Bayesian modeling of several covariance matrices and some results on propriety of the posterior for linear regression with correlated and/or heterogeneous errors

Michael J. Daniels
Department of Statistics
University of Florida
Gainesville, FL 32611-8545
mdaniels@stat.ufl.edu

Summary

We explore simultaneous modeling of several covariance matrices across groups using the spectral (eigenvalue) decomposition and modified Cholesky decomposition. We introduce several models for covariance matrices under different assumptions about the mean structure. We consider 'dependence' matrices, which tend to have many parameters, as constant across groups and/or parsimoniously modeled via a regression formulation. For 'variances', we consider both unrestricted across groups and more parsimoniously modeled via log-linear models. In all these models, we explore the propriety of the posterior when improper priors are used on the mean and 'variance' parameters (and in some cases, on components of the 'dependence' matrices). The models examined include several common Bayesian regression models, whose propriety has not been previously explored, as special cases. We propose a simple approach to weaken the assumption of constant dependence matrices in an automated fashion and describe how to compute Bayes Factors to test the hypothesis of constant 'dependence' across groups. The models are applied to data from two longitudinal clinical studies.

Key Words: Cholesky decomposition; Spectral decomposition; Variance-Correlation decomposition; Markov Chain Monte Carlo; Bayes Factor; Improper priors

1 Introduction

Consider the setting of modeling multivariate responses across several groups. Let Y_{ij} be a $p-$dimensional response vector for each of $j = 1, \ldots, n_i$ individuals in $i = 1, \ldots, C$ groups. Now, consider the following normal model,

$$Y_{ij} \sim N(\mu_{ij}, \Sigma_i), \tag{1}$$

where $\mu_{ij} = \mu_i$ or $\mu_{ij} = X_{ij}\alpha$ with X_{ij} a $p \times q$ covariate design matrix and α a $q \times 1$ vector of regression parameters. Pourahmadi, Daniels, and Park [26] discuss modeling covariance matrices across groups using several parameterizations of the matrix, Σ_i and likelihood methods. They also detail many settings where such models are useful including model-based clustering [2], analysis of financial data [5], quality control [18], and longitudinal clinical trials (studies). We will illustrate the latter application in this paper.

Simultaneous modeling of covariance matrices across groups has been addressed by numerous authors. Flury considers a spectral decomposition [14, 15, 16] and allows 'commonality' of the eigenvectors (and variations) across groups, while the eigenvalues are allowed to differ. Boik [4] generalized some of this work allowing finer models for the eigenvectors and structured models for the eigenvalues. Manly and Rayner [21], using a variance/correlation decomposition of the covariance matrix, develop a hierarchy of models for covariance matrices across groups, including proportional covariance matrices and a common correlation matrix across the groups. Barnard, McCulloch, and Meng [3] generalize Manly and Rayner's approach by modeling the variance using log-linear regressions and hierarchical priors in a Bayesian setting.

In this paper, we will examine Bayesian formulations of some of the models in Pourahmadi et al. [26] for the spectral decomposition and the generalized autoregressive parameter/innovation variance (GARP/IV) decomposition (also, sometimes called Modified Cholesky) which have not been given a full examination. In addition, we consider several extensions of these models which can be formulated and dealt with easily in the Bayesian framework.

We will quickly review some features of these two parameterizations/decompositions. The spectral decomposition of a matrix Σ_c is given by $\Sigma_c = P_c \Lambda_c P_c'$, where P_c is an orthogonal

matrix of eigenvectors and Λ_c is a diagonal matrix of eigenvalues. In the case of distinct eigenvalues, P_c can be expressed as $P_c = G_{c,12}G_{c,13}G_{c,23}\cdots G_{c,p-1,p}$ where $G_{c,ij}$ is a $p \times p$ matrix with $cos(\theta_{c,tl})$ in the tth and lth diagonal elements, $\pm sin(\theta_{c,tl})$ in the (t,l)th and (l,t)th elements, zeros on the rest of the off-diagonal elements and 1's on the rest of the diagonal. These $p(p-1)/2$ angles, $\theta_{c,tl}$, are called Givens angles [23, 10] and clearly represent rotations in the planes spanned by tth and lth component of the response vector Y. These can be useful for developing parsimonious models for P_c across groups as $\theta_{c,tl}$ are only restricted to lie in the interval $(-\pi/2, \pi/2)$ for uniqueness and positive definiteness of Σ_c. We also consider the GARP/IV decomposition, $\Sigma_c^{-1} = T_c D_c^{-1} T_c'$, where D_c is a diagonal matrix of (innovation) variances (IV) and T_c is a lower triangular matrix with a unit diagonal and regression coefficients (GARP) below the diagonal, $-\phi_{c,tl}$, $l = 1, \ldots, t-1$, $t = 2, \ldots, p$. These parameters are the negative of the regression coefficients in the following conditional means, $E[Y_{ijk}|Y_{ij1}, \ldots, Y_{ijk-1}] = \mu_{ijk} + \sum_{l=1}^{k-1} \phi_{i,kl}(y_{ijl} - \mu_{ijl})$. The innovation variances are given by the conditional variances, $Var[Y_{ijk}|Y_{ij1}, \ldots, Y_{ijk-1}] = \sigma_{ik}^2$.

Under these two decompositions, we will consider several models which allow specific components to be shared across groups and/or modeled with group-specific covariates. Under sensible (improper) priors, we will then examine the propriety of the posterior distribution. The models we examine contain as special cases Bayesian regression models with independent, heterogeneous errors, including having the error variances depend on covariates [31] and regression models with correlated errors. We will show that these models allow the use of improper priors on both regression coefficients and 'variance components' under reasonable conditions.

We remind the reader that in all of the following, we assume the $p \times 1$ vector of responses follow a multivariate normal distribution, as given in (1). In addition, the models based on a GARP/IV decomposition of Σ_c make the most sense when the components of Y_{ij} are ordered (as in longitudinal data) and the Givens angles decomposition of the orthogonal matrix P_c

is only unique if the p eigenvalues are distinct.

The outline of the paper follows. Section 2 will examine the posterior distribution when using improper priors under various assumptions about the mean and covariance matrix across groups for the model given in (1). Sections 3 and 4 propose some extensions of these models by shrinking components towards commonality using hierarchical priors. This avoids, to some extent, searching through a large space of models which correspond to individual components of the covariance matrix being shared or not shared across (a subset of) the groups [26]. In section 5, we present two examples of these models in longitudinal clinical trials/studies. Section 6 proposes some additional extensions for models on the 'variances'. We conclude and discuss open problems in Section 7.

2 Bayesian analogues

Among other models proposed in [26] were models that assumed common principal components, often abbreviated as CPC ($P_c = P$ for $c = 1, \ldots, C$), common GARP ($T_c = T$ or equivalently, $\phi_{c,tj} = \phi_{tj}$ for $c = 1, \ldots, C$), or common correlation ($R_c = R$, for $c = 1, \ldots, C$), while allowing the 'variance' parameters, i.e., the eigenvalues, innovation variances, and marginal variances respectively, to vary across groups. We first explore some Bayesian analogues of these models. Simple 'default' priors for such models would be to place priors on the mean regression coefficients and (innovation) variances of the form $p(\alpha, \sigma^2; a) \propto \frac{1}{(\sigma^2)^a}$ or on the mean regression coefficients and eigenvalues, $p(\alpha, \lambda; a) \propto \frac{1}{(\lambda)^a}$ [with the additional constraint that orders the eigenvalues]. Ignorance on the matrices P/T might be expressed as uniform priors over the appropriate space, i.e., for P, uniform on the (bounded) space of orthogonal matrices; for T, $p(\phi_{tj}) \propto 1$ (recall, these are unconstrained regression coefficients). Similar priors have been specified for correlation matrices (R) [3]; they assume $p(R) \propto 1$, i.e., a uniform prior on the compact subspace of the $p(p-1)/2$ dimensional cubic $[-1, 1]^{p(p-1)/2}$ such that R is positive definite. In addition, the reference prior derived in Yang

and Berger [33] implies a uniform prior on P and flat improper priors on the logarithm of the eigenvalues, i.e., $p(\lambda) \propto \frac{1}{\lambda}$. In Daniels [8], a flat prior distribution on the ϕ_{tj} is proposed. A recent paper by Sun and Ni [30] chose a constant prior on the (auto)-regressive coefficients in VAR models (equivalent to GARP here) and showed this prior to have good frequentist properties. The literature supports the use of these improper priors for these models and they are the typical default choices when no prior information is available.

We consider some extensions by including (structural or group specific) covariates in the covariance matrix (for details on structural covariates, see [24]). In particular, for the 'variances', $\log(\sigma^2_{ik}) = G_{ik}\eta$, for $i = 1, \ldots, C$ and $k = 1, \ldots, p$ (for the spectral decomposition, replace σ^2_{ik} with λ_{ik}) with $p(\eta) \propto 1$. We consider similar models for the 'dependence', specifically, the GARP, $\phi_{i,tj} = G_{i,tj}\gamma$, for $i = 1, \ldots, C$, $t = 2, \ldots, p$ and $j = 1, \ldots, t-1$ with $p(\gamma) \propto 1$ (cf. [25]).

With the specification of improper priors on the regression parameters and variance 'components', the posterior needs to be checked to determine whether it is a proper density. In the following, we present a theorem which gives sufficient conditions for the propriety of the posterior. Note that the models considered contain as special cases linear regression models with correlated and/or heterogeneous errors (that can depend on covariates). Thus, the propriety results will also hold for these models.

In the following, we provide conditions under which the posterior is proper for the below-specified structures for the mean, variance, and dependence under two decompositions/parameterizations of Σ_i, spectral and GARP/IV, which we will denote as P.I and P.II, respectively.

- mean

  M.I. $E[Y_{ij}] = \mu_i$

  M.II. $E[Y_{ij}] = X_{ij}\alpha$

- variance

  V.I. λ_{ik} under P.I; σ^2_{ik} under P.II

V.II. $\log(\lambda_{ik}) = G_{ik}\eta$ under P.I; $\log(\sigma^2_{ik}) = G_{ik}\eta$ under P.II

- dependence (P/T)

D.I. common across group, $P_i = P$ $i = 1, \ldots, C$ under P.I; $T_i = T$, $i = 1, \ldots, C$ under P.II

D.II. general, $\phi_{i,tj} = G_{i,tj}\gamma$ (only for P.II)

Theorem I: The posterior distribution of (Σ_i, μ_{ij}) for model (1), under each of the mean/variance/dependence assumptions given above and under the priors as described in this Section, will be proper under the following (sufficient) conditions.

- M.I/V.I: $(n_i - 1)/2 + a > 0$ for all i.
  - D.I (under P.I): $S_i = \sum_j (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)^T$ is positive definite for all i
  - D.I (under P.II): Using the first approach discussed in the appendix, we need $n_i > p - 1$ for at least one i, say i', and $\sum_{i \neq i'} n_i > p - 1$; also, $\frac{n_i - p}{2} + a > 0$ for at least one i and $\frac{n_i - 1}{2} + a > 0$ for the remaining i. Under the second approach $\sum_i n_i > p - 1$ and $\frac{n_i - p}{2} + a > 0$ for all i.
  - D.II (under P.II): $\frac{n_i - 1 - dim(\gamma)}{2} + a > 0$ for all i; $\sum_i Z_{ij}^T Z_{ij}$ (Z_{ij} is defined in (13)) is positive definite.
- M.II/V.I: $(n_i - q)/2 + a > 0$ for all i; $\sum_i \sum_j X_{ij}^T X_{ij}$ is positive definite
  - D.I (under P.I): $S_i = \sum_j (y_{ij} - X_{ij}\hat{\alpha})(y_{ij} - X_{ij}\hat{\alpha})^T$ is positive definite for all i ($\hat{\alpha}$ is the generalized least squares estimator).
  - D.I (under P.II): same as M.I. except $\frac{n_i - p - (q-1)}{2} + a > 0$ for at least one i, and $\frac{n_i - q}{2} + a > 0$ for the remaining i for the first approach and $\frac{n_i - p - (q-1)}{2} + a > 0$ for all i for the second approach.
  - D.II (under P.II): $\frac{n_i - q - dim(\gamma)}{2} + a > 0$ for all i; $\sum_i Z_{ij}^T Z_{ij}$ (Z_{ij} is defined in (13)) is positive definite.
- V.II: in addition to relevant conditions above for M.* and D.*, $\sum_i \sum_k G_{ik}^T G_{ik}$ is positive definite and $\frac{n_i - 1}{2} + 1 > 0$ for a subset of i (for details, see the appendix).

This theorem implies that improper priors on the 'regression' parameters and variances lead to proper posteriors under the conditions specified above; in fact, most of these conditions are quite intuitive. Details of the proof can be found in the appendix. The models considered in Theorem I include the following heterogeneous variance regression model [31], $Y_{ij} \sim N(X_{ij}\alpha, \sigma_i^2 I)$ where $\log(\sigma_i^2) = G_i\eta$ with improper priors on α and η as a special case. Thus, the Theorem also provides conditions for a proper posterior for this model.

3 Bayesian extensions

In the previous section, we gave conditions on prior distributions and the data in these covariance models to ensure the posterior is proper. As a next step, we will extend these models in several ways. Specifically, we will append to these models a prior on the 'dependence' parameters, which shrinks the 'dependence' matrices, P_c/T_c for group c, toward a common matrix. Such an approach will offer robustness to assuming equality across groups, but still borrow strength from a 'common matrix', which is important when some groups have small sample sizes. In addition, this can be viewed as an alternative to searching through a large class of models composed of only a subset of parameters being common across groups. This approach to parsimoniously model dependence is similar to the approach of hierarchical modeling often used on means. For each approach, we discuss a 'simple to compute' test for equality of the dependence matrices across groups.

3.1 Shrinking toward constant 'dependence' (P/T)

Within the Bayesian framework, it is natural to offer a compromise between common P/T and having them differ by group. We propose a hierarchical prior that shrinks the group-specific P/T to a common matrix. Consider the T matrix of GARP parameters, ϕ. We specify independent priors for these parameters of the form,

$$\phi_{i,tj} \sim N(\phi_{tj}, \tau^2), i = 1, \ldots, C. \tag{2}$$

These priors shrink the group-specific GARP parameters toward a common value, ϕ_{tj}. Clearly, the degenerate case, $\tau^2 = 0$, corresponds to common GARP. This approach provides some weakening of the common GARP assumption, but still shrinks (or borrows strength) from a common set of GARP parameters. Similar (in spirit) priors for the single group case have been proposed in Daniels and Pourahmadi [12].

The orthogonal matrix, P can be decomposed into Givens angles [10, 23] which all lie in $(-\pi/2, \pi/2)$. Using a logit transformation of these parameters ($h(a) = \frac{log(\pi/2-a)}{log(\pi/2+a)}$), we now have a set of unconstrained parameters that can be modeled similarly to GARP using a normal prior,

$$\mathrm{h}(\theta_{i,tj}) \sim N(h(\theta_{tj}), \tau^2), i = 1, \ldots, C. \tag{3}$$

Daniels and Kass [10] considered a variation of this prior in the single group case, but set $\theta_{tj} = 0$ (independence). We could develop similar priors for an appropriate transformation of the correlations as in Daniels and Kass [11]. However, these are highly constrained for the matrix to be positive definite so we will not discuss such priors here.

For both the GARP and Givens angle shrinkage priors, we place an improper flat prior on ϕ_{tj} and $h(\theta_{tj})$ respectively. Details on the priors for τ^2 are left for the next section.

Computations in the GARP case are straightforward [12] while they are more difficult for the Givens angles [10]. We provide details in Section 4.3.

3.2 Testing constant 'dependence' (P/T)

Given the models proposed in Section 3.1, either (2) or (3), the hypothesis of constancy of P/T across groups, e.g., $H_0 : P_1 = P_2 \cdots = P_C$, is equivalent to $H_0 : \tau^2 = 0$. A Bayes factor, B, for testing this hypothesis has the following form:

$$B = \frac{\int\int \prod_i \prod_j p(y_{ij}|\alpha, \Sigma_i) p(\Sigma_i|\tau^2 = 0) p(\alpha) d\Sigma_i d\alpha}{\int\int\int \prod_i \prod_j p(y_{ij}|\alpha, \Sigma_i) p(\Sigma_i|\tau^2) p(\tau^2) p(\alpha) d\Sigma_i d\tau^2 d\alpha}. \tag{4}$$

It turns out this can be computed with little difficulty by using the Savage-Dickey density ratio [13, 32]. In the next section, we will review this result and show that the required condition is satisfied for our prior specification.

3.2.1 Savage-Dickey Density Ratio

Consider a model with parameters (ω, ψ) and the following hypothesis test:

$$H_0 : \omega = \omega_0$$

$$H_a : \omega \neq \omega_0$$

where the prior under H_a is $p(\omega, \psi)$ and under H_0, $p_o(\psi)$. The Bayes factor, B for comparing these hypotheses is

$$B = \frac{\int L(\omega_0, \psi) p_o(\psi) d\psi}{\int \int L(\omega, \psi) p(\omega, \psi) d\omega d\psi}.$$

where $L(\cdot)$ is the likelihood. Dickey [13] showed that if $p(\psi|\omega_0) = p_0(\psi)$, then

$$B = \frac{p(\omega_0|y)}{p(\omega_0)}. \tag{5}$$

Verdinelli and Wasserman [32] coined (5) the 'Savage-Dickey density ratio'. In our models, $\omega = \tau^2$, $\omega_0 = 0$ and $\psi = (\theta_i, \theta, \alpha, \lambda)$ (or $\psi = (\phi_i, \phi, \alpha, \sigma^2)$). We specify the joint prior for ψ and τ^2 as $p(\psi, \tau^2) = p(\psi)p(\tau^2)$, where the prior on ψ is as specified in Section 2 and 3.1. The prior on τ^2, $p(\tau^2)$, is assumed to take the form $\frac{c}{(c+\tau^2)^2}$, which gives positive probability to the common matrix case and has been shown to have good operating characteristics [7]. This specification satisfies the condition given by Dickey. We will evaluate the sensitivity of the Bayes factor to the constant c, the prior median, in the example.

So, the Bayes factor for testing $H_0 : \tau^2 = 0$ vs. $H_a : \tau^2 \neq 0$ can be written as

$$BF(H_0 : H_a) = \frac{p_{\tau^2}(0|y)}{p_{\tau^2}(0)}$$

where $p_{\tau^2}(\cdot|y)$ is the marginal posterior distribution for τ^2 and $p_{\tau^2}(\cdot)$ is the marginal prior distribution. To compute this ratio, we only need to estimate the ordinate of a one-dimensional

posterior, $p_{\tau^2}(0|y)$, from the posterior sample. We use standard density estimation techniques to evaluate this ordinate [27].

An alternative to the Bayes Factor here, would be to compute $P(\tau^2 > a|y)$ for some constant a. a would be chosen as the smallest value corresponding to a 'practically non-zero' value of τ^2.

4 A more flexible shrinkage paradigm

As discussed in the introduction, searching through all models in terms of commonality of the ϕ or θ across the groups can be burdensome. We propose an extension of the shrinkage methods of the previous section that allows more flexibility. We will illustrate the following with the GARP parameters, though it will follow identically for the logit of the Givens angles. Consider the model

$$\phi_{i,tj} \sim N(\phi_{tj}, \tau^2_{tj}), i = 1, \ldots, C. \tag{6}$$

In (6), we have generalized (2) by allowing a separate shrinkage parameter for each ϕ_{tj}. This allows each parameter to be shrunk individually. However, there may not be much information to estimate these variance components individually, especially if C is small. So, we now place a shrinkage prior on the τ^2_{tj} as follows

$$1/\tau^2_{tj} \sim Gamma(\delta/\tau^2, \delta). \tag{7}$$

Now, we borrow strength on these shrinkage parameters across the $p(p-1)/2$ GARPs. A more parsimonious formulation of (6) and (7), which would be more practical and feasible for smaller C, and is natural in the case of GARP parameterization, would be to replace the separate τ^2_{tj} for each ϕ_{tj} to one for each set of GARP, i.e., a τ^2_t. This indicates a separate shrinkage parameter for each set of regression coefficients for the regression of y_{ijk} on its

predecessors, $y_{ij1}, \ldots, y_{ijk-1}$. We point out that there is not a similar such grouping of Givens angles.

4.1 Bayes Factor

We can again use a Bayes Factor to test for commonality of GARP (Givens angles). First, we determine what values of δ and τ^2 correspond to the null hypothesis of common matrices, i.e., $H_0 : T_1 = \cdots = T_C$. From (7), $E[1/\tau_{tj}^2] = \frac{1}{\tau^2}$ and $Var[1/\tau_{tj}^2] = \frac{1}{\delta}\frac{1}{\tau^2}$. So as $1/\delta \to 0$, this prior is degenerate at $1/\tau^2$. In addition, as $\tau^2 \to 0$, $E[1/\tau_{tj}^2] \to \infty$. Thus, the null hypothesis of common GARP here would correspond to $H_0 : \{\tau^2 = 0 \text{ and } 1/\delta = 0\}$. Similar to the single τ^2 model in Section 3, we place uniform shrinkage priors on these parameters, specifically, τ^2 and $1/\delta$, with the constant c chosen to represent a prior guess at the value of each of the parameters. The Bayes factor for testing common GARP (Givens) will take the form:

$$B = \frac{\int\int \prod_i \prod_j p(y_{ij}|\alpha, \Sigma_i) p(\Sigma_i|\tau^2 = 0, 1/\delta = 0) p(\alpha) d\Sigma_i d\alpha}{\int\int\int\int \prod_i \prod_j p(y_{ij}|\alpha, \Sigma_i) p(\Sigma_i|\tau^2, \delta) p(\tau^2) p(\delta) p(\alpha) d\Sigma_i d\tau^2 d\delta d\alpha}. \tag{8}$$

We can compute a BF using a modification of the approach in Section 3.2. Here, we would need to compute $p_{\tau^2,1/\delta}(0, 0|y)$. As it is difficult to estimate multi-dimensional densities non-parametrically, we will factor the joint density into a marginal and conditional distribution. Thus, the Bayes factor will be able to be computed by just dealing with two one-dimensional estimation problems; this is similar in spirit to Chib's approach for computing Bayes factors based on marginal distributions [6]. The details follow. First, using univariate density estimation techniques, we estimate $p_{1/\delta}(0|y)$ by running a Gibbs sampler on the full model. We then do a second run of the Gibbs sampler conditional on $1/\delta = 0$. This corresponds to the prior on $1/\tau_{tj}^2$ being degenerate at $1/\tau^2$. Thus, we are re-running the Gibbs sampler under the model/prior in (2) or (3). From this sample, we estimate $p_{\tau^2}(0|1/\delta = 0, y)$. We can then multiply these together to estimate $p_{\tau^2,1/\delta}(0, 0|y)$. We also point out that the Bayes

factor for comparing (2) and (6)-(7), given by $p_{1/\delta}(0|y)/p_{1/\delta}(0)$, which is computed in the process of computing the Bayes factor in (8), is just the ratio of the Bayes factors from the two models, i.e., (8) and (4).

As previously, an alternative to Bayes factors would be to compute a posterior probability such as $P(\tau^2 < \epsilon \text{ and } \delta > K)$ for suitably chosen small ϵ and large K.

4.2 Propriety of the Posterior

We now extend the propriety results from Section 2 to models for Σ_i using the shrinkage priors given by (2), (3), or (6)-(7).

Theorem II: The posterior distribution under the shrinkage priors will be proper given the conditions in Theorem I. However, the propriety conditions for the GARP shrinkage priors ((2) or (6)-(7)) can be weakened. See the appendix for details.

Proof: See appendix.

4.3 Computations

To sample from the posterior, we use Gibbs sampling with Metropolis-Hastings steps [28]. This involves sampling sequentially from the full conditional distributions of all the parameters in the specified model. We will also include details for missing data that is MAR; for this, we partition the complete data vector y into (y_{obs}, y_{mis}), the observed and missing responses. First, we will specify the forms of the full conditional distributions that hold for all the models: y_{mis} is (multivariate) normal, $\mu_i(\alpha)$ is normal, and $1/\sigma^2(1/\lambda)$ is (truncated) Gamma. In the single τ^2 models (Section 3), we sample from τ^2 using a random walk Metropolis-Hastings as the full conditional is not available in closed form. In the multiple τ^2's models (Section 4), $1/\tau_{tj}^2$ is Gamma and we need a random walk Metropolis-Hastings for (τ^2, δ). Further details to derive the exact forms of the full conditionals can be found in the Appendix.

4.3.1 GARP shrinkage priors

In the GARP setting, the full conditionals of the $\phi_{i,tj}$ are normally distributed. In particular, for the unrestricted GARP models, the covariance matrix of the GARP parameters will be block diagonal as described in [26]. The full conditional for the 'common' GARP's (ϕ_{tj}), will also be normally distributed.

4.3.2 Givens angles shrinkage priors

Computations are more difficult in the Givens angles models. The full conditional of the logit (h) transformation of the common angles, $h(\theta_{tj})$ will be normally distributed, but the transformation of the group-specific angles, $\theta_{i,tj}$, will not. We use random walk Metropolis-Hastings to sample from these parameters (cf. [10]). For computational efficiency, we also exploit the form of the orthogonal matrix P, with respect to the Givens angles, as described in the introduction, $P = G_{12}G_{13}G_{23}\cdots G_{p-1,p}$ and also use the fact that $G_{tj}G_{tj}^T = I_p$.

5 Examples

We illustrate these models on two datasets, one a longitudinal growth hormone trial [9] and the other, a longitudinal depression study [25]. For each, we had $C = 4$ groups; in the growth hormone trial, $p = 3$ and in the depression studies, $p = 8$. For the latter, if we just consider common or uncommon for each individual ϕ_{tj}, there would be over 200 million possible models (for the former, there are only 8 possible models). We will only fit the GARP models to these applications.

Both examples contained a lot of missing data. We assume the data are missing at random. Carefully modeling the covariance structure is especially important in making inferences in the presence of missing data (Daniels and Hogan, working paper).

Growth Hormone trial

Since there were only three time points in this trial, 0, 6 and 12 months, we only con-

sidered models with one τ^2 as specified in (2). The four 'groups' corresponded to the four treatments given to the subjects in the trials (the sample sizes per group were 41, 41, 40, and 38 respectively). As there was no apparent structure in the means over time, we allowed a separate mean for each time and group (i.e., μ_i). Table 1 contains the GARP for the four groups. In general, across all the parameters, there seem to only be a few major differences in the parameters.

Groups	T		
1	1	0	0
	-0.97	1	0
	-0.45	-0.65	1
	T		
2	1	0	0
	-0.90	1	0
	-0.26	-0.61	1
	T		
3	1	0	0
	-0.88	1	0
	-0.21	-0.59	1
	T		
4	1	0	0
	-0.73	1	0
	0.01	-0.78	1

Table 1: The T_c, $c = 1, \ldots, 4$ matrices for the growth hormone data fitting a distinct Σ for each of the four treatment groups. Recall, the components in the lower triangle of the T_c matrix are $-\phi_{c,tj}$, the negative of the GARP.

For the prior on τ^2, we set $c = .01$. The posterior mean of τ^2 was .0059 (.00028,.0225). The Bayes factor for testing $\tau^2 = 0$ vs $\tau^2 > 0$ was .57. Thus there was very little evidence against common GARP across the treatment groups. The posterior density of τ^2 evaluated

at 0 was not very sensitive to the prior. Increasing c to .1 would give a BF of about 5.7 in favor of $\tau^2 = 0$ and decreasing it to .001 would give a BF of about 15 in favor of $\tau^2 > 0$; not overwhelming evidence in either direction. (Note, the prior evaluated at $\tau^2 = 0$ is $1/c$.) The conclusions here are similar to the likelihood analysis in Pourahmadi et al. (2003).

Table 2 contains the posterior means at month 12 with lengths of 95% credible interval and change from baseline with lengths of 95% credible interval for three models: shrinkage model (2), common GARP model ($\phi_{c,tj} = \phi_{tj}$), and distinct (uncommon) GARP model ($\phi_{c,tj}$). The largest differences between the models is seen in Grp I (in terms of posterior means). In terms of the length of the credible intervals, we see in the change from baseline for Grp IV, there was a 14% increase going to the uncommon GARP model while there was a 10% increase in the month 12 intervals for the same group. These changes are intermediate to those found when comparing common versus uncommon GARP in which we see increases as large as 20%.

Month 12 mean and length of CI				
Model	Grp I	Grp II	Grp III	Grp IV
$\tau^2 > 0$	81.1 (34.9)	65.3 (22.7)	72.7 (26.6)	62.7 (26.8)
uncommon	78.9 (37.9)	65.3 (24.6)	72.6 (27.6)	63.1 (29.8)
common	81.2 (32.9)	65.0 (21.7)	72.7 (24.9)	62.6 (23.8)
Month 12 - Month 0 mean and length of CI				
Model	Grp I	Grp II	Grp III	Grp IV
$\tau^2 > 0$	11.8 (31.9)	-3.1 (20.7)	6.8 (23.3)	-2.5 (25.3)
uncommon	9.6 (33.4)	-3.2 (22.0)	6.7 (25.9)	-2.1 (29.5)
common	11.7 (30.4)	-3.3 (19.5)	6.8 (22.2)	-2.6 (23.8)

Table 2: Month 12 means and 95% credible interval **lengths**

Given the value of the Bayes factor here, we would recommend choosing the shrinkage model with τ^2 (which offers a nice compromise between the common and uncommon GARP models).

Depression trial

This trial was composed of a baseline + 16 weeks of 'active' treatment. Here, we focus

on the baseline week and the first 7 weeks of active treatment for illustration. As the trend over time appeared fairly linear over the first 7 weeks (however, it looked quadratic over the whole 16 week period), we assumed a linear trend for the mean structure. Previous analyses suggested major differences in dependence across four groups formed from combining the binary indicators for drug and for initial severity [25]; here, the sample sizes per group were much larger than the previous example, ranging from 98 to 250.

We set $c_{1/\delta} = .01$ and $c_{\tau^2} = .01$ for the priors. Since C was only four, but we had $p = 8$ times, we consider models that allowed the shrinkage parameter, τ^2 to vary across sets of GARP coefficients, τ_t^2, as specified in (6)-(7). The Bayes factor against common GARP in this model was > 2700. Again, the joint posterior density for δ and τ^2 was relatively insensitive to the priors. In order to get a BF that was close to 1, we would have needed to set c_{τ^2} to 25, which is a ridiculous value (especially, since the GARP tend to vary only between ± 1). So clearly, the data did not support common GARP for this data. Earlier work explored parsimonious GARP models for this data [25].

		Models	
	$\tau^2 > 0$	common	uncommon
int	17.24 (1.0)	17.27 (1.0)	17.24 (.99)
slope	-1.02 (.40)	-1.03 (.44)	-1.00 (.40)
int*Gp2	5.83 (1.50)	5.85 (1.51)	5.84 (1.49)
slope*Gp2	-.51 (.70)	-.50 (.71)	-.57 (.77)
int*Gp3	.17 (1.20)	.14 (1.18)	.18 (1.17)
slope*Gp3	-.15 (.59)	-.13 (.60)	-.13 (.62)
int*Gp4	6.83 (1.28)	6.78 (1.27)	6.86 (1.29)
slope*Gp4	-.95 (.52)	-.93 (.54)	-.99 (.56)

Table 3: Posterior means and lengths of 95% credible intervals for α.

Posterior means and lengths of 95% credible intervals for the mean regression coefficients, α appear in Table 3. The posterior means, relative to the lengths of the credible intervals are fairly similar across models. However, there was some variability in the lengths of the credible intervals across models. For example, for the coefficient for slope*Gp2, the credible interval was 10% longer in the shrinkage model versus the uncommon GARP model; the credible

interval for the 'baseline' slope increased by 10% as well when comparing the shrinkage model to the common GARP model. As in the growth hormone trial example, we would recommend inferences based on the shrinkage model over the common GARP model (based on Bayes factor results) and over the uncommon GARP model (based on parsimony as this model contains 112 dependence parameters).

6 Extension: Shrinking the 'variances'

Clearly, we can also shrink across groups for the 'variances'. However, this is of less concern than for 'dependence' parameters as there are $Cp(p-1)/2$ (on order of Cp^2) dependence parameters and only Cp variance parameters. Considering equal or not equal across groups corresponds to $2^{p(p-1)/2}$ possible models for the dependence parameters, but only 2^p for the variance parameters; for $p = 5$, that implies $2^{15} = 32678$ vs $2^5 = 32$. However, we will still discuss how this might be done below.

An approach for shrinking across groups with the marginal variances was proposed in [3] using log-normal priors on the variances, centered at values determined by group-specific covariates. An alternative approach to shrinking, particularly attractive from a computational perspective for the innovation variances and eigenvalues, would be to place Gamma priors on these 'variances'. For the diagonal elements of D_i, consider

$$1/\sigma_{ij}^2 \sim Gamma(\delta/\sigma_j^2, \delta), j = 1, \ldots, p, i = 1, \ldots, C$$

where $1/\sigma_j^2$ is the expectation of $1/\sigma_{ij}^2$; this is a similar parameterization to that used on the τ^2's in Section 4 (cf. (7)). Similar models have recently been proposed by Lin, Raz and Harlow [20] and Daniels [8]. By specifying Gamma priors, the full conditional distributions of $1/\sigma_{ij}^2$ will be Gamma which facilitates Gibbs sampling approaches for sampling from the posterior.

Similar to the previous sections, we can test for the equality of the 'variances' across groups using Bayes factors. As in Section 4, $1/\delta = 0$ corresponds to equality. Assuming proper priors on δ and $1/\sigma_j^2$, the propriety of the posterior for these models follows from Theorem I.

7 Discussion

We have shown, that under certain reasonable conditions, use of improper priors on means, variance, and dependence parameters in many correlated and heterogeneous error regression models results in proper posteriors. We have also proposed a simple, parsimonious way to model covariance matrices across groups and showed how a Bayes Factor to test for commonality can easily be computed.

Additional application areas for this methodology include pattern mixture models for non-ignorable missing data [19]; here, the patterns of missing data could be considered as 'groups'. Further issues arise in this setting due to non-identifiable parameters. This will be explored in future work.

Bayes factors were proposed to test for commonality of the 'dependence' matrices in the shrinkage models. Alternatives to Bayes Factors include the DIC [29] and posterior predictive loss [17]. Using heavier tailed distributions than the normal distribution for the priors (cf: (2)) will result in less overall shrinkage; for example, a straightforward extension could be developed using the representation of t-distributions as a gamma mixture of normals [1].

Future work will explore the propriety of the posterior in the common correlation model with and without restrictions on the marginal variances and in models with the Givens angles a function of group-specific covariates as in the GARP models. In addition, we will attempt to weaken the conditions for propriety in the models explored here as some of the current conditions are sufficient, but not necessary. Finally, we will explore approaches to search through the space of GARP models to facilitate determining which parameters vary across

groups and/or are equal across only a subset of the groups. This will allow more flexibility than the shrinkage approaches proposed here, but at a computational cost.

Appendix: Proof of Theorems I and II

Details for Theorems I and II

In the following, we will refer to the models for the covariance matrix under the GARP/IV decomposition as GARP models and under the spectral decomposition as PC (principal components) models. We remind the reader that to prove the propriety of the posterior distributions, it is sufficient to show that the marginal density of y is finite.

We first present a lemma which will be used in the proofs:

Lemma I: For two positive definite matrices, A_1 and A_2, $|A_1 + A_2| \geq |A_1| \geq \lambda_p(A_1)^p$, where $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \lambda_p(A)$, are the ordered eigenvalues of a $p \times p$ matrix A.

Proof:

$$\begin{aligned} |A_1 + A_2| &\geq \textstyle\prod_{i=1}^{p}[\lambda_i(A_1) + \lambda_i(A_2)] \\ &\geq \textstyle\prod_i[\lambda_p(A_1) + \lambda_i(A_2)] \\ &\geq \lambda_p(A_1)^p \end{aligned}$$

The first inequality appears in Marshall and Olkin [22].

We recall the priors on the 'variance' parameters, $p(\lambda_{ik}; a) \propto \frac{1}{(\lambda_{ik})^a}$, $k = 1, \ldots, p$ and $p(\sigma^2_{ik}; a) \propto \frac{1}{(\sigma^2_{ik})^a}$, $k = 1, \ldots, p$, $i = 1, \ldots, C$. We first integrate out the mean parameters (two cases).

Case I: $E[Y_{ij}] = \mu_i$. Assume $p(\mu_i) \propto 1$.

$$\begin{aligned} &\prod_i \int \frac{1}{|\Sigma_i|^{n_i/2}} \exp(-\frac{1}{2}\sum_j (y_{ij} - \mu_i)^T \Sigma_i^{-1}(y_{ij} - \mu_i)) d\mu_i \\ \propto &\prod_i \frac{1}{|\Sigma_i|^{(n_i-1)/2}} \exp(-\frac{1}{2}\sum_j (y_{ij} - \bar{y}_i)^T \Sigma_i^{-1}(y_{ij} - \bar{y}_i)) \\ = &\prod_i \frac{1}{|\Sigma_i|^{(n_i-1)/2}} \exp(-\frac{1}{2}\sum_j (y^\star_{ij})^T \Sigma_i^{-1}(y^\star_{ij})) \end{aligned} \tag{9}$$

where $y^\star_{ij} = y_{ij} - \bar{y}_i$.

Case II: $E[Y_{ij}] = X_{ij}\alpha$. Assume $p(\alpha) \propto 1$.

Define $\hat{\alpha} = (\sum_i \sum_j X_{ij}^T \Sigma_i^{-1} X_{ij})^{-1} \sum_i \sum_j X_{ij}^T \Sigma_i^{-1} Y_{ij}$ and $\lambda_{max} = max_i \lambda_{i1}$. Define P_k to be the kth column of the orthogonal matrix P.

$$
\begin{aligned}
& \int \frac{1}{\prod_i |\Sigma_i|^{n_i/2}} \exp(-\tfrac{1}{2} \sum_i \sum_j (y_{ij} - X_{ij}\alpha)^T \Sigma_i^{-1} (y_{ij} - X_{ij}\alpha)) d\alpha \\
\propto \; & \frac{1}{\prod_i |\Sigma_i|^{n_i/2} |\sum_i \sum_j X_{ij}^T \Sigma_i^{-1} X_{ij}|^{1/2}} \exp(-\tfrac{1}{2} \sum_i \sum_j (y_{ij} - X_{ij}\hat{\alpha})^T \Sigma_i^{-1} (y_{ij} - X_{ij}\hat{\alpha})) \\
= \; & \frac{1}{\prod_i |\Sigma_i|^{n_i/2} |\sum_i \sum_{k=1}^p \frac{1}{\lambda_{ik}} \sum_j X_{ij}^T P_k P_k^T X_{ij}|^{1/2}} \exp(-\tfrac{1}{2} \sum_i \sum_j (y_{ij} - X_{ij}\hat{\alpha})^T \Sigma_i^{-1} (y_{ij} - X_{ij}\hat{\alpha})) \\
= \; & \frac{1}{\prod_i |\Sigma_i|^{n_i/2} (1/\lambda_{max})^{q/2} |\sum_i \sum_k \frac{\lambda_{max}}{\lambda_{ik}} \sum_j X_{ij}^T P_k P_k^T X_{ij}|^{1/2}} \exp(-\tfrac{1}{2} \sum_i \sum_j (y_{ij} - X_{ij}\hat{\alpha})^T \Sigma_i^{-1} (y_{ij} - X_{ij}\hat{\alpha}))
\end{aligned}
$$

If we assume that $\sum_i \sum_j X_{ij}^T X_{ij}$ is positive definite, a standard assumption that can be checked for any dataset. then $|\sum_i \sum_k \frac{\lambda_{max}}{\lambda_{ik}} \sum_j X_{ij}^T P_k P_k^T X_{ij}|$ is positive. We use this fact to place an upper bound on the square root of its inverse, say $M^\star$. Thus,

$$
\begin{aligned}
\leq \; & M^\star \frac{1}{\prod_i |\Sigma_i|^{n_i/2} (1/\lambda_{max})^{q/2}} \exp(-\tfrac{1}{2} \sum_i \sum_j (y_{ij} - X_{ij}\hat{\alpha})^T \Sigma_i^{-1} (y_{ij} - X_{ij}\hat{\alpha})) \\
\propto \; & \frac{1}{\prod_i |\Sigma_i|^{n_i/2} (1/\lambda_{max})^{q/2}} \exp(-\tfrac{1}{2} \sum_i \sum_j (y_{ij} - X_{ij}\hat{\alpha})^T \Sigma_i^{-1} (y_{ij} - X_{ij}\hat{\alpha}))
\end{aligned}
$$

We remind the reader that only the result for Case I was exact.

To use the above result for the GARP models, just replace P_k with T_k (T_{ik}) where T_k is the kth column of the T matrix, and λ_{ik} with σ_{ik}^2. The same argument follows as was used for the spectral decomposition.

$$
\begin{aligned}
& \frac{1}{\prod_i |\Sigma_i|^{n_i/2} |\sum_i \sum_j X_{ij}^T \Sigma_i^{-1} X_{ij}|^{1/2}} \exp(-\tfrac{1}{2} \sum_i \sum_j (y_{ij} - X_{ij}\hat{\alpha})^T \Sigma_i^{-1} (y_{ij} - X_{ij}\hat{\alpha})) \\
\leq \; & M^\star \frac{1}{\prod_i |\Sigma_i|^{n_i/2} (1/\sigma_{max}^2)^{q/2}} \exp(-\tfrac{1}{2} \sum_i \sum_j (y_{ij} - X_{ij}\hat{\alpha})^T \Sigma_i^{-1} (y_{ij} - X_{ij}\hat{\alpha})) \\
\propto \; & \frac{1}{\prod_i |\Sigma_i|^{n_i/2} (1/\sigma_{max}^2)^{q/2}} \exp(-\tfrac{1}{2} \sum_i \sum_j (y_{ij} - X_{ij}\hat{\alpha})^T \Sigma_i^{-1} (y_{ij} - X_{ij}\hat{\alpha}))
\end{aligned}
$$

One final issue is that $\hat{\alpha}$ is a function of Σ_i. Define $S_i(\hat{\alpha}) = \sum_j (y_{ij} - X_{ij}\hat{\alpha})(y_{ij} - X_{ij}\hat{\alpha}))^T$, which is guaranteed to be positive definite under the previous conditions. We will remove the dependence of $S_i(\hat{\alpha})$ on Σ_i by bounding the exponential term. Re-write the exponential term in the above expression as

$$\begin{aligned} tr[\Sigma_i^{-1} S_i(\hat{\alpha})] &\geq \sum_{k=1}^{p} \lambda_k(\Sigma_i^{-1}) \lambda_{p-k+1}(S_i(\hat{\alpha})) \\ &\geq \sum_{k=1}^{p} \lambda_k(\Sigma_i^{-1}) \lambda_p(S_i(\hat{\alpha})) \end{aligned}$$

where $\lambda_k(\cdot)$ is defined as in Lemma I. The first inequality is from Ingram and Olkin [22]. Now define $\lambda_{min,i} = min_{\Sigma_i} \lambda_p(S_i(\hat{\alpha})) > 0$. Then,

$$\begin{aligned} tr[\Sigma_i^{-1} S_i(\hat{\alpha})] &\geq \sum_{k=1}^{p} \lambda_k(\Sigma_i^{-1}) \lambda_{min,i} \\ &= tr[\Sigma_i^{-1} \lambda_{min,i} I_p] \end{aligned}$$

Finally, for each i, simulate a 'new' set of data, $y_{ij}^{\star}$, from a normal distribution under the constraint that $\sum_j y_{ij}^{\star} y_{ij}^{\star T} = \lambda_{min,i} I_p$. So,

$$\begin{aligned} & \frac{1}{\prod_i |\Sigma_i|^{n_i/2} (1/\sigma_{max}^2)^{q/2}} \exp(-\frac{1}{2} \sum_i \sum_j (y_{ij} - X_{ij}\hat{\alpha})^T \Sigma_i^{-1} (y_{ij} - X_{ij}\hat{\alpha})) \\ \leq & \frac{1}{\prod_i |\Sigma_i|^{n_i/2} (1/\sigma_{max}^2)^{q/2}} \exp(-\frac{1}{2} \sum_i \sum_j (y_{ij}^{\star})^T \Sigma_i^{-1} (y_{ij}^{\star})) \end{aligned}$$

The rest of the proof is given under mean case I, but will also follow with mean case II (appropriate conditions for this case appear in the statement of the Theorem I in Section 2).

Theorem I details

Common GARP

For common GARP models, re-write Σ_i^{-1} as $T D_i^{-1} T^T$, note that $|\Sigma_i| = |D_i|$, and rewrite (9) as

$$\prod_i \frac{1}{|D_i|^{(n_i-1)/2}} \exp(-\frac{1}{2} \sum_j (y_{ij}^{\star} - Z_{ij}\phi)^T D_i^{-1} (y_{ij}^{\star} - Z_{ij}\phi))$$

where ϕ is the vector of common GARP parameters and Z_{ij} is the $p \times p(p-1)/2$ matrix defined below:

$$Z_{ij} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ y^\star_{ij1} & 0 & 0 & \cdots & 0 \\ 0 & y^\star_{ij1} & y^\star_{ij2} & \cdots & 0 \\ \vdots & & & & \\ 0 & \cdots & y^\star_{ij1} & \cdots & y^\star_{ij,p-1} \end{pmatrix}. \tag{10}$$

We now integrate out ϕ under $p(\phi) \propto 1$. Define $\hat{\phi} = (\sum_i \sum_j Z_{ij}^T D_i^{-1} Z_{ij})^{-1} \sum_i \sum_j Z_{ij}^T D_i^{-1} y^\star_{ij}$ and $SSE_i = \sum_j (y^\star_{ij} - Z_{ij}\hat{\phi})(y^\star_{ij} - Z_{ij}\hat{\phi})^T$.

$$\begin{aligned}
&\prod_i \int \frac{1}{|D_i|^{(n_i-1)/2}} \exp(-\frac{1}{2}\sum_j (y^\star_{ij} - Z_{ij}\phi)^T D_i^{-1}(y^\star_{ij} - Z_{ij}\phi))d\phi \\
\propto &\frac{1}{|\sum_i \sum_j Z_{ij}^T D_i^{-1} Z_{ij}|^{1/2} \prod_i |D_i|^{(n_i-1)/2}} \exp(-\frac{1}{2}\sum_i \sum_j (y^\star_{ij} - Z_{ij}\hat{\phi})^T D_i^{-1}(y^\star_{ij} - Z_{ij}\hat{\phi})) \\
= &\frac{1}{|\sum_i \sum_j Z_{ij}^T D_i^{-1} Z_{ij}|^{1/2} \prod_i |D_i|^{(n_i-1)/2}} \exp(-\frac{1}{2}\sum_i tr[D_i^{-1} SSE_i]) \\
= &\frac{1}{|\sum_i \sum_j Z_{ij}^T D_i^{-1} Z_{ij}|^{1/2} \prod_i |D_i|^{(n_i-1)/2}} \exp(-\frac{1}{2}\sum_i \sum_{k=1}^p \frac{1}{\sigma_{ik}^2} SSE_{ik})
\end{aligned} \tag{11}$$

where SSE_{ik} is the kth diagonal element of SSE_i. If $\sum_i n_i \geq p(p-1)/2 = dim(\phi)$ and $\sum_i \sum_j Z_{ij}^T Z_{ij}$ positive definite, then $SSE *_k (D_i) > 0$. However, we can weaken these conditions by recognizing that the terms $Z_{ij}^T D_i^{-1} Z_{ij}$ are block diagonal with blocks of the form $\frac{1}{\sigma_{ik}^2}\sum_j Z^{\star T}_{ijk} Z^\star_{ijk}$, for $k = 2, \ldots, p$ each of dimension $(k-1)$. Since the largest block is of dimension, $p-1$, only need the condition that $\sum_i n_i > p-1$. We also note that $|D_i| = \prod_k \sigma_{ik}^2$. To proceed, we have two choices: 1) apply Lemma I or 2) follow a similar argument to dealing with mean case II. First, we go through the application of Lemma I, (11) is equal to

$$\begin{aligned}
&\frac{1}{\prod_k |\sum_i \frac{1}{\sigma_{ik}^2} \sum_j Z^{\star T}_{ijk} Z^\star_{ijk}|^{1/2} \prod_i \prod_k \sigma_{ik}^{2\,(n_i-1)/2}} \exp(-\tfrac{1}{2}\textstyle\sum_i \sum_{k=1}^p \frac{1}{\sigma_{ik}^2} SSE_{ik}) \\
\leq &\frac{1}{\prod_k |\sum_j Z^{\star T}_{i'jk} Z^\star_{i'jk}|^{1/2}} \frac{1}{\prod_{k=2}^p \sigma_{i'k}^{2\,(n_{i'}-k)/2}} \frac{1}{\prod_{i \neq i'} \prod_k \sigma_{ik}^{2\,(n_i-1)/2}} \exp(-\tfrac{1}{2}\textstyle\sum_i \sum_{k=1}^p \frac{1}{\sigma_{ik}^2} SSE_{ik})
\end{aligned}$$

To apply Lemma 1 above to create the upper bound, we need, for at least one i, say i', the following two matrices to be full rank $\sum_j Z^{\star T}_{i'jp} Z^\star_{i'jp}$ and $\sum_{i \neq i'} \sum_j Z^{\star T}_{ijp} Z^\star_{ijp}$ (a sufficient

condition for these two matrices to be full rank is for $n_{i'} > p-1$ and $\sum_{i \neq i'} n_i > p-1$). If we use the second approach, an alternative set of conditions can be constructed. Following this approach, the conditions are $\sum_i n_i > p-1$, which implies that $\sum_i \sum_j Z_{ijp}^{\star T} Z_{ijp}^{\star}$ is full rank.

Common Principal Components (CPC)

For CPC (i.e., $P_i = P$, for $i = 1, \ldots, C$), (9) is equal to

$$\begin{aligned} & \frac{1}{\prod_i \prod_k \lambda_{ik}^{(n_i-1)/2}} \exp(-\frac{1}{2} \sum_i \sum_j (y_{ij}^{\star})^T P^T \Lambda_i^{-1} P (y_{ij}^{\star})) \\ = & \frac{1}{\prod_i \prod_k \lambda_{ik}^{(n_i-1)/2}} \exp(-\frac{1}{2} \sum_i tr[\Lambda_i^{-1} \sum_j P (y_{ij}^{\star})(y_{ij}^{\star})^T P^T]) \\ = & \frac{1}{\prod_i \prod_k \lambda_{ik}^{(n_i-1)/2}} \exp(-\frac{1}{2} \sum_i \sum_k \frac{1}{\lambda_{ik}} P_{ik}^{\star}) \end{aligned} \tag{12}$$

where $P_{ik}^{\star} = P_k^T S_i P_k$, with P_k the kth column of P and $S_i = \sum_j (y_{ij}^{\star})(y_{ij}^{\star})^T$.

General GARP models

For general GARP models, $\phi_{i,tl} = G_{i,tl}\gamma$. The proof will be similar to common GARP (though we no longer have block diagonality in this case). Define the kth column of the matrix Z_{ij} to be

$$Z_{ijk} = \sum_{l=1}^{k-1} G_{i,kl} y_{ijl}^{\star} \tag{13}$$

Note that,

$$\begin{aligned} & |\sum_i \sum_j Z_{ij}^T D_i^{-1} Z_{ij}| \\ = & |\sum_i \sum_k \tfrac{1}{\sigma_{ik}^2} \sum_j Z_{ijk} Z_{ijk}^T| \\ = & (1/\sigma_{max}^2)^{dim(\gamma)} |\sum_i \sum_k \tfrac{\sigma_{max}^2}{\sigma_{ik}^2} \sum_j Z_{ijk} Z_{ijk}^T| \end{aligned}$$

We can then proceed as with mean Case II under the condition that $\sum_i \sum_j Z_{ij}^T Z_{ij}$ is positive definite, so $\sum_i \sum_k \frac{\sigma_{max}^2}{\sigma_{ik}^2} \sum_j Z_{ijk} Z_{ijk}^T$ will be positive definite and we can use a finite upper bound for the reciprocal of its determinant. We use this result after integrating out α and

then $\gamma(\phi)$ (see common GARP case for details). We also point out that specific choices for $G_{i,tl}$ can result in even weaker conditions than those given above (for example, recall the conditions specific to the common GARP which are much weaker than those given here).

Unrestricted variance ($\sigma^2_{ik}/\lambda_{ik}$)

For common GARP model and mean case I, the result, after integrating out the mean and GARP parameters, looks like the product of independent inverse gamma distributions for each σ^2_{ik}. So sufficient conditions under the first approach described earlier (Lemma I) for propriety are $\frac{n_i-p}{2} + a > 0$ for at least one i and $\frac{n_i-1}{2} + a > 0$ for the other i. For mean case II, the latter two conditions are replaced with $\frac{n_i-p-(q-1)}{2} + a > 0$ for at least one i and $\frac{n_i-q}{2} + a > 0$ for the other i. Using the latter approach (analogy to dealing with mean case II), the conditions are $\frac{n_i-p}{2} + a > 0$ for all i for mean case I and $\frac{n_i-p-(q-1)}{2} + a > 0$ for all i for mean case II.

For CPC, we need S_i full rank for all i, which implies $P^\star_{ik} > 0$. We can then bound $\exp(-\frac{1}{2}\sum_i \sum_k \frac{1}{\lambda_{ik}} P^\star_{ik})$ by $\exp(-\frac{1}{2}\sum_i \sum_k \frac{1}{\lambda_{ik}} \epsilon_{ik})$, where $\epsilon_{ik} > 0$ is the lower bound on $P^\star_{ik}$. For mean case I, we need in addition $\frac{n_i-1}{2} + a > 0$; for mean case II, $\frac{n_i-q}{2} + a > 0$ for all i.

Using these conditions and the previous results, the posterior will be finite for the GARP models since the integral is just a mixture over finite, integrable densities over σ^2, with degrees of freedom of the inverse gamma distribution varying as specified above. For the models based on the spectral decomposition, the posterior will be integrable since $\int dP < \infty$ and the integral over the eigenvalues will be bounded by these inverse gamma distributions since the eigenvalues are ordered.

General case of $\log(\sigma^2_{ik}) = G_{ik}\gamma$, $\log(\lambda_{ik}) = G_{ik}\gamma$

First, assume $dim(\gamma) = q_g$. For ease, we will do the proof for $E[Y_{ij}] = \mu_i$ (mean Case I) and CPC. Other models follow similarly. Continuing from the CPC model with the mean integrated out, (12) is equal to

$$\prod_i [\frac{1}{\prod_k \lambda_{ik}^{(n_i-1)/2}}] \exp(-\frac{1}{2}\sum_i \sum_k \frac{1}{\lambda_{ik}} P_{ik}^\star)$$
$$= \prod_i [\frac{1}{\prod_k \exp(G_{ik}\gamma)^{(n_i-1)/2}}] \exp(-\frac{1}{2}\sum_i \sum_k \frac{1}{\exp(G_{ik}\gamma)} P_{ik}^\star) \quad (14)$$

Choose q_g vectors from the set $\{G_{ik} : k = 1, \ldots, p; i = 1, \ldots, C\}$, which we will denote as the set Q, such that $\sum_{i,k\in Q} G_{ik}^T G_{ik}$ is of full rank, q_g. Without loss of generality, assume $G_{ikj} \geq 0$. Then, (14) is equal to

$$\frac{1}{\prod_{i,k\in Q} \exp(G_{ik}\gamma)^{(n_i-1)/2}} \exp(-\frac{1}{2}\sum_{i,k\in Q} \frac{1}{\exp(G_{ik}\gamma)} P_{ik}^\star) \quad (15)$$
$$\times \frac{1}{\prod_{i,k\in Q^c} \exp(G_{ik}\gamma)^{(n_i-1)/2}} \exp(-\frac{1}{2}\sum_{i,k\in Q^c} \frac{1}{\exp(G_{ik}\gamma)} P_{ik}^\star) \quad (16)$$

We will first show the product over the terms not in Q is bounded. Rewrite (16) as

$$\prod_{i,k\in Q^c} \frac{\exp(-\frac{1}{2}\frac{P_{ik}^\star}{\exp(G_{ik}\gamma)})}{\exp(G_{ik}\gamma)^{(n_i-1)/2}}$$

Clearly, these terms are all bounded. Call the product of the upper bounds of all these terms $M^\star$. We now go back to the terms corresponding to the set Q. We do a linear transformation from γ to $G_{ik}\gamma$ for (i,k) in set Q. Since $\sum_{i,k\in Q} G_{ik}^T G_{ik}$ is of full rank, this is a full rank transformation with Jacobian, J_1, a function of the components of the G_{ik} vectors of Q. Denoting the previous linear transformation as $A\gamma$, we now do another transformation to $\exp(A\gamma) = \tau$, (where we define the exponential of a vector as the vector of exponentials) which will have Jacobian $J(\tau) = \frac{1}{\prod_{l=1}^{q_g} \tau_l}$. Define the set n_l' to be the set of q_q n_i corresponding to the G_{ik} in Q. Then, (14) is equal to

$$
\begin{aligned}
& \frac{1}{\prod_{i,k\in Q}\exp(G_{ik}\gamma)^{(n_i-1)/2}}\exp(-\tfrac{1}{2}\sum_{i,k\in Q}\frac{1}{\exp(G_{ik}\gamma)}P^{\star}_{ik}) \\
& \times\frac{1}{\prod_{i,k\in Q^c}\exp(G_{ik}\gamma)^{(n_i-1)/2}}\exp(-\tfrac{1}{2}\sum_{i,k\in Q^c}\frac{1}{\exp(G_{ik}\gamma)}P^{\star}_{ik}) \\
\leq\ & M^{\star}\frac{1}{\prod_{i,k\in Q}\exp(G_{ik}\gamma)^{(n_i-1)/2}}\exp(-\tfrac{1}{2}\sum_{i,k\in Q}\frac{1}{\exp(G_{ik}\gamma)}P^{\star}_{ik}) \\
\propto\ & \frac{1}{\prod_{l=1}^{qg}\tau_l^{(n'_l-1)/2}}\exp(-\tfrac{1}{2}\sum_{l=1}^{qg}\frac{1}{\tau_l}P^{\star}_l)J(\tau)J_1 \\
\propto\ & \frac{1}{\prod_{l}\tau_l^{(n'_l-1)/2+1}}\exp(-\tfrac{1}{2}\sum_{l}\frac{1}{\tau_l}P^{\star}_l)
\end{aligned}
$$

As previously, we put a lower bound on $P^{\star}_l$ (which under the current conditions is strictly greater than zero), $\epsilon_l > 0$. This looks like the product of gamma distributions in $\frac{1}{\tau_l}$. This will be integrable as long as $(n'_l-1)/2+1 > 0$ for all l and the integrated result will be finite with respect to the integral over P since $\int dP < \infty$.

Theorem II Details

Shrinkage prior on Givens angles

If the joint prior $p(\theta_i, \tau^2, \delta)$ is proper, then previous conditions for PC models are sufficient. We assume proper priors on parameters (τ^2, δ^2) as given in Sections 3 and 4 and a flat prior on $h(\theta)$ (where h is the 'logit' function specified in Section 3). If $C > 1$, then the prior $p(\theta_i, \tau^2, \delta)$ is proper.

Shrinkage prior on GARP

As in the Givens angle case, the joint prior $p(\phi_i, \tau^2, \delta)$ is proper, if $C > 1$ (Note: if C was not greater than 1, we would not even consider these models). If we ignore the proper prior on ϕ_i, then having a separate ϕ_i for each group can be expressed in term of the general GARP model with $\gamma^T = (\phi_1^T, \ldots, \phi_C^T)$, where ϕ_i is the $p(p-1)/2$ dimensional set of GARP parameter for group i. But the sufficient condition that $\frac{n_i-1-dim(\gamma)}{2} + a > 0$ is too strong (here, $dim(\gamma) = Cp(p-1)/2$). So we will weaken this condition below. After integrating out the mean (under Mean Case I, but this easily generalizes to Mean Case II), we re-write (9) as the following and then, integrate over ϕ_i,

$$\begin{aligned} & \prod_i \int \frac{1}{|D_i|^{(n_i-1)/2}} \exp(-\tfrac{1}{2}\sum_j (y^\star_{ij} - Z_{ij}\phi_i)^T D_i^{-1}(y^\star_{ij} - Z_{ij}\phi_i))d\phi_i \\ = & \prod_i \frac{1}{|D_i|^{(n_i-1)/2}|\sum_j Z_{ij}^T D_i^{-1} Z_{ij}|^{1/2}} \exp(-\tfrac{1}{2}\sum_j (y^\star_{ij} - Z_{ij}\hat{\phi}_i)^T D_i^{-1}(y^\star_{ij} - Z_{ij}\hat{\phi}_i)) \end{aligned}$$

where $\hat{\phi}_i = (\sum_j Z_{ij}^T D_i^{-1} Z_{ij})^{-1} \sum_j Z_{ij}^T D_i^{-1} y^\star_{ij}$ and Z_{ij} is given in (10). If we use the fact that Z_{ij} is block diagonal, as in the common GARP case, then we get the following condition: $n_i > p-1$ (which implies $\sum_j Z^{\star T}_{ijk} Z^\star_{ijk}$ is full rank for all i and k). If this integral is finite, then it will also be finite with a proper prior on ϕ_i.

Appendix: Computational details

Forms of the Normal likelihood

For deriving the full conditional distributions for α and for y_{mis}, under MAR, we use the following expression for the likelihood,

$$\frac{1}{\prod_i |\Sigma_i|^{n_i/2}} \exp(-\frac{1}{2}\sum_i \sum_j (y_{ij} - X_{ij}\alpha)^T \Sigma_i^{-1}(y_{ij} - X_{ij}\alpha)).$$

For deriving the full conditional distributions for covariance parameters in the GARP models, we re-express the likelihood as

$$\frac{1}{\prod_i \prod_{k=1}^p \sigma_{ik}^{2\ n_i/2}} \exp(-\frac{1}{2}\sum_i \sum_j ((y_{ij} - X_{ij}\alpha) - Z_{ij}\phi_i)^T D_i^{-1}((y_{ij} - X_{ij}\alpha) - Z_{ij}\phi_i)).$$

For deriving the full conditional distributions for covariance parameters in the spectral decomposition models, we re-express the likelihood as

$$\frac{1}{\prod_i \prod_{k=1}^p \lambda_{ik}^{n_i/2}} \exp(-\frac{1}{2}\sum_i \Lambda_i^{-1} \sum_j P_i(y_{ij} - X_{ij}\alpha)(P_i(y_{ij} - X_{ij}\alpha))^T).$$

where $P_i = G_{i,12}(\theta_{i,12})G_{i,13}(\theta_{i,13})G_{i,23}(\theta_{i,23}) \cdots G_{i,p-1,p}(\theta_{i,p-1,p})$

Also, the shrinkage prior in Section 3.1 (for GARP) is proportional to

$$[\prod_i \prod_{t=2}^p \prod_{j=1}^{t-1} (\frac{1}{\tau^2})^{1/2} \exp(-\frac{1}{2\tau^2}(\phi_{i,tj} - \phi_{tj})^2)] \frac{c}{(c+\tau^2)^2}$$

and for Section 3.2,

$$\prod_{t=2}^{p}\prod_{j=1}^{t-1}\prod_{i}[(\frac{1}{\tau_{tj}^2})^{1/2}\exp(-\frac{1}{2\tau_{tj}^2}(\phi_{i,tj}-\phi_{tj})^2)]\frac{(\frac{1}{\tau_{tj}^2})^{\delta/\tau^2-1}\exp(-\delta/\tau_{tj}^2)}{\Gamma(\delta)\delta^{\delta/\tau^2}}.$$

For the Givens angles, replace ϕ with $h(\theta)$.

The form of the full conditionals for all models are easily derived from these forms for the likelihood and priors.

Acknowledgments

This work was partially supported by NIH grants CA85295 and HL079457. The author would like to thank Professor Malay Ghosh for some helpful discussions regarding matrix manipulations, Professor Dongchu Sun for pointing out the text by Marshall and Olkin, Professor Brett Presnell on bounding some of the terms, and two referees whose comments improved the presentation. The author would also like to thank Dr. David MacLean (Memorial Hospital of Rhode Island and Pfizer) for providing the growth hormone data and Professor Joel Greenhouse for the depression data.

References

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

[2] J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

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

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

[5] T. Bollerslev. Modeling the coherence in short run nominal exchange rates: A multivariate generalized arch model. *The Review of Economics and Statistics*, 72:498–505, 1990.

[6] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

[7] M. J. Daniels. A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, 27:567–578, 1999.

[8] M. J. Daniels. Shrinkage priors for the dependence structure in longitudinal data. *Journal of Statistical Planning and Inference*, 127:119–130, 2005.

[9] M. J. Daniels and J. W. Hogan. Reparameterizing the pattern-mixture model for sensitivity analyses under informative drop-out. *Biometrics*, 56:1241–1248, 2000.

[10] M. J. Daniels and R. E. Kass. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94:1254–1263, 1999.

[11] M. J. Daniels and R. E. Kass. Shrinkage estimators for covariance matrices. *Biometrics*, 57(4):1173–1184, 2001.

[12] M. J. Daniels and M. Pourahmadi. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89(3):553–566, 2002.

[13] J. M. Dickey. The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42:204–223, 1971.

[14] B. N. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79:892–898, 1984.

[15] B. N. Flury. Asymptotic theory for common principal component analysis. *The Annals of Statistics*, 14:418–430, 1986.

[16] B. N. Flury. *Common principal components and related multivariate models*. John Wiley & Sons, 1988.

[17] A. E. Gelfand and S. K. Ghosh. Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85:1–11, 1998.

[18] D. M. Hawkins. Comment on "A review and analysis of the Mahalanobis-Taguchi system" (Pkg: p1-30). *Technometrics*, 45(1):25–29, 2003.

[19] J. W. Hogan and N. M. Laird. Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, 16:259–272, 1997.

[20] X. Lin, J. Raz, and S. D. Harlow. Linear mixed models with heterogeneous within-cluster variances. *Biometrics*, 53:910–923, 1997.

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

[22] A. W. Marshall and I. Olkin. *Inequalities – Theory of majorization and its applications*. Academic Press, 1979.

[23] J. C. Pinheiro and D. M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6:289–296, 1996.

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

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

[26] M. Pourahmadi, M. J. Daniels, and T. Park. Simultaneous modelling of the cholesky decomposition of several covariance matrices with applications. *Submitted*, 2005.

[27] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall Ltd, 1999.

[28] A. F. M. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (Disc: p53-102). *Journal of the Royal Statistical Society, Series B, Methodological*, 55:3–23, 1993.

[29] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (Pkg: p583-639). *Journal of the Royal Statistical Society, Series B, Methodological*, 64(4):583–616, 2002.

[30] D. Sun and S. Ni. Bayesian analysis of vector-autoregressive models with noninformative priors. *Journal of Statistical Planning and Inference, in press*, 2004.

[31] A. P. Verbyla. Modelling variance heterogeneity: Residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society, Series B, Methodological*, 55:493–508, 1993.

[32] I. Verdinelli and L. Wasserman. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90:614–618, 1995.

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