




Modeling Random Effects Using Global–Local Shrinkage Priors in Small Area Estimation

Xueying Tang, Malay Ghosh, Neung Soo Ha & Joseph Sedransk


To cite this article: Xueying Tang, Malay Ghosh, Neung Soo Ha & Joseph Sedransk (2018) Modeling Random Effects Using Global–Local Shrinkage Priors in Small Area Estimation, *Journal of the American Statistical Association*, 113:524, 1476–1489, DOI: 10.1080/01621459.2017.1419135

To link to this article: <https://doi.org/10.1080/01621459.2017.1419135>

 [View supplementary material](#) 

 [Published online: 11 Jul 2018.](#)

 [Submit your article to this journal](#) 

 [Article views: 1491](#)

 [View related articles](#) 

 [View Crossmark data](#) 

 [Citing articles: 11](#) [View citing articles](#) 



Modeling Random Effects Using Global–Local Shrinkage Priors in Small Area Estimation

Xueying Tang^a, Malay Ghosh^b, Neung Soo Ha^c, and Joseph Sedransk^d

^aDepartment of Statistics, Columbia University, New York, NY; ^bDepartment of Statistics, University of Florida, Gainesville, FL; ^cThe Nielsen Company, Seoul, South Korea; ^dJoint Program in Survey Methodology, University of Maryland, College Park, MD

ABSTRACT

Small area estimation is becoming increasingly popular for survey statisticians. One very important program is Small Area Income and Poverty Estimation undertaken by the United States Bureau of the Census, which aims at providing estimates related to income and poverty based on American Community Survey data at the state level and even at lower levels of geography. This article introduces global–local (GL) shrinkage priors for random effects in small area estimation to capture wide area level variation when the number of small areas is very large. These priors employ two levels of parameters, global and local parameters, to express variances of area-specific random effects so that both small and large random effects can be captured properly. We show via simulations and data analysis that use of the GL priors can improve estimation results in most cases. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY

Received November 2016
Revised December 2017

KEYWORDS

Bayesian model; Fay–Herriot model; Poverty rate; Spike-and-slab prior

1. Introduction

The demand for small area estimates is increasingly being felt both in the public and private sectors for policy formulation. One very important current day program undertaken by the United States Bureau of the Census is Small Area Income and Poverty Estimation. This program is targeted to provide estimates related to income and poverty based on American Community Survey (ACS) data at the state level and even at lower levels of geography such as counties, census tracts, and school districts. Examples include estimates of the total as well as the proportion of 5–17 years old children under poverty in such local areas.

Direct estimates for small areas are usually accompanied with large standard errors and coefficients of variation due to small sample sizes. This necessitates the use of models to link different small areas. Small area models are usually classified into area level models and unit level models. In this article, we focus on area level models. Some examples of discussions on unit level models are Battese, Harter, and Fuller (1988), Datta and Ghosh (1991), Tarozzi and Deaton (2009), and Rao and Molina (2015).



The classic area level model is due to Fay and Herriot (1979). It is a mixed effect model with known sampling variances, not necessarily all equal, and independent normal area-specific random effects with zero means and a common unknown variance. The simplicity and interpretability of the resulting small area estimates have made this model most popular in the small area literature with multiple applications.

Despite its wide popularity, questions have been raised regarding blanket application of the Fay–Herriot (FH) model


for all kinds of data. Up until now, the most common challenge to this model is its lack of robustness. The normality and homoscedasticity of random effects is always an unverifiable assumption, and no wonder, this assumption often fails in practice. There are several papers that dispensed with the assumption of normality of the random effects. Among others, we refer to Datta and Lahiri (1995), Lahiri and Rao (1995), Jiang, Lahiri, and Wan (2002), Datta, Rao, and Smith (2005), Jiang and Lahiri (2006a, 2006b), Li and Lahiri (2007), and Fabrizi and Trivisano (2010).


A different but equally pertinent issue has recently surfaced in the small area literature. This concerns the need for random effects in all small area problems, namely, whether even fixed effects models would be adequate for small area estimation in certain situations. Datta, Hall, and Mandal (2011), henceforth referred to as DHM, were the first to address this problem. They suggested a test-based approach where the null hypothesis is that the common random effect variance is zero. The test was based on a discrepancy statistic measuring the lack of fit of the multiple regression model of small area means on certain covariates. This work spurred further research with the same theme. Among others, one may refer to Molina, Nandram, and Rao (2014) and Morales, Pagliarella, and Salvatore (2015).

The DHM procedure performs well when the number of small areas is moderately large, but, as it often happens in practice, the number of small areas is very large, for example, when one considers all counties in the United States. In such situations, even if the regression estimates can describe the small area means very well in most of the small areas, the null hypothesis of

CONTACT Xueying Tang  xt2197@columbia.edu  Department of Statistics, Columbia University, Room 930 SSW, 1255 Amsterdam Avenue, New York, NY 10027.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

 These materials were reviewed for reproducibility.

no random effects is still very likely to be rejected, primarily due to the significant departure of direct estimates from the regression estimates caused by a few large residuals. This problem was first realized by Datta and Mandal (2015). They proposed to model random effects through a mixture of a point mass at zero and a zero-mean normal distribution. Such priors belong to the general class of spike-and-slab priors. The point mass part is suitable for areas where regression estimates are adequate, while the normal distribution part models random effects when the regression model as such is not adequate. Essentially, this model still assumes normally distributed random effects, but the variance can be either zero or a positive value.

In contrast to the spike-and-slab priors of Datta and Mandal (2015), the present article considers global–local (GL) shrinkage priors, which meet the same objective and beyond. In particular, these priors allow area-specific variance components for random effects. In areas where random effects are not needed, the variance component can be very small, while, in areas where regression estimates are not adequate, the variance component can be large so that random effects play a role. The proposed model is, especially advantageous when the number of small areas is very large, for example, when the small areas constitute 3000 plus counties of the United States, and a wide variation in area level effects are expected.

The GL shrinkage priors acquire the feature mentioned above by employing two levels of parameters to express prior variances of normally distributed random effects. The first level consists of the local shrinkage parameters, which are distinct for each area, while the other, the global shrinkage parameter, is common for all random effects. The global shrinkage parameter causes shrinkage on all random effects (global shrinkage effect) to capture the small random effects, while the local shrinkage parameters try to neutralize it for areas that need large random effects. The degree of this neutralizing effects is closely related with the tail of the priors on the local parameters. If it is appropriately heavy-tailed, both small and large random effects can be well-captured.

The GL shrinkage priors were introduced in a series of articles Carvalho, Polson, and Scott (2010), Polson and Scott (2009, 2010, 2012a, 2012b), and Scott (2011). They have been extended into a richer class. Some recent inventions are the three parameter beta normal (TPBN) priors (Armagan, Clyde, and Dunson 2011) and the generalized double Pareto (GDP) priors (Armagan, Dunson, and Lee 2012). The former itself is a big class and includes the now famous horseshoe (HS) priors (Carvalho, Polson, and Scott 2010), the normal-exponential-gamma (NEG) priors (Griffin and Brown 2005), and the Strawderman–Berger (SB) priors (Strawderman 1971; Berger 1980). In multiple testing (Datta and Ghosh 2013; Ghosh et al. 2016), and other contexts where the GL shrinkage priors have had success, they are often further classified into polynomial-tailed priors and exponential-tailed priors according to the tails of the priors of the local parameters. The former subclass is often preferred in these contexts because it enjoys asymptotic optimality when the underlying data are generated from the spike-and-slab model. The present article distinguishes from others on the GL shrinkage priors in that we do not treat them simply as an approximation of the spike-and-slab priors. We show that both subclasses of priors are useful depending on the

data and that the deviance information criterion (DIC) can be used to choose an appropriate one.

The outline of the remaining sections is as follows. In Section 2, we introduce the GL shrinkage priors in the area level model and explain their working mechanism to capture both small and large random effects. Section 3 presents a Gibbs sampler to perform Bayesian computation for the proposed model with some special choice of priors. In Section 4, we compare the performance of different models and priors through simulation in various situations. Some guidance of prior selection is provided in Section 5. Two real-data analyses are performed in Section 6 and we conclude with a discussion in Section 7.

2. Model

We begin with the standard area level model

$$y_i = \theta_i + e_i, \quad \theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, \quad i = 1, \dots, m, \quad (1)$$

where y_i is the direct estimate for small area i , \mathbf{x}_i is a p -dimensional covariate vector, and $\boldsymbol{\beta}$ is a p -dimensional vector of regression coefficients. We assume that $p < m$ and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)^T$ has rank p . The sampling error vector $\mathbf{e} = (e_1, \dots, e_m)^T$ and the random effect vector $\mathbf{u} = (u_1, \dots, u_m)^T$ are assumed to be independent. The elements of \mathbf{e} are independent and each e_i follows a normal distribution with mean 0 and variance V_i . Here, V_i is assumed to be known to avoid nonidentifiability and typically it can be obtained by modification of the standard survey estimate using generalized variance function of Fay and Graubard (2001).

The random effect u_i in (1) captures the variation of small area mean θ_i that cannot be explained by the covariates \mathbf{x}_i . In the FH model, the random effects are independent and identically distributed normal random variables

$$u_i | \sigma_{\text{FH}}^2 \stackrel{\text{iid}}{\sim} N(0, \sigma_{\text{FH}}^2), \quad i = 1, \dots, m. \quad (2)$$

In a hierarchical Bayesian formulation, priors are put on $\boldsymbol{\beta}$ and σ_{FH}^2 . Usually, $\pi(\boldsymbol{\beta}, \sigma_{\text{FH}}^2) \propto \pi(\sigma_{\text{FH}}^2)$ (Datta et al. 1996; You and Chapman 2006). The common variance σ_{FH}^2 for all random effects leads to the same area level variability across small areas, which is hardly realized in practice. For example, for county level estimation, counties usually differ widely in their variability.

The spike-and-slab prior of random effects in Datta and Mandal (2015), on the other hand, is given by

$$u_i = \delta_i v_i, \quad i = 1, \dots, m, \quad (3)$$

where $\delta_1, \dots, \delta_m$ are independent and identically distributed Bernoulli random variables with $P(\delta_i = 1) = \gamma$ and $v_i \sim N(0, \sigma_{\text{DM}}^2)$ for $i = 1, \dots, m$, independently. With this formulation, a random effect is included for small area i only when $\delta_i = 1$. Essentially, the Datta–Mandal (DM) model uses two variance parameters for the random effects. It is zero for the areas where random effects are not needed, and σ_{DM}^2 for all the small areas that need random effects. Independent priors are further assumed for $\boldsymbol{\beta}$, σ_{DM}^2 , and γ , with an improper uniform prior on $\boldsymbol{\beta}$, an inverse gamma (IG) prior on σ_{DM}^2 , and a beta prior, $\text{Beta}(a, b)$, on γ . If one suspects most of the local areas to have insignificant effects, then one chooses a and b in such a way that a/b is small.

Table 1. Some popular priors for local parameters.

Name	$\pi_{\lambda^2}(x)$	Class
Laplace	$\exp(-x)$	<i>E</i>
Normal gamma	$x^{a-1} \exp(-x)$	<i>E</i>
Horseshoe	$x^{-1/2}(1+x)^{-1}$	<i>P</i>
Strawderman-Berger	$(1+x)^{-3/2}$	<i>P</i>
Normal exponential gamma	$(1+x)^{-1-b}$	<i>P</i>
Three parameter beta normal	$x^{a-1}(1+x)^{-a-b}$	<i>P</i>
Generalized double Pareto	$\int_0^\infty y^{\alpha+1} \exp(-xy^2/2 - \eta y) dy$	<i>P</i>

NOTE: The first column gives the names of priors of u_i marginalized over λ_i^2 . The second column lists the corresponding priors of the local parameters λ_i^2 . In the third column, *E* and *P* stand for exponential-tailed priors and polynomial-tailed priors, respectively.

Unlike the FH model and the DM model, we use the GL shrinkage priors to model the random effects u_1, \dots, u_m . Specifically,

$$u_i | \lambda_i^2, \tau^2 \sim N(0, \lambda_i^2 \tau^2), \quad i = 1, \dots, m, \quad (4)$$

and the joint prior of β , τ^2 , and $\lambda^2 = (\lambda_1^2, \dots, \lambda_m^2)^T$ is

$$\pi(\beta, \tau^2, \lambda^2) \propto \pi_{\tau^2}(\tau^2) \prod_{i=1}^m \pi_{\lambda^2}(\lambda_i^2). \quad (5)$$

In the GL model (4), the variance parameters of random effects are area-specific. For each small area, it is expressed as the multiplication of a local parameter λ_i^2 and a global parameter τ^2 , characterizing the area-specific and overall variability of the random effects, respectively. Table 1 lists some of the popular priors for the local parameters used in other contexts. Among them, the Laplace (LA) prior is a special case of the normal-gamma (NG) priors (Fruehwirth-Schnatter and Wagner 2011; Griffin and Brown 2010) with $a = 1$. The HS, SB, and NEG priors are special cases of the TPBN priors. Despite their distinct forms, these GL shrinkage priors possess a common feature, the ability to assign nontrivial probability mass both near zero and in the tail, which enables our GL model to capture both small and large random effects based on data. To see this, first note that given the local and global parameters, the conditional posterior mean of the small area mean θ_i shrinks the direct estimate y_i toward the synthetic regression estimate $\mathbf{x}_i^T \beta$ as

$$E(\theta_i | \beta, \lambda^2, \tau^2, \mathbf{y}) = y_i - B_{GL,i} (y_i - \mathbf{x}_i^T \beta), \quad (6)$$

where $B_{GL,i} = V_i / (V_i + \lambda_i^2 \tau^2)$ is called a shrinkage factor. A larger (smaller) shrinkage factor causes more (less) shrinkage and produces an estimate closer to the synthetic estimate (direct estimate).

Theorem 2.1. Suppose $\pi_{\lambda^2}(\lambda_i^2)$ is either $\exp(-\lambda_i^2)$ or proportional to $\lambda_i^{a-1} (1 + \lambda_i^2)^{-a-b}$, or equivalently, either a LA prior or a TPBN prior is assumed for random effects. For any $\epsilon \in (0, 1)$, if $\tau^2 \rightarrow 0$

$$P(B_{GL,i} < \epsilon | \tau^2, \beta, y_i) \rightarrow 0. \quad (7)$$

More specifically, $P(B_{GL,i} < \epsilon | \tau^2, \beta, y_i) \asymp \tau^{2b}$ if $\pi_{\lambda^2}(\lambda_i^2) \propto (\lambda_i^2)^{a-1} (1 + \lambda_i^2)^{-a-b}$ and $P(B_{GL,i} < \epsilon | \tau^2, \beta, y_i) \asymp \exp(-\frac{V_i(1-\epsilon)}{\epsilon \tau^2})$ if $\pi_{\lambda^2}(\lambda_i^2) = \exp(-\lambda_i^2)$. On the other hand, if $|y_i - \mathbf{x}_i^T \beta| \rightarrow \infty$,

$$P(B_{GL,i} > \epsilon | \tau^2, \beta, y_i) \rightarrow 0. \quad (8)$$

We write $f_1(\tau^2) \asymp f_2(\tau^2)$ as $\tau^2 \rightarrow 0$ if there exist constants C_1 and C_2 such that $C_1 \leq \liminf_{\tau \rightarrow 0} f_1(\tau^2)/f_2(\tau^2) \leq \limsup_{\tau \rightarrow 0} f_1(\tau^2)/f_2(\tau^2) \leq C_2$.

Proof. See Appendix. \square

The first part of Theorem 2.1 shows that if the global parameter is small, then the posterior distribution of shrinkage factor $B_{GL,i}$ will concentrate near one and the posterior mean of the small area mean will be close to the synthetic estimate. Therefore, if the covariates can explain the small area means well in most of the areas, we would want to have a small τ^2 so that the posterior estimate of θ_i in these areas can be shrunk enough toward the synthetic estimate. It seems that a small τ^2 imposes strong shrinkage even for the areas where synthetic estimates are not good enough, but the second part of Theorem 2.1 demonstrates that this shrinkage can be offset if there is a large discrepancy between the direct estimate and the synthetic estimate. The interaction of the global shrinkage effect and the offset effect determines the total amount of shrinkage for each small area. This feature of the GL shrinkage priors resembles the working mechanism of the DM model. As pointed out by Datta and Mandal (2015), the conditional posterior mean of θ_i can be similarly expressed as a shrinkage estimate and the shrinkage factor is

$$B_{DM,i} = \frac{(1 - \tilde{\gamma}_i) \sigma_{DM}^2 + V_i}{\sigma_{DM}^2 + V_i}, \quad (9)$$

where

$$\tilde{\gamma}_i = \frac{\gamma}{\gamma + (1 - \gamma) \sqrt{\frac{\sigma_{DM}^2 + V_i}{V_i} \exp\left\{-\frac{1}{2} \frac{(y_i - \mathbf{x}_i^T \beta)^2 \sigma_{DM}^2}{V_i (\sigma_{DM}^2 + V_i)}\right\}}}. \quad (10)$$

The prior probability of a random effect being zero, γ , plays a similar role as our global parameter τ^2 . If γ is close to zero, the shrinkage factor will be close to 1, causing a shrinkage of the direct estimate toward the synthetic estimate for all small areas. However, if $|y_i - \mathbf{x}_i^T \beta|$ is large enough to make the second term in the denominator of $\tilde{\gamma}_i$ small compared with γ , then the shrinkage factor can be much smaller than one.

In our GL model, given $|y_i - \mathbf{x}_i^T \beta|$ and τ^2 , the extent of the offset effects is closely related with the heaviness of the tails of the priors for the local parameters. In general, a heavier tail has a stronger offset effect, producing smaller shrinkage factors and estimates closer to the direct estimates. Theorem 2.1 shows that the rate of $P(B_{GL,i} < \epsilon | \tau^2, \beta, y_i)$ going to zero is polynomial if TPBN priors are used for local parameters and it is exponential if LA priors are used. Hence, given $|y_i - \mathbf{x}_i^T \beta|$ and τ^2 , the shrinkage factors under LA priors are stochastically larger than those under TPBN priors thus causing more shrinkage.

Another way to look at the GL model is to treat the local parameters, λ_i^2 , as latent variables. After integrating them out, the random effects are independent heavy-tailed random variables with a common scale parameter. From this point of view, the local parameters help bring in extra variability for the random effects compared with the FH model and the amount of extra variability is closely related to the tail behavior of the distribution of the local parameters.

We have seen that the size of the global parameter is crucial to the level of global shrinkage. We put a prior π_{τ^2} on τ^2 so that it can be automatically learned from the data. Weakly informative

IG distributions have been widely used for variance parameters because of its conjugacy. Recently, Gelman (2006) and Polson and Scott (2012b) advocated half-Cauchy priors as a default choice for a top-level scale parameter in Gaussian hierarchical models because of its excellent risk properties. If an half-Cauchy prior is put on τ , then we will call the resulting prior on τ^2 the squared half-Cauchy (SHC) prior and its density is proportional to $(\tau^2)^{-1/2}(\sigma^2 + \tau^2)^{-1}$, where σ is a scale parameter. Both IG priors and SHC priors are considered for the global parameter τ^2 in our context and their performance is compared in Section 4.

3. Computation

Let $\mathbf{y} = (y_1, \dots, y_m)^T$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^T$, and $\mathbf{V} = \text{diag}(V_1, \dots, V_m)$. According to (1), (4), and (5), the joint posterior density of the GL model is

$$\pi(\boldsymbol{\beta}, \mathbf{u}, \lambda^2, \tau^2 | \mathbf{y}) \propto \exp\left[-\frac{1}{2} \sum_{i=1}^m \frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta} - u_i)^2}{V_i}\right] \times \prod_{i=1}^m \left\{ \pi_{\lambda^2}(\lambda_i^2) (\lambda_i^2 \tau^2)^{-\frac{1}{2}} \exp\left(-\frac{u_i^2}{2\lambda_i^2 \tau^2}\right) \right\} \pi_{\tau^2}(\tau^2). \quad (11)$$

Although we assume an improper uniform prior for $\boldsymbol{\beta}$, the integration of the right-hand side of (11) with respect to $\boldsymbol{\beta}, \mathbf{u}, \lambda^2$, and τ^2 can still be finite with some constraints in π_{λ^2} and π_{τ^2} . That is, the posterior distribution for our proposed model is proper under some conditions.

Theorem 3.1. The posterior distribution resulting from the GL shrinkage prior is proper if both $\pi_{\lambda^2}(\lambda_i^2)$ and $\pi_{\tau^2}(\tau^2)$ are proper.

Proof. See Appendix. □

All the priors for the local parameters listed in Table 1 are proper. Also, both IG and SHC priors are proper. Therefore, the propriety of posteriors for the GL model is guaranteed with such choices of priors.

A Gibbs sampler (Gelfand and Smith 1990) can be designed to sample from the posterior density (11). It is easy to find out the full conditionals as

1. $u_i | \boldsymbol{\beta}, \lambda^2, \tau^2, \mathbf{y} \stackrel{\text{ind}}{\sim} N((1 - B_{\text{GL},i})(y_i - \mathbf{x}_i^T \boldsymbol{\beta}), (1 - B_{\text{GL},i})V_i)$;
2. $\boldsymbol{\beta} | \mathbf{u}, \lambda^2, \tau^2, \mathbf{y} \sim N((\sum_{i=1}^m V_i^{-1} \mathbf{x}_i \mathbf{x}_i^T)^{-1} (\sum_{i=1}^m (y_i - u_i) V_i^{-1} \mathbf{x}_i), (\sum_{i=1}^m V_i^{-1} \mathbf{x}_i \mathbf{x}_i^T)^{-1})$;
3. $\pi(\tau^2 | \mathbf{u}, \boldsymbol{\beta}, \lambda^2, \mathbf{y}) \propto (\tau^2)^{-\frac{m}{2}} \exp(-\frac{1}{2\tau^2} \sum_{i=1}^m u_i^2 / \lambda_i^2) \pi_{\tau^2}(\tau^2)$;
4. $\pi(\lambda_i^2 | \mathbf{u}, \boldsymbol{\beta}, \tau^2, \mathbf{y}) \propto (\lambda_i^2)^{-1/2} \exp[-u_i^2 / (2\lambda_i^2 \tau^2)] \pi_{\lambda^2}(\lambda_i^2)$, $i = 1, \dots, m$.

Samples of \mathbf{u} and $\boldsymbol{\beta}$ can be drawn directly from 1 and 2. To draw samples of λ^2 and τ^2 , one needs to use a Metropolis–Hastings step in general. However, in some special cases, 3 and 4 can be expressed as the densities of well-known distributions, so samples of τ^2 and λ_i^2 can be drawn more directly. Specifically, if an exponential, IG or a TPB prior is used for the local parameters $\lambda_1^2, \dots, \lambda_m^2$ and/or the global parameter τ^2 , the corresponding conditional densities take a very simple form. We demonstrate this only for λ_i^2 . The corresponding results for τ^2 are very similar, needing only minor modifications.

Suppose first one uses the exponential prior $\pi_{\lambda^2}(\lambda_i^2) \propto \exp(-\lambda_i^2)$. Then, $\pi(\lambda_i^2 | \mathbf{u}, \boldsymbol{\beta}, \tau^2, \mathbf{y}) \propto (\lambda_i^2)^{-1/2} \exp[-u_i^2 /$

$(2\lambda_i^2 \tau^2) - \lambda_i^2]$. Recalling the definition of a generalized inverse Gaussian (GIG) density function $f(x) \propto x^{\lambda-1} \exp(-\chi/(2x) - \psi x/2)$ labeled as $\text{GIG}(\lambda, \chi, \psi)$, this conditional turns out to be $\text{GIG}(1/2, u_i^2/\tau^2, 2)$. Also, in this case, λ_i^{-2} has a straight inverse Gaussian distribution with mean $(2\tau^2/u_i^2)^{1/2}$ and shape parameter 2. If instead one uses an IG prior with shape parameter $s/2$ and rate parameter $t/2$, then $\pi(\lambda_i^2 | \mathbf{u}, \boldsymbol{\beta}, \tau^2, \mathbf{y}) \propto (\lambda_i^2)^{-(s+1)/2-1} \exp[-(u_i^2/\tau^2 + t)/(2\lambda_i^2)]$, which is IG with scale parameter $(s+1)/2$ and rate parameter $(u_i^2/\tau^2 + t)/2$. Finally, for a TPB prior $\pi_{\lambda^2}(\lambda_i^2) \propto (\lambda_i^2)^{a-1} (1 + \lambda_i^2)^{-a-b}$, introducing a latent parameter ξ_i with $\pi(\lambda_i^2 | \xi_i) \propto (\lambda_i^2)^{a-1} \xi_i^a \exp(-\xi_i \lambda_i^2)$ and $\pi(\xi_i) \propto \xi_i^{b-1} \exp(-\xi_i)$, one gets $\pi(\lambda_i^2 | \mathbf{u}, \boldsymbol{\beta}, \tau^2, \mathbf{y}) \propto (\lambda_i^2)^{a-1/2} \exp[-u_i^2/(2\lambda_i^2 \tau^2) - \xi_i \lambda_i^2]$, which is $\text{GIG}(a-1/2, u_i^2/\tau^2, 2\xi_i)$, and $\pi(\xi_i | \mathbf{u}, \boldsymbol{\beta}, \tau^2, \lambda^2, \mathbf{y}) \propto \xi_i^{a+b-1} \exp[-\xi_i(1 + \lambda_i^2)]$, which is a gamma distribution with shape parameter $a+b$ and rate parameter $\lambda_i^2 + 1$.

With an IG prior with shape c and rate d , the conditional posterior distribution of the global parameter τ^2 is an IG whose shape parameter is $m/2 + c$ and rate parameter is $\sum_{i=1}^m u_i^2/(2\lambda_i^2) + d$. To make sure the posterior distribution is not significantly affected by the prior, c and d should be much smaller than $m/2$ and $\sum_{i=1}^m u_i^2/(2\lambda_i^2)$, respectively. This is relatively easy to achieve for c since $m/2$ is known and usually large. For d , in practice, one can try a few test runs with different choices of d and choose the one that is small enough. For the two real-data examples presented in Section 6, we have checked that $c = d = 10^{-10}$ is small enough and a smaller value does not change results significantly. We use this choice in the simulation as well.

4. Simulations

In this section, we compare the performance of the FH model, the DM model, and our GL model via simulation studies. The data generation settings are adopted from Chakraborty, Datta, and Mandal (2016). The number of small areas, m , is set to be 100, 500, or 1000. For each choice of m , we generated data from model (1). The design matrix \mathbf{X} includes a column of ones and one explanatory variable sampled from $N(10, 2)$. The same \mathbf{X} is used to simulate data for a given m . The coefficient vector $\boldsymbol{\beta}$ is fixed at $(20, 1)^T$. The variances of the errors, V_i , are chosen from the set $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$ and each value in the set is allocated to the same number of small areas.

Five settings are considered for the random effects u_i , $i = 1, \dots, m$

$$u_i \sim N(0, 1), \quad (12)$$

$$u_i \sim \delta_i N(0, 5^2), \quad \delta_i \sim \text{Bern}(0.2), \quad (13)$$

$$u_i \sim \delta_i N(0, 5^2) + (1 - \delta_i) N(0, 1^2), \quad (14)$$

$$u_i \sim N(0, \sigma_i^2), \quad (15)$$

$$u_i \sim t_3. \quad (16)$$

Following Chakraborty, Datta, and Mandal (2016), in the normal mixtures setup (14), we set $\delta_i = 1$ for each i that is a multiple of 5 and keep rest of the $\delta_i = 0$. In the multivariate setup (15), the variances of random effects σ_i^2 are sampled with equal proportion from the set $\{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$.

By generating random effects from (12), we assume the true underlying model of the data is the FH model. Similarly, when the random effects are generated from (13), the true underlying

Table 2. Correlations between covariates and true small area means in different simulation settings.

m	Normal	Dirac–Normal	Normal mixture	Multivariate	T
100	0.85	0.60	0.55	0.42	0.71
500	0.82	0.54	0.51	0.39	0.65
1000	0.81	0.53	0.50	0.39	0.63

model is the DM model. The other three settings (14)–(16) violate the assumptions of the FH, the DM, and the GL models. For each of the five settings of random effects and each choice of the number of small areas, 100 datasets are generated. Table 2 gives the average of correlations between covariates and true small area means over the 100 datasets in different simulation settings.

For each dataset, we fit the FH model, the DM model, and our GL model. In the FH model, an improper uniform prior is assumed for σ_{FH}^2 to be consistent with Datta and Mandal (2015). This prior is also used in Morris and Tang (2011) and the propriety of the resulting posterior is well-known (Berger 2013). Some other choices, such as improper uniform priors for σ_{FH} and IG priors for σ_{FH}^2 , also exist in literature. Our exploration shows that results with various choices of $\pi(\sigma_{\text{FH}}^2)$ are not significantly different. For the DM model, our prior specifications are the same as those in Datta and Mandal (2015), where γ has a Beta(1, 4) prior, σ_{DM}^2 has an IG distribution with mean \bar{V} and variance \bar{V}^2 , and $\bar{V} = \frac{1}{m} \sum_{i=1}^m V_i$. For the GL model, we considered various choices for π_{λ^2} described in Table 1. Among them, HS, SB, NEG ($b = 0.75$), and GDP ($\alpha = \eta = 1$) are chosen as examples of polynomial-tailed priors, while LA and NG ($a = 0.5$) are examples of exponential-tailed priors. For the priors of the global parameter, we considered an IG prior with both shape and rate parameters being 10^{-10} and a SHC prior with scale parameter being 1.

For each fit, we estimate θ_i by its posterior mean and evaluate the fit by four deviance measures, average absolute deviation (AAD), average squared deviation (ASD), average absolute relative deviation (ARB), and average squared relative deviation (ASRB), which are defined as follows:

$$\begin{aligned} \text{AAD} &= \frac{1}{m} \sum_{i=1}^m |\hat{\theta}_i - \theta_i|, \quad \text{ASD} = \frac{1}{m} \sum_{i=1}^m (\hat{\theta}_i - \theta_i)^2, \\ \text{ARB} &= \frac{1}{m} \sum_{i=1}^m |(\hat{\theta}_i - \theta_i)/\theta_i|, \quad \text{ASRB} = \frac{1}{m} \sum_{i=1}^m (\hat{\theta}_i - \theta_i)^2/\theta_i^2. \end{aligned} \quad (17)$$

The empirical coverage rate of 95% credible interval of θ_i is also used to evaluate the interval estimates. For a given situation, we report the median of the five measures over the 100 datasets.

First, we compare the performance of our GL model with different priors for the global parameter τ^2 . Figure 1 presents the results for the GL model with HS or LA priors for the local parameters and IG or SHC priors for the global parameters. It is clear that the performance of the GL models is mostly governed by the priors on the local parameters. For a given choice of π_{λ^2} , the four deviation measures and the empirical coverage rate the credible intervals from GL models with different choices of priors on τ^2 vary only slightly, especially when the number of small areas is large. We will restrict our discussion on models with IG priors for the global parameter in the remainder of the article.

Next, we compare the GL model with the FH model and the DM model. The results are summarized in Figure 2. The first two columns in the figure show that the FH model performs the best when the random effects are generated from the normal setting (12) and the DM model performs the best when the random effects are generated from the dirac–normal mixture setting (13). This is not a surprise since the fitting model matches the underlying data generating model. Even so, the coverage rate for the DM model is not as good as those for GL models in the second column. When the random effects are generated from the other three settings, although the DM model performs much better than the FH model in terms of the four deviation measures, the empirical coverage rate of the 95% credible intervals is always lower than the nominal rate, especially when m is large. In these three settings, the model that has the smallest deviation measures is always a GL model. As for which GL model gives the best performance, it depends on the data generating model. In general, the GL models with the same type of priors for local parameters perform similarly. The exponential-tailed priors are better in settings (12), (15), and (16), while the polynomial-tailed priors are better in settings (13) and (14).

To take a closer look at the results, Figure 3 presents the deviation measure results in settings (13) and (14) stratified by the type of random effects. For example, for AAD, we compute

$$\text{AAD}_0 = \frac{1}{m_0} \sum_{\{i:\delta_i=0\}} |\hat{\theta}_i - \theta_i|,$$

and

$$\text{AAD}_1 = \frac{1}{m_1} \sum_{\{i:\delta_i=1\}} |\hat{\theta}_i - \theta_i|,$$

where m_j is the cardinality of $\{i : \delta_i = j\}$, $j = 0, 1$, AAD_0 is the averaged absolute deviation for estimates of areas with “small” random effects and AAD_1 is the averaged absolute deviation for estimates of area with “large” random effects. Similar stratification can be done for the other three deviation measures.

Figure 3 shows that polynomial and exponential GL models are excellent at capturing small and large random effects, respectively. Even when the data generating model is DM, GL models with exponential-tailed priors in general perform better than DM in estimating small area means for areas with large random effects. In terms of estimating means for areas with small random effects, the DM model is more sensitive to the true distributions of random effects than GL models. When the small random effects are normally distributed instead of exact zero, DM model loses its advantage in capturing small random effects over GL models with polynomial-tailed priors.

According to Figure 3, GL models with exponential-tailed priors consistently yield smaller deviation measures than the polynomial-tailed priors for areas with large random effects, whereas the reverse is true for areas with small random effects. Note that this does not contradict the observations we have made in Theorem 2.1 that the LA prior shrinks more than the TPBN priors since there the conclusions are made conditional on τ^2 while here inference about τ^2 is based on its posterior distribution. Table 3 gives medians of posterior means of τ^2 over 100 simulated datasets for the GL models with the HS prior and

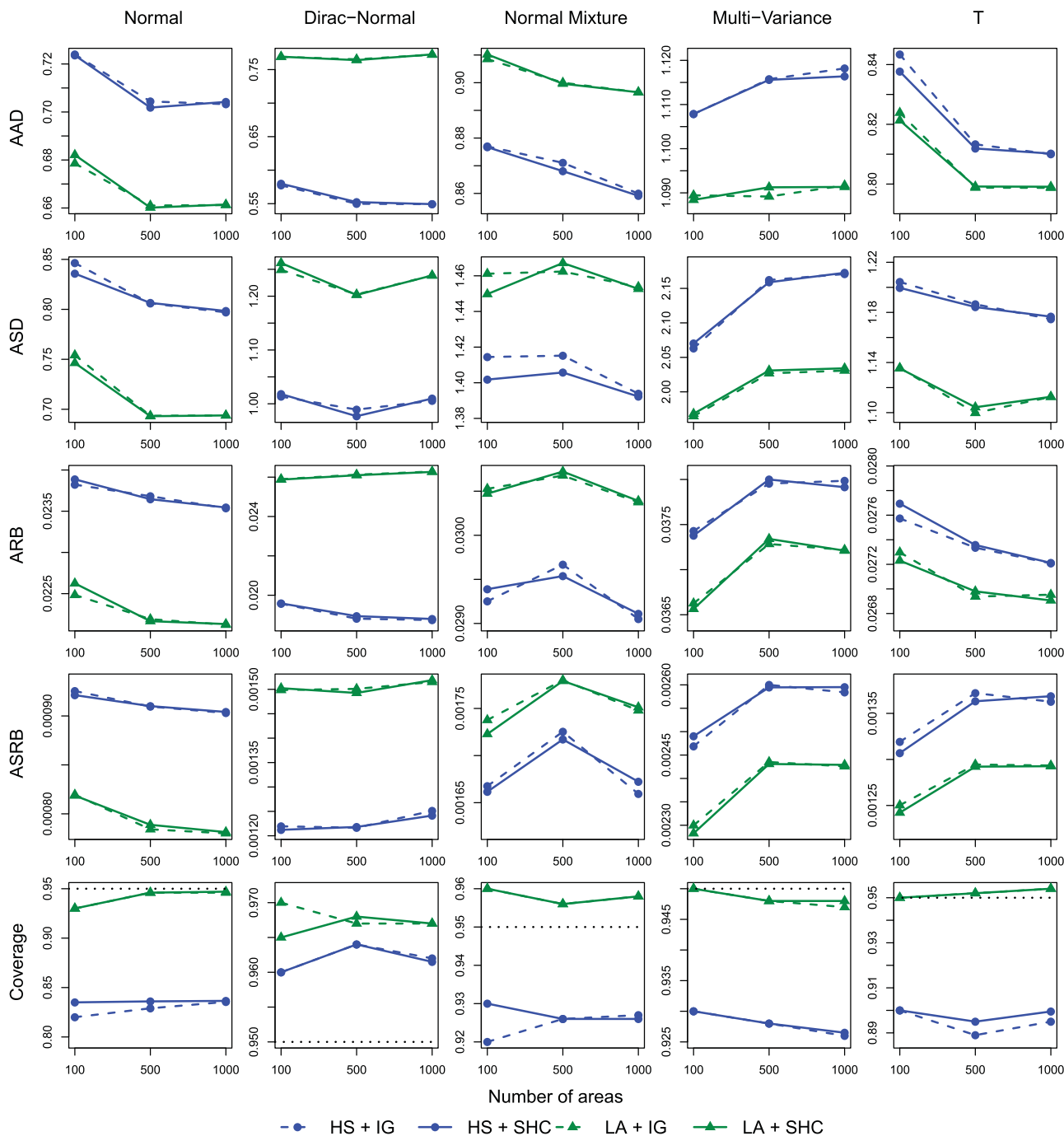


Figure 1. Deviation measures and coverage rates for GL models with different choices of priors for the global parameter. Note that the ranges of vertical axes are different across plots.

Table 3. Medians of posterior sample means of τ^2 over 100 datasets for the GL models with the HS prior and the LA prior for local parameters and the IG prior for the global parameter.

	Dirac-Normal					Normal mixture	Multivar	T
	Normal	$\gamma = 0.05$	$\gamma = 0.10$	$\gamma = 0.20$	$\gamma = 0.50$			
HS	0.08	0.04	0.11	0.30	2.65	0.59	3.45	0.32
LA	0.85	0.79	2.01	3.87	12.17	5.12	11.75	2.44

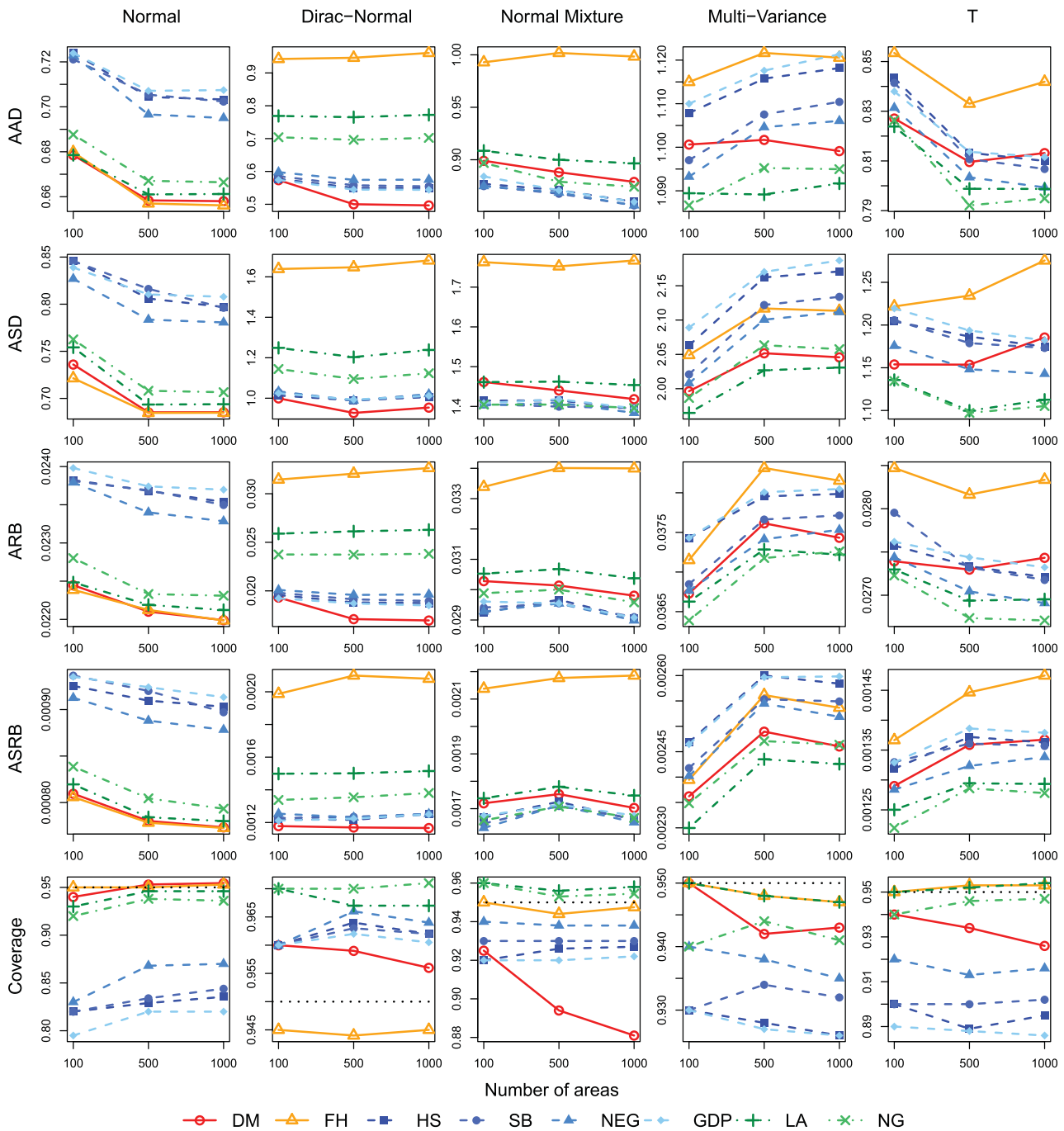


Figure 2. Deviation measures and coverage rates of 95% credible intervals for various models. Note that the ranges of vertical axes are different across plots.

the LA prior. The GL model with the LA prior consistently produces larger estimates of τ^2 than the GL model with the HS prior. This is because the relatively light exponential tail of the LA prior induces weaker offset effect. To remedy the extra shrinkage caused by this, a weaker global shrinkage effect (larger τ^2) is needed.

Table 3 also demonstrates the role of τ^2 in controlling the overall shrinkage level. In dirac-normal mixture settings, for both HS and LA priors, the estimates of τ^2 show an increasing trend as the proportion of nonzero random effects increases. Although the normal mixture setting has the same percentage and distribution of large random effects as the dirac-normal mixture setting with $\gamma = 0.2$, the global parameter is about

twice as large. This happens because small random effects are nonzero in the former setting and hence less global shrinkage is needed.

Another thing to be noted is that in the dirac-normal mixture setting, the DM model and the GL models with polynomial-tailed priors improved estimation by a large amount, while the improvement in other settings is not as large as in this setting. This difference is mainly due to the intrinsic features of the data generating models. The random effects setting (13) is more different from the FH model than the other settings because the majority of the random effects it generates are exact zeros. If the fitting model can grant exact zero random effects (DM model) or shrink small random effects aggressively to zeros

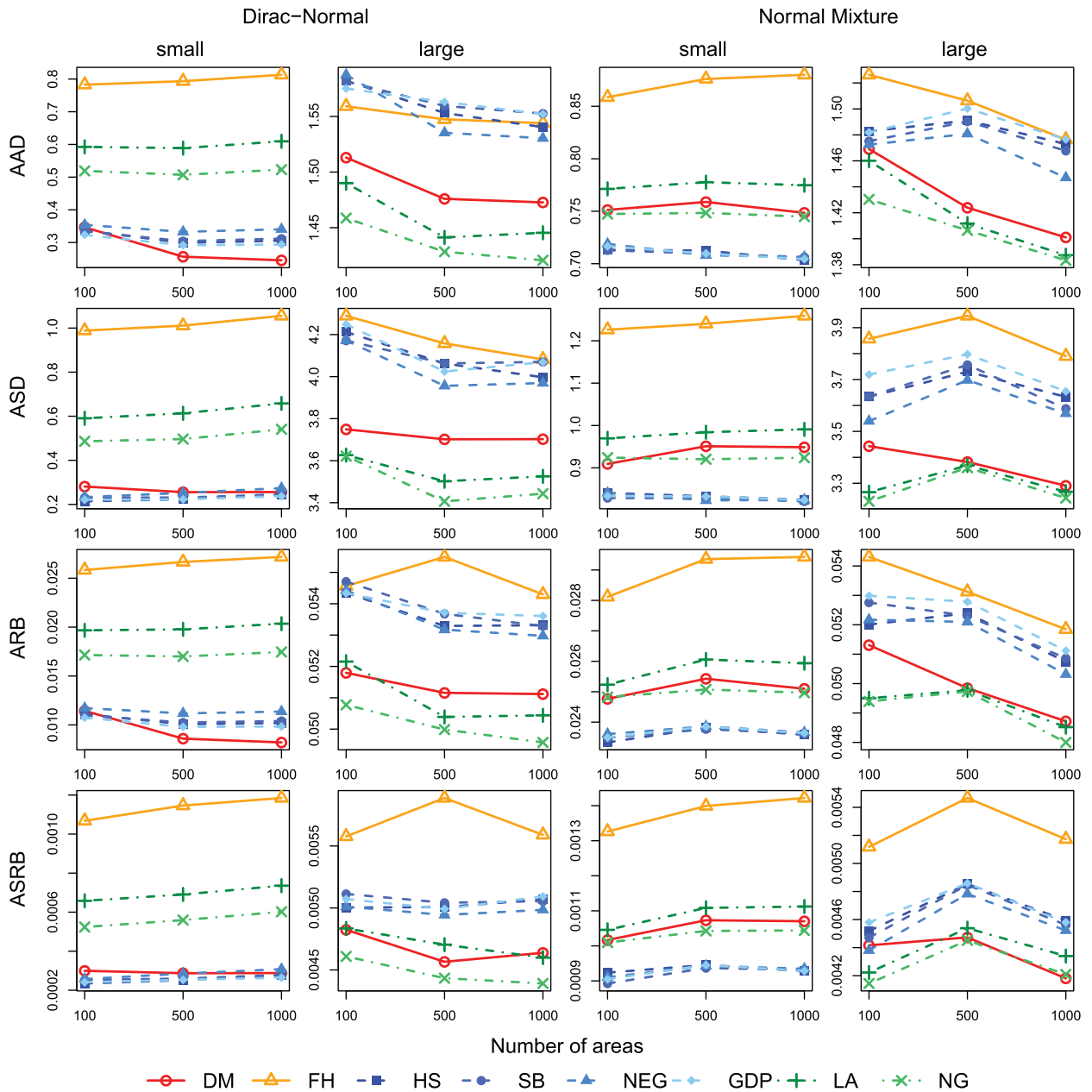


Figure 3. Stratified deviation measures for dirac-normal mixture model and normal mixture normal. Note that the ranges of vertical axes are different across plots.

(GL models with polynomial-tailed priors), then estimation performance will be greatly improved.

5. Prior Selection

We have seen from the simulation study that priors for random effects work differently in different scenarios and that there is no universal prior that works best in every situation. In practice, it is almost impossible to know the underlying true model, so how to choose an appropriate prior for random effects based on the data is an important question.

Among many other criteria for Bayesian model selection or comparison, we use the DIC proposed by Spiegelhalter et al. (2002) because of its simplicity and good empirical performance in our simulation studies. Similar to many other information criteria, DIC compares models by considering both goodness of fit and model complexity. Specifically, for a model with parameter θ ,

$$DIC = D(\bar{\theta}) + 2p_D,$$

where $p_D = \bar{D} - D(\bar{\theta})$ measures the model complexity, $D(\bar{\theta})$ is the deviance of the model evaluated at the posterior mean of model parameter θ , and \bar{D} is the posterior mean of the deviance. In our context, $D(\theta) = \sum_{i=1}^m (y_i - \theta_i)^2 / V_i$. Suppose we have a posterior sample $\theta^{(g)} = (\theta_1^{(g)}, \dots, \theta_m^{(g)})$, $g = 1, \dots, G$ of the small area means based on a model aforementioned. The DIC of this model can be easily estimated by

$$\widehat{DIC} = \frac{2}{G} \sum_{g=1}^G \sum_{i=1}^m \frac{(y_i - \theta_i^{(g)})^2}{V_i} - \sum_{i=1}^m \frac{(y_i - \hat{\theta}_i)^2}{V_i},$$

where $\hat{\theta}_i = \frac{1}{G} \sum_{g=1}^G \theta_i^{(g)}$ is the posterior sample mean of θ_i . Among all the candidate models, the one with the smallest DIC is chosen as the best one to perform final analysis. We applied this procedure to the simulated datasets in Section 4. The candidate models are all the models we have fitted there. The results

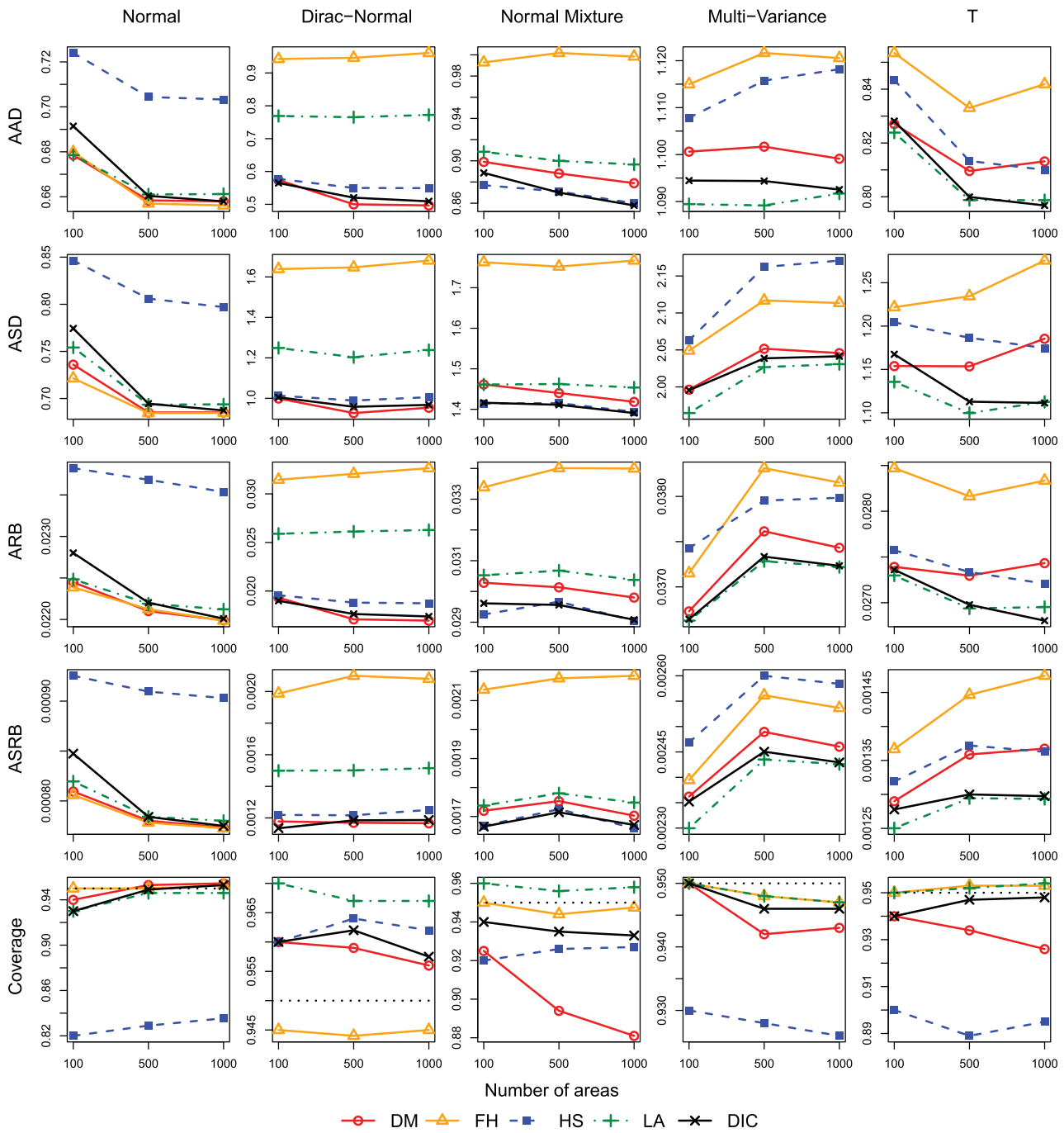


Figure 4. Deviation measures and coverage rates for the model selected by DIC. Note that the ranges of vertical axes are different across plots.

are presented in Figure 4. We can see from the figure that the DIC procedure performs very well, especially when the number of small areas is large. The selected model produces deviation measures close to the smallest among all candidate models and better coverage rate than the DM model.

6. Data Analyses

6.1. State Level Child Poverty Ratio

In this section, we use the child poverty ratio dataset discussed by Datta and Mandal (2015) to further demonstrate the performance our GL model and compare it with the DM model

and the FH model. The state-level direct estimates of the poverty ratio for age group 5–17 were obtained from 1999 Current Population Survey (CPS). Besides the intercept, three covariates are included in the regression part. They are the number of child exemptions, Internal Revenue Service nonfiler rate, and the residuals from fitting a model for the 1989 census poverty data on the previous covariates. As discovered in Datta and Mandal (2015), although the discrepancy test proposed in DHM suggested to include the random effects in the FH Model, the presence of a random effect is necessary only for Massachusetts.

We applied the FH model, the DM model, and the GL model to estimate the poverty ratio θ_i . The prior and hyperparameter specification of the FH model and the DM model is the same

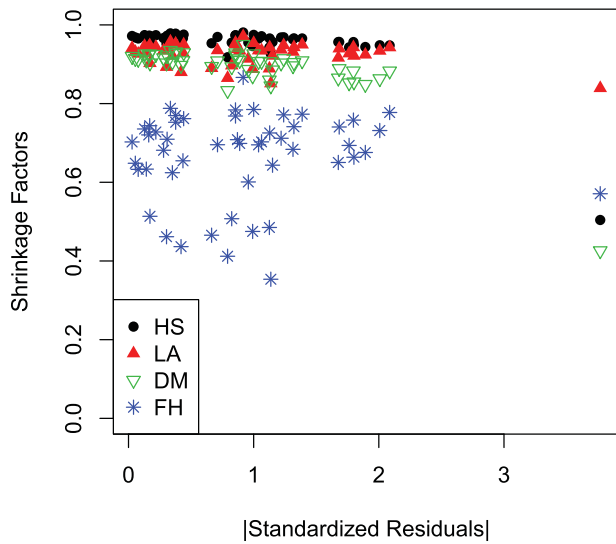


Figure 5. Posterior means of shrinkage coefficients of different states. The four points on the right denotes the shrinkage factors of Massachusetts.

as in Section 4. For the GL model, we used an IG prior for the global parameter τ^2 according to our findings in the simulation study. Both shape and rate parameters of the IG prior is set to 10^{-10} . We considered the HS and LA priors for the choice of π_{λ^2} . They are chosen as representatives of the polynomial-tailed and exponential-tailed priors. For each of these models, we estimated θ_i by its posterior mean.

First, we compare the shrinkage factors of different models. Recall that a shrinkage factor determines the amount of shrinkage of the direct estimate y_i toward the synthetic estimate $x_i^T \beta$. The larger the shrinkage factor, the more is the shrinkage toward the synthetic estimator. For DM model and our GL model, the shrinkage factors are given in (9) and (6), respectively. For the FH model, the shrinkage factor is

$$B_{FH,i} = \frac{V_i}{\sigma_{FH}^2 + V_i}.$$

Figure 5 reports the posterior means of the shrinkage factors of 51 states from various models. For all the states except for Massachusetts, both the GL model, with either the HS prior or the LA prior, and the DM model have much larger shrinkage factors than those produced by the FH model. The GL model with the HS prior shrinks the small area means of those states slightly more than those from the DM model. For Massachusetts, the DM model and the GL model with the HS prior produces less shrinkage than the FH model, while the GL model with the LA prior produces more shrinkage. Given the shrinkage factor B_i and the covariate coefficients β , the posterior variance of θ_i is $(1 - B_i)V_i$, so one can expect that a more shrunk estimate will have less uncertainty.

Figure 6 presents the posterior means of random effects u_i obtained from three different models. In general, the posterior means of the random effects from our GL model are closer to zero than those from the FH model since in the latter model, the variance parameter is overestimated because Massachusetts needs a large random effect. For Massachusetts, the estimated random effect from the FH model is smaller than that from the GL model with the HS prior, but is larger than

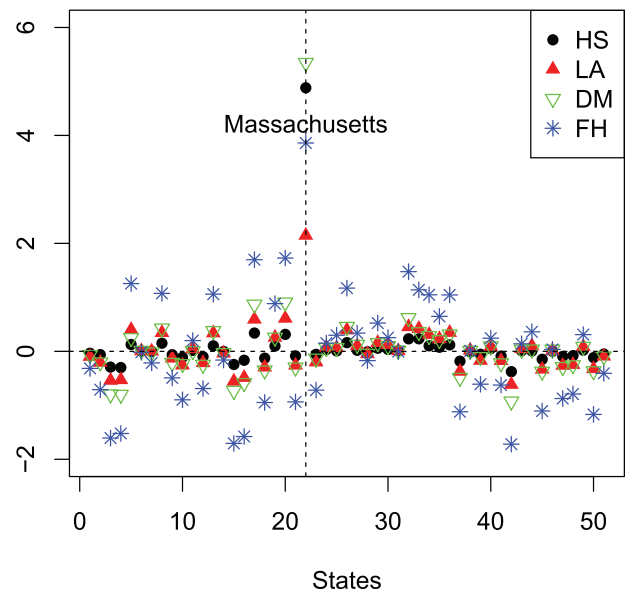


Figure 6. Posterior means of random effects u_i from various models.

that from the GL model with the LA prior. All these observations demonstrate that when compared with exponential-tailed priors, polynomial-tailed priors lead to less shrinkage for areas where synthetic estimates deviate from the direct estimates, but more shrinkage for areas where synthetic estimates are close to the direct estimates.

Following Datta and Mandal (2015), we use the ratio benchmarked state-level poverty ratios obtained from the 2000 census as the true values and compare the estimated θ_i with them through the four deviation measures defined in (17). Let $c_i, i = 1, \dots, 51$, be the state-level poverty ratio estimates for 5–17 age group obtained from the 2000 census without any adjustment and $c'_i, i = 1, \dots, 51$, be the ratio benchmarked poverty ratios. Then, $c'_i = Rc_i$, where $R = \frac{\sum_{i=1}^{51} (\text{pop})_i y_i}{\sum_{i=1}^{51} (\text{pop})_i c_i}$, $(\text{pop})_i$ is the estimated population of the 5–17 age group in the i th state and y_i is the direct estimate from 1999 CPS. The deviation measures are calculated by replacing θ_i in (17) by c'_i and the results are given in Table 4. As we can see, the GL models perform as well as or slightly better than the DM model, which is favored by the DIC. If the LA prior is used for the local parameters, the resulting deviance measurements are the smallest among the models we considered. The last column of Table 4 gives the number of states with absolute deviations of the estimated poverty ratios from the true ratios less than 2%. The deviation for Massachusetts drops below 2% if the LA prior or the fixed effect model is used.

Table 4. Comparison of the estimators obtained from various models.

Estimate	AAD	ASD	ARB	ASRB	DIC	# diff. < 2%
Direct	2.718	12.291	0.196	0.067	—	23
FH	1.192	2.551	0.080	0.010	273.29	40
DM	1.029	2.145	0.068	0.009	268.80	45
HS	1.011	2.041	0.067	0.008	273.09	45
LA	0.987	1.864	0.064	0.006	275.92	46
WLS	1.031	1.891	0.069	0.007	—	46

NOTE: WLS stands for the weighted least square model with no random effect for any state.

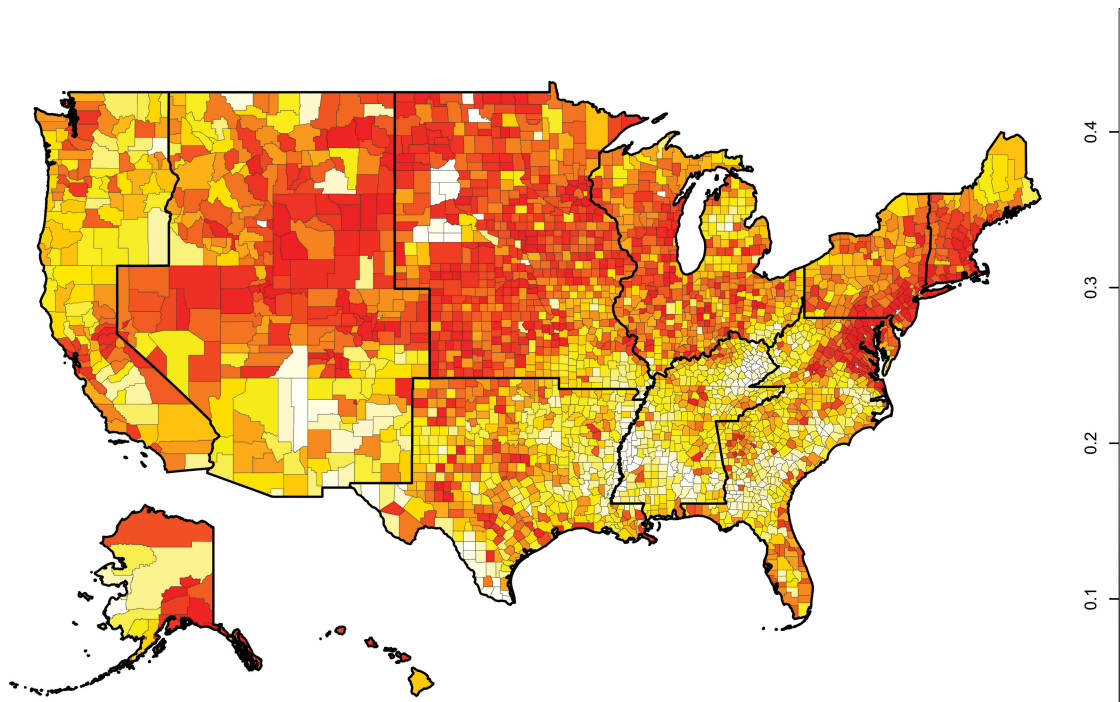


Figure 7. Posterior means of θ .

6.2. County Level Overall Poverty Rates

In this section, we analyze a dataset containing 5-year (2007–2011) county level pooled ACS estimates of overall poverty rates along with their associated design-based standard errors obtained from the “American Fact Finder” website maintained by the U.S. Census Bureau. There are 3141 counties in this dataset. Two counties were dropped from the total 3143 counties of United States (in 2011) due to lack of people in the poverty universe.

We used foodstamp participation rate and a column of ones as our covariate variables to produce model-based estimates of poverty rates. The correlation between foodstamp participation rate and overall poverty rate is 0.81. We fitted the FH, the DM, the GL model with the HS prior and the GL model with the LA prior. The prior for the global parameter is still an IG with both shape and rate parameter equal to 10^{-10} . The DIC for these models are -15883 , -15863 , -15752 , and -15947 , respectively. Since the GL model with the LA prior has the smallest DIC, we present the results from this model as our final analysis.

The estimated poverty rates range from 0.033 (Borden County, Texas) to 0.479 (Shannon County, South Dakota) and the median is 0.147. Figure 7 displays a map of the estimated poverty rates of 3141 counties. The counties with high poverty rates concentrate in East South Central Division and south part of South Atlantic Division. In Mississippi, Georgia, Alabama and New Mexico, 55% or more of their counties have poverty rates greater than the third quartile (0.189) of the estimated rates across counties. The counties with low poverty rates concentrate in New England, Middle Atlantic, north part of South Atlantic Division, and north part of Mountain Division. In New Hampshire, Connecticut, Rhode Island, Wyoming, Hawaii, and New Jersey, 70% or more of their counties have poverty rates lower than the first quartile (0.111).

Figure 8 reports the absolute values of the posterior means of the random effects. The estimated random effects vary greatly across counties. In about, 10% of the 3141 counties, the absolute value of the posterior mean of the random effect is less than 0.002, while in another 10% of counties, the value is greater than 0.038. East and west coasts are two regions that contain many counties with large random effects.

7. Discussion

The article considers GL shrinkage priors for random effects in the context of small area estimation when the variances of random effects differ widely. These priors assume the random effects are scale mixtures of normals. The variance parameters of the normal distributions are expressed as the multiplication of a local parameter and a global parameter. We find that the performance of the model is closely related to the tail of the priors for the local parameters. For a given global parameter τ^2 , exponential-tailed priors cause more shrinkage than polynomial-tailed priors. If an IG prior or a SHC prior is put on τ^2 , exponential-tailed priors and polynomial-tailed priors are good at estimating large and small random effects, respectively. Since different priors have their own best working scenarios, we choose to use DIC to perform model selection because of its simplicity and good performance in our simulation studies. Some other model selection criteria, such as Bayes factors (Kass and Raftery 1995) and predictive Bayesian information criterion (Ando 2007) can be applied, but they usually require more computational efforts.

Although the GL model uses area-specific variance parameters, it does not necessarily lead to over-fitting issues. As we have mentioned, the local parameters λ_i^2 can be treated as latent variables. Once they are integrated out, the random effects are essentially independent heavy-tailed random variables with a

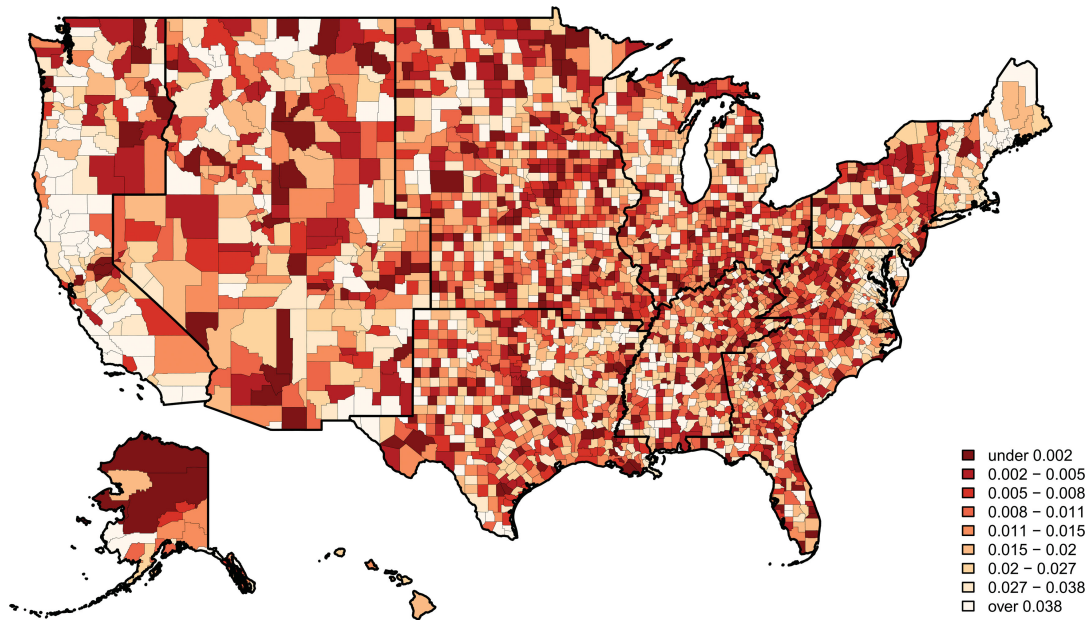


Figure 8. Absolute values of posterior means of random effects.

common scale parameter. Therefore, the local parameters just help bring in extra variability for the random effects compared with the FH model. On the other hand, we use DIC to choose models. It also penalizes the model complexity.

Although the GL model uses distinct variance parameters for each small area, the number of parameters sampled in one iteration of Markov chain Monte Carlo (MCMC) algorithm is the same for both the GL model and the DM model if latent variables are not introduced and the order of computational complexity is of order mp^2 . If latent variables are sampled, for example, in the HS model, the increased computational cost is of order m . The real-data analysis in this article was performed on an Intel Core i5 2.6GHz MacBook with 8 GB memory. The time consumed for fitting the FH model, the DM model, the GL model with the HS prior, and the GL model with the LA prior is 21, 45, 57, and 54 sec for the state level data and 91, 1152, 1916, and 1672 sec for the county level data. We used four covariates in the state level data analysis and two covariates in the county level data analysis. If more covariates are considered, the increased computational cost is mainly due to the calculation of $\tilde{X}^T (= V^{-\frac{1}{2}}X)$ and $(\tilde{X}^T \tilde{X})^{-1}$, which is needed for all the Bayesian models considered in this article.

When a SHC prior is assumed for τ^2 in Section 4, the scale parameter σ is fixed at 1. Gelman (2006) recommended a high but not off value for σ . We explore other values of σ ranging from 0.1 to 10 and find that the performance of the GL models on estimating the small area means merely changes.

Future work will direct toward further applications of these methods, for example, for unit level models as well for handling of multivariate small area data.

A. Proof of Theorems

A1. Proof of Theorem 3.1

Proof. To prove the propriety of the posterior distribution, we need to show the integral in the right-hand side of (11) with respect to β, u, λ^2 , and τ^2 is

finite. Since X is of full rank, $X^T V^{-1} X$ is nonsingular. Let $z_i = y_i - u_i, i = 1, \dots, m$ and $z = (z_1, \dots, z_m)^T$. We have

$$\begin{aligned} & \exp \left[-\frac{1}{2} \sum_{i=1}^m \frac{(y_i - x_i^T \beta - u_i)^2}{V_i} \right] \\ &= \exp \left[-\frac{1}{2} (z - X\beta)^T V^{-1} (z - X\beta) \right] \\ &= \exp \left[-\frac{1}{2} (\beta - \hat{\beta})^T X^T V^{-1} X (\beta - \hat{\beta}) - \frac{1}{2} z^T Pz \right], \end{aligned}$$

where $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} z, P = \tilde{X}^T (\tilde{X}^T \tilde{X})^{-1} \tilde{X}$, and $\tilde{X}^T = V^{-\frac{1}{2}} X$. Integrating the right-hand side of (11) with respect to β and noting $\exp(-\frac{1}{2} z^T Pz) \leq 1$, we have

$$\pi(u, \lambda^2, \tau^2 | y) \leq C \prod_{i=1}^m \left\{ \pi_{\lambda^2}(\lambda_i^2) (\lambda_i^2 \tau^2)^{-\frac{1}{2}} \exp\left(-\frac{u_i^2}{2\lambda_i^2 \tau^2}\right) \right\} \pi_{\tau^2}(\tau^2), \tag{18}$$

where C is a generic positive constant. Integrating out u from (18), we have

$$\bar{\pi}(\lambda^2, \tau^2 | y) \leq C \pi_{\tau^2}(\tau^2) \prod_{i=1}^m \pi_{\lambda^2}(\lambda_i^2).$$

Therefore, if the priors of τ^2 and λ_i^2 are proper, the integration of the posterior density with respect to β, u, λ^2 , and τ^2 is finite. \square

A2. Proof of Theorem 2.1

Proof.

$$\begin{aligned} & P(B_{GL,i} < \epsilon | \beta, \tau^2, y_i) \\ &= \frac{\int_{\frac{V_i}{\tau^2}(\frac{1}{\epsilon} - 1)}^{\infty} (V_i + \lambda_i^2 \tau^2)^{-1/2} \pi_{\lambda^2}(\lambda_i^2) \exp\left(-\frac{(y_i - x_i^T \beta)^2}{2(V_i + \lambda_i^2 \tau^2)}\right) d\lambda_i^2}{\int_0^{\infty} (V_i + \lambda_i^2 \tau^2)^{-1/2} \pi_{\lambda^2}(\lambda_i^2) \exp\left(-\frac{(y_i - x_i^T \beta)^2}{2(V_i + \lambda_i^2 \tau^2)}\right) d\lambda_i^2} \tag{19} \end{aligned}$$

If π_{λ^2} is proper, by Lebesgue dominated convergence theorem, the denominator in (19) converges to $V_i^{-1/2} \exp(-\frac{(y_i - x_i^T \beta)^2}{2V_i})$ as $\tau^2 \rightarrow 0$. Let N denote

the numerator, $r_i = |y_i - x_i^T \beta|$ and $c = V_i(1 - \epsilon)/\epsilon$. It is obvious that

$$\begin{aligned} \exp\left(-\frac{\epsilon r_i^2}{2V_i}\right) \int_{c/\tau^2}^{\infty} (V_i + \lambda_i^2 \tau^2)^{-1/2} \pi_{\lambda^2}(\lambda_i^2) d\lambda_i^2 &\leq N \\ &\leq (V_i/\epsilon)^{-1/2} \int_{c/\tau^2}^{\infty} \pi_{\lambda^2}(\lambda_i^2) d\lambda_i^2. \end{aligned}$$

If $\pi_{\lambda^2}(\lambda_i^2) = \exp(-\lambda_i^2)$, $N \leq (V_i/\epsilon)^{-1/2} \exp(-c/\tau^2)$. On the other hand,

$$\begin{aligned} N &\geq \exp\left(-\frac{\epsilon r_i^2}{2V_i}\right) \int_{c/\tau^2}^{\infty} (V_i + \lambda_i^2 \tau^2)^{-1/2} \exp(-\lambda_i^2) d\lambda_i^2 \\ &= \frac{1}{\tau^2} \exp\left(-\frac{\epsilon r_i^2}{2V_i}\right) \int_c^{\infty} (V_i + x)^{-1/2} \exp(-x/\tau^2) dx \\ &\geq \frac{1}{\tau^2} \exp\left(-\frac{\epsilon r_i^2}{2V_i}\right) \int_c^{2c} (V_i + x)^{-1/2} \exp(-x/\tau^2) dx \\ &\geq \frac{1}{\tau^2} \exp\left(-\frac{\epsilon r_i^2}{2V_i}\right) (V_i + 2c)^{-1/2} \int_c^{2c} \exp(-x/\tau^2) dx \\ &= \exp\left(-\frac{\epsilon r_i^2}{2V_i}\right) (V_i + 2c)^{-1/2} \exp(-c/\tau^2) (1 - \exp(-c/\tau^2)). \end{aligned}$$

Therefore, $P(B_{GL,i} < \epsilon | \beta, \tau^2, y_i) \asymp \exp(-c/\tau^2)$ as $\tau^2 \rightarrow 0$.
 If $\pi_{\lambda^2}(\lambda_i^2) = \frac{1}{B(a,b)} (\lambda_i^2)^{a-1} (1 + \lambda_i^2)^{-a-b}$, $0 < a, b \leq 1$, then

$$\begin{aligned} N &\leq \frac{(V_i/\epsilon)^{-1/2}}{B(a,b)} \int_{c/\tau^2}^{\infty} (\lambda_i^2)^{-a} (1 + \lambda_i^2)^{-a-b} d\lambda_i^2 \\ &= \frac{(V_i/\epsilon)^{-1/2}}{B(a,b)} \int_{c/(c+\tau^2)}^1 x^{a-1} (1-x)^{b-1} dx \\ &\leq \frac{(V_i/\epsilon)^{-1/2}}{B(a,b)} \left(\frac{c}{c+\tau^2}\right)^{a-1} \int_{c/(c+\tau^2)}^1 (1-x)^{b-1} dx \\ &= \frac{(V_i/\epsilon)^{-1/2}}{bB(a,b)} \left(\frac{c}{c+\tau^2}\right)^{a-1} \left(\frac{\tau^2}{c+\tau^2}\right)^b. \end{aligned}$$

On the other hand,

$$\begin{aligned} N &\geq \frac{\exp(-\epsilon r_i^2/(2V_i))}{B(a,b)} \int_{c/\tau^2}^{\infty} (V_i + \lambda_i^2 \tau^2)^{-1/2} (\lambda_i^2)^{a-1} (1 + \lambda_i^2)^{-a-b} d\lambda_i^2 \\ &= \frac{\exp(-\epsilon r_i^2/(2V_i))}{B(a,b)} \int_{c/(c+\tau^2)}^1 (V_i + x\tau^2 / (1-x))^{-1/2} x^{a-1} (1-x)^{b-1} dx \\ &\geq \frac{\exp(-\epsilon r_i^2/(2V_i))}{B(a,b)} \left(\frac{\tau^2}{c+\tau^2}\right)^{b-1} \int_{c/(c+\tau^2)}^1 (V_i + \tau^2/(1-x))^{-1/2} dx \\ &\geq \frac{\exp(-\epsilon r_i^2/(2V_i))}{B(a,b)} \left(\frac{\tau^2}{c+\tau^2}\right)^{b-1} \left(\frac{\tau^2 V_i}{c+\tau^2} + \tau^2\right)^{-1/2} \\ &\quad \times \int_{c/(c+\tau^2)}^1 (1-x)^{1/2} dx \\ &= \frac{\exp(-\epsilon r_i^2/(2V_i))}{(3/2)B(a,b)} (V_i + c + \tau^2)^{-1/2} \left(\frac{\tau^2}{c+\tau^2}\right)^b. \end{aligned}$$

Therefore, $P(B_{GL,i} < \epsilon | \beta, \tau^2, y_i) \asymp \tau^{2b}$ as $\tau^2 \rightarrow 0$.
 Similarly,

$$\begin{aligned} P(B_{GL,i} > \epsilon | \beta, \tau^2, y_i) &= \frac{\int_0^{c/\tau^2} (V_i + \lambda_i^2 \tau^2)^{-1/2} \pi_{\lambda^2}(\lambda_i^2) \exp\left(-\frac{r_i^2}{2(V_i + \lambda_i^2 \tau^2)}\right) d\lambda_i^2}{\int_0^{\infty} (V_i + \lambda_i^2 \tau^2)^{-1/2} \pi_{\lambda^2}(\lambda_i^2) \exp\left(-\frac{r_i^2}{2(V_i + \lambda_i^2 \tau^2)}\right) d\lambda_i^2}. \quad (20) \end{aligned}$$

Let N_2 and D_2 denote the numerator and denominator in (20). We have

$$N_2 \leq V_i^{-1/2} \exp(-\epsilon r_i^2/(2V_i)) \int_0^{c/\tau^2} \pi_{\lambda^2}(\lambda_i^2) d\lambda_i^2 \leq 1$$

and for any $0 < \epsilon_1 < \epsilon$

$$\begin{aligned} D_2 &\geq \int_{V_i(1-\epsilon_1)/\epsilon_1}^{\infty} (V_i + \lambda_i^2 \tau^2)^{-1/2} \pi_{\lambda^2}(\lambda_i^2) \exp\left(-\frac{r_i^2}{2(V_i + \lambda_i^2 \tau^2)}\right) d\lambda_i^2 \\ &\geq \exp(-\epsilon_1 r_i^2/(2V_i)) \int_{V_i(1-\epsilon_1)/\epsilon_1}^{\infty} (V_i + \lambda_i^2 \tau^2)^{-1/2} \pi_{\lambda^2}(\lambda_i^2) d\lambda_i^2. \quad (21) \end{aligned}$$

As a result, $P(B_{GL,i} > \epsilon | \beta, \tau^2, y_i) \leq C \exp\{-\epsilon(1-\epsilon_1)r_i^2/(2V_i)\}$, where C is a constant does not depend on r_i and $P(B_{GL,i} > \epsilon | \beta, \tau^2, y_i) \rightarrow 0$ as $r_i \rightarrow \infty$. \square

Acknowledgements

The authors thank Drs. G. Datta, A. Mandal, A. Chakraborty, and J. Maples for providing the datasets, and the editor, associate editors, and three anonymous reviewers for insightful comments and suggestions which improved the article significantly.

Funding

Ghosh's research was partially funded by SES-1327359.

References

Ando, T. (2007), "Bayesian Predictive Information Criterion for the Evaluation of Hierarchical Bayesian and Empirical Bayes Models," *Biometrika*, 94, 443–458. [1486]

Armagan, A., Clyde, M., and Dunson, D. B. (2011), "Generalized Beta Mixtures of Gaussians," in *Advances in Neural Information Processing Systems*, eds. J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger, Curran Associates, Inc., pp. 523–531. [1477]

Armagan, A., Dunson, D. B., and Lee, J. (2012), "Generalized Double Pareto Shrinkage," *Statistica Sinica*, 23, 119–143. [1477]

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28–36. [1476]

Berger, J. (1980), "A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean," *The Annals of Statistics*, 8, 716–761. [1477]

— (2013), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer Science & Business Media. [1480]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [1477]

Chakraborty, A., Datta, G. S., and Mandal, A. (2016), "A Two-Component Normal Mixture Alternative to the Fay–Herriot Model," *Statistics in Transition New Series*, 17, 67–90. [1479]

Datta, G. S., and Ghosh, M. (1991), "Bayesian Prediction in Linear Models: Applications to Small Area Estimation," *The Annals of Statistics*, 19, 1748–1770. [1476]

Datta, G. S., Ghosh, M., Nangia, N., and Natarajan, K. (1996), "Estimation of Median Income of Four-Person Families: A Bayesian Approach," in *Bayesian Analysis in Statistics and Econometrics*, eds. D. A. Berry, K. M. Chaloner, and J. K. Geweke, New York: Wiley, pp. 129–140. [1477]

Datta, G. S., Hall, P., and Mandal, A. (2011), "Model Selection by Testing for the Presence of Small-Area Effects, and Application to Area-Level Data," *Journal of the American Statistical Association*, 106, 362–374. [1476]

Datta, G. S., and Lahiri, P. (1995), "Robust Hierarchical Bayes Estimation of Small Area Characteristics in the Presence of Covariates and Outliers," *Journal of Multivariate Analysis*, 54, 310–328. [1476]

- Datta, G. S., and Mandal, A. (2015), "Small Area Estimation With Uncertain Random Effects," *Journal of the American Statistical Association*, 110, 1735–1744. [1477,1478,1480,1484,1485]
- Datta, G. S., Rao, J. N. K., and Smith, D. D. (2005), "On Measuring the Variability of Small Area Estimators Under a Basic Area Level Model," *Biometrika*, 92, 183–196. [1476]
- Datta, J., and Ghosh, J. K. (2013), "Asymptotic Properties of Bayes Risk for the Horseshoe Prior," *Bayesian Analysis*, 8, 111–132. [1477]
- Fabrizi, E., and Trivisano, C. (2010), "Robust Linear Mixed Models for Small Area Estimation," *Journal of Statistical Planning and Inference*, 140, 433–443. [1476]
- Fay, M. P., and Graubard, B. I. (2001), "Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators," *Biometrics*, 57, 1198–1206. [1477]
- Fay, R. E., and Herriot, R. A. (1979), "Estimates of Income for Small Places: An Application of James–Stein Procedures to Census Data," *Journal of the American Statistical Association*, 74, 269–277. [1476]
- Fruehwirth-Schnatter, S., and Wagner, H. (2011), "Bayesian Variable Selection for Random Intercept Modeling of Gaussian and Non-Gaussian Data," in *Bayesian Statistics (Vol. 9)*, eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, New York: Oxford University Press, pp. 165–200. [1478]
- Gelfand, A. E., and Smith, A. F. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409. [1479]
- Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1, 515–534. [1479,1487]
- Ghosh, P., Tang, X., Ghosh, M., and Chakrabarti, A. (2016), "Asymptotic Properties of Bayes Risk of a General Class of Shrinkage Priors in Multiple Hypothesis Testing Under Sparsity," *Bayesian Analysis*, 11, 753–796. [1477]
- Griffin, J. E., and Brown, P. J. (2005), "Alternative Prior Distributions for Variable Selection With Very Many More Variables Than Observations," Technical Report, University of Warwick. [1477]
- (2010), "Inference With Normal-Gamma Prior Distributions in Regression Problems," *Bayesian Analysis*, 5, 171–188. [1478]
- Jiang, J., and Lahiri, P. (2006a), "Estimation of Finite Population Domain Means: A Model-Assisted Empirical Best Prediction Approach," *Journal of the American Statistical Association*, 101, 301–311. [1476]
- (2006b), "Mixed Model Prediction and Small Area Estimation," *Test*, 15, 1–96. [1476]
- Jiang, J., Lahiri, P., and Wan, S.-M. (2002), "A Unified Jackknife Theory for Empirical Best Prediction With M-Estimation," *The Annals of Statistics*, 30, 1782–1810. [1476]
- Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [1486]
- Lahiri, P., and Rao, J. N. K. (1995), "Robust Estimation of Mean Squared Error of Small Area Estimators," *Journal of the American Statistical Association*, 90, 758–766. [1476]
- Li, Y., and Lahiri, P. (2007), "Robust Model-Based and Model-Assisted Predictors of the Finite Population Total," *Journal of the American Statistical Association*, 102, 664–673. [1476]
- Molina, I., Nandram, B., and Rao, J. N. K. (2014), "Small Area Estimation of General Parameters With Application to Poverty Indicators: A Hierarchical Bayes Approach," *The Annals of Applied Statistics*, 8, 852–885. [1476]
- Morales, D., Pagliarella, M. C., and Salvatore, R. (2015), "Small Area Estimation of Poverty Indicators Under Partitioned Area-Level Time Models," *Statistics and Operations Research Transactions*, 39, 19–34. [1476]
- Morris, C., and Tang, R. (2011), "Estimating Random Effects Via Adjustment for Density Maximization," *Statistical Science*, 26, 271–287. [1480]
- Polson, N. G., and Scott, J. G. (2009), "Alternative Global–Local Shrinkage Rules Using Hypergeometric–Beta Mixtures," Technical Report 14, Duke University, Department of Statistical Science. [1477]
- (2010), "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," *Bayesian Statistics*, 9, 501–538. [1477]
- (2012a), "Local Shrinkage Rules, Lévy Processes and Regularized Regression," *Journal of the Royal Statistical Society, Series B*, 74, 287–311. [1477]
- (2012b), "On the Half-Cauchy Prior for a Global Scale Parameter," *Bayesian Analysis*, 7, 887–902. [1477,1479]
- Rao, J. N., and Molina, I. (2015), *Small Area Estimation*, Hoboken, NJ: Wiley. [1476]
- Scott, J. G. (2011), "Bayesian Estimation of Intensity Surfaces on the Sphere Via Needlet Shrinkage and Selection," *Bayesian Analysis*, 6, 307–327. [1477]
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society, Series B*, 64, 583–639. [1483]
- Strawderman, W. E. (1971), "Proper Bayes Minimax Estimators of the Multivariate Normal Mean," *The Annals of Mathematical Statistics*, 42, 385–388. [1477]
- Tarozzi, A., and Deaton, A. (2009), "Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas," *The Review of Economics and Statistics*, 91, 773–792. [1476]
- You, Y., and Chapman, B. (2006), "Small Area Estimation Using Area Level Models and Estimated Sampling Variances," *Survey Methodology*, 32, 97. [1477]