Abstract Title Page

Not included in page count.

Title: Using Dirichlet Processes for modeling heterogeneous treatment effects across sites

Authors and Affiliations:
Luke Miratrix, Harvard University
Avi Feller, Harvard University
Natesh Pillai, Harvard University
Debdeep Pati, Florida State University

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Understanding site level variation in treatment effects is a problem growing in importance in the education field. In particular, weak or null global effects in large-scale randomized trials often mask important site-level variation. This suggests that, in the right setting, a given intervention might be much more effective than the global mean would otherwise suggest.

Given this, tools for describing the distribution of site-level impacts are of critical importance. Current methods (implicitly or explicitly) leverage Bayesian methods that assume that site-level effects are Normal (e.g., see Bloom, Raudenbush & Weiss, 2014). This assumption can yield misleading estimates of the site-level distribution (see Bloom and Weiland, 2015), potentially concealing features of these distributions such as skew or long tails. See for example, the small illustrative simulation on Figure 1: both EB estimates and using raw site-level estimates fails to recover the character of the true distribution. This project investigate this phenomenon and explores when more flexible models can lead to more credible inference.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

Modeling the variation of impacts of an intervention across sites is an important problem for understanding how an intervention might differentially operate in different contexts. For example, we might believe that site-level impacts would co-vary with site-level implementation fidelity. Moreover, the strength of this relationship is of interest on its own right. Similarly, we might wish to model the variation of impact across charter schools as we believe that different charter schools are going to vary substantially in their performance. To understand this variation, we would ideally be able to model a distribution of effects. Furthermore, as there is generally no substantive reason to believe that the distribution of site-level estimates is Normal, we need the model of the distribution of effects to be quite flexible, ideally nonparametric. For example, we might believe there could be a long tail of positive or negative effects, suggesting skew or thick tails.

Setting: Not applicable.

Population / Participants / Subjects: Not applicable.

Intervention / Program / Practice: Not applicable.

Significance / Novelty of study:

Description of what is missing in previous work and the contribution the study makes.

Modeling site-level effect distributions has been frequently approached using classic multilevel modeling strategies. For a canonical example for outcomes in general, see, for example, the *range of plausible outcomes* concept of Raudenbush and Bryk (2002). These strategies, while identifying the key questions of interest and identifying the importance of separating individual variation from site variation, arguably rely on normality assumptions in the distribution of random effects. We seek to examine how a relaxation of this distributional assumption could work in practice. To do this we use tools from the nonparametric Bayes literature. Alternative approaches exist, such as the "triple-goal estimators" of Shen et al. (1998) where they modify emperical Bayes estimation methods of site-level effects by changing the associated loss function in an attempt to balance trade-offs between individual estimates, overall density estimates, and estimating a good ranking of sites.

We are not the first to see the connection of the nonparametric Bayes literature and multilevel models. In particular, nonparametric Bayesian models have been used in multilevel regression contexts (e.g., Kyung et al., 2010, Li et al., 2011), but usually the focus is not on the distribution of the random effects but rather improved estimation of the fixed effects (e.g., Kyung et al., 2009). Some other recent work has explored modeling the random effects via a mixture of Normals (Walters, 2015). Feller and Gelman (2014) provide a brief review. We are hoping to directly investigate such random effects, both via finite mixtures and via nonparametric Bayes.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

Dirichlet Processes (Ferguson, 1973) are arguably the most famous of the nonparametric Bayesian tools currently in use. They come in two general forms, the Dirichlet Process and the Dirichlet Process Mixture (Antoniak, 1974).

Dirichlet Process. In this formulation, the researcher models the observed distribution as an infinite-bin multinomial with declining weights. This allows the researcher to put a prior on the distribution itself with two major components: a distribution of the weights and a base distribution over the outcome space. For example, we might observe a distribution F drawn from our DP, and the hierarchical model would then be:

$$F \sim DP(\alpha, G)$$
$$Z_1, \dots, Z_n \sim F$$

with *G* being, say, a zero-centered normal distribution with a large variance. Given *F*, our outcomes Z_i are i.i.d. In our problem the Z_i are conceptually the site-level treatment effects, *n* is the number of sites, and *F* is the distribution of these effects. *F,* the distribution of effects, is itself considered random with the prior defined by the DP. The DP prior allows for a nonparametric estimation of the shape and character of *F*.

Dirichlet Process Mixtures. The above formulation has a few shortcomings, the largest of which is that it does not allow for error in the estimate of *Y*. To be precise, the standard DP assumes that site-level treatment effects are measured exactly. This is clearly not the case, but the related sister of the DP, the Dirichlet Process Mixture model (DPM), allows this source of error. A variant of the above model is:

$$F \sim DP(\alpha, G)$$
$$Z_1, \dots, Z_n \sim F$$
$$T_j \sim N\left(Z_j, \theta_j\right), j = 1, \dots, n$$

Here the observed treatment estimates are considered to be normal draws from the true average effects in site j. This assumption is reasonable as the treatment estimates are averages of individual outcomes, and so the Central Limit Theorem renders this true even if individual outcome distributions are not normal.

The θ_j may vary depending on site characteristics, such as site-level sample size. A simplifying assumption would be to let the θ_j be functions of an overall individual variance term, sample size, and proportion of units treated. Our model here is a mild extension of the more common DPM model where θ_j would be a fully pooled theta. These are called infinite mixture models because the distribution of the T_j (under pooled theta) is a mixture of an infinite number of Gaussian bumps centered at the countably infinite support of the random distribution F.

Our mixture formulation allows for separating measurement error (due to the site level randomization and, potentially, observing only a sample at the site rather than full treatment at the site) from site level variation.

The key advantage of this approach as compared to classic multilevel modeling is we are directly measuring the distribution of the effects rather than attempting to recover the distribution by examining the emperical Bayes estimates of the individual sites in the sample.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

If this approach gives reasonable estimates of the distribution of the effects, then we would be able to answer questions such as what proportion of the sites had a negative impact, or what proportion of the sites had an impact larger than some threshold. Knowing whether there is a substantial skew in the impacts (e.g., generally small with some large, the reverse, or all sites generally the same) would be quite useful for formulating theories of the causes of the variation in site-level impacts.

However, there is reason to believe that the flexible models we propose might require sample sizes beyond what is currently the state-of-the-art. This too is important to identify, because it would suggest that answering questions about site level variation without using specific covariates to explicitly model such variation is impossible without strong structural assumptions such as normality. Relatedly, including covariates to increase precision (such as student-level demographics) could potentially influence needed sample size. This is another direction of inquiry of this project.

We investigate this method and explicitly compare it to the classic methods of multilevel modeling in a variety of contexts to see to what extent this and the classic methods works or fail. Our primary focuses are circumstances where the actual distribution of effects is non-normal. Our particular context of interest will be the case of a long right tail, i.e. where most treatment effects are small but some are large (similar to the rare but unusual effects of Rosenbaum (2007)). The results of this inquiry can inform when researchers could hope to detect and estimate aspects of site-level variation, and elucidate the strengths and vulnerabilities of the different approaches.

Research Design:

Description of the research design.
(May not be applicable for Methods submissions)

The mathematical guarantees of nonparametric Bayesian methods are generally asymptotic in nature, and it is known that these guarantees may require sample sizes that are quite large (i.e., the asymptotics do not kick in for moderate *n*). In our context, sites are generally small to moderate in number, usually reaching only to the hundreds. It is an open question whether we can reasonably expect to find decent estimates of distributions of effects in this scenario.

To explicitly grapple with the finite nature of our problem, we will turn to structured and systematic simulation studies to investigate the properties of these methods. In particular, we will generate data using existing data from randomized trials coupled with a specific model for average treatment effect and then attempt to recover the consequent distribution of site level effects, or summary statistics thereof.

We will measure quality of estimation in two ways. First, we will quantify different summary features of the effect distribution, such as proportion of units with negative effects and the skew of the estimated distribution as compared to the truth, and attempt to directly estimate these summary statistics. Second, and relatedly, we will use the Kolmogorov-Smirnov (KS) distance to measure overall departure of the estimated distribution of effects to the original.

Data Collection and Analysis: Not applicable.

Findings / Results: Not applicable.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

Modeling the distribution of site level effects is an important problem, but it is also an incredibly difficult one. Current methods rely on distributional assumptions in multilevel models for estimation. There it is hoped that the partial pooling of site level estimates with overall estimates, designed to take into account individual variation as compared to site level variation, does not distort the overall distribution of the Empirical Bayes site-level estimates too badly.

We plan on investigating this claim, and comparing the results of multilevel modeling to multilevel modeling where the assumption of the normally distributed site level effects is relaxed by replacing it with an unspecified distribution assigned a Dirichlet process prior.

We hope, via simulations and analysis of actual data sets, to identify where multilevel modeling does work in practice, where it is likely to lead to erroneous conclusions, and what requirements one needs to use these approaches with reasonable levels of confidence. It may be the case that instead of attempting to measure the distribution of site level effects, we should instead attempt to directly measure summary statistics of these distributions (such as skewness and variance). These are important next steps, once we establish how this approach generally works in practice.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*.

Bloom, H. S., Raudenbush, S. W., & Weiss, M. (2013). Estimating Variation in Program Impacts: Theory, Practice and Applications, 1–34.

Bloom, H. S., Raudenbush, S. W., & Weiss, M. (2014). Using Multi-site Evaluations to Study Variation in Effects of Program Assignment, 1–58.

Bloom, H. S., & Weiland, C. (2015). Quantifying Variation in Head Start Effects on Young Children's Cognitive and Socio-Emotional Skills Using Data from the National Head Start Impact Study. *Available at SSRN 2594430*.

Feller, A., & Gelman, A. (2014). Hierarchical Models for Causal Effects. *Emerging Trends in the Social and Behavioral Sciences: an Interdisciplinary, Searchable, and Linkable Resource*.

Ferguson, T. S. "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, 1973.

Kyung, M., Gill, J., & Casella, G. (2009). Characterizing the variance improvement in linear Dirichlet random effects models. *Statistics & Probability Letters*, *79*(22), 2343–2350.

Kyung, M., Gill, J., & Casella, G. (2010). Estimation in Dirichlet random effects models. *The Annals of Statistics*, *38*(2), 979–1009.

Li, Y., Müller, P., & Lin, X. (2011). Center-adjusted inference for a nonparametric Bayesian random effect distribution. *Statistica Sinica*, *21*.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (Vol. 1). Sage.

Rosenbaum, P. R. (2007). Confidence Intervals for Uncommon but Dramatic Responses to Treatment. *Biometrics, 63*(4), 1164–1171.

Shen, W., & Louis, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(2), 455–471.

Walters, C. R. (2015). Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start. *American Economic Journal: Applied Economics*, *7*(4), 76–102.

Appendix B. Tables and Figures

Not included in page count.

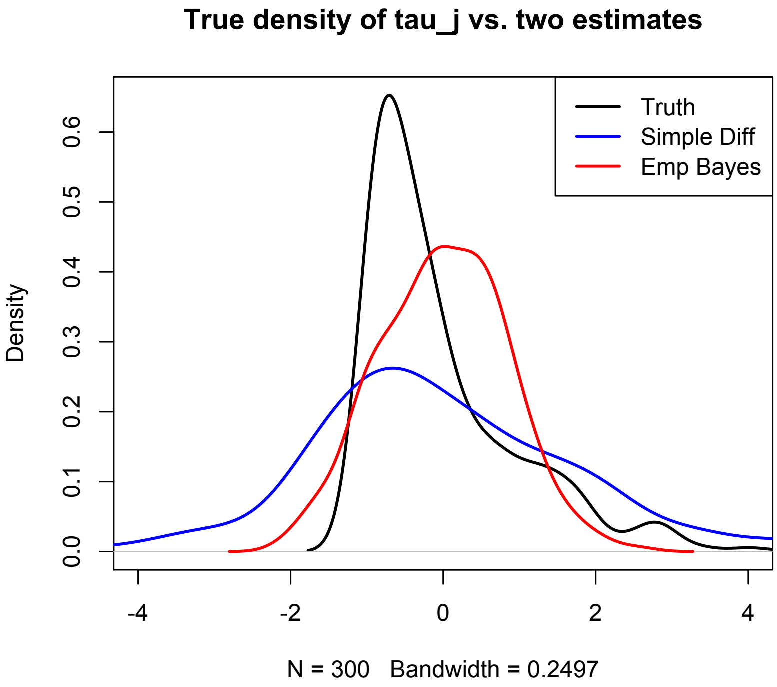

Figure 1: Simple simulation showing the failure of using Empirical Bayes estimates (under a multilevel model) or the raw site-level treatment estimates to estimate the original distribution of site-level effects. Note the Empirical Bayes estimates loses the skew, and the raw estimates are heavily over-dispersed due to not accounting for individual-level variation.