Avoiding Boundary Estimates in Linear Mixed Models Through Weakly Informative Priors

Yeojin Chung*, Sophia Rabe-Hesketh†

Andrew Gelman, Vincent Dorie, and Jinchen Liu‡

Abstract

Variance parameters in mixed or multilevel models can be difficult to estimate, especially when the number of groups is small. Here we address the problem that the group-level variance estimate is often on the boundary. We propose a maximum penalized likelihood approach which is equivalent to estimating the variance by its 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

Linear mixed models (e.g. Harville, 1977; Laird and Ware, 1982), also known as hierarchical or multilevel linear models, are widely used for longitudinal data, cross-sectional data on subjects nested in neighborhoods or institutions (hospitals, schools, firms), cluster-randomized trials, multi-site trials, and meta-analysis.

The models allow intercepts and sometimes coefficients to vary randomly between groups. However, when the number of groups is small, maximum likelihood estimates of group-level variance parameters can be noisy and can often be zero. In a multivariate setting, estimated

*Graduate School of Education, University of California, Berkeley, CA 94705, USA

†Graduate School of Education, University of California, Berkeley, CA 94705, USA, Institute of Education, University of London

‡Department of Statistics, Columbia University, New York, NY 10027, USA

covariance matrices can be degenerate non-positive-definite. In this paper we focus mostly on linear varying intercept models.

1.1 Motivation for avoiding boundary estimates

Zero group-level variance estimates can cause several problems. First, they can go against prior knowledge of researchers. Mixed models are typically used because it is known that there are processes operating at the group level that are not completely captured by the covariates. Omitted group-level covariates will lead to residual between-group variation.

A second problem with group-level variance estimates on the boundary is the resulting underestimation of uncertainty in fixed 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 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.

An argument against avoiding boundary estimates is that negative variance parameters should be permitted if the model is viewed as a marginal model for the responses given the covariates, in which case only the sum of the group-level and within-group variance must be positive (Verbeke and Molenberghs, 1997, p.52-53). However, we take a hierarchical perspective, where the intercepts vary due to omitted group-level variables, and therefore the group-level variance must be nonnegative.

In some settings, researchers are interested in testing null hypotheses that one or more variance parameters are zero. In other settings, a variance estimate at the boundary should not be viewed as non-rejection of a null hypothesis that was not of interest a-priori. Non-rejection of the null hypothesis cannot be viewed as evidence for zero variance, especially when the power of the test is low. This point is particularly important when zero variance leads to the smallest possible standard errors for parameters of interest as in meta-analysis where the practice of using tests of homogeneity as a basis for choosing between fixed and random-effects meta-analysis has been criticized (Hardy and Thompson, 1998; Curcio and Verde, 2011; Draper, 1995, p.52-53). Inclusion of varying intercepts can be viewed as a continuous model expansion (Draper, 1995) to allow for the possibility that there may be unexplained differences between groups (see also Gelman and Meng, 1996). Removing varying intercepts can then lead to an understatement of model uncertainty (Draper, 1995; Greenland, 2000).

We propose a method that pulls the variance estimates off the boundary and makes them more stable by maximizing the likelihood multiplied by a penalty function, or equivalently by assigning a prior distribution to the unknown variance parameters and finding the marginal posterior mode. 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 mixed models.

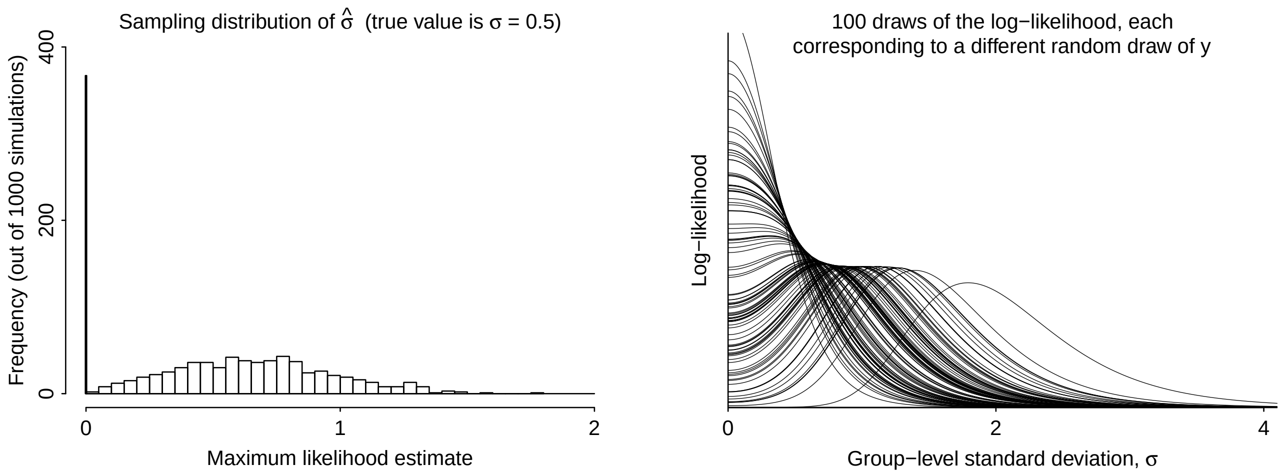
To avoid boundary estimates, we require a prior or penalty function that goes to zero at the boundary—but without requiring the sort of strong prior knowledge that would limit the routine use of this approach. We recommend a gamma prior with specific default parameter values that produce Bayes modal 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. If prior information is available, the parameters of the gamma prior can be chosen to appropriately instead of using default values.

The Bayes modal estimator with our recommended default prior has the same order and amount of asymptotic bias and variance with the maximum likelihood estimator while the bias has opposite direction. When the true group-level variance is not too close to zero, simulations show that our estimator tends to perform better than maximum likelihood and comparable to restricted maximum likelihood in terms of bias and mean squared error for the group-level standard deviation and coverage of confidence intervals for regression coefficients.

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. 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.

We begin by illustrating the boundary problem for a simple model in Section 1.3 and discuss Bayes modal estimation in Section 2. Section 3 proposes a gamma prior and consider properties of the resulting estimator. In Section 4 we apply the proposed method to a dataset and in Section 5 we perform simulations to compare performance of our method with maximum likelihood and restricted maximum likelihood in a range of situations. We end with a discussion in Section 6.



(a) (b)

Figure 1: *Simple varying intercept model with $\sigma_\theta = 0.5$ and $J = 10$ groups: (a) Sampling distribution of the maximum likelihood estimate $\widehat{\sigma}_\theta$, based on 1000 simulations of data from the model. (b) 100 simulations of the likelihood. In this example, the maximum likelihood estimate is extremely variable and the likelihood function is not very informative about σ_θ.*

1.2 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:

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

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 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 likelihood estimate of σ_θ. As expected, in more than a third of the simulations, the likelihood is maximized at $\widehat{\sigma}_\theta = 0$. Figure 1b displays 100 draws of the likelihood function, which shows in a different way that the maximum is likely to be on the boundary and that there is considerably uncertainty.

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

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

where y_{ij} is the response variable and $\boldsymbol{x}_{ij}$ is a p-dimensional vector of covariates for unit i in group j; $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients that do not vary between groups; $\theta_j \sim N(0, \sigma_\theta^2)$ 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 $(\boldsymbol{\beta}, \sigma_\theta, \sigma_\epsilon)$ are commonly estimated by maximum likelihood (ML). Another option is restricted or residualized maximum likelihood (REML, Patterson and Thompson, 1971), which is equivalent to specifying uniform priors for the regression coefficients $\boldsymbol{\beta}$ and maximizing the marginal posterior mode, integrated over θ_j and $\boldsymbol{\beta}$ (Harville, 1974). Unlike the ML estimator, the REML estimator of σ_θ^2 is unbiased in balanced designs (constant group-size) if it is allowed to be negative.

Discussion of small-sample inference for mixed models has largely focused on the covariance matrix of $\widehat{\boldsymbol{\beta}}$ (e.g., 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, $p(\boldsymbol{\beta}, \sigma_\epsilon) = 1$, on $\boldsymbol{\beta}$ and σ_ϵ.

The marginal log-posterior density with varying intercepts (θ_j) 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 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 likelihood.

Unlike posterior mean estimation, posterior modal estimation does not involve simulation and is computationally as efficient as maximum likelihood estimation. In addition, by modifying existing maximum likelihood estimation procedures, we can easily find the posterior mode. We have implemented Bayes modal estimation `gllamm` (Rabe-Hesketh et al., 2005; Rabe-Hesketh and Skrondal, 2008) in Stata and `lmer` in the `lme4` package (Bates and

Maechler, 2010) in R. In both programs, the user has the option to specify a prior and the corresponding log density is added to the log likelihood during optimization. (The modified `gllamm` is available from www.gllamm.org and the modified `lmer` can be found in the `blmer` package available from the Comprehensive R Archive Network.)

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 $p(\sigma_\theta)$ since our primary concern is to avoid 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. 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 Difference between ML estimator and Bayes modal estimator

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 profile log-likelihood can be approximated by a quadratic function in σ_θ around the ML estimator, $\hat{\sigma}_\theta^{ml}$,

$$\log L(\sigma_\theta) \approx -\frac{(\sigma_\theta - \hat{\sigma}_\theta^{ml})^2}{2 \cdot se_{ml}^2} + c_1. \tag{4}$$

Here $se_{ml} = \widehat{se}(\hat{\sigma}_\theta^{ml})$ represents the estimated asymptotic standard error of σ_θ (based on the observed information). This quadratic approximation of the profile log-likelihood function of σ_θ is reasonable because the first derivative of the profile log-likelihood (with respect to σ_θ, not σ_θ^2) at the ML estimate $\hat{\sigma}_\theta^{ml}$ is zero even when $\hat{\sigma}_\theta^{ml}$ is zero.

For example, consider a balanced random intercept model without covariates by setting $\boldsymbol{x}_{ij}^T\boldsymbol{\beta} = \mu$ and $n_i = n$ in the model (1). Then the profile log-likelood of σ_θ is given by

$$\log L_{\sigma_\theta}(\sigma_\theta) = -\frac{(n-1)J}{2}\log \hat{\sigma}_\epsilon^2 - \frac{J}{2}\log\left\{\hat{\sigma}_\epsilon^2 + n\sigma_\theta^2\right\} - \frac{1}{2}\left(\frac{SST}{\hat{\sigma}_\epsilon^2} - \frac{n\sigma_\theta^2}{\hat{\sigma}_\epsilon^2(\hat{\sigma}_\epsilon^2 + n\sigma_\theta^2)}SSB\right)$$

where

$$\hat{\sigma}_\epsilon^2 = \begin{cases} SSW/(n-1)J & \text{if } SSB \geq \frac{SSW}{n-1} \\ SST/nJ & \text{if } SSB < \frac{SSW}{n-1}, \end{cases}$$

$SST = \sum_j \sum_i (y_{ij} - \bar{y}_{\cdot\cdot})^2$, $SSB = n\sum_j (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$ and $SSW = SST - SSB$.

Taking the derivative of $\log L_{\sigma_\theta}$ with respect to σ_θ, we have

$$\frac{\partial \log L_{\sigma_\theta}}{\partial \sigma_\theta} = \left(-\frac{nJ}{2(\hat{\sigma}_\epsilon^2 + n\sigma_\theta^2)} + \frac{n \cdot SSB}{2(\hat{\sigma}_\epsilon^2 + n\sigma_\theta^2)^2} \right) \cdot 2\sigma_\theta. \tag{5}$$

When we have boundary estimates of σ_θ, it is possible that the log-likelihood function of σ_θ^2 has the maximum in the negative region, and so $\partial \log L_{\sigma_\theta} / \partial(\sigma_\theta^2)$ (the part in the parenthesis of the right-hand side in (5)) is negative at $\sigma_\theta^2 = 0$. In this case, the quadratic approximation of $\log L_{\sigma_\theta}$ in σ_θ^2 at the boundary will not be appropriate because the linear term still exists.

Even in this case, (5) will be zero because of the factor $2\sigma_\theta$. Therefore, in the Taylor expansion of $\log L_{\sigma_\theta}$ in σ_θ at 0, the linear term vanishes, the leading term becomes the quadratic (with negative coefficient when $\widehat{\sigma}_\theta = 0$) and the higher order terms are negligible around $\sigma_\theta = 0$. In Sections 4 and 5, we will confirm that the quadratic approximation fits well in an application and in simulations.

Using this quadratic approximation of the profile log-likelihood in σ_θ, we derive a number of properties of the gamma(α, λ) prior on σ_θ. (Derivations are in the supplementary materials.) In what follows, we discuss the behavior of $\hat{\sigma}_\theta$ for two cases: given under Property 1 for $\hat{\sigma}_\theta^{ml} = 0$ and Property 2 for $\hat{\sigma}_\theta^{ml} > 0$.

Property 1. *When $\hat{\sigma}_\theta^{ml} = 0$, for fixed $\alpha > 1$ and se_{ml}, the largest posterior mode is attained when $\lambda \to 0$ with the value*

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

When $\alpha = 2$, we obtain $\widehat{\sigma}_\theta = se_{ml}$. That is, when the ML 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 ML estimate. If the quadratic approximation in (4) holds and $\hat{\sigma}_\theta^{ml}$ is zero, the likelihood-ratio test

(LRT) statistic for $H_0 : \sigma_\theta = se_{ml}$ is $2(\log L(0) - \log L(se_{ml})) = 1$. For the null hypothesis $\sigma_\theta = 0$, it is known that asymptotic distribution (as J approaches infinity) of the test statistic is $0.5\chi_0^2 + 0.5\chi_1^2$ 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 $H_0 : \sigma_\theta = se_{ml}$ (> 0), the percentile will be larger because there is less point mass at zero (Crainiceanu et al., 2003). Therefore, a LRT statistic of 1 can be considered small.

Property 2. *When $\hat{\sigma}_\theta^{ml} > 0$, for fixed $\alpha > 1$ and se_{ml}, the largest possible posterior mode is attained when $\lambda \to 0$ with the value*

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

In addition, $\partial\widehat{\sigma}_\theta/\partial se_{ml}$ decreases in $\hat{\sigma}_\theta^{ml}$.

Similar to the case of $\hat{\sigma}_\theta^{ml} = 0$, $\hat{\sigma}_\theta$ is greater than $\hat{\sigma}_\theta^{ml}$ and is an increasing function of se_{ml}. The gradient $\partial\widehat{\sigma}_\theta/\partial se_{ml}$ has maximum $\sqrt{\alpha - 1}$ for $\hat{\sigma}_\theta^{ml} = 0$ that coincides with (6) and decreases as $\hat{\sigma}_\theta^{ml}$ increases. This implies that the gamma(α, λ) prior does not shift the posterior mode as much when $\hat{\sigma}_\theta^{ml} > 0$ as it does when $\hat{\sigma}_\theta^{ml} = 0$ when λ is close to zero. Therefore it has less influence on the estimate when the ML estimate is plausible than when the ML estimate is on the boundary.

3.3 Asymptotic properties

Although this paper is concerned with the problem of boundary estimates which occurs when J is small, it is important to investigate the asymptotic properties of the proposed estimator as $J \to \infty$ and compare them with the asymptotic properties of the ML estimator.

Consider a balanced random-intercept model with $\boldsymbol{x}_{ij}^T\boldsymbol{\beta} = \mu$ and $n_i = n$. For simplicity, we assume that μ and σ_ϵ^2 are known. Then the ML estimator of σ_θ is $\hat{\sigma}_\theta^{ml} = \left[\left(\sum_{j=1}^J(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2/J - \sigma_\epsilon^2/n\right)^+\right]^{1/2}$ where $(\cdot)^+ = \max(\cdot, 0)$.

When gamma(α, λ) is assigned to σ_θ, the posterior mode, say $\hat{\sigma}_\theta^{Bayes}$, is a root of a fifth order polynomial (See the supplementary materials). Therefore, we do not have simple formula for $\hat{\sigma}_\theta^{Bayes}$ but we can investigate its asymptotic properties using expansions of the log-posterior (or penalized log-likelihood) function.

The asymptotic distribution of the ML estimator in linear mixed models is shown in Miller (1977). To examine the asymptotic properties of estimator for σ_θ, it is sufficient to assume only $J \to \infty$ regardless of n. As $J \to \infty$, $\hat{\sigma}_\theta^{ml}$ is consistent with σ_θ^0 and $\sqrt{J}\left(\hat{\sigma}_\theta^{ml} - \sigma_\theta^0\right)$ follows $N(0, I(\sigma_\theta^0)^{-1})$ asymptotically where $I(\sigma_\theta^0)$ is the information matrix and σ_θ^0 is the true value of σ_θ.

Fu and Gleser (1975) show that the posterior mode is consistent and has the same limiting distribution as the ML estimator under some regularity conditions that are satisfied for our model. That is, as $J \to \infty$,

$$\sqrt{J}(\hat{\sigma}_\theta^{Bayes} - \sigma_\theta^0) \to N\left(0, I(\sigma_\theta^0)^{-1}\right).$$

Based on this result, we compare the higher order bias of the ML estimator and the Bayes modal estimator in the following theorem.

Theorem 3. *At the order of J^{-1}, the ML estimator and the Bayes modal estimator has the following bias.*

$$E(\hat{\sigma}_\theta^{ml}) = \sigma_\theta^0 - \frac{1}{4(\sigma_\theta^0)^3 J}\left(\frac{\sigma_\epsilon^2}{n} + (\sigma_\theta^0)^2\right)^2 + o(J^{-1})$$

$$E(\hat{\sigma}_\theta^{Bayes}) = \sigma_\theta^0 + \left(\frac{\alpha + \lambda\sigma_\theta^0 - 1}{2} - \frac{1}{4}\right)\frac{1}{(\sigma_\theta^0)^3 J}\left(\frac{\sigma_\epsilon^2}{n} + (\sigma_\theta^0)^2\right)^2 + o(J^{-1})$$

In addition, with the default prior ($\alpha = 2$ and $\lambda \to 0$), two estimators have the same *magnitude of bias but negative for* $\hat{\sigma}_\theta^{ML}$ *and positive for* $\hat{\sigma}_\theta^{Bayes}$.

Proof. Outline of the proof is in the supplementary materials and **(cite Dorie's dissertation)**. □

Not only $\hat{\sigma}_\theta^{Bayes}$ with the default prior is asymptotically unbiased and efficient as $\hat{\sigma}_\theta^{ml}$, but also they have the same magnitude of the bias at the higher order as seen in Theorem 3. Therefore, when J is large, the Bayes modal estimator will be as good as the ML estimator in addition that the Bayes modal estimator tend to be less biased for small J as will be shown in Section 5.

3.4 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. Note that while the ML estimator is invariant under transformations, the Bayes modal estimator is not due to the change in prior density when transforming σ_θ. Thus the transformation affects the posterior mode.

Consider the Box-Cox transformations (Box and Cox, 1964)

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

Property 4. *With a gamma*(α, λ) *prior on* σ_θ*, maximizing the posterior of* $g_\gamma(\sigma_\theta)$ *is equivalent to maximizing the posterior of* σ_θ *with a gamma*$(\alpha + 1 - \gamma, \lambda)$ *prior on* σ_θ.

For example, consider a special case with $\alpha = 1$, $\lambda \to 0$, and $\gamma = 0$, which implies the uniform (improper) prior on σ_θ and log transformation of σ_θ. With this prior, the marginal posterior density is just the 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. If we maximize the posterior density of $\log(\sigma_\theta)$, then the maximizer $\widehat{\log(\sigma_\theta)}$ will be the same as $\log(\hat{\sigma}_\theta)$ where $\hat{\sigma}_\theta$ is the maximizer of the posterior with gamma$(2, \lambda)$ prior on σ_θ.

We have discussed the gamma prior on the group-level standard deviation (σ_θ) since the profile log-likelihood as a function of σ_θ has a better quadratic approximation so it helps us to investigate the properties in Section 3.2. However, one might be still insterested in priors on the variance, σ_θ^2.

Property 5. *In the limit $\lambda \to 0$, a gamma(α, λ) prior on σ_θ^2 is equivalent to a gamma$(2\alpha - 1, \lambda)$ prior on σ_θ.*

Therefore, the properties of the gamma prior in this paper hold for the gamma prior on σ_θ^2 with α adjusted appropriately.

3.5 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, 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 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.

The difference between these two penalty terms is clear when σ_θ is close to zero. At $\sigma_\theta = 0$, the log of the gamma prior in (9) is $-\infty$ for $\alpha > 1$, whereas the REML penalty in (8) 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 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.

The REML penalty expression in (8) is derived for covariates with specific properties as described above. However, we found that the relationship between the REML and gamma penalty illustrated in this section holds more generally (see the supplementary materials.)

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 consist of an estimated treatment effect and associated standard error for each school (obtained by separate analyses of the data of each school) and have previously been analyzed by Rubin (1981) and Gelman et al. (2004).

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 measures of 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 estimated standard error for study i.

Figure 2 shows the profile log-likelihood (maximized with respect to μ) of σ_θ (left) and σ_θ^2 (right). On the left we see that the profile log-likelihood has its maximum at zero where the gradient is zero as discussed in Section 3.2. Further, the profile log-likelihood is quite flat.

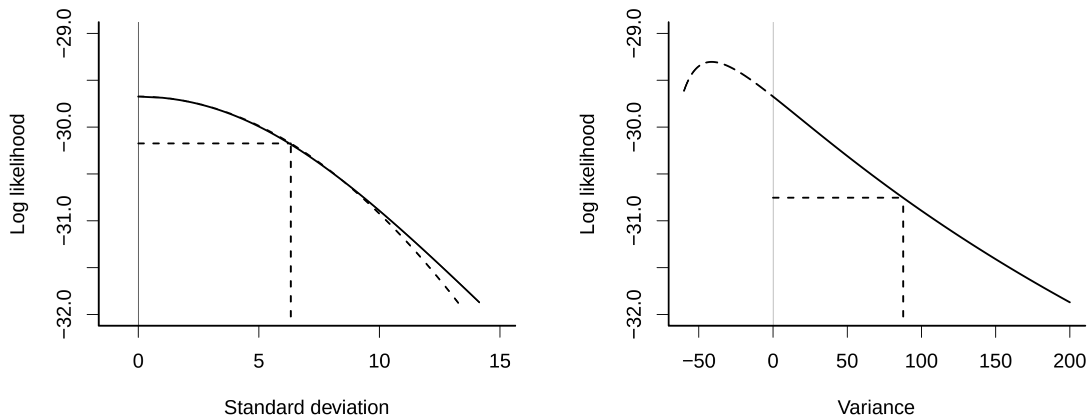


Figure 2: *Profile log likelihood as a function of σ_θ (left) and σ_θ^2 (right) for 8-schools data. The dashed curve on the left is the quadratic approximation at the mode, based on the estimated standard error. The vertical dashed line is the Bayes modal estimate for a gamma(2,λ) prior on σ_θ (left) 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 maximum is attained for a negative σ_θ^2, so the quadratic approximation at the ML estimate of zero is poor.*

We see in the right panel of Figure 2 that the profile log-likelihood has a negative gradient at zero as a function of σ_θ^2 so that the quadratic approximation for σ_θ^2 is poor at the maximum likelihood estimate of zero.

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.)

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

Prior Method	α	μ Est	μ SE	μ SE^R	σ_θ Est	σ_θ SE	σ_θ^2 Est	σ_θ^2 SE	Log-lik
ML		7.69	4.07	3.33	0	6.32	0	0.00	−29.67
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

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 $se_{ml}\sqrt{1-\alpha}$ and agree well with the posterior mode estimates with gamma$((\alpha+1)/2, \lambda)$ on σ_θ^2.

with these priors and maximum likelihood estimates are given in Table 1. The estimated standard error of $\hat{\sigma}_\theta^{ml}$ is 6.32 (which corresponds to se_{ml} in Section 3.2).

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

Specifying a gamma$(2, \lambda)$ prior on σ_θ^2 (row 5) gives estimates that agree well with those for a gamma$(3, \lambda)$ prior on σ_θ as expected (see Property 5). Similarly, a gamma$(1.5, \lambda)$ prior on σ_θ^2 (row 4) 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 decreases the log-likelihood by only 0.5.

Table 1 also reports model-based and robust standard error estimates for $\widehat{\mu}$ (SE^R). 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.

5 Simulation of balanced varying-intercept model

We consider a 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 1000 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. Although our method is based on the assumption that $\sigma_\theta > 0$, we include the condition $\sigma_\theta = 0$ as the worst-case scenario. We obtain posterior mode estimates with gamma(2, λ) and gamma(3, λ) priors on σ_θ, where $\lambda = 10^{-4}$. The REML penalty corresponds to $\alpha = 3$ since the model contains one group-level covariate. We compare posterior mode estimates with ML 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 $\sigma_\theta = 1/\sqrt{3}$, 47% of ML estimates and 45% 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 5% for ML and 4% for REML.

When $\sigma_\theta = 1$, the same pattern occurs but estimates are on the boundary less often for

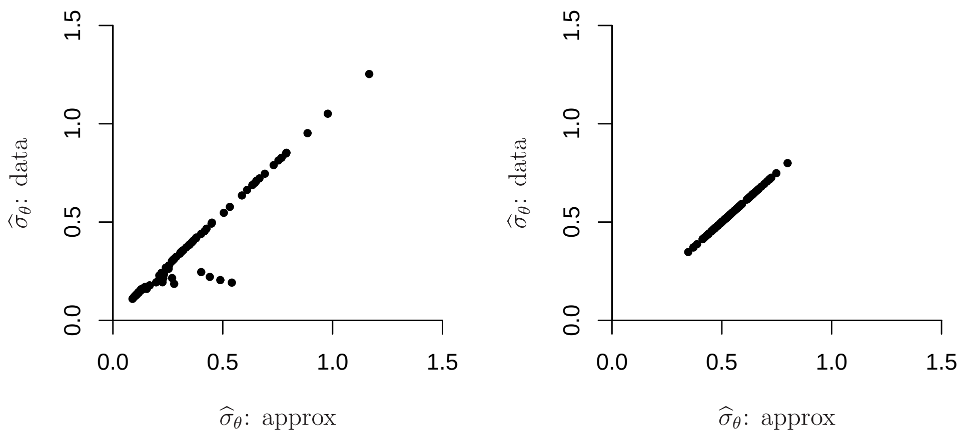


Figure 3: *Posterior mode estimates with a gamma(2,λ) prior on* σ_θ *for* $\rho = 0.25$ *and* $n = 30$ *for the first 100 replicates, compared with the posterior mode based on the quadratic approximation of the profile likelihood (see properties 1 and 2). Agreement is good, suggesting that the quadratic approximation is good. Dots on the left graph that fall off the line are due to a few samples that have uncommonly large estimated standard errors.*

a given condition. For $J = 3$ and $n = 5$, ML produces 34% of estimates on the boundary compared with 32% for REML. When J increases to 5 and n to 30, 1% of ML estimates and 0.7% of REML estimates are on the boundary. When $J = 30$, ML and REML yield no boundary estimates for either value of σ_θ.

In contrast to the ML and REML estimates, the Bayes modal estimates are never on the boundary in any of the simulation conditions. At the same time, the Bayes modal estimates do not differ significantly from the ML estimates. The likelihood ratio test statistic $-2\left[\log L(\widehat{\sigma}_\theta^{Bayes}) - \log L(\widehat{\sigma}_\theta^{ml})\right]$ for testing the restriction $\sigma_\theta = \widehat{\sigma}_\theta^{Bayes}$ was calculated for each replicate. When $J > 3$, the largest test statistic among all the replicates and simulation conditions is 2.60. Even for $J = 3$, the largest test statistic is 3.45. 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 3.2 by assuming that the profile log likelihood is quadratic. Figure 3 shows that the posterior modes calculated by the quadratic approximation of the profile log-likelihood (see properties 1 and 2) 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 $n = 30$.

Figure 4 summarizes the estimated bias and the root mean squared error (RMSE) of σ_θ, and the coverage of 95% confidence intervals (CI) for β_2 for the four methods for $n = 5$, $J = 3, 5, 10, 30$ and $\sigma_\theta = 0, 1/\sqrt{3}, 1$. Results for $n = 30$ are given in the supplementary materials.

Estimates of $\hat{\sigma}_\theta$ The first row of Figure 4 shows that the bias for σ_θ decreases as J increases and σ_θ decreases. Thus the differences between methods are most obvious with small J, and particularly when $\sigma_\theta > 0$.

For $\sigma_\theta > 0$, both REML and ML tend to underestimate σ_θ. Bayes modal estimates with gamma$(2, \lambda)$ also tend to be downward biased for σ_θ but not as much as the ML estimates. On the other hand, the Bayes modal estimator with gamma$(3, \lambda)$ produces the largest estimates among the four estimators so it often overestimates σ_θ. For $\sigma_\theta = 1$, the Bayes modal estimator with gamma$(3, \lambda)$ has the smallest bias for all J.

When $\sigma_\theta = 0$, as expected, the Bayes modal estimators assign more penalty on the values close to the boundary than REML, so the bias is larger than for REML as well as ML.

When $n = 30$ (given in the supplementary materials), the overall pattern is the same as when $n = 5$ but the Bayes modal estimates with gamma$(3, \lambda)$ are closer to REML for $\sigma_\theta > 0$. This confirms that the gamma penalty on σ_θ with $\alpha = 3$ agrees with the REML penalty when the model contains one group-level covariate, particularly with large n, as discussed in Section 3.5.

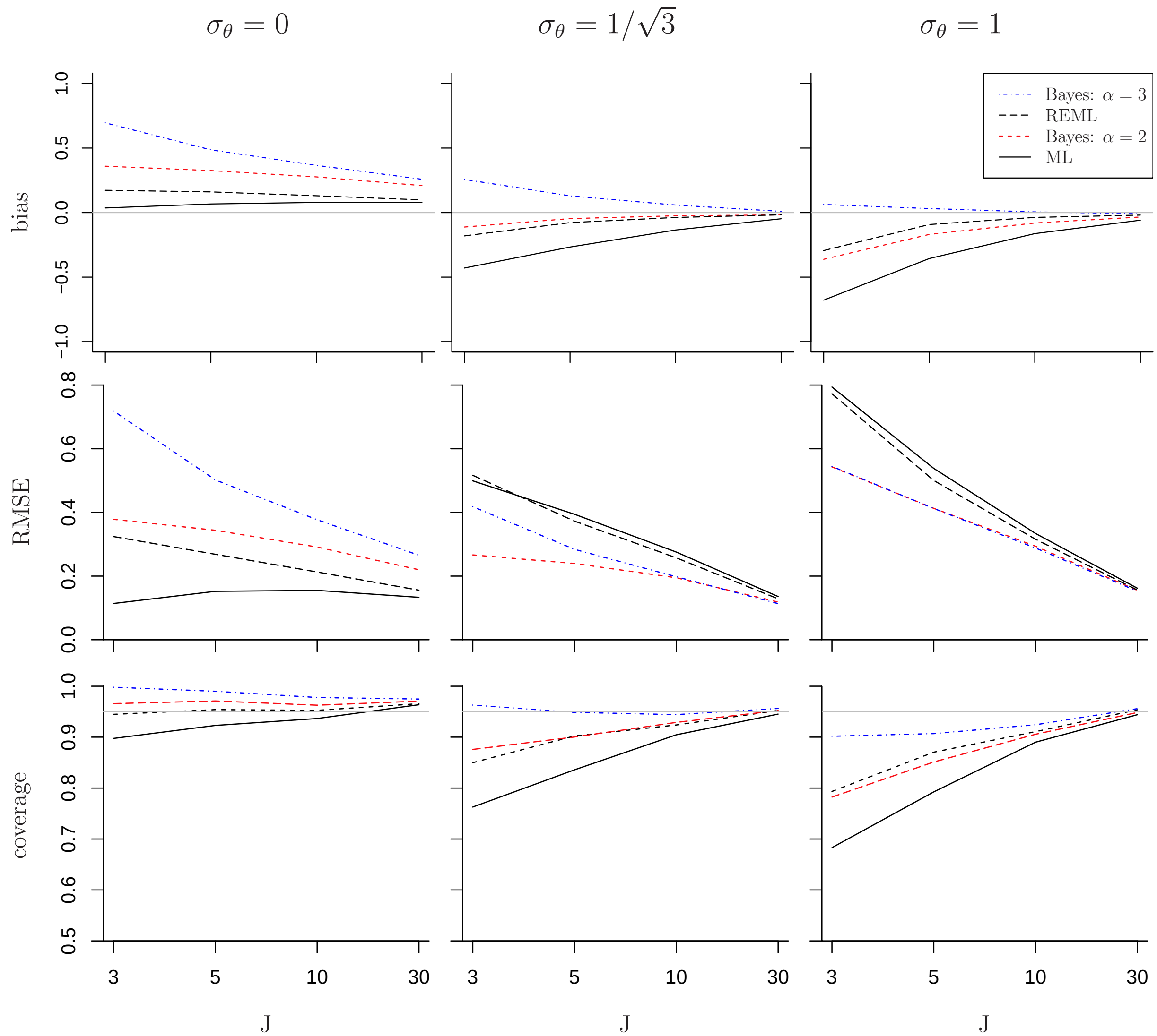


Figure 4: *Bias of* σ_θ *, RMSE of* σ_θ *and coverage of CI for* β_2 *for group size 5, standard deviation* $\sigma_\theta = 0, 1/\sqrt{3}$*, and 1 (columns) and number of groups* $J = 3, 5, 10, 30$ *(x-axis). Different estimators are represented by different line patterns. When* $\sigma_\theta > 0$*, all the methods outperform ML. Bias of the Bayes modal estimator is as low as REML depending on* α*. RMSE of the Bayes modal estimator with both* α *is smaller than REML and ML. Coverage of CI is best for the Bayes modal estimator with* $\alpha = 3$*.*

The root mean squared errors (RMSE) of both Bayes modal estimators are consistently smaller than for ML and REML when σ_θ is not zero (see second row of the figure). For $\sigma_\theta = 1/\sqrt{3}$ and $\sigma_\theta = 1$, REML has smaller bias than the Bayes modal estimator with gamma$(2, \lambda)$ but its RMSE is significantly larger because the REML estimates have the largest variance among the four estimators. The Bayes modal 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.

Coverage of CI for β_2 The standard error estimates 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 $\text{Var}(\widehat{\beta}_2) \approx (n\sigma_\theta^2 + \sigma_\epsilon^2)/nJs_{X_2}^2$ 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, the standard error of $\widehat{\beta}_2$ will be underestimated and the CI will be too narrow.

The third row of Figure 4 shows the proportions of 95% CI that cover the true value of β_2. The gray solid line shows the nominal coverage (0.95). For all values of σ_θ, ML gives CI with lower than nominal coverage. For $\sigma_\theta = 0$, all the methods except ML tend to have higher than nominal coverage.

When $\sigma_\theta > 0$, most of the methods have lower than nominal coverage, but the Bayes modal estimator with $\alpha = 3$ has the best coverage, particularly for $\sigma_\theta = 1/\sqrt{3}$. Although the Bayes modal estimator with $\alpha = 3$ tends to have large positive bias for σ_θ, it turns out to give better coverage. Recalling that gamma$(3, \lambda)$ is close to the REML penalty (discussed in Section 3.5) for large n, we found that the coverage for the Bayes modal estimator with $\alpha = 3$ is closer to REML for $n = 30$ (not shown here) than for $n = 5$. However, REML still shows significantly lower coverage than the Bayes modal estimator, particularly for small J.

In summary, when the true σ_θ is not zero, the bias of $\widehat{\sigma}_\theta$ is as low for the Bayes modal estimators as for REML depending on α. The RMSE of $\widehat{\sigma}_\theta$ is uniformly lower for the Bayes modal estimator with both gamma priors than for REML and the ML estimator. Coverage of the CI is best for the Bayes modal estimator with $\alpha = 3$. 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 ML and REML have quite a large proportion of boundary estimates, the Bayes modal estimator with a gamma prior is successful at avoiding boundary solutions and, at the same time, the estimates are not significantly different from ML 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 ML and REML estimates with posterior mode estimates with a gamma(2, λ) prior, which corresponds to the REML penalty when there is no group-level covariate. (Results are in the supplementary materials.)

Similar to the balanced case, when σ_θ is not zero, ML and REML tend to underestimate σ_θ and the RMSE tends 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 ML and REML when σ_θ is not zero while the posterior mode estimators perform better in this regard.

6 Discussion

In this paper, we considered linear varying intercept models and suggested specifying a gamma prior for the group-level standard deviation to avoid boundary estimates. We showed that our procedure guarantees non-zero estimates of the group-level variance, while maintaining statistical properties as good or better than maximum likelihood and restricted maximum likelihood when the true group-level variance is not too close to zero. The prior is only weakly informative in the sense that the log likelihood at the Bayes modal estimates is not much lower than the maximum.

As mentioned in the introduction, we acknowledge that there are situations where tests for zero variance correspond to a research question or where negative variances cannot be ruled out, and our method is clearly not appropriate in those situations. In other situations, it can however be problematic to react to boundary estimates by proceeding as if the true group-level variance is zero since this will often lead to the smallest possible estimated standard error for parameters of interest. We have shown that this strategy of accepting the maximum likelihood estimate results in under-coverage of confidence intervals for regression coefficients of group-level covariates. In datasets where boundary estimates occur, large range of values of the group-level standard deviation is often supported by the data, and our method provides one such value. Our approach is hence somewhere between purely data-based maximum likelihood estimation and setting the variance to a constant instead of estimating it, as suggested by Longford (2000) for the purpose of obtaining better standard errors and by Greenland (2000) when the variance is not identified.

Our idea can also be applied to models with varying intercepts and slopes where the problem is to regularize the covariance matrix, say Σ, away from its boundary, $|\Sigma| = 0$. In this case, the gamma prior can be naturally extended to the Wishart prior on Σ, which is equivalent to the product of gamma priors on the eigenvalues of $\Sigma^{1/2}$. Therefore the Wishart prior with a certain choice of parameters 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 ML estimates as did the gamma$(2,\lambda)$ in the univariate case.

Other applications of our approach include generalized linear mixed models, models with more hierarchical levels, 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.

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 inference, 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 (grant 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. “The effects of special preparation on SAT-Verbal scores.” *American Educational Research Journal*, 17(2):239–251 (1980).

Bates, D. and Maechler, M. *lme4: linear mixed-effects models using S4 classes* (2010). R

package version 0.999375-37.
URL `http://CRAN.R-project.org/package=lme4`

Box, G. and Cox, D. "An analysis of transformations." *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252 (1964).

Crainiceanu, C. and Ruppert, D. "Likelihood ratio tests in linear mixed models with one variance component." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):165–185 (2004).

Crainiceanu, C., Ruppert, D., and Vogelsang, T. "Some properties of likelihood ratio tests in linear mixed models." Technical report, available at http://www.orie.cornell.edu/∼davidr/papers (2003).

Curcio, D. and Verde, P. "Comment on: Efficacy and safety of tigecycline: a systematic review and meta-analysis." *Journal of Antimicrobial Chemotherapy Advanced Access, doi: 10.1093/jac/dkr353* (2011).

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

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

Draper, D. "Assessment and propagation of model uncertainty." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):45–97 (1995).

Drum, M. and McCullagh, P. "[Regression models for discrete longitudinal responses]: Comment." *Statistical Science*, 8(3):300–301 (1993).

Fu, J. and Gleser, L. "Classical asymptotic properties of a certain estimator related to the maximum likelihood estimator." *Annals of the Institute of Statistical Mathematics*, 27(1):213–233 (1975).

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

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

Gelman, A. "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1(3):515–533 (2006).

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

Gelman, A. and Meng, X. "Model checking and model improvement." In *Markov Chain Monte Carlo in Practice*, 189–201. Chapman and Hall (1996).

Greenland, S. "When should epidemiologic regressions use random coefficients?" *Biometrics*, 56(3):915–921 (2000).

Hardy, R. and Thompson, S. "Detecting and describing heterogeneity in meta-analysis." *Statistics in Medicine*, 17(8):841–856 (1998).

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

—. "Maximum likelihood approaches to variance components estimation and related problems." *Journal of the American Statistical Association*, 72:320–340 (1977).

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

Kenward, M. and Roger, J. H. “Small-sample inference for fixed effects from restricted maximum likelihood.” *Biometrics*, 53(3):983–997 (1997).

Laird, N. M. and Ware, J. H. “Random effects models for longitudinal data.” *Biometrics*, 38:963–974 (1982).

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

—. “On estimating standard errors in multilevel analysis.” *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):389–398 (2000).

Maris, E. “Estimating multiple classification latent class models.” *Psychometrika*, 64(2):187–212 (1999).

Miller, J. “Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance.” *The Annals of Statistics*, 5(4):746–762 (1977).

Mislevy, R. J. “Bayes modal estimation in item response models.” *Psychometrika*, 51(2):177–195 (1986).

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

O’Hagan, A. “On posterior joint and marginal modes.” *Biometrika*, 63(2):329–333 (1976).

Patterson, H. D. and Thompson, R. “Recovery of inter-block information when block sizes are unequal.” *Biometrika*, 58(3):545–554 (1971).

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

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

Rubin, D. B. "Estimation in parallel randomized experiments." *Journal of Educational Statistics*, 6(4):377–401 (1981).

Snijders, T. and Bosker, R. "Standard errors and sample sizes for two-level research." *Journal of Educational and Behavioral Statistics*, 18(3):237–259 (1993).

Swallow, W. and Monahan, J. "Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components." *Technometrics*, 26(1):47–57 (1984).

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

Tsutakawa, R. K. and Lin, H. Y. "Bayesian estimation of item response curves." *Psychometrika*, 51(2):251–267 (1986).

Verbeke, G. and Molenberghs, G. *Linear Mixed Models in Practice : A SAS-Oriented Approach*. Springer (1997).

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