Robust Empirical Bayes Small Area Estimation with Density Power Divergence

SHONOSUKE SUGASAWA
Risk Analysis Research Center, The Institute of Statistical Mathematics

Abstract. Empirical Bayes estimators are widely used to provide indirect and model-based estimates of means in small areas. The most common model is two-stage normal hierarchical model called Fay-Herriot model. However, due to the normality assumption, it can be highly influenced by the presence of outliers. In this article, we propose a simple modification of the conventional method by using density power divergence and derive a new robust empirical Bayes small area estimator. Based on some asymptotic properties of the robust estimator of the model parameters, we obtain an expression of second order approximation of the mean squared error of the proposed empirical Bayes estimator. We investigate some numerical performances of the proposed method through simulations and a real data application.

Key words: density power divergence; empirical Bayes; Fay-Herriot model; maximum likelihood estimation; mean squared error

1 Introduction

Direct survey estimators, based only on the area-specific sample data, are known to yield unacceptably large standard errors if the area-specifics sample sizes are small. Hence, it is necessary to “borrow strength” from related areas to increase the effective sample size and to provide indirect estimators with higher precision of estimates. Such indirect estimators are often based on mixed models and associated empirical Bayes estimators in which random effects represents area-specific effects. For comprehensive overviews and appraisals of models and methods for small area estimation, see Pfeffermann (2013) and Rao and Molina (2015).

The basic area-level model for small area estimation is a two-stage normal hierarchical model proposed in Fay and Herriot (1979), usually called Fay-Herriot model in small area estimation, described as

\[ y_i \sim N(\theta_i, D_i), \quad \theta_i \sim N(x_i^T \beta, A), \quad i = 1, \ldots, m, \tag{1} \]

where \( y_i \) is the direct estimator of the true small area mean \( \theta_i \), \( D_i \) is the sampling variances assumed to be known, \( x_i \) and \( \beta \) are, respectively, a vector of covariates and regression coefficients, and \( A \) is an unknown variance parameter. The hierarchical model [1] is also expressed as the simple random effect model:

\[ y_i = x_i^T \beta + v_i + \varepsilon_i, \quad v_i \sim N(0, A), \quad \varepsilon_i \sim N(0, D_i), \]

*This version: August 27, 2018
in which \( v_i \) corresponds to a random area effect and the small area mean \( \theta_i \) corresponds to \( x_i^T \beta + v_i \). Under the model (1), the Bayes estimator of \( \theta_i \) under squared error loss is given by

\[
\tilde{\theta}_i = y_i - \frac{D_i}{A + D_i} (y_i - x_i^T \beta),
\]

(2)

which shrinks the direct estimates \( y_i \) to the synthetic (regression) mean \( x_i^T \beta \) and the amount of shrinkage is determined by the random effect variance \( A \) and the sampling variance \( D_i \). Since the Bayes estimator (2) depends on unknown model parameters \( \phi = (\beta^T, A)^T \), we need to estimate \( \phi \) from the data. The standard method is the maximum likelihood (ML) method by maximizing the log-marginal likelihood function:

\[
\log f(y; \phi) = -\frac{m}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{m} \log(A + D_i) - \frac{1}{2} \sum_{i=1}^{m} \frac{(y_i - x_i^T \beta)^2}{A + D_i},
\]

(3)

with \( y = \{y_1, \ldots, y_m\} \), which comes from \( y_i \sim N(x_i^T \beta, A + D_i) \) under (1). The estimator \( \hat{\theta}_i \) by replacing \( \phi \) with \( \hat{\phi} \) is called empirical Bayes (EB) estimator.

Although the Fay-Herriot model and the associated EB estimator is useful in practice, several problems have been addressed so far and new approaches have been developed. They includes measurement errors in \( x_i \) (Ybarra and Lohr, 2008), non-parametric formulation for the regression part (Opsomer et al, 2008), pretest and model averaging estimator (Datta et al., 2011; Datta and Mandal, 2015), shrinkage estimation both means and variances (Sugasawa et al., 2016), spatio-temporal modeling (Marhuenda et al., 2013). Moreover, robust methods, in the case that there exist outlier samples in \( y \), have been recognized as one of the most important issues in small area estimation. So far, Ghost al. (2008) proposed a robust Bayes estimator in the Fay-Herriot model (1) based on influence functions, and Sinha and Rao (2009) proposed the use of Huber’s (1973) \( \psi \)-function in the general linear mixed models. Apart from the area-level data, other robust methods for unit-level data have been proposed in, for example, Chambers et al. (2014), Chambers and Tzavidis (2006) and Dongmo-Jiongo et al. (2013).

In this paper, we focus on the Fay-Herriot model, a basic area-level model, and develop a robust empirical Bayes method. We first remark that, under the model (1), the Bayes estimator (2) can be expressed as

\[
\tilde{\theta}_i = y_i + D_i \frac{\partial}{\partial y_i} \log f(y; \phi),
\]

(4)

thereby the empirical Bayes estimator is characterized by the log-marginal likelihood \( \log f(y; \phi) \). A new robust method proposed in this paper is based on the idea that the log-marginal likelihood \( \log f(y; \phi) \) is replaced with a robust alternative. To this end, we use density power divergence, a family of divergence including Kullback-Leibler divergence relating to ML estimators as a special case, proposed in Basu et al. (1998). We call robust marginal likelihood of the quasi-likelihood function from density power divergence, and it turns out that the robust marginal likelihood has a simple analytical expression. We derive a quasi-Bayes estimator based on the robust log-marginal likelihood and shows that the resulting quasi-Bayes estimator can be regarded as a natural generalization of the classical Bayes estimator (2) and has a nice property called tail robustness (Carvalho et al., 2010). Moreover, the model
parameters are estimated by maximizing the robust likelihood function and we reveal some asymptotic properties. We also derive an asymptotic expansion of the mean squared error (MSE) of the proposed robust empirical Bayes estimator and construct a second order unbiased estimator of MSE via parametric bootstrap.

This paper is organized as follows: In Section 2, we derive a robust Bayes estimator based on density power divergence and discuss some properties. In Section 3, we consider robust ML estimation of model parameters and derive an approximation of MSE of the robust empirical Bayes estimator based on the asymptotic properties of the robust ML estimator. We also provide a parametric bootstrap method for estimating MSE. In Section 4, we provide some results of simulation studies and an application to a real data set. In Section 5, we give some discussions. All the technical proofs are given in Appendix.

2 Density Power Divergence and Robust Bayes Estimator

2.1 Density power divergence

It is well-known that the maximum likelihood (ML) estimator minimizes the empirical estimates of Kullback-Leibler distance. However, it is often criticized that the ML estimator is sensitive to the distributional assumption and might perform poorly if the assumed distribution is misspecified or there exist outliers. To overcome the problem, Basu et al. (1998) introduced an estimating method based on density power divergence for identically distributed samples. However, in the Fay-Herriot model (1), we focus on, the samples \( y_i \), \( i = 1, \ldots, m \), are not identically distributed, thereby we need a small modification of the estimating method in Basu et al. (1998). To begin with, let \( y_i \), \( i = 1, \ldots, m \), be the independent random variable which has a density function \( f_i(y_i; \phi) \) with an unknown parameter \( \phi \). From Basu et al. (1998), the minimum density power divergence estimator of \( \phi \) based on a single sample \( y_i \) is obtained by maximizing

\[
\frac{1}{\alpha} f_i(y_i; \phi)^\alpha - \frac{1}{1 + \alpha} \int f_i(t; \phi)^{1+\alpha} dt.
\]

Thus, in this paper, we propose to use the function

\[
L_\alpha(y; \phi) = \frac{1}{\alpha} \sum_{i=1}^{m} f_i(y_i; \phi)^\alpha - \frac{1}{1 + \alpha} \sum_{i=1}^{m} \int f_i(t; \phi)^{1+\alpha} dt,
\]

instead of the log-likelihood function \( m^{-1} \sum_{i=1}^{m} \log f(y_i; \phi) \). Here \( \alpha \) controls the trade-off between robustness and asymptotic efficiency with \( 0 < \alpha < 1 \), noting that

\[
\lim_{\alpha \to 0} \left\{ L_\alpha(y; \phi) - m \left( \frac{1}{\alpha} - 1 \right) \right\} = \sum_{i=1}^{m} \log f_i(y_i; \phi),
\]

thereby the function \( L_\alpha \) is a natural extension of the log-likelihood function. Throughout the paper, we call \( L_\alpha \) robust likelihood function.
2.2 Robust Bayes estimator

When \( y_i \) follows the model (1), \( y_1, \ldots, y_m \) are independent and \( y_i \sim N(x_i^t \beta, A + D_i) \), thereby (5) can be expressed as

\[
L_\alpha(y; \phi) = \frac{1}{m} \sum_{i=1}^{m} \left\{ \frac{s_i(y_i; \phi)}{\alpha} - \frac{V_i^\alpha}{(1 + \alpha)^{3/2}} \right\},
\]

where \( V_i = 1/\sqrt{2\pi(A + D_i)} \) and

\[
s_i(y_i; \phi) = V_i^\alpha \exp \left\{ -\frac{\alpha(y_i - x_i^t \beta)^2}{2(A + D_i)} \right\}.
\]

We propose to use the function (6) instead of the log-marginal likelihood \( \log f(y; \phi) \) given by (3). We define quasi-Bayes estimator \( \tilde{\theta}_i^R \) of \( \theta_i \) as the estimator obtained by replacing \( L_\alpha(y; \phi) \) with \( \log f(y; \phi) \) in (2). Since we have

\[
\frac{\partial}{\partial y_i} L_\alpha(y; \phi) = \frac{\partial}{\partial y_i} s_i(y_i; \phi) = -\frac{D_i}{A + D_i} (y_i - x_i^t \beta) s_i(y_i; \phi),
\]

the quasi-Bayes estimator is

\[
\tilde{\theta}_i^R = y_i - \frac{D_i}{A + D_i} (y_i - x_i^t \beta) s_i(y_i; \phi),
\]

which we call robust Bayes estimator in this paper. It is observed that the shrinkage factor in the robust Bayes estimator \( \tilde{\theta}_i^R \) is \( s_i(y_i; \phi)D_i/(A + D_i) \) which depends on \( y_i \) while the shrinkage factor in the conventional Bayes estimator \( \hat{\theta}_i^R \) is \( D_i/(A + D_i) \). We also remark that \( \tilde{\theta}_i^R \) reduces to \( \hat{\theta}_i \) in (2) when \( \alpha = 0 \) since \( s_i(y_i; \phi)|_{\alpha=0} = 1 \).

The selection of \( \alpha \) controlling the robustness is often dictated by the user’s point of view. Here we propose a concrete selection method of \( \alpha \) based on the mean squared error (MSE) of the robust Bayes estimator \( \tilde{\theta}_i^R \), following Ghosh et al. (2008). The MSE formula is given in the following theorem, proved in the Appendix.

**Theorem 1.** Under the model (1), it holds

\[
\mathbb{E}[(\tilde{\theta}_i^R - \hat{\theta}_i)^2] = g_{1i}(A) + g_{2i}(A),
\]

where \( g_{1i}(A) = AD_i/(A + D_i) \) and

\[
g_{2i}(A) = \frac{D_i^2}{A + D_i} \left\{ \frac{V_i^{2\alpha}}{(2\alpha + 1)^{3/2}} - \frac{2V_i^\alpha}{(\alpha + 1)^{3/2} + 1} \right\}
\]

and \( g_{2i}(A) \) is increasing in \( \alpha \).

It is remarked that the MSE of the usual Bayes estimator \( \hat{\theta}_i \) given in (2) is \( g_{1i}(A) \), thereby the excess MSE of \( \tilde{\theta}_i^R \) over \( \hat{\theta}_i \) is \( g_{2i}(A) \), which clearly tends to 0 as \( \alpha \to 0 \) and is increasing in \( \alpha \) as shown in Theorem 1. Hence, there is a trade-off between robustness of \( \tilde{\theta}_i^R \) and the MSE evaluated under the model (1). We define \( \text{Ex}(\alpha) = \sum_{i=1}^{m} g_{2i}(A)/\sum_{i=1}^{m} g_{1i}(A) \) as the total excess MSE. We propose to select \( \alpha \) such that \( \text{Ex}(\alpha) \) does not exceed a user-specified percentage \( c\% \), namely we compute \( \alpha^* \) satisfying \( \text{Ex}(\alpha^*) = c/100 \). Such \( \alpha^* \) can not be obtained in an analytical way, but it can be easily obtained equation can be solved by simple numerical methods, for example, the bisectional method (Burden and Faires, 2010) that repeatedly bisects an interval and selects a subinterval in which a root exists until the process converges numerically.
2.3 Related robust Bayes small area estimators

As related robust (empirical) Bayes small area estimator under the model (1), Ghosh et al. (2008) proposed the form

$$\tilde{\theta}_i^G = y_i - \frac{D_i \sqrt{v_i(A)}}{A + D_i} \psi_K \left( \frac{y_i - \mathbf{x}_i \hat{\beta}(A)}{\sqrt{v_i(A)}} \right),$$  \hspace{1cm} (9)

where

$$\hat{\beta}(A) = \left( \sum_{i=1}^{m} \frac{\mathbf{x}_i \mathbf{x}_i^T}{A + D_i} \right)^{-1} \left( \sum_{i=1}^{m} \frac{\mathbf{x}_i y_i}{A + D_i} \right), \quad v_i(A) = A + D_i - \mathbf{x}_i^T \left( \sum_{i=1}^{m} \frac{\mathbf{x}_i \mathbf{x}_i^T}{A + D_i} \right)^{-1} \mathbf{x}_i,$$

and $\psi_K(t) = u \min(1, K/|u|)$ is the Huber’s $\psi$-function with a tuning constant $K > 0$. Similarly, Sinha and Rao (2009) used the Huber’s $\psi$-function to modify an estimating equation for $\hat{\theta}_i$, and suggested a robust estimator $\tilde{\theta}_i^{SR}$ as a solution of the equation:

$$\frac{1}{\sqrt{D_i}} \psi_K \left( \frac{y_i - \tilde{\theta}_i^{SR}}{\sqrt{D_i}} \right) - \frac{1}{\sqrt{A}} \psi_K \left( \frac{\tilde{\theta}_i^{SR} - \mathbf{x}_i \hat{\beta}}{\sqrt{A}} \right) = 0. \hspace{1cm} (10)$$

On the other hand, Datta and Lahiri (1995) developed a robust hierarchical Bayes method by using Cauchy distributions for random effects distribution in known outlying areas, and showed that the resulting Bayes estimator $\hat{\theta}_i$ holds that $\hat{\theta}_i - y_i \to 0$ as $|y_i| \to \infty$, which is known as tail robustness (Carvalho et al, 2010). Bayes estimators with tail robustness is desirable in this context since it does not over-shrink the direct estimator $y_i$ when $y_i$ is outlier. However, the Bayes estimator (9) is not tail robust since $|\tilde{\theta}_i^R - y_i| \to KD_i \sqrt{v_i(A)}/(A + D_i)$. Moreover, if $|y_i - \tilde{\theta}_i^{SR}| \to 0$ as $|y_i| \to \infty$, the left side of (10) reduces to $-K/\sqrt{A}$, so that $\tilde{\theta}_i^{SR}$ is not tail robust either. On the other hand, the proposed robust Bayes estimator (8) is clearly tail robust since $(y_i - \mathbf{x}_i \hat{\beta}) s_i(y_i; \phi) \to 0$ as $|y_i| \to \infty$. Hence, the proposed estimator would shrink less than $\tilde{\theta}_i^G$ and $\tilde{\theta}_i^{SR}$ when $y_i$ seems outlier.

3 Robust Empirical Bayes Estimation and Mean Squared Error Evaluation

3.1 Robust parameter estimation

For estimating the model parameter $\phi$, outliers should be omitted. To this end, we define the robust estimator $\tilde{\phi}$ of $\phi$ as $\tilde{\phi} = \arg\max L_\alpha(y; \phi)$. Hence, the robust estimator $\tilde{\phi}$ satisfies the estimating equation:

$$\frac{\partial L_\alpha}{\partial \beta} = \sum_{i=1}^{m} \frac{\mathbf{x}_i s_i(y_i; \phi)(y_i - \mathbf{x}_i \hat{\beta})}{A + D_i} = 0 \hspace{2cm} (11)$$

$$\frac{\partial L_\alpha}{\partial A} = \sum_{i=1}^{m} \left\{ \frac{(y_i - \mathbf{x}_i \hat{\beta})^2 s_i(y_i; \phi)}{(A + D_i)^2} - \frac{s_i(y_i; \phi)}{A + D_i} + \frac{\alpha V_i^\alpha}{(\alpha + 1)^{3/2}(A + D_i)} \right\} = 0.$$
It is noted that the straightforward evaluation calculation shows that the above estimating equation is unbiased under the model \([1]\). The equation \([11]\) can be numerically solved by the following fixed-point iteration process:

\[
\beta(t+1) = \left\{ \frac{m}{i=1} \frac{s_i(t)x_i'y_i}{A(t) + D_i} \right\}^{-1} \left\{ \frac{m}{i=1} \frac{s_i(t)x_i'y_i}{A(t) + D_i} \right\}
\]

where \(r_{i(t)} = s_i(t) - \alpha V_i(t)/(1 + \alpha)^{3/2}, s_i(t) = s_i(y_i; \phi(t)) \) and \(V_i(t) = 1/\sqrt{2\pi(A(t) + D_i)}, \) and the subscript \(^{(t)}\) denotes the time of iteration. A reasonable starting value would be the ML estimate obtained by maximizing \([3]\). We repeat \([12]\) until numerical convergence and obtain the estimator \(\hat{\phi} \). Substituting the robust estimator \(\hat{\phi} \) into the robust Bayes estimator \([8]\), we obtain the robust empirical Bayes estimator \(\tilde{\theta}^R = \tilde{\theta}^R(y_i, \hat{\phi}). \)

We next consider the asymptotic properties of the robust estimator under the model \([1]\). To this end, we assume the following regularity conditions:

(C1) \(0 < D_s \leq \min_{1 \leq i \leq m} D_i \leq \max_{1 \leq i \leq m} D_i \leq D^* < \infty. \)

(C2) \(\max_{1 \leq i \leq m} x_i'(X'X)^{-1}x_i = O(m^{-1}), \) where \(X = (x_1, \ldots, x_m)' \).

The assumption regarding the uniformly boundedness of the sampling variance \(D_i \) is required in deriving an asymptotic expression of mean squared error in small area estimation, for example, see Prasad and Rao (1990), Datta et al. (2005).

Concerning the asymptotic behavior of the robust estimator of \(\hat{\phi} \), we obtain the following results.

**Theorem 2.** Under conditions (C1) and (C2), \(\hat{\beta} \) and \(\hat{\alpha} \) are asymptotically independent and distributed as \(N(\beta, m^{-1}J^{-1}_iK^i \beta J^{-1}_i) \) and \(N(A, K_A/mJ_A^2) \), respectively, where

\[
J_\beta = \frac{1}{m(\alpha + 1)^{3/2}} \sum_{i=1}^{m} \frac{V_i(2 - \alpha)(\alpha^2 + \alpha + 1)}{(A + D_i)^2}, \quad J_A = \frac{1}{2m} \sum_{i=1}^{m} \frac{V_i(2 - \alpha)(\alpha^2 + \alpha + 1)}{(A + D_i)^2},
\]

\[
K_\beta = \frac{1}{m(2\alpha + 1)^{3/2}} \sum_{i=1}^{m} \frac{V_i(2 - \alpha)(\alpha^2 + \alpha + 1)}{(A + D_i)^2}, \quad K_A = \frac{1}{m} \sum_{i=1}^{m} \frac{V_i(2 - \alpha)(\alpha^2 + \alpha + 1)}{(A + D_i)^2} \left\{ \frac{2(2\alpha^2 + 1)}{(2\alpha + 1)^{5/2}} - \frac{\alpha^2}{(\alpha + 1)^3} \right\}.
\]

The proof is given in the Appendix. It is noted that, when \(\alpha = 0, \) the asymptotic (covariance matrix) variance is given by

\[
\left. J^{-1}_\beta K_\beta J^{-1}_\beta \right|_{\alpha=0} = \left( m^{-1} \sum_{i=1}^{m} \frac{x_i x_i'}{(A + D_i)} \right)^{-1}, \quad \left. K_A^{-1} \right|_{\alpha=0} = \left( 2m^{-1} \sum_{i=1}^{m} (A + D_i)^{-2} \right)^{-1},
\]

which coincides with the asymptotic (covariance matrix) variance of the maximum likelihood estimator of \(\beta \) and \(A \) (Datta and Lahiri, 2000). This is consistent to the fact that the robust likelihood \([6]\) reduces to the conventional likelihood function \([3]\).
### 3.2 Mean squared error evaluation

For risk evolution of the estimator $\hat{\theta}_i^R$, we consider the mean squared error (MSE), $\text{E}[(\hat{\theta}_i^R - \theta_i)^2]$, where the expectation is taken with respect to the joint distribution of $\theta_i$’s and $y_i$’s following the model $\{\}$. The MSE can be regarded as the integrated Bayes risk, this type of MSE has been used in the context of small area estimation (Prasad and Rao, 1990).

Since $\hat{\theta}_i^R$ depends on the estimator $\hat{\phi}$, the MSE $\text{E}[(\hat{\theta}_i^R - \theta_i)^2]$ takes the variability of $\hat{\phi}$ into account as well. Thus, we cannot exactly evaluate the MSE unlike Theorem 3. Conventionally, we use a second order approximation of the MSE, that is, we derive an asymptotic expansion of the expectation $\text{E}[(\hat{\theta}_i^R - \theta_i)^2]$ with an analytical expression of $m^{-1}$. The approximation formula is given in the following theorem proved in Appendix.

**Theorem 3.** Under (C1) and (C2), it holds

$$
\text{E}[(\hat{\theta}_i^R - \theta_i)^2] = g_{3i}(A) + g_{2i}(A) + \frac{g_{3i}(A)}{m} + \frac{g_{4i}(A)}{m} + \frac{2g_{5i}(A)}{m} + o(m^{-1}), \quad (13)
$$

where $g_{3i}(A)$ and $g_{2i}(A)$ are given in Theorem 7 and

$$
g_{3i}(A) = \frac{D^2V_{2\alpha}}{B_i^2(2\alpha + 1)^{3/2}}x_i^tJ^{-1}_\beta K_\beta J^{-1}_\beta x_i, \quad g_{4i}(A) = \frac{D^2V_{2\alpha}K_A}{B_i^3(2\alpha + 1)^{7/2}J_A}\left(\alpha^4 - \frac{1}{2} \alpha^2 + 1\right)
$$

$$
g_{5i}(A) = \frac{\alpha D^2x_i^tJ^{-1}_\beta K_\beta J^{-1}_\beta x_i}{2B_i^4(3B_iC_{11} - \alpha C_{21}) + \frac{D^2K_A}{24B_i^6J_A}\left\{3\alpha B_i^2C_{21} + (\alpha - 2)(3\alpha + 8)C_{11}\right\}}
$$

$$
+ \frac{D^2x_i^tJ^{-1}_\beta x_i}{B_i^4(3B_iC_{12} - \alpha C_{22}) + \frac{D^2}{2B_i^6J_A}\left\{\alpha C_{32} - 2B_iC_{22} + (2 - \alpha)B_i^2C_{12}\right\}}
$$

$$
+ \frac{D^2}{2B_i^4}\left\{b_A - \frac{\alpha V_{1\alpha}}{(\alpha + 1)^{3/2}B_iJ_A}\right\}\left\{(2 - \alpha)B_iC_{11} - \alpha C_{21}\right\},
$$

where $B_i = A + D_i$, $b_A = \lim_{m \to \infty} m\text{E}[^{\hat{\theta}_1^R - A}]$ is the first order bias of $\hat{\theta}_1^R$, and

$$
C_{jk} = (2j - 1)!!B_i^j\left\{V_{1\alpha}(k \alpha + 1)^{-j-1/2} - V_{1\alpha}(k \alpha + 1)^{-j-1/2}\right\}.
$$

The proof of Theorem 3 is given in the Appendix. It is noted that $C_{jk}|_{\alpha=0} = 0$, thereby it holds $g_{5i}(A)|_{\alpha=0} = 0$. Hence, the approximation of MSE given in (13) rescues to

$$
\text{E}[(\hat{\theta}_i^R - \theta_i)^2]|_{\alpha=0} = \frac{AD_i}{A + D_i} + \frac{D_i^2}{(A + D_i)^2}\sum_{i=1}^{m} \frac{x_i x_i^t}{A + D_i} - 1 x_i
$$

$$
+ \frac{2D_i^2}{(A + D_i)^2}\sum_{i=1}^{m} (A + D_i)^{-2} - 1 + o(m^{-1}),
$$

which corresponds to the approximation of MSE of the classical empirical Bayes estimator under the Fay-Herriot model, given in Datta and Lahiri (2000) and Datta et al. (2005).
3.3 Estimation of mean squared error

The approximation of the MSE given in Theorem [3] depends on unknown parameter \( A \), so that it cannot be directly used in practice. Then, we use a second order unbiased estimator of the MSE. Here a estimator \( \hat{T} \) is called second order unbiased if it satisfies \( E[\hat{T}] = T + o(m^{-1}) \). As shown in Theorem [3], \( g_{31}(A) \), \( g_{41}(A) \) and \( g_{51}(A) \) are smooth functions of \( A \), so that the plug-in estimators \( g_{31}(\hat{A}) \), \( g_{41}(\hat{A}) \) and \( g_{51}(\hat{A}) \) are second order unbiased. On the other hand, the plug-in estimators \( g_{11}(\hat{A}) \) and \( g_{21}(\hat{A}) \) have considerable bias since \( g_{11}(A) \) and \( g_{21}(A) \) are of \( O(1) \). Since the analytical derivation of these biases and bias corrected estimator of these terms requires tedious algebraic calculations, we here use the parametric bootstrap method, following Butar and Lahiri (2003). A second order unbiased MSE estimator we propose is given by

\[
\hat{M}_i = 2g_{11}(\hat{A}) + 2g_{21}(\hat{A}) - E^* \left[ g_{11}(\hat{A}) + g_{21}(\hat{A}) \right] + \frac{g_{31}(\hat{A})}{m} + \frac{g_{41}(\hat{A})}{m} + \frac{2g_{51}(\hat{A})}{m},
\]

where \( E^*[\cdot] \) denotes the expectation with respect to the parametric bootstrap samples defined as

\[
y_i^s = x_i^t\hat{\beta} + v_i^s + \varepsilon_i^s, \quad v_i^s \sim N(0, \hat{A}), \quad \varepsilon_i^s \sim N(0, D_i).
\]

As proved in Appendix, we obtain the following theorem.

**Theorem 4.** Under (C1) and (C2), the MSE estimator \( \hat{M}_i \) in (14) is second order unbiased, namely \( E[\hat{M}_i] = E[(\hat{\theta}_i^R - \theta_i)^2] + o(m^{-1}) \), where the expectations are taken with respect to the joint distribution of \( \theta_i \)'s and \( y_i \)'s following the model [1].

In (14), we used an additive form of bias correction for simplicity, but other forms of bias correction are available, see Hall and Maiti (2006). Concerning \( g_{51}(\hat{A}) \), we need to compute the estimate of \( b_A \), the first order bias of \( \hat{A} \), which can be calculated from the parametric bootstrap samples. Alternatively, one may use the parametric bootstrap method for directly computing \( g_{51}(\hat{A}) \). As show in the proof of Theorem [3], \( E[(\hat{\theta}_i^R - \hat{\theta}_i^R)(\hat{\theta}_i^R - \hat{\theta}_i)] = m^{-1}g_{51}(\hat{A}) + o(m^{-1}) \), thereby one can use

\[
E^* \left\{ \left( \hat{\theta}_i^R(y_i^s, \hat{\phi}) - \hat{\theta}_i^R(y_i^s, \hat{\phi}) \right) \left( \hat{\theta}_i^R(y_i^s, \hat{\phi}) - \hat{\theta}_i(y_i^s, \hat{\phi}) \right) \right\}
\]

instead of \( m^{-1}g_{51}(\hat{A}) \).

4 Numerical Studies

4.1 Prediction error comparison

We first investigate the prediction errors of the proposed robust method compared with some existing methods. To this end, we consider the following Fay-Herriot model:

\[
y_i = \theta_i + \varepsilon_i, \quad \theta_i = \beta_0 + \beta_1 x_i + \sqrt{A}u_i, \quad i = 1, \ldots, m
\]

with \( m = 30 \) and \( \varepsilon_i \sim N(0, D_i) \). The auxiliary variables \( x_i \) were initially generated from the uniform distribution on \((0, 1)\), which are fixed throughout the simulation experiments. The regression coefficients were fixed at \((\beta_0, \beta_1) = (0, 2)\) and we set \( A = 0.5 \). Concerning the sampling variance \( D_i \), we divided \( m \) areas into five groups with equal number of areas, and we set the same value of \( D_i \) within the same groups. The
group $D_i$-pattern we considered was $(0.2, 0.4, 0.6, 0.8, 1.0)$. Regarding the distribution of the standardized random area effect $u_i$, we considered the following scenarios:

(I) $u_i \sim N(0, 1)$,  
(II) $u_i \sim 0.85N(0, 1) + 0.15N(0, 10^2)$,  
(III) $u_i \sim 0.70N(0, 1) + 0.30N(0, 10^2)$,  
(IV) $u_i \sim 0.85N(0, 1) + 0.15t_5(0, 7^2)$,  
(V) $u_i \sim 0.85N(0, 1) + 0.15\chi^2_5(0, 7^2)$,

where $t_5(a, b)$ and $\chi^2_5(a, b)$ denote $t$- and chi-square distribution with 5 degrees of freedom, respectively, scaled to mean $a$ and variance $b$. In scenario (I), there is no outlying areas, while the random effect distribution $u_i$ is contaminated by distributions with large variances in the other scenarios.

As working methods for estimating $\theta_i$, we considered six methods: the proposed robust empirical Bayes estimator based on density power divergence with 1% (DPEB1) and 5% (DPEB2) excess MSE, the classical empirical Bayes (EB) estimator from Datta and Lahiri (2000), the robust Bayes estimator (10) from Sinha and Rao (2009) with model parameters estimated by the robust likelihood equation given in Sinha and Rao (2009) (REB1) or the ML method (REB2), and the robust Bayes estimator (9) from Ghosh et al. (2008) (GEB) with the ML estimator $\hat{A}$. Note that we used the ML estimator in applying the EB and GEB methods, and $K = 1.345$ for tuning parameters in Huber’s $\psi$-function. Based on $R = 5000$ iterations, we computed MSE as

$$MSE_i = \frac{1}{R} \sum_{r=1}^{R} \left( \hat{\theta}_i^{(r)} - \theta_i^{(r)} \right)^2,$$

where $\theta_i^{(r)}$ and $\hat{\theta}_i^{(r)}$ are estimated and true values of $\theta_i$, respectively, in the $r$th iteration. The resulting MSE values were averaged over the areas in the same groups, which are reported in Table 1 for five scenarios. Since there is no outlier in scenario (I), Table 1 shows that the classical EB estimator performs the best. However, it is important to point out that the proposed two methods, DPREB1 and DEREB2, provide similar MSE values to EB. For scenario (II)-(V) in which there exists some outliers, EB performs poorly as expected, compared with other robust methods except for REB1. Regarding the performances of REB1, we have found that the robust estimating equation for estimating model parameters tend to produce unstable estimates, which results relatively high MSE values in our study. Among robust methods, the proposed DPEB1 and DPEB2 perform better than the other existing methods in most cases, which might come from the tail robustness of the proposed method as discuss in Section 2.3. Concerning DPEB1 and DPEB2, we can observe from Table 1 that DPEB2 provides smaller MSE values than DPEB1 since DPEB2 allows larger excess MSE and more robust than DPEB1. On the other hand, in scenario (I), DPEB1 works better than DPEB2 due to the same reason.

4.2 Finite sample behavior of MSE estimators

We next investigate finite sample behaviors of MSE estimators for the proposed estimator. To this end, we considered the same data generating model used in Section 4.1 except that we considered $m = 20$ in this study. For the distribution of $u_i$, we used scenario (I), (II) and (III). To begin with, we computed the true values of MSE based on 5000 simulation runs in the same manner as in Section 4.1. As MSE estimators,
Table 1: Group averaged mean squared errors of the robust empirical Bayes estimator from the proposed method (DPREB1, DPREB2), method of Sinha and Rao (2009) (REB1, REB2), method of Ghosh et al. (2008) (GEB), and the classical (non-robust) empirical Bayes estimator (EB) from Datta and Lahiri (2000).

| Scenario | Group | DPEB1 | DPEB2 | EB | REB1 | REB2 | GEB |
|----------|-------|-------|-------|----|------|------|-----|
| (I)      | 1     | 0.158 | 0.159 | 0.159 | 0.175 | 0.161 | 0.158 |
|          | 2     | 0.252 | 0.256 | 0.254 | 0.279 | 0.261 | 0.257 |
|          | 3     | 0.322 | 0.328 | 0.323 | 0.349 | 0.331 | 0.331 |
|          | 4     | 0.366 | 0.376 | 0.366 | 0.389 | 0.373 | 0.382 |
|          | 5     | 0.391 | 0.406 | 0.387 | 0.409 | 0.394 | 0.417 |
| (II)     | 1     | 0.186 | 0.178 | 0.192 | 0.399 | 0.188 | 0.188 |
|          | 2     | 0.341 | 0.318 | 0.363 | 0.954 | 0.351 | 0.350 |
|          | 3     | 0.509 | 0.463 | 0.548 | 1.967 | 0.524 | 0.522 |
|          | 4     | 0.637 | 0.569 | 0.698 | 2.664 | 0.659 | 0.651 |
|          | 5     | 0.759 | 0.670 | 0.854 | 3.645 | 0.797 | 0.779 |
| (III)    | 1     | 0.195 | 0.191 | 0.197 | 0.240 | 0.195 | 0.195 |
|          | 2     | 0.382 | 0.368 | 0.390 | 0.954 | 0.384 | 0.384 |
|          | 3     | 0.563 | 0.537 | 0.580 | 1.601 | 0.567 | 0.567 |
|          | 4     | 0.736 | 0.695 | 0.762 | 1.615 | 0.743 | 0.741 |
|          | 5     | 0.888 | 0.829 | 0.926 | 2.226 | 0.895 | 0.892 |
| (IV)     | 1     | 0.181 | 0.175 | 0.185 | 0.254 | 0.181 | 0.181 |
|          | 2     | 0.336 | 0.321 | 0.352 | 0.713 | 0.338 | 0.337 |
|          | 3     | 0.475 | 0.446 | 0.507 | 1.345 | 0.480 | 0.474 |
|          | 4     | 0.576 | 0.536 | 0.629 | 1.976 | 0.588 | 0.575 |
|          | 5     | 0.707 | 0.655 | 0.777 | 2.327 | 0.744 | 0.703 |
| (V)      | 1     | 0.182 | 0.177 | 0.187 | 0.299 | 0.183 | 0.183 |
|          | 2     | 0.331 | 0.318 | 0.348 | 0.757 | 0.334 | 0.332 |
|          | 3     | 0.462 | 0.438 | 0.495 | 1.169 | 0.471 | 0.466 |
|          | 4     | 0.576 | 0.544 | 0.630 | 1.752 | 0.594 | 0.581 |
|          | 5     | 0.715 | 0.667 | 0.790 | 2.073 | 0.755 | 0.712 |

we used a second order unbiased MSE estimator $\hat{M}_i$ in (14) as well as the following two estimators for comparison:

\[(\text{nMSE}) \quad g_{1i}(\hat{A}) + g_{2i}(\hat{A}), \quad (15)\]

\[(\text{pMSE}) \quad g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + \frac{g_{3i}(\hat{A})}{m} + \frac{g_{4i}(\hat{A})}{m} + \frac{2g_{5i}(\hat{A})}{m}. \quad (16)\]

Note that nMSE is a plug-in estimator using the result given in Theorem [1] which ignores the variability of estimating model parameters, but nMSE is asymptotically unbiased. On the other hand, pMSE is the plug-in estimator of the second order approximation formula of MSE, given in Theorem [3] so that it ignores the bias of $g_{1i}(\hat{A}) + g_{2i}(\hat{A})$. Based on $S = 2000$ iterations, we computed the relative bias (RB)
and coefficient of variation (CV), defined as

\[
RB_i = \frac{1}{S} \sum_{s=1}^{S} \frac{\hat{MSE}_i^{(s)} - MSE_i}{MSE_i}, \quad \text{and} \quad CV_i = \sqrt{\frac{1}{S} \sum_{s=1}^{S} \left( \frac{\hat{MSE}_i^{(s)} - MSE_i}{MSE_i} \right)^2},
\]

where \(\hat{MSE}_i^{(s)}\) is the MSE estimate in the \(s\)th iteration and \(MSE_i\) is the true value.

In Table 2, we reported averaged RB and CV values within the same groups. Table 2 reveals that the proposed MSE estimator \(\hat{M}_i\) works quite well in this study. On the other hand, nMSE and pMSE tends to produce considerably smaller MSE estimates than the true MSE values, which are clearly undesirable in practice. Concerning CV values, it is revealed that CV of nMSE is larger than the other two methods while pMSE and bMSE produce similar CV values.

Table 2: Relative bias (RB) and coefficient of variation (CV) of MSE estimators of the naive estimator (15) (nMSE), the plug-in estimator (16) (pMSE) and the second order unbiased estimator (14) (bMSE).

| Scenario | Group | nMSE | pMSE | bMSE | MSE | nMSE | pMSE | bMSE |
|----------|-------|------|------|------|-----|------|------|------|
| I        | 1     | -30.86 | -14.22 | 1.25  | 42.23 | 19.12 | 18.32 |
|          | 2     | -34.60 | -19.99 | -4.53 | 46.85 | 31.32 | 31.02 |
|          | 3     | -37.28 | -23.46 | -8.00 | 49.82 | 36.86 | 36.49 |
|          | 4     | -34.75 | -25.88 | -9.15 | 50.08 | 42.89 | 43.44 |
|          | 5     | -36.81 | -26.30 | -9.45 | 51.74 | 43.51 | 44.39 |
| II       | 1     | -8.30  | -4.43  | 4.30  | 20.60 | 13.01 | 11.28 |
|          | 2     | -10.94 | -6.66  | 4.11  | 25.74 | 20.20 | 18.32 |
|          | 3     | -15.12 | -10.62 | 1.34  | 29.77 | 24.75 | 21.98 |
|          | 4     | -12.49 | -9.47  | 4.37  | 31.68 | 28.96 | 28.13 |
|          | 5     | -16.16 | -12.44 | 2.14  | 34.08 | 30.77 | 29.40 |
| III      | 1     | -2.78  | -1.66  | 3.76  | 12.34 | 9.14  | 9.00  |
|          | 2     | -5.19  | -3.79  | 3.91  | 15.95 | 13.58 | 12.46 |
|          | 3     | -7.06  | -5.43  | 4.10  | 18.50 | 16.20 | 14.88 |
|          | 4     | -6.43  | -5.33  | 6.02  | 20.37 | 19.05 | 18.72 |
|          | 5     | -9.68  | -8.25  | 4.25  | 22.62 | 20.93 | 19.66 |

4.3 Example: milk data

We consider an application to the milk data from the U.S. Bureau of Labor Statistics, which was used in Arora and Lahiri (1997) and You and Chapman (2006). In the data set, the estimated values of the average expenditure on fresh milk for the year 1989, denoted by \(y_i\), are available for 43 areas, with the sampling variances \(D_i\). Following You and Chapman (1997), we use \(x_i'\beta = \beta_j, j = 1, \ldots, 4\), if the \(i\)th area belongs to the \(j\)th major areas. The 4 major areas are

\[
M_1 = \{1, \ldots, 7\}, \quad M_2 = \{8, \ldots, 14\}, \quad M_3 = \{15, \ldots, 25\}, \quad M_4 = \{26, \ldots, 43\}.
\]
Hence, the Fay-Herriot model we consider is given by

$$y_i = \sum_{j=1}^{4} \beta_j I(i \in M_j) + v_i + \varepsilon_i, \quad i = 1, \ldots, 43,$$

where $v_i \sim N(0, A)$ and $\varepsilon_i \sim N(0, D_i)$, and area means are $\theta_i = \sum_{j=1}^{4} \beta_j I(i \in M_j) + v_i$. We first computed estimates of the model parameters from three methods, the ML method, the robust method (RML) of Sinha and Rao (2009), and the proposed robust ML method with density power divergence (DPD) defined as the solution of (11). Here 1% and 5% excess MSEs were used in DPD. The estimates are given in Table 3 and the estimated regression lines are presented in Figure 1. From Table 3 it is clear that the estimates of $\beta_3$ and $\beta_4$ from the four methods are quite similar, but $\beta_1$ and $\beta_2$ are not, which might come from some outliers in the two regions $M_1$ and $M_2$, as observed from Figure 1. Concerning the random effect variance $A$, the estimate from the classical ML method is overestimated, possibly due to the outliers in $M_1$ and $M_2$.

We next consider estimating $\theta_i$ and its risk assessment. For estimating $\theta_i$, we used the proposed robust empirical Bayes estimator with density power divergence (DPEB) with 5% excess MSE and the classical empirical Bayes (EB) estimator in which the ML estimator is used for estimating model parameters. The resulting values in selected 10 areas are presented in Table 4 in which the standardized residuals

$$r_i = \frac{y_i - \sum_{j=1}^{4} \hat{\beta}_j I(i \in M_j)}{\sqrt{\hat{A} + D_i}}$$

with ML estimates $\hat{\beta}_j$ and $\hat{A}$ are also reported. If $r_i$ is large, $y_i$ can be outlier. Regarding MSE estimation of EB and DPEB estimators, we used the second order unbiased estimator given in (14) for DPEB, and the estimator given in Datta and Lahiri (2000) for EB, which are shown in Table 3. From Table 3 it is revealed that the differences between DPEB an EB estimates are relatively large in areas with large standardized residuals, for example area 4 and 11. In such areas, we can observe that EB shrinks the direct estimate $y_i$ more than DPEB. On the other hand, in areas with small standardized residuals, both estimators produces similar estimates like in area 1 and 25. Hence, the proposed estimator prevent the classical EB estimator from over-shrinking the direct estimator $y_i$ in outlying areas. Concerning MSE estimates, we can observe that the MSE estimates for DPEB are uniformly larger than those for EB since the MSE is measured under the classical model (1).

### 5 Conclusion and Discussion

We have developed a robust empirical Bayes method for small area estimation by using a robust likelihood function based on density power divergence. We derived an asymptotic approximation of a mean square error (MSE) of the proposed robust empirical Bayes small area estimator based on the asymptotic properties of the robust estimator of the model parameters. Moreover, we have derived a second-order unbiased estimator of the MSE based on the parametric bootstrap.

This paper has been focused on the basic but extensively used area-level model, Fay-Herriot model. However, when unit level data is available, the nested error
Table 3: Estimates of model parameters in (17) from the four methods: maximum likelihood (ML) method, robust ML (RML) method from Sinha and Rao (2009), and RML based on density power divergence (DPD) with 1% and 5% excess MSE. The estimate of $A$ is multiplied by 100.

|       | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $A$  |
|-------|------------|------------|------------|------------|------|
| ML    | 0.968      | 1.096      | 1.194      | 0.725      | 1.552|
| RML   | 1.005      | 1.181      | 1.192      | 0.725      | 0.796|
| DPD (1%) | 0.974    | 1.115      | 1.194      | 0.727      | 1.498|
| DPD (5%) | 0.984    | 1.151      | 1.194      | 0.728      | 1.348|

Figure 1: Estimated regression lines from the four methods: maximum likelihood (ML) method, robust ML (RML) method from Sinha and Rao (2009), and RML based on density power divergence (DPD) with 1% and 5% excess MSE.

regression model (Battese et al, 1988) is useful, and our robust methods can be similarly applied to such a case. The detailed discussion is left to a valuable future study. On the other hand, in case of non-normal data, the area-level model based on natural exponential family with quadratic variance function (Ghosh and Maiti, 2004) is useful. However, it seems hard to extend this method to such a case since the Bayes estimator of the area mean under the model does not have a simple expression with
Table 4: Estimates of area means \( \theta_i \) from the classical empirical Bayes (EB) estimator and the proposed robust empirical Bayes estimator with density power divergence (DPEB) with associated MSE estimates in 10 selected areas.

| area | region | direct estimate | standardized residual estimate | DPEB estimate | MSE estimate | EB estimate | MSE estimate |
|------|--------|-----------------|--------------------------------|---------------|--------------|-------------|--------------|
| 1    | 1      | 1.099           | 0.640                          | 1.017         | 0.117        | 1.016       | 0.116        |
| 4    | 1      | 0.628           | -2.053                         | 0.763         | 0.094        | 0.775       | 0.092        |
| 5    | 1      | 0.753           | -1.247                         | 0.870         | 0.100        | 0.855       | 0.098        |
| 9    | 2      | 1.405           | 1.479                          | 1.235         | 0.120        | 1.205       | 0.119        |
| 11   | 2      | 0.615           | -3.009                         | 0.729         | 0.090        | 0.803       | 0.088        |
| 12   | 2      | 1.460           | 1.541                          | 1.239         | 0.130        | 1.197       | 0.128        |
| 20   | 3      | 1.292           | 0.475                          | 1.222         | 0.116        | 1.230       | 0.114        |
| 25   | 3      | 1.193           | -0.009                         | 1.194         | 0.092        | 1.194       | 0.090        |
| 31   | 4      | 0.886           | 0.625                          | 0.757         | 0.129        | 0.763       | 0.124        |
| 37   | 4      | 0.440           | -1.842                         | 0.538         | 0.082        | 0.541       | 0.080        |

the marginal likelihood function.

Acknowledgments. I am thankful to Yuki Kawakubo and Tatsuya Kubokawa for their helpful comments. This work was supported by JSPS KAKENHI Grant Numbers JP16H07406.

Appendix

In this Appendix, we use \( s_i = s_i(y_i; \phi) \) if there is no confusion.

A1. Useful Lemma

Lemma 1. When \( y_i \) follows the model (1), namely \( y_i \sim N(x_i^T \beta, A + D_i) \), it holds

\[
E[(y_i - x_i^T \beta)^2 s_i^k] = 0, \quad j, k = 1, 2, \ldots
\]

\[
E[(y_i - x_i^T \beta)^2 s_i^k] = V_i^{k\alpha} (k\alpha + 1)^{-j-1/2}(2j - 1)!!(A + D_i)^j, \quad j, k = 0, 1, 2, \ldots
\]

Proof. We first note that

\[
E[(y_i - x_i^T \beta)^c s_i^k] = \frac{V_i^{k\alpha}}{\sqrt{2\pi(A + D_i)^j}} \int_{-\infty}^{\infty} (t - x_i^T \beta)^c \exp \left\{ -\frac{(k\alpha + 1)(t - x_i^T \beta)^2}{2(A + D_i)} \right\} dt
\]

\[
= \frac{V_i^{k\alpha}}{\sqrt{k\alpha + 1}} E[Z^c],
\]

where \( Z \sim N(0, (A + D_i)/(k\alpha + 1)) \). Hence, the expectation is 0 when \( c \) is odd. On the other hand, when \( c = 2j, \quad j = 0, 1, 2, \ldots \), it follows that \( E[Z^{2j}] = (2j - 1)!!(A + D_i)^j(k\alpha + 1)^{-j} \), which completes the proof. \( \square \)
A2. Proof of Theorem 1

Noting that \( \tilde{\theta}_i = E[\theta_i | y_i] \), it follows that

\[
E[(\tilde{\theta}_R - \theta_i)^2] = E[(\tilde{\theta}_i - \theta_i)^2] + E[(\tilde{\theta}_R - \tilde{\theta}_i)^2]
\]

\[
= \frac{AD_i}{A + D_i} + E[(\tilde{\theta}_R - \tilde{\theta}_i)^2] \equiv g_{1i}(A) + g_{2i}(A)
\]

Since

\[
\tilde{\theta}_R - \tilde{\theta}_i = y_i - \frac{D_i}{A + D_i} (y_i - \mu_i) + \frac{D_i}{A + D_i} (y_i - \mu_i)
\]

we have

\[
g_{2i}(A) = \frac{D_i^2}{(A + D_i)^2} E \left[ (y_i - \mu_i)^2 (1 - s_i)^2 \right]. \tag{18}
\]

Using Lemma 1, we obtain the analytical expression of \( g_{2i}(A) \).

We next show that \( g_{2i}(A) \) is increasing in \( \alpha \in (0, 1) \). For notational simplicity, we put \( \mu_i = \mu_i(\gamma, \phi) \). Since \( (y_i - \mu_i)^2 (1 - s_i)^2 \) is a continuous and differentiable function of \( y_i \) and \( \alpha \), we have

\[
\frac{\partial g_{2i}(A)}{\partial \alpha} = -\frac{2D_i^2}{(A + D_i)^2} E \left[ (y_i - \mu_i)^2 (1 - s_i) \frac{\partial s_i}{\partial \alpha} \right].
\]

Note that \( s_i = f_i(y_i; \phi) \). If \( f(y_i; \phi) \leq 1 \), it holds \( 1 - s_i \geq 0 \) and \( s_i \) is decreasing with respect to \( \alpha \). Then, it follows that \( (1 - s_i) \partial s_i / \partial \alpha \leq 0 \). On the other hand, we have \( (1 - s_i) \partial s_i / \partial \alpha \leq 0 \) if \( f(y_i; \phi) \geq 1 \) by the similar argument. Hence, \( (1 - s_i) \partial s_i / \partial \alpha \leq 0 \) always follows, thereby we have \( \partial g_{2i}(A)/\partial \alpha \geq 0 \) for \( \alpha \in (0, 1) \), which completes the proof.

A3. Proof of Theorem 2

Under conditions (C1) and (C2), the theory of unbiased estimating equation (Godambe, 1960) shows that \( \hat{\phi} = (\hat{\beta}^T, \hat{A}) \) is consistent and asymptotically normal, with the asymptotic covariance matrix is given by

\[
\lim_{m \to \infty} E \left[ \frac{1}{m} \frac{\partial^2 L_\alpha}{\partial \phi \partial \phi^T} \right]^{-1} E \left[ \frac{1}{m} \frac{\partial L_\alpha}{\partial \phi} \frac{\partial L_\alpha}{\partial \phi^T} \right] E \left[ \frac{1}{m} \frac{\partial^2 L_\alpha}{\partial \phi \partial \phi^T} \right]^{-1}.
\]
From (11), straightforward calculation shows that

\[
\frac{1}{m} \frac{\partial^2 \hat{L}_A}{\partial \beta \partial A} = \frac{1}{m} \sum_{i=1}^{m} \frac{x_i \beta_i}{(A + D_i)^2} \left\{ \alpha (y_i - x_i \beta)^2 - (A + D_i) \right\}
\]

\[
\frac{1}{m} \frac{\partial^2 \hat{L}_A}{\partial A^2} = \frac{1}{m} \sum_{i=1}^{m} \frac{x_i s_i (y_i - x_i \beta)}{(A + D_i)^2} \left\{ \frac{\alpha}{2} (y_i - x_i \beta)^2 - \frac{\alpha}{2} (A + D_i) - 1 \right\}
\]

\[
\frac{1}{m} \frac{\partial^2 \hat{L}_A}{\partial \beta \partial A} = \frac{1}{m} \sum_{i=1}^{m} \left[ \frac{\alpha s_i (y_i - x_i \beta)^2}{(A + D_i)^2} \right] (y_i - x_i \beta)^2 - (A + D_i) \right\} - \frac{3 \alpha V^a_{\beta}}{(\alpha + 1)^{3/2} (A + D_i)^2}
\]

\[- \frac{s_i}{2 (A + D_i)^2} \left\{ (4 + \alpha) (y_i - x_i \beta)^2 - \alpha (A + D_i) \right\} + \frac{s_i}{(A + D_i)^2} \right]\]

Then, using Lemma 1, we have \(E[m^{-1} \partial^2 \hat{L}_A / \partial \beta \partial \beta] = J_\beta, E[m^{-1} \partial^2 \hat{L}_A / \partial A^2] = 0\) and \(E[m^{-1} \partial^2 \hat{L}_A / \partial A \partial \beta] = -J_A\). On the other hand, from (11) and Lemma 1, it holds that \(E[m^{-1} (\partial \hat{L}_A / \partial \beta) (\partial \hat{L}_A / \partial \beta')] = K_\beta, E[m^{-1} (\partial \hat{L}_A / \partial A) (\partial \hat{L}_A / \partial A')] = 0\) and \(E[m^{-1} (\partial \hat{L}_A / \partial A')^2] = K_A\). Hence, \(\hat{\beta}\) and \(\hat{A}\) is asymptotically independent and their asymptotic covariance matrices are \(J_\beta^{-1} K_\beta J_\beta^{-1}\) and \(J_A^{-1} K_A J_A^{-1}\), respectively, and the result has been established.

A4. Proof of Theorem 3

The MSE \(E[(\hat{\theta}_R^i - \theta_i)^2]\) can be decomposed as

\[E[(\hat{\theta}_R^i - \theta_i)^2] = E[(\hat{\theta}_R^i - \theta_i)^2] + 2E[(\hat{\theta}_R^i - \theta_i)(\hat{\theta}_R^i - \hat{\theta}_R^i)] + E[(\hat{\theta}_R^i - \hat{\theta}_R^i)^2],\]

where the first term reduces to \(g_{1i}(A) + g_{2i}(A)\) as shown in Theorem 1.

We first evaluate the third term. Taylor series expansion shows that

\[
\hat{\theta}_R^i - \hat{\theta}_R^i = \frac{\partial \hat{\theta}_R^i}{\partial \phi} (\phi - \hat{\phi}) + \frac{1}{2} (\hat{\phi} - \phi)^t \frac{\partial^2 \hat{\theta}_R^i}{\partial \phi \partial \phi} (\hat{\phi} - \phi),
\]

where \(\phi_\ast\) is on the line connecting \(\phi\) and \(\hat{\phi}\). Then, we get

\[E[(\hat{\theta}_R^i - \hat{\theta}_R^i)^2] = E \left[ \left( \frac{\partial \hat{\theta}_R^i}{\partial \phi} (\hat{\phi} - \phi) \right)^2 \right] + R_1 + R_2,
\]

where \(R_1 = E[(\hat{\phi} - \phi)^t (\hat{\theta}_R^i / \partial \phi)(\hat{\phi} - \phi)^t (\hat{\theta}_R^i / \partial \phi)(\hat{\phi} - \phi)^t] \) and \(R_2 = E[\{(\hat{\phi} - \phi)^t (\hat{\theta}_R^i / \partial \phi)(\hat{\phi} - \phi)^t\}^2] / 4\). Here we use the following lemma.

Lemma 2. Under (C1) and (C2), \(E[|\hat{\phi}_k - \phi_k|^r] = O(m^{-r/2})\) for any \(r > 0\) and \(k = 1, \ldots, p + 1\).

A rigorous proof of the lemma requires a uniform integrability, but intuitively, from Theorem 2 it holds that \(E[m^r |\hat{\phi}_k - \phi_k|^r] = O(1)\) under (C1) and (C2), which leads to Lemma 2.

In what follows, we use \(u_i = y_i - x_i^t \beta\) and \(B_i = A + D_i\) for notational simplicity. The straightforward calculation shows that

\[
\frac{\partial \hat{\theta}_R^i}{\partial \beta} = -\frac{D_i s_i x_i}{B_i^2} (\alpha u_i^2 - B_i), \quad \frac{\partial \hat{\theta}_R^i}{\partial A} = -\frac{D_i s_i u_i}{2 B_i^3} \left\{ \alpha u_i^2 - (2 - \alpha) B_i \right\}.
\]
Moreover, we have
\[
\frac{\partial^2 \hat{\theta}_i^R}{\partial \beta \partial \beta} = \frac{D_i s_i x_i x_i^T}{B_i^3} (\alpha u_i^3 - 3B_i u_i)
\]
\[
\frac{\partial^2 \hat{\theta}_i^R}{\partial A^2} = \frac{D_i s_i u_i}{12B_i^3} \left( -3\alpha^2 u_i^3 + 3\alpha^2 B_i u_i^2 + (4\alpha - 3\alpha^2)B_i u_i + (\alpha - 2)(3\alpha + 8)B_i^2 \right).
\]
Noting that
\[
R = \sum_{j=1}^{p+1} \sum_{k=1}^{p+1} \sum_{\ell=1}^{p+1} E \left[ \left( \frac{\partial \hat{\theta}_i^R}{\partial \phi_j} \right) \left( \frac{\partial^2 \hat{\theta}_i^R}{\partial \phi_k \partial \phi_\ell} \right) (\hat{\phi}_j - \phi_j)(\hat{\phi}_k - \phi_k)(\hat{\phi}_\ell - \phi_\ell) \right]
\]
we have
\[
|U_{1jk\ell}| \leq E \left[ \left( \frac{\partial \hat{\theta}_i^R}{\partial \phi_j} \right)^4 \left( \frac{\partial^2 \hat{\theta}_i^R}{\partial \phi_k \partial \phi_\ell} \right)^4 \left| \hat{\phi}_j - \phi_j \right|^4 \left| \hat{\phi}_k - \phi_k \right|^4 \left| \hat{\phi}_\ell - \phi_\ell \right|^4 \right]^{1/4}
\]
\[
\leq E \left[ \left( \frac{\partial \hat{\theta}_i^R}{\partial \phi_j} \right)^8 \frac{1}{8} \left( \frac{\partial^2 \hat{\theta}_i^R}{\partial \phi_k \partial \phi_\ell} \right)^8 \frac{1}{8} \left| \hat{\phi}_j - \phi_j \right|^4 \left| \hat{\phi}_k - \phi_k \right|^4 \left| \hat{\phi}_\ell - \phi_\ell \right|^4 \right]^{1/4}
\]
from Hölder’s inequality. Since \(E \left[ \left| \frac{\partial \hat{\theta}_i^R}{\partial \phi_j} \right|^8 \right] < \infty\) and \(E \left[ \left| \frac{\partial^2 \hat{\theta}_i^R}{\partial \phi_k \partial \phi_\ell} \right|^8 \right] < \infty\), it holds that \(R_1 = o(m^{-1})\) from Lemma 2. A quite similar evaluation shows that \(R_2 = o(m^{-1})\). On the other hand, using the similar argument given in the proof of Theorem 3 in Kubokawa et al. (2016), we have
\[
E \left\{ \left( \frac{\partial \hat{\theta}_i^R}{\partial \phi} \right)^2 (\hat{\phi} - \phi) \right\} = \text{tr} \left\{ E \left[ \frac{\partial \hat{\theta}_i^R}{\partial \phi} \frac{\partial \hat{\theta}_i^R}{\partial \phi} \right] E [ (\hat{\phi} - \phi)(\hat{\phi} - \phi) ] \right\} + o(m^{-1})
\]
\[
= \frac{1}{m} \text{tr} \left\{ E \left[ \frac{\partial \hat{\theta}_i^R}{\partial \beta} \frac{\partial \hat{\theta}_i^R}{\partial \beta} \right] J^{-1}_\beta K_\beta J^{-1}_\beta \right\} + \frac{1}{m} E \left[ \left( \frac{\partial \hat{\theta}_i^R}{\partial A} \right)^2 \right] J^{-1}_A K_A J^{-1}_A + o(m^{-1}),
\]
thereby, using Theorem 2 and
\[
E \left[ \frac{\partial \hat{\theta}_i^R}{\partial \beta} \frac{\partial \hat{\theta}_i^R}{\partial \beta} \right] = \frac{D_i^2 V_i^{2\alpha} x_i x_i^T}{(A + D_i^2)(2\alpha + 1)^{3/2}}
\]
\[
E \left[ \left( \frac{\partial \hat{\theta}_i^R}{\partial A} \right)^2 \right] = \frac{D_i^2 V_i^{2\alpha}}{(A + D_i)^2(2\alpha + 1)^{7/2}} \left( \alpha^4 - \frac{1}{2} \alpha^2 + 1 \right),
\]
we obtain \(E[(\hat{\theta}_i^R - \theta_i^R)^2] = m^{-1} g_{3i}(A) + m^{-1} g_{4i}(A) + o(m^{-1})\).

Concerning \(E[(\hat{\theta}_i^R - \theta_i)(\hat{\theta}_i^R - \theta_i^R)]\), we define \(\tilde{\theta}_i \equiv E[\theta_i | y_i] = y_i - D_i u_i / B_i\) as the classical Bayes estimator of \(\theta_i\). Then we have
\[
E[(\hat{\theta}_i^R - \tilde{\theta}_i)(\hat{\theta}_i^R - \tilde{\theta}_i^R)] = E[(\hat{\theta}_i^R - \theta_i)(\hat{\theta}_i^R - \theta_i^R)] = \frac{D_i}{B_i} E[(1 - s_i)u_i(\hat{\theta}_i^R - \tilde{\theta}_i^R)].
\]
Taylor series expansion shows that
\[ \hat{g}_i^R - \tilde{g}_i^R = \frac{\partial \hat{g}_i^R}{\partial \phi} (\hat{\phi} - \phi) + \frac{1}{2} (\hat{\phi} - \phi)^t \frac{\partial^2 \hat{g}_i^R}{\partial \phi \partial \phi^t} (\hat{\phi} - \phi) + R_3, \]
with
\[ R_3 = \frac{1}{6} \sum_{k=1}^{p+1} \sum_{j=1}^{p+1} \sum_{\ell=1}^{p+1} \frac{\partial^3 \hat{g}_i^R}{\partial \phi_k \partial \phi_j \partial \phi_\ell} (\hat{\phi}_k - \phi_k)(\hat{\phi}_j - \phi_j)(\hat{\phi}_\ell - \phi_\ell). \]

The similar evaluation to \( R_2 \) and \( R_3 \) shows that \( E[(1 - s_i)u_i R_3] = o(m^{-1}) \). Then, we have
\[
E[(\hat{g}_i^R - \theta_i)(\hat{g}_i^R - \tilde{g}_i^R)] = \frac{D_i}{B_i} E \left[ (1 - s_i)u_i \frac{\partial \hat{g}_i^R}{\partial \beta} (\hat{\phi} - \phi) \right] + \frac{D_i}{2B_i} E \left[ (1 - s_i)u_i (\hat{\phi} - \phi)^t \frac{\partial^2 \hat{g}_i^R}{\partial \phi \partial \phi^t} (\hat{\phi} - \phi) \right] + o(m^{-1})
\]
\[ \equiv T_1 + T_2 + o(m^{-1}). \]

It follows that
\[
T_2 = \frac{D_i}{2mB_i} \text{tr} \left\{ E \left[ (1 - s_i)u_i \frac{\partial^2 \hat{g}_i^R}{\partial \beta \partial \beta^t} J_\beta^{-1} K_{\beta} J_\beta^{-1} \right] \right\} + \frac{D_i K_A}{2mB_i J_A^2} E \left[ (1 - s_i)u_i \frac{\partial^2 \hat{g}_i^R}{\partial A \partial A^t} \right] + o(m^{-1}). \]  \( \quad \) (19)

From Lohr and Rao (2009), it follows that
\[
E[\beta - \beta|y_i] = b_\beta - m^{-1} B^{-1} J_\beta^{-1} x_i s_i u_i + o_p(m^{-1})
\]
\[
E[A - A|y_i] = b_A - m^{-1} J_A^{-1} \left\{ u_i^2 s_i B_i^2 - s_i + \frac{\alpha V_i^n}{(\alpha + 1)^{3/2} B_i} \right\} + o_p(m^{-1}),
\]
where \( b_\beta = \lim_{m \to \infty} m E[\beta - \beta] \) and \( b_A = \lim_{m \to \infty} m E[A - A] \), so that
\[
T_1 = -\frac{D_i}{mB_i} E \left[ s_i (1 - s_i)u_i \frac{\partial \hat{g}_i^R}{\partial \beta} J_\beta^{-1} x_i \right] - \frac{D_i b_A}{B_i} E \left[ s_i (1 - s_i)u_i \frac{\partial \hat{g}_i^R}{\partial A} (u_i^2 - B_i) \right] + \frac{D_i}{B_i} E \left[ (1 - s_i)u_i \frac{\partial \hat{g}_i^R}{\partial A} \right] \left\{ b_A - \frac{\alpha V_i^n}{m(\alpha + 1)^{3/2} B_i} \right\} + o(m^{-1}). \]  \( \quad \) (20)

Combining (20) and (19), and using Lemma 1, we obtain \( E[(\hat{g}_i^R - \theta_i)(\hat{g}_i^R - \tilde{g}_i^R)] = m^{-1} g_{5i}(A) + o(m^{-1}) \), which completes the proof.

A5. Proof of Theorem 4

For simplicity, let \( g_{12i}(A) = g_{1i}(A) + g_{2i}(A) \). Then, we have
\[
E[g_{12i}(\hat{A}) - g_{12i}(A)] = b_A \frac{\partial g_{12i}(A)}{\partial A} + \frac{1}{2m} \frac{\partial^2 g_{12i}(A)}{\partial A^2} K_{\beta} J_A^{-1} + \frac{1}{6} \frac{\partial^3 g_{12i}(A)}{\partial A^3} E[(\hat{A} - A)^3],
\]
where \( A^* \) is between \( A \) and \( \hat{A} \). From Lemma 2, it follows that \( E[g_{12i}(\hat{A}) - g_{12i}(A)] = m^{-1} c(A) + o(m^{-1}) \), where \( c(\cdot) \) is a smooth function. Hence, from Butar and Lahiri (2003), the second order unbiasedness of the bootstrap MSE estimator (14) follows.
References

[1] Arora, V. and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

[2] Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, 85, 549-559.

[3] 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.

[4] Burden, R. L. and Faires, J. D. (2010). *Numerical Analysis, 9th Edition*, Brooks-Cole Publishing.

[5] Butar, F. B. and Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference*, 112, 63-76.

[6] Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97, 465-480.

[7] Chambers, R. L., Chandra, H., Salvati, N. and Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B*, 76, 47-69.

[8] Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255-268.

[9] 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.

[10] 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.

[11] Datta, G. S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.

[12] Datta, G. S. and Mandal, A. (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, 110, 1735-1744.

[13] 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.

[14] Dongmo-Jiongo, V., Haziza, D. and Duchesne, P. (2013). Controlling the bias of robust small area estimators. *Biometrika*, 100, 843-858.

[15] 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.

[16] Ghosh, M. and Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, 91, 95-112.

[17] Ghosh, M., Maiti, T. and Roy, A. (2008). Influence functions and robust Bayes and empirical Bayes small area estimation. *Biometrika*, 95, 573-585.
[18] Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, **31**, 1208-1211.

[19] Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of Royal Statistical Society: Series B*, **68**, 221-238.

[20] Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, **1**, 799-821.

[21] Kubokawa, T., Sugasawa, S., Ghosh, M. and Chaudhuri, S. (2016). Prediction in heteroscedastic nested error regression models with random dispersions. *Statistica Sinica*, **26**, 465-492.

[22] 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.

[23] Lohr, S. L. and Rao, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, **96**, 457-468.

[24] Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics & Data Analysis*, **58**, 308-325.

[25] Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of Royal Statistical Society: Series B*, **70**, 265-286.

[26] Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, **28**, 40-68.

[27] Prasad, N. and Rao, J. N. K. (1990). The estimation of mean-squared errors of small-area estimators. *Journal of the American Statistical Association*, **90**, 758-766.

[28] Rao, J.N.K. and Molina, I. (2015) *Small Area Estimation, 2nd Edition*. Wiley.

[29] Sinha, S. K. and Rao, J. N. K. (2009). Robust small area estimation. *Canadian Journal of Statistics*, **37**, 381-399.

[30] Sugasawa, S. and Kubokawa, T. (2016). Bayesian estimators of small area models shrinking both mean and variances. *Scandinavian Journal of Statistics*, to appear.

[31] Ybarra, L. M. R. and Lohr, S. L. (2008) Small area estimation when auxiliary information is measured with error. *Biometrika*, **95**, 919-931.

[32] You, Y. and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, **32**, 97-103.