Running head: WEAKLY INFORMATIVE PRIOR FOR COVARIANCE MATRICES

Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models

Yeojin Chung

Kookmin University, Seoul, South Korea

Andrew Gelman

Columbia University, New York City, USA

Sophia Rabe-Hesketh

University of California, Berkeley, USA

Jinchen Liu

Columbia University, New York City, USA

Vincent Dorie

New York University, New York City, USA

Abstract

When fitting hierarchical regression models, maximum likelihood estimation has computational (and, for some users, philosophical) advantages compared with full Bayesian inference, but when the number of groups is small, estimates of the covariance matrix (Σ) of group-level varying coefficients are often degenerate. One can do better, even from a purely point-estimation perspective, by using a prior distribution or penalty function. In this paper, we use Bayes modal estimation to obtain positive definite covariance matrix estimates. We recommend a class of Wishart (not inverse-Wishart) priors for Σ with a default choice of hyperparameters: the degrees of freedom are set equal to the number of varying coefficients plus 2, and the scale matrix is the identity matrix multiplied by a value that is large relative to the scale of the problem. This prior is equivalent to independent gamma priors for the eigenvalues of Σ with shape parameter 1.5 and rate parameter close to 0. It is also equivalent to independent gamma priors for the variances with the same hyperparameters multiplied by a function of the correlation coefficients. With this default prior, the posterior mode for Σ is always strictly positive definite. Furthermore, the resulting uncertainty for the fixed coefficients is less underestimated than under classical maximum likelihood or restricted maximum likelihood. We also suggest an extension of our method that can be used when stronger prior information is available for some of the variances or correlations.

Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models

Hierarchical or mixed-effects regression models are increasingly popular in applied statistics and are Bayesian on two levels: A prior distribution is assigned to the varying coefficients, and the parameters of that prior distribution themselves are given a hyperprior. The family of models can be written in general terms as follows: data are in groups $j = 1, \ldots, J$. For each group j, there is a response vector $\boldsymbol{y}_j$ and two data matrices, X_j and Z_j, that have fixed and varying coefficients, respectively. The data model is $p(\boldsymbol{y}_j|X_j\boldsymbol{\beta} + Z_j\boldsymbol{b}_j)$, where $\boldsymbol{\beta}$ is the vector of fixed coefficients and $\boldsymbol{b}_j$ is the vector of regression coefficients that varies by group. The vectors $\boldsymbol{b}_j$ are modeled as independent draws from a prior distribution, $p(\boldsymbol{b}_j)$, given some hyperparameters. We shall assume a normal model for the varying coefficients, so that $\boldsymbol{b}_j \sim N(\mathbf{0}, \Sigma)$. The model could also include a nonzero mean vector or a group-level regression structure for the hyperprior distribution, but these can be folded into the fixed coefficients in the data model without loss of generality.

There is a rich literature on full Bayesian inference for hierarchical regressions. There is also an empirical Bayes version in which the hyperparameters (in this case, Σ) are estimated via maximum likelihood and then inference for the coefficients is performed conditional on the estimated Σ. From the Bayesian perspective, the empirical Bayes approach is suboptimal, both because it avoids the use of any prior information on Σ and because it understates posterior uncertainty. From a pragmatic perspective, however, we recognize that the point estimation approach has two advantages that give it great appeal to many users. First, existing software such as `lme4` in R and various commands in Stata allow such models to be fit fast and reliably for moderate-sized datasets, whereas MCMC software for full-Bayes inference is not yet so immediately practical. Second, the non-Bayesian motivation behind point estimation is attractive to practitioners who want the benefits of partial pooling and hierarchical modeling without needing to specify prior information or fully buy into the Bayesian paradigm.

The subject of the present article, as with its predecessor on varying-intercept models with constant coefficients (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013), is the use of Bayesian ideas and methods to produce better inferences for hierarchical models via better point estimates of the hyperparameters. In that sense, this work falls into a long tradition of Bayesian tools used for practical non-Bayesian inferences (e.g., Agresti & Coull, 1998). Bayes modal estimation (or penalized likelihood) has also been used to obtain more stable estimates in item response theory (e.g., Swaminathan & Gifford, 1985; Mislevy, 1986; Tsutakawa & Lin, 1986) and to avoid boundary estimates (or logit parameters tending to $\pm\infty$) in log-linear models (Galindo-Garre, Vermunt, & Bergsma, 2004), logistic regression (Gelman, Jakulin, Pittau, & Su, 2008), and latent class analysis (Maris, 1999; Galindo-Garre & Vermunt, 2006). Such an approach has also been used to obtain non-degenerate covariance matrices in finite mixtures of normal densities (Ciuperca, Ridolfi, & Idier, 2003; Vermunt & Magidson, 2005) and in multivariate regression (Warton, 2008).

The key problem solved by our method is the tendency of maximum likelihood estimates of Σ to be degenerate, that is, on the border of positive-definiteness, which corresponds to zero variance or perfect correlation among some linear combinations of the parameters. When the maximum likelihood estimate of a hierarchical covariance matrix is degenerate, this often arises from a likelihood that is nearly flat in the relevant dimension and just happens to have a maximum at the boundary.

Our solution is a class of weakly informative prior densities for Σ that go to zero on the boundary as Σ becomes degenerate, thus ensuring that the posterior mode (i.e., the maximum penalized likelihood estimate) is always nondegenerate. We recommend a class of Wishart priors with a default choice of hyperparameters: the degrees of freedom is the dimension of $\boldsymbol{b}_j$ plus two and the scale matrix is the identity matrix multiplied by a large enough number. This prior can be expressed as a product of gamma$(1.5, \theta)$ priors on the eigenvalues of Σ or as a product of gamma$(1.5, \theta)$ priors on variances of the varying effects with rate parameter $\theta \rightarrow 0$ and a function of the correlations (a beta prior in the two-dimensional case). In the varying-intercept model (Chung, Rabe-Hesketh, Dorie, et al., 2013) and random-effects meta-analysis model (Chung, Rabe-Hesketh, & Choi, 2013), the gamma$(1.5, \theta)$ prior successfully avoids boundary estimates while producing estimates that are consistent with the data. We show that this is also true for the default Wishart prior proposed in this paper for general varying coefficient models.

In a simulation study and an education example presented below, the default Wishart prior always gives nondegenerate estimates of Σ (in particular, non-perfect correlation coefficients) without decreasing the log-likelihood substantially. The standard deviations and the correlation between random effects estimators using the Wishart prior have better statistical properties than using (restricted) maximum likelihood.

When prior information is available for specific standard deviations or correlations, additional penalty functions may be included. Specifically, if the prior most plausible value for a standard deviation or correlation parameter is σ^* or ρ^* respectively, then we propose multiplying the Wishart prior by the gamma$(2, 2/\sigma^*)$ or $N(\rho^*, 0.25^2)$ densities to assign more prior probability around the preferred values while exploiting the property of the Wishart prior that it ensures that the estimates remain positive definite.

The outline of the paper is as follows. First, we illustrate the boundary estimation problems encountered in maximum likelihood estimation of hierarchical variance and covariance parameters. Then we introduce the default Wishart prior for Σ and investigate its properties. Next, additional penalty functions are proposed that incorporate further prior knowledge for some of the parameters. Finally, our method is applied to an example from education research and simulated data.

Boundary estimation problem

Consider the varying-coefficients model,

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{z}_{ij}^T \boldsymbol{b}_j + \epsilon_{ij}, \quad i = 1, \ldots, n_j, \ j = 1, \ldots, J, \tag{1}$$

where y_{ij} is the response variable for unit i in group j, $\boldsymbol{x}_{ij}$ is a p-dimensional covariate vector with constant coefficients $\boldsymbol{\beta}$, $\boldsymbol{z}_{ij}$ is a d-dimensional data vector with varying coefficients $\boldsymbol{b}_j \sim N(\boldsymbol{0}, \Sigma)$, and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is a residual for each observation. We further assume that $\boldsymbol{b}_j$ and ϵ_{ij} are independent.

Non-Bayesian point estimation

For each j, $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{n_j j}) \sim N(X_j \boldsymbol{\beta}, V_j)'$, where X_j is a $n_j \times p$ matrix with $\boldsymbol{x}_{ij}^T$ in the ith row, $V_j = Z_j \Sigma Z_j^T + \sigma_\epsilon^2 I$, and Z_j is a $n_j \times d$ matrix with $\boldsymbol{z}_{ij}^T$ in the ith row. The log-likelihood function is

$$\log p(y|\boldsymbol{\beta}, \Sigma, \sigma_\epsilon^2) = -\frac{1}{2}\left[\sum_{j=1}^{J} \log |V_j| + \sum_j (\boldsymbol{y}_j - X_j\boldsymbol{\beta})^T V_j^{-1} (\boldsymbol{y}_j - X_j \boldsymbol{\beta})\right] \tag{2}$$

where the constant term, $-(N/2)\log(2\pi)$, has been dropped. The maximum likelihood (ML) estimator is obtained by maximizing the log-likelihood function.

It is known that the ML estimator of the covariance matrix is biased for finite samples (Lehmann & Casella, 1998), and an often-preferred option is restricted maximum likelihood or REML (Patterson & Thompson, 1971), as it takes into account the degrees of freedom for the fixed coefficients $\boldsymbol{\beta}$. Harville (1974) showed that the REML estimator can be derived by specifying flat prior distributions for $\boldsymbol{\beta}$, marginalizing over $\boldsymbol{\beta}$, and maximizing the marginal (or restricted) likelihood with respect to Σ and σ_ϵ^2. The restricted log-likelihood function is given by

$$\begin{aligned}\log p_R(y|\Sigma,\sigma_\epsilon^2) &= -\frac{1}{2}\left[\log\left|\sum_{j=1}^{J} X_j^T V_j^{-1} X_j\right| + \sum_{j=1}^{J}\log|V_j|\right. \\ &\quad \left. + \sum_{j=1}^{J}(\boldsymbol{y}_j - X_j\hat{\boldsymbol{\beta}})^T V_j^{-1}(\boldsymbol{y}_j - X_j\hat{\boldsymbol{\beta}})\right]. \qquad (3)\end{aligned}$$

up to constant where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{j=1}^{J} X_j^T V_j^{-1} X_j\right)^{-1}\left(\sum_{j=1}^{J} X_j^T V_j^{-1}\boldsymbol{y}_j\right).$$

Singular estimates of Σ using ML and REML

ML and REML often yield singular (that is, non-positive-definite) estimates of Σ. This boundary includes the cases where some varying coefficients have zero variance or a varying coefficient is a linear combination of the other varying coefficients.

We present two simulation studies to demonstrate how often singular estimates of Σ occur in the varying-coefficients model. In the first study, we consider a model with two-dimensional varying coefficients: a varying intercept b_{0j} and a varying slope b_{1j}. We set the group size to $n = 10$ and the number of groups to $J = 5$ or 10. A covariate that varies within group only was generated from $N(0, 1)$ and group-mean centered. The varying coefficients (b_{0j}, b_{1j}) were generated from $N(\mathbf{0}, \sigma^2 I_2)$ with $\sigma = 0.25, 0.5, 0.75, 1$. Setting the correlation to 0 corresponds to the best-case scenario in the sense of being furthest from the boundary. The within-group variance σ_ϵ^2 was set to 1 and the fixed coefficients β_0 and β_1 were set to 0. For each of 1000 random samples of data from the model, we obtained ML and REML estimates using `lmer` (Bates & Maechler, 2010) in R.

Figure 1a shows the proportion of ML estimates of Σ on the boundary for the two-dimensional case. For $J = 5$ groups, 87% of ML estimates are singular when $\sigma = 0.25$

and the proportion decreases as σ increases but remains as high as 72% when $\sigma = 1$. For $J = 10$ groups, the proportions are smaller than those for $J = 5$ but still, in more than 40% of simulations, the likelihood is maximized at a singular $\hat{\Sigma}$. The REML estimator yields smaller proportions of singular estimates with a similar trend (not shown). For $J = 10$, 79% and 64% of REML estimates are singular when $\sigma = 0.25$ and $\sigma = 1$, respectively. For $J = 10$, the proportion is reduced to 69% and 35% when $\sigma = 0.25$ and $\sigma = 1$, respectively.

Our second simulation study considers various dimensions, from $d = 2$ to $d = 5$, each time with a varying intercept and $d - 1$ varying slopes for $n = 10$ and $J = 5$ or 10. The $d - 1$ covariates were independently drawn from $N(0, 1)$ and centered at their group means as in the previous simulation. The varying coefficients $\boldsymbol{b}_j$ were drawn from $N(\mathbf{0}, I_d)$ and σ_ϵ^2 was set to 1. Figure 1b presents the proportion of replicates where the ML estimate $\hat{\Sigma}$ is singular. As the number of dimensions increases, this proportion increases rapidly, exceeding 95% with five varying coefficients for both $J = 5$ and $J = 10$. For REML, the proportions of singular estimates are slightly lower than for ML but follow a similar pattern and exceed 35% across all simulation conditions.

Weakly informative Wishart prior for Σ

We propose posterior modal estimation with a prior on Σ, implicitly assuming uniform priors for the other parameters. With a prior $p(\Sigma)$, the log posterior function can be written as

$$\log p(\boldsymbol{\beta}, \Sigma, \sigma_\epsilon | \boldsymbol{y}) = \log p(y | \boldsymbol{\beta}, \Sigma, \sigma_\epsilon) + \log p(\Sigma) + c, \tag{4}$$

and we find the mode of $\log p(\boldsymbol{\beta}, \Sigma, \sigma_\epsilon | \boldsymbol{y})$. This approach can also be viewed as maximum penalized likelihood estimation where the $\log p(\Sigma)$ is a penalty function. We consider a family of Wishart (*not* inverse-Wishart) densities for the prior on Σ. The Wishart density function on Σ with hyperparameters ν and Ψ is defined by

$$p(\Sigma) = \frac{|\Sigma|^{(\nu-d-1)/2}\exp[-\frac{1}{2}\mathrm{tr}(\Psi^{-1}\Sigma)]}{2^{\nu d/2}|\Psi|^{\nu/2}\Gamma_d(\nu/2)}, \nu > d-1, \Psi > 0 \tag{5}$$

where $\Gamma_d(\nu/2) = \pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma(\nu/2 + (1-j)/2)$, ν is the degrees of freedom, and Ψ is a scale matrix with $E(\Sigma) = \nu\Psi$.

If we set Ψ to be a diagonal matrix $(1/2\theta)I_d$, the Wishart density of Σ in (5) can be written as

$$\begin{aligned}
p(\Sigma) &= \frac{\theta^{d\nu/2}}{\Gamma_d(\nu/2)}|\Sigma|^{(\nu-d-1)/2}\exp\left(-\theta\mathrm{tr}(\Sigma)\right) && (6)\\
&= \frac{\theta^{d\nu/2}}{\Gamma_d(\nu/2)}\prod_{r=1}^{d}\lambda_r^{(\nu-d-1)/2}\exp\left(-\theta\lambda_r\right) && (7)\\
&\propto \prod_{r=1}^{d} g\left(\lambda_r \left| \frac{\nu-d+1}{2}, \theta\right.\right), && (8)
\end{aligned}$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of Σ and $g(x|\alpha,\theta)$ is the gamma(α,θ) density with shape parameter α and rate parameter θ, $g(x|\alpha,\theta) = \frac{\theta^{\alpha-1}}{\Gamma(\alpha)}x^{\alpha-1}\exp(-x\theta)$. In the above equations, note that we do not transform the density of Σ to the density of eigenvalues, but just rewrite (5) as a function of eigenvalues without including a Jacobian term.

As a default choice, we propose $\nu = d+2$ and $\theta \to 0$. As $\theta \to 0, P(\Sigma)$ approaches zero. In practice, we can choose a sufficiently small number for θ, for example $\theta = 10^{-4}$. This is proportional to independent gamma$(1.5,\theta)$ densities of the eigenvalues as observed in (8). If Σ is a diagonal matrix, this prior implies gamma$(1.5,\theta)$ priors on the diagonal elements of Σ, which is equivalent to gamma$(2,\theta)$ priors on the standard deviations when $\theta \to 0$. If Σ is not diagonal, we obtain gamma$(1.5,\theta)$ priors on the variances and a function of the correlations.

The advantage of this family of density functions is that they equal zero at the boundary—thus the Bayes modal or penalized likelihood estimate for Σ will never be degenerate—but the densities move away from zero when off the boundary, so that the posterior mode can be arbitrarily close to degeneracy if this is what the data demand. In contrast, various other families of models do not have these properties, making them less desirable when used for the purpose of Bayes modal point estimation. The inverse-Wishart family of density, one of the most commonly used priors for Σ in the full Bayesian inference, is have also zero at the boundary. However, it tends to assign an excessive penalty near the boundary because it is a function of Σ^{-1} and $|\Sigma|^{-1}$ while the Wishart density is a function of Σ and $|\Sigma|$.

Priors on the covariance matrix in the varying-coefficients model have been investigated by several authors in the context of full Bayesian modeling. Daniels and Kass (1999) investigated nonconjugate Bayesian estimation of covariance matrices in hierarchical models including an inverse Wishart prior on covariance matrices with unknown scale and degrees of freedom and a normal prior on Fisher's z-transformed correlations. Barnard, McCulloch, and Meng (2000) decomposed $\Sigma = \mathrm{Diag}(\boldsymbol{s})\, R\, \mathrm{Diag}(\boldsymbol{s})$ where $\boldsymbol{s}$ is a vector of standard deviations and R is the correlation matrix, which is assigned marginal or jointly uniform priors. O'Malley and Zaslavsky (2005) propose a scaled inverse-Wishart, a decomposition similar to that of (Barnard et al., 2000) except that the central matrix R itself has an inverse-Wishart distribution rather than being constrained to be a covariance matrix. In addition, nonnegative definite covariance matrix estimators are also suggested by Srivastava and Kubokawa (1999) and Amemiya (1985), which allow $|\hat{\Sigma}|$ to be zero. Our approach is different from these others in being explicitly intended not for full Bayes inference but as a tool to obtain positive definite posterior modal estimates. As such, our concerns are different from those involved in constructing traditional Bayesian priors.

Unlike posterior mean estimation, Bayes modal estimation does not involve simulation and is computationally as efficient as maximum likelihood estimation. By modifying existing maximum likelihood estimation procedures, `gllamm` (Rabe-Hesketh, Skrondal, & Pickles, 2005) in Stata and `lmer` (Bates & Maechler, 2010) in R, we have developed software to find the maximum of the penalized likelihood. The modified `gllamm` is available from www.gllamm.org and `blmer`, the modified `lmer` function, can be found in the `blme` package available from the Comprehensive R Archive Network.

Varying-intercept models: $d = 1$

The varying-intercept model is a special case of the model in (1) with $d = 1$, given by

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + b_j + \epsilon_{ij}$$

where $b_j \sim N(0, \sigma_b^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The Wishart prior in (7) is equivalent to a gamma$(\nu/2, \theta)$ prior on σ_b^2. With the default choice of hyperparameters, $\nu = 3(= d + 2)$ and $\theta \to 0$, the Wishart prior coincides with a gamma$(1.5, \theta)$ prior on σ_b^2.

When $\theta \to 0$, the gamma$(1.5, \theta)$ prior on σ_b^2 has a density function proportional to σ_b. The gamma$(2, \theta)$ prior on σ_b is recommended as a weakly informative prior for avoiding estimates of σ_b equal to zero in the varying-intercept model (Chung, Rabe-Hesketh, Dorie, et al., 2013) and in random-effects meta-analysis models (Chung, Rabe-Hesketh, & Choi, 2013). Since the gamma$(2, \theta)$ prior is 0 at $\sigma_b = 0$, the posterior density is also 0 at $\sigma_b = 0$ and thus the posterior mode of σ_b is always strictly positive. In addition, since the gamma density has a positive constant derivative at $\sigma_b = 0$, the gamma$(2, \theta)$ density increases linearly at zero. It follows that the profile likelihood of σ_b (maximized over all the other parameters) dominates the posterior density of σ_b if the likelihood is strongly curved near $\sigma_b = 0$. That is, the prior does not rule out positive values near zero if they are supported by the likelihood. Chung, Rabe-Hesketh, Dorie, et al. (2013) show that the posterior mode is approximately one standard error away from zero when the ML estimate of σ_b is zero. Finally, the estimator behaves reasonably well in terms of mean squared error of parameter estimates and coverage of confidence intervals for fixed parameters.

In the context of small area estimation, strictly positive group-level variance estimators have been proposed for the Fay-Herriot model (1979), a varying-intercept model for aggregated group-level data and known heterogeneous within-group variances. *Adjustment for density maximization* (Morris, 2006; Li & Lahiri, 2010; Morris & Tang, 2011) applies a penalty term $\pi(\sigma_b^2) = (\sigma_b^2)^{c-1}$ to the likelihood, and this approach turns out to be equivalent to posterior modal estimation with a gamma(α, θ) prior on σ_b with $\alpha = 2c + 1$ and $\theta \to 1$. Therefore, for this specific varying-intercept model, our estimator shares the properties of adjustment for density maximization, such as predictions of the group means being minimax for mean squared-error loss when the within-group variances are equal and $c \leq 1$ (Morris & Tang, 2011).

Varying-intercept and varying-slope models: $d = 2$

When $d = 2$, the model includes a varying intercept and a varying slope of one covariate, written as

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + b_{0j} + b_{1j} z_{ij} + \epsilon_{ij}$$

where $(b_{0j}, b_{1j}) \sim N(\mathbf{0}, \Sigma)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

As shown in (8), with the default choice $\nu = d + 2$, the Wishart density can be written as a product of gamma$(1.5, \theta)$ densities on the eigenvalues λ_1 and λ_2. For the bivariate case, we can also express the default prior as a function of the variances (σ_1^2 and σ_2^2) and the correlation (ρ) between the two varying effects b_{0j} and b_{1j}, given by

$$p\,(\Sigma) \propto |\Sigma|^{1/2} = \sigma_1 \sigma_2 \sqrt{1 - \rho^2}. \tag{9}$$

This expression implies that Wishart$(4, (1/2\theta) I_d)$ with $\theta \to 0$ is equivalent to the joint density of independent gamma$(1.5, \theta)$ priors on both σ_1^2 and σ_2^2, and a beta(1.5,1.5) prior on $(\rho + 1)/2$.

Since the beta(1.5,1.5) prior on $(\rho + 1)/2$ is zero at the boundaries $\rho = \pm 1$, the posterior mode of Σ cannot be attained at any matrices with perfect correlation. In addition, the beta(1.5,1.5) density function increases rapidly as ρ approaches 0 from ± 1 and so does not rule out values close to ± 1. The left panel of Figure 2 shows the beta(1.5,1.5) density on $(\rho + 1)/2$. While gamma(2,θ) increases linearly at 0, the slopes of beta(1.5,1.5) at ± 1 are $\pm\infty$. Therefore, compared to the gamma$(2, \theta)$ prior for σ_1 and σ_2, the beta(1.5,1.5) for ρ is less informative with lower penalties on the values around the boundaries.

Beta priors have been used to avoid boundary estimates of the sample proportion to estimate the probability parameter p of the binomial distribution. When $\hat{p}$ is 0 or 1, the Wald confidence interval for p degenerates to the point estimate. To avoid such boundary estimates, (Agresti & Coull, 1998) specified a beta(2,2) prior on p. The posterior mean of p then is the sample proportion after adding two successes and two failures to the data. Compared with the beta(2,2), the beta(1.5,1.5) tends to assign less penalty at the boundaries and so is less informative.

Higher dimensional case: $d \geq 3$

Similar to the case $d = 2$, the default prior for $d \geq 3$ can be written as a product of $\sigma_r,\ r = 1, \ldots, d$ and a function of ρ_{rs}, the correlation between the r-th and s-th varying effects $(0 < r < s,\ s = 2, \ldots, d)$. For example with $d = 3$, the Wishart$(5, (1/2\theta)I_3)$ prior with $\theta \to 0$ can be written as

$$p(\Sigma) \propto |\Sigma|^{1/2} \propto \sigma_1\sigma_2\sigma_3\sqrt{1 - \rho_{12}^2 - \rho_{23}^2 - \rho_{13}^2 + 2\rho_{12}\rho_{23}\rho_{13}}. \tag{10}$$

This is a product of gamma$(1.5, \theta)$ priors on the variances and a function of the correlations. This function depends on the squares of the correlations, as in the two-dimensional case (9), but also contains the product of three correlations, which comes from the constraint $|\Sigma| > 0$ that defines the support of Wishart distributions. Because of this constraint, the Wishart prior automatically restricts the posterior mode of Σ to be strictly positive definite.

The graphs in Figure 2 show the conditional densities of ρ_{12} when Σ follows the Wishart$(d+2, (1/2\theta)I_d),\ \theta = 10^{-4}$. The curves are the density of ρ_{12} conditional on the other parameter values (standard deviations and the other correlations) that are randomly generated from Wishart$(d+2, (1/2\theta)I_d)$ with 20 replicates. When $d = 2$, the correlation follows beta$(1.5, 1.5)$ as discussed before. When $d = 3$, the curves have distinct supports, defined by $1 - (\rho_{12})^2 - (\rho^0_{23})^2 - (\rho^0_{13})^2 + 2\rho_{12}\rho^0_{23}\rho^0_{13} > 0$ where ρ^0_{13} and ρ^0_{23} for each replicate are given by randomly generated Σ. The curves for $d = 5$ are more scattered and the supports of the densities tend to be narrower than for $d = 2$ and 3 due to more restrictions required for the higher dimensional Σ to be positive definite.

The marginal prior densities of ρ_{rs} are displayed in Figure 3 for $d = 2, 5$, and 10. With 10000 replicates, d-dimensional matrices were randomly generated from the Wishart$(d+2, (1/2\theta)I)$ with $\theta = 10^{-4}$ and $10000(d-1)(d-2)/2$ correlation coefficients were used to construct the histograms. For $d = 2$ (left), the distribution of the correlation coefficient matches the beta(1.5,1.5) density, shown as a solid curve. As d increases, the marginal prior density of ρ_{rs} becomes more concentrated around zero because of the positive definiteness of Σ.

Incorporating additional prior information

In the previous section, we suggested the Wishart$(d+2, (1/2\theta)I)$ with $\theta \to 0$ as a default prior when no other information is available. If a researcher has additional prior knowledge about any specific standard deviations or correlations, he or she might want to adjust the prior to incorporate such information. In this section, we suggest multiplying the Wishart prior by functions of the parameters on which we have information. Because the Wishart density ensures that Σ is positive definite, we can choose the functions for the other parameters to be intuitive and easy to specify without regard for the parameter space.

If σ^* is a plausible value for σ_r, then the gamma$(2, 2/\sigma^*)$ density is recommended as a penalty. Recall that the default Wishart prior is proportional to gamma$(2, \theta)$ priors with $\theta \to 0$ on each standard deviation, multiplied by a function of the correlations. When the gamma$(2, 2/\sigma^*)$ density of σ is multiplied by the Wishart, the part including σ_r becomes $\sigma_r^2 \exp(-2\sigma_r/\sigma^*)$. This is proportional to the gamma$(3, 2/\sigma^*)$ density that has its mode at $\sigma_r = \sigma^*$. The gamma prior with shape parameter greater than two assigns more penalty near zero than for shape parameter equal to two. Therefore we have a more informative prior with mode at σ^*.

If any specific correlation ρ_{rs} is believed to be close to ρ^*, we can incorporate this prior information by multiplying the default Wishart prior by a N(ρ^*, τ^2) density. As usual, the scale parameter τ can be chosen depending on the prior uncertainty regarding ρ_{rs}. A possible default choice is $\tau = 0.25$ because it is the standard deviation of the beta(1.5,1.5) distribution. Figure 4 displays the shape of conditional prior densities of ρ_{12} with additional normal priors in the three dimensional case. When ρ_{13} and ρ_{23} are fixed at zero (left), the default Wishart$(5, (1/2\theta)I_3)$ prior (solid curve) is pretty flat. In order to incorporate the prior information, for example $\rho^* = -0.5$, the Wishart is multiplied by the $N(-0.5, 0.25^2)$ density, and then the prior mode moves toward -0.5 (dashed curve). When ρ_{13} and ρ_{23} are 0.5 (middle and right), the support of the Wishart for ρ_{12} is on $[-0.5, 1]$ because of the constraint of positive definiteness. When our prior value is on the boundary $\rho^* = -0.5$ (middle), the Wishart multiplied by $N(-0.5, 0.25^2)$ density is skewed toward -0.5, but still enforces positive definiteness. When the prior value is inside the support, $\rho^* = 0.5$, the resulting density is less skewed (right).

The default prior for ρ in the two dimensional case is beta(1.5,1.5), and so it would seem natural to use the beta family for ρ_{rs} when constructing an additional penalty. However, the parameters of the normal distribution are more intuitive because they represent the prior mean (and mode) and variance. In addition, since the positive definiteness of $\hat{\Sigma}$ is already guaranteed by the Wishart prior, estimates of Σ remain positive definite regardless of the type of additional penalties that multiply the Wishart prior. Furthermore, computation is no problem in any case; including any closed-form prior density adds essentially no cost to the optimization.

Example: a varying intercept, varying slope model in education research

We illustrate our approach using a study of Heller et al. (2007) on the effects of the Mathematics Pathways and Pitfalls (MPP) teacher professional development program on mathematics learning for students at different levels of English language proficiency. Half of 32 teachers were randomized to MPP and the other half to the control condition. Teachers randomized to the MPP condition were taught how to use the materials and then substituted MPP for part of their mathematics curriculum during the 2003–2004 school year, while control teachers used their regular mathematics curriculum. All students received an MPP test as a pre-test before the lessons and took the same test after the lessons as a post-test.

Post-test scores are regressed on the mean-centered pre-test scores, an indicator for treatment group (1 for MPP and 0 for control), English language learner (ELL) status (1 for ELL and 0 for non-ELL), and the treatment $\times$ ELL interaction term. A varying intercept and a varying slope for ELL status are included to allow for the cluster-randomized design. The model can be written as

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + b_{0j} + b_{1j} z_{ij} + \epsilon_{ij},$$

where y_{ij} is the post-test score for the i-th student of the j-th teacher, $\boldsymbol{x}_{ij}$ is the covariate vector that includes the mean-centered pre-test score, the treatment group indicator, ELL status, and the interaction between ELL status and treatment, and z_{ij} is ELL status. As usual, we assume $(b_{0j}, b_{1j}) \sim N(0, \Sigma)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. After dropping observations with missing values on any of the variables, data were available on 755 students and $J = 36$ teachers, with between 12 and 27 students per teacher. We fit the models by ML and REML using `lmer` in the `lme4` package and by BM using `blmer` in the `blme` package.

Table 1 includes parameter estimates with ML, REML, and the Bayes modal (BM) estimates with the default Wishart$(4, (1/2\theta)I_d)$ prior with $\theta = 10^{-4}$. Both ML and REML estimates of the correlation between b_{0i} and b_{1j} are -1. This implies an unrealistic perfect correlation between the teacher-level slopes and intercepts. The BM estimate of ρ is -0.32 and the standard deviation of the varying slope for ELL status increases from 0.71 for ML and 0.48 for REML to 3.64, a change that is within the uncertainty implied by the asymptotic standard error of 2.1 (ML) or 2.2 (REML) for that parameter. The standard deviation of the varying intercept stays similar for ML, REML, and BM.

The fixed coefficient estimates are similar across estimation methods. The coefficient for the interaction term between ELL and treatment changes the most among all the fixed coefficients, but the differences are negligible considering that the standard errors of the interaction term are greater than 4. The standard errors of the fixed coefficient estimates of Treatment, ELL, and Treatment by ELL are larger for BM than for ML or REML, suggesting that ML and REML underestimate the uncertainty.

The log likelihood at the BM estimates differs from the maximum by less than 1.

Figure 5 shows the profile likelihood of ρ (profiling out all the other parameters) divided by its maximum. Although the maximum likelihood is attained at $\rho = -1$, the profile likelihood is very flat and so the minimum (at $\rho = 1$) is attained with only an 8% decrement from the maximum. Therefore all the values of ρ including $\rho = -0.32$ are well supported by the data. As is typical in such settings, there is nothing special about the point estimate on the boundary, and it would be inappropriate for a researcher to use that estimate. Our Bayes modal approach gives a default procedure which allows a classical statistician to avoid the inappropriate degenerate estimate. A full Bayes approach using real prior information would do better, but Bayes modal approach takes us a bit in the right direction and has the advantage of being fast and easy to implement.

When a researcher is interested in comparing teacher-specific effects, b_{0j} and b_{1j} can be predicted using the conditional posterior mean (or mode) given the estimates of the model parameters and the data (called empirical Bayes prediction or best linear unbiased prediction.)

In Figure 6, scatter plots of empirical Bayes estimates of b_{1j} versus b_{0j} are displayed with the proportion of ELL students of each teacher represented by the gray scale: black indicates all the students are ELL and white indicates none are ELL. The sizes of the squares are proportional to the numbers of students for each teacher. For ML (left), due to the estimate $\hat{\rho} = -1$, the slopes b_{1j} are predicted perfectly linearly by the intercepts b_{0j}. In contrast, Bayes modal (right) shows more reasonable predictions for the varying slopes and intercepts. In addition, we can observe that 18 (out of 36) white squares with a gray border fall perfectly on a line—these are teachers without any ELL students in their classes. Four black squares (only three visible due to overlap) correspond to the teachers with only ELL students. The 18 groups without ELL students and the four groups with only ELL students do not provide any information about the slope variance and intercept variance, respectively, and none of the 22 groups provide information about the correlation between the varying slope and intercept. The lack of information could be one of the reasons we obtain the boundary estimates using ML and REML. As the group size increases (that is, the square increases) and the proportion of ELL students increases (that is, the square gets darker), the empirical Bayes predictions tend to be less shrunken toward the line formed by the white squares.

Using the fitted covariance matrix, we can calculate the marginal variances and correlations of the post-test score given their ELL status. The variance of the post-test scores for ELL students is $\mathrm{Var}(y_{ij}|z_{ij}=1) = \sigma_1^2 + \sigma_2^2 + 2\sigma_{12} + \sigma_\epsilon^2$ and, similarly, the variance for non-ELL student is $\mathrm{Var}(y_{ij}|z_{ij}=0) = \sigma_1^2 + \sigma_\epsilon^2$. The covariance between the post-test scores of two students of the same teacher is

$\mathrm{Cov}(y_{ij}, y_{i^*j}|z_{ij}=1, z_{i^*j}=1) = \sigma_1^2 + \sigma_2^2 + 2\sigma_{12}$ if both students are ELL,

$\mathrm{Cov}(y_{ij}, y_{i^*j}|z_{ij}=1, z_{i^*j}=0) = \sigma_1^2 + \sigma_{12}$ if one student is ELL, and

$\mathrm{Cov}(y_{ij}, y_{i^*j}|z_{ij}=0, z_{i^*j}=0) = \sigma_1^2$ if neither student is ELL.

Table 2 shows these model-implied marginal standard deviations and correlations with estimates from ML and BM substituted for the parameters. These standard deviation and correlation estimates are remarkably similar which also explains why the log-likelihood evaluated at the BM estimates is not much smaller than that evaluated at the ML estimates.

Simulation

We simulated data from the varying coefficient model as described in the preliminary simulation in the beginning of this paper but with only one covariate. We explored different values of the correlation ρ (0, 0.225, 0.450, 0.675, and 0.900), setting σ to be a moderate value of 0.5. With 1000 replicated samples generated in the same way as explained in Section with $J = 5$ and $n = 30$, we computed the bias and root mean squared error (RMSE) for σ_1, σ_2, and ρ. For ML and REML, the bias and RMSE of $\hat{\rho}$ are based on the replicates that generate legitimate estimates (that is, when neither $\hat{\sigma}_1$ nor $\hat{\sigma}_2$ is zero which happened in 1.2% of replicates for ML and 0.9% of replicates for REML). For Bayes modal estimation, we assigned a Wishart$(4, (1/2\theta)I)$ prior on Σ with $\theta = 10^{-4}$.

Figure 7 shows the proportion of boundary estimates of ρ $(1 - |\hat{\rho}| < 10^{-5})$. When ρ is 0, 21% of the ML estimates and 17% of the REML estimates have perfect correlations. As ρ increases, the proportion of $\hat{\rho}$ on the boundary also increases and reaches 60% for ML and 51% for REML. The BM method does not produce any boundary estimates of ρ for any of the simulation conditions.

In spite of no boundary estimates, the log-likelihood is not reduced substantially by using BM estimation. Investigating the difference in deviances $(= 2[\log L(\hat{\Sigma}_{\mathrm{ML}}) - \log L(\hat{\Sigma}_{\mathrm{BM}})])$ for all the replicates, the BM method does not reduce the log-likelihood by more than 2.2 from the maximum.

Figure 8 summarizes the bias and RMSE of $\hat{\rho}, \hat{\sigma}_1$, and $\hat{\sigma}_2$. When $\rho = 0$, the bias of $\hat{\rho}$ is almost zero for all three methods.

ML, REML, and BM all have some bias in estimating ρ, with BM having the most bias (that is, the most shrinkage toward 0), as would be expected given the regularization from the Wishart prior that squeezes $\hat{\rho}$ toward zero as seen in the shape of the prior density for ρ with $d = 2$ in Figure 3. However, BM gives the smallest RMSE of $\hat{\rho}$. The bias of $\hat{\sigma}_1$ and $\hat{\sigma}_2$ is similar across the different values of ρ for all the estimation methods. The BM estimates are less biased than ML and REML and the RMSE is smaller for BM.

The coverage of 95% confidence intervals for β_0 and β_1 does not change much with ρ. The average coverage of the BM confidence intervals is 0.940 for β_0 and 0.943 for β_1. The coverage by REML is about the same as that for BM, whereas ML shows slightly lower coverage with averages of 0.935 for β_0 and 0.937 for β_1.

Conclusion

For the hierarchical regression model, particularly with several varying coefficients, degenerate covariance matrix estimates do not have a practical interpretation but can commonly arise as point estimates in maximum likelihood estimation because there is often little information on these hyperparameters when performing inference for a hierarchical model with only a moderate number of groups. In addition, when $\hat{\Sigma}$ is singular, underestimated standard errors of the fixed coefficients make the researcher overconfident about the effect of the covariates. When a boundary estimate is attained but no prior information is available for Σ, the Bayes modal estimator using the default Wishart prior is recommended because it ensures strictly positive definite $\hat{\Sigma}$ and is weakly informative at the same time. The modified `gllamm` from www.gllamm.org for Stata and `blme` package for R enable easy application of our method for practitioners.

In varying-slope models, changing the location and scale of the covariates that have varying slopes implies that Σ must change to produce an equivalent model. For example, for longitudinal data, we might want to transform the time variable to have a value 0 at the initial time point. In this case, subtracting a constant from the covariate changes the variance of the varying intercepts and the correlation between intercepts and slopes. While ML and REML will yield equivalent models after linearly transforming the covariate, this is no longer true for Bayes modal estimation, which pulls the correlation towards 0. When using Bayesian regularization in this setting, it therefore becomes more important to choose meaningful centering points for the covariates with varying coefficients.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician, 52(2), 119–126.

Amemiya, Y. (1985). What should be done when an estimated between-group covariance matrix is not nonnegative definite? The American Statistician, 39(2), 112–117.

Barnard, J., McCulloch, R., & Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica, 10(4), 1281–1312.

Bates, D., & Maechler, M. (2010). lme4: linear mixed-effects models using s4 classes [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lme4 (R package version 0.999375-37)

Chung, Y., Rabe-Hesketh, S., & Choi, I.-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. Statistics in medicine, 32(23), 4071–4089.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. Psychometrika, 78(4), 685–709.

Ciuperca, G., Ridolfi, A., & Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. Skandinavian Journal of Statistics, 30, 45-59.

Daniels, M. J., & Kass, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. Journal of the American Statistical Association, 94(448), 1254-1263.

Fay, R. E., & Herriot, R. A. (1979, June). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366), 269–277.

Galindo-Garre, F., & Vermunt, J. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. Behaviormetrika, 33(1), 43–59.

Galindo-Garre, F., Vermunt, J., & Bergsma, W. (2004). Bayesian posterior mode estimation of logit parameters with small samples. Sociological Methods & Research, 33(1), 88–117.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 2, 1360-1383.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. Biometrika, 61(2), 383-385.

Lehmann, E. L., & Casella, G. (1998). Theory of point estimation. New York: Springer.

Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. Journal of multivariate analysis, 101(4), 882–892.

Maris, E. (1999). Estimating multiple classification latent class models. Psychometrika, 64(2), 187-212.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. Psychometrika, 51(2), 177-195.

Morris, C. (2006). Mixed model prediction and small area estimation (with discussions). Test, 15(1), 72–76.

Morris, C., & Tang, R. (2011). Estimating random effects via adjustment for density maximization. Statistical Science, 26(2), 271–287.

O'Malley, A. J., & Zaslavsky, A. M. (2005). Cluster-level covariance analysis for survey data with structured nonresponse (Tech. Rep.). Department of Health Care Policy, Harvard Medical School.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. Biometrika, 58(3), 545-554.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. Journal of Econometrics, 128(2), 301-323.

Srivastava, M., & Kubokawa, T. (1999). Improved nonnegative estimation of multivariate components of variance. Annals of Statistics, 27(6), 2008–2032.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. Psychometrika, 50(3), 349-364.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. Psychometrika, 51(2), 251-267.

Vermunt, J., & Magidson, J. (2005). Technical guide for latent gold 4.0: Basic and advanced (Tech. Rep.). Belmont Massachusetts: Statistical Innovations Inc.

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. Journal of the American Statistical Association, 103(481), 340–349.

Author Note

The research reported here was supported by the Institute of Education Sciences, U. S. Department of Education, through grant R305D100017 to University of California, Berkeley. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

 This data set is based upon work supported by the National Science Foundation under Grant No. 9911374, along with materials developed by WestEd. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and may not reflect the views, findings, or opinions of the National Science Foundation or WestEd.

	ML	REML	Bayes modal
Fixed effect			
Intercept	32.39 (2.01)	32.40 (2.07)	32.31 (2.11)
Pretest	0.56 (0.06)	0.56 (0.06)	0.56 (0.06)
Treatment	12.84 (3.15)	12.81 (3.24)	13.01 (3.30)
ELL	−2.46 (2.73)	-2.54 (2.77)	−2.66 (3.17)
ELL × Treatment	1.00 (4.13)	1.24 (4.19)	1.56 (4.84)
Varying effect (group: teacher)			
Intercept SD	8.31 (1.18)	8.62 (1.25)	8.50 (1.22)
ELL SD	0.71 (2.09)	0.48 (2.18)	3.64 (2.50)
Correlation	−1.00 (2.93)	−1.00 (0.00)	-0.32 (0.22)
Residual SD	226.5	227.3	226.3
Log likelihood	−3153.7	−3153.8	−3154.2

* The correlation is estimated on the arctan scale and the delta method is used for calculating the standard error.

Table 1

Parameter estimates for education example. The ML and REML estimates imply perfect correlation between the varying intercept and varying slope,whereas Bayes modal produces more reasonable estimates. The log-likelihood stays almost the same among the three methods. We present results here to more decimal places than would be recommended in practice in order to display the sometimes-small differences between the different estimates.

	ML	BM
Stdv. of ELL student	16.86	17.06
Stdv. of non-ELL student	17.19	17.26
Corr. of (ELL, ELL)	0.20	0.22
Corr. of (ELL, non-ELL)	0.22	0.21
Corr. of (non-ELL, non-ELL)	0.23	0.24

Table 2
Marginal standard deviations and correlations of post-test scores given ELL status. These values do not differ much between ML and BM although the slope standard deviation estimate and correlation estimate increased notably from ML to BM.

Figure Captions

Figure 1. Proportion of datasets, out of 1000, where the ML estimate of the covariance matrix is singular. (a) When $\sigma = 0.25$, 87% of ML estimates are singular for $J = 5$. As σ and J increase, the proportion decreases but is greater than 40% for the conditions considered. (b) As the dimension of Σ increases, there is a rapid increase in the probability of the estimate being degenerate.

Figure (a). Two-dimensional case

Figure (b). 2 to 5 dimensions; $\sigma = 1$

Figure 2. Conditional density of ρ_{ij} with Wishart$(d+2,(1/2\theta)I)$ on Σ, $\theta = 10^{-4}$, where the other parameters are randomly generated from the Wishart distribution for 20 replicates. When $d = 2$, the conditional density is beta(1.5,1.5), but for larger d, the curves are more scattered and the supports of the densities become narrower.

Figure 3. Marginal density of ρ_{rs} with Wishart$(d+2,(1/2\theta)I)$, $\theta = 10^{-4}$. When $d = 2$, the marginal density of ρ is equivalent to beta(1.5,1.5) on $(\rho+1)/2$ (solid curve). As d increases, the marginal density has more mass around 0 due to the positive semi-definite constraint of the covariance matrix.

Figure 4. Conditional prior density of ρ_{12} with additional $N(-0.5, 0.25^2)$ (left and middle) and $N(0.5, 0.25^2)$ (right) densities multiplying the default Wishart prior. The Wishart prior is on 3-dimensional Σ and ρ_{13} and ρ_{23} are fixed as 0 (left) and 0.5 (middle and right). The additional normal penalty makes the prior density skewed toward the prior value, but still enforces positive definiteness.

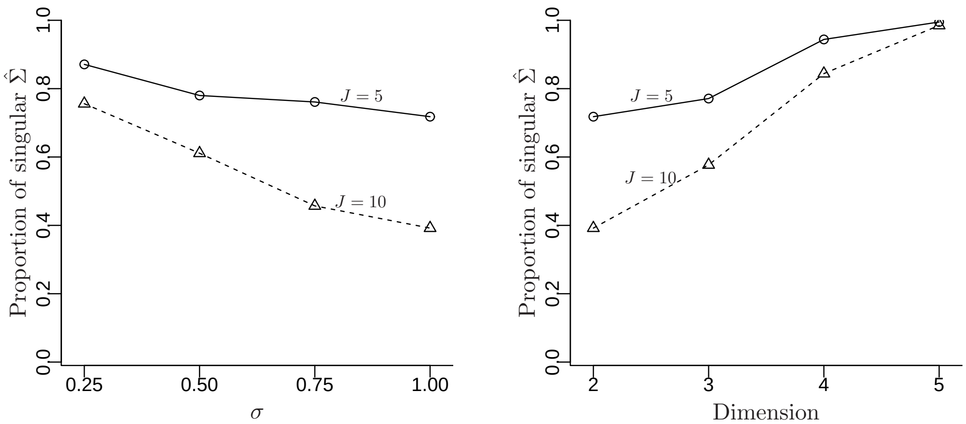
Figure 5. Profile likelihood of ρ. The ML estimate of ρ is -1 but the likelihood has very little information. Therefore the Bayes modal estimate of -0.3 is also well supported by

the data.

Figure 6. *Empirical Bayes predictions of varying effects. The size of each square represents* n_j *for the* j*-th teacher. The ratio of ELL students for each teacher is shown on a gray scale: black indicates all the students are ELL and white indicates none are ELL. For ML,* b_{1j} *are predicted perfectly linearly in* b_{0j}*. On the right graph, Bayes modal estimation shows more reasonable predictions for the varying slopes and intercepts.*

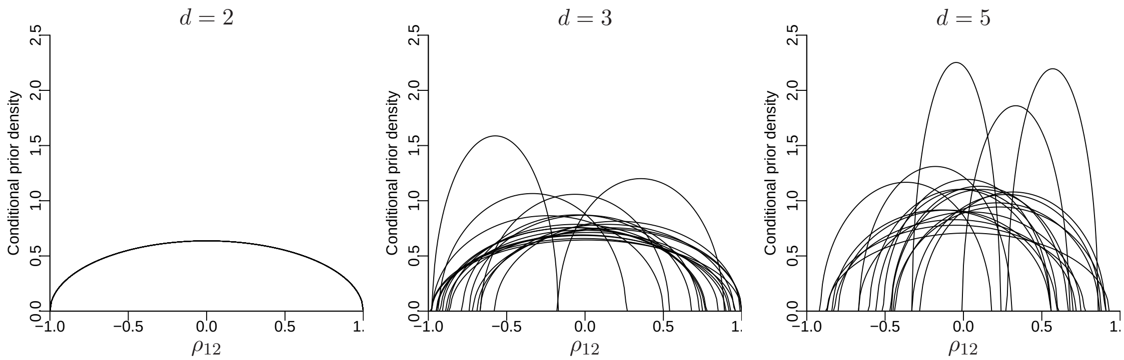
Figure 7. *Proportion of ML and REML estimates of* ρ *that are on the boundary. When* $\rho = 0$*, 21% of the ML estimates and 17% of the REML estimates are* ± 1*. As* ρ *increases, the proportion of estimates on the boundary,* $\hat{\rho}$ *equal to* ± 1*, also increases and reaches 60% for ML and 51% for REML when* $\rho = 0.9$*.*

Figure 8. *Bias and root mean squared error of* $\hat{\sigma}_1$*,* $\hat{\sigma}_2$ *and* $\hat{\rho}$ *with* $J = 5$ *and* $n = 30$ *of the varying-coefficient model with* $\sigma_1 = \sigma_2 = 0.5$ *and* ρ *in the grid. In our simulation, with* ρ *set to various positive values, the bias values are all negative, so we display absolute values to make the graphs easier to read given the convention that high values of bias are bad. BM has higher bias for* ρ *(that is, shrinking the estimate toward 0) compared to ML and REML, but the RMSE is smaller for BM. For both* σ_1 *and* σ_2*, BM has smaller bias and RMSE than ML and REML.*

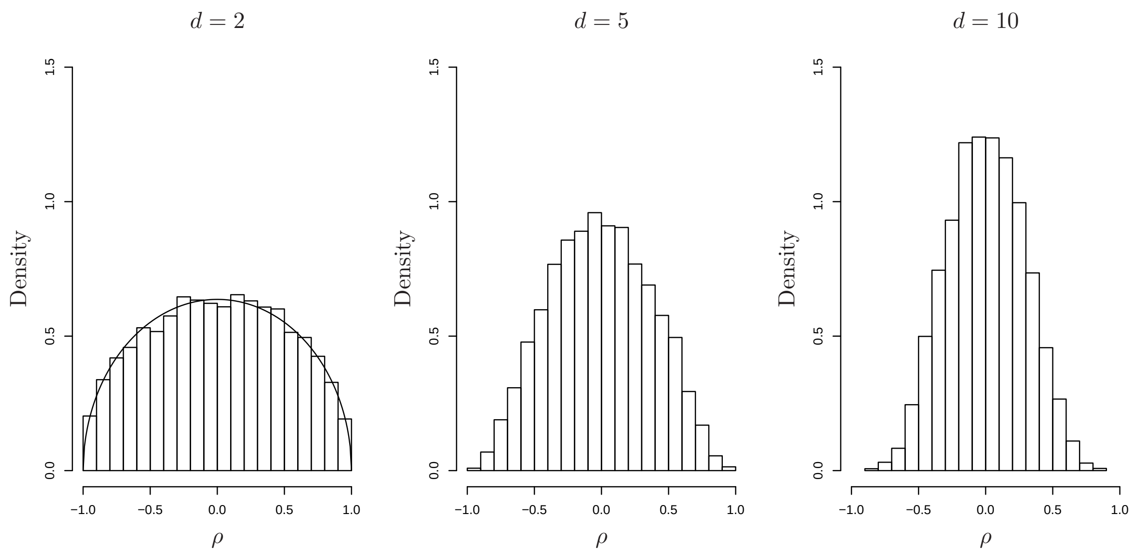


(a) Two-dimensional case

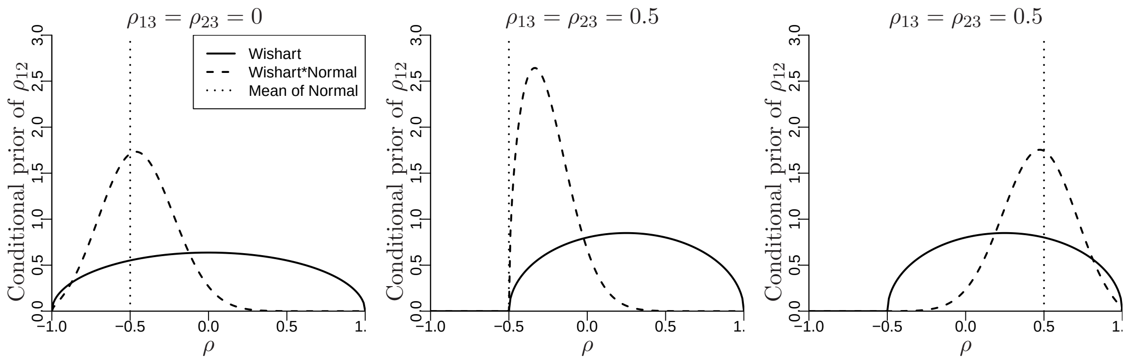
(b) 2 to 5 dimensions; $\sigma = 1$



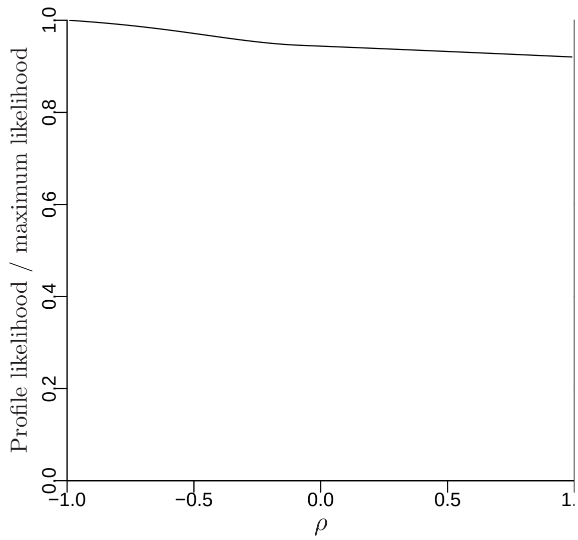
$d = 2$
$d = 3$
$d = 5$
Conditional prior density
ρ_{12}

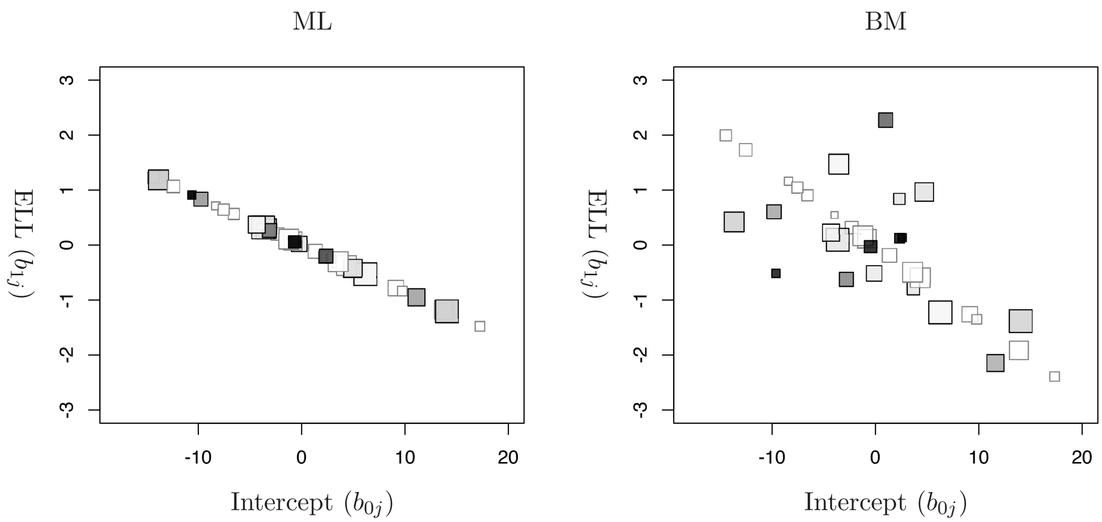


$d = 2$
$d = 5$
$d = 10$
Density
ρ
-1.0
-0.5
0.0
0.5
1.0
1.5

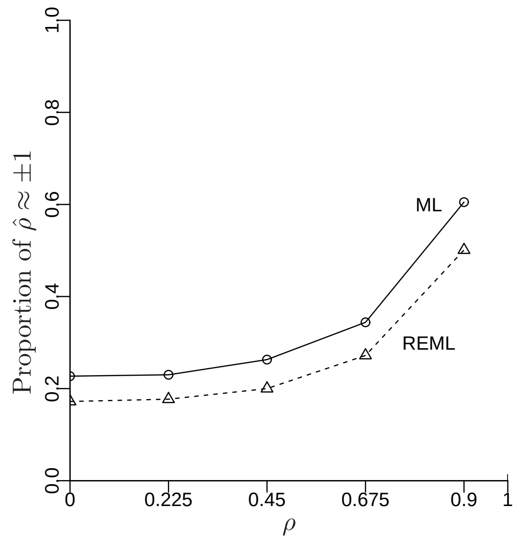


$\rho_{13} = \rho_{23} = 0$
$\rho_{13} = \rho_{23} = 0.5$
$\rho_{13} = \rho_{23} = 0.5$
Wishart
Wishart*Normal
Mean of Normal
Conditional prior of ρ_{12}
ρ

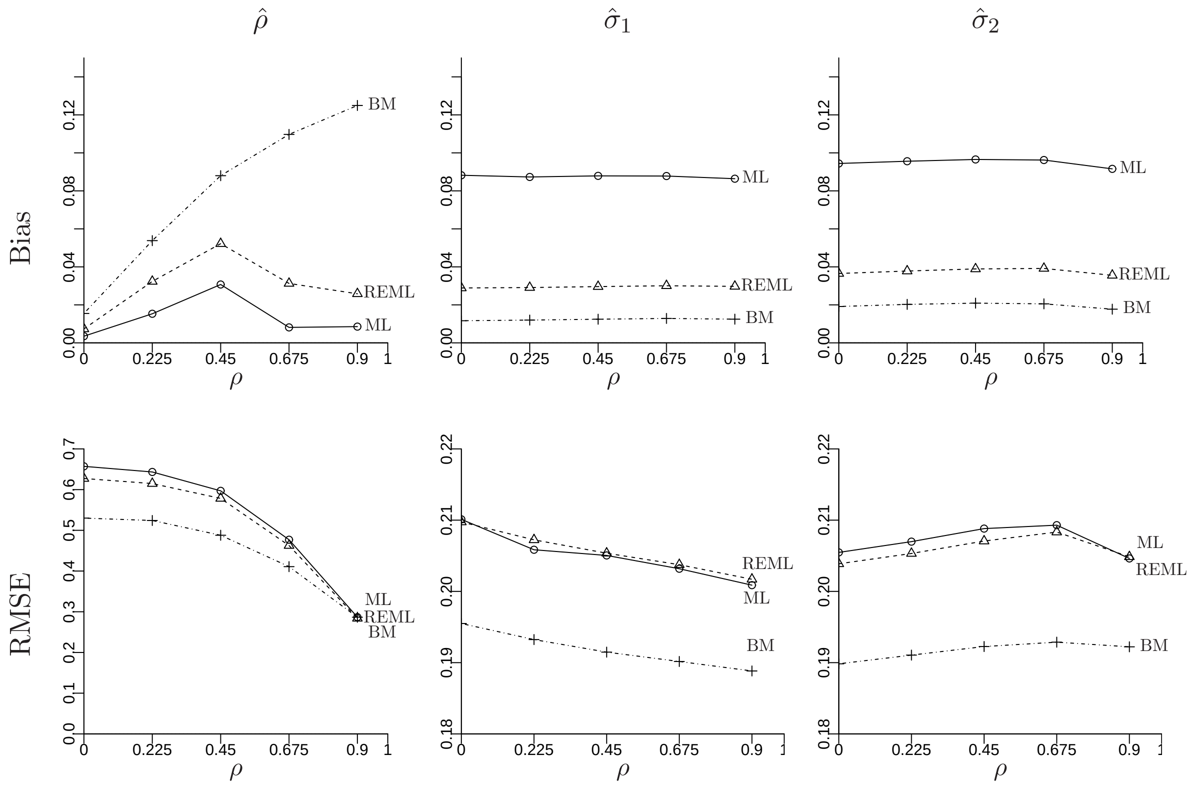

Profile likelihood / maximum likelihood
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
ρ



ML
BM
ELL (b_{1j})
Intercept (b_{0j})
3
2
1
0
-1
-2
-3
-10
0
10
20



Proportion of $\hat{\rho} \approx \pm 1$
1.0
0.8
0.6
0.4
0.2
0.0
0
0.225
0.45
0.675
0.9
1
ρ
ML
REML



$\hat{\rho}$
$\hat{\sigma}_1$
$\hat{\sigma}_2$
Bias
RMSE
BM
REML
ML
ρ
0
0.225
0.45
0.675
0.9
1
0.00
0.04
0.08
0.12
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.18
0.19
0.20
0.21
0.22