Abstract Title Page

Not included in page count.

Title: Avoiding Boundary Estimates in Hierarchical Linear Models Through Weakly Informative Priors

Authors and Affiliations:
Yeojin Chung (Graduate School of Education, University of California, Berkeley)
Sophia Rabe-Hesketh (Graduate School of Education, University of California, Berkeley, Institute of Education, University of London)
Andrew Gelman (Department of Statistics, Columbia University, New York)
Vincent Dorie (Department of Statistics, Columbia University, New York)
Jinchen Liu (Department of Statistics, Columbia University, New York)

Abstract Body

Background / Context:

Hierarchical or multilevel linear models (e.g., Raudenbush and Bryk, 2002), are widely used for longitudinal or cross-sectional data on students nested in classes and schools, and are particularly important for estimating treatment effects in cluster-randomized trials, multi-site trials, and meta-analyses. The models can allow for variation in treatment effects, as well as examination of the reasons for treatment effect variation. For example, random-effects meta-analysis (DerSimion and Laird, 1986) allows the treatment effect to vary between studies to accommodate differences in populations, treatment implementation, and measurement of outcomes. Meta-regression can then be used to investigate the sources of treatment-effect heterogeneity. In hierarchical linear models for cluster-randomized trials, differential effectiveness of treatments for different subpopulations, such as English language learner (ELL) and non-ELL students, can be investigated by including cross-level interactions between indicators for ELL status (level 1) and intervention group (level 2). Such models typically allow for residual between-school heterogeneity in the gap between ELL and non-ELLstudents, that is not explained by the intervention, by including a random coefficient of ELL status.

A practical problem often encountered when using these methods is that the number of groups (studies in meta-analysis, schools in cluster-randomized trials, and sites in multi-site trials)   is small and that cluster-level variance parameters are estimated as zero.

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

A second problem with boundary estimates is the resulting underestimation of uncertainty in fixed coefficient estimates. For instance, in a cluster-randomized study or meta-analysis, researchers might be overconfident in concluding that a treatment is effective. Similarly, overconfident conclusions regarding differential effectiveness of interventions for different subpopulations can result when variances of random coefficient are estimated as zero.

Third, group comparisons are often of interest to researchers, but when the group-level variance is estimated as zero, the resulting predictions of the group-level errors will all be zero, so one fails to find unexplained differences between groups.

Purpose / Objective / Research Question / Focus of Study:

We propose a method that pulls the group-level standard deviation estimate off the boundary while producing estimates that are consistent with the data. The idea is to specify a weakly informative prior distribution for the standard deviation and to maximize the resulting posterior distribution, a method that can also be viewed as penalized maximum likelihood estimation.

Significance / Novelty of study:

Bayes modal estimation has previously been used to obtain more stable estimates of item parameters in item response theory (e.g., Mislevy, 1986) and to avoid boundary estimates in log-linear and latent class analysis (Maris, 1999; Galindo-Garre and Vermunt, 2006). To our knowledge, this idea has not yet been applied to variance parameters in hierarchical models.

Statistical, Measurement, or Econometric Model:

For subject *i* in group *j*, we consider the model

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

where y_{ij} is the response variable, $\mathbf{x}_{ij}$ a p-dimensional vector of explanatory variables with regression coefficients β, $\theta_j \sim N(0, \sigma_\theta{}^2)$ is a group-level random intercept, and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon{}^2)$ is a residual. We further assume that θ_j and ε_{ij} are independent.

We specify a prior $p(\sigma_\theta)$ only for σ_θ, implicitly assuming a uniform prior, $p(\beta, \sigma_\varepsilon)$, on β and σ_ε.We find the parameters that maximize the marginal log-posterior density (with random intercepts integrated out). The marginal posterior density for $(\beta, \sigma_\theta, \sigma_\varepsilon)$ can equivalently be regarded as a penalized likelihood.

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

$$p(\sigma_\theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \sigma_\theta^{\alpha-1} e^{-\lambda \sigma_\theta}$$

with mean α / λ and variance α / λ^2, where α is the shape parameter and λ is the rate parameter.

For any $\alpha > 1$, the prior is zero at the origin and this ensures a positive estimate of the variance parameter, even when the maximum of the likelihood is at 0. If α is 2, the prior allows the likelihood to dominate if it is strongly curved near zero since the prior has a positive constant derivative at zero. Our *default choice* is gamma (α, λ) with $\alpha = 2$ and $\lambda \to 0$, which is the (improper) density $p(\sigma_\theta) \propto \sigma_\theta$. This default bounds the posterior mode away from zero while keeping it consistent with the likelihood.

Usefulness / Applicability of Method:

Theoretical results

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

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

Here $se_{ml} = se(\hat{\sigma}_\theta^{ml})$ represents the estimated asymptotic standard error of σ_θ (based on the observed information).

- With the default prior, the Bayes modal estimate is $\hat{\sigma}_\theta^{Bayes} = se_{ml}$ if the ML estimate is at the boundary. That is, the prior shifts the posterior mode away from zero but only by about one standard error. The resulting change in the log-likelihood is about 1.
- When the ML estimate is not at the boundary and the default prior is used, the difference between the Bayes modal and ML estimates is less than one standard error.
- When the posterior density is asymmetric, a transformation of σ_θ can make the density more symmetric so that the posterior mode will be located near the posterior mean which has good asymptotic properties. With a gamma prior on σ_θ, maximizing the posterior of a Box-Cox transformed σ_θ is equivalent to maximizing the posterior of σ_θ with a gamma prior with an adjusted value of α.
- A gamma prior (α, λ) on σ_θ^2 is equivalent to gamma $(2\alpha - 1, \lambda)$ prior on σ_θ.
- In a model with *r* group-level covariates, the gamma $((r+1)/2+1, \lambda)$ on σ_θ^2 (equivalently gamma $(r+2, \lambda)$ on σ_θ) approximately matches the restricted maximum likelihood (REML) penalty, particularly when the group-size *n* is large and λ is close to zero.

Data Analysis

Rubin (1981) analyzed 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).

The model for the estimated effect size y_i of study *i* can be written as

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

The ML estimate of the between-study standard deviation is at the boundary, $\hat{\sigma}_\theta^{ml} = 0$. With the default prior, the between-study standard deviation is estimated as $\hat{\sigma}_\theta^{Bayes} = 6.30$, close to the value $se_{ml} = 6.32$ that we expect based on the quadratic approximation of the profile likelihood. At the Bayes modal estimate, the log-likelihood is -30.18, only a little bit lower than the value -29.67 of the log-likelihood at the maximum likelihood estimate. Therefore, the Bayes modal estimate is consistent with the data.

Importantly, accepting the ML boundary estimate instead of using Bayes modal estimation, would lead to a much narrower estimated confidence interval (CI) for the main parameter estimate of interest, the overall effect size μ. Using ML gives an estimated 95% CI for μ from -0.3 to 15.7, compared with the 95% CI based on the Bayes modal estimate from -1.3 to 17.2. Using maximum likelihood estimates with robust standard errors (sandwich estimator), gives an estimated confidence interval from 1.2 to 14.2, even narrower than the interval using the model-based standard error.

Simulations

Simulations have been performed for model (1) with one covariate that varies only within groups (group-mean covariate constant across groups) and one covariate that varies only between groups. The number of groups was set to with *J*=3, 5, 10, 30, the group size (constant across groups) to *n*=5, 30, and the residual intraclass correlation to ρ=0, 0.25, and 0.5. For each combination of *J, n*, and ρ, 1000 datasets were generated. For each dataset, we obtained ML, REML, and posterior mode estimates with gamma$(2,\lambda)$ and gamma$(3,\lambda)$ priors on σ_θ, where $\lambda = 10^{-4}$. The REML penalty corresponds to $\alpha = 3$ since the model contains one group-level covariate.

When $\rho > 0$, the bias of $\hat{\sigma}_\theta$ is as low for the Bayes modal estimators as for REML depending on α. The RMSE of $\hat{\sigma}_\theta$ is uniformly lower for the Bayes modal estimator with both gamma priors than for the REML and the ML estimators. Coverage of the CI for the regression coefficient of the covariate that varies between groups is best for the Bayes modal estimator with $\alpha = 3$ and comparable to REMLwith $\alpha = 2$. Importantly, neither prior ever produces a boundary estimate ($\hat{\sigma}_\theta < 10^{-5}$), whereas ML and REML have quite a large proportion of boundary estimates. At the same time, the change in log-likelihood between ML and Bayes modal estimates tends to be small.

Conclusions:

We considered linear varying intercept models and suggested specifying a gamma prior for the group-level standard deviation to avoid boundary estimates. We showed that this prior is only weakly informative and that the Bayes modal estimator has good frequency properties.

Acknowledgment:
The research was supported by the IES (grant R305D100017).

Appendices

Appendix A. References

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

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

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

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

Raudenbush, S. W. and Bryk, A. S. (2002), Hierarchical Linear Models, Sage, Thousand Oaks, CA.

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