

A non-degenerate estimator for hierarchical variance parameters via  
penalized likelihood estimation

Yejin Chung\*, Sophia Rabe-Hesketh†

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

†Institute of Education, University of London

Andrew Gelman, Jingchen Liu and Vincent Dorie

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

February 10, 2012

**Abstract**

We propose a maximum penalized likelihood approach for estimating group-level variance parameters in mixed or multilevel models, equivalent to estimating variance parameters by their posterior mode, given a weakly informative prior distribution. By choosing the prior from the gamma family with shape parameter greater than 1, we ensure that the estimated variance will be positive. When the maximum likelihood estimate is at zero, our maximum penalized likelihood estimator is approximately one standard error from zero and thus remains consistent with the likelihood while being non-degenerate. (In contrast, lognormal and inverse-gamma penalty functions effectively bound the estimated variance parameter away from zero, which can result in regularized estimates that are inconsistent with the likelihood.) We also discuss the use of the gamma family to convey substantive prior information. In either case—pure penalization or prior information—our recommended procedure gives non-degenerate estimates when the number of groups is small and in the limit coincides with maximum likelihood as the number of groups increases.

**Keywords:** Hierarchical Linear Model, Multilevel Model, Penalized likelihood, Variance Estimation

---

<sup>§</sup>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).

# 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 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 variability is known to exist a priori due to omitted group-level covariates. Maximum likelihood is a useful way to estimate variance parameters in mixed models. But when the number of groups is small, estimates of group-level variance parameters can be noisy and can often be zero. In a multivariate setting, estimated covariance matrices can be singular. Estimating a variance to be zero causes underestimation of uncertainty in the parameter estimates of the model.

Most clearly, degenerate variance estimates lead to complete shrinkage of predictions for new and existing groups and yield estimated prediction standard errors that understate uncertainty. For example, Gelman et al. (2007) fit a multilevel model predicting voter choice given income, with the intercept and slope for income varying by state. They found that richer voters tended to support Republican candidates but with a slope that varied depending on some state-level predictors. For one election year, the fitted model had a zero value for the point estimate of the variance of the state-level errors for the slopes. In the resulting inferences, the state-level slopes were perfectly predicted by the state-level predictors. There is no reason to believe this—the perfect prediction is merely an artifact of a variance estimate that happened to be zero—and it is awkward to graph these results, showing an estimated perfect fit that we do not and should not believe. A related difficulty arises when comparing instances of a model that is repeatedly fit to similar data from different surveys or different years, yielding zero variance estimates some of the time.

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 above, discarding a variance component or setting the variance to zero understates uncertainty. 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.

Importantly, 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 (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).

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, 2000, 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.

Several authors (e.g., Kubokawa and Tsai, 2006; Srivastava and Kubokawa, 1999; Mathew and Niyogi, 1994; Kelly and Mathew, 1994) have developed nonnegative (definite) estimators

of variance components to improve standard estimation methods such as analysis of variance, maximum likelihood (ML) or restricted maximum likelihood (REML) in terms of Stein loss or mean-squared error (MSE). Our estimators are developed to avoid boundary estimates while respecting the data, and they turns out to perform as well as the standard methods such as ML or REML in terms of MSE.

## 1.1 Outline of our approach

In this paper we develop a non-degenerate estimator by maximizing the likelihood\* multiplied by 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.

Our primary aim is to develop a default penalization that gives non-degenerate variance estimators in multilevel models, bounded away from zero but automatically respecting the data. In particular, we recommend a class of gamma priors (for unidimensional problems) and Wishart (for multidimensional) that produce maximum penalized likelihood 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 prior has little influence when the number of groups is large or when the data are informative about the variance.

Penalized likelihood estimation has 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

---

\*We use the term likelihood to refer to the integrated or marginal likelihood, with random effects integrated out.

et al., 2004) and latent class analysis (Maris, 1999; Galindo-Garre and Vermunt, 2006). To our knowledge, this idea has not previously been applied to variance parameters in multilevel models.

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

Although our main objective is to avoid boundary estimates, we also compare the bias and MSE 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.

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

We begin by illustrating the boundary problem for a simple model in Section 1.2. In Section 2, we discuss Bayes modal estimation and propose a gamma prior as a weakly informative prior. Section 3 shows theoretical 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.

## 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:

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

In our simulation, we set the group-level standard deviation  $\sigma_\theta$  to 0.5. From this model, we create 1000 simulated datasets and estimate  $\sigma_\theta$  by maximum likelihood by solving for  $\hat{\sigma}_\theta$  in the equation  $1 + \hat{\sigma}_\theta^2 = \frac{1}{J} \sum_{j=1}^J y_j^2$ , with the boundary constraint that  $\hat{\sigma}_\theta = 0$  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 about here.]

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

Figure 1b displays 100 draws of the 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  $\sigma_\theta$  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  $\sigma_\theta$  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.

## 2 Maximum penalized likelihood estimation of $\sigma_\theta$

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

We consider the model

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

where  $y_{ij}$  is the response variable and  $\mathbf{x}_{ij}$  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  $\theta_j$  and  $\epsilon_{ij}$  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  $\theta_j$  and  $\boldsymbol{\beta}$  (Harville, 1974). Unlike the ML estimator, the REML estimator of  $\sigma_\theta^2$  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 Maximum penalized likelihood estimation

In the present article, we are particularly concerned with the group-level standard deviation, and we specify a penalty or a prior only for  $\sigma_\theta$ , implicitly assuming a uniform prior,  $p(\boldsymbol{\beta}, \sigma_\epsilon) = 1$ , on  $\boldsymbol{\beta}$  and  $\sigma_\epsilon$ .

The penalized log-likelihood function can be written as

$$\log l_p(\sigma_\theta, \boldsymbol{\beta}, \sigma_\epsilon; \mathbf{y}) = \log l(\sigma_\theta, \boldsymbol{\beta}, \sigma_\epsilon; \mathbf{y}) + \log p(\sigma_\theta) + c, \quad (2)$$

where the first term of the right hand side is the log-likelihood,  $\log p(\sigma_\theta)$  is an additive penalty term or log prior, and  $c$  is a constant. We find the maximum penalized likelihood (MPL) estimator that attains the maximum of (2). The penalized log-likelihood can be regarded as the marginal log-posterior density with varying intercepts ( $\theta_j$ ) integrated out and MPL estimates are equivalent to posterior modal estimates. By integrating the posterior over  $\theta_j$ , we avoid the incidental parameter problem (Neyman and Scott, 1948; O’Hagan, 1976; Mislevy, 1986).

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 maximum penalized likelihood estimation for `gllamm` (Rabe-Hesketh et al., 2005; Rabe-Hesketh and Skrondal, 2008) in Stata and for `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](http://www.gllamm.org) and the modified `lmer` can be found in the `blme` package available from the Comprehensive R Archive Network.)

### 2.3 Desired properties of a weakly informative prior

Our goal is to find a penalty or a prior for  $\sigma_\theta$  so that MPL estimates are off the boundary, but with the penalty being weak enough so that inferences are consistent with the data. For our purpose, we desire  $p(\sigma_\theta)$  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 a prior is linear at zero and there is no “dead zone” in the penalty near zero—that is, the penalty does not rule out positive values near zero if they are supported by the likelihood.

For our default choice of penalty, 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 the posterior mode. 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 non-degenerate 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  $\sigma_\theta$  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.

## 2.4 Gamma penalty function

We propose the logarithm of a gamma (not inverse-gamma) density as a penalty function of  $\sigma_\theta$  or equivalently, assign a gamma prior on  $\sigma_\theta$  : defined by

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

with mean  $\alpha/\lambda$  and variance  $\alpha/\lambda^2$ , where  $\alpha$  is the shape parameter and  $\lambda$  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  $\alpha > 1$ ,  $\text{gamma}(\alpha, \lambda)$  satisfies the first condition that  $p(0) = 0$ . In order to have a positive constant derivative at zero (the second condition),  $\alpha$  can be chosen to be 2.

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

- Our *default choice* is  $\text{gamma}(\alpha, \lambda)$  with  $\alpha = 2$  and  $\lambda \rightarrow 0$ , which is the (improper) density ( $p(\sigma_\theta) \propto \sigma_\theta$ ). As we discuss shortly, this default bounds the MPL estimate 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( $\alpha, \lambda$ ) with  $1/\lambda$  set to the prior estimate of  $\sigma_\theta$ .

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

### 3 Theoretical properties

#### 3.1 Difference between ML estimator and MPL estimator

To examine the effect of  $\alpha$  and  $\lambda$  on the MPL estimator 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  $\sigma_\theta$  around the ML estimator,  $\hat{\sigma}_\theta^{\text{ML}}$ ,

$$\log L(\sigma_\theta) \approx -\frac{(\sigma_\theta - \hat{\sigma}_\theta^{\text{ML}})^2}{2 \cdot \widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})^2} + c_1. \quad (4)$$

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

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

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

where

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

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

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

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

When we have boundary estimates of  $\sigma_\theta$ , it is possible that the log-likelihood function of  $\sigma_\theta^2$  has its 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  $\sigma_\theta^2$  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  $\sigma_\theta$  at 0, the linear term vanishes, the leading term becomes the quadratic (with negative coefficient when  $\hat{\sigma}_\theta = 0$ ) 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  $\sigma_\theta$ , we derive a number of properties of the log-gamma( $\alpha, \lambda$ ) penalty of  $\sigma_\theta$ . (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  $\hat{\sigma}_\theta^{\text{ML}} = 0$  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  $\lambda \rightarrow 0$  with the value*

$$\hat{\sigma}_\theta = \widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}}) \sqrt{\alpha - 1}. \quad (6)$$

When  $\alpha = 2$ , we obtain  $\hat{\sigma}_\theta = \widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$ . 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 (4) 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  $2(\log L(0) - \log L(\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}}))) = 1$ . For the null hypothesis  $\sigma_\theta = 0$ , it is known that 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 and 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, a 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  $\lambda \rightarrow 0$  with the value*

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

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

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  $\partial\hat{\sigma}_\theta/\partial\widehat{\text{se}}(\hat{\sigma}_\theta^{\text{ML}})$  has maximum  $\sqrt{\alpha - 1}$  for  $\hat{\sigma}_\theta^{\text{ML}} = 0$  that coincides with (6) and decreases as  $\hat{\sigma}_\theta^{\text{ML}}$  increases. This implies that the log-gamma( $\alpha$ ,  $\lambda$ ) 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  $\lambda$  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.

### 3.2 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  $J \rightarrow \infty$  and compare them with the asymptotic properties of the ML estimator.

Consider a balanced varying-intercept model with  $\mathbf{x}_{ij}^T \boldsymbol{\beta} = \mu$  and  $n_i = n$ . For simplicity, we assume that  $\mu$  and  $\sigma_\epsilon^2$  are known. Then the ML estimator of  $\sigma_\theta$  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 log-gamma( $\alpha, \lambda$ ) penalty is applied to  $\sigma_\theta$ , the MPL estimator, say  $\hat{\sigma}_\theta^{\text{MPL}}$ , is a root of a fifth order polynomial (See the supplementary materials B). Therefore, we do not have a simple formula for  $\hat{\sigma}_\theta^{\text{MPL}}$  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  $\sigma_\theta$ , it is sufficient to assume only  $J \rightarrow \infty$  regardless of  $n$ . As  $J \rightarrow \infty$ ,  $\hat{\sigma}_\theta^{\text{ML}}$  is consistent with  $\sigma_\theta^0$  and  $\sqrt{J} (\hat{\sigma}_\theta^{\text{ML}} - \sigma_\theta^0)$  follows  $N(0, I(\sigma_\theta^0)^{-1})$  asymptotically where  $I(\sigma_\theta^0)$  is the information matrix and  $\sigma_\theta^0$  is the true value of  $\sigma_\theta$ .

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  $J \rightarrow \infty$ ,

$$\sqrt{J}(\hat{\sigma}_\theta^{\text{MPL}} - \sigma_\theta^0) \rightarrow N(0, I(\sigma_\theta^0)^{-1}).$$

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

**Theorem 3.** *At the order of  $J^{-1}$ , the ML estimator and the MPL estimator have the fol-*

lowing bias.

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

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

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

*Proof.* An outline of the proof is in the supplementary materials and Dorie (2012).  $\square$

The MPL estimator of  $\sigma_\theta^{\text{MPL}}$  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 3. In addition, the MPL estimator tend to be less biased for small  $J$  as will be shown using simulations in Section 5.

### 3.3 Transformation of $\sigma_\theta$

In the Bayesian point of view, when the posterior density of  $\sigma_\theta$  is asymmetric, a transformation of  $\sigma_\theta$  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 posterior modal estimator is not due to the change in prior density when transforming  $\sigma_\theta$ . 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( $\alpha, \lambda$ ) prior on  $\sigma_\theta$ , maximizing the posterior of  $g_\gamma(\sigma_\theta)$  is equivalent to maximizing the posterior of  $\sigma_\theta$  with a gamma( $\alpha + 1 - \gamma, \lambda$ ) prior on  $\sigma_\theta$ .

For example, consider a special case with  $\alpha = 1$ ,  $\lambda \rightarrow 0$ , and  $\gamma = 0$ , which implies the uniform (improper) prior on  $\sigma_\theta$  and log transformation of  $\sigma_\theta$ . 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  $\sigma_\theta$  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  $\text{gamma}(2, \lambda)$  prior on  $\sigma_\theta$ .

We have discussed the gamma prior on the group-level standard deviation ( $\sigma_\theta$ ) since the profile log-likelihood as a function of  $\sigma_\theta$  has a better quadratic approximation so it helps us to investigate the properties in Section 3.1. However, one might still be interested in priors on the variance,  $\sigma_\theta^2$ .

**Property 5.** *In the limit  $\lambda \rightarrow 0$ , a  $\text{gamma}(\alpha, \lambda)$  prior on  $\sigma_\theta^2$  is equivalent to a  $\text{gamma}(2\alpha - 1, \lambda)$  prior on  $\sigma_\theta$ .*

Therefore, the properties of the gamma prior in this paper hold for the gamma prior on  $\sigma_\theta^2$  with  $\alpha$  adjusted appropriately.

### 3.4 Connection to REML

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

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

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

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

In order to compare the REML penalty and log-gamma penalty, 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  $\mathbf{z}_1, \dots, \mathbf{z}_q$  of the design matrix, consist of the same elements for each group and satisfy  $\mathbf{1}^T \mathbf{z}_u = 0$ ,  $\mathbf{z}_u^T \mathbf{z}_{u'} = 1$  if  $u = u'$ , and 0 otherwise for  $u = 1, \dots, q$ . The level-2 covariates are assumed to be dummy variables for the first  $r (< J - q - 2)$  groups. Then the REML penalty becomes

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

where  $c_1$  is a constant. The proof is provided in the supplementary materials.

Recall that, when  $\lambda \rightarrow 0$ , the  $\text{gamma}(\alpha, \lambda)$  prior on  $\sigma_\theta^2$  (equivalently  $\text{gamma}(2\alpha - 1, \lambda)$  on  $\sigma_\theta$ ) has log density,

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

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

The difference between these two penalty terms is clear when  $\sigma_\theta$  is close to zero. At  $\sigma_\theta = 0$ , the log-gamma penalty term in (9) is  $-\infty$  for  $\alpha > 1$ , whereas the REML penalty in (8) approaches  $-\infty$  only if  $\sigma_\epsilon \rightarrow 0$  or  $n \rightarrow \infty$ . This explains why REML can produce

boundary estimates. Further, it implies that the log-gamma penalty assigns more penalty on  $\sigma_\theta$  close to zero than REML for small  $n$  and large  $\sigma_\epsilon$ . Otherwise, REML can approximately be viewed as a special case of our method with a log-gamma penalty.

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 (DerSimonian 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

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

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

[Figure 2 about here.]

Figure 2 shows the profile log-likelihood (maximized with respect to  $\mu$ ) of  $\sigma_\theta$  (left) and

$\sigma_\theta^2$  (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 3.1. Further, the profile log-likelihood is quite flat. We see in the middle panel of Figure 2 that the profile log-likelihood has a negative gradient at zero as a function of  $\sigma_\theta^2$  so that the quadratic approximation for  $\sigma_\theta^2$  is poor at the maximum likelihood estimate of zero.

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

Inference for  $\sigma_\theta$  is also important because it affects both the point estimate and estimated standard error of the overall effect size  $\mu$ ,

$$\widehat{\text{se}}(\hat{\mu}) = \left[ \sum_i \frac{1}{s_i^2 + \hat{\sigma}_\theta^2} \right]^{-1/2}. \quad (11)$$

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  $\mu$  are 7.7 and 8.1, respectively.)

[Table 1 about here.]

For the model in (10), we consider four different penalties: log-gamma(2,  $\lambda$ ) and log-gamma(3,  $\lambda$ ) on  $\sigma_\theta$  and log-gamma(1.5,  $\lambda$ ) and log-gamma(2,  $\lambda$ ) on  $\sigma_\theta^2$ , where  $\lambda = 10^{-4}$ . MPL estimates with these penalties and ML estimates are given in Table 1. 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 3.1).

When the penalty is assigned on  $\sigma_\theta$  (rows 2 and 3),  $\hat{\sigma}_\theta^{\text{MPL}}$  is 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 quadratic in  $\sigma_\theta$ , as it appears to be in the left panel of Figure 2. In both cases, the log-likelihood at the MPL estimate is only a little bit lower than the maximum log-likelihood.

Specifying a log-gamma(2,  $\lambda$ ) penalty on  $\sigma_\theta^2$  (row 5) gives estimates that agree well with those for a log-gamma(3,  $\lambda$ ) penalty on  $\sigma_\theta$  as expected (see Property 5). Similarly, a log-gamma(1.5,  $\lambda$ ) on  $\sigma_\theta^2$  (row 4) gives MPL estimates that are close to the estimates with log-gamma(2,  $\lambda$ ) on  $\sigma_\theta$ . A log-gamma penalty on  $\sigma_\theta^2$  with  $\alpha = 1.5$  corresponds to REML with no level-2 covariates. While REML gives  $\hat{\sigma}_\theta = 0$  (not shown here), a log-gamma penalty 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  $\hat{\mu}$  ( $\text{se}^R$ ). We see that the estimated model-based standard error of the estimated overall effect size  $\mu$  increases with  $\sigma_\theta$  as implied by (11), whereas the robust standard errors, based on the sandwich estimator, change very little.

## 5 Simulation study: balanced varying-intercept model

We consider a varying-intercept model,

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

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 generated 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 MPL estimates with  $\log\text{-gamma}(2, \lambda)$  and  $\log\text{-gamma}(3, \lambda)$  penalties on  $\sigma_\theta$ , where  $\lambda = 10^{-4}$ . 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  $\sigma_\theta$  that are on the boundary (less than  $10^{-5}$ ) when the true  $\sigma_\theta$  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 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  $\sigma_\theta$ .

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  $-2 [\log L(\hat{\sigma}_\theta^{\text{MPL}}) - \log L(\hat{\sigma}_\theta^{\text{ML}})]$  for testing the restriction  $\sigma_\theta = \hat{\sigma}_\theta^{\text{MPL}}$  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

[Figure 3 about here.]

We now assess how well some of the relationships hold that were derived in Section 3.1 by assuming that the profile log-likelihood is quadratic. Figure 3 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, $\lambda$ ) penalty on  $\sigma_\theta$  for  $J = 3$  (left) and  $J = 30$  (right) when  $\rho = 0.25$  and  $n = 30$ .

[Figure 4 about here.]

Figure 4 summarizes the estimated bias and the root mean squared error (RMSE) of  $\sigma_\theta$ , and the coverage of 95% confidence intervals (CI) for  $\beta_2$  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  $\sigma_\theta$  decreases as  $J$  increases and  $\sigma_\theta$  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  $\sigma_\theta$ . MPL estimates with a log-gamma(2, $\lambda$ ) penalty also tend to be downward biased for  $\sigma_\theta$  but not as much as the ML estimates. On the other hand, the MPL estimator with log-gamma(3, $\lambda$ ) produces the largest estimates among the four estimators so it often overestimates  $\sigma_\theta$ . For  $\sigma_\theta = 1$ , the MPL estimator with log-gamma(3, $\lambda$ ) 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.

When  $n = 30$  (given in the supplementary materials), the overall pattern is the same as when  $n = 5$  but the MPL with log-gamma(3, $\lambda$ ) are closer to REML for  $\sigma_\theta > 0$ . This confirms that the log-gamma penalty on  $\sigma_\theta$  with  $\alpha = 3$  agrees with the REML penalty

when the model contains one group-level covariate, particularly with large  $n$ , as discussed in Section 3.4.

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

**Coverage of CI for  $\beta_2$**  The standard error estimates of the estimated coefficient of the group-level covariate ( $\hat{\beta}_2$ ) is greatly influenced by  $\hat{\sigma}_\theta$ . The squared asymptotic standard error of  $\hat{\beta}_2$  from the Hessian matrix is  $\text{Var}(\hat{\beta}_2) \approx (n\sigma_\theta^2 + \sigma_\epsilon^2)/nJs_{X_2}^2$  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  $\hat{\sigma}_\theta$  is on the boundary, the standard error of  $\hat{\beta}_2$  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  $\beta_2$ . The gray solid line shows the nominal coverage (0.95). For all values of  $\sigma_\theta$ , 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 MPL estimator with  $\alpha = 3$  has the best coverage, particularly for  $\sigma_\theta = 1/\sqrt{3}$ . Although the MPL estimator with  $\alpha = 3$  tends to have large positive bias for  $\sigma_\theta$ , it turns out to give better coverage. Recalling that  $\text{log-gamma}(3, \lambda)$  is close to the REML penalty (discussed in Section 3.4) for large  $n$ , we found that the coverage for the MPL 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 MPL estimator, particularly for small  $J$ .

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  $\sigma_\theta$  is not zero, RMSE of  $\hat{\sigma}_\theta$  is uniformly lower for the MPL estimator with both log-gamma penalties than for REML and the ML estimator and coverage of the CI for the fixed coefficient for the group-level covariate is best for the MPL estimator with  $\alpha = 3$ . Although there is no obvious winner between gamma with  $\alpha = 2$  and  $\alpha = 3$ , neither penalty ever produces a boundary estimate ( $\hat{\sigma}_\theta < 10^{-5}$ ).

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 MPL estimates with a gamma(2,  $\lambda$ ) 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  $\sigma_\theta$  is not zero, ML and REML tend to underestimate  $\sigma_\theta$  and the RMSE tends to be larger than for the MPL 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  $\sigma_\theta$  is not zero while the MPL estimators perform better in this regard.

## 6 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 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 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 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, 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.

Our idea can also be applied to models with varying intercepts and slopes where the problem is to regularize the covariance matrix, say  $\Sigma$ , away from its boundary,  $|\Sigma| = 0$ . In this case, the gamma prior can be naturally extended to the Wishart prior on  $\Sigma$ , 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  $\Sigma$  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  $\text{gamma}(2, \lambda)$  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.

### Supporting Information

A supplementary material contains derivation of properties in Section 3, proofs of Theorem 3 and equation (7), REML and gamma priors in general cases, and additional simulation results.

### 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 Eigen and Eigen* (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).
- DerSimonian, R. and Laird, N. “Meta-analysis in clinical trials.” *Controlled Clinical Trials*, 7:177–188 (1986).

- Dorie, V. “Mixed Methods for Mixed Models: Bayesian Point Estimation and Classical Uncertainty Measures in Multilevel Models.” Ph.D. thesis, Columbia University (2012).
- 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).
- Gelman, A., Shor, B., Bafumi, J., and Park, D. “Rich state, poor state, red state, blue state: What’s the matter with Connecticut?” *Quarterly Journal of Political Science*, 2(4):345–367 (2007).
- 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).
- Kelly, R. and Mathew, T. “Improved nonnegative estimation of variance components in some mixed models with unbalanced data.” *Technometrics*, 171–181 (1994).
- Kenward, M. and Roger, J. H. “Small-sample inference for fixed effects from restricted maximum likelihood.” *Biometrics*, 53(3):983–997 (1997).
- Kubokawa, T. and Tsai, M. “Estimation of covariance matrices in fixed and mixed effects linear models.” *Journal of Multivariate Analysis*, 97(10):2242–2261 (2006).
- Laird, N. M. and Ware, J. H. “Random effects models for longitudinal data.” *Biometrics*, 38:963–974 (1982).
- Longford, N. T. “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).
- Mathew, T. and Niyogi, A. “Improved nonnegative estimation of variance components in balanced multivariate mixed models.” *Journal of Multivariate Analysis* (1994).
- 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).

- Raudenbush, S. and Bryk, A. “Empirical Bayes meta-analysis.” *Journal of Educational and Behavioral Statistics*, 10(2):75 (1985).
- 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).
- Srivastava, M. and Kubokawa, T. “Improved nonnegative estimation of multivariate components of variance.” *Annals of Statistics*, 2008–2032 (1999).
- 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 for Longitudinal Data*. Springer Verlag (2000).
- White, H. “A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity.” *Econometrica*, 48:817–838 (1990).

Yejin Chung, Graduate School of Education, University of California, Berkeley, 4623 Tolman Hall, Berkeley, CA 94720  
E-mail: ychung@berkeley.edu

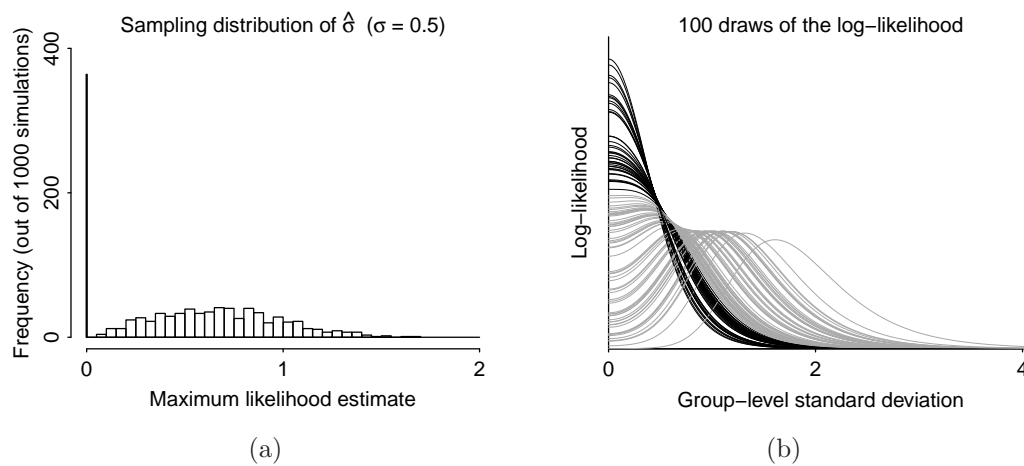


Figure 1: *From a simple one-dimensional hierarchical model with scale parameter 0.5 and data in 10 groups: (a) Sampling distribution of the maximum likelihood estimate  $\hat{\sigma}_\theta$ , based on 1000 simulations of data from the model. (b) 100 simulations of the log-likelihood. The dark lines are the log-likelihoods with maximum at 0 and the grey lines are the others. The maximum likelihood estimate is extremely variable and the likelihood function is not very informative about  $\sigma_\theta$ .*

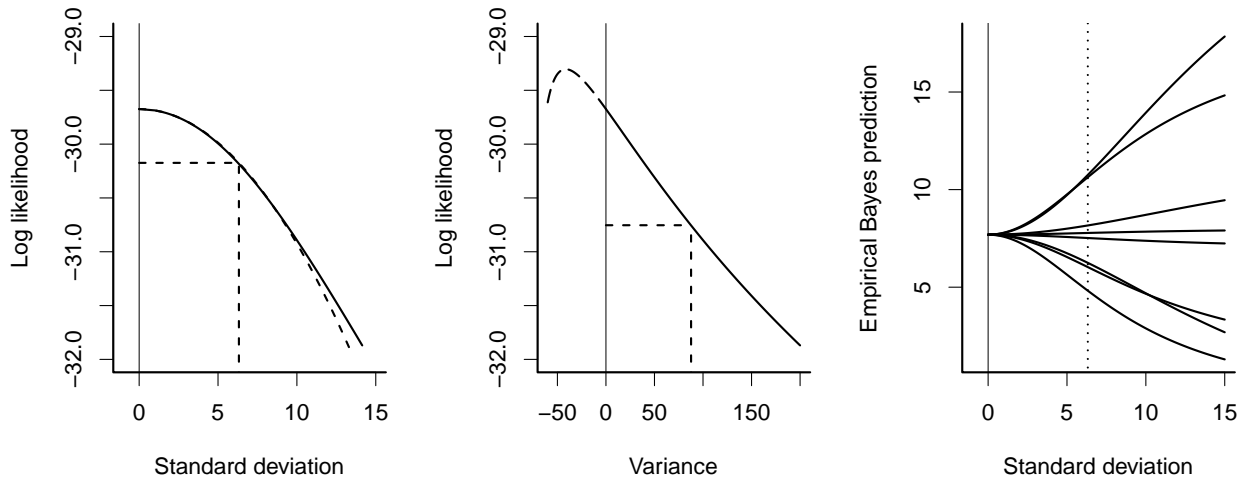


Figure 2: Profile log-likelihood as a function of  $\sigma_\theta$  (left) and  $\sigma_\theta^2$  (middle) and empirical Bayes prediction (right) for 8-schools data. The dashed curve on the left is the quadratic approximation at the mode, based on the estimated standard error. The vertical dashed line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

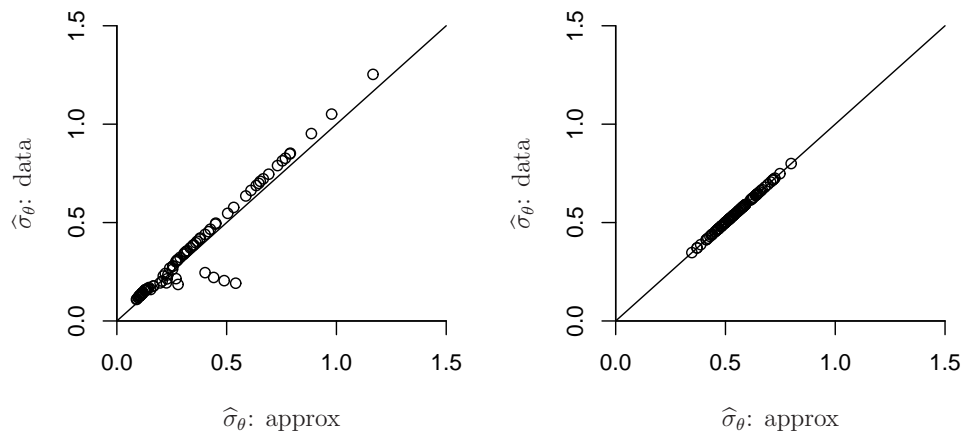


Figure 3: *MPL estimates with log-gamma(2, $\lambda$ ) penalty on  $\sigma_\theta$  for  $J = 3$  (left) and  $J = 30$  (right),  $\rho = 0.25$  and  $n = 30$  for the first 100 replicates, compared with the MPL estimates based on the quadratic approximation of the profile log-likelihood (see properties 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.*

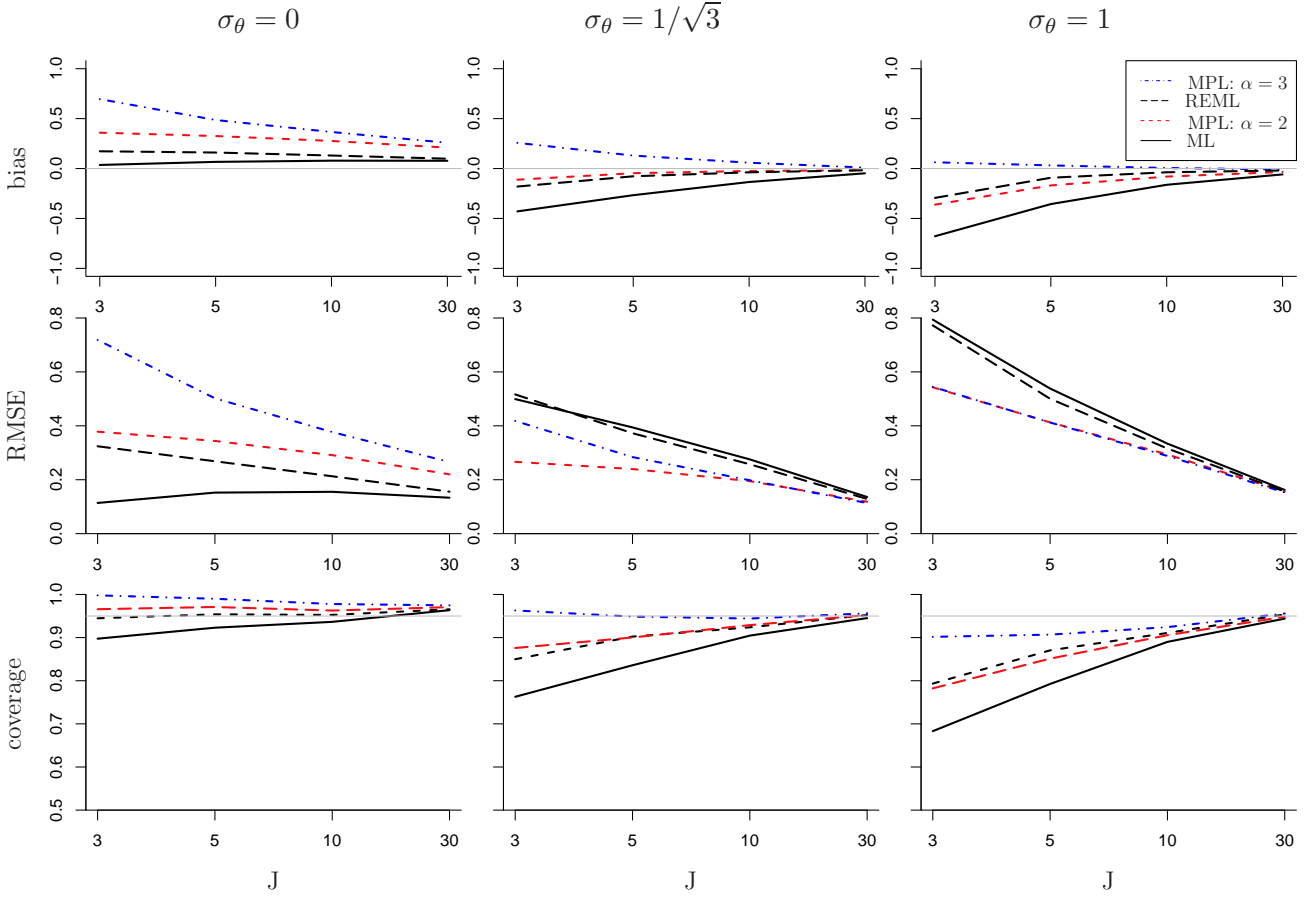


Figure 4: Bias of  $\sigma_\theta$ , RMSE of  $\sigma_\theta$  and coverage of CI for  $\beta_2$  for group size 5, standard deviation  $\sigma_\theta = 0, 1/\sqrt{3}$ , and 1 (columns) and number of groups  $J = 3, 5, 10, 30$  (x-axis). Different estimators are represented by different line patterns. When  $\sigma_\theta > 0$ , all the methods outperform ML. Bias of the MPL estimator is as low as REML depending on  $\alpha$ . RMSE of the MPL estimator with both  $\alpha$  is smaller than REML and ML. Coverage of CI is best for the MPL estimator with  $\alpha = 3$ .

Table 1: ML and MPL estimates for the 8 schools data, where the penalty is  $\log\text{-gamma}(\alpha, \lambda)$  on  $\sigma_\theta$  or  $\sigma_\theta^2$ , with  $\lambda = 10^{-4}$ . With  $\log\text{-gamma}(\alpha, \lambda)$  penalty on  $\sigma_\theta$ , the MPL estimates are approximately at  $\widehat{\text{se}}(\widehat{\sigma}_\theta^{\text{ML}})\sqrt{1-\alpha}$  and agree well with the MPL estimates with  $\log\text{-gamma}((\alpha+1)/2, \lambda)$  on  $\sigma_\theta^2$ .

Method	$\mu$			$\sigma_\theta$		Log-lik
	est	se	se <sup>R</sup>	est	se	
ML	7.69	4.07	3.33	0	6.32	-29.67
MPL: $\text{gamma}(2, \lambda)$ on $\sigma_\theta$	7.92	4.72	3.39	6.30	4.61	-30.18
MPL: $\text{gamma}(3, \lambda)$ on $\sigma_\theta$	8.10	5.38	3.43	9.42	5.34	-30.76
MPL: $\text{gamma}(1.5, \lambda)$ on $\sigma_\theta^2$	7.92	4.72	3.38	6.28	4.79	-30.18
MPL: $\text{gamma}(2, \lambda)$ on $\sigma_\theta^2$	8.09	5.37	3.42	9.37	5.30	-30.75

se<sup>R</sup>: robust (sandwich) standard error.