PSYCHOMETRIKA
2013
DOI: 10.1007/s11336-013-9328-2

A NONDEGENERATE PENALIZED LIKELIHOOD ESTIMATOR FOR VARIANCE PARAMETERS IN MULTILEVEL MODELS

YEOJIN CHUNG

SCHOOL OF BUSINESS ADMINISTRATION, KOOKMIN UNIVERSITY

SOPHIA RABE-HESKETH

GRADUATE SCHOOL OF EDUCATION, UNIVERSITY OF CALIFORNIA, BERKELEY

INSTITUTE OF EDUCATION, UNIVERSITY OF LONDON

VINCENT DORIE, ANDREW GELMAN, AND JINGCHEN LIU

DEPARTMENT OF STATISTICS, COLUMBIA UNIVERSITY

Group-level variance estimates of zero often arise when fitting multilevel or hierarchical linear models, especially when the number of groups is small. For situations where zero variances are implausible a priori, we propose a maximum penalized likelihood approach to avoid such boundary estimates. This approach is equivalent to estimating variance parameters by their posterior mode, given a weakly informative prior distribution. By choosing the penalty from the log-gamma family with shape parameter greater than 1, we ensure that the estimated variance will be positive. We suggest a default log-gamma$(2, \lambda)$ penalty with $\lambda \to 0$, which ensures that the maximum penalized likelihood estimate is approximately one standard error from zero when the maximum likelihood estimate is zero, thus remaining consistent with the data while being nondegenerate. We also show that the maximum penalized likelihood estimator with this default penalty is a good approximation to the posterior median obtained under a noninformative prior.

Our default method provides better estimates of model parameters and standard errors than the maximum likelihood or the restricted maximum likelihood estimators. The log-gamma family can also be used to convey substantive prior information. In either case—pure penalization or prior information—our recommended procedure gives nondegenerate estimates and in the limit coincides with maximum likelihood as the number of groups increases.

Key words: Bayes modal estimation, hierarchical linear model, mixed model, multilevel model, penalized likelihood, variance estimation, weakly informative prior.

1. Introduction

Linear mixed models (e.g., Harville, 1977; Laird & 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, multisite trials, and meta-analysis. The models include random intercepts and sometimes random coefficients that vary among groups and that we will refer to as varying intercepts and coefficients. We consider the situation where some unexplained group-level variability is known to exist a priori. Maximum likelihood (ML) is a useful way to estimate the variance parameters. But when the number of groups is small, estimates of the group-level variance can be noisy and can often be zero.

Requests for reprints should be sent to Yeojin Chung, School of Business Administration, Kookmin University, Seoul, South Korea. E-mail: jini.y.chung@gmail.com

Published online: 12 March 2013

Zero group-level variance estimates can cause several problems. Zero variance can go against prior knowledge of researchers and results in underestimation of uncertainty in fixed coefficient estimates. Inferences for groups 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. In addition, uncertainty in predictions for new and existing groups is also understated. In Section 2.1, we further discuss problems caused by the boundary estimate.

In this paper, we develop a nondegenerate estimator by maximizing the log-likelihood plus a penalty function, or equivalently by assigning a prior distribution to the unknown variance parameters and finding the posterior mode. It is possible to do this without requiring strong prior knowledge. But our functional form is general enough that it can also be applied when real prior information is available.

Penalized likelihood or Bayes modal estimation has been used to obtain more stable estimates in item response theory (Swaminathan & Gifford, 1985; Mislevy, 1986; Tsutakawa & Lin, 1986) and to avoid boundary estimates (or logit parameters tending to $\pm\infty$) in log-linear models (Galindo-Garre, Vermunt, & Bergsma, 2004), logistic regression (Gelman, Jakulin, Pittau, & Su, 2008), and latent class analysis (Maris, 1999; Galindo-Garre & Vermunt,2006). Such an approach has also been used to obtain nondegenerate covariance matrices in finite mixtures of normal densities (Ciuperca, Ridolfi, & Idier, 2003; Vermunt & Magidson, 2005) and in multivariate regression (Warton, 2008). Our penalized likelihood approach to avoid boundary estimates for variance parameters in multilevel models turns out to be similar to, but more general than, the independently developed adjustment for density maximization approach by Morris and Tang (2011). They apply the idea to the Fay and Herriot (1979) model for small-area estimation using area-level (group-level) data and focus on the problem of predicting area-level means (see also Li & Lahiri, 2010). In contrast, we consider unit-level data and focus on estimation of the model parameters and standard errors of regression coefficients. By adjusting two parameters of the log-gamma penalty, our method can take into account prior knowledge about the group-level variance.

We recommend a class of log-gamma penalties (or gamma priors) that in our default setting (the log-gamma(2, λ) penalty with $\lambda \to 0$) produce maximum penalized likelihood (MPL) estimates (or 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. The penalty has little influence when the number of groups is large or when the data are informative about the variance, and the asymptotic mean squared error of the proposed estimator is the same as that of the maximum likelihood estimator. The default penalization can also be interpreted as equivalent to specifying a noninformative prior for the group-level standard deviation and applying a log transformation to this parameter to symmetrize the posterior distribution. We compare the bias and mean squared error of our estimator to maximum likelihood and restricted maximum likelihood in simulations across a wide range of conditions. Our method performs well and also provides better estimates of standard errors of regression coefficients.

Compared with full Bayes or posterior mean (or median) 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 additional convergence checking is required and there is no need to specify priors for all model parameters.

We have implemented penalized likelihood 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 the parameters of the penalty function, the programs automatically find the maximum penalized likelihood estimate of the variance parameter and provide inferences for the coefficients conditional on that estimate.

In Section 2, we discuss our motivation for avoiding boundary estimates for group-level variance parameters. In Section 3, we introduce our maximum penalized likelihood approach and our recommended default penalty function or weakly informative prior distribution. Section 4 shows theoretical properties of the resulting estimator. In Section 5, we apply the proposed method to a dataset, and in Section 6 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 7.

2. Motivation for Avoiding Boundary Estimates

2.1. Problems with Boundary Estimates

When a variance parameter is estimated as zero, there is typically a large amount of uncertainty about this variance. One possibility is to declare in such situations that not enough information is available to estimate a multilevel model. However, the available alternatives can be unappealing since, as noted in the introduction, discarding a variance component or setting the variance to zero understates the uncertainty. In particular, standard errors for coefficients of covariates that vary between groups will be too low as we will see in Section 2.2. The other extreme is to fit a regression with indicators for groups (a fixed-effects model), but this will overcorrect for group effects (it is mathematically equivalent to a mixed-effects model with variance set to infinity), and also does not allow predictions for new groups.

Degenerate variance estimates lead to complete shrinkage of predictions for new and existing groups and yield estimated prediction standard errors that understate uncertainty. This problem has been pointed out by Li and Lahiri (2010) and Morris and Tang (2011) in small area estimation.

Here is an example. Using multilevel modeling of data on US voters' choice of candidates, Gelman, Shor, Bafumi, and Park (2007) found that richer voters tended to support Republican candidates but with a slope that varied depending on some state-level covariates. For some models fit to some elections, the estimated variance of the residuals for the state-level slopes was zero. In the resulting inferences, the slopes were perfectly predicted by the state-level covariates. 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 or attempt to directly interpret 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, as found by Bell (1999) when estimating annual poverty rates of school-aged children for the US states using data from the Current Population Survey.

If zero variance is not a null hypothesis of interest, a boundary estimate, and the corresponding zero likelihood ratio test statistic, should not necessarily lead us to accept the null hypothesis and to proceed as if the true variance is zero. This point is particularly important when zero variance leads to the smallest possible standard errors for parameters of interest as in random-effects 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 (e.g., Hardy & Thompson, 1998; Borenstein, Hedges, Higgins, & Rothstein, 2009; Curcio & Verde, 2011; Draper, 1995, pp. 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 & Meng, 1996).

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 & Molenberghs, 2000, pp. 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.

2.2. *Example: Meta-analysis*

Two important examples where the number of groups is often small and the estimate of the group-level variance affects the standard errors for the coefficients of interest are meta-analyses and cluster-randomized trials. Here, we briefly consider a meta-analysis example that we return to in Section 5 where we also analyze data from a cluster-randomized trial.

A classic example where the maximum likelihood estimate of the group-level variance is zero is a meta-analysis of randomized experiments of coaching for the Scholastic Aptitude Test (SAT) conducted in eight schools (Alderman & Powers, 1980; Rubin, 1981; Gelman, Carlin, Stern, & Rubin, 2004). 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.

Meta-analysis with varying intercepts (DerSimonian & 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

$$y_i = \mu + \theta_i + \epsilon_i, \quad \theta_i \sim N\big(0, \sigma_\theta^2\big),\ \epsilon_i \sim N\big(0, s_i^2\big), \tag{1}$$

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

The ML estimate of σ_θ is 0, which implies that the treatment effect μ is the same for all schools, and that the studies conducted in the different schools are "functionally equivalent" (Borenstein et al., 2009, p. 83) in terms of populations, interventions, and measures of outcomes. Because studies conducted by different researchers in different settings are usually known a priori to be heterogeneous, there has been much criticism of the practice of testing the null hypothesis of homogeneity and proceeding as if the variance is zero when the null hypothesis is not rejected (e.g., Borenstein et al., 2009; Draper, 1995; Hardy & Thompson, 1998; Overton, 1998; Viechtbauer, 2005). For instance, Higgins, Thompson, and Spiegelhalter (2009, p. 149) argue that "such a null hypothesis is usually untenable." Here, we propose not to proceed as if the variance is zero when its point estimate is zero, if the data are consistent with larger values of the variance (see also Curcio & Verde, 2011).

The range of values of σ_θ that is supported by the data can be assessed by considering the estimated standard error of the estimate of σ_θ and by plotting the profile log likelihood of σ_θ (maximized over μ). For the 8-schools data, the estimated standard error of 6.32 is substantial and the profile log-likelihood (the left plot in Figure 1) is quite flat, showing that large values of σ_θ (e.g., $\sigma_\theta = 6.30$) are supported.

Inference for σ_θ is important because it affects both the point estimate $\hat{\mu}$ of the overall effect size and its estimated standard error, $\widehat{\text{se}}(\hat{\mu}) = [\sum_i (s_i^2 + \hat{\sigma}_\theta^2)^{-1}]^{-1/2}$, which increases with $\hat{\sigma}_\theta$. For example, the estimated standard error is 4.1 for $\hat{\sigma}_\theta = 0$, compared with 5.5 for $\hat{\sigma}_\theta = 10$ (the corresponding estimates of μ are 7.7 and 8.1, respectively).

To compare study-specific effects, we can predict θ_i using the empirical Bayes predictor, $\tilde{\theta}_i = (1 - \hat{B}_i) y_i + \hat{B}_i \hat{\mu}$ where $\hat{B}_i = s_i^2/(\hat{\sigma}_\theta^2 + s_i^2)$ (e.g., Raudenbush & Bryk, 1985). When σ_θ is estimated as zero, all the studies have the same predicted value $\hat{\mu}$. The right panel of Figure 1 shows that predictions change rapidly with increasing $\hat{\sigma}_\theta$. The widths of the empirical Bayes prediction intervals also increase with increasing $\hat{\sigma}_\theta$, so that the uncertainty of the predictions is understated whenever σ_θ is underestimated.

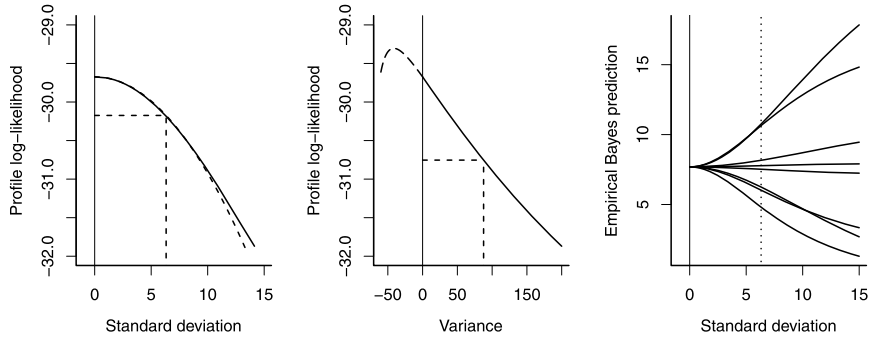
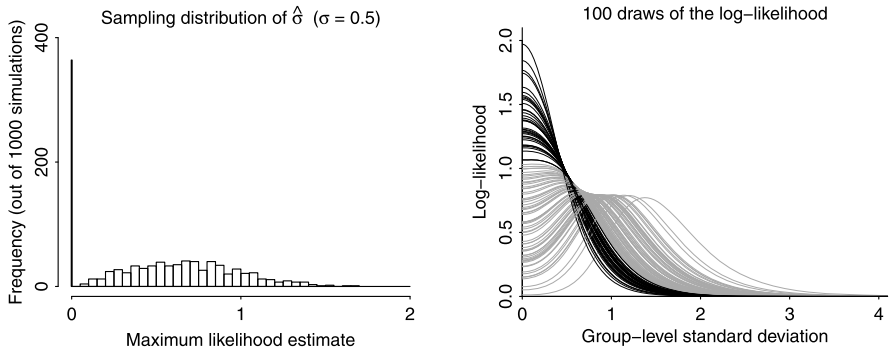


FIGURE 1.
Profile log-likelihood as a function of σ_θ (*left*) and σ_θ^2 (*middle*) and empirical Bayes predictions (*right*) for 8-schools data. The *dashed curve* on the *left* is the quadratic approximation at the mode, based on the estimated standard error. The *vertical dashed line* is the MPL estimate for a log-gamma(2, 0) penalty on σ_θ (*left*) or σ_θ^2 (*middle*). The *vertical dotted line* on the *right panel* indicates one standard error of $\hat{\sigma}_\theta^{\mathrm{ML}}$.



(a) (b)

FIGURE 2.
For a simple varying-intercept model with $\sigma_\theta = 0.5$ and $J = 10$ groups: (**a**) Sampling distribution of the maximum likelihood estimates $\widehat{\sigma}_\theta$, based on 1,000 simulations of data from the model. (**b**) Log-likelihood functions for 100 simulated datasets. When the maximum is at 0, curves are shown in *black* and otherwise in *gray*. The maximum likelihood estimates are extremely variable and the likelihood function is not very informative about σ_θ.

2.3. Simulation: Boundary Problems for a Simple Model

To demonstrate that boundary estimates occur frequently with a small number of groups and that large values of the variance also tend to be supported in such situations, we simulate data from a varying-intercept model with $J = 10$ groups. To keep things simple, we do not include covariates, treat the mean and within-group variance as known, and set the group size to $n = 1$:

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

From this model, with $\sigma_\theta = 0.5$, we create 1,000 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 2(a) shows the empirical sampling distribution of the maximum likelihood estimates of σ_θ. As expected, in more than a third of the simulations, the 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 2(b) displays 100 draws of the 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. Setting σ_θ to zero would be a mistake, and it would also be wrong to say that the likelihood offers no information at all. In particular, it bounds σ_θ on the high end. A fair point summary would be somewhere in the range supported by the likelihood, with a standard error high enough to acknowledge the uncertainty in the inference.

3. Maximum Penalized Likelihood Estimation of σ_θ

3.1. A Brief Review of Maximum Likelihood and Restricted Maximum Likelihood

We consider the model

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \theta_j + \epsilon_{ij}, \quad i = 1, \ldots, n_j, \ j = 1, \ldots, J, \ \sum_{j=1}^{J} n_j = N, \tag{2}$$

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. Another option is restricted or residualized maximum likelihood (REML, Patterson & Thompson, 1971), which is equivalent to specifying uniform priors for the regression coefficients $\boldsymbol{\beta}$ and finding 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 & 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, Ruppert, and Vogelsang (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.

3.2. Maximum Penalized Likelihood Estimation

In the present article, we are particularly concerned with the group-level standard deviation, and we specify a penalty for σ_θ. The penalized log-likelihood function can be written as

$$\log l_p(\sigma_\theta, \boldsymbol{\beta}, \sigma_\epsilon; \boldsymbol{y}) = \log l(\sigma_\theta, \boldsymbol{\beta}, \sigma_\epsilon; \boldsymbol{y}) + \log p(\sigma_\theta), \tag{3}$$

where the first term of the right-hand side is the log-likelihood and $\log p(\sigma_\theta)$ is an additive penalty term. We find the maximum penalized likelihood (MPL) estimator that maximizes (3).

The exponential of the penalty term can be regarded as a Bayesian prior density for σ_θ. Assuming a uniform prior, $p(\boldsymbol{\beta}, \sigma_\epsilon) = 1$, for $\boldsymbol{\beta}$ and σ_ϵ, the penalized log-likelihood is (up to an

additive constant) the marginal log-posterior density with varying intercepts (θ_j) integrated out. Therefore, the MPL estimates can be viewed as posterior modal estimates. By integrating the posterior over θ_j, we avoid the incidental parameter problem (Neyman & Scott, 1948; O'Hagan, 1976; Mislevy, 1986).

Unlike posterior mean estimation, maximum penalized likelihood (or posterior modal) estimation does not involve simulation and is computationally as efficient as maximum likelihood estimation. By modifying existing maximum likelihood estimation procedures, `gllamm` (Rabe-Hesketh, Skrondal, & Pickles, 2005, Rabe-Hesketh & Skrondal, 2012) in Stata and `lmer` (Bates & Maechler, 2010) in R, we have developed software to find the maximum of the penalized likelihood. The modified `gllamm` is available from www.gllamm.org and `blmer`, the modified `lmer` function, can be found in the `blme` package available from the Comprehensive R Archive Network. In both programs, the user has the option to specify a penalty that is added to the log-likelihood during optimization.

3.3. Log-Gamma Penalty Function

We propose the logarithm of a gamma density as a penalty function of σ_θ, which is equivalent to assigning a gamma (*not* inverse-gamma) prior on σ_θ,

$$p(\sigma_\theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\sigma_\theta^{\alpha-1}e^{-\lambda\sigma_\theta}, \quad \alpha > 0,\ \lambda > 0 \tag{4}$$

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).

As a default choice of the parameters, we suggest $\alpha = 2$ and $\lambda \to 0$. Since the gamma density with $\alpha = 2$ is 0 at the origin, the MPL estimate of σ_θ is always positive even when the maximum of the likelihood is at 0. In addition, with $\lambda \to 0$, the gamma density function has a positive constant derivative at zero, which allows the likelihood to dominate if it is strongly curved near zero. The positive constant derivative implies that the prior is linear at zero so that there is no dead zone near zero. The top-left panel of Figure 3 shows that the gamma$(2, 0.1)$ density increases linearly from zero with a gentle slope. The shape will be even flatter with a smaller rate parameter.

Other values than zero can also be used for λ when a researcher has prior knowledge about the group-level variance. For example, we can set $1/\lambda$ to the prior estimate of σ_θ since $1/\lambda$ is the mode of gamma$(2, \lambda)$. Choosing $\lambda \to 0$ as a default has the advantage that it does not depend on the scale of the response variable.

Various reasonable-seeming choices of priors are not useful for avoiding boundary estimates using the MPL approach. 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 or posterior mean inference (see Gelman, 2006), but do not work if the goal is to get a nondegenerate posterior mode estimate.

The *lognormal* and *inverse-gamma* densities are 0 at the boundary but all their derivatives are zero at the origin, essentially ruling out small estimates of σ_θ no matter what the data suggest. This can be seen in Figure 3 where both the inverse-gamma$(2, 5)$ (bottom-left) and lognormal$(1, 0.5)$ (top-right) have a cutoff below which the prior will dominate. The shape of the inverse-gamma changes dramatically depending on the choice of hyperparameters, as seen by comparing the inverse-gamma$(2, 5)$ with the inverse-gamma$(0.01, 0.01)$ (bottom-right). An inverse-gamma prior with small hyperparameters is often used as a noninformative prior for variances in multilevel models because it is flat apart from the spike near zero; but posterior mean inferences can be sensitive to the choice of hyperparameters (Gelman, 2006), unless the likelihood is concentrated away from zero (Browne & Draper, 2006). The posterior will have its mode close to the mode of the prior as long as the likelihood has moderate curvature, so that

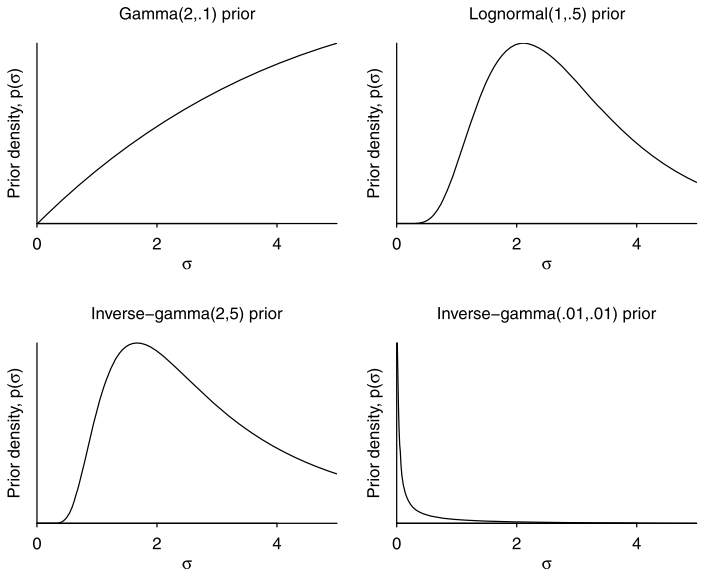


FIGURE 3.
Inverse-gamma, log-normal, and gamma priors. Both log-normal$(1, 0.5)$ (*top-right*) and inverse-gamma$(2, 5)$ (*bottom-left*) have a cutoff below which the prior will dominate. Inverse-gamma$(0.01, 0.01)$ (*bottom-right*) has a sharp peak at 0.01. Therefore, these priors will dominate the likelihood when the likelihood has a gentle curvature. In contrast, gamma$(2, 0.1)$ (*top-left*) increases slowly and linearly from zero.

the inverse-gamma prior with small hyperparameters becomes informative for posterior mode inference. Thus, the log-normal and inverse-gamma can only be used when there is real prior information to guide the choices of their two parameters; they cannot be a default choice of the sort we are seeking here.

4. Theoretical Properties

4.1. Difference Between Maximum Likelihood and Maximum Penalized Likelihood

To examine the effect of α and λ on the MPL estimator analytically, we treat $(\boldsymbol{\beta}, \sigma_\epsilon)$ as nuisance parameters and assume that the profile log-likelihood of σ_θ can be approximated by a quadratic function in σ_θ around the ML estimator, $\hat{\sigma}_\theta^{\mathrm{ML}}$,

$$\log L(\sigma_\theta) \approx -\frac{(\sigma_\theta - \hat{\sigma}_\theta^{\mathrm{ML}})^2}{2 \cdot \widehat{\mathrm{se}}(\hat{\sigma}_\theta^{\mathrm{ML}})^2} + c_1. \tag{5}$$

Here, $\widehat{\mathrm{se}}(\hat{\sigma}_\theta^{\mathrm{ML}})$ 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 $\hat{\sigma}_\theta^{\mathrm{ML}}$ is zero even when $\hat{\sigma}_\theta^{\mathrm{ML}}$ is zero.

For example, consider a balanced varying-intercept model without covariates by setting $\boldsymbol{x}_{ij}^T\boldsymbol{\beta} = \mu$ and $n_i = n$ in model (2). For convenience, we assume that σ_ϵ is known as σ_0. Then the profile log-likelihood of σ_θ is given by

$$\log L_{\sigma_\theta}(\sigma_\theta) = -\frac{J}{2}\log\left(\sigma_0^2 + n\sigma_\theta^2\right) - \frac{1}{2}\left(\frac{\mathrm{SST}}{\sigma_0^2} - \frac{n\sigma_\theta^2}{\sigma_0^2(\sigma_0^2 + n\sigma_\theta^2)}\mathrm{SSB}\right)$$

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

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

$$\frac{\partial \log L_{\sigma_\theta}}{\partial \sigma_\theta} = \left(-\frac{nJ}{2(\sigma_0^2 + n\sigma_\theta^2)} + \frac{n \cdot \text{SSB}}{2(\sigma_0^2 + n\sigma_\theta^2)^2} \right) \cdot 2\sigma_\theta. \tag{6}$$

When we have a boundary estimate of σ_θ, the log-likelihood function of σ_θ^2 usually has its maximum in the negative region, and so $\partial \log L_{\sigma_\theta}/\partial(\sigma_\theta^2)$ (the part in the parentheses on the right-hand side in (6)) 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, (6) 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 $\hat{\sigma}_\theta^{\text{ML}} = 0$) and the higher order terms are negligible around $\sigma_\theta = 0$. In Sections 5 and 6, we will confirm that the quadratic approximation fits well in two applications and in simulations.

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

Property 1. When $\hat{\sigma}_\theta^{\text{ML}} = 0$, for fixed $\alpha > 1$ and $\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$, the largest possible MPL estimate is attained when $\lambda \to 0$ with the value

$$\widehat{\sigma}_\theta = \widehat{\text{se}}\big(\hat{\sigma}_\theta^{\text{ML}}\big)\sqrt{\alpha - 1}. \tag{7}$$

When $\alpha = 2$, we obtain $\widehat{\sigma}_\theta = \widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$. That is, when the ML estimate is on the boundary, the log-gamma$(2, \lambda)$ penalty shifts the MPL estimate away from zero but not more than one estimated standard error.

One standard error can be regarded as a statistically insignificant distance from the ML estimate. If the quadratic approximation in (5) holds and $\hat{\sigma}_\theta^{\text{ML}}$ is zero, the likelihood-ratio test (LRT) statistic for $H_0 : \sigma_\theta = \widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$ is $2(\log L(0) - \log L(\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}}))) = 1$. Testing $H_0 : \sigma_\theta = 0$ is not a standard problem because the null value is on the boundary of the parameter space and this problem has been investigated by several authors (Self & Liang, 1987; Stram & Lee, 1994). The 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 & Ruppert, 2004). For testing $H_0 : \sigma_\theta = \widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$ (> 0), the percentile will be larger because there is less point mass at zero (Crainiceanu et al., 2003). Therefore, an LRT statistic of 1 can be considered small.

Property 2. When $\hat{\sigma}_\theta^{\text{ML}} > 0$, for fixed $\alpha > 1$ and $\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$, the largest possible MPL estimate is attained when $\lambda \to 0$ with the value

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

In addition, $\partial\widehat{\sigma}_\theta/\partial\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$ decreases with $\hat{\sigma}_\theta^{\text{ML}}$.

Similar to the case of $\hat{\sigma}_\theta^{\text{ML}} = 0$, $\hat{\sigma}_\theta$ is greater than $\hat{\sigma}_\theta^{\text{ML}}$ and is an increasing function of $\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$. The gradient $\partial\widehat{\sigma}_\theta/\partial\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$ has maximum $\sqrt{\alpha - 1}$ for $\hat{\sigma}_\theta^{\text{ML}} = 0$ that coincides with (7) and decreases as $\hat{\sigma}_\theta^{\text{ML}}$ increases. This implies that the log-gamma(α, λ) penalty does not shift the MPL estimate as much when $\hat{\sigma}_\theta^{\text{ML}} > 0$ as it does when $\hat{\sigma}_\theta^{\text{ML}} = 0$ 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.

We have discussed the log-gamma penalty on the group-level standard deviation (σ_θ) since the profile log-likelihood as a function of σ_θ has a better quadratic approximation, and thus helps us to investigate the properties in Section 4.1. However, one might still be interested in penalties on the variance, σ_θ^2.

Property 3. In the limit as $\lambda \to 0$, a log-gamma(α, λ) penalty on σ_θ^2 is equivalent to a log-gamma$(2\alpha - 1, \lambda)$ penalty on σ_θ.

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

4.2. Asymptotic Properties

Although this paper is concerned with the problem of boundary estimates which occur 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 varying-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 $\hat{\sigma}_\theta^{\mathrm{ML}} = \{\max(\frac{1}{J}\sum_{j=1}^{J}(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 - \frac{1}{n}\sigma_\epsilon^2, 0)\}^{1/2}$.

When the log-gamma(α, λ) penalty is applied to σ_θ, the MPL estimator, say $\hat{\sigma}_\theta^{\mathrm{MPL}}$, is a root of a fifth order polynomial (see Appendix B). Therefore, we do not have a simple formula for $\hat{\sigma}_\theta^{\mathrm{MPL}}$, but we can investigate its asymptotic properties using expansions of the penalized log-likelihood (or the log-posterior) function.

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

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

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

Theorem 4. *At the order of J^{-1}, the ML estimator and the MPL estimator have the following bias equations, respectively,*

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

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

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

Proof: An outline of the proof is in Appendix B and Dorie (2013). □

The MPL estimator of σ_θ with the default penalty is not only asymptotically unbiased and as efficient as the ML estimator, but also has the same magnitude of bias at the higher order as

seen in Theorem 4. In addition, the MPL estimator tends to be less biased for small J as will be shown using simulations in Section 6.

4.3. *Bayesian Point Estimation and Transformation of σ_θ*

From a Bayesian point of view, 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 or median which have good asymptotic properties. The ML estimator is invariant under transformations, but the posterior modal estimator is not because of the change in prior density when transforming σ_θ. Thus, the transformation affects the posterior mode.

Property 5. The posterior with a gamma$(2, \lambda)$ prior on σ_θ and with $\lambda \to 0$ is the same (as a function of σ_θ) as the posterior of $\log(\sigma_\theta)$ with a (improper) uniform prior $p(\sigma_\theta) = 1$. If the log transformation symmetrizes the posterior density, the posterior mode of σ_θ with a gamma$(2, \lambda)$ prior is equal to the posterior median of σ_θ with a uniform prior.

This implies that the posterior mode $\widehat{\log(\sigma_\theta)}$ of $\log(\sigma_\theta)$ with a uniform prior on σ_θ is the same as $\log(\hat{\sigma}_\theta)$ where $\hat{\sigma}_\theta$ is the posterior mode with a gamma$(2, \lambda)$ prior on σ_θ.

With a uniform prior on σ_θ, the profile posterior density of σ_θ is just the profile likelihood of σ_θ, 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, so that the posterior mode of $\log(\sigma_\theta)$ is close to the posterior median of $\log(\sigma_\theta)$. Since the log transformation is strictly increasing, the posterior median of $\log(\sigma_\theta)$ is the same as $\log(\tilde{\sigma}_\theta)$ where $\tilde{\sigma}_\theta$ is the median of the original posterior of σ_θ with the uniform prior on σ_θ. Because the posterior of $\log \sigma_\theta$ with the uniform prior on σ_θ has the same functional form as the posterior of σ_θ with the gamma$(2, \lambda)$ prior on σ_θ, they are maximized at the same σ_θ, which is also close to $\tilde{\sigma}_\theta$.

This relationship between the log transformation and the gamma prior can be extended to general power transformations. Consider the Box–Cox transformations (1964), $g_\gamma(\sigma_\theta) = (\sigma_\theta^\gamma - 1)/\gamma$ if $\gamma \neq 0$ and $g_\gamma(\sigma_\theta) = \log(\sigma_\theta)$ if $\gamma = 0$.

Property 6. With $\lambda \to 0$, maximizing the posterior of $g_\gamma(\sigma_\theta)$ with a gamma(α, λ) prior on σ_θ is equivalent to maximizing the posterior of σ_θ with a gamma$(\alpha + 1 - \gamma, \lambda)$ prior on σ_θ. If g_γ symmetrizes the posterior density, the posterior mode of σ_θ with a gamma$(\alpha + 1 - \gamma, \lambda)$ prior on σ_θ is the same as the posterior median of σ_θ.

It follows that the shape parameter α can be chosen to attain a more symmetric posterior density so that the posterior mode is close to the posterior median with the uniform prior.

4.4. *Connection to REML*

Patterson and Thompson (1971) describe the REML log-likelihood, say $\log L_R$, in terms of the original log-likelihood, L, and an additive penalty term,

$$\log L_R = \log L - \frac{1}{2} \log\bigl\{\det\bigl(X^T V^{-1} X\bigr)\bigr\}, \tag{8}$$

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 (2), V is a block-diagonal matrix with $n_j \times n_j$ blocks, V_j, $j = 1, \ldots, J$, where V_j contains $\sigma_\theta^2 + \sigma_\epsilon^2$ on the diagonal and σ_θ^2 on the off-diagonals. Recalling that the penalized log-likelihood in (3) is the sum of the log-likelihood

and the log-gamma density, the second term in (8), denoted by $\log p_R(\sigma_\theta^2)$, is analogous to the log of the gamma density function.

In order to compare the REML and log-gamma penalties, we consider a special case of model (2) with balanced group size n, q level-1 covariates, and r level-2 covariates. The level-1 covariates, written as columns $z_1, \ldots, z_q$ of the design matrix, consist of the same elements for each group and satisfy $\mathbf{1}^T z_u = 0$, $z_u^T z_{u'} = 1$ 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

$$\log p_R\left(\sigma_\theta^2\right) = \frac{r+1}{2} \log\left(\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n}\right) + c_1, \tag{9}$$

where c_1 is a constant. The proof is provided in Appendix C.

Recall that when $\lambda \to 0$, the log-gamma(α, λ) penalty on σ_θ^2 (equivalently log-gamma$(2\alpha - 1, \lambda)$ on σ_θ) can be written as

$$\log p\left(\sigma_\theta^2\right) = (\alpha - 1) \log \sigma_\theta^2 + c_2. \tag{10}$$

Ignoring the constant terms that have no influence on the MPL estimate, we see that the log-gamma$(\frac{r+3}{2}, \lambda)$ on σ_θ^2 (equivalently log-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-gamma penalty term in (10) is $-\infty$ for $\alpha > 1$, whereas the REML penalty in (9) 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 log-gamma penalty assigns more penalty on σ_θ close to zero than REML for small n and large σ_ϵ.

The REML penalty expression in (9) is derived for covariates with specific properties as described above. However, we found that the relationship between the REML and log-gamma penalties illustrated in this section holds more generally (see Appendix D).

4.5. Connection to Adjustment for Density Maximization

Adjustment for density maximization (ADM; Morris, 2006; Li & Lahiri, 2010; Morris & Tang, 2011) has been proposed for obtaining strictly positive group-level variance estimates in the context of small area estimation. The Fay–Herriot model (1979) considered in these papers is equivalent to random-effects meta-regression, the model in (1) but with covariates. The focus is on prediction of θ_i and, therefore, on estimation of the shrinkage factor $B_i = s_i^2/(s_i^2 + \sigma_\theta^2)$ because the conditional means and variances of θ_i are linear in B_i, not in σ_θ^2. With a prior $\pi(\sigma_\theta^2)$, Morris and Tang (2011) approximate the posterior of B_i by a beta distribution and adjust the posterior by multiplying by $B_i(1 - B_i)$. The mode of the adjusted posterior density approximates the posterior mean of B_i and the posterior variance can also be approximated using the second derivative of the adjusted density at the maximum. This procedure leads to the maximization of the adjusted (profile) likelihood of σ_θ^2, $L_a(\sigma_\theta^2) = \sigma_\theta^2 \pi(\sigma_\theta^2) L(\sigma_\theta^2)$.

Based on the restriction of $\pi(\sigma_\theta^2)$ to scale-invariant improper priors $\pi(\sigma_\theta^2) = (\sigma_\theta^2)^{c-1}$, this method is equivalent to MPL estimation with a log-gamma(α, λ) penalty on σ_θ with $\lambda \to 0$ and $\alpha = 2c + 1$. Therefore, MPL also shares the properties of ADM for meta-regression and the Fay–Herriot model (when the within-group variances are treated as known), such as predictions of θ_i being minimax for mean squared-error loss when the within-group variances are equal and $c \leq 1$ (Morris & Tang, 2011).

Morris and Tang's proposal $\pi(\sigma_\theta^2) = 1$ corresponds to MPL with a log-gamma$(2, \lambda)$ penalty on σ_θ^2, equivalently a log-gamma$(3, \lambda)$ on σ_θ with $\lambda \to 0$. With constant variances $s_i^2 = s^2$, MPL therefore produces the James–Stein shrinkage constant as does ADM. Morris and Tang (2011)

TABLE 1.
ML and MPL estimates for the meta-analysis, where the penalty is log-gamma$(\alpha, 0)$ on σ_θ. The MPL estimates are approximately at $\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})\sqrt{1-\alpha}$.

Method	μ			σ_θ		Log-lik
	est	se	se^R	est	se	
ML	7.69	4.07	3.33	0.00	6.32	−29.67
MPL: log-gamma$(2, 0)$ on σ_θ	7.92	4.72	3.39	6.30	4.61	−30.18
MPL: log-gamma$(3, 0)$ on σ_θ	8.10	5.38	3.43	9.42	5.34	−30.76

se^R: robust standard error (sandwich estimator).

also mentioned the case $c = 1/2$, which is equivalent to our default penalization. However, they support the flat prior, showing that the corresponding posterior mode approximates the posterior mean and variance of B_i well in simulation studies.

Different from the focus on prediction of θ_i, we are interested in estimating the model parameters and the standard errors in the setting of linear mixed effect models. The MPL method with a log-gamma penalty covers the ADM but is more flexible by allowing different choices for the shape and rate parameter based on prior knowledge.

5. Examples

5.1. Meta-analysis

In this section, we apply MPL estimation with the log-gamma penalty function to the 8 schools data introduced in Section 2.2.

In this and the following sections, we use the expression "log-gamma$(\alpha, 0)$" though gamma is defined only for $\lambda > 0$. Log-gamma$(\alpha, 0)$ refers to the function $(\alpha - 1)\log\sigma_\theta$, which is the same as log-gamma(α, λ) up to a constant when $\lambda \to 0$. Using a penalty function with a very small value of λ, for example log-gamma$(2, 10^{-4})$, gives very close results to log-gamma$(2, 0)$ for the examples and simulations in the following sections. For the model in (1), we consider two different penalties: log-gamma$(2, 0)$ and log-gamma$(3, 0)$ on σ_θ. MPL estimates with these penalties and ML estimates are given in Table 1. The ML estimate of σ_θ is zero as mentioned in Section 2.2, and the estimated standard error of $\hat{\sigma}_\theta^{\text{ML}}$ is 6.32 (which corresponds to $\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$ in Section 4.1).

The MPL estimates $\hat{\sigma}_\theta^{\text{MPL}}$ are 6.30 and 9.42 for $\alpha = 2$ and $\alpha = 3$, respectively. These are close to the values $\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})\sqrt{\alpha - 1}$ with $\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}}) = 6.32$, which we expect with $\hat{\sigma}_\theta^{\text{ML}} = 0$ if the profile log-likelihood is approximately quadratic in σ_θ, as it appears to be in the left panel of Figure 1. In both cases, the log-likelihood at the MPL estimate is only a little bit lower than the maximum log-likelihood.

Table 1 also reports model-based and robust (based on the sandwich estimator) standard error estimates for $\widehat{\mu}$. We see that the model-based standard error increases with $\hat{\sigma}_\theta$ as mentioned in Section 2.2, whereas the robust standard error change very little.

Figure 1 shows the profile log-likelihood (maximized with respect to μ) of σ_θ (left) and σ_θ^2 (middle). On the left we see that the profile log-likelihood has its maximum at zero where the gradient is zero as discussed in Section 4.1. Further, the profile log-likelihood is quite flat. We see in the middle panel of Figure 1 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. In contrast, the profile log likelihood as a function of σ_θ is well-approximated by a quadratic at the mode (dashed curve in left figure).

TABLE 2.
Parameter estimates of the varying-intercept model for the cluster-randomized trial data. ML gives $\hat{\sigma}_\theta = 0$ but MPL with log-gamma(2, 0) gives 0.26, which is close to the standard error of the ML estimate. The log-likelihood is reduced by only 0.4. The standard errors of the fixed coefficient estimates are larger for MPL as expected.

	ML		MPL	
	est	se	est	se
Intercept	19.51	0.86	19.35	0.88
Meat	0.50	0.40	0.53	0.46
Milk	−0.60	0.39	−0.60	0.45
Calorie	−0.28	0.39	−0.21	0.46
Age at time 0	0.02	0.11	0.03	0.11
$\hat{\sigma}_\theta$	0.00	0.31	0.26	0.18
$\hat{\sigma}_\epsilon$	3.07	0.10	3.07	0.10
Log-likelihood	−1260.2		−1260.6	

5.2. *Cluster-Randomized Trial*

Whaley, Sigman, Neumann, Bwibo, Guthrie, Weiss, Alber, and Murphy (2003) present a cluster-randomized trial from rural Kenya that was designed to explore the impact of three different diets on the cognitive development of school children. Twelve schools were randomized to provide meals with different dietary supplements: Meat, Milk, Energy, or No meal. Cognitive assessments (Raven's score) were made at 5 time-points, and we analyze the last observation made after 21 months of treatment for a total of 496 children. The data were provided by Weiss (2005). Three school-level dummy variables for the three dietary treatments and one child-level covariate (age at the beginning of treatment) are included in the varying-intercept model, written as

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4ij} + \theta_j + \epsilon_{ij}, \quad j = 1, \ldots, 12, \ i = 1, \ldots, n_j \tag{11}$$

where $\theta_j \sim N(0, \sigma_\theta^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

Table 2 presents ML and MPL estimates with a log-gamma(2, 0) penalty for the model in (11). The ML estimate of the school-level standard deviation is on the boundary, whereas the MPL estimate is 0.26, close to 0.31, the estimated standard error of the ML estimate. The log-likelihood decreases negligibly, by only 0.4, at the MPL estimate. Figure 4 shows that the profile log-likelihood of σ_θ is approximately quadratic in σ_θ near the mode. As expected, the estimated standard errors of the treatment effect estimates are larger for MPL than for ML because the treatments vary between groups.

6. Simulation Study: Balanced Varying-Intercept Model

We consider a varying-intercept model,

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

with $J = 3, 5, 10, 30$ groups and $n = 5, 30$ observations per group. Three groups represent an extreme situation that is unlikely to occur often in practice, and 30 is the largest number of groups considered because the penalty term is not likely to have much influence for more than 30 groups. Five units per group is small and occurs, for instance, in longitudinal data, whereas 30 units per group is quite large and fairly common in cross-sectional settings. This model includes two

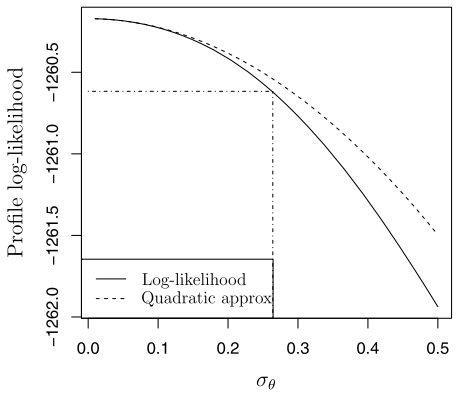


FIGURE 4.
Profile log-likelihood curve of σ_θ and its quadratic approximation at the maximum for the cluster-randomized trial data. The *dash-dotted line* indicates the MPL estimate, 0.264.

covariates: $x_{1ij} = i$ varies within groups only (its mean is constant across groups), and $x_{2ij} = j$ varies between groups only. The coefficients β_0, β_1, β_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 generated 1,000 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 MPL estimates with log-gamma$(2, 0)$ and log-gamma$(3, 0)$ penalties on σ_θ. The REML penalty corresponds to $\alpha = 3$ since the model contains one group-level covariate. We compare MPL 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 $\sigma_\theta = \frac{1}{\sqrt{3}}$, 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 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 MPL estimates are never on the boundary in any of the simulation conditions. At the same time, the likelihood at the MPL estimates does not differ considerably from the maximum. The likelihood ratio test statistic $-2[\log L(\widehat{\sigma}_\theta^{\text{MPL}}) - \log L(\widehat{\sigma}_\theta^{\text{ML}})]$ for testing the restriction $\sigma_\theta = \widehat{\sigma}_\theta^{\text{MPL}}$ 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 4.1 by assuming that the profile log-likelihood is quadratic. Figure 5 shows that the MPL estimates calculated by the quadratic approximation of the profile log-likelihood (see Properties 1 and 2) agree well with the MPL estimates with a log-gamma$(2, 0)$ penalty on σ_θ for $J = 3$ (left) and $J = 30$ (right) when $\rho = 0.25$ and $n = 30$.

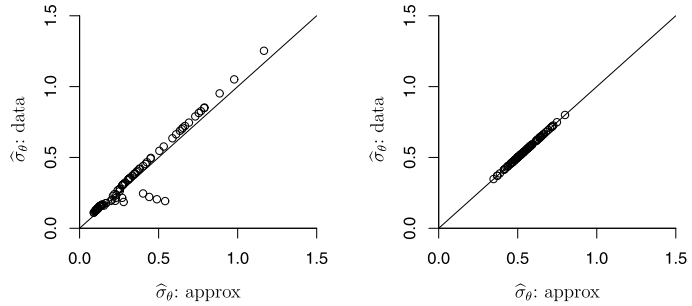


FIGURE 5.
MPL estimates with a log-gamma(2, 0) penalty on σ_θ for $J = 3$ (*left*) and $J = 30$ (*right*), $\rho = 0.25$ and $n = 30$ for the first 100 replicates, compared with the MPL estimates based on the quadratic approximation of the profile log-likelihood (see Properties 1 and 2). Agreement is 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.

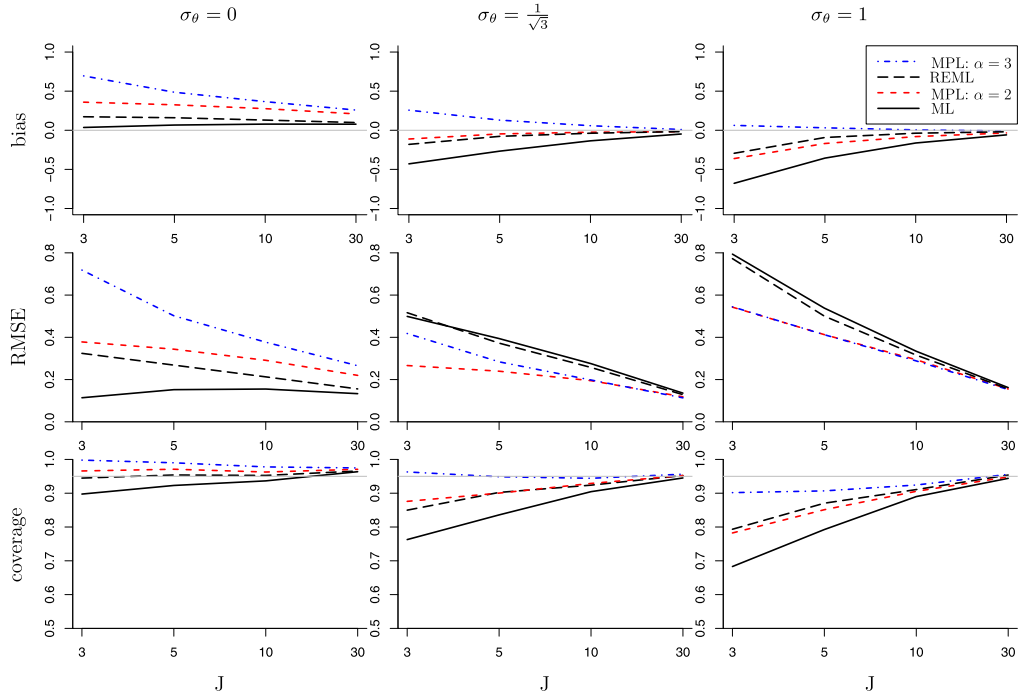


FIGURE 6.
Bias of σ_θ, RMSE for σ_θ, and coverage of confidence intervals for β_2 for group size $n = 5$, standard deviation $\sigma_\theta = 0, \frac{1}{\sqrt{3}}$, and 1 (columns) and number of groups $J = 3, 5, 10, 30$ (x-axis). Different estimators are represented by different line patterns. When $\sigma_\theta > 0$, all the methods outperform ML and the bias of the MPL estimator is as low as REML depending on α. Also, when $\sigma_\theta > 0$, the RMSE of the MPL estimator with both values of α is smaller than for REML and ML and coverage is best for the MPL estimator with $\alpha = 3$.

Figure 6 summarizes the estimated bias and the root mean squared error (RMSE) of σ_θ, and the coverage of 95 % confidence intervals for β_2 for the four methods for $n = 5$, $J = 3$, 5, 10, 30, and $\sigma_\theta = 0, \frac{1}{\sqrt{3}}$, 1 and Figure 7 gives the results for $n = 30$.

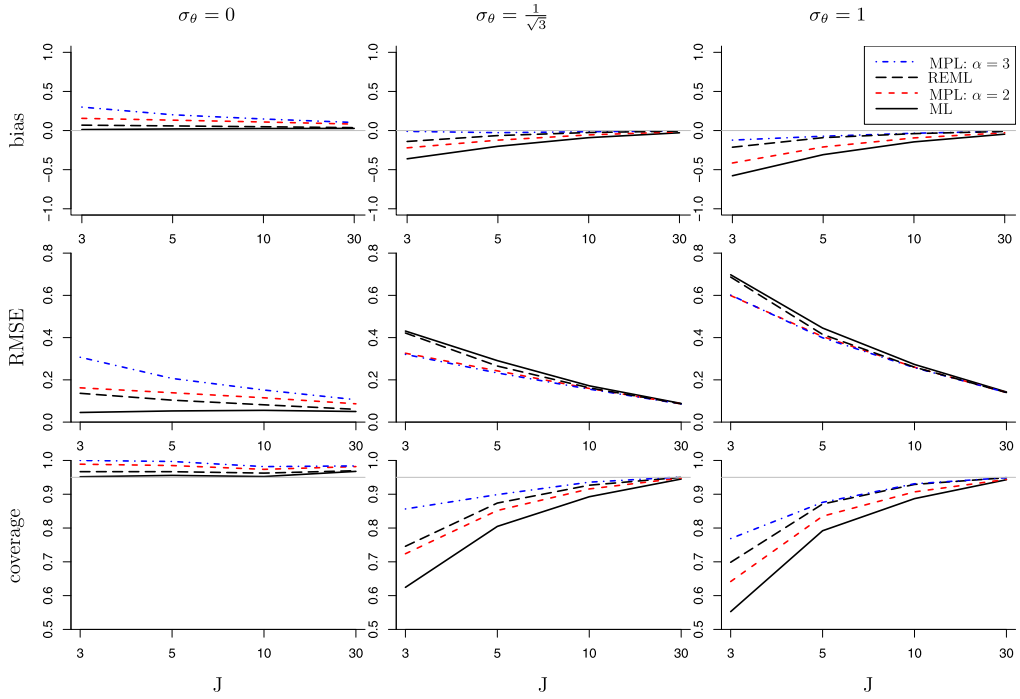


FIGURE 7.
Bias, RMSE of σ_θ and coverage of CI for β_2 for group size $n = 30$, standard deviation $\sigma_\theta = 0$, $\frac{1}{\sqrt{3}}$, and 1 (columns) and number of groups $J = 3, 5, 10, 30$ (x-axis). Different estimators are represented by different line patterns. When $\sigma_\theta > 0$, all the methods outperform ML and the bias of the MPL estimator is as low as REML depending on α. Also, when $\sigma_\theta > 0$, the RMSE of the MPL estimator with both values of α is smaller than for REML and ML and coverage is best for the MPL estimator with $\alpha = 3$.

Estimates of σ_θ The first rows of Figures 6 and 7 show 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. MPL estimates with a log-gamma$(2, 0)$ penalty also tend to be downward biased for σ_θ but not as much as the ML estimates. On the other hand, the MPL estimator with log-gamma$(3, 0)$ produces the largest estimates among the four estimators so it often overestimates σ_θ. For $\sigma_\theta = 1$, the MPL estimator with log-gamma$(3, 0)$ has the smallest bias for all J. When $\sigma_\theta = 0$, as expected, the MPL estimators assign more penalty on the values close to the boundary than REML, so the bias is larger than for REML and ML.

The overall pattern is the same for $n = 5$ and $n = 30$, but for $n = 30$ the MPL with log-gamma$(3, 0)$ is closer to REML for $\sigma_\theta > 0$ than for $n = 5$. This confirms that the log-gamma penalty on σ_θ with $\alpha = 3$ is similar to the REML penalty when the model contains one group-level covariate, particularly with large n, as discussed in Section 4.4.

The root mean squared errors (RMSE) of both MPL estimators are consistently smaller than for ML and REML when σ_θ is not zero (see middle rows of Figures 6 and 7). For $\sigma_\theta = \frac{1}{\sqrt{3}}$ and $\sigma_\theta = 1$, REML has similar or smaller bias than MPL with a log-gamma$(2, 0)$ penalty, but its RMSE is appreciably larger for small J because the REML estimator has the largest variance among the four estimators. The MPL estimator tends to have smaller RMSE with a log-

gamma(2, 0) penalty than with a log-gamma(3, 0) penalty but the difference decreases as n, J, and σ_θ increase.

Coverage of Confidence Intervals for β_2 The standard error estimates of the estimated coefficient of the group-level covariate ($\widehat{\beta}_2$) are 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 & 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 confidence intervals will be too narrow.

The bottom rows of Figures 6 and 7 show the proportions of 95 % confidence intervals 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 confidence intervals 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 MPL estimator with $\alpha = 3$ has the best coverage, particularly for $\sigma_\theta = \frac{1}{\sqrt{3}}$ and $n = 5$. Although the MPL estimator with $\alpha = 3$ tends to have large positive bias for σ_θ, it turns out to give better coverage. Recalling that log-gamma(3, 0) is close to the REML penalty (discussed in Section 4.4) for large n, the coverage for the MPL estimator with $\alpha = 3$ is closer to REML for $n = 30$ than for $n = 5$. However, REML still shows significantly lower coverage than the MPL estimator, particularly for small J.

We also considered the average ratio of the widths of the MPL versus REML confidence intervals when a log-gamma(3, 0) penalty is used for MPL. The largest increase in the widths of the confidence intervals occurs in the extreme situation when $\sigma_\theta = 0$ and J is small, with an average increase of 60 % for $J = 3$ and 30 % for $J = 5$ when $n = 5$.

However, we are interested in the situation where zero variance is considered unreasonable a priori; and for a moderate standard deviation $\sigma_\theta = \frac{1}{\sqrt{3}}$, the MPL confidence intervals are on average 20 % wider for $J = 5$, 8 % wider for $J = 10$, and 2 % wider for $J = 30$ when $n = 5$. In these situations, MPL improves the coverage while it increases the width of confidence intervals by a moderate amount.

In summary, the MPL estimator with a log-gamma penalty is successful at avoiding boundary solutions; and, at the same time, the likelihood does not change substantially most of the time. Furthermore, the MPL method performs as well as or better than ML or REML: if σ_θ is not zero, the RMSE for σ_θ is uniformly lower for the MPL estimator with both penalties than for the REML and ML estimators. When there are very few groups ($J = 3$ and $J = 5$), the MPL estimators have greater bias than REML even when σ_θ is not 0. With such a small number of groups, it would be preferable to use an informative prior that incorporates prior knowledge. However, the default priors have smaller RMSE than REML, even for these extreme cases. Comparing the MPL estimator with $\alpha = 2$ and $\alpha = 3$, $\alpha = 2$ appears better in terms of bias and RMSE, whereas $\alpha = 3$ produces better coverage. Although there is no obvious winner, both penalties successfully avoid boundary estimates.

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 $\sigma_\theta = 0, \frac{1}{\sqrt{3}}, 1$, we compared ML and REML estimates with MPL estimates with a log-gamma(2, 0) penalty, which corresponds to the REML penalty when there is no group-level covariate. (Results are in Appendix E.)

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 MPL estimates. The advantage of the log-gamma penalty 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 MPL estimators perform better in this regard.

7. Discussion

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

We have shown that the strategy of accepting the maximum likelihood estimate results in undercoverage of confidence intervals for regression coefficients of group-level covariates. In datasets where boundary estimates occur, a 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.

We proposed log-gamma(α, λ) with $\alpha = 2$ and $\lambda \to 0$ as our default choice for the penalty function, but sometimes weak prior information is available about a variance parameter. When $\alpha = 2$, the gamma density has its mode at $1/\lambda$, and so one can 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, the gamma function assigns greater penalty on the boundary, but this is acceptable if it represents real information about σ_θ.

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 log-gamma penalty can be naturally extended to the log-Wishart penalty on Σ, which is equivalent to the sum of log-gamma penalties on the eigenvalues of $\Sigma^{1/2}$. Therefore, the log-Wishart penalty with a certain choice of parameters will shift the MPL estimate of each eigenvalue away from 0, or equivalently move the MPL estimate of Σ away from singularity. At the same time, it moves the eigenvalues approximately at most one standard error away from the ML estimates as did the log-gamma$(2, 0)$ 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 as 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 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, Liu, & Gelman, 2013). 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.

Acknowledgements

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).

Appendix A. Derivation of Properties in Section 4

Here, we derive Properties 1 and 2.

Properties 1 and 2: With the quadratic approximation of the profile log-likelihood in Section 3.2 using Equation (5), the MPL estimator is given by

$$\widehat{\sigma}_\theta = -\frac{\lambda \cdot \widehat{\mathrm{se}}(\hat{\sigma}_\theta^{\mathrm{ML}})^2}{2} + \frac{\hat{\sigma}_\theta^{\mathrm{ML}}}{2} + \frac{1}{2}\sqrt{\left(\lambda \cdot \widehat{\mathrm{se}}\left(\hat{\sigma}_\theta^{\mathrm{ML}}\right)^2 - \hat{\sigma}_\theta^{\mathrm{ML}}\right)^2 + 4(\alpha-1)\widehat{\mathrm{se}}\left(\hat{\sigma}_\theta^{\mathrm{ML}}\right)^2}. \quad \text{(A.1)}$$

With a simple calculation, we can show that $\partial\widehat{\sigma}_\theta/\partial\lambda \le 0$. Therefore, as $\lambda \to 0$ for fixed α and $\widehat{\mathrm{se}}(\hat{\sigma}_\theta^{\mathrm{ML}})$, the MPL estimate increases monotonically to the maximum. When $\hat{\sigma}_\theta^{\mathrm{ML}} = 0$, the maximum is $\widehat{\mathrm{se}}(\hat{\sigma}_\theta^{\mathrm{ML}})\sqrt{\alpha-1}$. When $\hat{\sigma}_\theta^{\mathrm{ML}} > 0$, (A.1) is reduced into

$$\widehat{\sigma}_\theta = \frac{\hat{\sigma}_\theta^{\mathrm{ML}}}{2} + \frac{\hat{\sigma}_\theta^{\mathrm{ML}}}{2}\sqrt{1 + 4(\alpha-1)\widehat{\mathrm{se}}\left(\hat{\sigma}_\theta^{\mathrm{ML}}\right)^2 \Big/ \left(\hat{\sigma}_\theta^{\mathrm{ML}}\right)^2} > \hat{\sigma}_\theta^{\mathrm{ML}}.$$

In addition, $\partial\widehat{\sigma}_\theta/\partial\widehat{\mathrm{se}}(\hat{\sigma}_\theta^{\mathrm{ML}})$ becomes

$$\frac{\partial\widehat{\sigma}_\theta}{\partial\widehat{\mathrm{se}}(\hat{\sigma}_\theta^{\mathrm{ML}})} = \frac{\alpha-1}{\sqrt{\alpha - 1 + (\hat{\sigma}_\theta^{\mathrm{ML}})^2/\{4\widehat{\mathrm{se}}(\hat{\sigma}_\theta^{\mathrm{ML}})^2\}}},$$

which decreases as $\hat{\sigma}_\theta^{\mathrm{ML}}$ increases.

Property 3: If we assign the log-gamma(α, λ) penalty on σ_θ^2 instead of σ_θ, the penalty 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 the log-gamma$(2\alpha-1, \lambda)$ penalty on σ_θ.

Property 6: Let $t = g_\gamma(\sigma_\theta)$. Then the Jacobian is $\partial g_\gamma^{-1}(t) = (\gamma t + 1)^{1/\gamma - 1}$, which is $\sigma_\theta^{1-\gamma}$ when written as a function of σ_θ. Therefore, 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$(\alpha - \gamma + 1, \lambda)$.

Appendix B. Proof of Theorem 4

Proof: Let $S_{nJ} = (\sum_j (\bar{y}_{\cdot j} - \mu)^2)/n^2 J$ and $T_{nJ} = S_{nJ} - (\sigma_\epsilon^0)^2/n - \sigma_\theta^2$. Then S_{nJ} follows $(\sigma_\epsilon^2/n + (\sigma_\theta^0)^2)\chi_J^2/J$, $T_{nJ} = O_p(J^{-1/2})$, $E(T_{nJ}) = 0$ and $Var(T_{nJ}) = 2(\sigma_\epsilon^2 + (\sigma_\theta^0)^2)^2/J$. Using these terms, we can expand $\hat{\sigma}_\theta^{\mathrm{ML}}$ as

$$\hat{\sigma}_\theta^{\mathrm{ML}} = \sqrt{T_{nJ} + \left(\sigma_\theta^0\right)^2} = \sigma_\theta^0 + \frac{1}{2\sigma_\theta^0}T_{nJ} - \frac{1}{8(\sigma_\theta^0)^3}T_{nJ}^2 + O_p\left(J^{-3/2}\right).$$

Therefore, we have

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

For the asymptotic bias of $\hat{\sigma}_\theta^{\mathrm{MPL}}$, here we describe the outline of the proof. Details are in Dorie (2013). We will work with an estimating equation $\psi_{nJ}(\sigma_\theta)$, given by

$$\psi_{nJ}(\sigma_\theta) = \sigma_\theta^5 + \sigma_\theta^4\left(\frac{J}{\lambda} - \frac{\alpha-1}{\lambda}\right) + 2\sigma_\theta^3\frac{\sigma_\epsilon^2}{n} + \sigma_\theta^2\left(\frac{J}{\lambda}\frac{\sigma_\epsilon^2}{n} - \frac{J}{\lambda}S_{nJ} - 2\frac{\alpha-1}{\lambda}\frac{\sigma_\epsilon^2}{n}\right) + \sigma_\theta\frac{\sigma_\epsilon^4}{n^2} - \frac{\alpha-1}{\lambda}\frac{\sigma_\epsilon^4}{n^2},$$

and $\hat{\sigma}_\theta^{\mathrm{MPL}}$ will be a root of $\psi_{nJ}(\sigma_\theta) = 0$. The expression above Theorem 4 gives $\hat{\sigma}_\theta^{\mathrm{MPL}} - \sigma_\theta^0 = O_p(J^{-1/2})$. Therefore, the Taylor expansion of ψ_{nJ} around σ_θ^0 is given by

$$\psi_{nJ}\big(\hat{\sigma}_\theta^{\mathrm{MPL}}\big) = \psi_{nJ}\big(\sigma_\theta^0\big) + \psi'_{nJ}\big(\sigma_\theta^0\big)\big(\hat{\sigma}_\theta^{\mathrm{MPL}} - \sigma_\theta^0\big) + \frac{1}{2}\psi''_{nJ}\big(\sigma_\theta^0\big)\big(\hat{\sigma}_\theta^{\mathrm{MPL}} - \sigma_\theta^0\big)^2 + o_p\big(J^{-1}\big).$$

As the left-hand side of the approximation is 0, we can complete the square to obtain:

$$\hat{\sigma}_\theta^{\mathrm{MPL}} - \sigma_\theta^0 = \frac{-\psi'_{nJ}(\sigma_\theta^0) \pm \sqrt{\psi'_{nJ}(\sigma_\theta^0)^2 - 2\psi_{nJ}(\sigma_\theta^0)\psi''_{nJ}(\sigma_\theta^0) - 2\psi''_{nJ}(\sigma_\theta^0)o_p(J^{-1})}}{\psi''_{nJ}(\sigma_\theta)}.$$

Note that each of ψ, ψ' and ψ'' are of $O_p(J)$, so that when we pass in $1/J$ under the root we make each term $O_p(1)$,

$$\hat{\sigma}_\theta^{\mathrm{MPL}} - \sigma_\theta^0 = \frac{-\frac{1}{J}\psi'_{nJ}(\sigma_\theta^0) \pm \sqrt{\frac{1}{J^2}\psi'_{nJ}(\sigma_\theta^0)^2 - \frac{1}{J^2}2\psi_{nJ}(\sigma_\theta^0)\psi''_{nJ}(\sigma_\theta^0) - \frac{1}{J^2}2\psi''_{nJ}(\sigma_\theta^0)o_p(J^{-1})}}{\frac{1}{J}\psi''_{nJ}(\sigma_\theta^0)}.$$

The difference $\sqrt{J}(\hat{\sigma}_\theta^{\mathrm{MPL}} - \sigma_\theta^0)$ will blow up unless we take the positive root so that the leading terms cancel. Using the expansions of ψ, ψ' and ψ'' and the expansion of the square root, we can reduce the numerator to

$$a_1 T_{nJ} + a_2 J^{-1} + a_3 T_{nJ}^2 + o_p\big(J^{-1}\big) \tag{B.1}$$

with some constants a_1, a_2, and a_3.

Similarly, Taylor expansion of the reciprocal of the denominator is written as

$$b_1 + b_2 T_{nJ} + o_p\big(J^{-1/2}\big) \tag{B.2}$$

with constants b_1 and b_2. Multiplication of (B.1) by (B.2) gives the bias up to the order of J^{-1} and it follows that

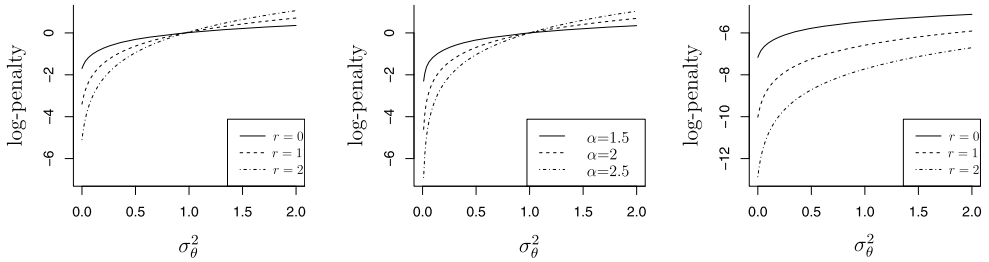
$$\hat{\sigma}_\theta^{\mathrm{MPL}} - \sigma_\theta^0 = \frac{1}{2\sigma_\theta^0}T_{nJ} - \frac{1}{8(\sigma_\theta^0)^3}T_{nJ}^2 + \frac{\lambda}{2J(\sigma_\theta^0)^3}\left(\frac{\sigma_\epsilon^2}{n} + \big(\sigma_\theta^0\big)^2\right)^2\left(\frac{\alpha-1}{\lambda} - \sigma_\theta^0\right) + O_p\big(J^{-3/2}\big).$$

Since $\hat{\sigma}_\theta^{\mathrm{MPL}}$ is uniformly integrable, the expectation of the above is

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

Appendix C. Proof of Equation (9)

The model in (2) can be written as $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where X is a covariate matrix, $\boldsymbol{\epsilon}$ follows $N(0, V)$, V is a block-diagonal matrix with $n \times n$ blocks V_j, and each V_j contains $\sigma_\theta^2 + \sigma_\epsilon^2$ on the diagonal and σ_θ^2 on the off-diagonals. As noted in Section 4.4, the REML log-likelihood can be written as the log-likelihood with an additive penalty term, $-\log\{\det(X^T V^{-1} X)\}/2$.



(a) REML penalty in (8) with $n = 30$, $\sigma_\epsilon = 1$ (b) log-gamma$(\alpha, 0)$ penalty (c) REML penalty for data with $n = 30, q = 1, J = 5$

FIGURE 8.
REML log-penalty function compared with log-gamma$((r+1)/2+1, 0)$ penalty. The shapes of the curves agree quite well, except when σ_θ^2 is close to 0 where the log-gamma penalty tends to 0.

The inverse of V is also block-diagonal of the same structure as V but with $\{\sigma_\epsilon^2 + (n_j - 1)\sigma_\theta^2\}/\sigma_\epsilon^2(\sigma_\epsilon^2 + n_j\sigma_\theta^2)$ in the diagonals and $-\sigma_\theta^2/\sigma_\epsilon^2(\sigma_\epsilon^2 + n_j\sigma_\theta^2)$ in the off-diagonals.

Let the columns of X consist of a vector of ones, q level-1 covariates $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q)$ and r level-2 covariates $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r)$. When we assume that $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r$ are dummy variables for the first r groups and $\boldsymbol{z}_i^T\boldsymbol{z}_i = 1$ and $\boldsymbol{z}_i^T\boldsymbol{z}_j = 0$ for all $i \neq j$ and the data are balanced, $X^TV^{-1}X$ can be simplified to a block-diagonal with

$$\frac{1}{\sigma_\epsilon^2(\sigma_\epsilon^2 + n_j\sigma_\theta^2)}\begin{bmatrix} Jn_j\sigma_\epsilon^2 & n\sigma_\epsilon^2\mathbf{1}_{1\times r} \\ n_j\sigma_\epsilon^2\mathbf{1}_{r\times 1} & n_j\sigma_\epsilon^2 I_r \end{bmatrix}$$

and $\frac{J}{\sigma_\epsilon^2}I_{q\times q}$.

Therefore it follows that

$$\det\left(X^TV^{-1}X\right) = \left(\frac{n}{\sigma_\epsilon^2 + n\sigma_\theta^2}\right)^{(r+1)}(J-r)\left(\frac{J}{\sigma_\epsilon^2}\right)^q$$

and

$$-\frac{1}{2}\log\left\{\det\left(X^TV^{-1}X\right)\right\} = \frac{r+1}{2}\log\left(\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n}\right) + \text{constant}.$$

Appendix D. REML and Log-Gamma Penalty in General Cases (Referred in Section 4.4)

Figure 8 compares the REML penalty function in (9), the log of the gamma density with corresponding $\alpha = (r+1)/2 + 1$, and the REML penalty function in the second term of (8) for a dataset with $n = 30$, $J = 5$, $q = 1$, $r = 0$, 1, or 2, which does not have the form assumed when deriving (9). For evaluating the REML penalty term in (8), 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 $w_{1j} = j$ for all $j = 1, \ldots, J$ 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 Figures 8(a) and (c), the penalties differ by a constant which does not affect the mode, so formula (9) appears to hold more generally.

For Figures 8(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 $\alpha = 2$, 3, and 4, respectively, except where σ_θ^2 is close to zero. At $\sigma_\theta^2 = 0$, the log-gamma penalty 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

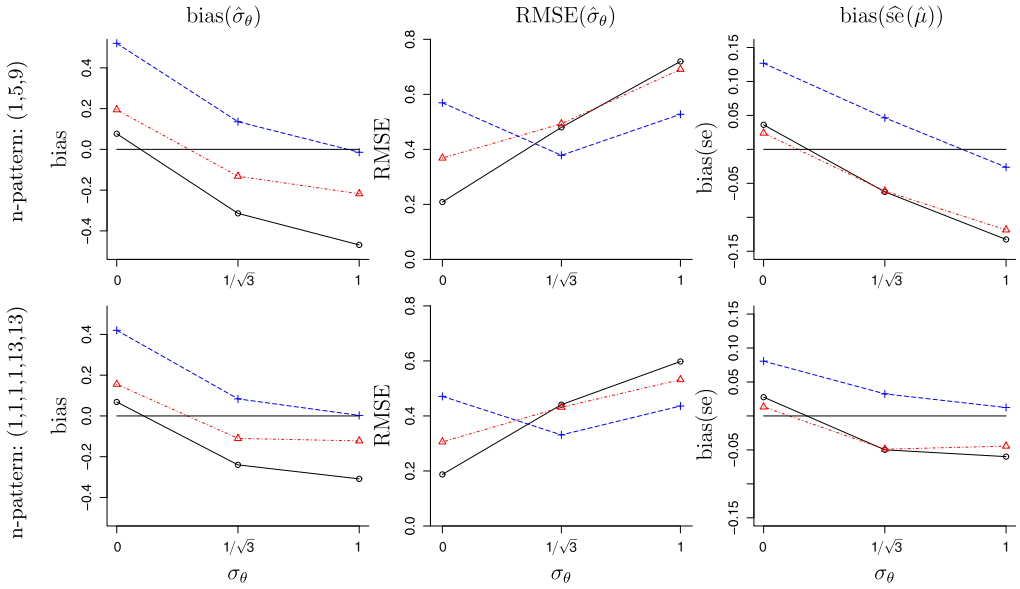


FIGURE 9.
Bias and RMSE of $\hat{\sigma}_\theta$, and bias of the standard error of $\hat{\mu}$. + is MPL, △ is REM and ∘ is ML.

that the log-gamma penalty 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 log-gamma penalty.

Appendix E. Simulation of Unbalanced Variance Component Model

Swallow and Monahan (1984) compared several variance estimation methods for the one-way model, given by

$$y_{ij} = \mu + \theta_j + \epsilon_{ij}, \quad i = 1, \ldots, n_j, \ j = 1, \ldots, J \tag{E.1}$$

where $\theta_j \sim N(0, \sigma_\theta^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. They considered unbalanced data with eight different patterns of group sizes $(n_1, \ldots, n_J)$, and compared the bias and RMSE of estimators of $\boldsymbol{\sigma}_\theta$ using simulated datasets.

In this appendix, we picked two of the patterns Swallow and Monahan (1984) considered, $(n_1, \ldots, n_J) = (1, 5, 9)$ and $(1, 1, 1, 1, 13, 13)$ with $\sigma_\epsilon = 1$, and compared ML and REML with the performance of the MPL estimates with log-gamma$(2, 0)$ penalty on σ_θ, which approximates the REML penalty for this model.

As for the balanced case in Section 6, both ML and REML tend to underestimate σ_θ for $\sigma_\theta > 0$. (See the left column of Figure 9.) On the other hand, MPL tends to overestimate σ_θ but the magnitude of the bias decreases as σ_θ increases. For $\sigma_\theta = 1$, the MPL estimator has the smallest bias for both patterns of group sizes. The RMSE is smallest for the MPL estimator when $\sigma_\theta > 0$ as shown in the middle column of Figure 9.

The last column in Figure 9 shows the estimated bias of the standard error of $\hat{\mu}$. When σ_θ is zero, there is almost no difference in the bias between the ML and REML estimators. As σ_θ increases, the bias for the MPL estimator becomes increasingly smaller than the bias for the other estimators.

References

Alderman, D., & Powers, D. (1980). The effects of special preparation on SAT-verbal scores. *American Educational Research Journal*, *17*(2), 239–251.

Bates, D., & Maechler, M. (2010). *lme4: Linear mixed-effects models using S4 classes*. R. package version 0.999375-37.

Bell, W. (1999). Accounting for uncertainty about variances in small area estimation. In *Bulletin of the International Statistical Institute, 52nd session, Helsinki*.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

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

Browne, W., & Draper, D. (2006). A comparison of Bayesian and likelihood methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514.

Ciuperca, G., Ridolfi, A., & Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Skandinavian Journal of Statistics*, *30*(1), 45–59.

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

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

Curcio, D., & Verde, P. (2011). Comment on: Efficacy and safety of tigecycline: a systematic review and meta-analysis. *Journal of Antimicrobial Chemotherapy*, *66*(12), 2893–2895.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188.

Dorie, V. (2013). *Mixed methods for mixed models: Bayesian point estimation and classical uncertainty measures in multilevel models*. PhD thesis, Columbia University.

Dorie, V., Liu, J., & Gelman, A. (2013). *Bridging between point estimation and Bayesian inference for generalized linear models* (Technical report). Department of Statistics, Columbia University.

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

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

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

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

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

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

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

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall/CRC.

Gelman, A., Jakulin, A., Pittau, M.G., & Su, Y.S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Gelman, A., & Meng, X. (1996). Model checking and model improvement. In *Markov chain Monte Carlo in practice* (pp. 189–201). London: Chapman & Hall.

Gelman, A., Shor, B., Bafumi, J., & Park, D. (2007). Rich state, poor state, red state, blue state: what's the matter with Connecticut? *Quarterly Journal of Political Science*, *2*(4), 345–367.

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

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

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

Harville, D.A. (1977). Maximum likelihood approaches to variance components estimation and related problems. *Journal of the American Statistical Association*, *72*(358), 320–338.

Higgins, J.P.T., Thompson, S.G., & Spiegelhalter, D.J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A*, *172*(1), 137–159.

Huber, P.J. (1967). The behavior of maximum likelihood estimation under nonstandard condition. In L.M. LeCam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 221–233). Berkeley: University of California Press.

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

Laird, N.M., & Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

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

Longford, N.T. (2000). On estimating standard errors in multilevel analysis. *Journal of the Royal Statistical Society. Series D*, *49*(3), 389–398.

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

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

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195.
Morris, C. (2006). Mixed model prediction and small area estimation (with discussions). *Test*, *15*(1), 72–76.
Morris, C., & Tang, R. (2011). Estimating random effects via adjustment for density maximization. *Statistical Science*, *26*(2), 271–287.
Neyman, J., & Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*(1), 1–32.
O'Hagan, A. (1976). On posterior joint and marginal modes. *Biometrika*, *63*(2), 329–333.
Overton, R. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods*, *3*(3), 354.
Patterson, H.D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.
Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata* (3rd ed.). College Station: Stata Press.
Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*(2), 301–323.
Raudenbush, S., & Bryk, A. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*(2), 75–98.
Rubin, D.B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, *6*(4), 377–401.
Self, S.G., & Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605–610.
Snijders, T., & Bosker, R. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational and Behavioral Statistics*, *18*(3), 237–259.
Stram, D.O., & Lee, J.W. (1994). Variance components testing in the logitudinal mixed effects model. *Biometrics*, *50*(4), 1171–1177.
Swallow, W., & Monahan, J. (1984). Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics*, *26*(1), 47–57.
Swaminathan, H., & Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*(3), 349–364.
Tsutakawa, R.K., & Lin, H.Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, *51*(2), 251–267.
Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Berlin: Springer.
Vermunt, J., & Magidson, J. (2005). *Technical guide for Latent Gold 4.0: basic and advanced* (Technical report). Statistical Innovations Inc., Belmont, Massachusetts.
Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293.
Warton, D.I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, *103*(481), 340–349.
Weiss, R.E. (2005). *Modeling longitudinal data*. New York: Springer.
Whaley, S., Sigman, M., Neumann, C.G., Bwibo, N.O., Guthrie, D., Weiss, R.E., Alber, S., & Murphy, S.P. (2003). Animal source foods improve dietary quality, micronutrient status, growth and cognitive function in Kenyan school children: background, study design and baseline findings. *The Journal of Nutrition*, *133*(11), 3965–3971.
White, H. (1990). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838.

Manuscript Received: 12 JUN 2012
Final Version Received: 8 OCT 2012