{\Sigma_1, \dots, \Sigma_M\}, \{\hat{\Sigma}_1, \dots, \hat{\Sigma}_M\}) = \sum_{m=1}^M \frac{n_m}{N} L(\Sigma_m, \hat{\Sigma}_m), \\ L(\Sigma_m, \hat{\Sigma}_m) &= \text{tr}(\Sigma_m^{-1} \hat{\Sigma}_m) - \log |\Sigma_m^{-1} \hat{\Sigma}_m| - T, \end{aligned}$$

where $N = \sum_m n_m$ and $\hat{\Sigma}_m = [\mathbb{E}(\mathbf{T}_m \mathbf{G}_m \mathbf{T}_m^\top | \mathcal{D})]^{-1}$ is the Bayes estimator for group m under $L(\Sigma_m, \hat{\Sigma}_m)$, the standard log-likelihood loss function for a single covariance matrix (Yang and Berger, 1994). $L_{\text{all}}(\Sigma, \hat{\Sigma})$ measures the loss to estimating the whole collection of covariance matrices by taking a weighted average with weights proportional to group size. The estimated risk is the average loss over the 100 data sets.

To compare our new methodology with commonly used methods, we consider ten prior structures: the distribution on the partitions given by (4) with the prior for q uniform on $\mathcal{Q}$, called the ‘covariance partition prior’; the partitions are random but $q = 0$ is fixed, i.e., the ‘independent-uniforms prior’; $\wp_t$ is fixed as $\wp_{\text{ind}}$ for all t (all groups are treated independently); $\wp_t$ is fixed as $\wp_{\text{pool}}$ for all t (data is pooled into a single group); and an ‘oracle’ that uses $\wp_t = \wp_t^{\text{true}}$ for each t . With each of these five choices for the partitions, we consider two priors on the Cholesky parameters. Shrinkage is the sparse prior discussed in Section 2.2. We also consider a non-shrinkage prior defined by equations (6) and (7) where $\mathbf{I}_t = \sigma^2 \mathbf{I}_{t-1}$. The parameter σ^2 is the prior variance of the standardized GARPs and is given an InvGamma(0.1, 0.1) prior. As a final competitor, we include a Bayesian version of the model from (Pourahmadi et al., 2007) that assumes a common $\mathbf{T}_m$ across groups while allowing the IVs $\mathbf{G}_m$ to be group-specific; we denote this by “PDP-2007.” The middle columns of Table 1 contain the estimated risk of estimating the collection of covariance matrices and the risk for estimating the final group under each of these eleven prior specifications.

As expected, the oracle with shrinkage produces the lowest estimated risk. But as the oracle estimator is not available in real situations (we never know the true partitions $\wp_t^{\text{true}}$), we use this only as a benchmark for comparison. Of those remaining, methods that adaptively search for an appropriate partition structure (covariance partition and independent-uniforms) have lower risk, and for each choice of partition prior, GARP shrinkage beats non-shrinkage. Neither $\wp_{\text{pool}}$ prior works well because they produce inconsistent estimators. Similarly, PDP-2007 performs poorly as the assumption of common $\mathbf{T}_m$ is violated. Further, the $\wp_{\text{ind}}$ prior without shrinkage has too many parameters to estimate efficiently; adding shrinkage alleviates this to somewhat. The improvement due to the covariance partition prior with shrinkage is particularly evident for Σ_8 , demonstrating that we are able to leverage information across the data to improve estimation in the smaller groups.

Next, we consider a simulation study where the dimension of the outcome is increased to $T = 50$. The true partitions change at every 5th time point (t divisible by 5) and cycles between

the four partitions $\{\{1, 2, 3, 4\}, \{5, 6, 7, 8\}\}$, $\{\{1, 2\}, \{3, 4\}, \{5, 6, 7, 8\}\}$, $\{\{1, 2, 3, 4\}, \{5, 6, 7, 8\}\}$, and $\{\{1, 2, 3, 4\}, \{5, 6\}, \{7, 8\}\}$. The Cholesky parameters are drawn from the sparse model presented in Table A.1 of the Supplementary Materials. We run an MCMC chain for each of 50 generated data sets with our eleven prior choices for 5000 iterations after 1000 burn-in and retain every 5th iteration. With the larger dimension we now perform $K = 15$ sweeps of the neighborhood step so that updates are proposed for φ_t under this step with a similar frequency to the single updates. Risk estimates are available on right side of Table 1.

With the larger dimension we continue to see strong performance of our methodology. The covariance partition prior and independent-uniform priors both dominate the default choices, showing risk reductions greater than 50%. With shrinkage the independent-uniform has 5% larger risk than the choice allowing q to be random. The partition prior with shrinkage has 14% larger risk than the oracle, compared to 140% larger for independence and over 18,000% larger for group-pooling. The conclusions for estimating the smaller group 8 are more extreme but qualitatively similar.

In addition to these two simulations, we perform others under varying choices of the true data distribution. The covariance partition prior continues to exhibit strong risk performance under situations where the groups are unrelated ($\varphi_t = \varphi_{\text{ind}}$ for all t), where there is only a single group ($\varphi_t = \varphi_{\text{pool}}$ for all t), and where the partition φ_t varies sharply from the previous partition φ_{t-1} . The interested reader can find these additional results in the Supplementary Materials, along with boxplots comparing the losses incurred in these two simulation studies.

5. Depression data example

We now apply our methodology to data from a series of studies on the efficacy of pharmaceutical depression treatment. This data, originating from Thase et al. (1997), has also been analyzed in Pourahmadi and Daniels (2002) and Gaskins and Daniels (2013). Weekly measurements of depression scores are recorded over a period of 17 weeks. The goal of the study was to compare symptom improvement under pharmacotherapy and psychotherapy treatments versus psychotherapy-only treatments. Previous analyses have shown that the baseline depression level strongly influences the rate and variability of improvement; hence, the data have previously been considered to fall into four groups defined by crossing the two treatments with a binary indicator of high or low initial severity. Here we choose to further stratify each of these four groups by gender for a total of $M = 8$ groups. Each group is assumed to have a unique, quadratic mean function, and the group-specific regression coefficients are given independent flat priors.

The sample sizes of the first seven groups range from 28 to 73, with group 8 (drug treatment, high baseline severity, female) much larger at 193 patients. Around 30% of measurements are missing, ranging from 20 to 56% by group. Much of the missingness is structural since patients were not always scheduled to be seen each week; consequently, missingness is assumed to be ignorable (Daniels and Hogan, 2008). To assist in computation, we include data augmentation in Step 5 into the MCMC algorithm.

5.1. Computational details

As in the risk simulation, we run an MCMC chain for both the shrinkage and non-shrinkage prior with each of the four choices of partition structures: covariance partition prior, independent-uniforms prior, group independence, and group pooling, along with the PDP-2007 model. We do not know the true partition structure, so there is no oracle estimator. For each prior choice, a single chain is run for 25,000 iterations after a burn-in of 5000, and the values from every fifth iteration are retained for posterior inference. The covariance partition and independent-uniforms prior use $K = 5$ neighborhood moves per iteration and took roughly 9 hours in R on a desktop PC.

First, we consider the effect of the choice of sampling algorithm in the fitting of this data to the covariance partition prior with shrinkage. As is often the case when sampling across partitions, the MCMC chain only encounters a small subset of partitions in Ω , and it is necessary to make sure that the sampler does not get stuck in local modes. Inspecting the trace plots of the IVs from a sampler that uses only the biased random walk (BRW) step shows such a problem. Likewise, a sampling algorithm that uses the neighborhood step without the single partition moves does not accept the candidate partitions $\mathcal{P}_{t:(t+k)}^*$ often enough to traverse the parameter support. Along with thorough inspection of trace plots, we computed simulation inefficiency factors (SIF), the ratio of the number of iterations (discarding burn-in, without thinning) to the effective sample size (Robert and Casella, 2004, Section 12.3), for each of the GARPs and IVs. Using the full sampling scheme of Section 3.3, the median SIF (over the 1224 covariance parameters) is 4.76 compared 5.41 for the single partition scheme; 60% of parameters have greater efficiency under the full sampler. Only 92 parameters (7.5%) have SIF over 50 (corresponding to an effective sample size less than 500), which quite good considering the multi-modal nature of our partition model. We

also compute the SIF for the observed data log-likelihood, $\sum_{m=1}^M \loglik(\beta_m, \sum_m |\mathcal{Z}_m|)$, as a univariate measure of the overall performance of the sampler. The log-likelihood SIFs are 2.49 for the full sampler, 2.61 for BRW-SM mixture, 3.22 for BRW, and 2.79 for SM.

Considering the full algorithm, we may also look at how well the partitions are mixing. As the partition is not a numeric quantity, this is more difficult as SIFs and trace plots are not available. In Figure 1, we display trace plots for the degree of the partition and the number of group pairs for three representative times $t = 1, 5, 9$. From these, we see how the partition is varying by both degree and the number of pairs, but it is important to note that these plots do not display the full story as the partition may change without changing either the degree or number of pairs. The MH acceptance rates in the second row are lower than desired (especially for the later times), but this is partly a consequence of having a large discrete support. Based on the mixing of the partitions, covariance parameters, and data likelihood, we conclude that the algorithm of Section 3.3 is able to adequately explore the regions of the parameter space with high posterior density.

Additional adaptations to our sampling algorithm could be implemented to further improve mixing such as tempering (Geyer and Thompson, 1995), using a launch state for the SM moves (Jain and Neal, 2004), adding regeneration to the single partition update, among

others. But it is unclear whether the extra computational complexity would outweigh any improvements in mixing.

5.2. Model selection

To determine the model that best fits the data, we use the deviance information criterion (DIC; Spiegelhalter et al., 2002). The DIC is defined as $DIC = Dev + 2p_D$, where Dev is the deviance evaluated at the posterior expectation of the parameters and p_D is a model complexity term. Usually interpreted as the effective number of model parameters, p_D is computed as the difference between the expected deviance and Dev . We use the observed data likelihood to compute the deviances (Wang and Daniels, 2011). Models with smaller DIC are considered to better balance the model fit and complexity. Table 2 contains the model comparison statistics.

Based on the DIC criterion, the covariance partition prior with the shrinkage Cholesky structure is found to be the overall best model for this data. This is followed by the non-shrinkage covariance partition prior and the two independent-uniform models. The greater the number of parameters to estimate (larger p_D) the more pronounced is the effect of the shrinkage prior. With the independence model the shrinkage prior estimates around one hundred fewer parameters than the non-sparse counterpart, but with ϕ_{pool} the shrinkage and non-shrinkage Cholesky structures give practically indistinguishable models. The independent-uniforms prior requires 80 to 100 more parameters than the covariance partition prior. These additional parameters arise due to partitions tending to have a larger degree under the independent-uniforms prior and the additional variability in the partition process from assuming independence across partitions. As in the previous risk studies, the PDP-2007 model is too restrictive, leading to poor estimation. Our proposed method using either the shrinkage or non-shrinkage Cholesky prior and with q either stochastic or fixed at 0 produce markedly better models than the standard, naive methods.

As discussed in Section 1, the choice of covariance model can have an important impact on mean estimation. For instance, the treatment effect, $\mu_{m1} - \mu_{mT}$, for group 6 (drug, low severity, female) is estimated to be 8.7 under the best fitting modeling with a 95% credible interval of (5.9, 11.4). The estimates under ϕ_{pool} and DPD-2007 are 18% and 16% larger, respectively. While these differences are not statistically significant (the credible intervals overlap), they may have practical implications on the conclusions reached from the study. In other groups such as group 7 (no drug, high severity, female), treatment effect estimates differ across models by less than 3%. Across all groups the ϕ_{pool} choice (and to a lesser extent, PDP-2007) produces estimates that are systematically larger than the other covariance models, indicating that failing to consider a sufficiently flexible prior for Σ may lead to overstating the reduction in depression symptoms.

5.3. Estimated covariance structure

We finally examine the structure that the covariance partition prior with Cholesky shrinkage induces for this data. The mean of the partition process smoothness parameter q is 6.91 with 95% credible interval (4.75, 9.75). Figure 2 shows the clustering properties of our prior from baseline through week 11; the remaining, undisplayed weeks behave similarly to weeks 8–

11. Each panel depicts the pairwise probability that m_1 and m_2 share a sequential distribution at time t (week $t - 1$), that is, the probability that there exists a set $S \in \mathcal{P}_t$ with $\{m_1, m_2\} \subseteq S$. At baseline and the next two weeks groups tend to fall into two sets defined by the initial severity indicator with a low severity set $\{1, 2, 5, 6\}$ and a high severity set $\{3, 4, 7, 8\}$. From weeks 3 through 6 there is less stability in the partitions, but from week 7 on, there are two strong clusters: one containing groups 4 (drug, high severity, male), 5 (no drug, low, female), and 6 (drug, low, female) and the other containing groups 2 (drug, low, male), 7 (no drug, high, female), and 8 (drug, high, female). The set memberships of groups 1 and 3 (no drug, low/high, male) are less stable as t varies.

While our model makes use of the partitions primarily as a tool for sharing information about covariance across groups, the matching structure that our model estimates may be clinically meaningful and provide subject-matter experts new intuition about similarities and differences across groups. This is the case for this data; for instance, during the first three weeks, the high severity set $\{3, 4, 7, 8\}$ has innovation variances more than twice as large as the low severity set $\{1, 2, 5, 6\}$. Knowing that improvement is less predictable for high severity patients (across both genders and treatments) during the initial weeks may suggest a need for closer follow-up in these early weeks of treatment. In the later weeks, we are able to similarly interpret the set $\{4, 5, 6\}$ as containing the more stable groups (lower IV) than the groups in the set $\{2, 7, 8\}$, with group 3 being the least stable (highest IV). Again, this may potentially have implications on patient management.

6. Discussion

In this article we have introduced a novel solution to the simultaneous covariance estimation problem for longitudinal data. The proposed methodology involves partitioning the groups so that those in a common set share the values of dependence parameters in $\ell(Y_{mi1}, \dots, Y_{mi,t-1})$. Additionally, we introduce a sparse structure on the Cholesky parameters by adapting a Bayesian Lasso with increasing shrinkage for higher lag GARPs. The covariance partition prior stochastically considers a large model space for the set of covariance matrices while providing a more computationally attractive alternative to Gaskins and Daniels (2013).

In order to efficiently estimate the GARPs, we use the Bayesian Lasso (Park and Casella, 2008), but there are a number of alternative shrinkage priors in the literature. Many methods (including ours) fall into the category of global-local shrinkage (Polson and Scott, 2010), where each parameter has its own shrinkage factor. It is well known that the Bayesian Lasso tends to over-shrink very large parameter values, and recent contributions develop methods to mitigate this (Carvalho et al., 2010; Griffin and Brown, 2010; Bhattacharya et al., 2012; Armagan et al., 2013). But in our case the parameters we are shrinking are standardized regression coefficients with the predictor and response on the same scale, so $|\phi_{mjt}|/\sqrt{\gamma_{mt}}$ will typically be less than one. Hence, over-shrinking GARPs will rarely be a major concern. Further, while some global-local priors have efficient Gibbs steps for both $[C_\ell]$ and $[C_\ell^{-1}]$ (C_ℓ generically denotes the local shrinking factors), our situation requires that we update C_ℓ jointly with the partitions (marginal over C_ℓ), which may be challenging with these priors. Using the normal-gamma model of Griffin and Brown (2010) would be the most accessible

extension; the exponential distribution in line (5) would be replaced by a gamma distribution with rate depending on lag (Gaskins et al., 2014).

In our simulations and data analysis, we have considered situations where the number of groups is fairly small $M = 8$ and the dimension is small-to-moderate (T up to 50). We, therefore, provide a few comments about scaling of our work to larger dimensions. The biggest problem in terms of scaling is the number of groups M . For M larger than 10, it is likely not reasonable to compute and store $a_q(\mathcal{P})$ for all $\mathcal{P} \in \Omega$ (e.g., $B_{10} = 115, 975$). In this case one could use the independent-uniforms prior which has known attractiveness of B_M^{-1} . However, for larger M the exploration of the large support Ω may be quite difficult. The dimension of the response T is less of an issue as the algorithm will scale linearly in T . While the code we have used does not incorporate it, it is also possible to update non-adjacent partitions in parallel.

Throughout this work, we have only considered sharing of information across the dependence structures. This choice was made so as to flexibly model the covariance without restricting the choice of mean structure used. However, if the analyst wishes to also include a clustering of the mean, it can be easily incorporated into our methodology. One choice would be to add the mean μ_{mt} to be grouped along with (ϕ_{mb}, γ_{mb}) in $\mathcal{P}_t$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by NIH grants CA-85295 and CA-183854.

References

- Arabie P, Boorman SA. Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology*. 1973; 10:148–203.
- Armagan A, Dunson DB, Lee J. Generalized double pareto shrinkage. *Statistica Sinica*. 2013; 23(1): 119–143. [PubMed: 24478567]
- Bhattacharya A, Pati D, Pillai NS, Dunson DB. Bayesian shrinkage. 2012 arXiv:1212.6088.
- Boik RJ. Spectral models for covariance matrices. *Biometrika*. 2002; 89(1):159–182.
- Boik RJ. Principal component models for correlation matrices. *Biometrika*. 2003; 90(3):679–701.
- Booth JG, Casella G, Hobert JP. Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society, Series B*. 2008; 70(1):119–139.
- Carvalho CM, Polson NG, Scott J. The horseshoe prior for sparse signals. *Biometrika*. 2010; 97(2): 465–480.
- Chiu TYM, Leonard T, Tsui KW. The matrix-logarithmic covariance model. *Journal of the American Statistical Association*. 1996; 91:198–210.
- Crowley EM. Product partition models for normal means. *Journal of the American Statistical Association*. 1997; 92(437):192–198.
- Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*. 2013 In press.
- Daniels MJ. 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*. 2006; 97(5):1185–1207.

- Daniels, MJ.; Hogan, JW. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Chapman & Hall; 2008.
- Daniels MJ, Pourahmadi M. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*. 2002; 89:553–566.
- Day WHE. The complexity of computing metric distances between partitions. *Mathematical Social Sciences*. 1981; 1(3):269–287.
- Denoeud, L.; Guénoche, A. Comparison of distance indices between partitions. In: Batagelj, V.; Bock, H-H.; Ferligoj, A.; Žiberna, A., editors. *Data Science and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization. Springer; Berlin Heidelberg: 2006. p. 21–28.
- Dunson DB, Xue Y, Carin L. The matrix stick-breaking process: Flexible Bayes meta-analysis. *Journal of the American Statistical Association*. 2008; 103(481):317–327.
- Fox E, Dunson DB. Bayesian nonparametric covariance regression. 2011 arXiv: 1101.2017.
- Gaskins JT, Daniels MJ. A nonparametric prior for simultaneous covariance estimation. *Biometrika*. 2013; 100(1):125–138.
- Gaskins JT, Daniels MJ, Marcus BH. Bayesian methods for non-ignorable dropout in joint models in smoking cessation studies. 2014 Submitted.
- Geyer CJ, Thompson EA. Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*. 1995; 90(431):909–920.
- Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*. 1995; 82(4):711–732.
- Griffin JE, Brown PJ. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*. 2010; 5(1):171–188.
- Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. *Biometrika*. 2011; 98(1):1–15. [PubMed: 23049124]
- Hartigan JA. Partition models. *Communications in Statistics, Part A – Theory and Methods*. 1990; 19(8):2745–2756.
- Hoff PD. A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society, Series B*. 2009; 71(5):971–992.
- Hoff PD, Niu X. A covariance regression model. *Statistica Sinica*. 2012; 22:729–753.
- Huang JZ, Liu N, Pourahmadi M, Liu L. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*. 2006; 93(1):85–98.
- Jain S, Neal RM. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*. 2004; 13(1):158–182.
- Manly BFJ, Rayner JCW. The comparison of sample covariance matrices using likelihood ratio tests. *Biometrika*. 1987; 74:841–847.
- McNicholas PD, Murphy TB. Model-based clustering of longitudinal data. *The Canadian Journal of Statistics*. 2010; 38(1):153–168.
- Mykland P, Tierney L, Yu B. Regeneration in Markov chain samplers. *Journal of the American Statistical Association*. 1995; 90(429):233–241.
- Neal RM. Slice sampling. *The Annals of Statistics*. 2003; 31(3):705–767.
- Park T, Casella G. The Bayesian lasso. *Journal of the American Statistical Association*. 2008; 103(482):681–686.
- Polson, NG.; Scott, J. Shrink globally, act locally: Sparse Bayesian regularization and prediction. In: Bernardo, JM.; Bayarri, MJ.; Berger, JO.; Dawid, AP.; Heckerman, D.; Smith, AFM.; West, M., editors. *Bayesian Statistics*. Vol. 9. 2010. p. 501–538.
- Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*. 1999; 86(3):677–690.
- Pourahmadi M. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*. 2000; 87(2):425–435.
- Pourahmadi M, Daniels MJ. Dynamic conditionally linear mixed models for longitudinal data. *Biometrics*. 2002; 58(1):225–231. [PubMed: 11890319]
- Pourahmadi M, Daniels MJ, Park T. Simultaneous modelling of the Cholesky decomposition of several covariance matrices. *Journal of Multivariate Analysis*. 2007; 98(3):568–587.

- Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*. 1997; 59(4):731–792.
- Robert, CP.; Casella, G. *Monte Carlo Statistical Methods*. 2. Springer-Verlag; 2004.
- Smith M, Kohn R. Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*. 2002; 97(460):1141–1153.
- Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*. 2002; 64(4):583–639.
- Stanley, RP. *Enumerate Combinatorics*. Vol. 1. Canbridge University Press; 1997.
- Thase M, Greenhouse J, Frank E, Reynolds C, Pilkonis P, Hurley K, Grochocinski V, Kupfer D. Treatment of major depression with psychotherapy or psychotherapy-pharmacotherapy combinations. *Archives of General Psychiatry*. 1997; 54:1009–1015. [PubMed: 9366657]
- Tierney L. Markov chains for exploring posterior distributions. *The Annals of Statistics*. 1994; 22(4): 1701–1728.
- Wang C, Daniels MJ. 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*. 2011; 67(3):810–818. [PubMed: 21361893]
- Yang R, Berger JO. Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*. 1994; 22:1195–1211.
- Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*. 2007; 94(1):19–35.

Figure 1.

Each column contains trace plots of a function of the partition at times $t = 1, 5, 9$ for the first 10,000 iterations after burn-in. The first row displays the degree of the partition, and the second the number of pairs of groups in the partition.

Figure 2.

The posterior probability groups m_1 and m_2 share a sequential distribution at week t for baseline through week 11 ($t = 1, \dots, 12$). The size of the box is proportional to $\text{pr}(\exists S \in \mathcal{P}_t : \{m_1, m_2\} \subseteq S)$, and the boxes on the main diagonal have area one.

Table 1

Risk estimates from simulation studies of Section 4.

Partition Prior	Cholesky Prior	T = 10 simulation			T = 50 simulation		
		Risk under $L_{all}(\Sigma, \hat{\Sigma})$	Risk under $L(\Sigma_8, \hat{\Sigma}_8)$	Risk under $L_{all}(\Sigma, \hat{\Sigma})$	Risk under $L(\Sigma_8, \hat{\Sigma}_8)$	Risk under $L_{all}(\Sigma, \hat{\Sigma})$	Risk under $L(\Sigma_8, \hat{\Sigma}_8)$
oracle	shrinkage	0.247	0.310	4.54	6.08		
cov. partition prior	shrinkage	0.326	0.432	5.18	7.49		
indep.-uniforms prior	shrinkage	0.332	0.435	5.45	7.97		
oracle	non-shrinkage	0.371	0.463	6.66	8.73		
cov. partition prior	non-shrinkage	0.410	0.507	6.70	8.82		
indep.-uniforms prior	non-shrinkage	0.414	0.519	6.78	8.97		
$\wp_I = \wp_{ind}$	shrinkage	0.504	0.726	10.88	15.2		
$\wp_I = \wp_{ind}$	non-shrinkage	0.772	1.090	14.89	20.0		
	PDP-2007	0.823	1.207	631	1086		
$\wp_I = \wp_{pool}$	shrinkage	0.806	1.194	860	1100		
$\wp_I = \wp_{pool}$	non-shrinkage	0.829	1.195	850	1100		

Table 2

Model selection statistics for depression study.

Partition Prior	Cholesky Prior	<i>Dev</i>	<i>p_D</i>	DIC
cov. partition prior	shrinkage	39,020	513	40,046
cov. partition prior	non-shrinkage	38,987	544	40,074
indep.-uniforms prior	shrinkage	38,934	595	40,123
indep.-uniforms prior	non-shrinkage	38,884	643	40,170
$\wp_t = \wp_{\text{ind}}$	shrinkage	38,658	814	40,286
$\wp_t = \wp_{\text{pool}}$	non-shrinkage	39,879	213	40,306
$\wp_t = \wp_{\text{pool}}$	shrinkage	39,908	200	40,309
	PDP-2007	39,586	368	40,323
$\wp_t = \wp_{\text{ind}}$	non-shrinkage	38,549	912	40,374