Avoiding Boundary Estimates in Linear Mixed Models Through Weakly Informative Priors

Yeojin Chung*, Sophia Rabe-Hesketh**
* **Graduate School of Education, University of California, Berkeley, CA 94705, USA
**Institute of Education, University of London

Andrew Gelman***, Jingchen Liu*** and Vincent Dorie***
***Department of Statistics, Columbia University, New York, NY 10027, USA

Abstract

Variance parameters in mixed or multilevel models can be difficult to estimate, especially when the number of groups is small. We propose a maximum penalized likelihood approach which is equivalent to estimating variance parameters by their marginal posterior mode, given a weakly informative prior distribution. By choosing the prior from the gamma family with at least 1 degree of freedom, we ensure that the prior density is zero at the boundary and thus the marginal posterior mode of the group-level variance will be positive. The use of a weakly informative prior allows us to stabilize our estimates while remaining faithful to the data.

1 Introduction

Maximum marginal likelihood is a useful way to estimate variance parameters in mixed models. But when the number of groups is small, estimates of group-level variance parameters can be noisy and can often be zero. In a multivariate setting, estimated covariance matrices can be degenerate non-positive-definite.

We propose a method that pulls the variance estimates off the boundary and makes them more stable by maximizing the marginal likelihood multiplied by a penalty function, or equivalently by assigning a prior distribution to the unknown variance parameters and finding the marginal posterior mode. Such a regularized approach will solve our problem as long as we have a prior or penalty function that goes to zero at the boundary—but without requiring the sort of strong prior knowledge which would limit the routine use of this approach. While Bayes estimates are good if real prior information is available, here we are in the more common statistical problem of searching for a generic Bayes inference with good frequency properties.

Is this possible? Is there a default choice of prior that gives reasonable, stable inference for variance parameters in multilevel models, priors that are zero at the boundary yet automatically respect the data? Amazingly, it turns out the answer is yes.

In this paper, we aim at developing a default choice of prior that gives reasonable stable inference for variance parameters in multilevel models, priors that are zero at the boundary yet automatically respect the data. In particular, we recommend a class of gamma priors (for unidimensional problems) and Wishart (for multidimensional) that produce Bayes modal estimates approximately one standard error away from zero when the maximum likelihood estimate is at zero. We consider these priors to be weakly informative in the sense that they supply some direction but still allow inference to be driven by the data.

1.1 Background

Linear mixed models, also known as hierarchical or multilevel linear models, are widely used in the biomedical and social sciences. Typical applications include longitudinal data, observational data on subjects nested in institutions (hospitals, schools, firms) or in neighborhoods, cluster-randomized trials, multi-site trials, and meta-analysis.

Most statistical software packages, including Stata, R and SAS, fit the models by (restricted) maximum likelihood. The resulting estimates of group-level variance-covariance parameters are often on the boundary of the parameter space. For example, in a cluster-randomized trial of a hospital-level intervention, the between-hospital variance in patient outcomes may be known to be large, but the estimate may be zero, particularly if there are only a small number of hospitals in the trial.

The problem of zero estimates does not arise in single-level models, where the log-likelihood approaches $-\infty$ as the variance parameter approaches zero. In multilevel models, however, there are no direct data on the group-level variance; as a result, the likelihood does not rule out zero variances, and when noise levels are high, the marginal likelihood can happen to have a maximum at zero.

Variance components estimated as zero can cause practical problems. First, zero variance estimates can go against prior knowledge of researchers. For example, Gelman et al. (2007) fit a multilevel model predicting voter choice given income, with the intercept and slope for income varying by state. They found that richer voters tended to support Republican candidates but with a slope that varied depending on some state-level predictors. For one election year, the fitted model had a zero value for the point estimate of the variacne of the state-level errors for the slopes. In the resulting inferences, the state-level slopes were perfectly predicted by the state-level predictors. There is no reason to believe this—the perfect prediction is merely an artifact of a variance estimate that happened to be zero—and it is awkward to graph these results, showing an estimated perfect fit that we do not and

should not believe. A related difficulty arises when comparing instances of a model that is repeatedly fit, to similar data from different surveys or different years, yielding zero variance estimates some of the time.

A second problem with boundary estimates of variance components is the resulting underestimation of uncertainty in coefficient estimates. For instance, in a cluster-randomized study or meta-analysis, researchers might be overconfident in concluding that a treatment is effective.

Third, group comparisons are often of interest to researchers, but zero group-level variance estimates make it impossible to get useful inferences. When the group-level variance is estimated as zero, the resulting predictions of the group-level errors will all be zero, so one fails to find unexplained differences between groups.

Finally, convergence problems can occur for optimization algorithms that is performed on the scale of log-variance.

In models with varying intercepts and slopes, boundary estimates for correlations can also cause problems. For example, correlations of 1 or -1 between varying slopes and intercepts imply that the slope is perfectly predicted from the intercept and vice versa which rarely makes substantive sense. And in practice such a conclusion is usually not strongly supported by the data; rather, what typically happens is that the likelihood is close to being flat and happens to have a maximum at the boundary. When such estimates are found, it is common to simplify the model and remove one of the varying dimensions, which can lead to poorly estimated standard errors. Convergence problems can again occur if the covariance matrix is constrained to be positive-definite, and some software in this case automatically reverts to the model with constant slopes.

These problems have not been studied much in the past. Most of the literature on variance estimation focuses on the covariance matrix of observed variables, whereas we are interested in the covariance matrix of *latent* variables (coefficients that vary by group). With observed variables, zero variances and degenerate covariance matrices are ruled out by the likelihood (except when the data themselves happen to fall on a linear subspace), whereas with latent variables, some amount of the observed variation in the data can be explained by the within-group variance, and as a result the likelihood never rules out degenerate possibilities. In fact, as we shall see, degenerate point estimates can happen frequently.

1.2 Outline of our approach

We propose a maximum penalized marginal likelihood estimator based on a weakly informative prior distribution for the group-level variance parameters. We maximize the marginal posterior distribution, with the varying intercepts integrated out.

Bayes modal estimation has previously been used to obtain more stable estimates of item

parameters in item response theory (Swaminathan and Gifford, 1985; Mislevy, 1986; Tsutakawa and Lin, 1986) and to avoid boundary estimates in log-linear models (Galindo-Garre et al., 2004) and latent class analysis (Maris, 1999; Galindo-Garre and Vermunt, 2006). To our knowledge, this idea has not yet been applied to variance parameters in multilevel models.

For varying-intercept models, we propose a prior distribution that prevents boundary estimates but increases the variance estimate by no more than about one standard error, and hence has little influence when the data are informative about the variance.

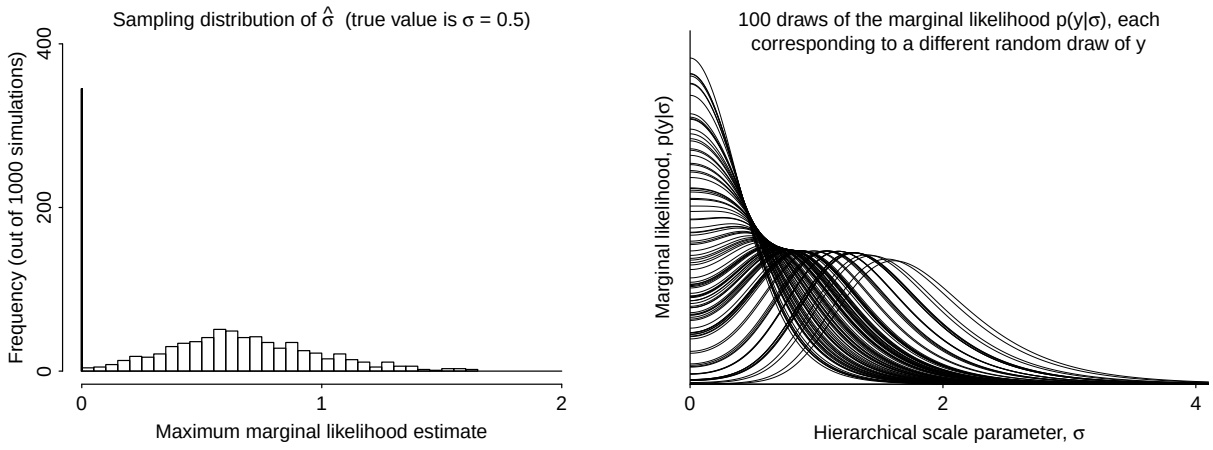
We compare the inferences from our procedure to those under maximum likelihood and restricted maximum likelihood estimation in simulations across a wide range of conditions. Our method performs well, not only in terms of parameter estimation but also in providing better estimates of standard errors of regression coefficients in many situations.

In addition, our method to avoid boundary estimates is flexible enough to include stronger prior information when available, by specifying a scale parameter in the prior density.

Our method can be considered as posterior modal estimation with a uniform prior for the group-level variance after applying a log transformation to make the posterior distribution more symmetric and the posterior mode closer to the posterior mean. Bayes modal inference for other Box-Cox transformations of the group-level variance can be achieved by tuning the shape parameter of the prior.

Compared with full Bayes or posterior mean estimation, our approach does not require simulation and is computationally as efficient as maximum likelihood estimation, in fact potentially more efficient as it avoids the slow convergence that can occur if the maximum likelihood estimate is on the boundary. No elaborate convergence checking is required and there is no need to specify priors for all model parameters. We have implemented posterior modal estimation in Stata and R with only minor modifications of existing software for maximum likelihood estimation of linear mixed models. Given user-specified or default choices of hyperparameters, the programs automatically find the posterior mode of the variance parameter and provide inferences for the coefficients conditional on that estimate.

Our method has natural extensions to models beyond the linear mixed model with a varying intercept. For the model with varying intercept and slopes, the proposed posterior modal estimation method can be generalized using the relationship between the gamma and Wishart distributions. Since we propose a principled method to avoid boundary estimates, we can extend it to other models in which there are variance parameters that could be estimated at zero including generalized linear mixed models and hierarchical models with more than two levels.



(a) (b)

Figure 1: *From a simple one-dimensional hierarchical model with scale parameter 0.5 and data in 10 groups: (a) Sampling distribution of the maximum marginal likelihood estimate $\widehat{\sigma}_\theta$, based on 1000 simulations of data from the model. (b) 100 simulations of the marginal likelihood, $p(y|\sigma_\theta)$. In this example, the maximum marginal likelihood estimate is extremely variable and the likelihood function is not very informative about σ_θ.*

1.3 Boundary problem for a simple model

We demonstrate the problem with a varying-intercept model with $J = 10$ groups and a single group-level variance parameter. To keep things simple, we do not include covariates and treat the mean and within-group variance as known:

$$y_j \sim \mathrm{N}(\theta_j, 1), \quad \theta_j \sim \mathrm{N}(0, \sigma_\theta^2), \qquad \text{for } j = 1, \ldots, J.$$

In our simulation, we set the group-level standard deviation σ_θ to 0.5. From this model, we create 1000 simulated datasets and estimate σ_θ by maximum marginal likelihood by solving for $\widehat{\sigma}_\theta$ in the equation $1 + \widehat{\sigma}_\theta^2 = \frac{1}{J}\sum_{j=1}^J y_j^2$, with the boundary constraint that $\widehat{\sigma}_\theta = 0$ if $\frac{1}{J}\sum_{j=1}^J y_j^2 < 1$. In this simple example, it is easy to derive the probability of obtaining a boundary estimate as $\Pr(\chi^2(J) < \frac{J}{1+\sigma_\theta^2}) = 0.37$.

Figure 1(a) shows the sampling distribution of the maximum marginal likelihood estimate of σ_θ. As expected, in more than a third of the simulations, the marginal likelihood is maximized at $\widehat{\sigma}_\theta = 0$. The noise is so much larger than the signal here that it is impossible to do much more than bound the group-level variance; the data do not allow an accurate estimate.

Figure 1b displays 100 draws of the marginal likelihood function, which shows in a different way that the maximum is likely to be on the boundary, with there being quite a bit of

uncertainty. We want a point estimator that is positive while being consistent with the data.

2 Bayes modal estimation with weakly informative priors

2.1 A brief review of the maximum likelihood and restricted maximum likelihood estimation

We consider the model

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

where $\boldsymbol{x}_{ij}$ and y_{ij} are observed; $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients that do not vary by group; $\theta_j \sim N(0, \sigma_\theta^2)$ is a group-level error; and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is a residual for each observation. We further assume that θ_j and ϵ_{ij} are independent.

The parameters $(\boldsymbol{\beta}, \sigma_\theta, \sigma_\epsilon)$ in (1) are commonly estimated by maximizing the marginal likelihood (the integral of the joint likelihood, averaging over the varying intercepts θ_j). Another option is restricted or residualized maximum likelihood (REML, Patterson and Thompson, 1971), which is equivalent to the marginal posterior mode, averaging also over a uniform prior on $\boldsymbol{\beta}$ (Harville, 1974). Unlike the maximum likelihood estimator, the REML estimator of σ_θ^2 is unbiased in balanced designs if it is allowed to be negative.

Discussion of small-sample inference for hierarchical models has largely focused on the covariance matrix of $\widehat{\boldsymbol{\beta}}$, say $V(\widehat{\boldsymbol{\beta}})$ (Kenward and Roger, 1997). Longford (2000) points out that this covariance matrix is often poorly estimated because variance components are estimated inaccurately. The sandwich estimator (Huber, 1967; White, 1990) is asymptotically consistent even if the distributional assumptions are violated. However, as Drum and McCullagh (1993) note, it can perform poorly when the sample size is small. Crainiceanu et al. (2003) derive a general expression for the probability that the (local) maximum of the marginal (or restricted) likelihood is at the boundary for linear mixed models and Crainiceanu and Ruppert (2004) discuss the finite-sample distribution of the likelihood ratio statistic for testing null hypotheses regarding the group-level variance.

2.2 Bayes modal estimation

In the present article we are particularly concerned with the group-level standard deviation, and we specify a prior $p(\sigma_\theta)$ only for σ_θ, implicitly assuming a uniform prior, $p(\boldsymbol{\beta}, \sigma_\epsilon)$, on $\boldsymbol{\beta}$ and σ_ϵ.

The marginal log-posterior density (with varying coefficients integrated out) can be written as

$$\log p(\sigma_\theta, \boldsymbol{\beta}, \sigma_\epsilon | \boldsymbol{y}) = \log p(\boldsymbol{y} | \sigma_\theta, \boldsymbol{\beta}, \sigma_\epsilon) + \log p(\sigma_\theta) + c, \tag{2}$$

where the first term of the right hand side is the marginal log-likelihood and c is a constant. We find the parameters that maximize (2). By integrating the posterior over θ, we avoid the incidental parameter problem (Neyman and Scott, 1948; O'Hagan, 1976; Mislevy, 1986).

The marginal posterior density for $(\boldsymbol{\beta}, \sigma_\theta, \sigma_\epsilon)$ can equivalently be regarded as a penalized marginal likelihood.

2.3 Desired properties of a weakly informative prior

Our goal is to find a prior or penalty function for σ_θ so that the posterior mode is off the boundary, but with the prior being weak enough so that inferences are consistent with the data.

For our purpose, we desire a prior on σ_θ that

(i) is zero at the origin and

(ii) has a positive constant derivative at zero.

Condition (i) ensures a positive estimate of the variance parameter, even when the maximum of the likelihood is at 0. Condition (ii) allows the likelihood to dominate if it is strongly curved near zero. The positive constant derivative implies that there is no "dead zone" in the prior near zero—that is, the prior does not rule out positive values near zero if they are supported by the likelihood.

For our default choice of prior we do not impose any restriction on the right tail of $p(\sigma_\theta)$: our primary concern here lies in the boundary estimates and the right tail has little impact on that. If the number of groups is small and we want to further control the estimate, it would make sense to assign a finite scale to the prior to constrain the right tail.

Various reasonable-seeming choices of priors do not satisfy both the above conditions. The *exponential* and *half-Cauchy* families, for example, do not decline to zero at the boundary, so they do not rule out posterior mode estimates of zero. Such priors can be excellent weakly informative priors for full Bayesian (posterior mean) inference (see Gelman, 2006) but do not work if the goal is to get a stable and reasonable posterior mode estimate.

The *lognormal* and *inverse-gamma* densities satisfy condition (i) but not condition (ii). They have a zero derivative at the origin, essentially ruling out low estimates of σ_θ no matter what the data suggest. For any choice of prior in either of these families, there is some ϵ below which the prior is essentially zero, and you can find data for which the posterior mode is inconsistent with the data. Thus, the lognormal can only be used when there is real prior information to guide the choices of its two parameters; it cannot be a default choice of the sort we are seeking here.

3 Gamma prior

We propose a gamma (not inverse-gamma) prior on σ_θ: defined by

$$p(\sigma_\theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\sigma_\theta^{\alpha-1}e^{-\lambda\sigma_\theta}, \quad \alpha > 0, \lambda > 0 \tag{3}$$

with mean α/λ and variance α/λ^2, where α is the shape parameter and λ is the rate parameter (the reciprocal of the scale parameter).

With an appropriate choice of parameters, the gamma satisfies the two conditions for the weakly informative prior listed in the previous section. For any $\alpha > 1$, gamma(α, λ) satisfies the first condition that $p(0) = 0$. In order to have a positive constant derivative at zero (the second condition), α can be chosen to be 2.

3.1 Default choice and other options

We consider three ways to apply the gamma prior as penalty function:

- Our *default choice* is gamma(α, λ) with $\alpha = 2$ and $\lambda \to 0$, which is the (improper) density ($p(\sigma\theta) \propto \sigma_\theta$). As we discuss shortly, this default bounds the posterior mode away from zero while keeping it consistent with the likelihood.

- Sometimes we have *weak prior information* about a variance parameter that we would like to include in our model. When $\alpha = 2$, the gamma density has its mode at $1/\lambda$, and so our recommendation is to use the gamma(α, λ) prior with $1/\lambda$ set to the prior estimate of σ_θ.

- If *strong prior information* is available, then both parameters of the gamma density can be set to encode this. If α is given a value higher than 2, property (ii) above will no longer hold, but this is acceptable if this represents real information about σ_θ.

3.2 The effect of the prior on the posterior mode

To examine the effect of α and λ on the posterior mode analytically, we treat $(\boldsymbol{\beta}, \sigma_\epsilon)$ as nuisance parameters and assume that the marginal profile log-likelihood can be approximated by a quadratic function in σ_θ,

$$\log L(\sigma_\theta) \approx -\frac{(\sigma_\theta - \mu)^2}{2\tau^2} + c_1. \tag{4}$$

The mode μ corresponds to the maximum likelihood estimate of σ_θ, and the standard deviation τ corresponds to the estimated asymptotic standard error of σ_θ (based on the observed

information). Under this assumption, we derive a number of properties of the gamma(α, λ) prior on σ_θ.

Property 1. *The posterior mode is*

$$\widehat{\sigma}_\theta \;=\; -\frac{\lambda\tau^2}{2} + \frac{\mu}{2} + \frac{1}{2}\sqrt{(\tau^2\lambda - \mu)^2 + 4(\alpha-1)\tau^2}. \tag{5}$$

In what follows, we discuss the behavior of $\hat{\sigma}_\theta$ for two cases: given under Property 2 for $\mu = 0$ and Property 3 for $\mu > 0$.

Property 2. *When $\mu = 0$, for fixed $\alpha > 1$ and τ, the largest posterior mode is attained when $\lambda \to 0$ with the value*

$$\widehat{\sigma}_\theta = \tau\sqrt{\alpha - 1}. \tag{6}$$

With a simple calculation, we can show that $\partial\widehat{\sigma}_\theta/\partial\lambda \leq 0$. Therefore, as $\lambda \to 0$ for fixed α and τ, the posterior mode increases monotonically to the maximum. With $\alpha = 2$, the largest possible posterior mode is τ. That is, when the maximum likelihood estimate is on the boundary, the gamma$(2, \lambda)$ prior shifts the posterior mode away from zero but not more than one standard error.

One standard error can be regarded as a statistically insignificant distance from the maximum likelihood estimate. If the quadratic approximation in (4) holds and the maximum likelihood estimate μ is zero, the likelihood-ratio test statistic for $H_0 : \sigma_\theta = \tau$ is $2(\log L(0) - \log L(\tau)) = 1$. For the null hypothesis $\sigma_\theta = 0$, it is known that asymptotic distribution (as J approaches infinity) of the test statistic is $0.5\chi^2_0 + 0.5\chi^2_1$ with 99th percentile 5.41. In finite samples, the mass at zero is larger and the 99th percentile is smaller, but even with $J = 5$, the 99th percentile is as large as 3.48, in a model without covariates and large cluster size (Crainiceanu and Ruppert, 2004). For testing the null hypothesis that $\sigma_\theta = \tau > 0$, the percentile will be larger because there is less point mass at zero (Crainiceanu et al., 2003). Therefore, a likelihood ratio test statistic of 1 can be considered small.

Property 3. *If $\mu > 0$ and $\alpha > 1$, the largest possible posterior mode is attained when $\lambda \to 0$ with the value*

$$\widehat{\sigma}_\theta = \frac{\mu}{2} + \frac{\mu}{2}\sqrt{1 + 4(\alpha-1)\tau^2/\mu^2} > \mu.$$

In addition, $\partial\widehat{\sigma}_\theta/\partial\tau$ decreases in μ.

The gradient $\partial\widehat{\sigma}_\theta/\partial\tau = \frac{\alpha-1}{\sqrt{\alpha-1+\mu^2/(4\tau^2)}}$ is always less than $\sqrt{\alpha-1}$. (Recall that $\partial\hat{\sigma}_\theta/\partial\tau = \sqrt{\alpha-1}$ for $\mu = 0$.) In addition, this derivative becomes smaller as μ increases. This implies that, when λ is close to zero, the gamma(α, λ) prior does not shift the posterior mode as much when the maximum likelihood estimate is away from zero as it does when the maximum likelihood estimate is zero. That is, while the gamma prior helps us avoid boundary estimates

when the maximum likelihood estimate is on the boundary, it has less influence on the estimate when the maximum likelihood estimate is plausible.

3.3 Transformation of σ_θ

When the posterior density of σ_θ is asymmetric, a transformation of σ_θ can make the density more symmetric so that the posterior mode will be located near the posterior mean which has good asymptotic properties.

Property 4. *As $\lambda \to 0$, a gamma$(2, \lambda)$ prior on σ_θ is equivalent to a uniform$(0, \infty)$ prior on σ_θ after log transformation of σ_θ.*

With the uniform (improper) prior on σ_θ in the range of $(0, \infty)$, the log-transformed σ_θ has the prior $p(\log \sigma_\theta) \propto \sigma_\theta$ and this is equivalent to gamma$(2, \lambda)$ on σ_θ as $\lambda \to 0$.

With the uniform prior on σ_θ, the marginal posterior density is just the marginal likelihood, which is often right-skewed or even has its mode at $\sigma_\theta = 0$ (where the boundary estimation problem occurs). In this case, the log transformation of σ_θ can make the shape of the posterior more symmetric.

The equivalence of changing α and transforming σ_θ can be generalized to a family of power transformations beyond the log. Consider the Box and Cox (1964) transformations, defined by

$$g_\gamma(\sigma_\theta) = \begin{cases} \frac{\sigma_\theta^\gamma - 1}{\gamma} & \text{if } \gamma \neq 0; \\ \log(\sigma_\theta) & \text{if } \gamma = 0 \end{cases}$$

Property 5. *A gamma(α, λ) prior on σ_θ is equivalent to a gamma$(\alpha + 1 - \gamma, \lambda)$ prior on $g_\gamma(\sigma_\theta)$.*

From the Jacobian $\sigma_\theta^{1-\gamma}$ of the inverse transformation, the prior $p(g_\gamma(\sigma_\theta))$ of $g_\gamma(\sigma_\theta)$ is proportional to $\sigma_\theta^{\alpha-\gamma} e^{-\lambda\sigma_\theta}$, which is proportional to gamma$(\alpha - \gamma + 1, \lambda)$. Therefore, any order of power transformation of σ_θ for obtaining a more symmetric posterior density is equivalent to adjusting α to $\alpha + 1 - \gamma$ without transforming σ_θ. Property 4 is a special case with $\alpha = 1$, $\gamma = 0$ and $\lambda \to 0$.

Although we have discussed the gamma prior on the group-level standard deviation (σ_θ), one might want to consider priors on the variance, σ_θ^2. If we assign the gamma(α, λ) on σ_θ^2 instead of σ_θ, the logarithm of the prior becomes $\log p(\sigma_\theta^2) = 2(\alpha-1)\log\sigma_\theta - \lambda\sigma_\theta^2$. In the limit $\lambda \to 0$, the term $2(\alpha-1)\log\sigma_\theta$ is the same as the corresponding term of gamma$(2\alpha-1, \lambda)$ on σ_θ. Therefore a gamma(α, λ) prior on σ_θ^2 has almost the same effect on the posterior mode as a gamma$(2\alpha - 1, \lambda)$ prior on σ_θ when $\lambda \to 0$.

3.4 Connection to REML

In Section 2.1, we mentioned that REML gives an unbiased estimate for variance components in the balanced case (when negative variance estimates are permitted). In this section, we regard REML as a penalized likelihood estimator and compare the REML penalty with the log of the gamma density, considered as a penalty on the log-likelihood.

Longford (1993) describes the REML log-likelihood, say $\log L_R$, in terms of the original log-likelihood, L, and an additive penalty term,

$$\log L_R = \log L - \frac{1}{2}\log\left(\det(X^T V^{-1} X)\right) \tag{7}$$

where V is the $N \times N$ covariance matrix of the vector of all responses $\boldsymbol{y}$ and X is the design matrix with rows $\boldsymbol{x}_{ij}^T$. In the varying-intercept model in (1), V is a block-diagonal matrix with $n_j \times n_j$ blocks, $V_j,\ j = 1, \ldots, J$, where V_j contains $\sigma_\theta^2 + \sigma_\epsilon^2$ in the diagonal and σ_θ^2 in the off-diagonals. Recalling that the log-posterior density is the sum of the log-likelihood and the log-prior density in (2), the second term in (7), denoted by $\log p_R(\sigma_\theta)$, is analogous to the log of the gamma prior.

In order to compare the REML penalty and log gamma density in a simple closed form, we consider a special case of model (1) with balanced group size n, q level-1 covariates, and r level-2 covariates. The level-1 covariates, written as columns $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q$ of the design matrix, consist of the same elements for each group and satisfy $\mathbf{1}^T \boldsymbol{z}_u = 0$, $\boldsymbol{z}_u^T \boldsymbol{z}_{u'} = 1$ if $u = u'$, and 0 otherwise for $u = 1, \ldots, q$. The level-2 covariates are assumed to be dummy variables for the first $r (< J - q - 2)$ groups. Then the REML penalty becomes

$$\log p_R(\sigma_\theta) \quad = \quad \frac{r+1}{2}\log\left(\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n}\right) + c_1 \tag{8}$$

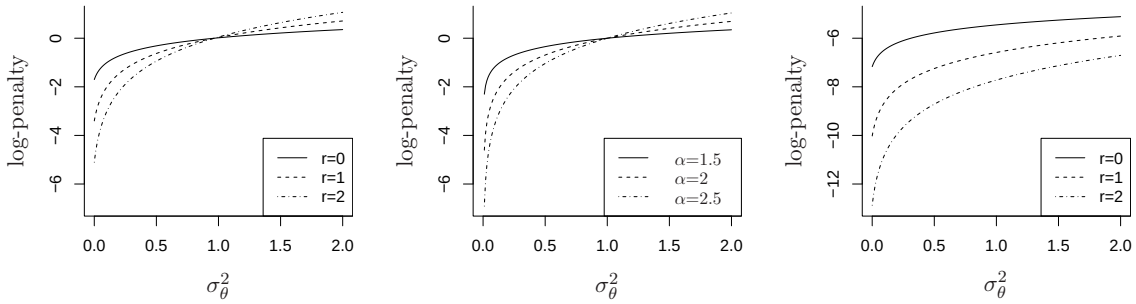
where c_1 is a constant. The proof is provided in the web-based supplementary materials.

Recall that, when $\lambda \to 0$, the gamma(α, λ) prior on σ_θ^2 (equivalently gamma$(2\alpha - 1, \lambda)$ on σ_θ) has log density,

$$\log p(\sigma_\theta^2) = (\alpha - 1)\log \sigma_\theta^2 + c_2. \tag{9}$$

Ignoring the constant terms that have no influence on the posterior mode, we see that the gamma$((r+1)/2 + 1, \lambda)$ on σ_θ^2 (equivalently gamma$(r + 2, \lambda)$ on σ_θ) approximately matches the REML penalty, particularly when the group-size n is large and λ is close to zero.

Figure 2 compares the REML penalty function in (8), the log of the gamma density with corresponding $\alpha = (r+1)/2 + 1$, and the REML penalty function in the second term of (7) for a dataset with $n = 30$, $J = 5$, $q = 1$, $r = 0$, 1, or 2, which does not have the form assumed when deriving (8). For the latter, the columns of the covariate matrix X consist of a vector of ones, a level-1 covariate $\boldsymbol{z}_1$ with $z_{1ij} = i$ and two level-2 covariates $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ where $w_{1j} = j$



(a) REML penalty in (8) with $n = 30$, $\sigma_\epsilon = 1$

(b) gamma(α, λ) prior

(c) REML penalty for data with $n = 30, q = 1, J = 5$

Figure 2: *REML log-penalty function compared with gamma*$(\alpha = (r+1)/2 + 1, \lambda)$ *prior, with* $\lambda = 10^{-4}$. *The shapes of the curves agree quite well except when* σ_θ *is close to 0 where the gamma prior tends to 0.*

for all $j = 1, \ldots, j = J$ and $\boldsymbol{w}_2$ is the same as $\boldsymbol{w}_1$ except that the values for the last group are 0 instead of J. Comparing Figure 2(a) and (c), the penalties differ by a constant which does not affect the mode, so formula (8) appears to hold more generally.

For Figure 2(a) and (b), the constant terms were ignored to make the figures easier to compare. The REML penalty functions with $r = 0$, 1, and 2 look very similar to the gamma penalty on σ_θ^2 with α=1.5, 2.0, and 2.5, respectively, except where σ_θ is close to zero. At $\sigma_\theta = 0$, the log of the gamma prior is $-\infty$ for $\alpha > 1$, whereas the REML penalty approaches $-\infty$ only if $\sigma_\epsilon \to 0$ or $n \to \infty$. This explains why REML can produce boundary estimates. Further, it implies that the gamma prior assigns more penalty on σ_θ^2 close to zero than REML for small n and large σ_ϵ. Otherwise, REML can approximately be viewed as a special case of our method with a gamma prior.

3.5 Implementation in Stata and R

We have implemented posterior modal estimation with a family of gamma priors in Stata and R. Compared with posterior mean estimation, our method is less computationally intensive and is easy to implement by modifying existing maximum likelihood estimation procedures for (generalized) linear mixed models such as `gllamm` (Rabe-Hesketh et al., 2005; Rabe-Hesketh and Skrondal, 2008) in Stata and `lme4` (Bates and Maechler, 2010) in R.

In Stata, `gllamm` maximizes the marginal log-likelihood by the Newton-Raphson algorithm. We add the logarithm of the prior to the marginal log-likelihood so that the posterior mode can be estimated by the Newton-Raphson algorithm. `gllamm` with a gamma prior as an option is available from the Boston College Department of Economics Statistical Software

Components (SSC).

In R, the `lmer` function in the `lme4` package directly computes the marginal and residualized log-likelihood using sparse matrix decomposition techniques (Bates, 2005). Our program `blmer` mildly alters `lmer` by adding a default or user-specified penalty function. The program can be found in the `arm` package and is available on the Comprehensive R Archive Network.

4 Application: meta-analysis of 8-schools data

Alderman and Powers (1980) report the results of randomized experiments of coaching for the Scholastic Aptitude Test (SAT) conducted in eight schools. The data for the meta-analysis consist of an estimated treatment effect and associated standard error for each school (obtained by separate analyses of the data of each school). The data have previously been analyzed by Rubin (1981) and Gelman et al. (2004) to compare different approaches for estimating effects in parallel studies. Meta-analyses such as this one, with a small number of studies, pose a challenge for variance estimation, leading to poor coverage of estimated confidence intervals for the overall treatment effect (e.g., Brockwell and Gordon, 2001).

Meta-analysis with varying intercepts (DerSimion and Laird, 1986), typically called random-effects meta-analysis, allows for heterogeneity among studies due to differences in populations, interventions, and outcomes. The model for the effect size y_i of study i can be written as

$$y_i = \mu + \theta_i + \epsilon_i, \quad \theta_i \sim N(0, \sigma_\theta^2), \quad \epsilon_i \sim N(0, s_i^2), \tag{10}$$

and allows the true effect $\mu + \theta_i$ of study i to deviate from the overall effect size μ by a study-specific amount θ_i. In addition, the estimated effect y_i for study i differs from its true value by an estimation error ϵ_i with standard deviation set equal to the standard error for study i.

Figure 3 shows the marginal profile log-likelihood (maximized with respect to μ) of σ_θ (left) and σ_θ^2 (right). On the left we see that the maximum likelihood estimate of σ_θ is zero and that the marginal profile log-likelihood function decreases slowly as σ_θ increases from zero. For instance, at $\sigma_\theta = 10$ the log-likelihood is only 1.2 less than at the maximum, suggesting that the data are consistent with such large values of the standard deviation.

Inference for σ_θ is important because it affects both the point estimate and estimated standard error of the overall effect size μ,

$$\widehat{se}(\widehat{\mu}) = \left[\sum_i \frac{1}{s_i^2 + \sigma_\theta^2} \right]^{-1/2}. \tag{11}$$

For example, the estimated standard error is 4.1 for $\sigma_\theta = 0$, compared with 5.5 for $\sigma_\theta = 10$ (the corresponding estimates of μ are 7.7 and 8.1, respectively.)

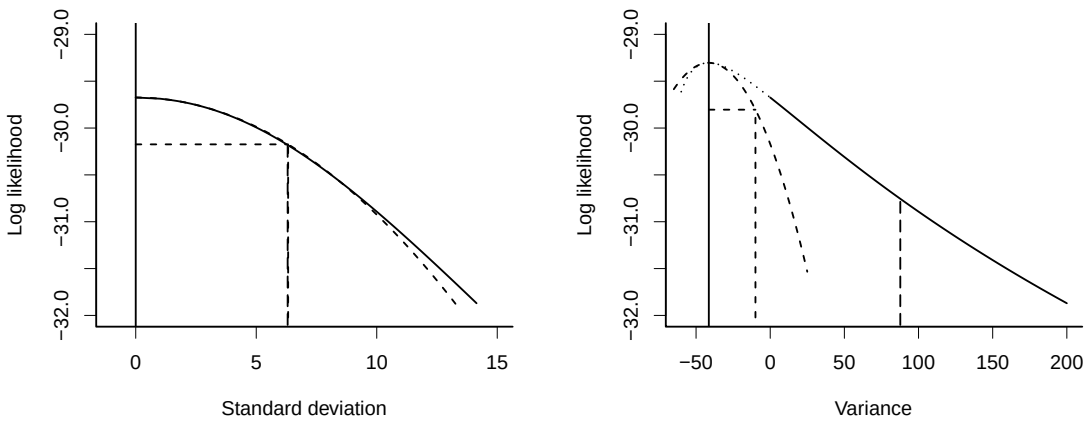


Figure 3: *Log marginal profile log-likelihood as a function of σ_θ (left) and σ_θ^2 (right) for 8-schools data. The dotted curve is the quadratic approximation at the mode, based on the estimated standard error. The vertical dotted line is one standard error away from the mode and the vertical dashed line is the Bayes modal estimate for a gamma(2,λ) prior on σ_θ (left) or σ_θ^2 (right). The quadratic approximation is good as a function of σ_θ (left) and consequently the Bayes modal estimate is one standard error away from the maximum likelihood estimate of zero. As a function of σ_θ^2 (right), the quadratic approximation is poor.*

Prior		μ			σ_θ		σ_θ^2		
Method	α	Est	SE	SE^R	Est	SE	Est	SE	Log-lik
ML		7.69	4.07	3.33	0	6.32	0	0.00	−29.67
ML[1]		7.13	3.19				−41.40	31.44	−29.30
gamma on σ_θ	2	7.92	4.72	3.39	6.30	4.61	39.73	58.15	−30.18
gamma on σ_θ	3	8.10	5.38	3.43	9.42	5.34	88.65	100.62	−30.76
gamma on σ_θ^2	1.5	7.92	4.72	3.38	6.28	4.79	39.42	57.65	− 30.18
gamma on σ_θ^2	2	8.09	5.37	3.42	9.37	5.30	87.71	99.23	−30.75

[1] allow $\sigma_\theta^2 < 0$

SE^R: robust (sandwich) standard error.

Table 1: Maximum likelihood and posterior mode estimates for the 8 schools data, where the prior is gamma(α, λ) on σ_θ or σ_θ^2, with $\lambda = 10^{-4}$. With gamma(α, λ) priors on σ_θ, the posterior mode estimates are approximately at $\tau\sqrt{1-\alpha}$ and agree well with the posterior mode estimates with gamma($(\alpha+1)/2, \lambda$) on σ_θ^2.

For the model in (10), we consider four different priors: gamma$(2, \lambda)$ and gamma$(3, \lambda)$ on σ_θ and gamma$(1.5, \lambda)$ and gamma$(2, \lambda)$ on σ_θ^2, where $\lambda = 10^{-4}$. Posterior mode estimates with these priors and maximum likelihood estimates are given in Table 1. The estimated standard error of the maximum likelihood estimate of σ_θ is 6.32 (which corresponds to τ in Section 3.2).

When the prior is on σ_θ (rows 3 and 4), the posterior mode estimates of σ_θ are at 6.30 and 9.42 for $\alpha = 2$ and $\alpha = 3$, respectively. These are close to the values $\tau\sqrt{\alpha - 1}$ with $\tau = 6.32$, which we expect with $\mu = 0$ if the marginal posterior log-likelihood is quadratic in σ_θ, as it appears to be in the left panel of Figure 3. In both cases, the marginal log-likelihood is only a little bit lower than the maximum.

The maximum likelihood estimate of σ_θ^2 is -41.40 when it is allowed to be negative. Specifying a gamma$(2, \lambda)$ prior on σ_θ^2 (row 6) gives estimates that agree well with those for a gamma$(3, \lambda)$ prior on σ_θ as expected (see Section 3.2). Similarly, a gamma$(1.5, \lambda)$ prior on σ_θ^2 (row 5) gives posterior mode estimates that are close to the estimates with gamma$(2, \lambda)$ on σ_θ. A gamma prior on σ_θ^2 with $\alpha = 1.5$ corresponds to REML with no level-2 covariates. While REML gives $\widehat{\sigma}_\theta = 0$ (not shown here), a gamma prior with $\alpha = 1.5$ gives a legitimate estimate and at the same time it reduces the marginal log-likelihood by only 0.5.

We see in the right panel of Figure 3 that the quadratic approximation at the mode is not a good approximation to the profile log likelihood as a function of σ_θ^2 and that the posterior mode estimate (for $\alpha = 2$) is considerably more than one standard error away from the maximum likelihood estimate.

Table 1 also reports model-based and robust standard error estimates for $\widehat{\mu}$. We see that the estimated model-based standard error of the estimated overall effect size μ increases with σ_θ as implied by (11), whereas the robust standard errors, based on the sandwich estimator, change very little. (The data and code for this example can be downloaded from the journal website.)

5 Simulation of balanced varying-intercept model

We consider a simple varying-intercept model,

$$y_{ij} = \beta_0 + \theta_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \epsilon_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, J \tag{12}$$

with $J = 3, 5, 10, 30$ groups and $n = 5, 30$ observations per group. This model includes two covariates: $x_{1ij} = i$ varies within groups only (its mean is constant across groups), and $x_{2ij} = j$ varies between groups only. The coefficients $\beta_0, \beta_1, \beta_2$ are fixed parameters, $\theta_j \sim N(0, \sigma_\theta^2)$ is a varying intercept for each group, and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is an error for each observation.

For each combination of J and n, we generate 100 datasets with true parameter values,

$\beta_0 = 0$, $\beta_1 = \beta_2 = 1$, $\sigma_\epsilon = 1$, and $\sigma_\theta = 0, 1/\sqrt{3}$, or 1, which correspond to intra-class correlations $\rho = 0, 0.25$ and 0.5, respectively. We obtain posterior mode estimates with gamma(2, λ) and gamma(3, λ) priors on σ_θ, where $\lambda = 10^{-4}$. The REML penalty correponds to $\alpha = 3$ since the model contains one group-level covariate. We compare posterior mode with maximum likelihood and REML estimates.

Boundary estimates Here we report the proportion of estimates of σ_θ that are on the boundary (less than 10^{-5}) when the true σ_θ is not zero ($1/\sqrt{3}$ and 1). For $\sigma_\theta = 1/\sqrt{3}$, 67% of maximum likelihood estimates and 46% of REML estimates are zero for $J = 3$ and $n = 5$. As J or n increases, the proportion decreases, but for $J = 5$ and $n = 30$, the proportion of estimates on the boundary is still 7% for maximum likelihood and 5% for REML.

When $\sigma_\theta = 1$, the same pattern occurs but estimates are on the boundary less often for a given condition. For $J = 3$ and $n = 5$, maximum likelihood produces 49% of estimates on the boundary compared with 34% for REML. When J increases to 5 and n to 30, 4% of maximum likelihood estimates and 1% of REML estimates are on the boundary. When $J = 30$, maximum likelihood and REML yield no boundary estimates for either value of σ_θ while $n = 30$ still gives some boundary estimates depending on J. Therefore, we can infer that the number of groups is more critical than the cluster size for avoiding boundary estimates.

In contrast to the maximum likelihood and REML estimates, the posterior mode estimates are never on the boundary in any of the simulation conditions. At the same time, the posterior mode estimates do not differ significantly from the maximum likelihood estimates. The likelihood ratio test for the model that restricts σ_θ to the posterior mode estimate $\widehat{\sigma}_\theta^{Bayes}$, is based on the test statistics

$$-2\left[\log L(\widehat{\sigma}_\theta^{Bayes}) - \log L(\widehat{\sigma}_\theta^{ML})\right]. \tag{13}$$

and this test statistics was calculated for each replicate. When $J > 3$, the largest test statistics among all the replicates and simulation condition is 2.6. Even for $J = 3$, the largest test statistic is 3.4. As discussed in Section 3.2, these values are not large.

Quadratic approximation We now assess how well some of the relationships hold that were derived in section 3 by assuming that the marginal profile log likelihood is quadratic. Figure 4 shows that the posterior modes calculated by the quadratic approximation of the marginal profile log-likelihood (see properties 2 and 3, where μ and τ are the maximum likelihood estimate of σ_θ and its standard error, respectively) agree well with the posterior mode estimates with a gamma(2,λ) prior on σ_θ for $J = 3$ and $J = 30$ when $\rho = 0.25$ and $n = 30$. However, when a gamma prior is specified on σ_θ^2 for $J = 3$ (not shown here), the quadratic approximation tends to underestimate the posterior mode. These findings are consistent with

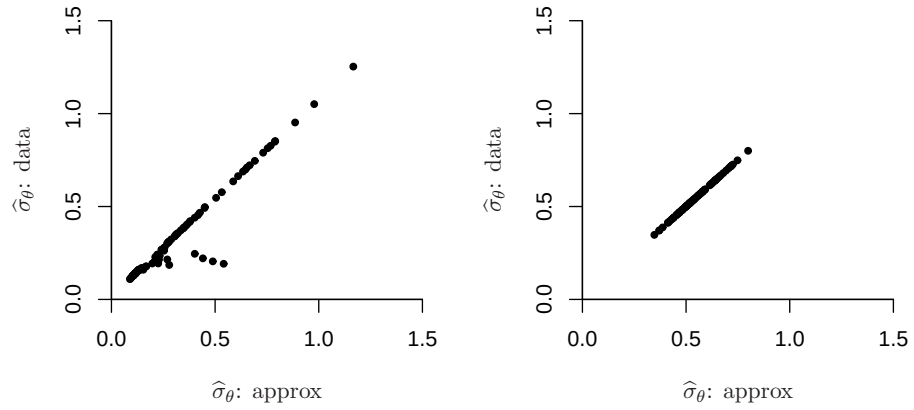


Figure 4: *Posterior mode estimates with a gamma(2,λ) prior on σ_θ for $\rho = 0.25$ and $n = 30$, compared with the posterior mode based on the quadratic approximation of the marginal profile likelihood (see properties 2 and 3, where μ and τ are the maximum likelihood estimate and standard error, respectively). Agreement is good, suggesting that the quadratic approximation is good. Dots on the left graph that fall off the line are due to a few samples that have uncommonly large standard errors.*

the observation, for the 8 schools example in Section 4, that the quadratic approximation of the marginal profile log-likelihood was a better approximation when considered as a function of σ_θ than when considered as a function of σ_θ^2. When $J = 30$, the quadratic approximation works well for a gamma prior on σ_θ or on σ_θ^2.

Estimation of σ_θ Figure 5 summarizes the bias of the maximum likelihood, REML, and posterior mode estimators of σ_θ. As J or n increase and as σ_θ decreases, the bias decreases. Thus the differences between methods are most obvious with small J or n, and particularly when the true σ_θ is not zero.

For $\sigma_\theta = 1/\sqrt{3}$ and 1, REML has the smallest bias in general and maximum likelihood tends to underestimate σ_θ. Posterior mode estimates with gamma$(2, \lambda)$ tends to be downward biased for σ_θ but not as much as maximum likelihood estimates. On the other hand, the posterior mode estimator with gamma$(3, \lambda)$ produces the largest estimates among the four estimators so it often overestimates σ_θ, but the amount of bias decreases as n increases. For $\sigma_\theta = 1$, the posterior mode estimator with gamma$(3, \lambda)$ is less biased than REML for both $n = 5$ and $n = 30$. In addition, the posterior mode estimator with gamma$(3, \lambda)$ is close to REML when $n = 30$ and σ_θ is not zero. This confirms that the gamma penalty on σ_θ with $\alpha = 3$ (equivalently gamma penalty on σ_θ^2 with $\alpha = 2$) agrees with the REML penalty when the model contains one group-level covariate, particularly with large n. When $\sigma_\theta = 0$, as

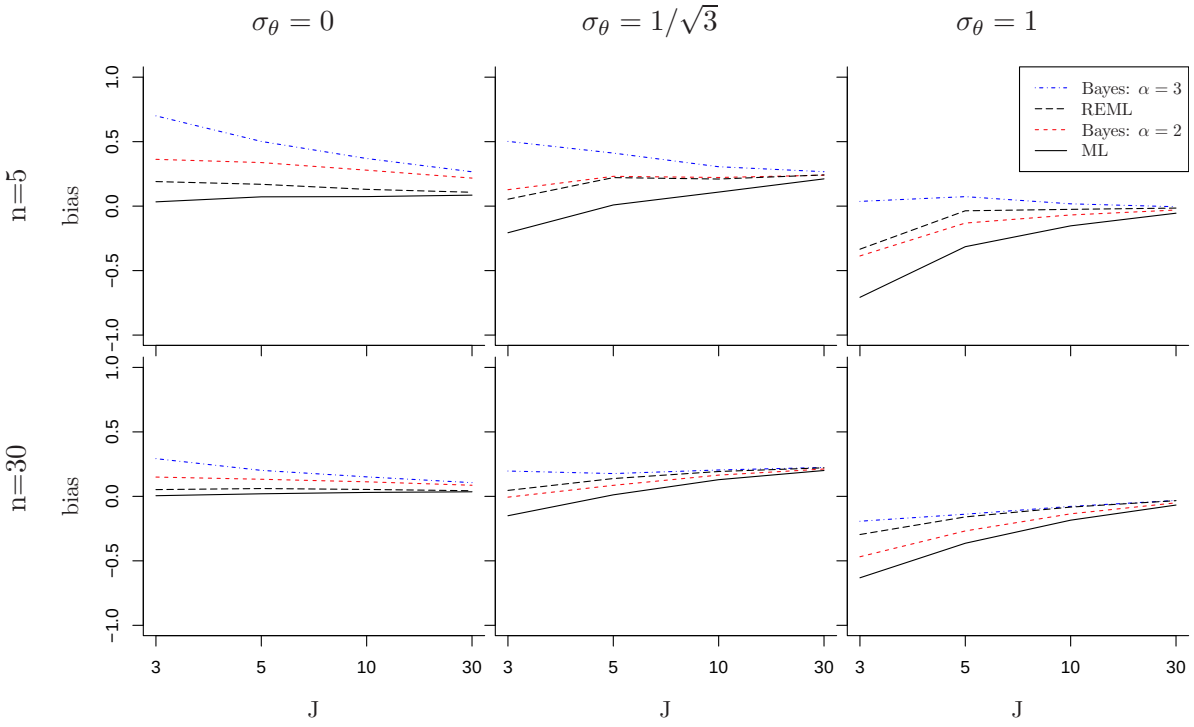


Figure 5: *Estimated bias of $\widehat{\sigma}_\theta$ for group sizes 5 and 30 (rows), standard deviations $\sigma_\theta = 0$, $1/\sqrt{3}$, and 1 (columns) and number of groups $J = 3, 5, 10, 30$ (x-axis). Different estimators are represented by different line patterns as shown in the legend in the top-right graph. When $\sigma_\theta > 0$, all estimators outperform maximum likelihood. Posterior mode with gamma$(3, \lambda)$ on σ_θ performs similarly to REML for $n = 30$ and better than REML for $\sigma_\theta = 1$.*

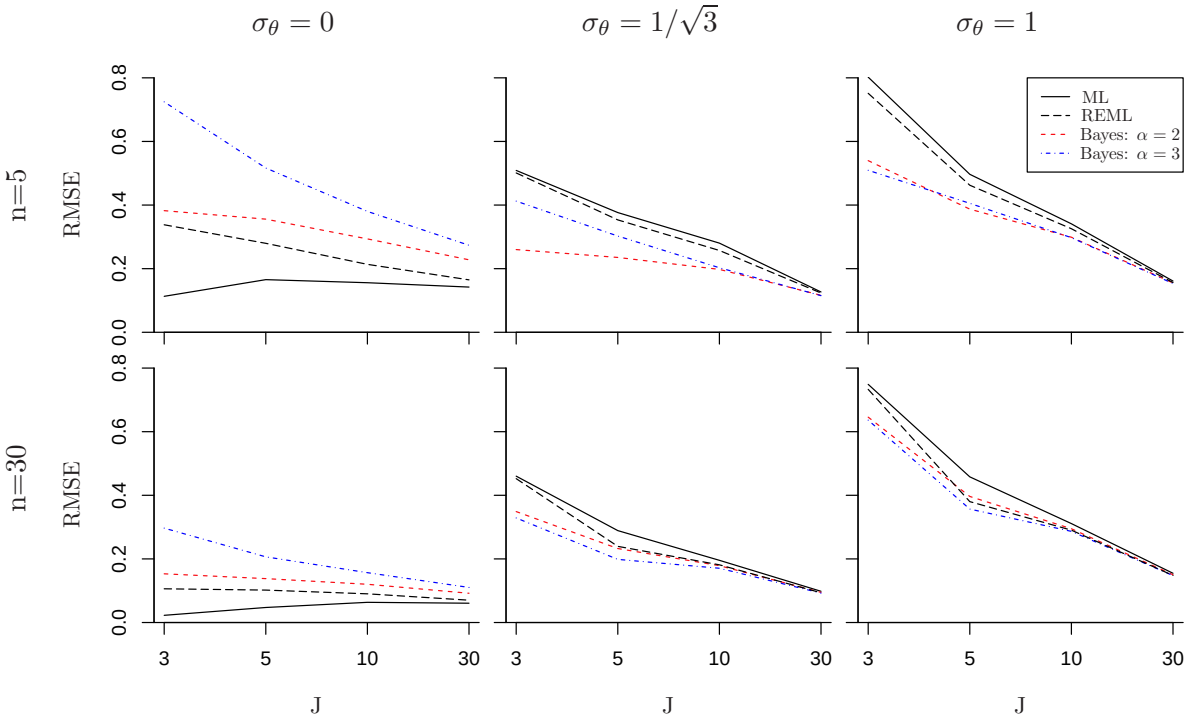


Figure 6: *RMSE of* $\widehat{\sigma}_\theta$*. For* $\sigma_\theta > 0$*, posterior mode with* $\alpha = 2$ *and* $\alpha = 3$ *performs better than maximum likelihood or REML.*

expected, the posterior mode estimator with gamma(3, λ) assigns more penalty on the values close to the boundary than REML, so the bias is larger than for REML.

Figure 6 shows the root mean squared errors (RMSE) of $\widehat{\sigma}_\theta$. When the true σ_θ is not zero, both posterior mode estimators ($\alpha = 2, 3$) have smaller RMSE than REML and maximum likelihood. For $\sigma_\theta = 1/\sqrt{3}$ and $\sigma_\theta = 1$, REML has smaller bias than the posterior mode estimator with gamma$(2, \lambda)$ but its RMSE is significantly larger because the REML estimates have the largest variance among the four estimators. The posterior mode estimator tends to have smaller RMSE with gamma$(2, \lambda)$ than with gamma$(3, \lambda)$ but the difference decreases as n, J and σ_θ increase.

Estimation of standard error of $\widehat{\beta}_2$ The estimated standard error of the estimated coefficient of the group-level covariate ($\widehat{\beta}_2$) is greatly influenced by $\widehat{\sigma}_\theta$. The squared asymptotic standard error of $\widehat{\beta}_2$ from the Hessian matrix is

$$Var(\widehat{\beta}_2) \approx \frac{n\sigma_\theta^2 + \sigma_\epsilon^2}{nJs_{X_2}^2} \tag{14}$$

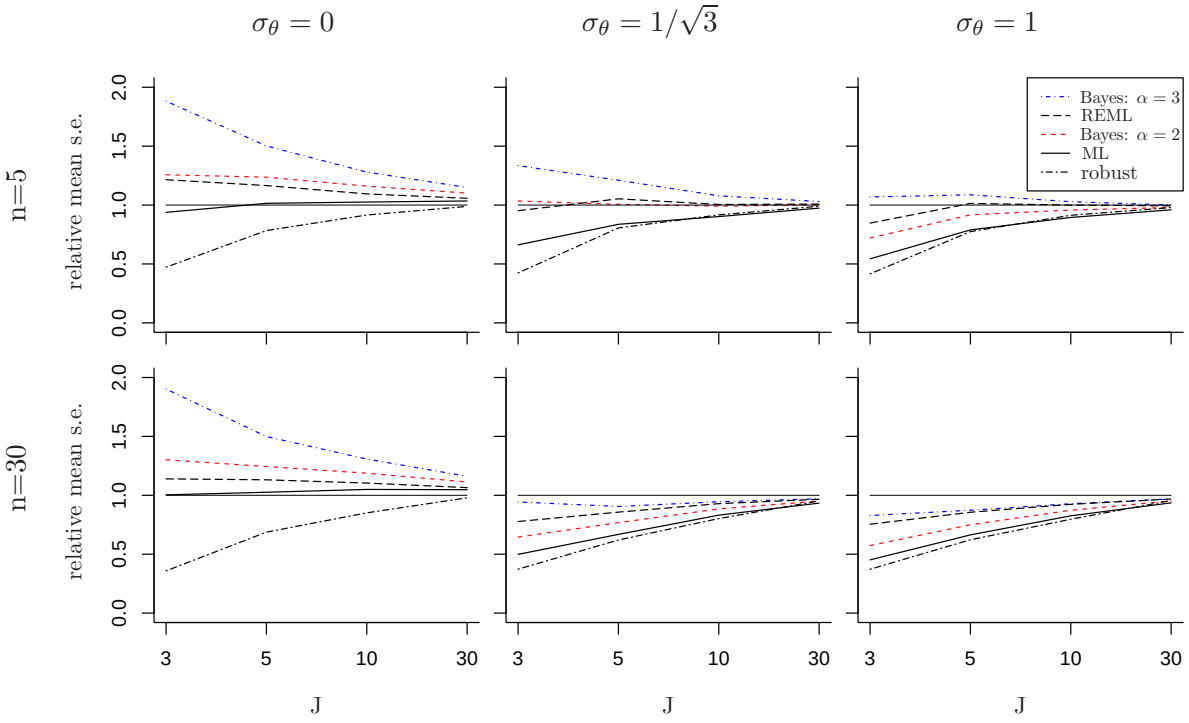


Figure 7: *Mean of* $s.e(\widehat{\beta}_2)$ *divided by the asymptotic standard error. For* $\sigma_\theta > 0$*, the posterior mode estimator performs better than the maximum likelihood estimator and is close to REML.*

where s_{X_2} is the standard deviation of the group-level covariate X_2 (Snijders and Bosker, 1993). When the true variance is not zero but $\widehat{\sigma}_\theta$ is on the boundary, (14) implies that the standard error of $\widehat{\beta}_2$ will be underestimated.

Figure 7 shows ratios of the mean estimated standard error divided by the theoretical standard error in (14). When the true variance is zero, standard errors for maximum likelihood match the theoretical standard errors well but when the true variance is not zero, maximum likelihood badly underestimates the standard errors. Especially in the case $\sigma_\theta = 1$, maximum likelihood underestimates standard errors by nearly 50% for $J = 3$, regardless of cluster size n. Compared to maximum likelihood, REML performs consistently well regardless of the true variance. When $\sigma_\theta > 0$, the posterior mode estimator with a gamma$(2, \lambda)$ prior is close to REML when $n = 5$, and the posterior mode estimator with a gamma$(3, \lambda)$ is close to REML for $n = 30$. The robust standard errors (also known as the sandwich variance estimator) are worse than any of model-based standard errors. As noted in Section 2.1, robust standard errors are known to perform poorly in small samples.

In summary, when the true σ_θ is not zero, the bias of $\widehat{\sigma}_\theta$ is as low for the posterior mode estimator with both gamma priors as for REML. The RMSE of $\widehat{\sigma}_\theta$ is lower for the posterior

mode estimator with both gamma priors than for REML and the maximum likelihood estimator. Regarding the standard error estimates for $\widehat{\beta}_2$, the posterior mode with a gamma$(2, \lambda)$ performs well for small n and the posterior mode with gamma$(3, \lambda)$ works better for large n (again when $\sigma_\theta > 0$). Although there is no obvious winner between gamma with $\alpha = 2$ and $\alpha = 3$, neither prior ever produces a boundary estimate ($\widehat{\sigma}_\theta < 10^{-5}$). Recalling that maximum likelihood and REML have quite a large proportion of boundary estimates, the posterior mode estimator with a gamma prior is successful at avoiding boundary solutions and, at the same time, the estimates are not significantly different from maximum likelihood estimates for most cases.

We also performed a simulation study for unbalanced variance component models without any covariates, following Swallow and Monahan (1984). For two different unbalanced patterns with $\sigma_\theta = 0, 1/\sqrt{3}, 1$, we compared maximum likelihood and REML estimates with posterior mode estimates with a gamma$(2,\ \lambda)$ prior, which corresponds to the REML penalty when there is no group-level covariate. (Results are in the web-based supplementary materials.)

Similar to the balanced case, when σ_θ is not zero, maximum likelihood and REML tend to underestimate σ_θ and the RMSEs tend to be larger than for the posterior mode estimates. The advantage of the gamma prior in terms of the RMSE is more obvious for $\sigma_\theta = 1$. The standard errors of the fixed intercept estimate are also underestimated by maximum likelihood and REML when σ_θ is not zero while the posterior mode estimators perform better in this regard.

6 Discussion

6.1 Multivariate extension using the Wishart distribution

In the previous sections, we discussed the properties of gamma(2,λ) as a weakly informative prior for the group-level standard deviation. In a model with varying intercepts and slopes,

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

where $\boldsymbol{\theta}_j \sim N(0, \Sigma)$ is a d-dimensional vector that varies between groups, we would like to regularize the variance-covariance matrix Σ away from its boundary, $|\Sigma| = 0$. If $|\Sigma| = 0$, at least one eigenvalue of Σ is zero. We consider gamma priors on each of the eigenvalues, say $\lambda_1, \ldots, \lambda_d$.

Then the prior on Σ can be written as

$$p(\Sigma) \quad \propto \quad \prod_{i=1}^{d} \lambda_i \exp(-\delta\lambda_i). \tag{15}$$

The Wishart density function is defined by

$$p(W|\nu, S) = \frac{|W|^{(\nu-d-1)/2} \exp[-\frac{1}{2}tr(S^{-1}W)]}{2^{\nu d/2}|S|^{n/2}\Gamma_d(\nu/2)}, \nu > d-1, S > 0 \tag{16}$$

where $\Gamma_d(\nu/2) = \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma(\nu/2 + (1-j)/2)$, ν is degrees of freedom, and S is a scale matrix with $E(W) = \nu S$. With $\nu = d+3$ and $S = \frac{1}{2\delta}I$, the right side of (15) is proportional to the Wishart density.

Therefore the Wishart$(d+3, \frac{1}{2\delta}I)$ prior will shift the posterior mode of each eigenvalue away from 0, or equivalently move the posterior mode of Σ away from the singularity. At the same time, it moves the eigenvalues approximately at most one standard error away from the maximum likelihood estimates as did the gamma(2,λ) in the univariate case.

The Wishart prior on Σ corresponds to a gamma prior on σ_θ^2. If we consider the gamma prior on σ_θ, this can be extended to the Wishart prior on $\Sigma^{\frac{1}{2}} = D\Lambda^{\frac{1}{2}}D^T$, where D is a matrix of eigenvectors of Σ and $\Lambda^{\frac{1}{2}}$ is a diagonal matrix with $\sqrt{\lambda_i}$ for the i-th diagonal element.

6.2 Other extensions

We anticipate that our approach could be applied to many extensions in the world of multilevel models, including generalized linear mixed models, models with multiple variance parameters (nested or non-nested), and latent variable models of all sorts—basically, any models in which there are variance parameters that could be estimated at zero.

Another generalization arises when there are many variance parameters—either from a large group-level covariance matrix, several different levels of variation in a multilevel model, or both. In any of these settings, it can make sense to stabilize the estimated variance parameters by modeling them together, adding another level of the hierarchy to allow partial pooling of estimated variances.

6.3 Connections to other inferential approaches

One might argue that the proclivity for the standard estimates to be degenerate is a feature, not a bug. But testing is a separate issue and we do not recommend mixing it with estimation. For the purpose of estimating coefficients and their uncertainties, we want to allow the possibility of positive σ_θ even if we cannot reject the null hypothesis that it is zero.

Finally, from a computational as well as an inferential perspective, a natural interpretation of a posterior mode is as a starting point for full Bayes inferenc, in which informative priors are specified for all parameters in the model and Metropolis or Gibbs jumping is used to capture uncertainty in the coefficients and the variance parameters (Dorie et al., 2011). For reasons discussed above, it can make sense to switch to a different class of priors when moving

to full Bayes: once modal estimation is abandoned, there is no general reason to work with priors that go to zero at the boundary.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences (R305D100017) and the National Science Foundation (SES-1023189), the Department of Energy (DE-SC0002099), and National Security Agency (H98230-10-1-0184). The opinions expressed are those of the authors and do not represent the views of the funding agencies.

References

Alderman, D. and Powers, D. (1980). The effects of special preparation on sat-verbal scores. *American Educational Research Journal* **17,** 239–251.

Bates, D. and Maechler, M. (2010). *lme4: Linear Mixed-Effects Models Using S4 Classes*. R package version 0.999375-37.

Box, G. and Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* **26,** 211–252.

Brockwell, S. E. and Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine* **20,** 825–840.

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66,** 165–185.

Crainiceanu, C., Ruppert, D., and Vogelsang, T. (2003). Some properties of likelihood ratio tests in linear mixed models. Technical report.

DerSimion, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7,** 177–188.

Dorie, V., Liu, J., and Gelman, A. (2011). Bridging between point estimation and Bayesian inference for generalized linear models. Technical report, Department of Statistics, Columbia University.

Drum, M. and McCullagh, P. (1993). [Regression Models for Discrete Longitudinal Responses]: Comment. *Statistical Science* **8,** 300–301.

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

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1,** 515–533.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis. second edition.* Champan and Hall/CRC.

Gelman, A., Shor, B., Bafumi, J., and Park, D. (2007). Rich state, poor state, red state, blue state: Whats the matter with Connecticut? *Quarterly Journal of Political Science* **2,** 345–367.

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

Huber, P. J. (1967). The behavior of maximum likelihood estimation under nonstandard condition. In LeCam, L. M. and Neyman., J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1*, pages 221–233, Berkeley. University of California Press.

Kenward, M. and Roger, J. H. (1997). Small-sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53,** 983–997.

Longford, N. T. (1993). *Random Coefficient Models.* Oxford University Press.

Longford, N. T. (2000). On estimating standard errors in multilevel analysis. *Journal of the Royal Statistical Society: Series D (The Statistician)* **49,** 389–398.

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

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

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16,** 1–32.

O'Hagan, A. (1976). On posterior joint and marginal modes. *Biometrika* **63,** 329–333.

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

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata, second ed.* Stata Press, College Station, TX.

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

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics* **6,** 377–401.

Snijders, T. and Bosker, R. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational and Behavioral Statistics* **18,** 237–259.

Swallow, W. and Monahan, J. (1984). Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics* **26,** 47–57.

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

Tsutakawa, R. K. and Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika* **51,** 251–267.

White, H. (1990). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* **48,** 817–838.