Mixed Methods for Mixed Models

Vincent Dorie

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Abstract

Mixed Methods for Mixed Models

Vincent Dorie

This work bridges the frequentist and Bayesian approaches to mixed models by borrowing the best features from both camps: point estimation procedures are combined with priors to obtain accurate, fast inference while posterior simulation techniques are developed that approximate the likelihood with great precision for the purposes of assessing uncertainty. These allow flexible inferences without the need to rely on expensive Markov chain Monte Carlo simulation techniques. Default priors are developed and evaluated in a variety of simulation and real-world settings with the end result that we propose a new set of standard approaches that yield superior performance at little computational cost.

Contents

List of Figures

List of Tables

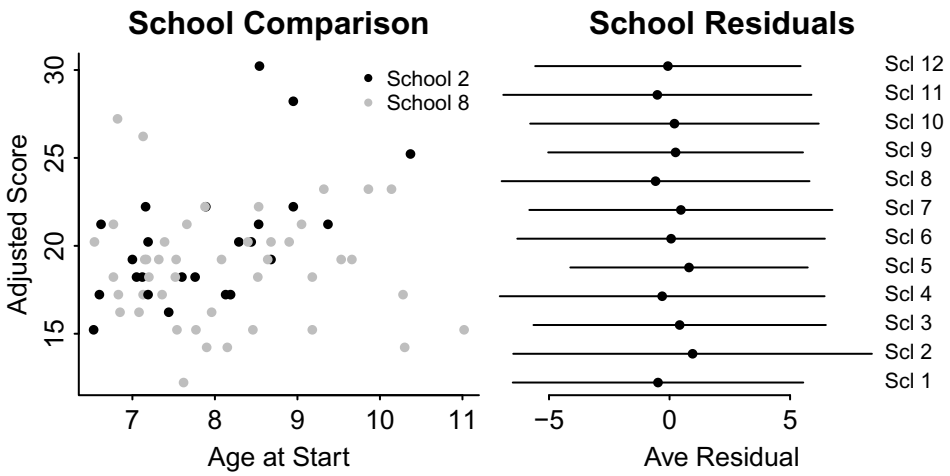

Figure 1: Cognitive assessment (Raven's scores) after controlling for the effect of treatment and starting age by fitting a simple linear model. Adjusted scores are the raw scores minus the estimate of the treatment effect from the linear model. The left figure shows two schools, the scores in which seem to be drawn from different distributions, with school 2 on average having a higher mean than school 8. On the right is the collection of the average residuals within the schools taken from the linear model and 95% confidence intervals for the corresponding expected value.

1 Introduction

1.1 Motivating Example

Whaley et al. (2003) report the results of a cluster-randomized experiment of the impact of diet on cognitive development in schools in rural Kenya. Twelve schools were randomly assigned one of four different treatments and children within the schools took cognitive assessments (Raven's score) at multiple time points over the course of several months. As the intervention was designed to last 21 months, we consider only the last observation for those that completed the treatment. In terms of data, for each child we have their Raven's score, what treatment the school received, and an additional child-level covariate of their age at the beginning of the study.

We can fit a simple linear model to attempt to predict a child's score as a function of the treatment they received and our child-level covariate. After having done so, one might ask what variation remains and if any might be due to differences in schools. Different

schools in different areas of rural Kenya should be expected to represent different populations. Figure 1 starts to address this by showing the scores after controling for treatment using the aforementioned linear model. While a great deal of uncertainty exists with regards to any estimate of the true average for a school, there is considerable evidence that the school themselves vary.

Unfortunately, for data sets such as this the traditional methods break down. The problem itself suggests a hierarchical/mixed effects model, particularly as there are colinearity issues with a fixed-effects approach and some of the schools have small sample sizes. Maximum likelihood estimation of a hierarchical linear model, however, yields an estimate of 0 variation at the school level. While we can use this to estimate the true average of any school, if we carry it to its logical conclusion then all schools have the same true average and this is stated with 100% certainty. If the goal had been to understand how the effects of the treatment vary across different populations, using the maximum likelihood estimate we would be unable to say anything constructive.

In light of all of this, the statistical concerns that we wish to address are simply to estimate the difference between the schools and to quantify our uncertainty in the estimate. To do so, the problem can be decomposed into "point" and "interval" estimation tasks. In point estimation, we are trying to obtain an accurate, non-zero estimate of the hierarchical variance parameters. Using this, one can obtain an estimate for estimands such as the difference between schools that is also non-zero. Interval estimation corresponds to assessing the uncertainty in the fitted parameters, which can be used to create intervals for the quantity of interest.

To this, we further add the constraints that our methods must be of low computational cost and easy to use. We alternatively use maximization and marginalization when addressing the former and have developed open-source, well documented software for the latter.

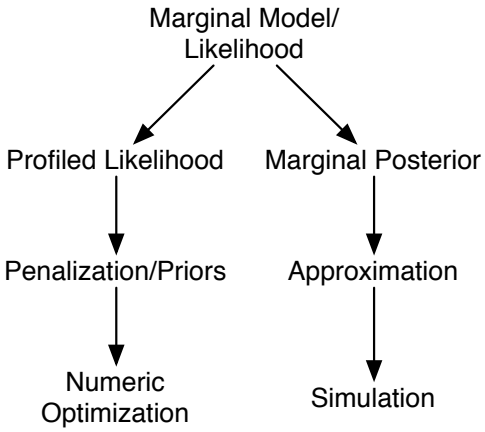

Figure 2: Graphical representation of approach to the twin problems of point (left branch) and interval estimation (right). From the same base object - the likelihood or marginal distribution of the data - successive analytic optimization yields a profiled function while successive integration yields a marginal posterior.

1.2 Overview

We tackle the point and interval estimation problems in turns. Figure 2 provides an outline of our twin approaches. For the first, we investigate classes of penalty functions on the covariance of the hierarchical components that produce estimates that are away from the boundary, yet are close to the MLE when it is positive definite. In the Bayesian paradigm, this amounts to applying a prior and estimating the posterior mode.

As for the second concern, we pivot entirely and apply flat priors to all of the parameters of the model, marginalize the resulting posterior, and develop a high quality/low computational cost approximation. This approximation can be used to simulate independent realizations of the covariance matrix, which can in turn be used to simulate the other parameters. Given a single set of simulations, any estimand of interest can be computed. Given a collection of simulations, their empirical quantiles represent intervals with good frequentist coverage probabilities.

We call these "mixed methods" because they sit between standard Bayesian and frequentist practice. In neither case are the priors applied those that a Bayesian might use when deriving a full posterior. Alternatively, they represent new penalty functions applied to the

likelihood. At the same time, after applying a prior a Bayesian statistician does not often settle for a point estimate, while conversely a frequentist generally obtains replications from an empirical distribution and not the likelihood itself.

And yet, what we propose not only works, but is justifiable from either perspective. For a frequentist, with moderately large samples the effects of a suitably chosen, weak prior will be small and leave the posterior mode close to the maximum likelihood estimate. In addition, we directly show that our simulation approach performs well under frequentist evaluation. A Bayesian, however, might use either method as a quick approximation or a set of starting points to a posterior distribution sampler.

1.3 Related Work

Variance Priors

The concept of placing a prior over the covariance of the modeled coefficients in a hierarchical model certainly is not new - every fully Bayesian approach requires this and a healthy debate exists over prior choice. However, most Bayesian solutions are centered around deriving the whole posterior, particularly a sequence of samples from a Gibbs or Markov Chain-Monte Carlo (MCMC) sampler. In contrast, selecting a prior based on deriving the posterior mode seems to be rare within the literature. A thorough investigation of producing simulations from a marginal posterior in a hierarchical model also does not exist. For the most part, what distinguishes this work from its predecessors is that the goals differ although the framework is the same. These different goals have motivated novel methods.

In the history of priors on the hierarchical variance in multilevel models, the use of an inverse-gamma/inverse-Wishart has long been a popular choice, in large part due to its conditional conjugacy (Hill, 1965; Gelfand et al., 1990; Spiegelhalter et al., 1995). Conversely, there has been considerable effort to develop a suitably vague, "objective" or reference prior including the uniform shrinkage prior (Daniels, 1999; Natarajan and Kass, 2000), a Jeffreys or proper Jeffreys prior (Berger and Deely, 1988; Berger and Strawderman, 1996), and a form

between the two (DuMouchel, 1994). Other approaches include uniform on variance scale with bounded support, or similarly uniform on the scale of the logarithm of the variance (Spiegelhalter, 2001). More recently, has been the use of half-t prior on standard deviations (Gelman, 2006) and a generalization to a scale-mixture of inverse-Wisharts (Huang and Wand, 2013).

This last method of Gelman, Huang, and Wand deserves additional attention as the priors used are simplifications of the matrix variate beta prime that we introduce in section 8.2. However those authors use the prior in the traditional Bayesian context of a posterior sampler, while our focus is on approximating the marginal posterior distribution of the hierarchical variance parameter under flat priors. It may be that this close connection makes the distribution a natural choice for the full Bayes solution.

Finally, the gamma prior proposed in section 3 was first proposed in the related work, (Chung et al., 2013). Theorem 1 appears there as well, albeit without proof. The multivariate generalization to the Wishart is forthcoming.

Marginal Simulations

The use of simulations from marginal distributions has a long history although it is not often explicated directly. Within Bayesian contexts, it arises most frequently as an embedded component of a Gibbs sampler, also known as part of a "collapsing" strategy (Liu, 1994; Van Dyk and Park, 2008). When the posterior of a set of parameters can be reduced to a marginal and a sequence of conditional distributions, all of which are inexpensive to sample, then a Bayesian would simply do so without much comment.

Another technique used is sampling/reimportance sampling, abbreviated as SIR (Rubin, 1987, 1988). SIR increases the accuracy of an approximate distribution that is easy to sample from by simply increasing the pool of available simulations.

In some sense, the approach used herein is a spiritual successor to Gelman and Rubin (1992), which tackles the problem from an iterative perspective and uses a MCMC sampler.

Despite that difference, both make an extensive effort to approximate the target distribution for the sake of easing subsequent computation.

2 Hierarchical Models

As all of our work takes place in the setting of hierarchical models, we start by defining them and discussing their relevant features.

2.1 Background

Overview

A hierarchical model is one in which the response variables exhibit some sort of natural grouping and that the groups themselves have structured variation. In the example of section 1.1, there were students grouped in schools and it is plausible that there are average cognitive abilities for schools such that these averages are draws from a common distribution.

Under the interpretation of a linear model as attempting to fit a straight line through the data, a hierarchical model consists of varying the intercept and slope of this line in a particular fashion, depending on group membership. If y_i is the ith observed response, x_i a corresponding covariate, and $j[i]$ the group number of the ith observation, a comparison between simple and hierarchical linear models is outlined by:

$$\begin{array}{cc} \text{Simple linear model} & \text{“Fixed effect” model} \\ y_i = \beta_1 + \beta_2 x_i + \epsilon_i & y_i = (\beta_1 + \beta_{1,j[i]}) + (\beta_2 + \beta_{2,j[i]})x_i + \epsilon_i \end{array}$$

$$\text{Hierarchical linear model}$$

$$y_i = (\beta_1 + \theta_{1,j[i]}) + (\beta_2 + \theta_{2,j[i]})x_i + \epsilon_i \text{ and } \theta_j \sim F$$

where F stands for an arbitrary distribution. In the typical hierarchical linear model, the error terms and the hierarchical variation are assumed to be Gaussian.

Hierarchical models and their components are known by different names in different contexts. The models themselves are also called “multilevel” or “mixed effect.” We will stick with “hierarchical models,” although a strict hierarchy is not necessarily implied. The coefficients that are not under any modeling assumption, β above, are commonly called

"fixed effects." The modeled coefficients, θ, go by "random effects" and are a form of "latent variables." As we move futher into the Bayesian paradigm, all of the coefficients and other parameters in the model may be given prior distributions. In spite of this, we continue to use the term "unmodeled coefficients" when referring to β and "modeled coefficients" when referring to θ, so that our treatment is not limited by context.

The linear hierarchical model as it is used in this context was introduced by Laird and Ware (1982). Modern references include the books (Skrondal and Rabe-Hesketh, 2004), (Gelman and Hill, 2007), and (Goldstein, 2011).

Modeled Coefficients

Our primary interest is in the modeled coefficients themselves, although the methods we develop have wider applicability. Inferences for the modeled coefficients are somewhat complicated by the fact that they are not parameters in the classical sense, nor are they observed quantities. Instead, estimation typically proceeds by first averaging out the modeled coefficients, leaving a "marginal likelihood" for the response. This integration step and the use of a marginal distribution are at the root of the point and interval estimation problems.

The marginal model could have been defined simply by itself or even arisen from a different integration process, so that the marginal likelihood remains well defined even if the matrix that corresponds to the covariance of the modeled coefficients is negative definite. It is the insistence on a hierarchical interpretation that constrains this matrix to be a valid covariance - *i.e.* positive semidefinite - and introduces a boundary into the parameter space.

The integration step also compounds the interval estimation problem by requiring that inferences about the modeled coefficients be based on an estimate of their posterior distribution. This leads to a procedure known as "Empirical Bayes," or EB. Building intervals for EB estimates is unusually difficult, as the "naive" approach of using plug-in estimators ignores uncertainty in how those estimates were obtained in the first place and produces results that are on average too narrow. Recapturing this estimation uncertainty is the key

to building intervals with good coverage properties.

2.2 Simple Model

Here we introduce a simple hierarchical model with most of the complexity stripped away for the sake of inspiration and explanation.

Definition

Suppose that we have a random sample of N observations evenly divided into J groups, with $n = N/J$ in each group. All of the observations are normally distributed around some common mean, although individuals with the same group have an offset particular to that group. The offsets are all also normally distributed with a mean of 0. Specifically:

$$\begin{aligned} y_{ij} \mid \theta_j &\overset{\text{iid}}{\sim} \mathrm{N}(\mu + \theta_j, \sigma_y^2) \qquad i = 1, \ldots, n, \\ \theta_j &\overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_y^2 \sigma_\theta^2) \qquad j = 1, \ldots, J. \end{aligned} \tag{1}$$

Furthermore, observations between groups are assumed to be independent. Note that we have adopted the convention of using the residual variance (σ_y^2) when modeling the hierarchical coefficients, such that it can be called a "common scale" factor. This is exclusively for mathematical convenience.

Within a frequentist framework, the parameters of this model that need to be estimated are μ - the overall average, σ_y^2 - the residual variance or common scale, and σ_θ^2 - the scaled variance of the modeled coefficients. From the Empirical Bayes perspective, these are the "hyperparameters."

The point estimation problem is when the MLE of σ_θ^2 is zero, which requires developing a class of estimators that are non-degenerate or strictly positive. The interval estimation problem is to quantify the uncertainty in any estimate of the parameters, which can then be

used to build intervals for estimates of θ.

Likelihood

In order to obtain the likelihood, the joint distribution of the response and the modeled coefficients, y and θ respectively, must first be integrated with respect to θ. It is:

$$p(y,\theta;\sigma_y^2,\sigma_\theta^2,\mu) = (2\pi\sigma_y^2)^{-(N+J)/2}(\sigma_\theta^2)^{-J/2}\exp\left\{-\frac{1}{2\sigma_y^2}\sum_j\left[\sum_i(y_{ij}-\theta_j-\mu)^2+\frac{1}{\sigma_\theta^2}\theta_j^2\right]\right\}. \tag{2}$$

From this it is apparent that each θ_j has the conditional distribution:

$$\theta_j \mid y \overset{\text{ind}}{\sim} \mathrm{N}\left((\bar{y}_j-\mu)\frac{\sigma_\theta^2}{\sigma_\theta^2+1/n}, \sigma_y^2\frac{\sigma_\theta^2/n}{\sigma_\theta^2+1/n}\right) \qquad j=1,\ldots,J, \tag{3}$$

where $\bar{y}_j = \frac{1}{n}\sum_i y_{ij}$ is the average within group j. When the average for those in group j needs to be estimated, the weighted average $\mathsf{E}[\theta_j \mid y] + \mu = \frac{\sigma_\theta^2}{\sigma_\theta^2+1/n}\bar{y}_j + \frac{1/n}{\sigma_\theta^2+1/n}\mu$ is useful.

Integrating the joint distribution with respect to these conditionals produces the likelihood, which is:

$$L(\sigma_y^2,\sigma_\theta^2,\mu) = (2\pi)^{-N/2}(\sigma_y^2)^{-N/2}n^{-J/2}(\sigma_\theta^2+1/n)^{-J/2} \exp\left\{-\frac{1}{2\sigma_y^2}\sum_j\left[\sum_i(y_{ij}-\bar{y}_j)^2+\frac{1}{\sigma_\theta^2+1/n}(\bar{y}_j-\mu)^2\right]\right\}. \tag{4}$$

This is equivalent to the marginal model:

$$y_{ij} \sim \mathrm{N}\left(\mu, \sigma_y^2\sigma_\theta^2+\sigma_y^2\right) \qquad i=1,\ldots n, j=1,\ldots,J, \tag{5}$$

$$\mathsf{COV}(y_{ij},y_{kl}) = \begin{cases} \sigma_y^2\sigma_\theta^2 & i\neq k, j=l \\ 0 & j\neq l \end{cases}. \tag{6}$$

Maximum Likelihood

With the introduction of the simplified model, we formally describe how the point and interval problems arise. With regards to the first, we note that the marginal model is valid for a greater range of values for σ_θ^2 as the parameter is merely a "variance component." Specifically, the hierarchical interpretation constrains the within-group covariance to be non-negative while within the marginal this is a possibiltiy.

We can go further and derive the maximum likelihood estimator $\hat{\sigma}_\theta^2$. If $S_w^2 = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2$ is the sum of the squares within groups, $S_b^2 = \sum_j (\bar{y}_j - \bar{y})^2$ is the between groups sum of squares, and $S_t^2 = S_w^2 + nS_b^2 = \sum_j \sum_i (y_{ij} - \bar{y})^2$ is the total sum of squares, then $\hat{\sigma}_\theta^2 = \left[\frac{N-J}{J}\frac{nS_b^2}{S_w^2} - \frac{1}{n}\right]^+$. As the different sums of squares result from successive projecting the data onto a vector and then projecting the residuals, this quantity is related to an F statistic and has an easily calculated probability of being 0.

Put another way, provided that the within group sum of squares is proportionally no more than $\frac{N}{N-J}$ of the total sum of squares, then the MLE of the residual variance is $\frac{1}{N-J}S_w^2$, while the MLE of hierarchical variance - the product $\sigma_\theta^2\sigma_y^2$ - is $\frac{1}{N}S_t^2 - \frac{1}{N-J}S_w^2$. If not, $\hat{\sigma}_y^2 = \frac{1}{N}S_t^2$. A competition of sorts exists between the two parameters to partition the total sum of squares with priority given to the residual variance.

Finally, as the within group sample size n increases, the probability of a zero estimate arising decreases, provided that the hierarchical model assumption is correct and $\sigma_\theta > 0$.

Returning to the interval estimation problem, using the MLE of the model parameters it is possible to estimate the posterior means of θ_j. To create an interval, one can "naively" use the MLE to estimate the posterior variance. When the $\hat{\sigma}_\theta^2$ is 0, so is this estimate. When $\hat{\sigma}_\theta^2$ is not, a consequence of theorem 1 in section 3 is that it can be shown that it is biased downwards.

The variance estimate is biased first because the MLEs themselves are biased, but also due to failing to incorporate all sources of uncertainty. Considering the decomposition $\mathsf{VAR}(\hat{\theta}) = \mathsf{E}[\mathsf{VAR}(\hat{\theta} \mid y)] + \mathsf{VAR}[\mathsf{E}(\hat{\theta} \mid y)]$, the naive approach uses an estimate of the first quantity on

the right hand side while ignoring the second.

2.3 General Model

A hierarchical linear model in its full generality expands on the simple model introduced above to allow multiple intercepts and slopes, all of which can vary at more than one grouping factor. To extend the example of section 1.1, the schools could have been grouped within regions which themselves had structured variation, or the children could have been grouped non-hierarchically based on individual characteristics.

To write out the model concisely, we use matrix notation. Specifically, let y be an N dimensional vector, X an $N \times P$ dimensional matrix of response level covariates, and Z a group level covariate matrix with N rows and a number of columns determined by the structure of the hierarchy. β is a P-vector of unmodeled coefficients and θ a vector of modeled coefficients.

The structure of Z and θ is determined by the number of grouping factors, the number of groups within each factor, and the number of items that can vary at each level. For every group within a factor, we can have an intercept and as many varying slopes as there are groups. Z is a sparse matrix consistenting of mostly 0s that "selects" and multiplies in the covariates for the appropriate elements of θ. Let there be K grouping factors and, within the kth factor, J_k groups and Q_k different types of varying coefficients. Then the modeled coefficients at the kth level form J_k different vectors each of length Q_k - *i.e.* a vector θ_{kj}, so that there are $Q = \sum_k Q_k \times J_k$ modeled coefficients in total. This also gives the number of columns of Z and the length of θ, itself a concatenation of all of the various θ_{kj}s.

Given all of this, the model is specified as:

$$y \mid \theta \sim \mathrm{N}(X\beta + Z\theta, \sigma_y^2 \mathrm{I}_N),$$
$$\theta \sim \mathrm{N}(0, \sigma_y^2 \Sigma_\theta).$$

In more verbose fashion, the second line can also be written as:

$$\theta_{kj} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_y^2 \Sigma_k) \quad j = 1, \ldots, J_k, k = 1 \ldots, K,$$

This lets us highlight the structure of Σ_θ as:

$$\Sigma_\theta = \begin{bmatrix} \mathrm{I}_{J_1} \otimes \Sigma_1 & 0 & \cdots & 0 \\ 0 & \mathrm{I}_{J_2} \otimes \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{I}_{J_K} \otimes \Sigma_K \end{bmatrix}.$$

Working with the first specification is more convenient for intermediate calculations, but in future sections it will be necessary to map from the block diagonal version of the covariance matrix back down to the free parameters. When useful, these parameters will be denoted as the vector σ_θ.

3 Boundary Avoiding Prior

The simplified model described in section 2.2 is small enough to be mathematically tractable. In this section, we present an asymptotic expansion for the maximizer in σ_θ under an arbitrary prior/penalty function as the number of groups increases.

3.1 Optimal Prior

Assuming that the simple model specified by equation 1 is true, then we can apply a prior to σ_θ and obtain an asymptotic expansion for the posterior mode as the number of groups increases.

Theorem 1. *Under an arbitrary prior $p(\sigma_\theta)$ that does not depend on the data and whose log-density is twice differentiable, the following asymptotic expansion for the posterior mode $\hat{\sigma}_\theta$ holds:*

$$\hat{\sigma}_\theta - \sigma_\theta = -\frac{1}{2\sigma_\theta}T - \frac{1}{8\sigma_\theta^3}T^2 - \frac{1}{2\sigma_\theta^2}\frac{n}{n-1}\left(\sigma_\theta^2 + 1/n\right)^2 U(\sigma_\theta)\frac{1}{J} + O_p(J^{-3/2}) \tag{7}$$

where $U(\sigma_\theta) = \frac{\mathrm{d}}{\mathrm{d}\sigma_\theta}\log p(\sigma_\theta)$ and $T = \sigma_\theta^2 + \frac{1}{n} - \frac{n-1}{n}\frac{nS_b^2}{S_w^2}$. Furthermore, as J varies $\hat{\sigma}_\theta$ is uniformly integrable.

The proof of this is in the appendix, section 12.2. This expansion can be used to derive optimal penalty functions/priors for different criterion. An example is given by taking the expected value of both sides, which shows

$$O(J^{-2}) = 3\frac{n-2}{2} + \frac{1}{2}\frac{1}{\sigma_\theta^2} - \frac{n}{\sigma_\theta}\left(\sigma_\theta^2 + 1/n\right)U(\sigma_\theta). \tag{8}$$

Isolating $U(\sigma_\theta)$, integrating, and exponentiating yields the result:

Corollary 1. *The bias term at order $1/J$ is eliminated by the improper prior*

$$p(\sigma_\theta) \propto \sigma_\theta^{1/2}\frac{(\sigma_\theta^2 + 1/n)^{1/2}}{(\sigma_\theta^2 + 1/n)^{3/2n}}.$$

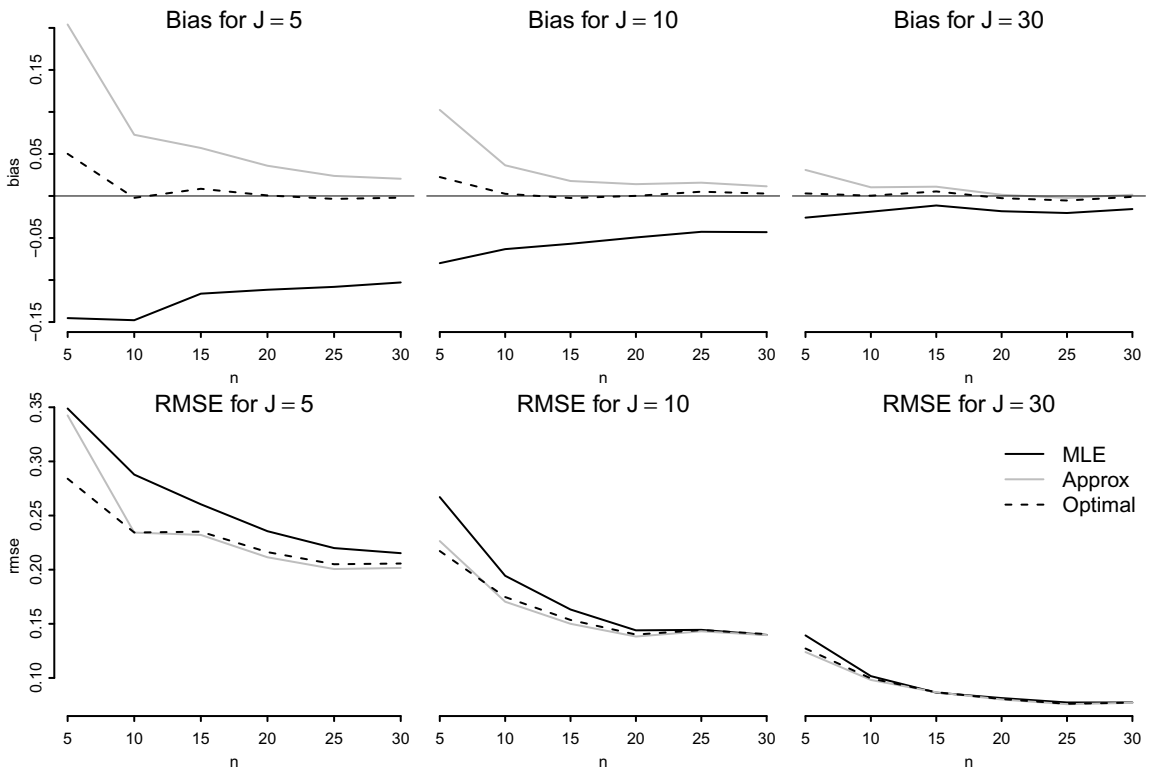

Figure 3: Simulation study assessing the bias in the maximum likelihood estimate and the correction given by the corollary to 1. Each point on graph corresponds to an average of 1000 repetitions for an experiment with the given values of n and J, as well as $\sigma_\theta = 0.5$, $\sigma_y = 1$, and $\mu = 0$.

For large n, the bias correcting prior limits to $p(\sigma_\theta) \propto \sigma_\theta^{3/2}$.

3.2 Validation

In order to test the sensitivity of theorem 1 to sample sizes, we conducted a simulation study and calculated bias and root-mean squared error in estimating σ_θ. The results are shown in figure 3. It appears that the first order bias correction is effective even for small numbers of groups ($J = 5$).

In addition, the approximation that results as we take n tending to ∞, $p(\sigma_\theta) \propto \sigma_\theta^{3/2}$, enjoys performance superior to the maximum likelihood estimate and quickly reduces the bias. Examining equation 8 shows that this prior consistently overstates σ_θ at order $1/J$, yielding conservative estimates of uncertainty. In addition, it enjoys the results of the next

param	n	β_1	β_2	β_3	β_4	σ_y	σ_θ^i	ρ_θ^{i1}	ρ_θ^{i2}	ρ_θ^{i3}
value	25	-4	0.5	0.2	-0.4	1	2	0.2	-0.5	0
							1.5		0.3	0.15
							1			-0.2
							0.75			
cov		x_1	x_2	x_3	x_4					
dist		1	N(0, 1)	N(0, 1)	Bern(0.5)					

Table 1: Parameters for a multivariate, hierarchical model simulation. And equal number of observations per group were used with group sizes varying from 5 to 25. Σ_θ is listed in the form of its standard deviations and correlations under σ_θ^i and ρ_θ^{ij} respectively.

section in that it is easily generalizabile.

3.3 Generalization

The bias correcting penalty of $p(\sigma_\theta) \propto \sigma^{3/2}$ corresponds to an improper gamma prior on σ_θ with shape equal to 2.5. The results of section 7.3 show that the posterior is proper in the vast majority of cases, although when it is not using a proper gamma prior with a large mode produces similar results.

The gamma family of distributions can be extended to covariance matrices by the use of the Wishart distribution. We propose using an improper prior with degrees of freedom set to the dimension of the covariance matrix plus 2.5. While theorem 1 was proven for estimates of standard deviations, the use of an improper prior eliminates the ambiguity of the choice of scale. For the full model of section 2.3, this is equivalent to using the penalty term $p(\Sigma_k) \propto |\Sigma_k|^{3/4}$.

Extending theorem 1 to higher dimensions has thus far proven to be difficult, but we can assess the utility of the Wishart prior through the use of simulation. We arbitrarily set the parameters for a 4 dimensional, single grouping factor according to table 1. For a chosen number of groups with 25 observations in each group, and then for 500 repetitions, covariates are simulated, modeled coefficients and a response are created, and the maximum likelihood estimate and posterior mode are obtained and evaluated.

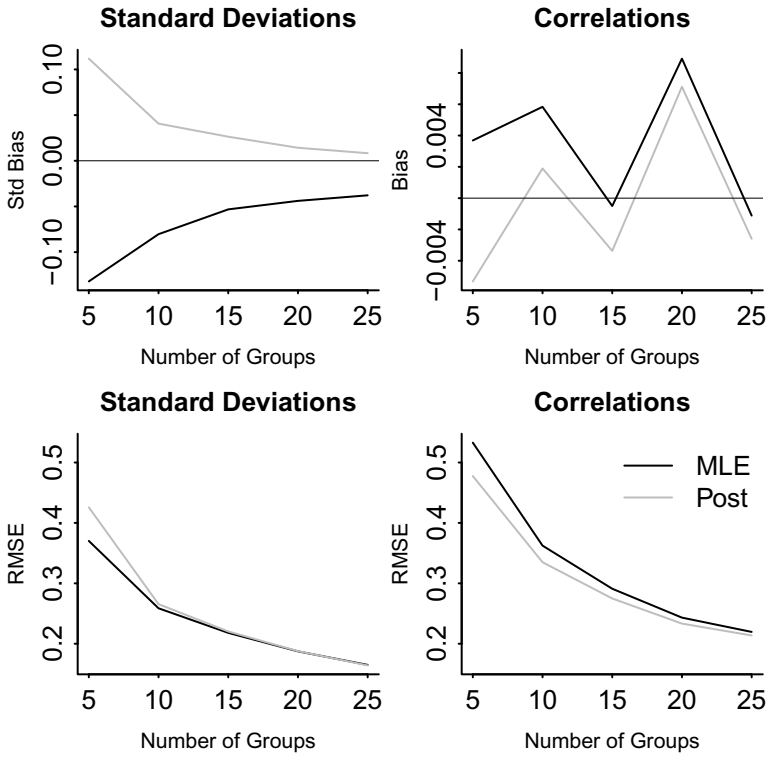

Figure 4: Average of biases (top) and root-mean squared errors (bottom) for the simulation study of section 3.3. Results were computed for each coordinate separately and then averaged. Standard deviations were rescaled before doing so.

In terms of bias and root-mean squared error, the results are reported in figure 4 separately for standard deviations and correlations. In terms of loss, the two methods are comparable. In terms of bias, the posterior mode estimates the standard deviations more accurately and in a positively biased, conservative fashion. Correlations seem to be estimated more-or-less consistently by both methods, as their bias is small compared to the scale of the parameters.

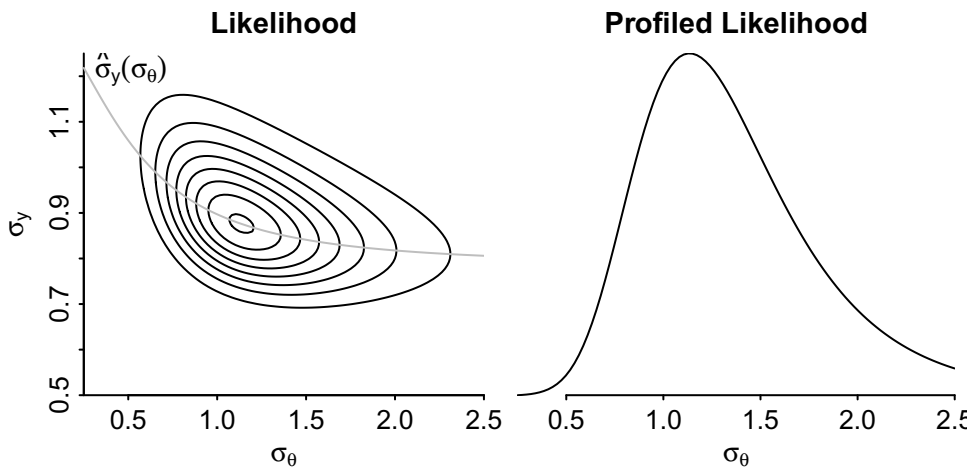

Figure 5: Illustration of a profiled likelihood. The left panel shows the likelihood for a simple hierarchical linear model (section 2.2) as a function of σ_θ and σ_y (μ has already been maximized). The gray line corresponds to the maximum in σ_y as a function of σ_θ, and following the contour along this line produces the profiled likelihood of σ_θ in the right panel. The profiled likelihood goes through the joint mode, so that maximizing it is sufficient to maximize the likelihood.

4 Profiled Posterior

Section 3 describes an approach to produce superior estimates of the modeled coefficient covariance in a hierarchical model. In this section, we discuss how to efficiently derive the maximum likelihood estimate and the possible priors that lead to similarly computationally simple posteriors. This permits implementation of the bias-reducing penalty function, as well as the application of priors to all of the model components for a more fully-Bayesian solution.

4.1 Profiling

A profiled likelihood is one over several parameters, some of which have been optimized analytically. Once a maximizer for one or more parameters has been derived, these estimators can be plugged back into the likelihood and the resulting equation optimized instead.

Figure 5 demonstrates this phenomenon for a simple hierarchical model of the form of section 2.2. The likelihood is a function of the three parameters μ, σ_θ, and σ_y, but by

equation 4 the maximizer of this function in terms of μ is always the total average of the data, $\bar{y}$. $\bar{y}$ is plugged in, yielding a likelihood with μ profiled out. In turn, σ_y can be profiled out although this time as a function of σ_θ. After these two steps a profiled likelihood over only the hierarchical variance components remains.

4.2 General Model

For the simple model, the profiled likelihood of σ_θ can be solved directly. For the general model of section 2.3 we will show that it is again possible to profile down to the hierarchical variance but that numeric techniques are necessary thereafter.

To show this, we need first to derive the likelihood by integrating θ from the joint distribution. That density is

$$p(y, \theta; \Sigma_\theta, \beta, \sigma_y^2) = (2\pi\sigma_y^2)^{-(N+Q)/2} |\Sigma_\theta|^{-1/2} \exp\left\{ -\frac{1}{2\sigma_y^2} \left[\|y - X\beta - Z\theta\|^2 + \theta^\top \Sigma_\theta^{-1} \theta \right] \right\}. \quad (9)$$

Making the change of variables $\theta = L_\theta \theta'$, where $L_\theta L_\theta^\top = \Sigma_\theta$ is a Cholesky factorization yields

$$\begin{aligned} p(y, \theta'; \Sigma_\theta, \beta, \sigma_y^2) &= (2\pi\sigma_y^2)^{-(N+Q)/2} \exp\left\{ -\frac{1}{2\sigma_y^2} \left[\|y - X\beta - ZL_\theta\theta'\|^2 + \|\theta'\|^2 \right] \right\} \\ &= (2\pi\sigma_y^2)^{-(N+Q)/2} \exp\left\{ -\frac{1}{2\sigma_y^2} \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \theta' \\ \beta \end{bmatrix} \right\|^2 \right\}. \end{aligned}$$

Writing the joint distribution in this fashion demonstrates that θ' and β together enjoy the standard role of coefficients in a linear model with an "augmented" design and response matrix, $\begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix}$ and $\begin{bmatrix} y \\ 0 \end{bmatrix}$ respectively.

To integrate out θ', we first opt to complete the square with regards to both θ' and β, as

their joint mode has subsequent utility. We denote these quantities as $\tilde{\theta}$ and $\tilde{\beta}$ and they are obtained in the standard fashion of inverting the crossproduct of the now augmented design matrix, detailed in appendix section 12.3. This operation changes the joint distribution to

$$p(y, \theta'; \Sigma_\theta, \beta, \sigma_y^2) = (2\pi\sigma_y^2)^{-(N+Q)/2} \exp\left\{ -\frac{1}{2\sigma_y^2} \left[\begin{bmatrix} \theta' - \tilde{\theta} \\ \beta - \tilde{\beta} \end{bmatrix}^\top \begin{bmatrix} L_\theta^\top Z^\top Z L_\theta + \mathrm{I}_Q & L_\theta^\top Z^\top X \\ X^\top Z L_\theta & X^\top X \end{bmatrix} \begin{bmatrix} \theta' - \tilde{\theta} \\ \beta - \tilde{\beta} \end{bmatrix} + \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} Z L_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} \right\|^2 \right] \right\}.$$

We now block-wise decompose the crossproduct of the augmented design matrix in the following fashion. Let

$$\begin{aligned} L_Z L_Z^\top &= L_\theta^\top Z^\top Z L_\theta + \mathrm{I}_Q, \\ L_{ZX} &= X^\top Z L_\theta L_Z^{-\top}, \\ L_X L_X^\top &= X^\top X - L_{ZX} L_{ZX}^\top. \end{aligned}$$

so that

$$\begin{bmatrix} L_Z & 0 \\ L_{ZX} & L_X \end{bmatrix} \begin{bmatrix} L_Z^\top & L_{ZX}^\top \\ 0 & L_X^\top \end{bmatrix} = \begin{bmatrix} L_\theta^\top Z^\top Z L_\theta + \mathrm{I}_Q & L_\theta^\top Z^\top X \\ X^\top Z L_\theta & X^\top X \end{bmatrix}.$$

L_Z depends on Σ_θ, so if not for brevity we would write $L_Z(\sigma_\theta)$.

Rotating θ's covariance with β into its mean by letting $\mu_{\theta|\beta} = \tilde{\theta} - L_Z^{-\top} L_{ZX}^\top (\beta - \tilde{\beta})$ yields

$$p(y,\theta;\Sigma_\theta,\beta,\sigma_y^2) = (2\pi\sigma_y^2)^{-Q/2}\exp\left\{-\frac{1}{2\sigma_y^2}\begin{bmatrix}\theta'-\mu_{\theta|\beta}\\ \beta-\tilde{\beta}\end{bmatrix}^\top\begin{bmatrix}L_Z L_Z^\top & 0\\ 0 & L_X L_X^\top\end{bmatrix}\begin{bmatrix}\theta'-\mu_{\theta|\beta}\\ \beta-\tilde{\beta}\end{bmatrix}\right\}\times$$
$$(2\pi\sigma_y^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma_y^2}\left\|\begin{bmatrix}y\\0\end{bmatrix}-\begin{bmatrix}ZL_\theta & X\\ \mathrm{I}_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{\theta}\\ \tilde{\beta}\end{bmatrix}\right\|^2\right\}.$$

Finally, we arive at the likelihood which is:

$$p(y;\Sigma_\theta,\beta,\sigma_y^2) = (2\pi\sigma_y^2)^{-N/2}|L_Z|^{-1}\exp\left\{-\frac{1}{2\sigma_y^2}\left[(\beta-\tilde{\beta})^\top L_X L_X^\top(\beta-\tilde{\beta})+\left\|\begin{bmatrix}y\\0\end{bmatrix}-\begin{bmatrix}ZL_\theta & X\\ \mathrm{I}_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{\theta}\\ \tilde{\beta}\end{bmatrix}\right\|^2\right]\right\}. \quad (10)$$

Noting that y appears in the exponential as a quadratic, it must have a distribution that is Gaussian. Taking its expected value and covariance yields the marginal model:

$$y \sim \mathrm{N}\left(X\beta, \sigma_y^2\mathrm{I}_N + \sigma_y^2 Z\Sigma_\theta Z^\top\right). \quad (11)$$

4.3 Profiled Likelihood

Given the equation of the likelihood for the general model, it is possible to profile it. Direct examination of (10) demonstrates that the global mode with respect to β is that of the joint mode, $\tilde{\beta}$. Plugging in this produces the first-stage profiled equation

$$p(y;\Sigma_\theta,\hat{\beta},\sigma_y^2) = (2\pi\sigma_y^2)^{-N/2}|L_Z|^{-1}\exp\left\{-\frac{1}{2\sigma_y^2}\left\|\begin{bmatrix}y\\0\end{bmatrix}-\begin{bmatrix}ZL_\theta & X\\ \mathrm{I}_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{\theta}\\ \tilde{\beta}\end{bmatrix}\right\|^2\right\}.$$

The mode in σ_y^2 can again be determined by inspection, namely $\hat{\sigma}_y^2 = \frac{1}{N} \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} \right\|^2$. Finally, we obtain the profiled likelihood

$$p(y; \Sigma_\theta, \hat{\beta}, \hat{\sigma}_y^2) = \left(2\pi\hat{\sigma}_y^2(\sigma_\theta)\right)^{-N/2} |L_Z(\sigma_\theta)|^{-1} e^{-N/2}, \tag{12}$$

where we have highlighted the dependencies on the free parameters of Σ_θ.

The derivation of this function hints at an efficient algorithm for calculating the maximum likelihood estimate for the general model. For any value of σ_θ, first compute the joint mode in θ' and β, with a side effect of obtaining the maximal value of β. Use this mode to compute the augmented sum of squared residuals and thus calculate the maximizer of σ_y^2. Finally, compute the profiled likelihood and use this in numerical optimization routine.

4.4 First Bayesian Extensions

Now that the general formula for profiling the likelihood has been outlined, it is possible to consider extensions to the model. The application of various priors represent different amounts of work, and here we discuss those those that can be applied with minimal additional complexity. Other priors are considered in section 12.4 of the appendix.

As priors are successively applied to model components, it becomes important to emphasize the quantity of interest. For example, after placing a prior over β but not the other parameters, the posterior distribution $p(\theta, \beta \mid y; \Sigma_\theta, \sigma_y^2)$ represents a traditional Bayesian estimand. Point estimation in this setting corresponds to estimating the posterior means of θ, β, or $E[(\theta, \beta) \mid y; \widehat{\Sigma}_\theta, \hat{\sigma}_y^2]$.

Conversely, the "likelihood" could be redefined. Once β has been modeled, to be strictly frequentist it should be integrated out from the joint distribution to yield $p(y; \Sigma_\theta, \sigma_y^2)$. For linear coefficients there is justification for taking this integral - the family of Restricted Maximum Likelihood (REML) estimates arise from applying and averaging β over flat prior.

However, little precedent exists for doing similarly with the variance components so we tend to avoid this approach. REML estimates are discussed in the appendix, also in section 12.4.

Alternatively, if the perspective of the penalized likelihood is accepted, then putting a prior on β is equivalent to maximizing $p(\beta \mid y; \Sigma_\theta, \sigma_y^2)$ to find $p(\hat{\beta} \mid y; \widehat{\Sigma}_\theta, \hat{\sigma}_y^2)$. While this hybrid quantity may be unusual to the Bayesian, it represents a stop-gap on the way to a full posterior mode obtained by placing priors over all model components. It is this penalized approach that we adopt, treating the marginal likelihood of equation 10 and model of equation 11 as the fundamental objects.

Unmodeled Coefficients Priors

A Gaussian prior over β can be incorporated with minimial difficulty by treating it as "pseudo data" and futher augmenting the design matrix. For example, if we assume that $\beta \sim \mathrm{N}(0, \sigma_y^2 \Sigma_\beta)$ for Σ_β known and $L_\beta L_\beta^\top = \Sigma_\beta$, then the joint distribution becomes

$$p(y, \theta, \beta; \Sigma_\theta, \sigma_y^2) = (2\pi\sigma_y^2)^{-(N+Q+P)/2} |\Sigma_\beta|^{-1/2} \times \exp\left\{ -\frac{1}{2\sigma_y^2} \left[\|y - X\beta - Z L_\theta \theta'\|^2 + \theta^\top \Sigma_\theta^{-1} \theta + \beta^\top \Sigma_\beta^{-1} \beta \right] \right\}.$$

Making the same change of variables to $\theta = L_\theta \theta'$ yields

$$p(y, \theta', \beta; \Sigma_\theta, \sigma_y^2) = (2\pi\sigma_y^2)^{-(N+Q+P)/2} |\Sigma_\beta|^{-1/2} \exp\left\{ -\frac{1}{2\sigma_y^2} \left\| \begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} Z L_\theta & X \\ 0 & L_\beta^{-1} \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \theta' \\ \beta \end{bmatrix} \right\|^2 \right\}.$$

At this point, we procede as before with the differences that $L_X L_X^\top$ is now set to $X^\top X + \Sigma_\beta^{-1} - L_{ZX} L_{ZX}^\top$ and that the degrees of freedom for σ_y^2 have increased. The marginal posterior that is to be optimized is:

$$p(\beta \mid y; \Sigma_\theta, \sigma_y^2) \propto (\sigma_y^2)^{-(N+P)/2} |L_Z|^{-1} \times$$
$$\exp\left\{ -\frac{1}{2\sigma_y^2} \left[(\beta - \tilde{\beta})^\top L_X L_X^\top (\beta - \tilde{\beta}) + \left\| \begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ 0 & L_\beta^{-1} \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} \right\|^2 \right] \right\},$$

where $\tilde{\theta}$ and $\tilde{\beta}$ are, as before, the joint mode.

This section highlights an unfortunate consequence of modeling the hierarchical variance as being times the common scale, *i.e.* $\mathsf{VAR}(\theta) = \sigma_y^2 \Sigma_\theta$, namely that how other components are modeled now requires similar treatment or else the optimization problem becomes considerably more difficult. If substantive knowledge about the distribution of β is to be incorporated, say that the intercept has a standard deviation of 2 or correlation with a slope of 0.2, to specify this requires knowledge of σ_y^2. This can either be estimated, or in the appendix we consider how to relax this requirement and perform efficient numeric optimization of σ_y^2 in an inner-loop.

Common Scale Priors

After substituting in the maximizer of unmodeled coefficients, the profiled likelihood as a function of σ_y^2 is of the form:

$$(\sigma_y^2)^{-\mathrm{df}/2} e^{-\frac{1}{2\sigma_y^2} S^2}$$

for "df" a degrees of freedom term and S^2 a sum of squares. Taking the derivative of the logarithm of this with respect to σ_y^2 yields a linear equation so that optimizing is straightforward.

A conjugate prior, $\sigma_y^2 \sim \mathrm{Inv} - \mathrm{Gamma}$, again yields a linear optimization by adjusting the degree of freedom and sum of squares by the shape and scale of the prior respectively.

Specifically, if the shape of the prior is α and the scale is γ, then $\hat{\sigma}_y^2 = \frac{1}{\mathrm{df}+2\alpha+2}(S^2 + 2\gamma)$.

For two additional prior distribution types optimization of the log-posterior is quadratic. If $\sigma_y^2 \sim$ Gamma with a shape of α and a scale of γ, then the profiled log-posterior is of them form:

$$(\sigma_y^2)^{-\mathrm{df}/2+\alpha-1} \exp\left\{-\frac{1}{2\sigma_y^2}S^2 - \frac{\sigma_y^2}{\gamma}\right\}$$

and the posterior mode of σ_y^2 is

$$\hat{\sigma}_y^2 = \frac{\gamma}{4}\left(\sqrt{(\mathrm{df} - 2\alpha + 2)^2 + 8S^2/\gamma} - (\mathrm{df} - 2\alpha + 2)\right).$$

Treating the parameter as a standard deviation and placing an inverse gamma prior on σ_y with shape α and scale γ also yields a quadratic optimization. The respective profiled posterior and modes are:

$$(\sigma_y^2)^{-(\mathrm{df}+\alpha+1)/2} \exp\left\{-\frac{1}{2\sigma_y^2}S^2 - \frac{\gamma}{\sigma_y}\right\},$$
$$\hat{\sigma}_y = \frac{\gamma + \sqrt{\gamma^2 + 4(\mathrm{df} + \alpha + 1)S^2}}{2(\mathrm{df} + \alpha + 1)}.$$

Covariance Priors

The first concern in imposing a prior over the covariance of the modeled coefficients is that Σ_θ is itself comprised of many submatrices, as outlined in section 2.3. For each grouping factor in the model, a different prior can be used so that a prior over Σ_θ is really a collection of priors over $\Sigma_1, \ldots, \Sigma_K$. How relationships between the levels can be modeled is a topic for future work.

For the most part, placing a prior on any Σ_k is no more complicated than applying a penalty function to the likelihood after the other parameters have been profiled out. The profiled likelihood of equation 12 becomes

$$p(\Sigma_\theta \mid y; \hat{\beta}, \hat{\sigma}_y^2) \propto \hat{\sigma}_y^2(\sigma_\theta)^{-N/2}|L_Z(\sigma_\theta)|^{-1}p(\Sigma_\theta).$$

As optimization typically proceeds numerically, one need only modify the result passed to the optimizer to include the presence of the prior.

Modeling becomes slightly more difficult if substantive information about the distribution of the covariance of the modeled coefficients is to be incorporated. As it appears in the likelihood only as a covariance modulo the common scale, to impose a real-world valued prior is equivalent to imposing a prior on both parameters.

For example, if $\tilde{\Sigma} = \sigma_y^2\Sigma = \sigma_y^2\Sigma_1$ is the absolute covariance of the modeled coefficients for a model with a single grouping factor and of dimension $Q = Q_1$, and furthermore it is desired to model $\tilde{\Sigma}$ as having an inverse Wishart distribution with degrees of freedom ν and scale matrix Φ, then posterior to be optimized is:

$$\begin{aligned}
p(\tilde{\Sigma} \mid y; \sigma_y^2, \beta) &\propto (\sigma_y^2)^{-N/2}|L_Z|^{-1}|\tilde{\Sigma}|^{-(\nu+Q+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Phi\tilde{\Sigma}^{-1}\right)\right\}\times \\
&\qquad \exp\left\{-\frac{1}{2\sigma_y^2}\left[(\beta-\tilde{\beta})^\top L_X L_X^\top(\beta-\tilde{\beta}) + \left\|\begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix}\begin{bmatrix} \theta \\ \beta \end{bmatrix}\right\|\right]\right\}, \\
&= (\sigma_y^2)^{-(N+\nu+Q+1)/2}|L_Z|^{-1}|\Sigma|^{-(\nu+Q+1)/2}\times \\
&\qquad \exp\left\{-\frac{1}{2\sigma_y^2}\left[(\beta-\tilde{\beta})^\top L_X L_X^\top(\beta-\tilde{\beta}) + \left\|\begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix}\begin{bmatrix} \theta \\ \beta \end{bmatrix}\right\| + \mathrm{tr}\left(\Phi\Sigma^{-1}\right)\right]\right\}.
\end{aligned}$$

The essential point of this equation is that, when taken point-wise, an inverse Wishart distribution on $\tilde{\Sigma}$ increases the degrees of freedom for σ_y^2 and adjusts its scale - a change that is equivalent to an inverse gamma prior. Straightfoward profiling steps for priors on the covariance of the modeled coefficients when not on the common scale exist only in the cases of the preceeding section, namely if that matrix has an inverse or regular Wishart distribution.

Modeled Coefficient Covariance Σ_k	Unmodeled Coefficients β	Common Scale σ_y^2
$p(\Sigma_k)$ arbitary	$\beta \sim \mathrm{N}(0, \sigma_y^2 \Sigma_\beta)$	$\sigma_y^2 \sim \mathrm{Inv-Gamma}$
		$\sigma_y \sim \mathrm{Inv-Gamma}$
$\tilde{\Sigma}_k = \sigma_y^2 \Sigma_k$,		$\sigma_y^2 \sim \mathrm{Gamma}$
$\tilde{\Sigma}_k \overset{\text{ind}}{\sim} \mathrm{Wish/Inv-Wish}$		
$\tilde{\Sigma}_k^{1/2} \overset{\text{ind}}{\sim} \mathrm{Inv-Wish}$		

Table 2: Priors for hierarchical model components that present minimal complication for profiled optimization.

Additionally, imposing an inverse Wishart on the matrix's square root works as well. Finally, as there may be more than one covariance matrix, when mixing and matching different kinds of priors it is important to be careful to leave a straightforward optimization in σ_y^2.

4.5 Summary

Table 2 condenses the results of the previous sections to those priors which can be applied without drastically altering the approach. Choices are constrained by the necessity for conjugacy that parameters have variances that factor in the common scale. This can be easily relaxed only when the resulting prior leaves a linear or quadratic optimization in σ_y^2. In all cases, the profiling algorithm is given by:

1. Determine the maximizer of the joint density in (θ', β) for $\theta = L_Z \theta'$.
2. After plugging in the maximizer $\hat{\beta}$, compute the mode in σ_y^2. This will be either a linear or quadratic optimization depending on the choices of priors.
3. Plug in $\hat{\sigma}_y^2$ and numerically optimize over the resulting function.

4.6 Generalized Linear Hierarchical Models

In general, analytic profiling is impossible for generalized linear hierarchical models. There is no direct benefit in including a common scale factor so it is not traditionally done, and

for the unmodeled coefficients to have a simple maximizer requires an exponential form that is quadratic - *i.e.* the simple linear hierarchical model. Consequently, optimization is done over the entire parameter set after approximating the integral over the modeled coefficients and the posterior is obtained by directly penalizing this function.

5 Optimization Software

We have written a software package blme for the R programming language that does profiled posterior maximization in hierarchical linear and generalized linear models. blme extends the popular package lme4 by Bates et al. (2012).

5.1 Calling `blmer`

blme was designed to be familiar to users of lme4. A `bmer` S4 object extends the `merMod` class, and consequently inherits all of the same functionality. Hierarchical linear models are fit using the `lmer` function, while generalized linear models are fit with `glmer`. Fitting a model in blmer is achieved by modifying a call to one of these two function with the addition of several new arguments, or simply replicating the call and using the default priors.

The prototype for the `blmer` and `bglmer` functions are given below:

```
blmer(formula, data, REML = TRUE,
      control = lmerControl(), start = NULL, verbose = 0L,
      subset, weights, na.action, offset, contrasts = NULL,
      devFunOnly = FALSE, cov.prior = wishart,
      fixef.prior = NULL, resid.prior = NULL, ...)
bglmer(formula, data, family = gaussian,
       control = glmerControl(), start = NULL, verbose = 0L,
       nAGQ = 1L, subset, weights, na.action, offset,
       contrasts = NULL, mustart, etastart,
       devFunOnly = FALSE, cov.prior = wishart,
       fixef.prior = NULL, ...)
```

All but the last lines are identical to the prototypes for `lmer` and `glmer` respectively.

The formats for the new arguments are all delayed function calls of syntax that will be described below, but as an overview, the new arguments are:

family	restriction mode	posterior	options & defaults
$p(\Sigma_k) \propto 1$	none	Σ_k	none
`gamma`	$\dim \Sigma_k = 1$ $\Sigma_k := \sigma_k^2$	σ_k^2 or σ_k	`shape = 2.5,` `rate = 0,` `posterior.scale = 'sd',` `common.scale = TRUE,`
`invgamma`			`shape = 0.001,` `scale = shape + 0.05,` `posterior.scale = 'var',` `common.scale = TRUE`
`wishart`	$\dim \Sigma_k > 1$	Σ_k	`df = dim(sigma.k) + 2.5,` `scale = Inf,` `posterior.scale = 'cov',` `common.scale = TRUE`
`invwishart`			`df = level.dim + 1 - 0.02,` `scale = diag(df + 1, level.dim),` `posterior.scale = 'cov',` `common.scale = TRUE`

Table 3: Types, families, and options for priors on the covariance of the modeled coefficients. Rates/scales are chosen so that the mode of the prior is at 100 for standard deviations and 10^4 for variances.

- `cov.prior`: priors on the covariance matrices of the modeled coefficients, Σ_θ. Applies to calls to both `blmer` and `bglmer`.
- `fixef.prior`: a prior on the unmodeled regression coefficients, β. Also applies to both functions.
- `resid.prior`: prior on the residual variance or common scale, σ_y^2. Only applies to calls to `blmer` as generalized linear models do not have this parameter.

As priors require hyperparameters of their own, we have adopted a delayed function evaluation system so that they can be provided in the named-list syntax.

5.2 Covariance Priors

From an interface perspective, the principal difficulty in specifying a prior on Σ_θ is that there are separately estimated covariance matrices for each of the K grouping factors. To afford the user more flexibility, priors can be specified as a default that will apply to all grouping factors or applied directly by the level name. To specify a prior that applies to a specific, named grouping factor, one passes to `blmer` an argument conforming to the format:

`factor.name ~ distribution.name(options.list)`

Conversely, to specify a default prior that should apply to all grouping factors, the correct format is simply: `distribution.name(options.list)`. The different prior types, options, and defaults are enumerated in table 3 and described below

Specifying a direct, multivariate prior type with the default parameterization enables that prior to apply to the univariate case as well. Conseqently, the Wishart can be used to specify a Gamma distribution and the inverse-Wishart the inverse-Gamma, although this implies that the posterior mode to be calculated on the variance scale.

blme permits fine tuning of the prior specification via a list of options placed in parentheses subsequent to naming the prior type. For a direct prior, the options should consist of prior parameters, such as the shape or scale of the distribution. For decompositions, the options should specify the names of the families to be applied to the individual components. These distributional families can be further controled in the same fashion as a directly-applied prior.

The options for covariance priors are:

1. `shape`/`df`/`scale`/`rate` - standard parameters for named distributions
2. `posterior.scale` - whether or not the prior is on the scale of a standard deviation/square root or a variance/covariance matrix, and consequently how the posterior should be interpretted. In the univariate setting, this is a choice between `'sd'` and `'var'`, while for the multivariate the options are named `'sqrt'` and `'cov'`
3. `common.scale` - `TRUE` or `FALSE` determining if the prior is to be interpretted modulo residual variance or is specified in an absolute sense

5.3 Covariance Examples

For the purposes of illustrating the means by which a covariance prior is specified in blme, suppose that we have data consisting of a single predictor and two grouping factors. The

outcomes are stored in `y`, the predictor in `x`, and the grouping factors are `g.1` and `g.2`. Further suppose that it makes sense to model a different intercept for each of the different groups in the first factor, while we expect the intercept and the slope to vary in the second.

Likelihood fit

To fit a model in lme4, one might call:

```
lmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2));
```

By default, `blmer` applies a prior to the covariance of the modeled coefficients, so to recover the likelihood fit the following call must be used:

```
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = NULL);
```

Univariate, default prior, standard deviation scale

The following places a univariate prior on the standard deviation of the contributions to the intercept for the first grouping factor:

```
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = gamma);
```

As gamma distributions are univariate, the prior does not extend to the second factor, which instead receives no modeling. If a third grouping factor with a single varying coefficient existsed, the gamma prior would apply to that as well.

Multivariate, default prior, variance scale

If we install a Wishart prior as a default, it will downgrade to a gamma for the univariate case and consequently apply to the variances for groups 1 and 2.

```
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = wishart);
```

Named grouping factors

We can mix the above by naming the grouping factors. We also change the prior on the first group so that this differs from the previous example.

```
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = list(g.1 ~ invgamma, g.2 ~ wishart));
```

Default priors with options

For univariate families, it is easy to specify options for the default that will apply in more than one case.

```
blmer(y ~ 1 + x + (1 | g.1) + (1 | g.2),
      cov.prior = gamma(shape = 3, rate = 1,
                        posterior.scale = 'var'));
```

Expressions as parameters

Finally, the function to create a prior is evaluated in the environment that calls `blmer`. It is thus possible to pass in variables or entire computations.

```
test.covar <- rwish(df = 3, scale = diag(2));
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = g.2 ~ wishart(scale = test.covar)));
```

A few variables are defined for when expressions are evaluated, so that complex defaults can be defined. These include

- `level.dim` - dimension of Σ_k, or how many items varying at the kth grouping factor.
- `n`/`n.obs` - number of observations
- `p`/`n.fixef` - number of unmodeled coefficients

In addition, as the priors are created by function calls, it is possible for arguments to refer to each other. For the priors above, these include

- `gamma` - `shape` and `rate`
- `invgamma` - `shape` and `scale`
- `wishart` - `df` and `scale`
- `invwishart` - `df` and `scale`

For example, to change the default degrees of freedom for the Wishart:

```
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = wishart(df = level.dim + 2));
```

5.4 Unmodeled Coefficient Priors

At present, only two different types of priors exist for priors on the unmodeled coefficients, the flat prior and multivariate normal. In the future, the addition of t priors may be investigated.

Unmodeled coefficient priors are specified in similar fashion to those over covariance matrices - by character strings containing a named distribution with an optional named-list of hyper parameters. That is, by passing the `fixef.prior` argument to `blmer` or `bglmer` a string containing:

```
family.name(options.list)
```

A normal prior has the following options:

1. `common.scale` - true or false, depending on whether or not the prior's covariance should be multiplied by the common scale factor, σ_y^2, when it exists

2. `sd` - either a single value to be used for all unmodeled coefficients, a vector with two values where the first is for the intercept and the second is replicated for all slopes, or a vector of length equal to the number of unmodeled coefficients that contains the standard deviations for each coefficient

3. `cov` - largely equivalent to the `sd` option but can also be used to specify a full matrix

The default is a normal prior with a standard deviation of 10 for the intercept and 2.5 for all slopes, that is then multiplied by the common scale and scaled further by the data. If the prior is not on the common scale, as discussed in section 4.4 and the appendix 12.4, profiling the common scale requires numeric techniques.

By default `lmer` peforms restricted maximum likelihood, or REML estimation. This is equivalent to imposing a flat prior on the unmodeled coefficients and integrating them out of the likelihood. If the REML option stays at the default and a Gaussian prior is used, the integration will still take place. If this is undesirable, pass the argument of `REML = FALSE` to `blmer`.

Example

An example demonstrating most of the options for a normal prior is:

```
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2), REML = FALSE,
      fixef.prior = normal(cov = diag(c(8^2, 2^2)),
                           common.scale = FALSE));
```

5.5 Common Scale Priors

Common scale priors follow the formula above, specified by giving a text string of the form: `family.name(options.list)`. Potential family names are specified in table 4. Note that, in accordance with the discussion of section 4.4, imposing a gamma prior on σ_y without setting the rate to 0 will result in a non-trivial profiling step.

family	options & defaults
`none`	none
`gamma`	`shape = 0,`
	`rate = 0,`
	`posterior.scale = 'var'`
`invgamma`	`shape = 0,`
	`scale = 0,`
	`posterior.scale = 'var'`
`point`	`value = 1.0,`
	`posterior.scale = 'sd'`

Table 4: Families and options for priors on the residual variance parameter/common scale, σ_y^2.

6 Covariance Priors Compared

In this section we consider real and simulated data sets under a variety of priors on the covariance of the modeled coefficients. Lambert et al. (2005) made a similar comparison for hierarchical models when fit using MCMC, and this is designed to serve a similar role for the posterior mode setting. As that analysis was done for meta-analysis in which the residual variance, σ_y^2 is known, we first operate under those assumptions. The question of using REML estimates with covariance penalties has not yet satisfactorily been addressed, so we also make an initial investigation into that direction.

Our main goal in this analysis is to provide practical advice for the fitting of hierarchical models. In connection with the preceeding results, we focus primarily on point estimation. However, as future sections are concerned with uncertainty estimation, we include some results on posterior credible intervals as well.

The main results of the investigation can be broken down into cases. When it is plausible that the variance of the modeled coefficients is close to 0, maximum likelihood or restricted ML (REML) produce perfectly adequate point estimates. No prior tested produces an adequate credible interval in this scenario, and further investigation is required. When 0 estimates are unacceptable or it is believed that the variance is positive, the gamma family of priors that penalize by $\sigma_\theta^{\alpha-1}$, for α between 2 and 3 either reduce the bias or at least make it positive, so that they enable conservative inference. The various priors based on the notion of shrinkage or Jeffrey's invariance do a good job capturing the uncertainty for large parameter values, but some additional work is necessary with regards to the behavior near the origin. Finally, REML estimates are in general superior to maximum likelihood, however they are not sufficient to remove the downward bias nor necessarily pull the estimate away from the boundary.

6.1 Prior Scale

Before any comparison can be made, the issue of prior and posterior scaling needs to be addressed. For example, in the study done by Lambert uniform priors are compared when applied to the logarithm of the modeled coefficient variance, the variance itself, and to the square root/standard deviation. In order to make the results of the various priors comparable, the posteriors are all computed in the same parameterization. When using MCMC, this happens automatically: a Markov chain producing samples from the posterior $\sigma_\theta \mid y$ can be transformed to one sampling from $\sigma_\theta^2 \mid y$ by simply squaring the samples.

The situation for maximization is slightly more complicated. In general, the posterior mode under a prior on one scale cannot simply be transformed to obtain the posterior mode under another. Consider a simple hierarchical model with a prior on $p(\sigma_\theta^2)$ on the variance scale. The function which is numerically optimized, the log profiled posterior, is:

$$\log p(\sigma_\theta^2 \mid y) \propto l(\sigma_\theta^2) + \log p(\sigma_\theta^2).$$

To obtain a posterior mode for the standard deviation we would need to multiply by a term corresponding to a change of variables and instead be plugging in values to the equation:

$$\log p(\sigma_\theta \mid y) \propto l\left((\sigma_\theta)^2\right) + \log p\left((\sigma_\theta)^2\right) + \log \sigma_\theta.$$

From the perspective of usable software, it would be awkward if the supposedly optimal variance is not the square of the best standard deviation, not to mention the computational costs incurred by having to optimize a new set of equations for every desired transformation.

With this in mind, we opt to express all priors in a common parameterization and focus on the functional form of the penalty term. While treating all priors on the scale of standard deviation has an intuitively understandable appeal, we opt for variances as they enable future comparisons in higher dimensions.

One side-effect of this treatment is that a prior that was originally expressed on the scale

study	log odds-ratio	standard error
1	-0.05	0.45
2	-0.22	0.29
3	1.02	0.52
4	0.96	0.27
5	0.42	0.24

Table 5: Log odds ratios for a meta-analysis of 5 studies on the rate of failure when comparing two treatment regimes and their associated standard errors.

of a standard deviation may become unusuable when transformed to a prior on a variance. This is particularly true of the half-distributions, such as half-Cauchy or half-normal. Those distributions limit to a finite point at 0, but after including the Jacobian of the transformation behave like $1/\sigma_\theta$ near the origin. To circumvent this, we assume that it was the functional form itself that recommended these priors to their supporters and include them without the change-of-variables adjustment. This is backed up by the subsequent results, which were of a uniformly inferior quality when the priors were interpretted literally.

Finally, while Lambert includes a variety of uniform priors on different scales, the above discussion demonstrates that these will all produce identical results - assuming that the maximum of the likelihood is not too extreme. As our primary goal is to investigate priors with universal applicability, we omit uniform priors or consider their unbounded, transformed versions instead.

6.2 Meta-Analysis Study

Lambert et al. (2005) start with a meta-analysis study of log-odds ratio data for failure between two treatment regimes taken from Glasziou et al. (2004). The data are reproduced in table 5. The model used is:

$$
\begin{aligned}
y_j \mid \theta_j &\overset{\text{ind}}{\sim} \mathrm{N}(\theta_j, \sigma^2_{y,j}), \quad j = 1, \ldots, 5, \\
\theta_j &\overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2_\theta), \\
\mu &\sim \mathrm{N}(0, 100^2),
\end{aligned}
$$

with $p(\sigma^2_\theta)$ taking on different forms and $\sigma^2_{y,j}$ known.

In the original work, 13 different priors were implemented. Draws from the joint posterior of σ^2_θ and μ were obtained for each by using the BUGS program, which implements a Gibbs sampler (Spiegelhalter et al., 1996). Point estimates and 95% credible intervals for μ and σ_θ were created from these samples by taking their mean and using empirical quantiles.

Taking inspiration from Lambert et al. and various other works in the hierarchical model literature, we selected 12 priors of our own which are subsequently described. For each, we calculate the joint posterior mode for μ and σ^2_θ. To obtain uncertainty estimates, we use a slice-sampler on the marginal posterior of $\sigma^2_\theta \mid y$. As such, the width of intervals should be directly comparable to Lambert et al.'s analysis when the priors are the same. In fact, those samples could be used to find the posterior mean which would then make the point estimates comparable as well.

This raises the issue that many of the priors tested were originally recommended for their utility in full-posterior estimation. As such, there is no expectation that they will be useful in maximization - in fact, the opposite supposition is more likely. We include these priors here to provide some continuity with previous work and to hopefully gain insight into what an optimal prior might look like.

Priors

Prior 0

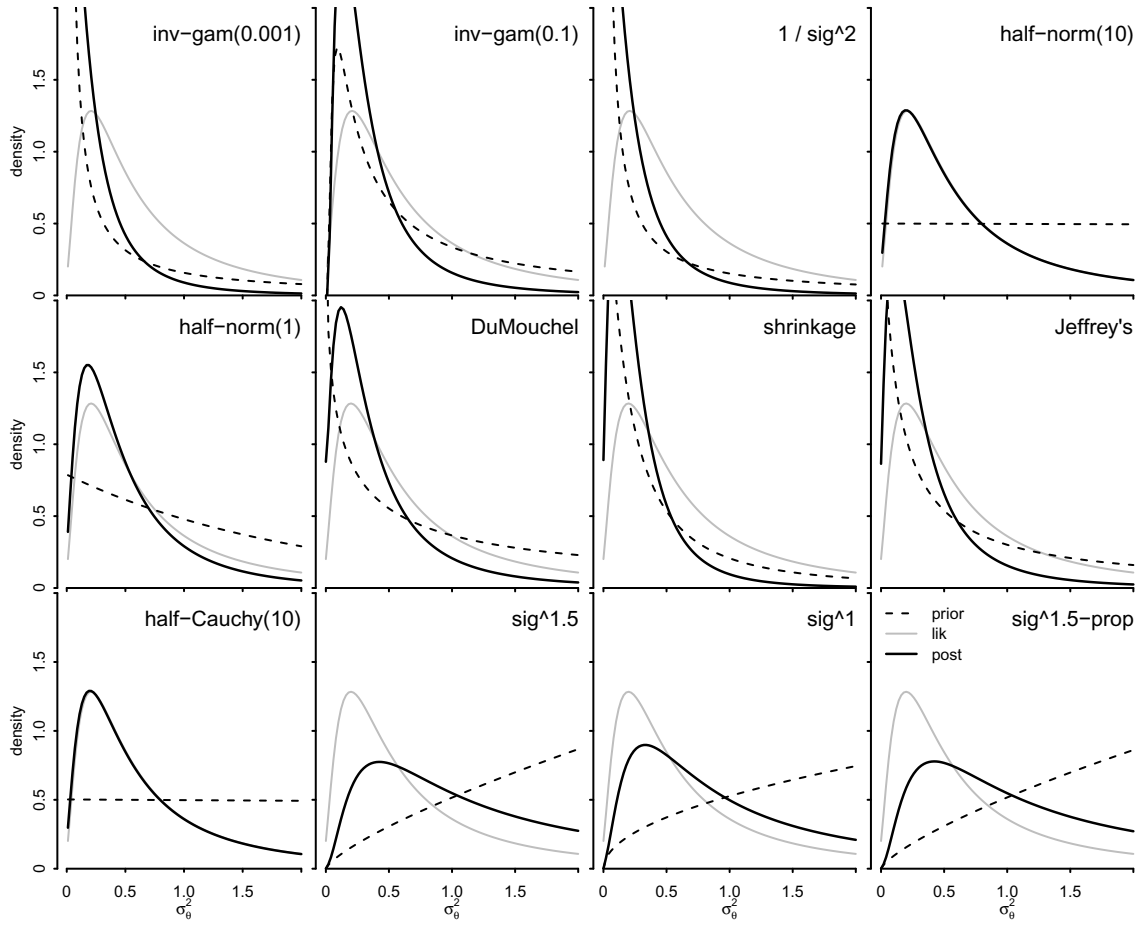

Figure 6: From left to right, top to bottom: priors, "likelihood" (profiled posterior under flat prior), and *profiled* posterior of $\sigma^2_\theta, \hat{\mu} \mid y$ for priors 1 through 12 in the meta-analysis study of section 6.2. The horizontal scale is that of a variance, so that the priors are characterized by their penalty terms. The vertical axis is taken so that the visible graph integrates approximately to 1.

$$p(\sigma^2_\theta) \propto 1$$

We start with a "flat" prior corresponding to no penalty or maximum likelihood. This is also equivalent to a uniform prior on the variance scale, except that it does not arbitrarily exclude extreme parameters values. It produces a proper posterior given that there are more than 2 groups, and thus is a perfectly valid option for consideration. Being proportional to the likelihood (rather, here the posterior $\mu \mid y; \sigma^2_\theta$), it also represents a baseline on which every other prior must operate.

Prior 1

$$p(\sigma_\theta^2) \propto \left(\sigma_\theta^2\right)^{-(0.001+1)} e^{-\frac{0.001}{\sigma_\theta^2}}$$

This corresponds to an inverse-gamma distribution on the variance scale and also results from rescaling a gamma prior on the precision scale. It is notable for being used in WinBUGS examples (Spiegelhalter et al., 1995), and can be considered an example of the inverse gamma (ϵ, ϵ) family in which both hyperparameters are the same. The prior technically goes to 0 at the origin, but has a large amount of probability density on small values of σ_θ^2.

Prior 2

$$p(\sigma_\theta^2) \propto \left(\sigma_\theta^2\right)^{-(0.1+1)} e^{-\frac{0.1}{\sigma_\theta^2}}$$

This is a simple variant on the first prior, included by Lambert et al. as a test of the inverse-gamma's sensitivity to its hyperparameters. For simulated data, it is particularly useful as it demonstrates how the two tails of the distribution impact the posterior, as the true value moves above and below the prior mode.

Prior 3

$$p(\sigma_\theta^2) \propto 1/\sigma_\theta^2$$

A standard non-informative prior for variance components in simple linear models, it has been noted to yield improper posteriors if there are an insufficient number of groups (DuMouchel and Waternaux, 1992). It arises from a uniform prior on the scale of the logarithm of the variance/standard deviation. It is the only prior considered here that is not bounded near 0, and is included largely as a cautionary tale.

Prior 4

$$p(\sigma_\theta^2) \propto e^{-\frac{1}{2}\frac{\sigma_\theta^2}{10^2}}$$

A gamma prior on σ_θ^2 (shape = 1, scale = 200), it also corresponds to placing a half-normal distribution on σ_θ centered at 0 with a standard deviation of 10 and then ignoring the Jacobian. Given that it is effectively flat near the origin, it should have little impact on point estimation. However, the tails of the distribution may be useful in penalizing large estimates, provided external knowledge exists.

Prior 5

$$p(\sigma_\theta^2) \propto e^{-\frac{1}{2}\sigma_\theta^2}$$

A variant on the previous, it arises from a half-normal distribution on σ_θ with a variance of 1. While of little practical use as a default for optimization, the shape of the tail plays a role in sampling.

Prior 6

$$p(\sigma_\theta^2) \propto \frac{h}{(h+\sigma_\theta)^2}$$

where $h^2 = J/\sum \sigma_j^{-2}$ is the harmonic mean of the observed variances. This is equivalent to a log-logistic distribution on the standard deviation and has been used by DuMouchel and Normand (2000). While not deriving from any formal source it has the advantage of a proven track record.

Prior 7

$$p(\sigma_\theta^2) \propto \frac{h^2}{(h^2+\sigma_\theta^2)^2}$$

with h as defined in prior 6.

This is an approximate uniform shrinkage prior. In the empirical-Bayes literature, the hierarchical structure "shrinks" the estimate of a group's mean towards the prior mean. The estimate of any θ_j is taken from the posterior mean of $\theta_j \mid y$, which has the form of a weighted average between prior and data means:

$$\mathsf{E}(\theta_j \mid y) = \frac{\sigma_\theta^2}{\sigma_{y,j}^2+\sigma_\theta^2} y_j + \frac{\sigma_{y,j}^2}{\sigma_{y,j}^2+\sigma_\theta^2}\mu$$

When all $\sigma_{y,j}^2$ are equal, placing a uniform $(0,1)$ prior on the coefficent of the prior mean and transforming to the variance scale yields the uniform shrinkage prior. The use of the harmonic mean for unequal variances is suggested by DuMouchel (1994) and the form itself appears in a work by Daniels (1999).

Prior 8

$$p(\sigma_\theta^2) \propto \prod_{j=1}^{J} \frac{1}{\left(\sigma_{y,j}^2+\sigma_\theta^2\right)^{1/J}}$$

A variant on the Jeffrey's prior -to which it reduces if $\sigma_{y,j}$ are all equal - it was discussed by Berger and Deely (1988).

Prior 9

$$p(\sigma_\theta^2) \propto \left(1+\frac{\sigma_\theta^2}{10^2}\right)^{-1}$$

The half-Cauchy prior on the standard deviation of the modeled coefficients is recommended by Gelman (2006). There it appears with a scale of 25, which unnecessarily vague

for the standard errors here.

Prior 10

$$p(\sigma_\theta^2) \propto \sigma_\theta^{3/2}$$

An improper gamma distribution on the variance with a shape of 7/4, and our recommendation from section 3. It also corresponds to an improper gamma prior on σ_θ with a shape of 1.5.

Priors 11 and 12

To test the sensitivity of the gamma to its shape parameter and the impact of using an improper prior, we include the following:

$$p(\sigma_\theta^2) \propto \sigma_\theta$$

$$p(\sigma_\theta^2) \propto \sigma_\theta^{3/2} e^{-\sigma_\theta^2/100}$$

The first is equivalent to a gamma prior on the scale of σ_θ with a shape of 2, while the second is a vaguely proper version of prior 10.

Methods

For each of the 12 priors above, as well as the likelihood, the joint posterior mode of $\sigma_\theta^2, \mu \mid y$ was calculated by numeric optimization. For priors that are directly supported by blme, it was used without modification by fixing the residual standard deviation to 1 with a point prior, apply a Gaussian to the unmodeled coefficient, and adding observation weights equal to the inverse of the study's squared standard error.

As when the residual standard deviation is known a prior on the covariance of the modeled coefficients influences the estimating equation strictly as a penalty, it is possible to use `blmer`

to calculate the likelihood portion directly (here, actually $p(\mu \mid y; \sigma_\theta^2)$). Wrapping this in a function that includes the penalty term permits numeric optimization, and was completed using a quasi-Newton method. For some data sets, the likelihood can contain multiple extrema with one at 0. `blmer` contains an automatic correction for this scenario, but to prevent the generic optimizer from getting stuck starting values were used that were either the true value or obtained from other, similar methods.

To obtain uncertainty estimates, a slice sampler was implemented and used on the marginal posterior of $\sigma_\theta^2 \mid y$, obtained by integrating out μ from the joint. When the joint mode of σ_θ^2 was away from 0, samples were obtained from the marginal posterior of $\log \sigma_\theta^2$ instead. The number of iterations varied depending on the context, but in all cases a burn-in of 100 samples was used. Once samples of σ_θ^2 were obtained, $\mu \mid y, \sigma_\theta^2$ was drawn directly. Posterior 95% credible intervals for each parameter were obtained by looking at the empirical quantiles of the samples.

Results

Figure 7 shows the point estimates and 95% credible intervals for the various priors in the meta analysis study. Intervals were based on the empirical quantiles of 500 samples.

While the prior on σ_θ^2 has little impact on the point estimate of the log-odds ratio μ, the uncertainty in the posterior over the first parameter translates into significant variation in the posterior over the second. Accurate estimation of the uncertainty in σ_θ^2 is important not just for its owns sake, but also because it may change the statistical significance of the rest of the model.

Furthermore, the choice of prior had significant impact on the posterior mode. In the original study, priors uniform on the variance scale produced posterior means around 0.8, while ones uniform on the standard deviation scale were close to 0.5. Here, with the exception of the gamma family priors (10-12), the inverse gamma with a small prior mode (1), and prior 3, most of the modes are tightly distributed around a between study standard deviation of

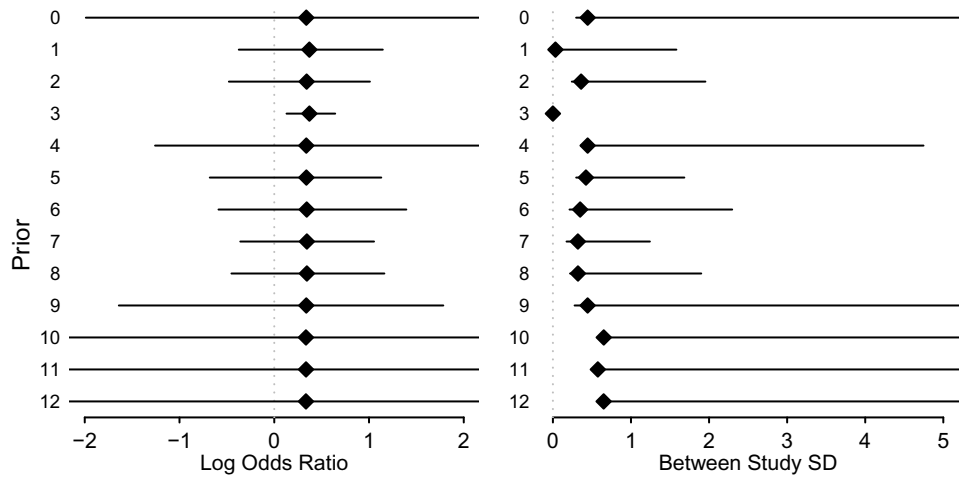

Figure 7: From left to right: point estimates and 95% credible intervals for the grand mean parameter μ and the modeled coefficient standard deviation, σ_θ, for the meta-analysis study of section 6.2. Error bars for the mean extending beyond the plot range are $(-2.0, 2.7)$ for prior 0, $(-39, 53)$ for 10, $(-6.6, 9.3)$ for 11, and $(-3.2, 4.2)$ for 12. Unplotted upper bounds for the standard deviation are 8.2 for prior 0, 288 for 10, 34 for 11, and 11 for 12. Draws from the posterior under prior 3 were unable to be obtained in a reasonable amount of time, as the sampler got stuck near the origin. For this case, we assume that every posterior sample is 0 and build intervals accordingly.

0.5. We forgo discussion of the subjective quality of these results for the following objective comparison using simulated data, but this roughly shows how to obtain an modal estimate close to a mean estimate for different priors.

6.3 Meta-Analysis Simulation

Again following Lambert et al. (2005), we conduct a simulation study of meta-analyses, varying the number of studies and the between study standard deviation. For each of $\sigma_\theta = 0.001, 0.3$, or 0.8 and each of $J = 5, 10$, or 30, data are generated according to the following model:

$$
\begin{aligned}
\theta_j &\overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\theta^2) \quad j = 1, \ldots, J, \\
\mathrm{logit}(p_{0j}) &= 0, \\
\mathrm{logit}(p_{1j}) &= \mu + \theta_j, \\
r_{0j} &\sim \mathrm{Bin}(n_{0j}, p_{0j}), \\
r_{1j} &\sim \mathrm{Bin}(n_{1j}, p_{1j}), \\
y_j &= \log \frac{r_{1j}/(n_{1j} - r_{1j})}{r_{0j}/(n_{0j} - r_{0j})}, \\
\sigma_{y,j} &= \sqrt{\frac{1}{r_{0j}} + \frac{1}{n_{0j} - r_{0j}} + \frac{1}{r_{1j}} + \frac{1}{n_{1j} - r_{1j}}},
\end{aligned}
$$

where $\mu = 0.323$. The number of subjects per trial runs from 100 to 500 in 100 unit increments, as per the original design. As it was not specified, we split subjects evenly between the two treatment arms so that n_{kj} varies from 50 to 250. The data represent various experimental runs with similar, but varying probabilities of "successes" or events. Hierarchical linear models are fit to the resulting log-odds ratios, weighted by the inverse of the squares of the estimated standard errors.

For each experimental condition, 1000 data sets were generated. For each data set, and for each prior from 0 to 12 above, the profiled posterior mode of $\sigma_\theta^2, \mu \mid y$ was calculated. Additionally, 200 simulations were drawn from the marginal posterior of $\sigma_\theta^2 \mid y$ using a slice sampler, again using a 100 sample burn-in. As before, the empirical quantiles of these were transformed into a 95% credible interval for σ_θ, and the samples themselves used to do similar for μ.

Sampling Distributions

To assess the bias and variability in the estimates, the modes themselves were recorded for each of the 1000 simulations. Figure 8 summarizes the point estimates for the underlying log-

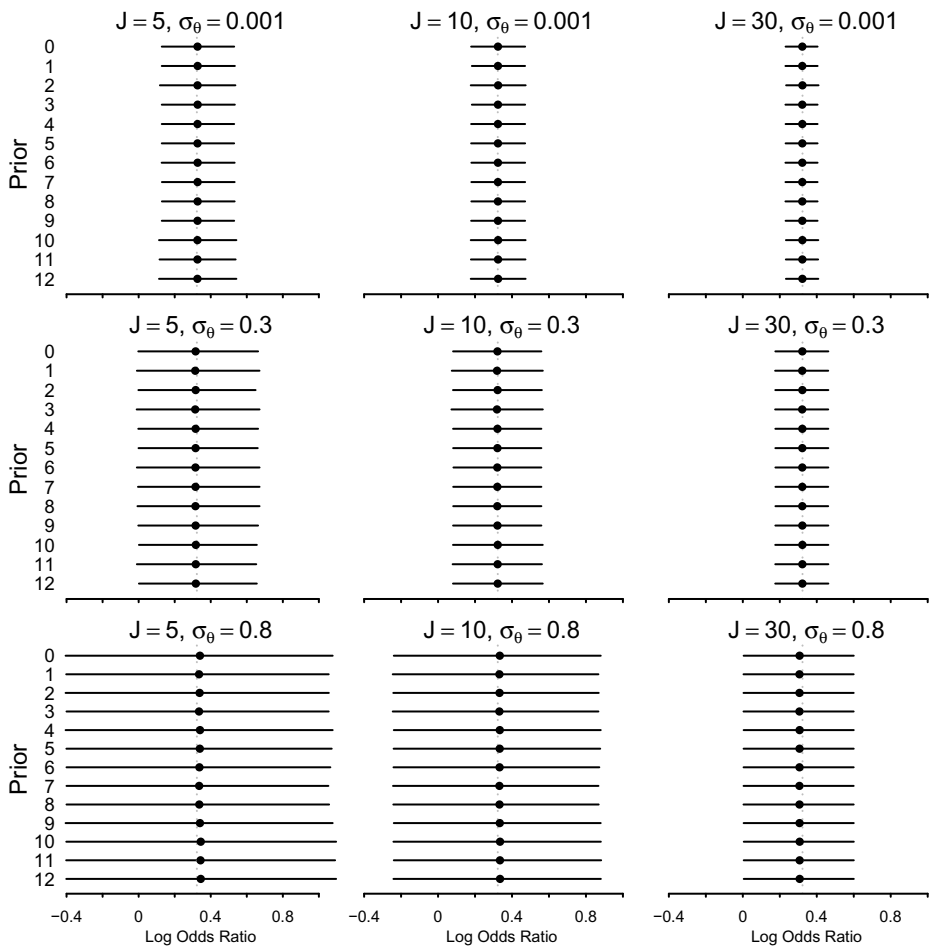

Figure 8: *Sampling distributions* of $\hat{\mu}$ under the various experimental conditions of section 6.3. The points are the mean of posterior modes for 1000 simulated data sets and are an estimate of its expected value. The error bars are empirical 95% intervals. The gray dotted line at 0.383 corresponds to the true value of μ.

odds ratio across all experimental conditions and various priors, while figure 9 does similar for the between study standard deviation. Corresponding with the results of Lambert et al., regardless of estimation technique, $\hat{\mu}$ appears to be relatively stable. Furthermore, the various gamma priors significantly reduce the bias in the estimate when the true between-study standard deviation is away from the boundary. Note that these graphs show the sampling distributions of estimates under the model, and not what any individual fit says about the uncertainty in the parameter.

The poor performance of the inverse gamma priors, priors 1 and 2, highlight a major

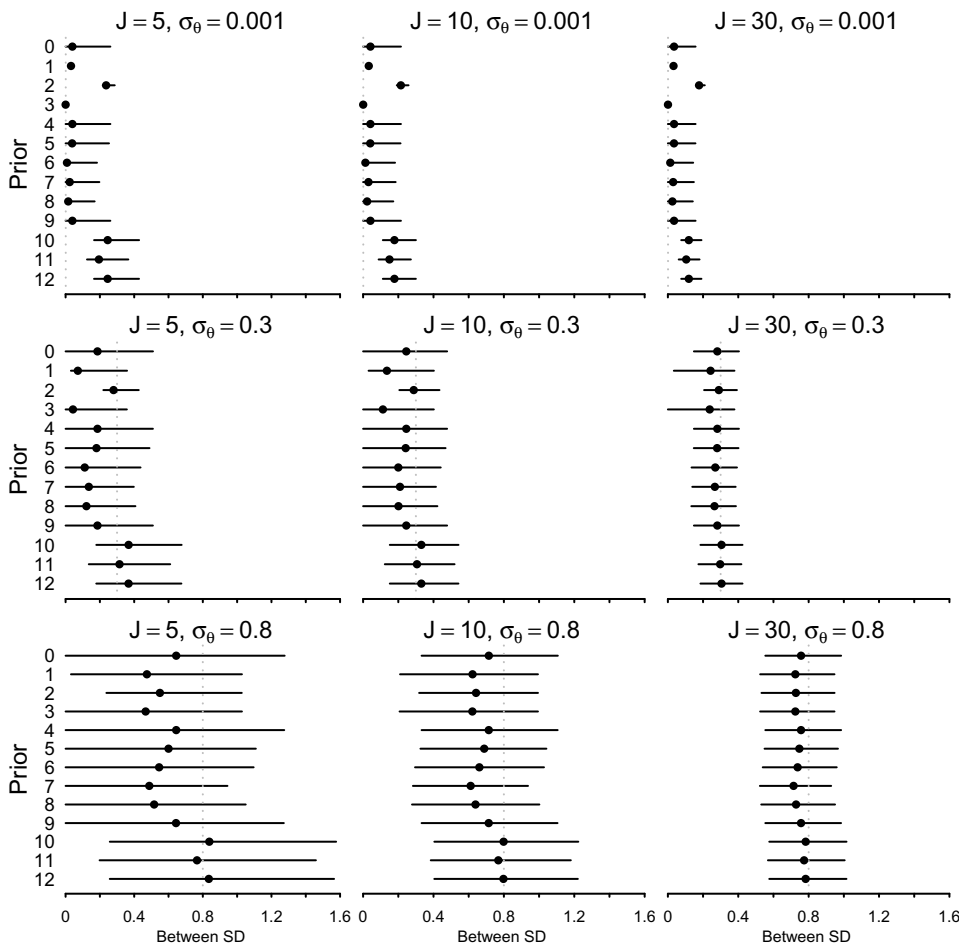

Figure 9: Similar to figure 8, the *sampling distributions* of $\hat{\sigma}_\theta$ under the various experimental conditions. The gray dotted lines correspond to the true values of σ_θ.

drawback to their use and echoes the concerns of Gelman (2006). For small values of the parameter, the prior decreases to 0 at a rate faster than that for which the likelihood can compensate. Figure 10 shows this effect up close, wherein the maximizer of the likelihood is on a strongly negative slope of the prior. Increasing the certainty of the 0 estimate, *e.g.* increasing the number of studies, has little effect so that to use an inverse gamma prior is to effectively specify a soft minumum for the between study standard deviation. Were that the only effect the prior might have its uses, but it also exhibits a strong pull when the likelihood peaks to the right of the prior mode.

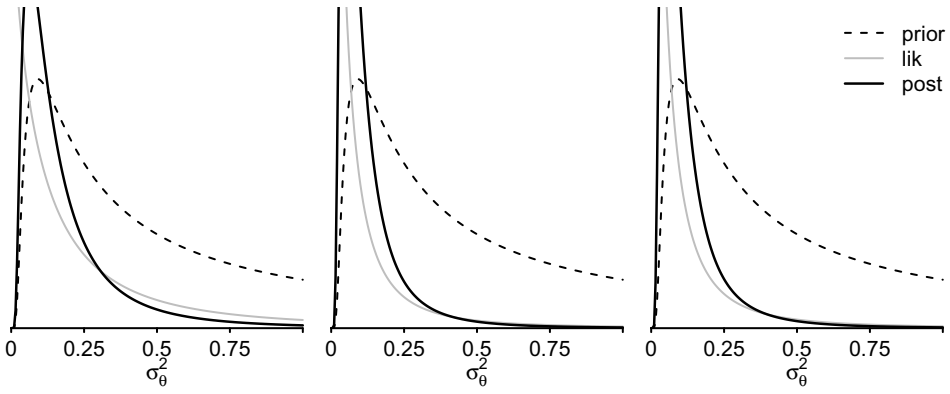

Figure 10: Prior, profiled likelihood, and posterior for 3 simulation runs of section 6.3 with $\sigma_\theta = 0.001$, $J = 5$, and prior 2 ($\sigma^2_\theta \sim \Gamma^{-1}(0.1, 0.1)$). The vertical axis is scaled so that the likelihood and posterior integrate to 1, while the prior is 10 times the density so as to facilitate visual comparison.

Posterior Uncertainty

Figures 11 and 12 show the coverage rates and average interval widths for 95% intervals for the parameters based off of the posterior samples under various priors.

The results raise a few issues. The first of which is that the only prior to obtain any coverage for the case in which $\sigma_\theta = 0.001$ was the otherwise-unusuable $p(\sigma^2_\theta) \propto 1/\sigma^2_\theta$. In order for central 95% intervals to capture this value, a full 2.5% of the probability mass for a posterior would have to lie to the left of 0.001^2. To build intervals that are valid for values close to or at 0, the posterior should include some positive probability of generating 0, and hence a "spike and slab" style prior might be appropriate.

For the case where the true between study standard deviation is away from 0, we see a clustering of sorts among the various priors:

- The "flat" priors, including the likelihood (0), the half-normal with standard deviation 10 (4), and the half-Cauchy (9) all tend to miss-estimate the variability in σ_θ for small sample sizes. Analysis of the intervals shows that they tend to encapsulate values that are too large, *i.e.* when the value σ_θ lies outside of the interval it does so more often on the left.

- The gamma family priors (10-12) all significantly over-estimate the probability of large

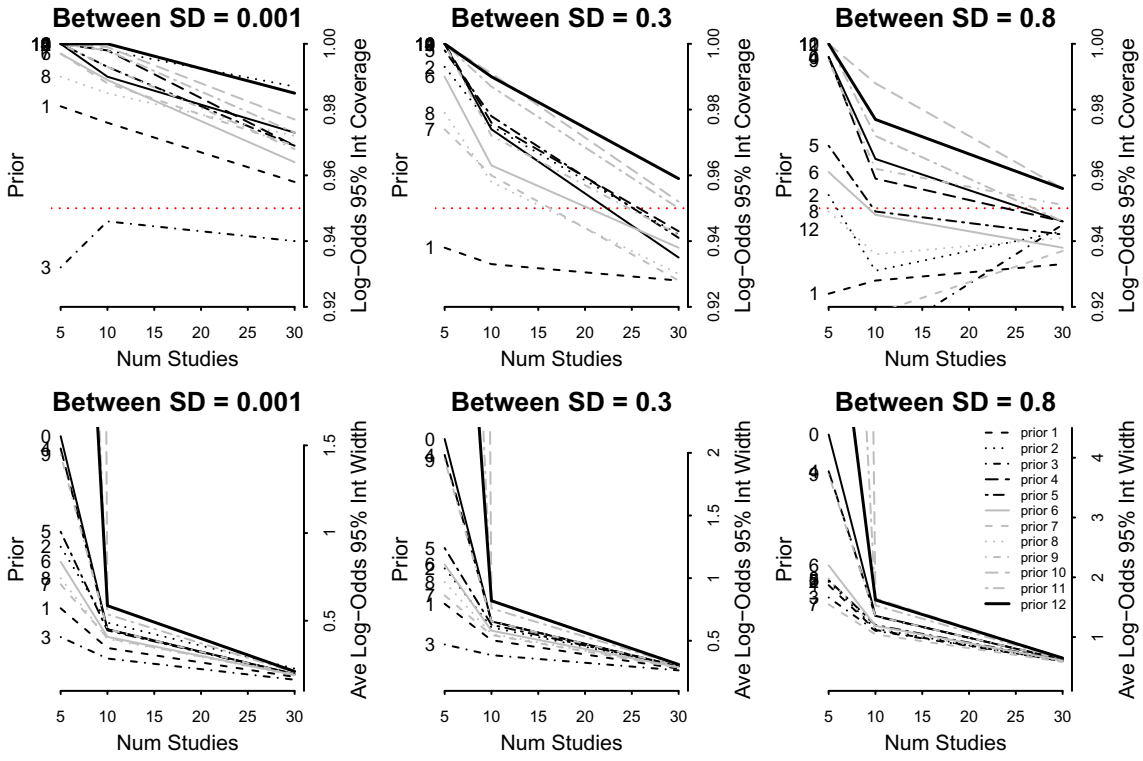

Figure 11: Coverage rates and average interval widths for 95% credible intervals for the true log-odds ratio, μ, as the number of groups increases in the simulation study of section 6.3.

values of the between study standard deviation. Poor coverage in small samples for σ_θ is obtained due to large tails of the posterior pulling the bulk of probability mass far away from 0. For the log-odds ratio, this over-estimation yields overly-wide intervals and excessive coverage probability.

- The various subjective priors, including the half-normal with standard deviation 1 (5), inverse gammas (1 and 2) can all be situationally useful but in the general case caution must be exercised.

- Approaches based around Jeffrey's and the uniform shrinkage priors (6, 7, and 8) all enjoy superior coverage for both model parameters.

To further illustrate these points, figure 13 shows the marginal posteriors for select priors using multiple draws from the model.

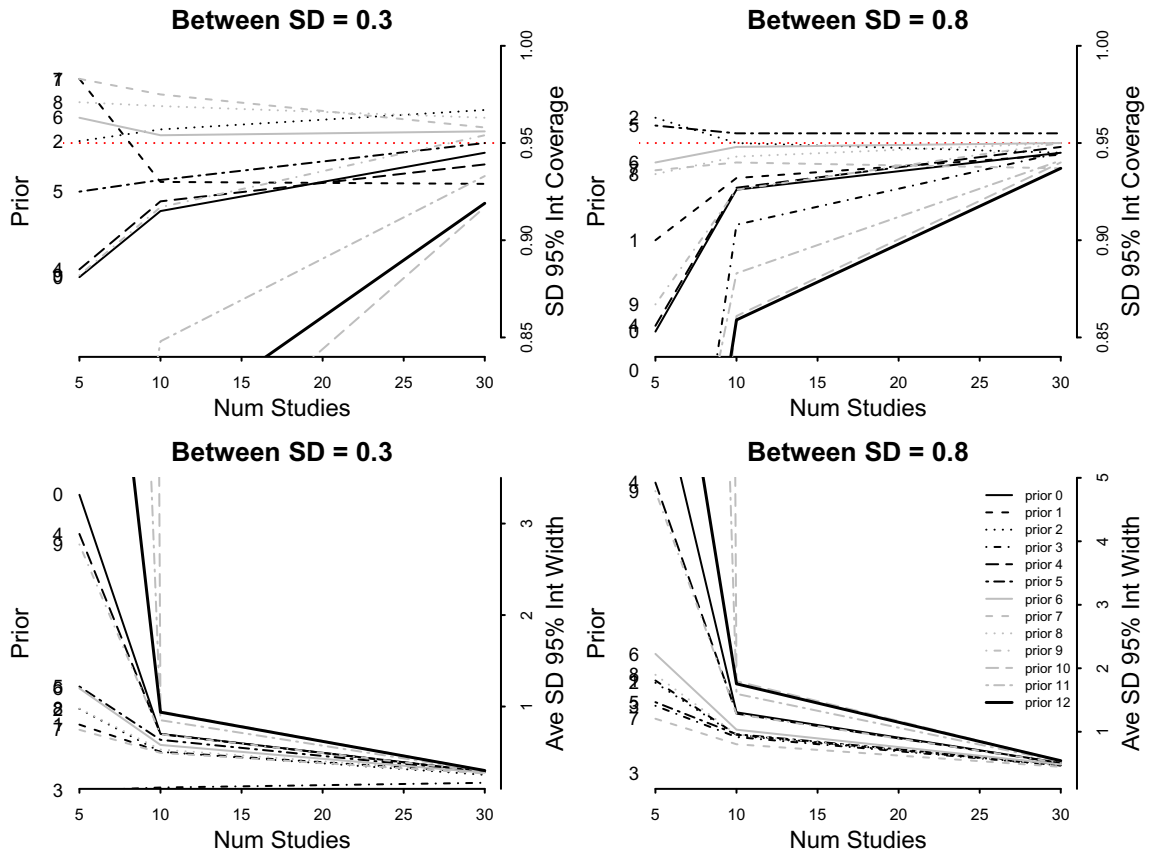

Figure 12: Coverage rates and average interval widths for 95% credible intervals for the true between study standard deviation, σ_θ, as the number of groups increases. The only prior to achieve any coverage when σ_θ was 0.001 was $p(\sigma_\theta^2) \propto 1/\sigma_\theta^2$, so that plot was suppressed.

6.4 Restricted Maximum Likelihood

The restricted maximum likelihood (REML) method for fitting hierarchical models incorporates a penalty term that corresponds to the estimation of the unmodeled coefficients, β, and serves to reduce the bias in the estimation of the residual variance, σ_y^2 (Lindstrom and Bates, 1990; Harville, 1974). A side effect of using REML is that the penalty also reduces the bias in estimation of σ_θ^2. Penalties being equivalent to priors, here we broaden our comparison to include the REML estimate. Replacing the maximization step for μ with integration in the proof of theorem 1 shows that the REML estimate is also biased downward. Given this and the correcting tendencies of the improper gamma prior we include that as a baseline.

The REML penalty itself can be seen as arising from an integration of the modeled

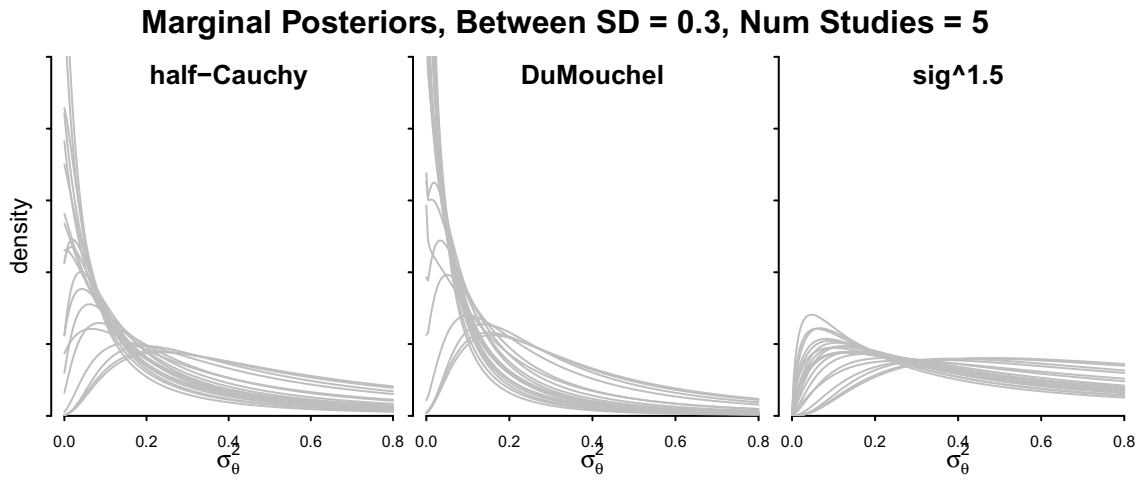

Figure 13: 20 draws from the marginal posterior corresponding to the simulation of section 6.3 for priors representative of different approaches in hierarchical modeling.

coefficients out of the likelihood using a flat prior. For the general model of section 2.3, as well as adjusting the degrees of freedom this introduces the normalizing constant:

$$\left| X^\top \left(\mathrm{I} + Z \Sigma_\theta Z^\top \right)^{-1} X \right|.$$

In fact, directly adding this term and including the prior $p(\sigma_y^2) \propto \sigma_y^P$ yields a model with an identical mode, albeit a different interpretation.

As before, we conduct a simulation study. Here, our main interest is point estimation so we forgo addressing the impact of REML on posterior credible intervals. We consider 5 different models:

1. ML - maximum likelihood
2. REML - restricted maximum likelihood
3. σ_θ - ML penalized by $p(\sigma_\theta^2) \propto \sigma_\theta$
4. σ_θ + REML - REML penalized by $p(\sigma_\theta^2) \propto \sigma_\theta$
5. σ_θ + DF - ML penalized by $p(\sigma_\theta^2) \propto \sigma_\theta$, but also the degree of freedom adjustment $p(\sigma_y^2) \propto \sigma_y^P$

While the REML estimate can be computed for a meta-analysis, it primarily is used when σ_y^2 is unknown. Thus, we move away from the meta-analysis scenario in our simulations to a full-fledged hierarchical linear model and now include two additional axes of variation: a within group sample size that largely determines the accuracy in estimating of σ_y^2, and additional linear coefficients that serve to introduce bias. The data generating model is a limited expansion of the balanced case from section 2.2 and is given by:

$$\begin{aligned} y_{ij} \mid \theta_j &\overset{\text{ind}}{\sim} \mathrm{N}(\beta_1 + x_{i1}\beta_2 + \cdots x_{iP}\beta_P + \theta_j, \sigma_y^2) \qquad i = 1, \ldots, N, \\ \theta_j &\overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_y^2 \sigma_\theta^2) \qquad j = 1, \ldots, J. \end{aligned}$$

As before, J is one of 5, 10, or 30 and σ_θ is one of 0.001, 0.3, or 0.8. n takes on values 5, 10, or 30 as well, while P can be 2, 4, or 6. σ_y is fixed at 1, as any change to it will only linearly scale the results. Finally, β itself takes on the first P values of the vector $(0.5, -0.5, 0.1, 1, -1.25, -1.5)$. For each simulation setting, 300 data sets were generated. Each covariate was drawn as i.i.d. standard normals. The five different models all were fit directly by `blmer` and the estimates of the parameters recorded.

The average of the point estimates of σ_θ over simulation runs are presented in figures 14 through 16. For the most part, the estimates follow a strict ordering across all simulation settings. Furthermore, when σ_θ is positive the REML estimate is in general negatively biased, while the various σ_θ penalized estimates are show a positive bias. A similar kind of ordering takes place for σ_y. The use of the degree of freedom correction in prior 5 seems to have a significant impact on any given estimate estimate of β, however it does not appear to introduce any bias. Representative plots for these parameters are shown in figure 17.

6.5 Recommendations

Point Estimation

For the purposes of point estimation, when the true hierarchical/modeled coefficient variation is 0 the best course of action is naturally to fix the parameter to that value and drop the corresponding coefficients from the model. This situation being unknowable, when 0 is both acceptable and plausible there is little reason to prefer an alternative to the maximum likelihood estimator. If estimating the residual standard deviation in an unbiased fashion is also important, the restricted maximum likelihood estimator recommends itself.

However, as soon as the 0 variance estimate becomes implausible or impractial, an improper gamma prior can serve to reduce the bias in the estimate of the modeled coefficient variance. Shape parameters producing priors between $p(\sigma_\theta^2) \propto \sigma_\theta$ and $p(\sigma_\theta^2) \propto \sigma_\theta^{1.5}$ are all viable. Borrowing the notion from REML estimation that it is important to compensate for a loss in degrees of freedom leads to further improvement. Combining this with our generalization to the Wishart in section 3.3, we recommend the following default prior

$$p(\Sigma_1, \ldots, \Sigma_k, \sigma_y^2) \propto \sigma_y^P \prod_{k=1}^{K} |\Sigma_k|^{\nu/2},$$

where P is the number of unmodeled coefficients and ν is between 1 and 1.5. Setting ν to 1 is not enough to guarantee a conservative estimate, while the result of setting ν to 1.5 is often too much so.

Interval Estimation

With regards to interval estimation, the situation becomes murkier. In the context of a single varying coefficient and a meta-analysis, the approximate uniform shrinkage prior, Jeffrey's, or DuMouchel and Normand's all yield reasonably accurate intervals when the true standard variance is away from the boundary. In order to come up with a "one-size fits all" approach, the following issues have to be addressed:

- multiple varying coefficients at any level
- multiple levels of variation
- having a positive probability of 0 in the posterior, not just positive density

The following sections detail an approximation that enables the efficien sampling from the posterior of an arbitrary hierarchical linear model under a flat prior. This will hopefully serve as a first step, as any prior can be tacked onto the approximation and the efficiency gains preserved.

For discussion on possible multivariate priors, Natarajan and Kass (2000) construct an approximate uniform shrinkage and an approximate Jeffrey's prior. They subsequently recommend the use of an inverse Wishart distribution using the covariates to set the scale (Kass and Natarajan, 2006). Huang and Wand (2013) generalize the inverse Wishart by using a mixture of scale parameters.

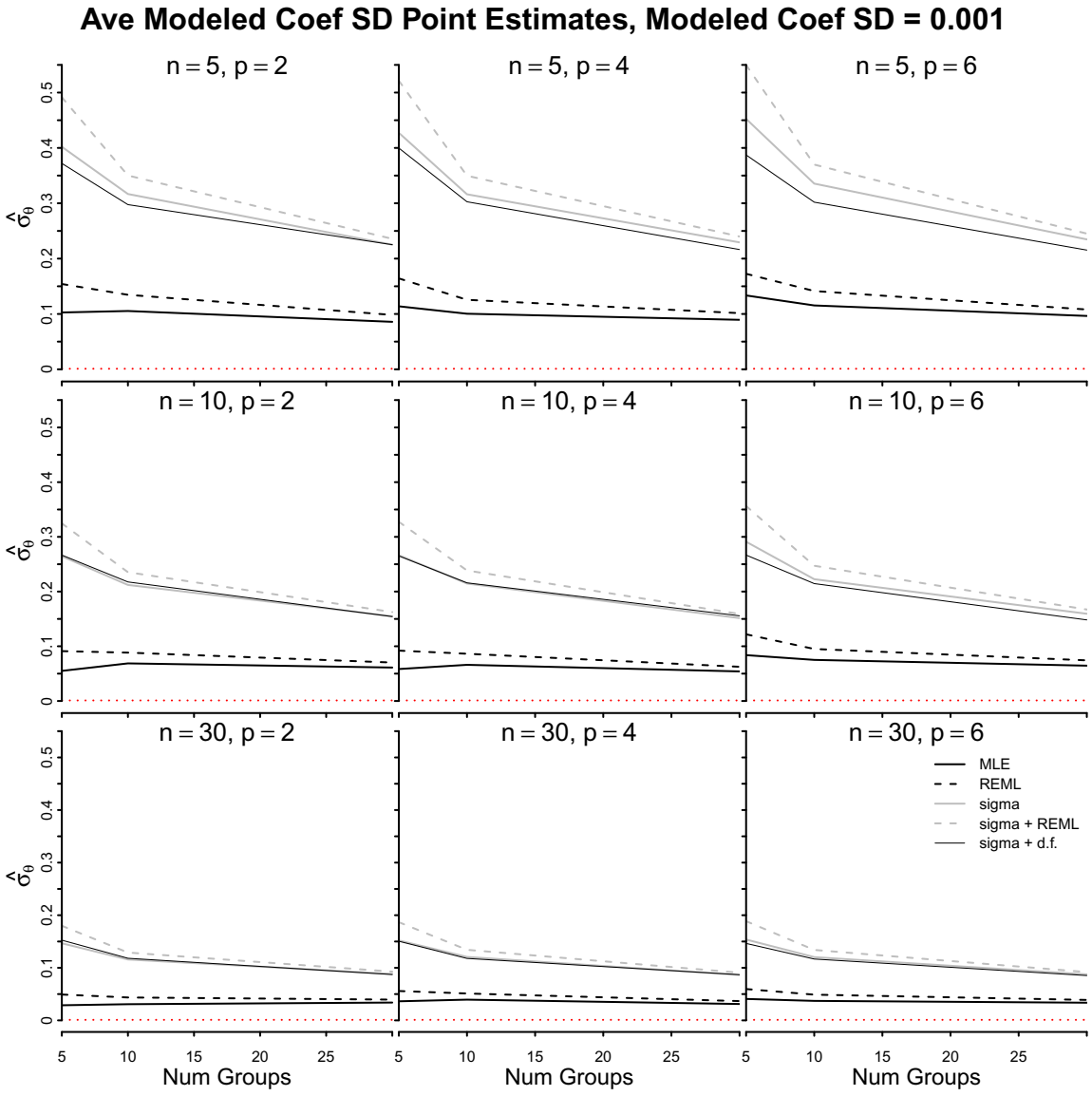

Figure 14: Average of point estimates for the standard deviation of the modeled coefficients, σ_θ, corresponding to the simulation of section 6.4. For this set of results, the true value is $\sigma_\theta = 0.001$, and is highlighted by the dotted red line. At number of groups equal to 5, 10, and 30, the line is estimated using 300 randomly generated data sets.

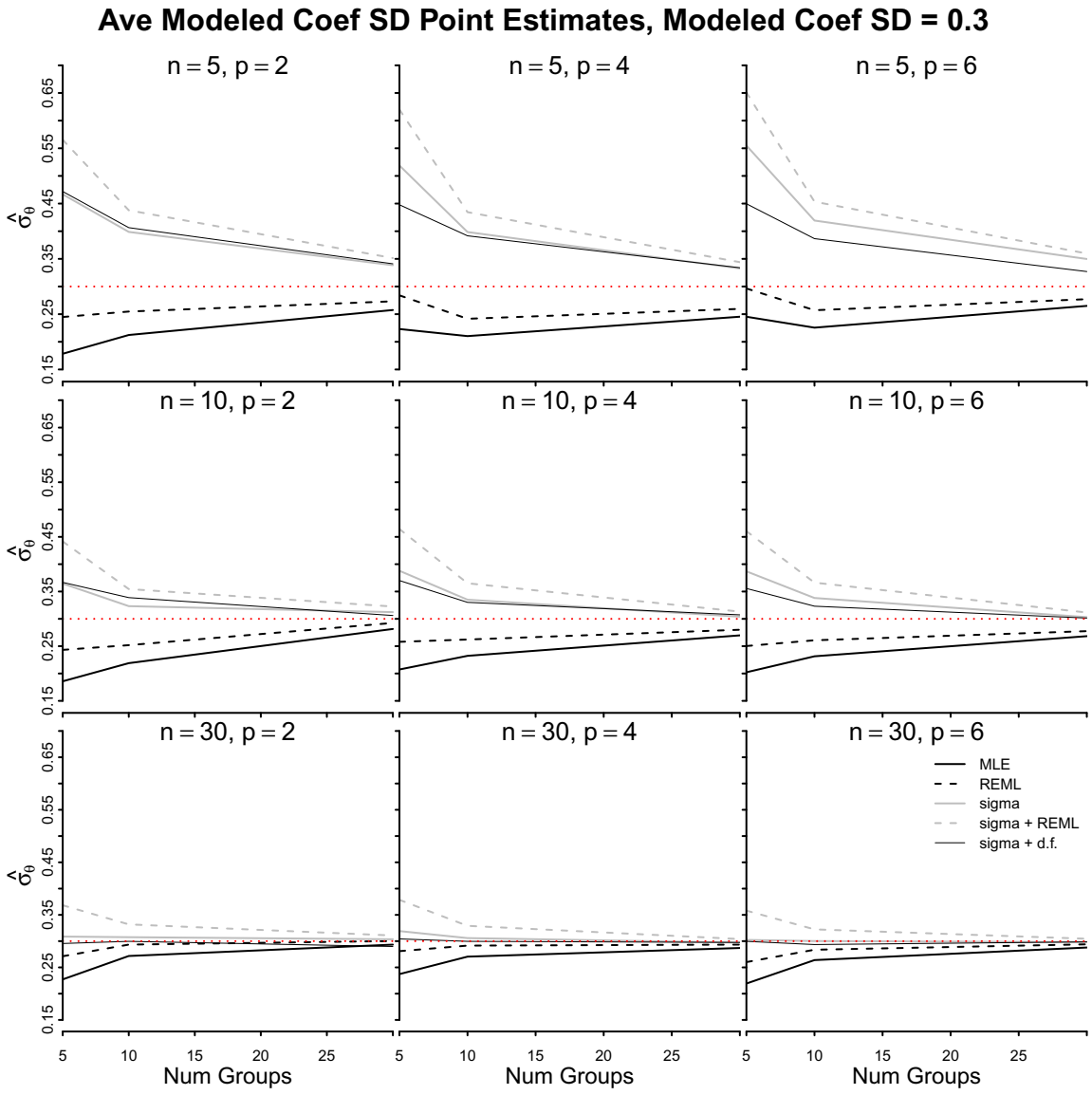

Figure 15: Average of point estimates of σ_θ for when the true value, highlighted in red, is $\sigma_\theta = 0.3$.

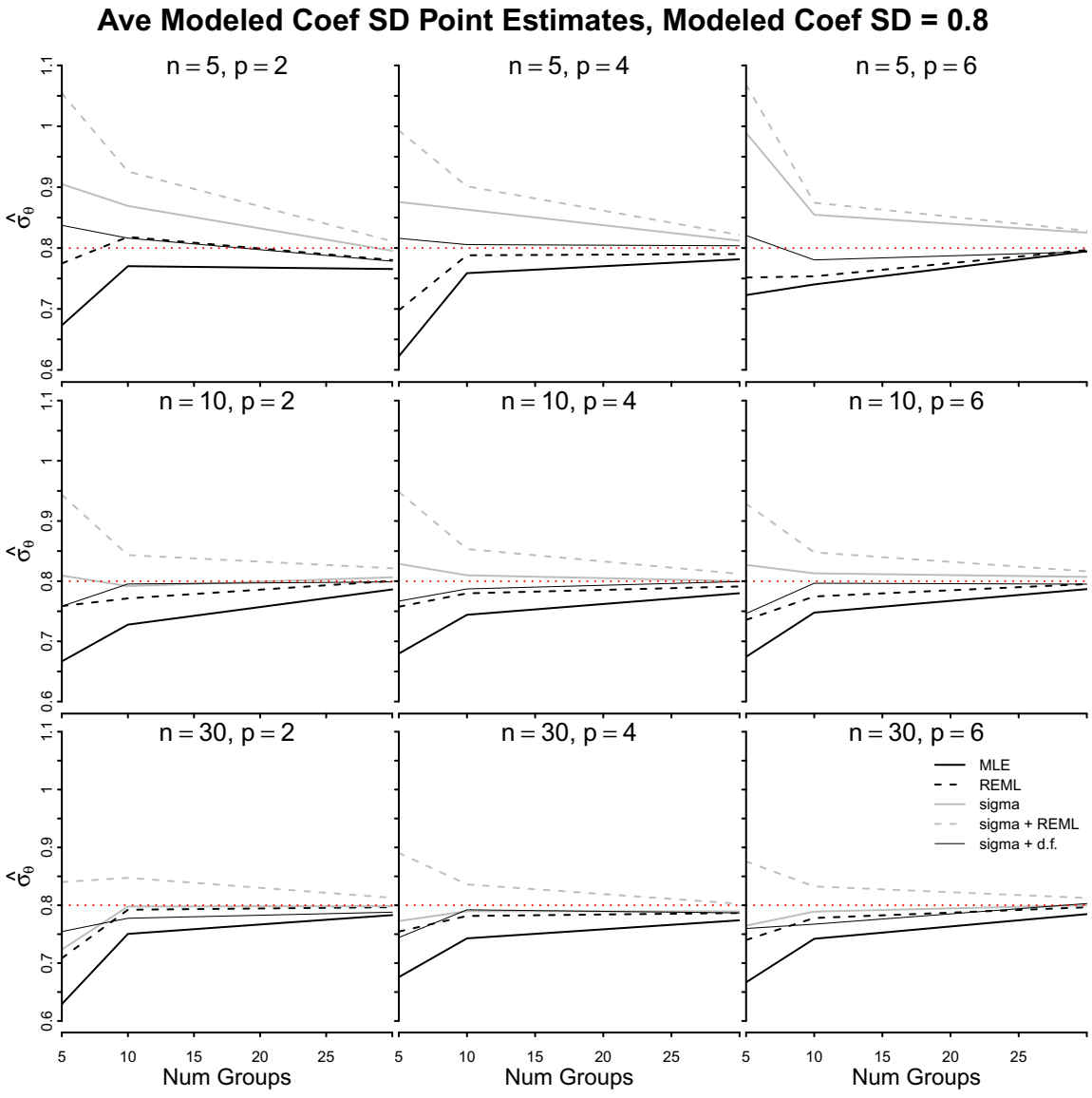

Figure 16: Average of point estimates of σ_θ for when the true value, highlighted in red, is $\sigma_\theta = 0.8$.

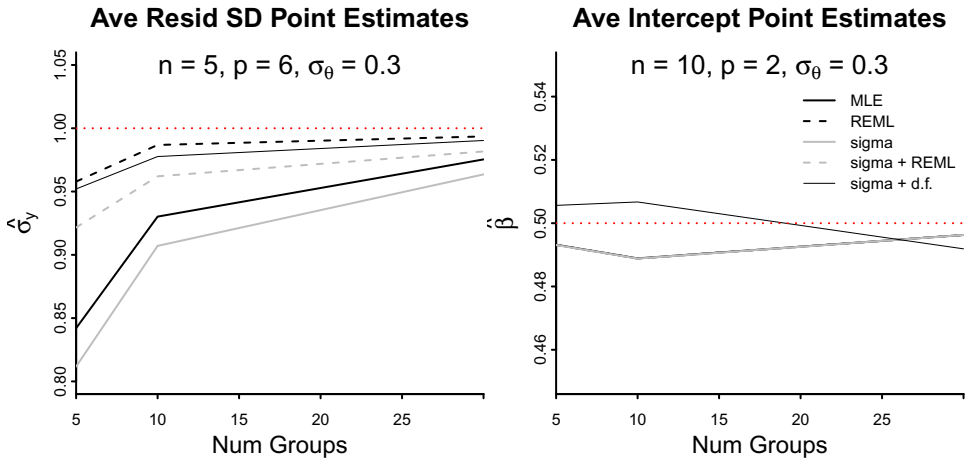

Figure 17: Representative plots of the expected value of the estimators for the residual standard deviation, σ_y, (left) and the intercept, β_1, (right).

7 Marginal Posterior/Simulation

Simulations from posterior distributions of parameters are a convenient way to assess uncertainty in a model fit, while marginal posteriors are a low-cost way to derive such simulations. Here we discuss both concepts and consider their application to the simple hierarchical linear model of section 2.2 under flat priors on the parameters.

7.1 Overview

The traditional Bayesian estimand for data modeled parametrically is the entire posterior distribution of the parameters given the data. That is, if y is the data, θ is a set of parameters, $y \mid \theta$ has density $p(y \mid \theta)$, and θ has the prior $p(\theta)$, then the quantity of interest is the function $p(\theta \mid y)$. This is typically summarized by some set of draws from this distribution, from which further inferences can be made. For example, if θ is univariate and $\tilde{\theta}^{(1)}, \ldots, \tilde{\theta}^{(L)}$ are L such draws, then a 95% credible interval for θ is given by the empirical 95% quantiles of the samples and an estimate of the posterior mean $E[\theta \mid y]$ is the average of the samples.

Considerable Bayesian literature exists concerning finding "non-informative" priors, so that the prior is in some sense the least subjective and the posterior the most faithful to the data. A good overview of the various principles or rules one might adopt in selecting a prior is given by Kass and Wasserman (1996).

Once a prior has been chosen, and hence a posterior is calculable, marginalization represents a simple way to draw samples from the posterior when there is more than one parameter. In the context of the simple hierarchical linear model, if one can obtain a draw from the marginal posterior $\sigma_\theta^2 \mid y$, it is possible to use this value to obtain a draw from $\sigma_y^2 \mid y, \sigma_\theta^2$. In turn, it is possible to plug in this value as well to obtain a draw from $\mu \mid y, \sigma_\theta^2, \sigma_y^2$, and so forth. This sequence of samples combined is a draw from the full posterior distribution $\theta, \mu, \sigma_y^2, \sigma_\theta^2 \mid y$ and hence a complex, high-dimensional operation is reducible to simple sampling operations.

This procedure rests on being able to efficiently sample from some marginal posterior at the end of a chain of integrals. The distribution that we have chosen is that of the covariance of the modeled coefficients, Σ_θ in the general case. For hierarchical linear models, conditioned on this covariance matrix the other parameters have simple distributions. For generalized linear models, a Gaussian approximation can be used. The main complication in drawing samples from this marginal posterior is largely due to the structure that the covariance has as a function of its free parameters: in general, the marginal posterior does not seem to match any known distribution.

Our approach is to approximate the marginal posterior using a class of distributions that we develop. This class is inspired by results in the simplified model for which an exact solution can be obtained, namely the beta prime distribution. A generalization of this, the Matrix-Variate Beta Prime (MVBP), has similar tail behavior as the marginal posterior. To provide a good approximation about the mode, both it and the second derivative at the mode are matched to the target distribution, yielding the Curvature-adjusted Matrix-Variate Beta Prime (CMVBP).

Once we have a high quality approximation, we can use ideas such as sampling/importance resampling to obtain independent simulations from the full posterior, or any number of Monte-Carlo based techniques if some degree of dependence is acceptable.

7.2 Simplified Model

As under an arbitrary hierarchical linear model the distribution of interest is unwieldy, we first look to the simplified model of section 2.2 for inspiration.

The joint posterior density under flat priors has the same functional form as the joint distribution of θ and y - that is equation 2. Integrating out θ from this function, we find that the joint posterior of $\mu, \sigma_y^2, \sigma_\theta^2 \mid y$ is proportional to the likelihood, equation 4. That is:

$$p(\mu, \sigma_y^2, \sigma_\theta^2 \mid y) \propto (\sigma_y^2)^{-N/2} \left(\sigma_\theta^2 + 1/n\right)^{-J/2} \times \exp\left\{-\frac{1}{2}\frac{1}{\sigma_y^2}\sum_j \left[\sum_i (y_{ij} - \bar{y}_j)^2 + \frac{1}{\sigma_\theta^2 + 1/n}(\bar{y}_j - \mu)^2\right]\right\}.$$

To make the above more concise, we use our notation of $S_w^2 = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2$ as the within-group sum of squares and $S_b^2 = \sum_j (\bar{y}_j - \bar{y})^2$ as the between-groups sum of squares. After integrating out μ as a Gaussian we obtain:

$$p(\sigma_y^2, \sigma_\theta^2 \mid y) \propto (\sigma_y^2)^{-(N-1)/2} \left(\sigma_\theta^2 + 1/n\right)^{-(J-1)/2} \exp\left\{-\frac{1}{2}\frac{1}{\sigma_y^2}\left[S_w^2 + \frac{1}{\sigma_\theta^2 + 1/n} S_b^2\right]\right\}.$$

At this point, σ_y^2 can be integrated out as having an inverse gamma distribution. We finally obtain, after some rearrangement,

$$p(\sigma_\theta^2 \mid y) \propto \frac{(\sigma_\theta^2 + 1/n)^{(N-J)/2-1}}{(\sigma_\theta^2 + 1/n + \frac{S_b^2}{S_w^2})^{(N-1)/2-1}}. \tag{13}$$

7.3 Beta-Prime Distribution

The marginal posterior distribution can be identified as a shifted, truncated beta prime. As $n \to \infty$, the shift and truncation both vanish. The beta prime family is also known as a Pearson Type VI or beta distribution of the second kind.

For our purposes, there are two key ways of thinking about a random variable with such a distribution. The beta prime arises when one takes a random variable with a beta distribution and divides it by 1 minus itself. Equivalently, the beta prime also arises as a sample from an inverse gamma distribution whose scale is itself sampled from a gamma, or to say a scale-mixture of inverse gammas. As this relates to the simple hierarchical model,

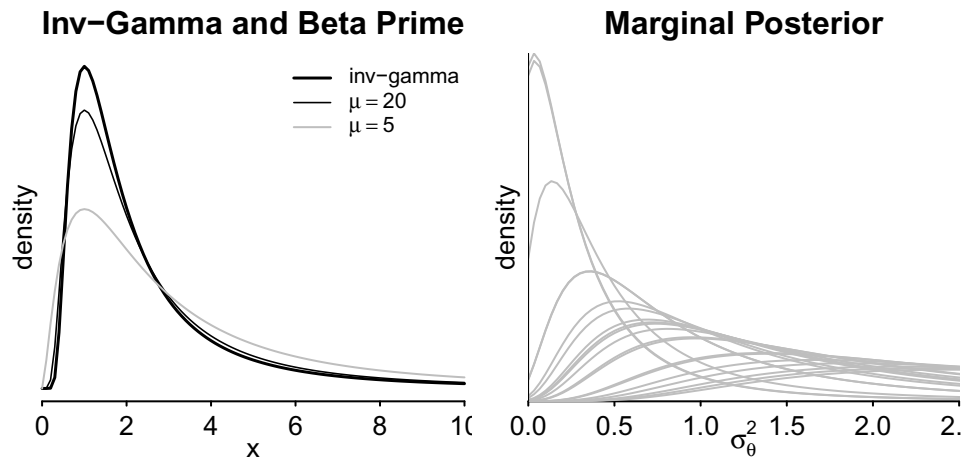

Figure 18: a) Left - Comparison of inverse gamma with a shape parameter of 2 and a mode of 1 (thick black line) and beta prime distributions (thin black and gray). For the beta prime, the shape parameter at the inverse gamma level is fixed and the scale determined to fix the mode at 1, while the remaining shape parameter, μ, varies. b) Right - 20 simulated marginal posteriors, $p(\sigma_\theta^2 \mid y)$, for $n = 5$, $J = 8$ and $\sigma_\theta = \sigma_y = 1$.

$$\begin{aligned}
\beta &\sim \text{beta}\left(\frac{N-J}{2}, \frac{J-1}{2} - 1\right), \\
\sigma^2 &= \frac{S_b^2}{S_w^2}\frac{\beta}{1-\beta}, \\
\sigma_\theta^2 \mid y &= \sigma^2 - 1/n \quad \text{given } \sigma^2 \geq 1/n,
\end{aligned}$$

or,

$$\begin{aligned}
\gamma &\sim \text{gamma}\left(\frac{N-J}{2}, \text{scale} = \frac{S_b^2}{S_w^2}\right), \\
\sigma^2 \mid \gamma &\sim \text{inv} - \text{gamma}\left(\frac{J-1}{2} - 1, \text{scale} = \gamma\right), \\
\sigma_\theta^2 \mid y &= \sigma^2 - 1/n \quad \text{given } \sigma^2 \geq 1/n.
\end{aligned}$$

The first perspective is useful in mathematical proofs, as integrals over the marginal posterior density can be related to the familiar form of the beta function. It also permits

inverse-CDF sampling when desired. The second gives a simple, generalizable approach that is useful when considering more complicated models. Figure 18a highlights the connection with the inverse gamma, while part b shows several samples of the marginal posterior for repetitions of a simulation study.

After having identified the marginal posterior, it can be noted that it will be proper under mild conditions, namely that the sample size exceed the number of groups, and that the number of groups is greater than 3.

7.4 Simplified Model Simulation Procedure

Given that we can sample directly from the marginal posterior in this case, that suggests the following procedure to obtain samples from the full posterior:

0. Calculate the sums of squares, S_w^2 and S_b^2.

1. Sample a scale γ from a gamma distribution with shape equal to $(N-J)/2$ and scale equal to S_b^2/S_w^2.

2. Sample $\sigma_\theta^2 \mid y$ by drawing from an inverse gamma distribution with shape equal to $(J-1)/2-1$ and scale equal to γ. Subtract from this $1/n$. If the result is less than or equal to 0, go back to 1.

3. Sample a value of $\sigma_y^2 \mid \sigma_\theta^2, y$ from an inverse gamma distribution with shape $(N-1)/2-1$ and scale $\frac{1}{2}\left(S_w^2 + (\sigma_\theta^2 + 1/n)^{-1} S_b^2\right)$.

4. Sample $\mu \mid \sigma_\theta^2, \sigma_y^2, y$ as normal with mean $\bar{y}$ and variance $\sigma_y^2(\sigma_\theta^2 + 1/n)/J$.

5. Finally, sample each $\theta_j \mid \sigma_\theta^2, \sigma_y^2, \mu, y$ as independently normal with a mean of $(\bar{y}_j - \mu)\frac{\sigma_\theta^2}{\sigma_\theta^2+1/n}$ and a variance of $\frac{\sigma_y^2}{n}\frac{\sigma_\theta^2}{\sigma_\theta^2+1/n}$.

The last two steps can be easily combined by sampling μ and θ together from their jointly normal, conditional posterior.

8 General Model Posterior Simulations

The results of the preceeding section let us to sample directly from the marginal posterior in the simplified case and guarantee that the operation has a valid statistical interpretation provided there are a sufficient number of observations. In order to do similar for the general model, we derive the marginal posterior and several of its properties, after which we scale up the beta prime generate a reasonable approximation. We can then pin this approximating distribution so that it has the same mode as the marginal posterior, and adjust its second derivative at the mode in the same fashion.

8.1 Marginal Posterior

Once again applying flat priors to the parameters, the marginal posterior of Σ_θ can be obtained through a sequence of integrations. Starting with the likelihood in equation 10 we can proceed as in the simple model, first integrating out β as Gaussian and then σ_y^2 as an inverse gamma. The details of this are in the appendix. The result, after some careful rearrangements, is of the form:

$$p(\Sigma_\theta \mid y) \propto \frac{\left|\Sigma_\theta^{-1}\right|^{1/2}}{\left|\Sigma_\theta^{-1} + Z^\top(\mathrm{I} - X(X^\top X)^{-1}X^\top)Z\right|^{1/2}} \times \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} Z & X \\ \Sigma_\theta^{-1/2} & 0 \end{bmatrix} \begin{bmatrix} Z^\top Z + \Sigma_\theta^{-1} & Z^\top X \\ X^\top Z & X^\top X \end{bmatrix}^{-1} \begin{bmatrix} Z^\top & \Sigma_\theta^{-\top/2} \\ X^\top & 0 \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|^{-(n-p-2)}$$

To make this expression cleaner and to assist analysis, we denote as H_x the matrix that enables one to project a vector onto the column space of its subscript, *e.g.* $H_X = X(X^\top X)^{-1}X^\top$. This is also known as the "hat" matrix. Let $H_x^\perp = \mathrm{I} - H_x$ be the orthogonal complement to H_x. Furthermore, let $\tilde{y}$ be the augmented response vector, $(y, 0)$, and $\tilde{X}(\sigma_\theta)$ be the augmented design matrix above. Then we have:

$$p(\Sigma_\theta \mid y) \propto \frac{|\Sigma_\theta^{-1}|^{1/2}}{|\Sigma_\theta^{-1} + Z^\top H_X^\perp Z|^{1/2}} \|\tilde{y} - H_{\tilde{X}(\sigma_\theta)}\tilde{y}\|^{-(n-p-2)}. \tag{14}$$

At this point, the posterior is comprised of two distinct parts that we will refer to as, on the left, a determinant term and on the right a sum of squared residuals term.

The residual term arises from projecting the vector $\tilde{y}$ onto the column space of the augmented design, so it is bounded above by the sum of the squares of y and below by a constant that is only zero if the observations are perfectly predictable - a measure 0 set that we ignore. Taking limits as the eigenvalues of Σ_θ go to infinity show that it is the determinant term that drives the behaviors in the tails of the distribution.

Considering this determinant term, it bears similarities to a simple generalization of the beta prime distribution that we consider in the next section. The generalization would be exactly equal to the determinant term if it were not for the complicated way in which Σ_θ depends on its free parameters.

Finally, whether or not this marginal posterior is proper depends, in a non-trivial fashion, on $Z^\top H_X^\perp Z$. Results from simulation show that the main criterion seems to be that there are a sufficient number of groups at any level to estimate the number of varying coefficients.

Marginal Posterior Derivatives

In order to approximate the marginal posterior, we calculate the first and second derivatives of its logarithm. We have opted for an analytic result due to our experience that there is often little information about the components of Σ_θ, leading to poor numeric estimates. The corresponding calculations are contained in the appendix, but we provide an overview due to the wider applicability of some of the intermediate steps. Practically speaking the main consequence is that we are able to obtain these matrices in an accurate and efficient fashion.

The derivatives are taken in two broad stages, the first corresponding to Σ_θ as an abstract quantity and the second determining how the components of Σ_θ vary along with its free parameters. For the first step, given any function $f(\Sigma_\theta)$, $\mathrm{d}f/\mathrm{d}\Sigma_\theta$ follows directly from any

suitable definition of a matrix derivative. Furthermore, the result is specific to the specific marginal model and thus has little general applicability.

The second calculation, $\mathrm{d}\Sigma_\theta/\mathrm{d}\sigma_\theta$, depends only on the structure of Σ_θ and is consequently a sparse matrix of 0s and 1s. To go from taking a derivative with respect to Σ_θ to one with respect to σ_θ then involves multiplying by this matrix, which is equivalent to taking particular summations of the elements of the former. This derivative can be used in any problem with an equivalent matrix structure.

$\mathrm{d}\Sigma_\theta/\mathrm{d}\sigma_\theta$ also procedes in stages. If "diag" is an operator constructing a block diagonal matrix of its arguments, from the model specification we have that $\Sigma_\theta = \mathrm{diag}\left(\mathrm{I}_{J_1} \otimes \Sigma_1, \ldots, \mathrm{I}_{J_K} \otimes \Sigma_K\right)$. To take a derivative with respect to each Kronecker product-block as a whole is relatively straightforward. With the context of the model, this corresponds to taking a derivative of the joint covariance with respect to the covariance of a grouping factor.

Within any grouping factor, by writing out the pattern of repetitions of the elements of Σ_k, the appropriate pattern of 0s and 1s can be derived. Combining this with the above then yields the derivative.

$\mathrm{d}\Sigma_\theta/\mathrm{d}\sigma_\theta$ itself is a large, sparse matrix that could itself be coded up and multiplied directly. Obtaining the correct indicies for non-zeroes is largely a matter of careful calculation. However, as multiplication by this matrix results in a summation, further simplification is possible. Within the derivatives of the marginal posterior, two kinds of terms exist: Kronecker products of arbitrary matrices, $A \otimes B$, and the outer product of a vector, $vv^\top$. As a final step in calculating the total derivative, we determine what summations are required to compute $[\mathrm{d}\Sigma_\theta/\mathrm{d}\sigma_\theta]^\top (A \otimes B)\mathrm{d}\Sigma_\theta/\mathrm{d}\sigma_\theta$ and $[\mathrm{d}\Sigma_\theta/\mathrm{d}\sigma_\theta]^\top vv^\top \mathrm{d}\Sigma_\theta/\mathrm{d}\sigma_\theta$. This allows us to avoid having to explicitly compute the Kronecker and outer products respectively.

8.2 Matrix-Variate Beta Prime

To approximate the marginal posterior, we generalize the beta prime distribution to the class of covariance matrices by utilizing the scale-mixture of inverse-gammas interpretation. After

having done so we consider a change of variables that permits us to generate samples with the same mode and second derivative at the mode as the marginal posterior.

Definition

Much as the beta prime distribution can be seen as a scale mixture of inverse-gamma random variables, a generalization of the beta prime distribution can be given by a scale mixture of inverse Wishart random variables. An early instance and generalization of this distribution is given by Mathai (2005). If $\tilde{\Sigma}$ is a $d \times d$ covariance matrix and

$$\tilde{\Sigma} \mid \Psi \sim \text{inv} - \text{Wishart}\left(\nu, \text{scale} = \Psi\right),$$
$$\Psi \sim \text{Wishart}\left(\mu, \text{scale} = C\right),$$

then

$$p(\tilde{\Sigma}) = \frac{|\tilde{\Sigma}^{-1}|^{(\nu+d+1)/2}}{|\tilde{\Sigma}^{-1} + C^{-1}|^{(\nu+\mu)/2}} \frac{1}{|C|^{\mu/2}\mathrm{B}_d(\nu/2, \mu/2)}.$$

In this last expression, the normalizing constant is a multivariate generalization of the beta function, $\mathrm{B}_d(a, b) = \Gamma_d(a)\Gamma_d(b)/\Gamma_d(a+b)$. $\Gamma_d(a)$ is the normalizing constant for the Wishart distribution, the multivariate gamma function. We call a random covariance matrix with this density a Matrix-Variate Beta Prime (MVPB).

Figure 19 shows the marginal distributions over the variances and the correlation induced by an MVBP. The two degrees of freedom parameters permit more mass to be placed in the tails of the distribution than the inverse Wishart otherwise allows. As the degrees of freedom at the scale level (μ) go to infinity, the scale matrix itself is drawn with infinite precision and the simple inverse Wishart is recovered.

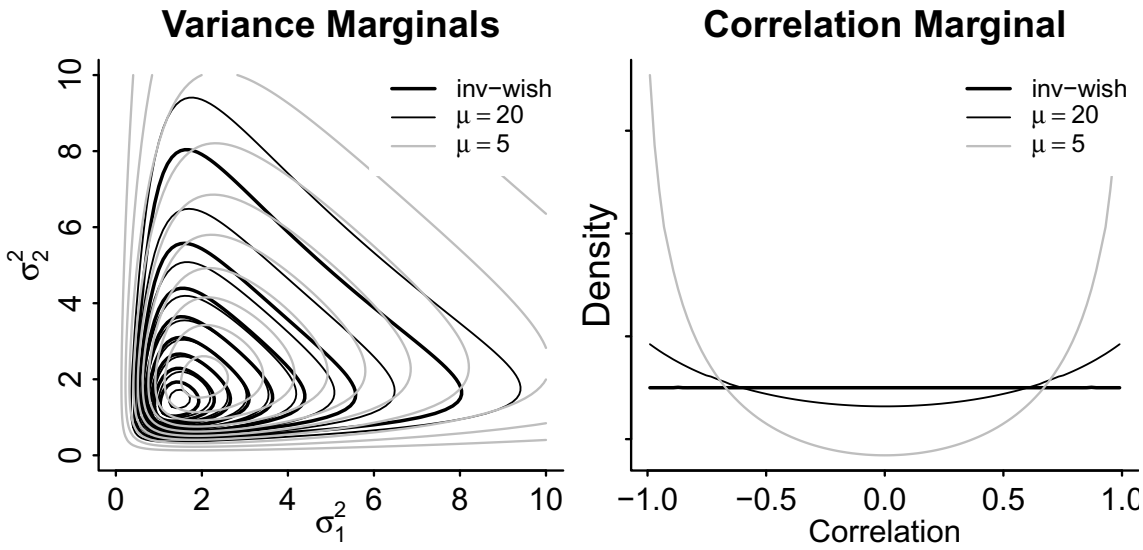

Figure 19: Marginal distributions of variances (left) and correlation (right) for the inverse Wishart distribution and matrix-variate beta prime (MVBP). The inverse Wishart shown is over a two dimensional covariance with a mode at the identity matrix and has three degrees of freedom. The MVBP similarly has its mode at the identity and three degrees of freedom at the inverse Wishart level (ν), but has varying degrees of freedom at the scale level (μ).

Matching the Marginal Posterior

In the following, we use the term "curvature" to refer to the second derivative of the log density at its mode. In order to make the MVBP match the marginal posterior as closely as possible, we match the two at their modes and adjust the curvature. By taking the gradient of the log-density and setting it equal to the 0 vector, we find that the matrix-variate beta prime is maximized at $M = \frac{\mu-d-1}{\nu+d+1}C$. For convenience in matching modes we reparameterize in terms of this matrix.

At this point, we would like to be able to set the curvature of the MVBP to that of the marginal posterior at the mode and somehow solve the resulting equation. Setting aside for the moment concerns over the parameterization of Σ as a covariance, the second derivative of the log density is a $d^2 \times d^2$ matrix. Having fixed the mode at M, the MVBP has only two free parameters with which to solve a system of d^4 equations. To address this, we consider a change of variables.

A curvature adjusting transformation for a random vector is any one such the second

derivative of the log density at the mode is set to a specific matrix. This broad category of transformations can be limited by a sequence of assumptions, in practice generally being reduced to a linear function. If $\tilde{\eta}$ is a random vector whose distribution is maximized at $\hat{\eta}$, then for the appropriate choice of matrix A, $\eta = A(\tilde{\eta} - \hat{\eta}) + \hat{\eta}$ will have the desired distribution. This procedure is also known as a "whitening" transformation, followed by a "coloring" one.

Attempting to apply a linear transformation to Σ as a matrix only nets an additional d^2 parameters - not enough to equate curvatures. In order to apply a full-rank linear transformation to a matrix-valued random variable, the matrix must first be vectorized. That is, we find the matrix A such that

$$\operatorname{vec}\Sigma = A\operatorname{vec}(\tilde{\Sigma} - M) + \operatorname{vec} M$$

has a distribution whose second derivative at M is equal to a fixed matrix.

This transformation gives us the Curvature-Adjusted Matrix Variate Beta Prime, or CMVBP. If A is invertible and $\tilde{\Sigma}(\Sigma)$ is the inverse transformation, the CMVBP has the density

$$p(\Sigma) = \frac{\left|\tilde{\Sigma}(\Sigma)^{-1}\right|^{(\nu+d+1)/2}}{\left|\tilde{\Sigma}(\Sigma)^{-1} + \frac{\mu-d-1}{\nu+d+1}M^{-1}\right|^{(\nu+\mu)/2}} \frac{|A|^{-1}\left(\frac{\mu-d-1}{\nu+d+1}\right)^{d\mu/2}}{|M|^{\mu/2}\mathrm{B}_d(\nu/2,\mu/2)}. \tag{15}$$

As in the simplified model the result was exact, attempting to match curvatures will leave the distribution unchanged. By adopting the transformation in the general case, we remain able to sample directly from the marginal posterior. As before, it is also the case that Σ may not be positive semi-definite so that in sampling, a rejection step is required.

8.3 Applying the CMVBP

Grouping Factors

As for hierarchical linear model there is no restriction on the number of grouping factors, the posterior over Σ_θ is often over multiple matrices: $\Sigma_1, \ldots, \Sigma_k$. On the other hand, the CMVBP enables us to produce samples for only single covariance matrices.

To bridge the gap, we propose two different methods. In the first, we relax our requirement that samples be derived independently and implement a Gibbs-within-Metropolis sampler, taking draws succesively from the conditional distributions $\Sigma_k \mid \Sigma_{[-k]}, y$.

The second method is to independently draw K different matrices from MVBPs, vectorize them all, and then simultaneously adjust their curvatures. We adopt this method when describing a simulation procedure, and detail its specifics below.

Degrees of Freedom

In having gone from the matrix-variate beta prime on a $d \times d$ covariance matrix to the curvature adjusted version, an additional d^4 parameters were introduced. Consequently, the degrees of freedom parameters ν and μ now are free to be specified.

In picking them, we again look to the simplified model. For that case, and using the scale-mixture of inverse gammas interpretation, the scale level variate had a gamma distribution with a shape parameter of $(N-J)/2$. Given this, the variance was then drawn with this scale from an inverse gamma with shape $(J-1)/2-1$. Here the 1 in $J-1$ arose from integrating out the mean parameter which generalizes to the number of unmodeled coefficients, P. In terms of the full model, there is one grouping factor having J_1 members and $Q_1 = 1$ types of modeled coefficients.

One possible generalization would be to generate a scale matrix from a Wishart distribution with degrees of freedom $\mu = N - Q_1 \times J_1 + Q_1 - 1$ and then a covariance from an inverse Wishart with $\nu = J_1 - P - Q_k - 1$. As we are approximating at each of K levels,

this gives the decision rules to set parameters as:

$$\mu_k := N - Q_k \times (J_k - 1) - 1,$$
$$\nu_k := J_k - Q_k - P - 1.$$

As a consequence of this approach, the simulation procedure is only valid provided that $J_k > Q_k + P + 1$, or that for each factor, there are a sufficient number of groups. The efficacy of setting the parameters in this fashion is evaluated in section 8.5.

8.4 Full Model Simulation Procedure

We arrange σ_θ so that it is of the form $(\text{vec}(\Sigma_1)^\top, \ldots, \text{vec}(\Sigma_K)^\top)^\top$. Using the second method described above, we can obtain importance samples from the marginal posterior by taking the following steps:

0. Find the marginal posterior mode, $\hat{\sigma}_\theta$, and calculate the curvature:

$$-I(\hat{\sigma}_\theta) = \frac{\mathrm{d}^2}{(\mathrm{d}\sigma_\theta)^2} \log p(\sigma_\theta \mid y)\Big|_{\sigma_\theta = \hat{\sigma}_\theta}.$$

   The mode can be obtained by first finding the MLE of σ_θ and using that as a starting point for an optimization routine.

1. Sample K scale matrices from Wishart distributions with degrees of freedom $\mu_k = N - Q_k \times (J_k - 1) - 1$ and mode $\hat{\Sigma}_k$.

2. Sample matrices $\tilde{\Sigma}_1, \ldots, \tilde{\Sigma}_K$ independently from inverse Wisharts with degrees of freedom $\nu_k = Q_K \times (J_k - 1) - P - 1$.

3. Adjust the curvature of the MVBP random variables just sampled.

(a) Create the matrix

$$\Psi = \operatorname{diag}\left(\tfrac{2}{\nu_1+Q_1+1}\tfrac{\nu_1+\mu_1}{\mu_1-Q_1-1}\hat{\Sigma}_1 \otimes \hat{\Sigma}_1, \ldots, \tfrac{2}{\nu_K+Q_K+1}\tfrac{\nu_K+\mu_K}{\mu_K-Q_K-1}\hat{\Sigma}_K \otimes \hat{\Sigma}_K\right),$$

deriving from the curvature of the MVBP.

(b) Create the vector $\tilde{\sigma} = (\operatorname{vec}(\tilde{\Sigma}_1)^\top, \ldots, \operatorname{vec}(\tilde{\Sigma}_K)^\top)^\top$.

(c) Create the matrix $A = I(\hat{\sigma}_\theta)^{-1/2}\Psi^{-1/2}$.

(d) Apply the transformation

$$\sigma = A(\tilde{\sigma} - \hat{\sigma}) + \hat{\sigma}.$$

(e) Reconstitute sampled matrices $\Sigma_1, \ldots, \Sigma_k$ from the vector σ.

4. Ensure that all the resulting matrices are positive semi-definite. If not, go back to step 1.

5. Compute the importance weight $\log p(\Sigma_\theta \mid y)/\log p(\Sigma_\theta)$.

From this, we can generate a large number of importance samples and their weights. To produce a draw from the full posterior, we then

1. Sample a $\Sigma_\theta \mid y$ from the pool of importance samples.

2. Sample $\sigma_y^2 \mid \Sigma_\theta, y$ from an inverse gamma distribution with shape $(N-P)/2-1$ and inverse-scale $1/2$ times the sum of the squared residuals for that Σ_θ.

3. Sample $(\beta, \theta)^\top \mid \sigma_y^2, \Sigma_\theta, y$ as jointly normal with a mean of the projection of $\tilde{y}$ onto the column space of $\tilde{X}(\sigma_\theta)$ and a covariance of $\sigma_y^2\left(\tilde{X}^\top(\sigma_\theta)\tilde{X}(\sigma_\theta)\right)^{-1}$.

For a model fit under a boundary-avoiding prior, as recommended previously, the above procedure will always be valid. When fit against, the likelihood alone, it is possible that marginal mode or the information fail to be invertible. Handling these cases is an issue of ongoing concern, however simulating from a slightly penalized version may be viable.

8.5 Simulation Study

To test our procedure, we conduct a simulation study and compare the approximate with the exact posterior. In addition, as the setting is fabricated it permits us to evaluate inferences made using the simulations in an exact fashion. We compare coverage rates for interval estimates of the modeled coefficients using our simulation technique with other methods.

Popular interval estimation techniques for hierarchical models often fall under the name of empirical Bayes confidence intervals. A catalog of early efforts including a reasonably accurate, *ad hoc* approach is given by Morris (1983). Laird and Louis (1987) presents the parametric bootstrap, which combines resampling with the parametric model. An offshoot of this literature is the estimation of uncertainty for small-area estimators, including Prasad and Rao (1990); Jiang et al. (2002); Hall and Maiti (2006). The most fundamental approach is the naive EB estimate.

Many techniques are designed around simple sampling models and cannot directly be extended to the case with mulitple grouping factors. With this in mind, we limit ourselves the naive EB estimator and the type III parametric bootstrap of Laird and Louis. For the parametric bootstrap, we use the refitted parameter estimates when plugging into the EB covariance estimate, which the accounts for the uncertainty in estimating Σ_θ. Because these methods rely on point estimates that are known to be down-ward biased, we also include variants that use the mode calculated from a penalized likelihood with the recommended prior from section 6.5, $p(\sigma_y^2, \Sigma_\theta) \propto \sigma_y^P |\Sigma_\theta|^{1/2}$. For comparison with simulation based techniques, we include an approach that does not account for the uncertainty in Σ_θ, and thus requires no approximation. We call this an "estimated posterior" approach. Finally we include the full simulation technique proposed here.

We use a simple varying intercepts/varying slopes model with a single covariate and balanced group membership. Rather than write θ as the totality of the modeled coefficients,

		EB	EB+	Par Boot	PBS+	Est Post	CMVBP
Coverage (95%)	Intercept	0.75	0.75	0.72	0.73	0.91	0.96
	Slope	0.77	0.84	0.67	0.73	0.85	0.96
Ave Width	Intercept	2.06	2.08	1.95	1.98	3.22	4.22
	Slope	1.77	1.94	1.46	1.61	2.14	2.87

Table 6: Results of a simulation study showing coverage percentages of 95% intervals as well as average interval widths for various techniques. "EB" are naive empirical Bayes intervals, "Par Boot" is a type III parametric boot strap, and "Est Post" estimates the posterior by using the maximum likelihood estimate of Σ_θ. The "+" versions of the first two techniques use estimates of parameters from a penalized version using the recommendations of section 6.5. Coverages and widths were averaged together for the multiple $J = 10$ intercept and slope coefficients.

we denote the slopes by γ. Then we have

$$\begin{aligned} y_{ij} \mid \theta, \gamma &\overset{\text{ind}}{\sim} \mathrm{N}\left(\beta_0 + \theta_j + (\beta_1 + \gamma_j)x_{ij}, \sigma_y^2\right) \quad i = 1, \ldots, n, \\ \begin{pmatrix} \theta_j \\ \gamma_j \end{pmatrix} &\overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_y^2 \Sigma_\theta) \qquad\qquad\qquad\qquad j = 1, \ldots, J. \end{aligned}$$

In arbitrary fashion, we construct Σ_θ from standard deviations $\sigma_1 = 1.5$ and $\sigma_2 = 0.75$ with correlation $\rho = 0.16$. x is generated for each sample run as independent standard normals. $n = 8$, $J = 10$, and finally, $\beta_0 = 3, \beta_1 = -0.5$, and $\sigma_y = 1.5$. 500 repetitions in total are generated. To assess the practical utility of the procedure, we recorded coverage rates of 95% intervals for the modeled coefficients. For any method that requires it, 100 iterations were used.

Figure 20 demonstrates the quality of the approximate solution by overlaying the marginal distributions of the variances for both the posterior and the CMVBP.

Table 6 contains the results of coverage rates as well as the average interval width. Results are further broken down into those for intercept coefficients and those for slopes, as the type of quantity had a significant impact on coverage. Unsurprisingly, the additional uncertainty added by drawing simulations over the covariance of the modeled coefficients increased both the widths of the intervals, and the coverage probabilities.

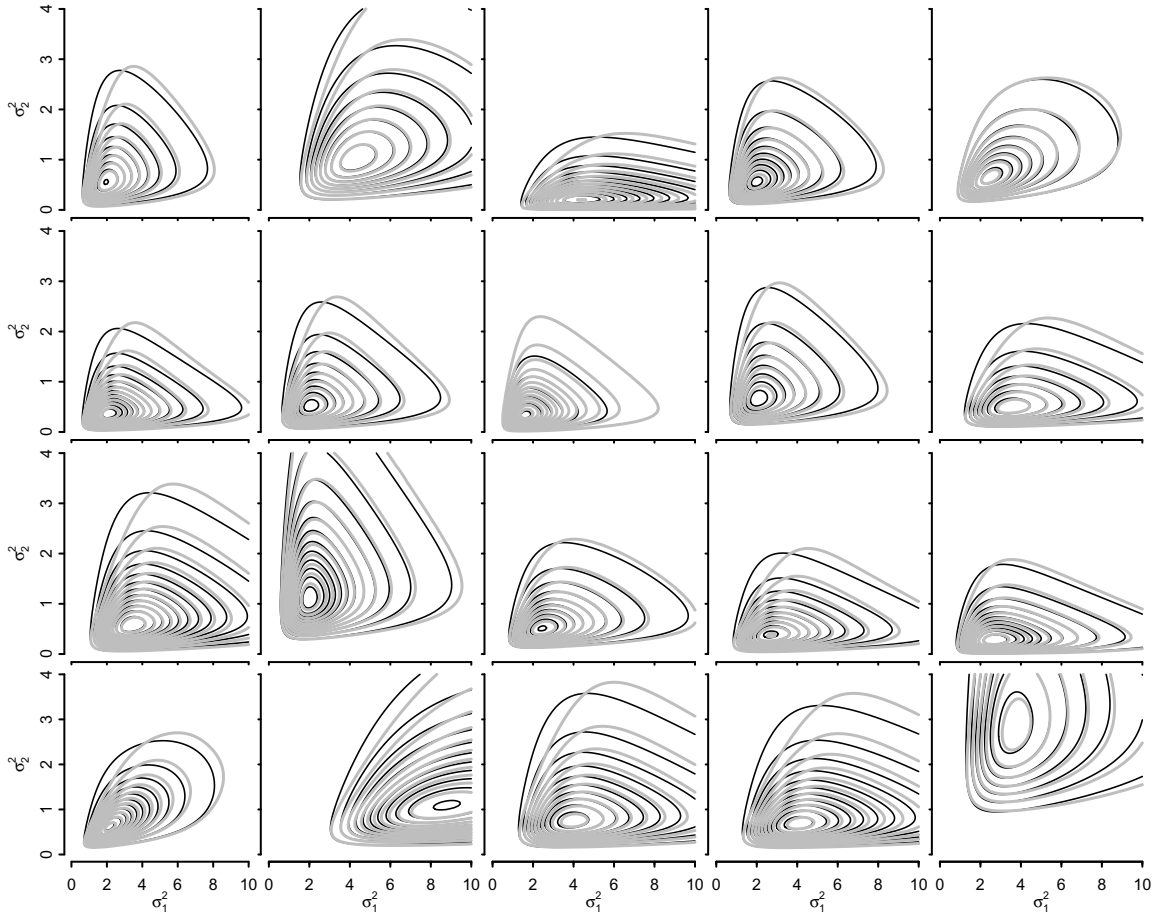

Figure 20: Comparison of marginal posterior (black lines) and CMVBP (gray lines) for 20 random samples from the model in section 8.5. In addition to integrating out the other model parameters, the correlation in Σ_θ was marginalized over as well so that the contours represent the distributions over variances alone.

The peformance of the naive empirical Bayes confidence interval is not surprising, it being well known that the method fails to account for uncertainty in estimating the parameters. The results for the type III parametric bootstrap are less expected. By design, it should account for uncertainty in estimating Σ_θ by refitting the parameters from fake data generated using the maximum-likelihood fit. Since the MLE is downward biased, using the corrective measures of previous sections can, and does, improve the quality of the approximation. We can further improve accuracy by using these for the refit steps as well, yielding coverages of 0.74 and 0.82 for the intercept and slope respectively. Since these remain deficient, we are left noting that the procedure of bootstrapping the EB intervals is itself guaranteed

to under-estimate the uncertainty, and that bootstrapping a corrected interval would yield better results.

Finally, we wish to note that by chosing essentially a point-prior for Σ_θ and then assuming a flat one, we have constructed as adversarial a simulation setting as possible. As the truth becomes more like our model, we would expect even better results.

9 Simulation Software

The `sim` function in the R programming language implements the hierarchical model posterior simulation technique of section 8 and is available in the `arm` package.

9.1 Calling `sim`

The `sim` function is a generic operating on different kinds of model fits. For those fit by `lme4`, it has two arguments: 1) a fitted model S4 object of class `mer`, and 2) optionally the number of simulations required, defaulting to 100. It returns an object of class `sim.mer` with the following attributes:

1. `fixef` - a matrix of random samples of the unmodeled coefficients with dimension equal to the number of samples times the number of unmodeled coefficients

2. `ranef` - a named list of elements corresponding to the grouping factors in the model, each being a array of samples of the modeled coefficients with dimension equal to the number of samples times the number of groups within the factor times the number of coefficients varying at that level

3. `resid.var` - a vector of samples of the residual variance, when applicable and `NULL` otherwise

4. `ranef.cov` - a named list of with elements corresponding to the grouping factors and each item being an array with the leading dimension being of size equal to the number of samples and each sub-indexed item being a randomly sampled matrix

9.2 Examples

We illustrate the use of `sim` in the context of a simple hierarchical linear model. To generate the data:

```
> set.seed(0);
> J <- 8;
> n <- 5;
> N <- n * J;

> beta <- c(3, -0.5);
> g    <- rep(1:J, rep(n, J));
> sigma.y   <- 1.5;
> sigma.the.11 <- 2^2;
> sigma.the.22 <- 0.75^2;
> rho          <- 0.2;
> sigma.the.21 <- rho / sqrt(sigma.the.11 * sigma.the.22);
> Sigma.the <- matrix(c(sigma.the.11, sigma.the.21,
                        sigma.the.21, sigma.the.22), 2, 2);

> theta <- sigma.y * t(chol(Sigma.the)) %*% matrix(rnorm(2 * J), 2, J);
> x <- rnorm(N);
> y <- (beta[1] + theta[1,g]) + (beta[2] + theta[2,g]) * x +
       rnorm(N, 0, sigma.y);
```

To obtain the maximum likelihood fit:

```
> M1 <- lmer(y ~ 1 + x + (1 + x | g), REML = FALSE);
> display(M1);
lmer(formula = y ~ 1 + x + (1 + x | g), REML = FALSE)
            coef.est coef.se
(Intercept)  3.47     0.84
x           -0.24     0.61
```

```
Error terms:
 Groups   Name        Std.Dev. Corr
 g        (Intercept) 2.30
          x           1.58     0.12
 Residual             1.05
---
number of obs: 40, groups: g, 8
AIC = 170.9, DIC = 159
deviance = 158.9
```

To generate 100 samples from the marginal posterior, simply pass the `mer` object to `sim()`.

```
> M1.sim <- sim(M1);
```

`sim()` Output

To access the results from `sim`, use the operator.

```
> print(names(attributes(M1.sim)));
[1] "fixef"     "ranef"     "resid.var" "ranef.cov"
[5] "class"
```

By illustration, the first 5 rows of the samples of the unmodeled coefficients are given by:

```
> M1.sim@fixef[1:5,];
     (Intercept)      x
[1,]         1.4 -0.028
```

```
[2,]         4.0  2.155
[3,]         2.4 -1.117
[4,]         1.5 -3.284
[5,]         6.3  0.513
```

The modeled coefficients are accessed similarly:

```
> names(M1.sim@ranef);
[1] "g"

> dim(M1.sim@ranef$g);
[1] 100   8   2

> M1.sim@ranef$g[1:5,,"(Intercept)"];
         1     2      3     4     5     6     7      8
[1,]  4.29  5.01  2.817 -0.57  0.14  5.37 -0.47  0.017
[2,]  2.39  1.66  0.011 -3.28 -1.61  2.42 -2.78 -2.590
[3,]  3.09  3.48  0.923 -2.57 -0.60  4.33 -1.61 -0.479
[4,]  3.79  5.28  2.841 -1.10 -0.36  5.96  0.14  0.015
[5,] -0.18 -0.58 -2.819 -6.21 -3.95 -0.18 -5.86 -4.119

> M1.sim@ranef$g[1:5,,"x"];
          1    2     3     4     5     6     7     8
[1,]  0.066 1.90 -2.14 -0.43  2.75  1.19 -1.66 -0.66
[2,] -2.561 0.75 -4.82 -1.64 -0.04 -2.05 -2.74 -2.76
[3,]  1.224 3.39 -1.66  2.54  3.70  1.60 -0.25  0.37
[4,]  1.841 5.33  0.53  2.99  6.03  3.88  2.32  2.58
```

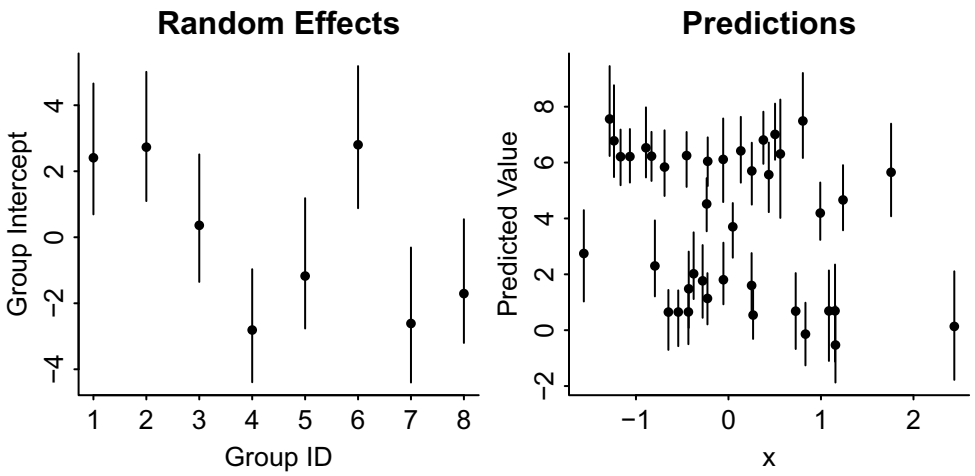

Figure 21: a) Estimates and 95% credible intervals for random effects, and b) estimates and 95% credible intervals for predictions.

```
[5,] -2.491 1.44 -3.46  0.21  1.22 -0.88 -2.04 -1.69
```

9.3 Assessing Uncertainty in Parameters

Using the simulations, we can directly assess our uncertainty in the estimates of the modeled coefficients. For example, to construct a 95% confidence interval for the difference in averages for the first two groups:

```
> quantile(M1.sim@ranef$g[,1,"(Intercept)"] -
           M1.sim@ranef$g[,2,"(Intercept)"],
           c(0.025, 0.975));
2.5%  98%
-1.6  1.6
```

As an example of the utility of posterior samples, the following code snippet estimates the standard errors of latent intercepts to produce Figure 21a.

```
> lower <- apply(M1.sim@ranef$g[,,"(Intercept)"], 2, quantile, 0.025);
> upper <- apply(M1.sim@ranef$g[,,"(Intercept)"], 2, quantile, 0.975);
> estimates <- ranef(M1)$g[,"(Intercept)"];

> numGroups <- length(estimates);
> xValues <- rbind(1:numGroups, 1:numGroups, rep(NA, numGroups));
> yValues <- rbind(       ower,       upper, rep(NA, numGroups));

> plot(xValues, yValues, type = "l",
       main = "Random Effects", xlab = "Group ID", ylab = "Group Intercept");
> points(1:numGroups, estimates);
```

Assessing Uncertainty in Predictions

When passed a `sim.mer` object and the `mer` object used to create it, the `fitted` function uses the samples to generate a collection of model predictions corresponding to the observed covariates. For as many simulations were generated, there are as many predictions as observations, so that in the example above with 40 data points:

```
> M1.fitted <- fitted(M1.sim, M1);
> dim(M1.fitted);
[1]  40 100
```

We can quickly and simply display the estimates and their uncertainty as a function of the continuous covariate. The result of the following is in Figure 21b.

```
> lower <- apply(M1.fitted, 1, quantile, 0.025);
> upper <- apply(M1.fitted, 1, quantile, 0.975);
> estimates <- fitted(M1);
```

```
> numObsv <- length(estimates);
> xValues <- rbind(    x,     x, rep(NA, numObsv));
> yValues <- rbind(lower, upper, rep(NA, numObsv));


> plot(xValues, yValues, type = "l",
       main = "Predictions", xlab = "x", ylab = "Predicted Value");
> points(x, estimates);
```

10 Example

In order to demonstrate the results of the preceeding sections, we work through a comprehensive example. For this, a working copy of R is required as are the blme and arm packages, as well as all of their prerequisites.

10.1 Cognitive Assessments in Rural Kenya

We return to the motivating example of section 1.1, of using cognitive assessments to evaluate the effect of dietary treatments in rural Kenya as published by Neumann et al. (2003). The data have been made available by Weiss (2005).

Cleaning the Data

```
> dataURL <-
    "http://rem.ph.ucla.edu/mld/data/tabdelimiteddata/cognitive.txt"
> kenya <- read.csv(dataURL, sep = "\t");
```

To replicate the 0 variance estimate, we clean the data to include only those children who have been in the study for more than 20 months.

```
> ids         <- unique(kenya$id);
> numChildren <- length(ids);


> lastObsRows <- sapply(1:numChildren, function(i) {
    childRows <- kenya$id == ids[i];
    child <- kenya[childRows,];

    maxTime <- max(child$relmonth, na.rm = TRUE);
    if (maxTime <= 20) return(rep(FALSE, nrow(child)));
```

```
    return (child$relmonth == maxTime);
  });


> lastObsRows <- unlist(lastObsRows);
> children <- kenya[lastObsRows & !is.na(kenya$ravens),];
```

Finally, we standardize the regression inputs.

```
> standardize <- function(x) (x - mean(x)) / sd(x);
> children$ravens.z <- standardize(children$ravens);
> children$age_at_time0.z <- standardize(children$age_at_time0);
```

Model Fitting

In order to make a direct comparison with the pure-likelihood inference from section 1.1, we start by fitting a varying intercepts model with "dummy variables" for the different treatments and a covariate for the age at the beginning of the study. We can obtain the maximum likelihood estimate by running:

```
> M0 <-  lmer(ravens.z ~ treatment + age_at_time0.z + (1 | schoolid),
              children, REML = FALSE);
> display(M0);
lmer(formula = ravens.z ~ treatment + age_at_time0.z + (1 | schoolid),
    data = children, REML = FALSE)
                coef.est coef.se
(Intercept)      -0.05     0.09
treatmentcontrol  0.09     0.13
treatmentmeat     0.25     0.13
treatmentmilk    -0.10     0.12
```

```
age_at_time0.z    0.01     0.04


Error terms:
 Groups   Name        Std.Dev.
 schoolid (Intercept) 0.00
 Residual             0.99
---
number of obs: 496, groups: schoolid, 12
AIC = 1412.1, DIC = 1398
deviance = 1398.1
```

Σ_θ is reported under error terms above, specifically `schoolid (Intercept)`.

To find a non-degenerate estimate, we can use `blmer`. As we are not interested in priors on the unmodeled coefficients, we leave that as flat and use the default of $p(\sigma_y^2, \Sigma_\theta) \propto \sigma_y^2 \sigma_\theta^{1.5}$. Given the high sample size and low number of unmodeled coefficients, the prior on the modeled coefficient covariance alone is probably sufficient.

```
> M1 <- blmer(ravens.z ~ treatment + age_at_time0.z + (1 | schoolid),
              children, REML = FALSE,
              cov.prior = gamma(2.5, 0),
              resid.prior = gamma(3, 0, 'sd;'));


> display(M1);
blmer(formula = ravens.z ~ treatment + age_at_time0.z + (1 |
    schoolid), data = children, REML = FALSE, cov.prior = "gamma(shape = 2.5, rate =
    fixef.prior = NULL, var.prior = "gamma(3, 0, 'sd')")
                coef.est coef.se
(Intercept)     -0.03     0.11
```

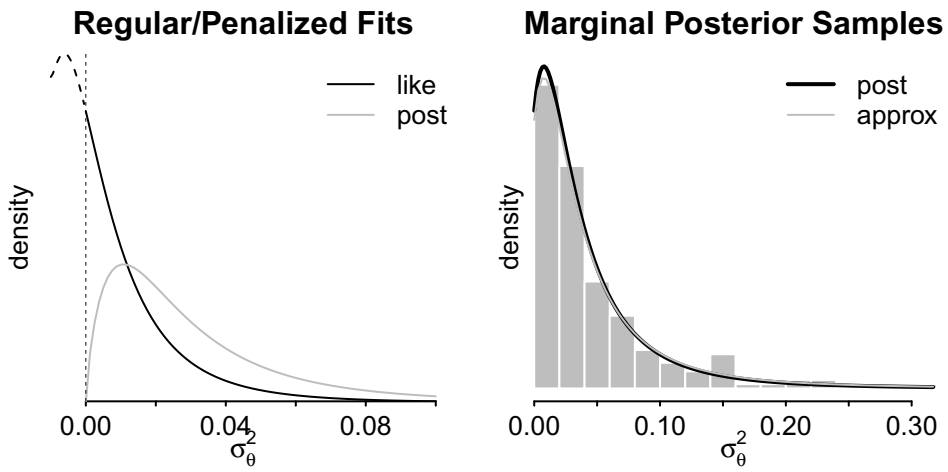

Figure 22: Left: the profiled likelihood and posterior as a function of Σ_θ, which here reduces to a single variance and right: the marginal posterior, the matrix-variate beta prime approximation, and importance samples.

```
treatmentcontrol  0.06     0.16

treatmentmeat     0.23     0.16

treatmentmilk    -0.13     0.15

age_at_time0.z    0.01     0.05


Error terms:

 Groups   Name        Std.Dev.

 schoolid (Intercept) 0.10

 Residual             0.99

---

number of obs: 496, groups: schoolid, 12

AIC = 1413.5, DIC = 1400

deviance = 1399.5
```

The hierarchical standard deviations and variances reported above have the common scale factor multiplied in so that the can be interpreted directly, that is $0.1 = \sigma_y \Sigma_\theta^{1/2}$.

The estimates are quite similar except for the now-positive variance component. A visual comparison of the profiled likelihood and posterior of Σ_θ is shown in the left panel of figure 22, which demonstrates that the posterior mode is consistent and that the estimate of 0 is premature. In fact, if we conduct a sequence of point-null hypothesis tests against an otherwise unconstrained alternative, negative 2 times the difference between the two log likelihoods will have an approximate chi-squared distribution with one degree of freedom. Inverting this yields a 95% confidence interval which we truncate to $[0, 0.026]$. The posterior mode found by `blme` is $\Sigma_\theta = 0.011$.

Uncertainty

The marginal posterior, inherently representing the uncertainty accumulated from the estimation of the other parameters, has considerably heavier tails. It, and the approximating distribution used in simulation is graphed in figure 22b. We can take samples from this using `sim` to generate a to obtain a one-sided 95% credible interval for Σ_θ in the following fashion:

```
> M0.sims <- sim(M0);
> quantile(M0.sims@ranef.cov$schoolid, 0.95);

 95%
0.18
```

Inferences

Figure 23a shows the prediction lines for the two schools originally highlighted in figure 1. The posterior mode leads to similar inference in this sense, but enables further comparisions. A graph similar to figure 23b for the maximum likelihood model would consist of 12 dots at 0 with 0 width intervals. In similar fashion, we can now make statements about the hierarchical components that would before have been non-sensical. For example, a 95% credible interval for the difference in intercepts between school 2 and school 8 is given by $[-0.14, 0.69]$.

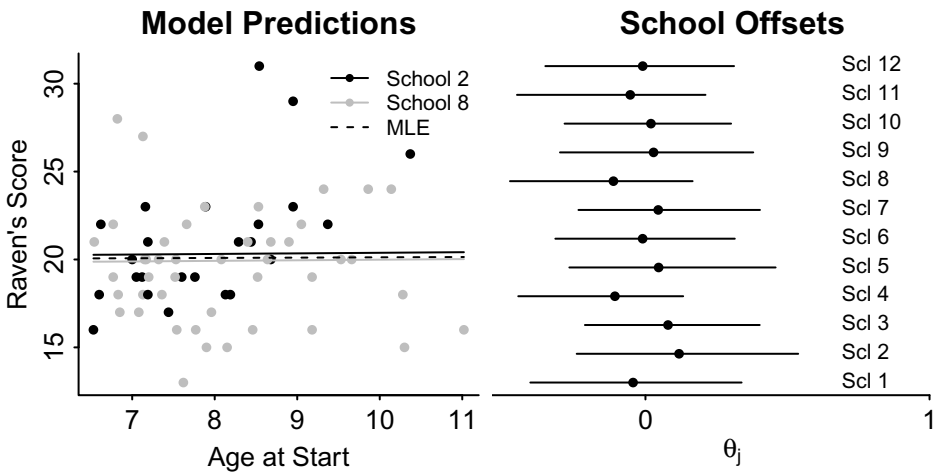

Figure 23: a) Comparison of the maximum likelihood estimated regression and the posterior mode. The MLE is unable to capture that a difference between might exist, while the Bayesian method finds them to be similar, but distinct. b) The estimated school offsets and uncertainties. These compare favorably to the empirical averages shown in figure 1.

Varying Intercepts/Varying Slopes

The simply varying intercepts model is used largely to show the procedure for fitting in blme and using `sim`. More practically of interest, albeit less graphically compelling, is the varying intercept and slope model. Proceeding as before with the maximum likelihood fit:

```
> M0 <-  lmer(ravens.z ~ treatment + age_at_time0.z +
              (1 + age_at_time0.z | schoolid),
              children, REML = FALSE);
> display(M0);
lmer(formula = ravens.z ~ treatment + age_at_time0.z + (1 + age_at_time0.z |
    schoolid), data = children, REML = FALSE)
                coef.est coef.se
(Intercept)      -0.04     0.10
treatmentcontrol  0.09     0.13
treatmentmeat     0.26     0.14
treatmentmilk    -0.07     0.13
```

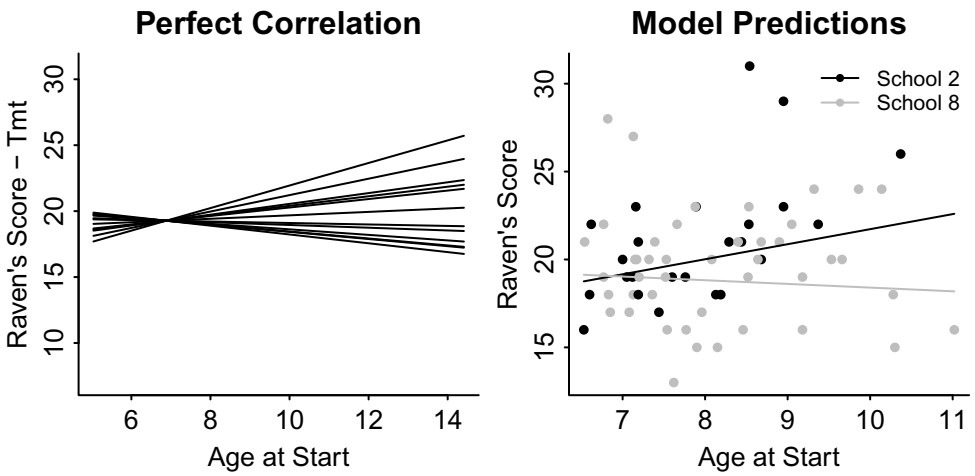

Figure 24: a) Fitted regression lines from the degenerate, maximum likelihood varying intercept/varying slope model exhibiting the fitted perfect correlation. To demonstrate this more effectively, the treatment coefficient is ignored. b) The MLE lines for groups 2 and 8.

```
age_at_time0.z    0.05     0.08


Error terms:
 Groups   Name           Std.Dev. Corr
 schoolid (Intercept)    0.11
          age_at_time0.z 0.20     1.00
 Residual                0.97
---
number of obs: 496, groups: schoolid, 12
AIC = 1410.5, DIC = 1392
deviance = 1392.5
```

This estimate is also degenerate, although in a less-obvious fashion. Correlations of ± 1 correspond to covariances on the boundary but lack the simple interpetation of an estimate of zero variance. Instead, they imply that as a group's intercept "moves up" in value, its slope moves up in a perfectly proportional fashion. Similar to the zero variance case, this yields an unwieldy regression object and overstates the confidence in some comparisons.

This perfect correlation is exemplified in figure 24a, in which the different group regressions are graphed while having subtracted out the treatment effect. Figure 24b shows the lines for just groups 2 and 8, along with the data. While a positive correlation seems justified, a perfect one is extreme.

Finally, the `blmer` fit slightly walks back from perfect correlation and assessing uncertainty is straightforward with `sim`:

```
> M1 <- blmer(ravens.z ~ treatment + age_at_time0.z +
              (1 + age_at_time0.z | schoolid), children, REML = FALSE,
              cov.prior = wishart(3.5, Inf),
              resid.prior = gamma(3, 0, 'sd'));


> display(M1);
blmer(formula = ravens.z ~ treatment + age_at_time0.z + (1 +
    age_at_time0.z | schoolid), data = children, REML = FALSE,
    cov.prior = "wishart(df = 3.5, scale = Inf)", fixef.prior = NULL,
    var.prior = "gamma(3, 0, 'sd')")
                 coef.est coef.se
(Intercept)      -0.03     0.10
treatmentcontrol  0.09     0.14
treatmentmeat     0.25     0.15
treatmentmilk    -0.07     0.14
age_at_time0.z    0.05     0.08

Error terms:
 Groups   Name           Std.Dev. Corr
 schoolid (Intercept)    0.12
          age_at_time0.z 0.21     0.93
```

```
 Residual                0.97
---
number of obs: 496, groups: schoolid, 12
AIC = 1411, DIC = 1393
deviance = 1393.0


> M0.sims <- sim(M0);
> quantile(sqrt(M0.sims@ranef.cov$schoolid[,1,1]), 0.95);
> quantile(sqrt(M0.sims@ranef.cov$schoolid[,2,2]), 0.95);
 95%
0.49
 95%
0.63
```

The approximation used to obtain simulations is detailed in figure 25.

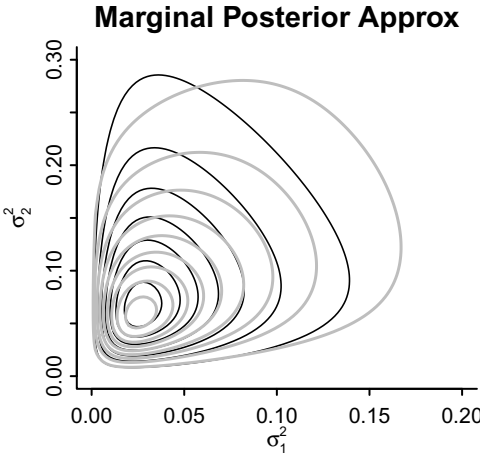

Figure 25: Matrix-variate beta prime approximation to the marginal posterior of the variances for the bivariate Raven's score hierarhical model. Contours are obtained by numerically integrating out the correlation. In this case, as the marginal posterior mode is at the boundary the approximation was hand-tuned in the following fashion: 1) a marginal mode was estimated by penalizing the likelihood with a term of $|\Sigma_1|^{0.05}$, 2) the second derivative was obtained using this penalized model, and 3) the maximum eigenvalue from the second derivative was divided by 10 so that its magnitude was roughly that of the second derivative obtained from the marginal mode in the un-penalized model.

11 References

References

Douglas Bates, Martin Maechler, and Ben Bolker. lme4: Linear mixed-effects models using s4 classes. 2012.

James O Berger and John Deely. A bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology. *Journal of the American Statistical Association*, 83(402):364–373, 1988.

James O Berger and William E Strawderman. Choice of hierarchical priors: admissibility in estimation of normal means. *The Annals of Statistics*, 24(3):931–951, 1996.

Yeojin Chung, Sophia Rabe-Hesketh, Vincent Dorie, Andrew Gelman, and Jingchen Liu. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, Forthcoming, 2013. URL `http://www.stat.columbia.edu/~gelman/research/published/draft4_12.pdf`.

Michael J Daniels. A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, 27(3):567–578, 1999.

W. DuMouchel and C. Waternaux. Discussion of "hierarchical models for combining information and for meta-analysis" by cn morris and sl normand. *Bayesian statistics*, 4: 321–344, 1992.

William DuMouchel. Hierarchical bayes linear models for meta-analysis. Technical report, 1994.

William DuMouchel and Sharon-Lise Normand. Computer-modeling and graphical strategies for meta-analysis. *Meta-analysis in medicine and health policy*, pages 127–178, 2000.

Alan E. Gelfand, Susan E. Hills, Amy Racine-Poon, and Adrian F. M. Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, 85(412):pp. 972–985, 1990. ISSN 01621459. URL `http://www.jstor.org/stable/2289594`.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

Paul P Glasziou, Chris Del Mar, Sharon L Sanders, and M Hayem. Antibiotics for acute otitis media in children [review]. *The Cochrane Database of Systematic Reviews*, (1), 2004.

Harvey Goldstein. *Multilevel statistical models*, volume 922. Wiley, 2011.

Peter Hall and Tapabrata Maiti. On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):221–238, 2006.

David A Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383–385, 1974.

Bruce M Hill. Inference about variance components in the one-way model. *Journal of the American Statistical Association*, 60(311):806–825, 1965.

Alan Huang and MP Wand. Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2):439–452, 2013.

Jiming Jiang, Partha Lahiri, and Shu-Mei Wan. A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics*, 30(6):1782–1810, 2002.

Robert E Kass and Ranjini Natarajan. A default conjugate prior for variance components in generalized linear mixed models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):535–542, 2006.

Robert E. Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):pp. 1343–1370, 1996. ISSN 01621459. URL `http://www.jstor.org/stable/2291752`.

Nan M Laird and Thomas A Louis. Empirical bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82(399):739–750, 1987.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

Paul C Lambert, Alex J Sutton, Paul R Burton, Keith R Abrams, and David R Jones. How vague is vague? a simulation study of the impact of the use of vague prior distributions in mcmc using winbugs. *Statistics in medicine*, 24(15):2401–2428, 2005.

Mary J Lindstrom and Douglas M Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687, 1990.

Jun S Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

AM Mathai. A pathway to matrix-variate gamma and normal densities. *Linear Algebra and Its Applications*, 396:317–328, 2005.

Carl N Morris. Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.

Ranjini Natarajan and Robert E Kass. Reference bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95(449):227–237, 2000.

Charlotte G Neumann, Nimrod O Bwibo, Suzanne P Murphy, Marian Sigman, Shannon Whaley, Lindsay H Allen, Donald Guthrie, Robert E Weiss, and Montague W Demment. Animal source foods improve dietary quality, micronutrient status, growth and cognitive function in kenyan school children: background, study design and baseline findings. *the Journal of Nutrition*, 133(11):3941S–3949S, 2003.

NGN Prasad and JNK Rao. The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*, 85(409):163–171, 1990.

Donald B Rubin. A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the sir algorithm. *Journal of the American Statistical Association*, 82 (398):543–546, 1987.

Donald B Rubin. Using the sir algorithm to simulate posterior distributions. *Bayesian statistics*, 3:395–402, 1988.

Anders Skrondal and Sophia Rabe-Hesketh. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC, 2004.

David Spiegelhalter, Andrew Thomas, Nicky Best, and Wally Gilks. Bugs 0.5* examples volume 1 (version ii). *MRC Biostatistics Unit, Cambridge, UK*, 1995.

David Spiegelhalter, Andrew Thomas, Nicky Best, and Wally Gilks. Bugs 0.5: Bayesian inference using gibbs sampling manual (version ii). *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK*, 1996.

David J Spiegelhalter. Bayesian methods for cluster randomized trials with continuous responses. *Statistics in medicine*, 20(3):435–452, 2001.

David A Van Dyk and Taeyoung Park. Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796, 2008.

A. M. Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(1):pp. 80–88, 1969. ISSN 00359246. URL `http://www.jstor.org/stable/2984328`.

Robert E Weiss. *Modeling longitudinal data*. Springer, 2005. URL `http://rem.ph.ucla.edu/mld/`.

Shannon E Whaley, Marian Sigman, Charlotte Neumann, Nimrod Bwibo, Donald Guthrie, Robert E Weiss, Susan Alber, and Suzanne P Murphy. The impact of dietary intervention on the cognitive development of kenyan school children. *The Journal of nutrition*, 133 (11):3965S–3971S, 2003.

12 Appendix

12.1 Distribution of Simple Model MLE

The MLE for the simple model is used in subsequent proofs and an analysis of it promotes insight into other useful quantities. Equation 4 gives the likelihood for μ, σ_y^2, and σ_θ^2 for the simple model. From it, the maximizer in μ is seen to be $\bar{y}$, and the maximizer in σ_y^2 $\frac{1}{N}\left(S_w^2 + (\sigma^2 + 1/n)^{-1} S_b^2\right)$. Plugging these in yields the profiled likelihood:

$$p(\sigma_\theta^2, \hat{\mu}, \hat{\sigma}_y^2; y) = (2\pi)^{-N/2} \left[\frac{1}{N}\left(S_w^2 + \frac{1}{\sigma_\theta^2 + 1/n} S_b^2\right)\right]^{-N/2} n^{-J/2} (\sigma_\theta^2 + 1/n)^{-J/2} e^{-N/2}. \quad (16)$$

Taking a derivative of the logarithm of this with respect to σ_θ^2 and setting it equal to 0 leads to

$$\hat{\sigma}_{\theta,\mathrm{MLE}}^2 = \left[\frac{N-J}{N}\frac{S_b^2}{S_w^2/n} - \frac{1}{n}\right]^+. \quad (17)$$

At this point, it is more convenient to consider the version that is not constrained to be positive. After deriving the distribution of this it can be seen that, so long as σ_θ is greater than 0, the probability of truncation goes to 0 as n or J go to infinity.

Sums of Squares

A sufficient statistic for σ_θ is the ratio of sums of squares $R = \frac{S_b^2}{S_w^2/n}$. To determine the individual distributions of the sums involved, first condition on θ and consider the vector composed of $y_{ij} - \theta_j$, $u = y - [\theta \otimes 1_n]$. By the model, every element is independently and identically normal with mean μ and variance σ_y^2. If we take this vector and project it onto the matrix $X = \mathrm{I}_J \otimes 1_n$ (block repetitions of a 1s vector), we obtain:

$$
\begin{aligned}
\hat{u} &= X(X^\top X)^{-1}X^\top(y - \theta \otimes 1_n) \\
&= X\left((\mathrm{I}_J \otimes 1_n^\top)(\mathrm{I}_J \otimes 1_n)\right)^{-1} X(y - \theta \otimes 1_n), \\
&= \frac{1}{n}(\mathrm{I}_J \otimes 1_n)(\mathrm{I}_J \otimes 1_n^\top)(y - \theta \otimes 1_n), \\
&= \frac{1}{n}(I_j \otimes 1_n 1_n^\top)y - \theta \otimes 1_n.
\end{aligned}
$$

This produces a vector of residuals

$$
\begin{aligned}
u - \hat{u} &= y - \theta \otimes 1_n - \left(\frac{1}{n}(I_j \otimes 1_n 1_n^\top)y - \theta \otimes 1_n\right), \\
&= y - (\bar{y}_1, \ldots, \bar{y}_J) \otimes 1_n.
\end{aligned}
$$

This vector simply contains the elements $y_{ij} - \bar{y}_j$, so that summing the squares yields S_w^2. By the standard theory for linear models, this quantity is σ_y^2 times a random variable with a chi-squared distribution and $N - J$ degrees of freedom. All of these calculations took place given θ, but as the residuals don't involve θ they have the same distribution unconditionally.

Furthermore, the residuals are independent of their projections, but the projections are just repetitions of group averages. Given θ, the vector of group averages is independent of the residuals, but the residuals are independent of θ so unconditional independence holds as well.

Finally, to determine the distribution of the between group sum of squares, marginally each $\bar{y}_j$ is independent of the others and is distributted normally with mean μ and variance $\sigma_y^2(\sigma_\theta^2 + 1/n)$.

In summary,

$$
\begin{aligned}
S_w^2 &= \sum_{j=1}^{J}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2, \\
&\sim \sigma_y^2\chi_{N-J}^2, \\
S_b^2 &= \sum_{j=1}^{J}(\bar{y}_j-\bar{y})^2, \\
&\sim \sigma_y^2(\sigma_\theta^2+1/n)\chi_{J-1}^2, \\
S_w^2 &\perp S_b^2.
\end{aligned}
$$

Ratio of Sums of Squares

As the ratio of scaled sums of squares, themselves independent χ^2 random variables, R has a scaled F distribution. Specifically,

$$
\begin{aligned}
R &= \frac{S_b^2}{S_w^2/n}, \\
&= \frac{S_b^2}{J}\Big/\frac{S_w^2}{N}, \\
&= \frac{\sigma_y^2(\sigma_\theta^2+1/n)}{\sigma_y^2(\sigma_\theta^2+1/n)}\frac{J-1}{J}\frac{S_b^2}{J-1}\Big/\frac{\sigma_y^2}{\sigma_y^2}\frac{N-J}{N}\frac{S_w^2}{N-J}, \\
&\stackrel{\mathrm{d}}{=} \left(\sigma_\theta^2+\frac{1}{n}\right)\frac{N-n}{N-J}F_{J-1,N-J},
\end{aligned}
$$

where $F_{J-1,N-J}$ is a random variable with the corresponding F distribution.

MLE

Putting this together with the functional form of the MLE, we find

$$
\hat{\sigma}_{\theta,\mathrm{MLE}}^2 \stackrel{\mathrm{d}}{=} \left[\left(\sigma_\theta^2+\frac{1}{n}\right)\frac{J-1}{J}F_{J-1,N-J}-\frac{1}{n}\right]^+.
$$

Of principal interest are the asymptotics of $\hat{\sigma}_{\theta,\text{MLE}}$. Considering the right-hand side of equation 17 without the truncation and appliying the law of large numbers as J goes to infinity, we see that it has the almost-sure limit of σ_θ^2. Instead applying the central limit theorem and multiplying by $\sqrt{J}$, we obtain an asymptotically normal distribution with a stable variance. Taking this and using Slutsky's theorem to map back to the original estimator, we have that $\hat{\sigma}^2_{\theta,\text{MLE}} \in O_p(J^{-1/2})$. Finally, the transformation that results from taking the square root adds a term that is $O_p(J^{-1})$, so that $\hat{\sigma}_{\theta,\text{MLE}} \in O_p(J^{-1/2})$ as well.

12.2 Proof of Theorem 1

Overview

We obtain an asymptotic expansion for the posterior mode under the frequentist assumptions that there exists a true parameter value and that it is greater than 0. To do this, we take a sequence of Taylor series approximations to an estimating equation for σ_θ, $\Psi(\sigma_\theta)$. If we expand the equation at its root in a neighborhood of the true value of the parameter, then we can complete the square to isolate $\hat{\sigma}_\theta - \sigma_\theta$ in:

$$\Psi(\hat{\sigma}_\theta) = \Psi(\sigma_\theta) + \Psi'(\sigma_\theta)(\hat{\sigma}_\theta - \sigma_\theta) + \frac{1}{2!}\Psi''(\sigma_\theta)(\hat{\sigma}_\theta - \sigma_\theta)^2 + \frac{1}{3!}\Psi'''(\sigma_\theta^*)(\hat{\sigma}_\theta - \sigma_\theta)^3,$$

for σ_θ^* in between $\hat{\sigma}_\theta$ and σ_θ.

To achieve this, we need to be able to bound the remainder term, which involves determining the asymptotic order of $\hat{\sigma}_\theta$. Further determining the order of $\Psi''(\sigma_\theta)$ enables a more precise, higher order expansion. Finally, consistency of $\hat{\sigma}_\theta$ is necessary to take the appropriate root after completing the square.

As in the end we are interested not so much in the expansion as in the expected value of functions of it, it is helpful to have it in a form that lends itself to further analysis. This is done by focusing on a particular statistic with a straightforward distribution, itself a function

of R detailed above.

Estimating Equation

Penalizing this by an prior $p(\sigma_\theta)$ produces the profiled posterior $p(\sigma_\theta \mid y; \hat{\mu}, \hat{\sigma}_y^2)$. For simplicity, we assume that the prior is:

1. not data-dependent, including y, n, and J
2. four times log continuously differentiable in a neighborhood of σ_θ
3. with support in a neighborhood of σ_θ

The first two of these can easily be weakened, but without greatly benefitting our purpose.

Taking a logarithm of the posterior followed by a derivative yields the initial estimating equation:

$$\begin{aligned}
\log p(\sigma_\theta \mid y; \hat{\mu}, \hat{\sigma}_y^2) &\propto -\frac{N}{2}\left(S_w^2 + \frac{1}{\sigma_\theta^2 + 1/n} S_b^2\right) - \frac{J}{2}(\sigma_\theta^2 + 1/n) + \log p(\sigma_\theta), \\
\frac{\mathrm{d}}{\mathrm{d}\sigma_\theta} \log p(\sigma_\theta \mid y; \hat{\mu}, \hat{\sigma}_y^2) &= N \frac{\sigma_\theta (\sigma_\theta^2 + 1/n)^{-2} S_b^2}{S_w^2 + (\sigma_\theta^2 + 1/n)^{-1} S_b^2} - J \frac{\sigma_\theta}{\sigma_\theta^2 + 1/n} + U(\sigma_\theta),
\end{aligned}$$

where $U(\sigma_\theta) = \frac{\mathrm{d}}{\mathrm{d}\sigma_\theta} \log p(\sigma_\theta)$. As multiplying this by a non-zero, finite expression does not change its roots, we do so with $-\frac{1}{J}\frac{1}{\sigma_\theta}\frac{1}{S_w^2}(\sigma_\theta^2 + 1/n)^2 \left(S_w^2 + (\sigma_\theta^2 + 1/n)^{-1} S_b^2\right)$ and proceed ignoring the possible trivial root at 0. After some rearrangement, we obtain the estimating equation:

$$\Psi(\sigma_\theta) = \left(\sigma_\theta^2 + 1/n\right) - (n-1)\frac{S_b^2}{S_w^2} - \frac{1}{J}\frac{1}{\sigma_\theta} U(\sigma_\theta) \left(\sigma_\theta^2 + 1/n\right) \left(\sigma_\theta^2 + 1/n + \frac{S_b^2}{S_w^2}\right).$$

To simplify this, we note the reappearance of the ratio $R = \frac{S_b^2}{S_w^2/n}$ and define the statistic $T = \sigma_\theta^2 + \frac{1}{n} - \frac{n-1}{n} R$. Recalling the discussion above, we have that $T \in O_p(J^{-1/2})$. Using

this, we can decompose Ψ into its asymptotically distinct components:

$$\Psi(\sigma_\theta) = T - \frac{1}{J}\frac{1}{\sigma_\theta}\frac{n}{n-1}(\sigma_\theta^2 + 1/n)^2 U(\sigma_\theta) + \frac{T}{J}\frac{1}{\sigma_\theta}\frac{1}{n-1}(\sigma_\theta^2 + 1/n)U(\sigma_\theta). \tag{18}$$

One feature of this equation that we wish to highlight is that it has a constant part (which happens to be zero), a term times T that is thus $O_p(J^{-1/2})$, one times $1/J$, and finally a T/J term in $O_p(J^{-3/2})$. In addition, its derivatives with respect to σ_θ have the same arrangement. For simplicity denote these coefficients as a_k, b_k, c_k, and d_k respectively, where k refers to the order of derivative. More specifically, these are functions of σ_θ although that dependence is ignored for the moment. Filling in only the most simple of coefficients, we have:

$$\begin{aligned}
\Psi(\sigma_\theta) &= 0 \quad + T + c_0\frac{1}{J} + d_0\frac{T}{J}, \\
\Psi'(\sigma_\theta) &= 2\sigma_\theta + 0 + c_1\frac{1}{J} + d_1\frac{T}{J}, \\
\Psi''(\sigma_\theta) &= 2 \quad + 0 + c_2\frac{1}{J} + d_2\frac{T}{J}.
\end{aligned}$$

The remaining coefficients can be derived by calculation, although we note that $c_0 = -\frac{1}{\sigma_\theta}\frac{n}{n-1}(\sigma_\theta^2 + 1/n)^2 U(\sigma_\theta)$.

The lack of σ_θ in the leading terms of the second derivative of the estimating equation implies that all future derivatives are $O_p(J^{-1})$. By assumption, Ψ and its first three derivatives are continuous functions of σ_θ, so that by the continuous mapping theorem $J\Psi'''(\sigma_\theta^*)$ converges in probability to a constant and $\Psi'''(\sigma_\theta^*) \in O_p(J^{-1})$. As such, a Taylor series expansion that truncates at the third term is valid to a higher order. This requires the assumption of the existence fourth derivative of the log of the prior. Without it, the expansion remains valid but only to a lower asymptotic order.

Finally, under our restrictions on priors the posterior mode will have the same asymptotics as the MLE. To apply a result such as Walker (1969), we note that even though y_{ij} are not independent, they are equal in distribution to random variables that can be expressed as

independent and identically distributed observations. Consequently, $\hat{\sigma}_\theta \in O_p(J^{-1/2})$ and $\hat{\sigma}_\theta$ is consistent for σ_θ.

Taylor Series Approximation

Expanding $\Psi(\hat{\sigma}_\theta)$ around σ_θ produces:

$$\begin{aligned}\Psi(\hat{\sigma}_\theta) &= \Psi(\sigma_\theta) + \Psi'(\sigma_\theta)(\hat{\sigma}_\theta - \sigma_\theta) + \frac{1}{2!}\Psi''(\sigma_\theta)(\hat{\sigma}_\theta - \sigma_\theta)^2 + \frac{1}{3!}\Psi'''(\sigma_\theta^*)(\hat{\sigma}_\theta - \sigma_\theta)^3, \\ 0 &= \Psi(\sigma_\theta) + \Psi'(\sigma_\theta)(\hat{\sigma}_\theta - \sigma_\theta) + \frac{1}{2}\Psi''(\sigma_\theta)(\hat{\sigma}_\theta - \sigma_\theta)^2 + O_p(J^{-5/2}),\end{aligned}$$

where the truncation follows from $\Psi'''(\sigma_\theta^*) \in O_p(J^{-1})$, $\hat{\sigma}_\theta \in O_p(J^{-1/2})$, and $|\hat{\sigma}_\theta - \sigma_\theta|^3 \in O_p(J^{-3/2})$.

For now we omit the notational dependence on σ_θ and complete the square to find

$$\hat{\sigma}_\theta - \sigma_\theta = \frac{-\Psi' \pm \sqrt{\Psi'^2 - 2\Psi\Psi'' - 2\Psi'' O_p(J^{-5/2})}}{\Psi''}.$$

From here we successively approximate the square root, and then divide by Ψ''.

Denoting the term under the square root as Δ and noting that $\Psi''(\sigma_\theta) \in O_p(1)$,

$$\begin{aligned}\Delta &= \Psi'^2 - 2\Psi\Psi'' + O_p(J^{-5/2}), \\ &= \left(a_1 + c_1\frac{1}{J} + d_1\frac{T}{J}\right)^2 - 2\left(b_0 T + c_0\frac{1}{J} + d_0\frac{T}{J}\right)\left(a_2 + c_2\frac{1}{J} + d_2\frac{T}{J}\right) + O_p(J^{-5/2}), \\ &= a_1^2 + 2a_1c_1\frac{1}{J} + 2a_1d_1\frac{T}{J} + c_1^2\frac{1}{J^2} - \\ &\quad 2a_2b_0T - 2a_2c_0\frac{1}{J} - 2a_2d_0\frac{T}{J} - 2b_0c_2\frac{T}{J} - 2c_0c_2\frac{1}{J^2} - 2b_0d_2\frac{T^2}{J} + O_p(J^{-5/2}), \\ &= a_1^2 - 2a_2b_0T + 2\left(a_1c_1 - a_2c_0\right)\frac{1}{J} + 2\left(a_1d_1 - a_2d_0 - b_0c_2\right)\frac{T}{J} + 2\left(0.5c_1^2 - c_0c_2\right)\frac{1}{J^2} - \\ &\quad 2b_0d_2\frac{T^2}{J} + O_p(J^{-5/2}).\end{aligned}$$

For this, we have used that $a_0 = b_1 = b_2 = 0$.

Using the expansion about 0

$$\begin{aligned} f(x) &= \sqrt{a^2 + x}, \\ &= a + \frac{1}{2}\frac{x}{a} - \frac{1}{8}\frac{x^2}{a^3} + \frac{1}{16}\frac{x^3}{a^5} - \frac{5}{128}\frac{x^4}{a^7} + o(x^5) \end{aligned}$$

we can produce an expansion for $\sqrt{\Delta}$. However, additionally noting that $\hat{\sigma}$ is consistent, we determine that we must have $\hat{\sigma}_\theta - \sigma_\theta = \Psi''^{-1}(-\Psi' + \sqrt{\Delta})$. Writing this out produces:

$$\begin{aligned} -\Psi' + \sqrt{\Delta} = &-\frac{a_2 b_0}{a_1}T - \frac{a_2 c_0}{a_1}\frac{1}{J} - \left(\frac{a_2 d_0}{a_1} + \frac{b_0 c_2}{a_1}\right)\frac{T}{J} + \left(\frac{1}{2}\frac{c_1^2}{a_1} - \frac{c_0 c_2}{a_1}\right)\frac{1}{J^2} - \frac{1}{2}\frac{a_2^2 b_0^2}{a_1^3}T^2 + \\ &\left(\frac{a_2 b_0 c_0}{a_1^2} - \frac{a_2^2 b_0 c_0}{a_1^3}\right)\frac{T}{J} - \left(\frac{1}{2}\frac{c_1^2}{a^1} - \frac{a_2 c_0 c_1}{a_1^2} + \frac{1}{2}\frac{a_2^2 c_0^2}{a_1^3}\right)\frac{1}{J^2} + \\ &\left(\frac{a_2 b_0 d_1}{a_1^2} - \frac{a_2^2 b_0 d_0}{a_1^3} - \frac{a_2 b_0^2 c_2}{a_1^3} - \frac{b_0 d_2}{a_1}\right)\frac{T^2}{J} - \frac{1}{2}\frac{a_2^3 b_0^3}{a_1^5}T^3 + \\ &\frac{3}{2}\left(\frac{a_2^2 b_0^2 c_1}{a_1^4} - \frac{a_2^3 b_0^2 c_0}{a_1^5}\right)\frac{T^2}{J} - \frac{5}{8}\frac{a_2^4 b_0^4}{a_1^7}T^4 + O_p(J^{-5/2}), \\ = &-\frac{a_2 b_0}{a_1}T - \frac{1}{2}\frac{a_2^2 b_0^2}{a_1^3}T^2 - \frac{a_2 c_0}{a_1}\frac{1}{J} - \frac{1}{2}\frac{a_2^3 b_0^3}{a_1^5}T^3 + \left(\frac{a_2 b_0 c_1}{a_1^2} - \frac{a_2 d_0}{a_1} - \frac{b_0 c_2}{a_1} - \frac{a_2^2 b_0 c_0}{a_1^3}\right)\frac{T}{J} - \\ &\frac{5}{8}\frac{a_2^4 b_0^4}{a_1^7}T^4 + \left(\frac{a_2 c_0 c_1}{a_1^2} - \frac{1}{2}\frac{a_2^2 c_0^2}{a_1^3} - \frac{c_0 c_2}{a_1}\right)\frac{1}{J^2} + \\ &\left(\frac{a_2 b_0 d_1}{a_1^2} - \frac{a_2^2 b_0 d_0}{a_1^3} - \frac{a_2 b_0^2 c_2}{a_1^3} + \frac{3}{2}\frac{a_2^2 b_0^2 c_1}{a_1^4} - \frac{3}{2}\frac{a_2^3 b_0^2 c_0}{a_1^5} - \frac{b_0 d_2}{a_1}\right)\frac{T^2}{J} + O_p(J^{-5/2}). \end{aligned}$$

Expanding the denominator around 0,

$$
\begin{aligned}
f(x) &= \frac{1}{a+x}, \\
&= \frac{1}{a} - \frac{x}{a^2} + o(x^2), \\
\frac{1}{\Psi''} &= \frac{1}{a_2 + c_2\frac{1}{J} + d_2\frac{T}{J}}, \\
&= \frac{1}{a_2} - \frac{c_2}{a_2^2}\frac{1}{J} - \frac{d_2}{a_2^2}\frac{T}{J} + O_p(J^{-2}),
\end{aligned}
$$

we arrive at:

$$
\begin{aligned}
\hat{\sigma}_\theta - \sigma_\theta &= \frac{-\Psi' + \sqrt{\Delta}}{\Psi''}, \\
&= -\frac{b_0}{a_1}T - \frac{1}{2}\frac{a_2 b_0^2}{a_1^3}T^2 - \frac{c_0}{a_1}\frac{1}{J} - \frac{1}{2}\frac{a_2^2 b_0^3}{a_1^5}T^3 + \left(\frac{b_0 c_1}{a_1^2} - \frac{d_0}{a_1} - \frac{a_2 b_0 c_0}{a_1^3}\right)\frac{T}{J} - \frac{5}{8}\frac{a_2^3 b_0^4}{a_1^7}T^4 + \\
&\quad \left(\frac{c_0 c_1}{a_1^2} - \frac{1}{2}\frac{a_2 c_0^2}{a_1^3}\right)\frac{1}{J^2} + \left(\frac{b_0 d_1}{a_1^2} - \frac{a_2 b_0 d_0}{a_1^3} - \frac{1}{2}\frac{b_0^2 c_2}{a_1^3} + \frac{3}{2}\frac{a_2 b_0^2 c_1}{a_1^4} - \frac{3}{2}\frac{a_2^2 b_0^2 c_0}{a_1^5}\right)\frac{T^2}{J} + O_p(J^{-5/2}).
\end{aligned}
$$

If we plug in our definitions of a and b, we have:

$$
\begin{aligned}
\hat{\sigma}_\theta - \sigma_\theta &= -\frac{1}{2\sigma_\theta}T - \frac{1}{8\sigma_\theta^3}T^2 - \frac{c_0}{2\sigma_\theta}\frac{1}{J} - \frac{1}{16\sigma_\theta^5}T^3 + \left(\frac{c_1}{4\sigma_\theta^2} - \frac{d_0}{2\sigma_\theta} - \frac{c_0}{4\sigma_\theta^3}\right)\frac{T}{J} - \frac{5}{128\sigma_\theta^7}T^4 + \\
&\quad \left(\frac{c_0 c_1}{4\sigma_\theta^2} - \frac{c_0^2}{8\sigma_\theta^3}\right)\frac{1}{J^2} + \left(\frac{d_1}{4\sigma_\theta^2} - \frac{d_0}{4\sigma_\theta^3} - \frac{c_2}{16\sigma_\theta^3} + \frac{3c_1}{16\sigma_\theta^4} - \frac{3c_0}{16\sigma_\theta^5}\right)\frac{T^2}{J} + O_p(J^{-5/2}).
\end{aligned}
$$

Dropping the terms that are $O_p(J^{-3/2})$ and plugging in the definition of c_0 yields the expansion in 3.

Expected Value

Uniform integrability of $\hat{\sigma}_\theta$ is obtained by bounding the distance between it and $\hat{\sigma}_{\theta,\text{MLE}}$. If we write $\Psi(\sigma_\theta) = m(\sigma_\theta) + u(\sigma_\theta)$ such that $m(\hat{\sigma}_{\theta,\text{MLE}}) = 0$ and take the expansions:

$$
\begin{aligned}
m(\sigma) + u(\sigma) &= m(\hat{\sigma}_\theta) + u(\hat{\sigma}_\theta) + r_1(\sigma)(\sigma - \hat{\sigma}_\theta), \\
m(\sigma) &= m(\hat{\sigma}_{\theta,\mathrm{MLE}}) + r_2(\sigma)(\sigma - \hat{\sigma}_{\theta,\mathrm{MLE}}), \\
\hat{\sigma}_\theta - \hat{\sigma}_{\theta,\mathrm{MLE}} &= \frac{m(\sigma)}{r_2(\sigma)} - \frac{m(\sigma) + u(\sigma)}{r_1(\sigma)}
\end{aligned}
$$

we obtain an expression that is a polynomial in T with coefficients given by the unknown remainder functions. As T is just a linear transformation of a random variable with an F distribution, which in turn has moments of all orders, so does the difference between the estimators. As $\hat{\sigma}_{\theta,\mathrm{MLE}}$ similarly has moments of all orders, $\hat{\sigma}_\theta$ must as well and is consequently uniformly integrable.

Armed with the ability to pass to the limit under the integral, it is possible to obtain meaningful expressions for the expected value of the asymptotic expansion solely in terms of the expected value of the powers of T. They are:

$$
\begin{aligned}
\mathsf{E}[T] &= \left(\sigma_\theta^2 + \frac{1}{n}\right)\frac{n-3}{N-J-2}, \\
&= \left(\sigma_\theta^2 + \frac{1}{n}\right)\left[\frac{n-3}{n-1}\frac{1}{J} + 2\frac{n-3}{(n-1)^2}\frac{1}{J^2}\right] + O(J^{-3}), \\
\mathsf{E}[T^2] &= \left(\sigma_\theta^2 + \frac{1}{n}\right)^2\left[\frac{2Jn(n-1)}{(N-J-2)(N-J-4)} - \frac{n^2+6n-15}{(N-J-2)(N-J-4)}\right], \\
&= \left(\sigma_\theta^2 + \frac{1}{n}\right)^2\left[2\frac{n}{n-1}\frac{1}{J} - \frac{n^2-6n-15}{(n-1)^2}\frac{1}{J^2}\right] + O(J^{-3}), \\
E[T^3] &= \left(\sigma_\theta^2 + \frac{1}{n}\right)^3\left[\frac{-2J(n-1)n(n+13) + 3n^3 + 9n^2 + 45n - 105}{(N-J-2)(N-J-4)(N-J-6)}\right], \\
&= -2\left(\sigma_\theta^2 + \frac{1}{n}\right)^3\frac{n(n+13)}{(n-1)^2}\frac{1}{J^2} + O(J^{-3}), \\
E[T^4] &= \left(\sigma_\theta^2 + \frac{1}{n}\right)^4\frac{(n-1)^4(12J^2+4J-15) + (n-1)^3(24J^2+52J-46)}{(N-J-2)(N-J-4)(N-J-6)(N-J-8)}\times \\
&\quad \frac{(n-1)^2(12J^2+416J-288) + (n-1)(368J-768) + 384}{(N-J-2)(N-J-4)(N-J-6)(N-J-8)}, \\
&= 12\left(\sigma_\theta^2 + \frac{1}{n}\right)^4\frac{n^2}{(n-1)^2}\frac{1}{J^2} + O(J^{-3}).
\end{aligned}
$$

We can plug this into the expansion for the estimator to obtain:

$$
\begin{aligned}
\mathsf{E}[\hat{\sigma}_\theta - \sigma_\theta] = &-\frac{1}{J}\frac{1}{2\sigma_\theta}\left[\frac{1}{n-1}\left(\sigma_\theta^2 + \frac{1}{n}\right)\left(3\frac{n-2}{2} + \frac{1}{2}\frac{1}{\sigma_\theta^2}\right) + c_0\right] - \\
&\frac{\sigma_\theta}{32(n-1)^2}\left(\sigma_\theta^2 + \frac{1}{n}\right)\left[15n^2 - n\left(40 - \frac{41}{\sigma_\theta^2}\right) - 156 - \frac{128}{\sigma_\theta^2} + \frac{41}{\sigma_\theta^4} - \frac{1}{n\sigma_\theta^2}\left(60 + \frac{52}{\sigma_\theta^2} - \frac{15}{\sigma_\theta^4}\right)\right]\frac{1}{J^2} + \\
&\frac{2c_0c_1\sigma_\theta - c_0^2}{8\sigma_\theta^3}\frac{1}{J^2} + \frac{n-3}{n-1}\left(\sigma_\theta + \frac{1}{n\sigma_\theta}\right)\left(\frac{c_1\sigma_\theta - 2d_0\sigma_\theta^2 - c_0}{4\sigma_\theta^2}\right)\frac{1}{J^2} + \\
&\frac{1}{8}\frac{n}{n-1}\left(\sigma_\theta + \frac{1}{n\sigma_\theta}\right)^2\left[4d_1\sigma_\theta^3 - 4d_0\sigma_\theta^2 - c_2\sigma_\theta^2 + 3c_1\sigma_\theta - 3c_0\right]\frac{1}{J^2} + O(J^{-5/2}).
\end{aligned}
$$

Truncating at $O(J^{-1})$ produces equation 8, while the solving for the prior produces corollary 1.

12.3 Joint Mode Calculation

As one of the principal advantages to using hierarchical models is the ability to use sparse design matrices for the modeled coefficients, the matrix inversion step used to calculate the mode in θ' of the joint distribution y, θ' must be completed carefully. In section 4.2 we have the following:

$$p(y, \theta'; \Sigma_\theta, \beta, \sigma_y^2) = (2\pi\sigma_y^2)^{-(N+Q)/2} \exp\left\{ -\frac{1}{2\sigma_y^2} \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \theta' \\ \beta \end{bmatrix} \right\|^2 \right\},$$

and we derive the decompositions

$$\begin{bmatrix} L_Z & 0 \\ L_{ZX} & L_X \end{bmatrix} \begin{bmatrix} L_Z^\top & L_{ZX}^\top \\ 0 & L_X^\top \end{bmatrix} = \begin{bmatrix} L_\theta^\top Z^\top & \mathrm{I}_Q \\ X^\top & 0 \end{bmatrix} \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix}.$$

Z can be stored as a sparse matrix given that each row has $Q = \sum_{k=1}^K Q_k J_k$ elements but only $\sum_{k=1}^K Q_k$ are non-zero. Similar L_θ is sparse as it is a left factor of a block diagonal matrix. The pattern of non-zeroes in each makes their product have a structure similar to Z itself. Given this, it is possible to compute L_Z using sparse matrix decomposition techniques. L_{ZX} and L_X are generally dense.

For the ordinary linear regression problem of minimizing $\|u - A\eta\|^2$, we have that $\hat{\eta} = (A^\top A)^{-1} A^\top u$. Similarly,

$$
\begin{aligned}
\begin{bmatrix}\tilde{\theta}\\ \tilde{\beta}\end{bmatrix} &= \begin{bmatrix} L_Z^{-\top} & -L_Z^{-\top}L_{ZX}^{\top}L_X^{-\top} \\ 0 & L_X^{-\top}\end{bmatrix}\begin{bmatrix} L_Z^{-1} & 0 \\ -L_X^{-1}L_{ZX}L_Z^{-1} & L_X^{-1}\end{bmatrix}\begin{bmatrix} L_\theta^{\top}Z^{\top} & \mathrm{I}_Q \\ X^{\top} & 0\end{bmatrix}\begin{bmatrix} y \\ 0\end{bmatrix}, \\
&= \begin{bmatrix} L_Z^{-\top} & -L_Z^{-\top}L_{ZX}^{\top}L_X^{-\top} \\ 0 & L_X^{-\top}\end{bmatrix}\begin{bmatrix} L_Z^{-1}L_\theta^{\top}Z^{\top}y \\ L_X^{-1}X^{\top}y - L_X^{-1}L_{ZX}L_Z^{-1}L_\theta^{\top}Z^{\top}y\end{bmatrix}, \\
&= \begin{bmatrix} L_Z^{-\top} & -L_Z^{-\top}L_{ZX}^{\top}L_X^{-\top} \\ 0 & L_X^{-\top}\end{bmatrix}\begin{bmatrix} \underset{\sim}{\theta} \\ L_X^{-1}\left(X^{\top}y - L_{ZX}\underset{\sim}{\theta}\right)\end{bmatrix}, \qquad \underset{\sim}{\theta} = L_Z^{-1}L_\theta^{\top}Z^{\top}y, \\
&= \begin{bmatrix} L_Z^{-\top} & -L_Z^{-\top}L_{ZX}^{\top}L_X^{-\top} \\ 0 & L_X^{-\top}\end{bmatrix}\begin{bmatrix} \underset{\sim}{\theta} \\ \underset{\sim}{\beta}\end{bmatrix}, \qquad \underset{\sim}{\beta} = L_X^{-1}\left(X^{\top}y - L_{ZX}\underset{\sim}{\theta}\right), \\
&= \begin{bmatrix} L_Z^{-\top}\left(\underset{\sim}{\theta} - L_{ZX}^{\top}L_X^{-\top}\underset{\sim}{\beta}\right) \\ L_X^{-\top}\underset{\sim}{\beta}\end{bmatrix}.
\end{aligned}
$$

To find the joint mode from the block-wise decomposition, one then computes in order $\underset{\sim}{\theta}, \underset{\sim}{\beta}, \tilde{\beta}$, and finally $\tilde{\theta}$.

$\underset{\sim}{\theta}$ and $\underset{\sim}{\beta}$ are further useful for calculating the penalized sum of squared residuals. Again analogizing to an ordinary linear regression, with $\underset{\sim}{\eta} = L_A^{-1}A^{\top}u$, $L_AL_A^{\top} = A^{\top}A$:

$$
\begin{aligned}
u^{\top}u - \underset{\sim}{\eta}^{\top}\underset{\sim}{\eta} &= u^{\top}u - u^{\top}AL_A^{-\top}L_A^{-1}A^{\top}u, \\
&= u^{\top}u - u^{\top}A\hat{\eta}, \\
&= \|u - A\hat{\eta}\|^2.
\end{aligned}
$$

Consequently,

$$
y^{\top}y - \underset{\sim}{\theta}^{\top}\underset{\sim}{\theta} - \underset{\sim}{\beta}^{\top}\underset{\sim}{\beta} = \left\| \begin{bmatrix} y \\ 0\end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{\theta}\\ \tilde{\beta}\end{bmatrix}\right\|^2.
$$

If one is profiling σ_y^2 linearly, it is thus possible to reuse this calculation.

Finally, we note that these steps are applicable for simple linear regressions that include potentially sparse and dense design matrices, as occurs in fixed effect models.

12.4 Additional Optimization Schemes

Completely arbitrary priors can be placed on any model parameter and optimization done at the possible expense of the two profiling steps, β, and σ_y^2.

Unmodeled Coefficient Prior

In section 4.4 a Gaussian prior was placed on β that utilized σ_y^2 and hence was "on the common scale". It is possible to place a Gaussian prior with a real-world, absolute covariance if one is willing to add σ_y^2 to the optimization parameter set.

Assume that $\beta \sim \mathrm{N}(0, \Sigma_\beta)$ with Σ_β known and $L_\beta L_\beta^\top = \Sigma_\beta$. The joint distribution in spherical modeled coefficients given by:

$$p(y, \theta', \beta; \Sigma_\theta, \sigma_y^2) = (2\pi\sigma_y^2)^{-(N+Q+P)/2} |\Sigma_\beta|^{-1/2} \exp\left\{ -\frac{1}{2\sigma_y^2} \left\| \begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ 0 & \sigma_y L_\beta^{-1} \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \theta' \\ \beta \end{bmatrix} \right\|^2 \right\}.$$

Defining L_Z and L_{ZX} as before and setting $L_X L_X^\top = X^\top X + \sigma_y^2 \Sigma_\beta^{-1} - L_{ZX} L_{ZX}^\top$, it is possible to integrate out θ' and obtain the posterior:

$$p(\beta \mid y; \Sigma_\beta, \sigma_y^2) \propto (\sigma_y^2)^{-(N+P)/2} |L_Z|^{-1} \times$$
$$\exp\left\{ -\frac{1}{2\sigma_y^2} \left[(\beta - \tilde{\beta}(\sigma_y^2))^\top L_X(\sigma_y^2) L_X(\sigma_y^2)^\top (\beta - \tilde{\beta}(\sigma_y^2)) + \left\| \begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ 0 & \sigma_y L_\beta^{-1} \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta}(\sigma_y^2) \\ \tilde{\beta}(\sigma_y^2) \end{bmatrix} \right\|^2 \right] \right\}.$$

As before, $\tilde{\beta}$ and $\tilde{\theta}$ are the joint modes although now they depend on the value of σ_y^2. β can be profiled directly, which leaves for optimization:

$$p(\hat{\beta} \mid y; \Sigma_\beta, \sigma_y^2) \propto (\sigma_y^2)^{-(N+P)/2} |L_Z|^{-1} \times \exp\left\{ -\frac{1}{2\sigma_y^2} \left\| \begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ 0 & \sigma_y L_\beta^{-1} \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta}(\sigma_y^2) \\ \tilde{\beta}(\sigma_y^2) \end{bmatrix} \right\|^2 \right\}.$$

REML Estimates

Restricted maximum likelihood can be viewed as a particular form of penalty function, or prior. In the most direct sense, it is what arises from *modeling* β with a flat distribution. From the likelihood perspective, inference requires that β be integrated out. Having previously found the joint mode of θ' and β in our profiling step, it is trivial to compute this integral.

Returning to equation 10 in section 4.2, the conditional distribution of β under a flat prior is Gaussian. Integrating it out yields:

$$p(y; \Sigma_\theta, \sigma_y^2) = (2\pi\sigma_y^2)^{-(N-P)/2} |L_Z|^{-1} |L_X|^{-1} \exp\left\{ -\frac{1}{2\sigma_y^2} \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} \right\|^2 \right\}.$$

$\hat{\sigma}_y^2$ can be directly computed as the exponential term divided by $N - P$. This leaves the

profiled REML equation:

$$p(y; \Sigma_\theta, \hat{\sigma}_y^2) = \left(2\pi\hat{\sigma}_y^2(\sigma_\theta)\right)^{-(N-P)/2} |L_Z(\sigma_\theta)|^{-1} |L_X(\sigma_\theta)|^{-1} e^{-(N-P)/2}.$$

It is furthermore possible to extend these considerations to the prior $\beta \sim \mathrm{N}(0, \sigma_y^2 \Sigma_\beta)$, yielding the likelihood:

$$p(y; \Sigma_\theta, \sigma_y^2) = (2\pi\sigma_y^2)^{-N/2} |L_Z|^{-1} |L_X|^{-1} \exp\left\{ -\frac{1}{2\sigma_y^2} \left\| \begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ 0 & \mathrm{L}_\beta^{-1} \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} \right\|^2 \right\}.$$

Thus, $\hat{\sigma}_y^2$ is the exponential term divided by N and the profiled likelihood is:

$$p(y; \Sigma_\theta, \hat{\sigma}_y^2) = \left(2\pi\hat{\sigma}_y^2(\sigma_\theta)\right)^{-N/2} |L_Z(\sigma_\theta)|^{-1} |L_X(\sigma_\theta)|^{-1} e^{N/2}.$$

If we assume $\beta \sim \mathrm{N}(0, \Sigma_\beta)$, we obtain a likelihood similar to the non-REML case but with the addition of the penalty $|L_X(\sigma_\theta, \sigma_y^2)|^{-1}$ and an adjustment in degrees of freedom.

12.5 Marginal Posterior Derivation

Starting from the joint distribution of y and θ' for the full model and assuming flat priors on all parameters, we have

$$p(\sigma_y^2, \beta, \Sigma_\theta, \theta' \mid y) \propto (\sigma_y^2)^{-(N+Q)/2} \exp\left\{ -\frac{1}{2\sigma_y^2} \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \beta \end{bmatrix} \right\|^2 \right\}.$$

Let $A = \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \beta \end{bmatrix}$, and $u = (y, 0)^\top$, and $\eta = (\theta', \beta)^\top$. Then we have:

$$p(\sigma_y^2, \beta, \Sigma_\theta, \theta' \mid y) \propto (\sigma_y^2)^{-(N+Q)/2} \exp\left\{-\frac{1}{2\sigma_y^2}\left(\eta - (A^\top A)^{-1}A^\top u\right)^\top A^\top A \left(\eta - (A^\top A)^{-1}A^\top u\right)\right\} \times$$
$$\exp\left\{-\frac{1}{2\sigma_y^2}\left[u^\top u - u^\top A(A^\top A)^{-1}A^\top u\right]\right\},$$
$$= (\sigma_y^2)^{-(N+Q)/2} \frac{|A^\top A|^{1/2}}{|A^\top A|^{1/2}} \exp\left\{-\frac{1}{2\sigma_y^2}\left[(\eta - \hat{\eta})^\top A^\top A(\eta - \hat{\eta}) + u^\top H_A^\perp u\right]\right\}.$$

Simultaneously integrating out θ' and β as jointly normal leaves

$$p(\sigma_y^2, \Sigma_\theta \mid y) \propto (\sigma_y^2)^{-(N-P)/2} |A^\top A|^{-1/2} \exp\left\{-\frac{1}{2\sigma_y^2}\|u - A\hat{\eta}\|^2\right\},$$
$$= (\sigma_y^2)^{-(N-P-2)/2-1} \frac{\|u - A\hat{\eta}\|^{N-P-2}}{\|u - A\hat{\eta}\|^{N-P-2}}, \exp\left\{-\frac{1}{2\sigma_y^2}\|u - A\hat{\eta}\|^2\right\} \times |A^\top A|^{-1/2}.$$

From here, σ_y^2 has an inverse gamma distribution with a shape of $(N-P)/2$ and a scale of one half of the sum of squared residuals. Integrating it yields

$$p(\Sigma_\theta \mid y) \propto \left|\begin{matrix} L_\theta^\top Z^\top Z L_\theta + \mathrm{I}_Q & L_\theta^\top Z^\top X \\ X^\top Z L_\theta & X^\top X \end{matrix}\right|^{-1/2} \times$$
$$\left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} ZL_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} L_\theta^\top Z^\top Z L_\theta + \mathrm{I}_Q & L_\theta^\top Z^\top X \\ X^\top Z L_\theta & X^\top X \end{bmatrix}^{-1} \begin{bmatrix} L_\theta^\top Z^\top & \mathrm{I}_Q \\ X^\top & 0 \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|^{-(N-P-2)}$$

Determinant Rearrangement

Denote the determinant term as D. So long as Σ_θ is positive definite we have:

$$
\begin{aligned}
D &= \left| \begin{bmatrix} L_\theta^\top & 0 \\ 0 & \mathrm{I}_P \end{bmatrix} \begin{bmatrix} Z^\top Z + \Sigma_\theta^{-1} & Z^\top X \\ X^\top Z & X^\top X \end{bmatrix} \begin{bmatrix} L_\theta & 0 \\ 0 & \mathrm{I}_P \end{bmatrix} \right| \\
&= \left| \begin{bmatrix} L_\theta L_\theta^\top & 0 \\ 0 & \mathrm{I}_P \end{bmatrix} \right| \left| \begin{bmatrix} Z^\top Z + \epsilon \mathrm{I}_Q & Z^\top X \\ X^\top Z & X^\top X \end{bmatrix} + \begin{bmatrix} \Sigma_\theta^{-1} - \epsilon \mathrm{I}_Q & 0 \\ 0 & 0 \end{bmatrix} \right|,
\end{aligned}
$$

where ϵ is chosen so that $\epsilon > 0$ and $\Sigma_\theta - \epsilon \mathrm{I}_Q$ is still invertible. This is necessary as $Z^\top Z$ may be of less than full rank. Denote the Cholesky factors

$$
\begin{aligned}
L_\epsilon^{-\top} L_\epsilon^{-1} &= \Sigma_\theta^{-1} - \epsilon \mathrm{I}_Q, \\
A_\epsilon^\top A_\epsilon &= \begin{bmatrix} Z^\top Z + \epsilon \mathrm{I} & Z^\top X \\ X^\top Z & X^\top X \end{bmatrix}.
\end{aligned}
$$

$$
\begin{aligned}
D &= |\Sigma_\theta||A_\epsilon^\top A_\epsilon| \left| \mathrm{I}_{Q+P} + A_\epsilon^{-\top} \begin{bmatrix} L_\epsilon^{-\top} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} L_\epsilon^{-1} & 0 \\ 0 & 0 \end{bmatrix} A_\epsilon^{-1} \right|, \\
&= |\Sigma_\theta||A_\epsilon^\top A_\epsilon| \left| \mathrm{I}_{Q+P} + \begin{bmatrix} L_\epsilon^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z^\top Z + \epsilon \mathrm{I}_Q & Z^\top X \\ X^\top Z & X^\top X \end{bmatrix}^{-1} \begin{bmatrix} L_\epsilon^{-\top} & 0 \\ 0 & 0 \end{bmatrix} \right|, \quad \text{Slyvester's thm,} \\
&= |\Sigma_\theta||A_\epsilon^\top A_\epsilon| \left| \mathrm{I}_{Q+P} + L_\epsilon^{-1} \left(Z^\top Z + \epsilon \mathrm{I}_Q - Z^\top X (X^\top X)^{-1} X^\top Z \right)^{-1} L_\epsilon^{-\top} \right|, \qquad \text{block inv,} \\
&= |\Sigma_\theta| \frac{|A_\epsilon^\top A_\epsilon|}{|Z^\top H_X^\perp Z + \epsilon \mathrm{I}_Q|} |Z^\top H_X Z + \epsilon \mathrm{I}_Q + L_\epsilon^{-\top} L_\epsilon^{-1}|, \qquad \text{Slyvester,} \\
&= |\Sigma_\theta||\Sigma_\theta^{-1} + Z^\top H_X^\perp Z| \frac{\begin{vmatrix} Z^\top Z + \epsilon \mathrm{I}_Q & Z^\top X \\ X^\top Z & X^\top X \end{vmatrix}}{|Z^\top H_X^\perp Z + \epsilon \mathrm{I}_Q|}, \\
&= |\Sigma_\theta||\Sigma_\theta^{-1} + Z^\top H_X^\perp Z||X^\top X| \frac{|Z^\top H_X^\perp Z + \epsilon \mathrm{I}_Q|}{|Z^\top H_X^\perp Z + \epsilon \mathrm{I}_Q|}, \\
&= |\Sigma_\theta||\Sigma_\theta^{-1} + Z^\top H_X^\perp Z||X^\top X|.
\end{aligned}
$$

If necessary, this can be re-written as $|\mathrm{I}_Q + L_\theta^\top Z^\top H_X^\perp Z L_\theta||X^\top X|$, an expression that, by continuity, remains valid even at the boundary of the parameter space.

This yields equation 14.

Sum of Squares Rearrangement

For the purposes of taking derivatives, we write the sum of squares in a more-tractible expression.

Let S be the Schur complement of the lower-right block for matrix that needs to be inverted, $S = L_\theta^\top Z^\top Z L_\theta + \mathrm{I}_Q - L_\theta^\top Z^\top X (X^\top X)^{-1} X^\top Z L_\theta = L_\theta^\top Z^\top H_X^\perp Z L_\theta + \mathrm{I}_Q$. Taking a block-wise inverse,

$$\begin{bmatrix} L_\theta^\top Z^\top Z L_\theta + \mathrm{I}_Q & L_\theta^\top Z^\top X \\ X^\top Z L_\theta & X^\top X \end{bmatrix}^{-1} =$$
$$\begin{bmatrix} S^{-1} & -S^{-1} L_\theta^\top Z^\top X (X^\top X)^{-1} \\ -(X^\top X)^{-1} X^\top Z L_\theta S^{-1} & (X^\top X)^{-1} + (X^\top X)^{-1} X^\top Z L_\theta S^{-1} L_\theta^\top Z^\top X (X^\top X)^{-1} \end{bmatrix}.$$

If we denote $Z L_\theta S^{-1} L_\theta^\top Z^\top$ as M, we have

$$\begin{bmatrix} Z L_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} L_\theta^\top Z^\top Z L_\theta + \mathrm{I}_Q & L_\theta^\top Z^\top X \\ X^\top Z L_\theta & X^\top X \end{bmatrix}^{-1} \begin{bmatrix} L_\theta^\top Z^\top & \mathrm{I}_Q \\ X^\top & 0 \end{bmatrix} =$$
$$\begin{bmatrix} H_X + H_X^\perp M H_X^\perp & H_X^\perp Z L_\theta S^{-1} \\ S^{-1} L_\theta^\top Z^\top H_X^\perp & S^{-1} \end{bmatrix}.$$

Finally, incorporating the vector $(y, 0)^\top$:

$$\left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} Z L_\theta & X \\ \mathrm{I}_Q & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} \right\|^2 = y^\top H_X^\perp y - y^\top H_X^\perp Z \left(Z^\top H_X^\perp Z + \Sigma_\theta^{-1} \right)^{-1} Z^\top H_X^\perp y.$$

If we were to replace both y and Z with their residuals from a projection onto the column space of X, *i.e.* $H_X^\perp y$ and $H_X^\perp Z$, this is the sum of the squares of the residuals when projected again while including a penalty term.

12.6 Marginal Posterior Derivatives

We take derivatives first with respect to Σ_θ, and later with respect to $\Sigma_1, \ldots, \Sigma_k$. Finally, we change to the free parameters σ_θ.

Matrix derivatives are taken by column-vectorizing the target function and variable being differenced. The result is a matrix with the number of rows equal to the length of the function output and columns equal to the length of the input.

Determinant Derivative

Let $f(\Sigma) = \log|\Sigma| + \log|\Sigma^{-1} + A|$ where A is a symmetric matrix of suitable dimension, then

$$\begin{aligned}
\frac{\mathrm{d}f}{\mathrm{d}\Sigma} &= \mathrm{vec}(\Sigma^{-1})^\top - \mathrm{vec}\left(\left(\Sigma^{-1} + A\right)^{-1}\right)^\top (\Sigma^{-1} \otimes \Sigma^{-1}), \\
&= \left[\mathrm{vec}(\Sigma^{-1}) - (\Sigma^{-1} \otimes \Sigma^{-1})\,\mathrm{vec}\left(\left(\Sigma^{-1} + A\right)^{-1}\right)\right]^\top, \\
&= \mathrm{vec}\left[\Sigma^{-1} - \Sigma^{-1}\left(\Sigma^{-1} + A\right)^{-1}\Sigma^{-1}\right]^\top.
\end{aligned}$$

If A is invertible, this equals $\mathrm{vec}\left((\Sigma + A^{-1})^{-1}\right)^\top$.

Taking another derivative yields:

$$\begin{aligned}
\frac{\mathrm{d}^2 f}{\mathrm{d}\Sigma \mathrm{d}\Sigma} &= -(\Sigma^{-1} \otimes \Sigma^{-1}) + \left(\Sigma^{-1}(\Sigma^{-1} + A)^{-1}\Sigma^{-1} \otimes \Sigma^{-1}\right) + \left(\Sigma^{-1} \otimes \Sigma^{-1}(\Sigma^{-1} + A)^{-1}\Sigma^{-1}\right) - \\
&\qquad \left(\Sigma^{-1}(\Sigma^{-1} + A)^{-1}\Sigma^{-1} \otimes \Sigma^{-1}(\Sigma^{-1} + A)^{-1}\Sigma^{-1}\right), \\
&= -\left[\Sigma^{-1} - \Sigma^{-1}(\Sigma^{-1} + A)^{-1}\Sigma^{-1} \otimes \Sigma^{-1} - \Sigma^{-1}(\Sigma^{-1} + A)^{-1}\Sigma^{-1}\right].
\end{aligned}$$

And again, if A is invertible this reduces to $-\left[(\Sigma + A^{-1})^{-1} \otimes (\Sigma + A^{-1})^{-1}\right]$.

In the context of a hierarchical model, we have:

$$\frac{\mathrm{d}}{\mathrm{d}\Sigma_\theta}\log\frac{|\Sigma_\theta^{-1}|^{1/2}}{|\Sigma_\theta^{-1}+Z^\top H_X^\perp Z|^{1/2}} = \mathrm{vec}\left[\Sigma_\theta^{-1}-\Sigma_\theta^{-1}\left(\Sigma_\theta^{-1}+Z^\top H_X^\perp Z\right)^{-1}\Sigma_\theta^{-1}\right]^\top,$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\Sigma_\theta\mathrm{d}\Sigma_\theta}\log\frac{|\Sigma_\theta^{-1}|^{1/2}}{|\Sigma_\theta^{-1}+Z^\top H_X^\perp Z|^{1/2}} =$$

$$-\left[\Sigma_\theta^{-1}-\Sigma_\theta^{-1}\left(\Sigma_\theta^{-1}+Z^\top H_X^\perp Z\right)^{-1}\Sigma_\theta^{-1}\otimes\Sigma_\theta^{-1}-\Sigma_\theta^{-1}\left(\Sigma_\theta^{-1}+Z^\top H_X^\perp Z\right)^{-1}\Sigma_\theta^{-1}\right].$$

Sum of Squares Derivative

For $g(\Sigma)=\log\left(a-b^\top(\Sigma^{-1}+C)^{-1}b\right)$ where a is a constant, b is vector, and C is a symmetric matrix, the derivative is given by:

$$\begin{aligned}\frac{\mathrm{d}g}{\mathrm{d}\Sigma} &= -\frac{\left(b^\top\otimes b^\top\right)\left((\Sigma^{-1}+C)^{-1}\otimes(\Sigma^{-1}+C)^{-1}\right)\left(\Sigma^{-1}\otimes\Sigma^{-1}\right)}{a-b^\top(\Sigma^{-1}+C)^{-1}b},\\ &= -\frac{\mathrm{vec}\left[\Sigma^{-1}(\Sigma^{-1}+C)^{-1}bb^\top(\Sigma^{-1}+C)^{-1}\Sigma^{-1}\right]^\top}{a-b^\top(\Sigma^{-1}+C)^{-1}b}.\end{aligned}$$

Denoting $s=\exp(g(\Sigma))$ as the sum of squares and $d=\Sigma^{-1}(\Sigma^{-1}+C)^{-1}b$, the second derivative is:

$$\begin{aligned}\frac{\mathrm{d}^2g}{\mathrm{d}\Sigma\mathrm{d}\Sigma} &= -\frac{1}{s}\frac{\mathrm{d}}{\mathrm{d}\Sigma}\left[\Sigma^{-1}(\Sigma^{-1}+C)^{-1}bb^\top(\Sigma^{-1}+C)^{-1}\Sigma^{-1}\right]-\\ &\quad\frac{1}{s^2}\mathrm{vec}\left[\Sigma^{-1}(\Sigma^{-1}+C)^{-1}bb^\top(\Sigma^{-1}+C)^{-1}\Sigma^{-1}\right]\mathrm{vec}\left[\Sigma^{-1}(\Sigma^{-1}+C)^{-1}bb^\top(\Sigma^{-1}+C)^{-1}\Sigma^{-1}\right]^\top,\\ &= -\frac{1}{s}\left[-(\Sigma^{-1}-\Sigma^{-1}(\Sigma^{-1}+C)^{-1}\Sigma^{-1}\otimes dd^\top)-(dd^\top\otimes\Sigma^{-1}-\Sigma^{-1}(\Sigma^{-1}+C)^{-1}\Sigma^{-1})\right]-\\ &\quad\frac{1}{s^2}\left(dd^\top\otimes dd^\top\right),\\ &= -\frac{1}{s^2}\left[\left(dd^\top\otimes dd^\top-s\left(\Sigma^{-1}-\Sigma^{-1}(\Sigma^{-1}+C)^{-1}\Sigma^{-1}\right)\right)+\right.\\ &\qquad\left.\left(dd^\top-s\left(\Sigma^{-1}-\Sigma^{-1}(\Sigma^{-1}+C)^{-1}\Sigma^{-1}\right)\otimes dd^\top\right)\right].\end{aligned}$$

For hierarchical models, we have:

$$
\begin{aligned}
a &= y^\top H_X^\perp y, \\
b &= Z^\top H_X^\perp y, \\
C &= Z^\top H_X^\perp Z, \\
d &= \Sigma_\theta^{-1}\left(\Sigma_\theta^{-1} + Z^\top H_X^\perp Z\right)^{-1} Z^\top H_X^\perp y.
\end{aligned}
$$

12.7 Matrix-Variate Beta Prime

Density

Assume the following model for a d dimensional covariance matrix:

$$
\begin{aligned}
\tilde{\Sigma} \mid \Psi &\sim \text{inv} - \text{Wishart}(\nu, \Psi), \\
\Psi &\sim \text{Wishart}(\nu, C).
\end{aligned}
$$

The joint density of Σ and Ψ is:

$$
\begin{aligned}
p(\tilde{\Sigma}, \Psi) &= \frac{|\Psi|^{\nu/2}|\tilde{\Sigma}|^{-(\nu+d+1)/2}}{2^{\nu d/2}\Gamma_d(\nu/2)}\text{etr}\left\{-\frac{1}{2}\Psi\tilde{\Sigma}^{-1}\right\} \times \frac{|C|^{-\mu/2}|\Psi|^{(\mu-d-1)/2}}{2^{\mu d/2}\Gamma_d(\mu/2)}\text{etr}\left\{-\frac{1}{2}\Psi C^{-1}\right\}, \\
&= \frac{|\Psi|^{(\nu+\mu-d-1)/2}|\Sigma^{-1}+C^{-1}|^{(\nu+\mu)/2}}{2^{(\nu+\mu)d/2}\Gamma_d\left((\nu+\mu)/2\right)}\text{etr}\left\{-\frac{1}{2}\Psi(\tilde{\Sigma}^{-1}+C^{-1})\right\} \times \\
&\quad \frac{\Gamma_d\left((\nu+\mu)/2\right)}{|\tilde{\Sigma}^{-1}+C^{-1}|^{(\nu+\mu)/2}}\frac{|\tilde{\Sigma}^{-1}|^{(\nu+d+1)/2}}{|C|^{\mu/2}\Gamma_d(\nu/2)\Gamma_d(\mu/2)}.
\end{aligned}
$$

The conditional distribution of $\Psi \mid \tilde{\Sigma}$ is Wishart with $\nu + \mu$ degrees of freedom and a scale of $(\tilde{\Sigma}^{-1} + C^{-1})^{-1}$ so that the leading part integrates to 1. Consequently,

$$p(\tilde{\Sigma}) = \frac{|\tilde{\Sigma}^{-1}|^{(\nu+d+1)/2}}{|\tilde{\Sigma}^{-1} + C^{-1}|^{(\nu+\mu)/2}} \frac{1}{|C|^{\mu/2}\mathrm{B}_d(\nu/2, \mu/2)}.$$

First Derivative

Denote $l(\tilde{\Sigma}) = \log p(\tilde{\Sigma})$. Then,

$$\frac{\mathrm{d}l(\tilde{\Sigma})}{\mathrm{d}\tilde{\Sigma}} = -\frac{\nu+d+1}{2}\operatorname{vec}(\Sigma^{-1})^\top + \frac{\nu+\mu}{2}\operatorname{vec}\left(\Sigma^{-1}(\Sigma^{-1}+C^{-1})^{-1}\Sigma^{-1}\right)^\top.$$

Setting this equal to the 0 vector demonstrates that it is maximized at $\widehat{\tilde{\Sigma}} = \frac{\mu-d-1}{\nu+d+1}C$. Had we wished to parameterize the distribution by its mode, say M, we would have:

$$\begin{aligned}
\tilde{\Sigma} \mid \Psi &\sim \mathrm{inv} - \mathrm{Wishart}(\nu, \Psi), \\
\Psi &\sim \mathrm{Wishart}\left(\mu, \frac{\nu+d+1}{\mu-d-1}M\right), \\
p(\tilde{\Sigma}) &= \frac{|\tilde{\Sigma}^{-1}|^{(\nu+d+1)/2}}{|\tilde{\Sigma}^{-1} + \frac{\mu-d-1}{\nu+d+1}M^{-1}|^{(\nu+\mu)/2}} \frac{\left(\frac{\mu-d-1}{\nu+d+1}\right)^{\mu d/2}}{|M|^{\mu/2}\mathrm{B}_d(\nu/2, \mu/2)}.
\end{aligned}$$

Second Derivative

$$\begin{aligned}
\frac{\mathrm{d}^2 l(\tilde{\Sigma})}{\mathrm{d}\tilde{\Sigma}\mathrm{d}\tilde{\Sigma}} = \frac{\nu+d+1}{2}(\tilde{\Sigma}\otimes\tilde{\Sigma}) - \frac{\nu+\mu}{2}\Big[&\left(\tilde{\Sigma}^{-1}\otimes\tilde{\Sigma}^{-1}(\tilde{\Sigma}^{-1}+C^{-1})^{-1}\tilde{\Sigma}^{-1}\right) + \\
&\left(\tilde{\Sigma}^{-1}(\tilde{\Sigma}^{-1}+C^{-1})^{-1}\tilde{\Sigma}^{-1}\otimes\tilde{\Sigma}^{-1}\right) - \left(\tilde{\Sigma}^{-1}(\tilde{\Sigma}^{-1}+C^{-1})^{-1}\tilde{\Sigma}^{-1}\otimes\tilde{\Sigma}^{-1}(\tilde{\Sigma}^{-1}+C^{-1})^{-1}\tilde{\Sigma}^{-1}\right)\Big].
\end{aligned}$$

When evaluated at the mode, we have

$$\begin{aligned}
l''(M) &= \frac{\nu+d+1}{2}(M^{-1}\otimes M^{-1}) - \frac{\nu+\mu}{2}\left[2\frac{\nu+d+1}{\nu+\mu}(M^{-1}\otimes M^{-1}) - \left(\frac{\nu+d+1}{\nu+\mu}\right)^2(M^{-1}\otimes M^{-1})\right], \\
&= -\frac{\nu+d+1}{2}\frac{\mu-d-1}{\nu+\mu}(M^{-1}\otimes M^{-1}).
\end{aligned}$$

12.8 Change of Variables

In this section, we demonstrate the specific set of assumptions that lead to a linear transformation for the purposes of adjusting the second derivative of the MVBP.

Derivation

Suppose that Σ is a $d \times d$ dimensional covariance matrix and it has the density $p_\Sigma(\Sigma)$, while $\Psi = \Psi(\Sigma)$ is a one-to-one transformation with density $p_\Psi(\Psi)$. Similarly, the logarithms of the density will be l_Σ and l_Ψ respectively. Denote differentiation with respect to Ψ by a '′' and differentation with respect to Σ by an over-set '$\cdot$'. Writing $\Sigma(\Psi) = \Sigma$, we have:

$$
\begin{aligned}
l_\Psi(\Psi) &= l_\Sigma(\Sigma) + \log|\Sigma'|, \\
l'_\Psi(\Psi) &= \dot{l}_\Sigma(\Sigma)\Sigma' + \operatorname{vec}\left(\Sigma'^{-1}\right)^\top \Sigma'', \\
l''_\Psi(\Psi) &= \Sigma'^\top \ddot{l}_\Sigma(\Sigma)\Sigma' + (\dot{l}_\Sigma(\Sigma) \otimes \mathrm{I}_{d^2}) T_{d^4,d^4}\Sigma'' + \Sigma''^\top(\Sigma'^{-\top} \otimes \Sigma'^{-1})\Sigma'' + \\
&\qquad \left(\operatorname{vec}(\Sigma'^{-1})^\top \otimes \mathrm{I}_{d^2}\right) T_{d^6,d^6}\Sigma'''.
\end{aligned}
$$

In the above, $T_{m,n}$ is a permutation matrix with the property that for A with dimensions $m \times n$, $\operatorname{vec}(A)^\top = T_{m,n}\operatorname{vec}(A)$.

If $\dot{l}(\hat{\Sigma}) = 0$ is the unique mode of p_Σ and we make the following assumptions,

1. $\Psi(\hat{\Sigma}) = \hat{\Sigma}$ - *i.e.* the transformed variable has the same mode,
2. $\Sigma''(\hat{\Sigma}) = 0$ - the transformed variable has the same information,
3. $\Sigma'''(\Psi) = 0$,

then the linear transformation follows and we are able to directly adjust the second derivative of the transformed density at its mode. In fact, by 1 and 2,

$$
l'_\Psi(\hat{\Sigma}) = \dot{l}_\Sigma(\hat{\Sigma})\Sigma'(\hat{\Sigma}) + \operatorname{vec}(\Sigma'(\hat{\Sigma}))^\top \Sigma''(\hat{\Sigma}) = 0.
$$

By 1 and 3,

$$l''_{\Psi}(\hat{\Sigma}) = \Sigma'(\hat{\Sigma})^{\top} \ddot{l}(\hat{\Sigma}) \Sigma'(\hat{\Sigma}).$$

In deriving a transformation, we find the first assumption to be natural. It alone prompts that $\text{vec}(\Sigma'(\hat{\Sigma}))^{\top} \Sigma''(\hat{\Sigma}) = 0$. The first term cannot be 0 as that would imply that Ψ is actually a constant. While components of each term could be zero such that the total product is as well, *a-priori* no rationale for doing so seems to exist. That leaves that the second term must be zero, or assumption 2. Finally, the third assumption is simply a matter of convenience.

Specificaly, the third assumption makes the adjusting the second derivative at the mode a simple matter of solving a bi-linear equation. If we want $l''_{\Psi}(\hat{\Sigma}) = C$,

$$\begin{aligned} C &= \Sigma'(\hat{\Sigma})^{\top} \ddot{l}_{\Sigma}(\hat{\Sigma}) \Sigma'(\hat{\Sigma}), \\ C^{1/2} &= \ddot{l}_{\Sigma}(\hat{\Sigma})^{1/2} \Sigma'(\hat{\Sigma}), \\ \Sigma'(\hat{\Sigma}) &= \ddot{l}(\hat{\Sigma})^{-1/2} C^{1/2}. \end{aligned}$$

Because $\Sigma(\Psi)$ must be a linear transformation, we then have $\text{vec}(\Sigma(\Psi)) = \ddot{l}_{\Sigma}(\hat{\Sigma})^{-1/2} C^{1/2} \,\text{vec}(\Psi) +$ const. In order to preserve the mode,

$$\begin{aligned} \text{vec}(\Sigma(\Psi)) &= \ddot{l}_{\Sigma}(\hat{\Sigma})^{-1/2} C^{1/2} \,\text{vec}(\Psi - \hat{\Sigma}) + \text{vec}(\hat{\Sigma}), \\ \text{vec}(\Psi(\Sigma)) &= C^{-1/2} \ddot{l}_{\Sigma}(\hat{\Sigma})^{1/2} \,\text{vec}(\Sigma - \hat{\Sigma}) + \text{vec}(\hat{\Sigma}). \end{aligned}$$

Application

Finally, we can connect this to the previous section and utilize the second derivative of the MVBP at its mode. If I is the second derivative of the log marginal posterior at its mode, or any desired second derivative for that matter, and M is that mode, then desired

transformations are:

$$\operatorname{vec}(\tilde{\Sigma}(\Sigma)) = \sqrt{\frac{2}{\nu+d+1}\frac{\nu+\mu}{\mu-d-1}}(M^{1/2}\otimes M^{1/2})I^{1/2}\operatorname{vec}(\Sigma - M) + \operatorname{vec}(M),$$
$$\operatorname{vec}(\Sigma(\tilde{\Sigma})) = \sqrt{\frac{\nu+d+1}{2}\frac{\mu-d-1}{\nu+\mu}}I^{-1/2}(M^{-1/2}\otimes M^{-1/2})\operatorname{vec}(\tilde{\Sigma} - M) + \operatorname{vec}(M),$$

and we have moved from $\tilde{\Sigma}$ to Σ while maintaining the desired properties that $l''_\Sigma(M) = I$ and $l'_\Sigma(M) = 0$.