Robust Hierarchical Bayes Small Area Estimation for Nested Error Regression Model

Adrijo Chakraborty¹, Gauri Sankar Datta²,³ and Abhyuday Mandal²

¹NORC at the University of Chicago, Bethesda, MD 20814, USA
²Department of Statistics, University of Georgia, Athens, GA 30602, USA
³Center for Statistical Research and Methodology, US Census Bureau
E-mails: adrijo.chakraborty@gmail.com, gauri@stat.uga.edu and amandal@stat.uga.edu

Summary

Standard model-based small area estimates perform poorly in presence of outliers. Sinha and Rao (2009) developed robust frequentist predictors of small area means. In this article, we present a robust Bayesian method to handle outliers in unit-level data by extending the nested error regression model. We consider a finite mixture of normal distributions for the unit-
level error to model outliers and produce noninformative Bayes predictors of small area means. Our solution generalizes the solution by Datta and Ghosh (1991) under the normality assumption. Application of our method to a data set, which is suspected to contain an outlier, confirms this suspicion and correctly identifies the suspected outlier, and produces robust predictors and posterior standard deviations of the small area means. Evaluation of several procedures via simulations shows that our proposed procedure is as good as the other procedures in terms of bias, variability, and coverage probability of confidence or credible intervals, when there are no outliers. In presence of outliers, while our method and Sinha-Rao method perform similarly, they improve over the other methods. This superior performance of our procedure shows its dual (Bayes and frequentist) dominance, and is attractive to all practitioners, Bayesians and frequentists, of small area estimation
Key words: Normal mixture; outliers; prediction intervals and uncertainty; robust empirical best linear unbiased prediction; unit-level models.

1 Introduction

The nested error regression (NER) model with normality assumption for both the random effects or model error terms and the unit-level error terms has played a key role in analyzing unit-level data in small area estimation. Many popular small area estimation methods have been developed under this model. In the frequentist approach, Battese et al. (1988), Prasad and Rao (1990), Datta and Lahiri (2000), for example, derived empirical best linear unbiased predictors (EBLUPs) of small area means. These authors used various estimation methods for the variance components and derived approximately accurate estimators of mean squared error (MSEs) of the EBLUPs. On the other hand, Datta and Ghosh (1991) followed the hierarchical Bayesian (HB) approach to derive posterior means as HB predictors and variances of the small area means. While the underlying normality assumptions for all the random quantities are appropriate for regular data, they fail to adequately accommodate outliers. Consequently, these frequentist/Bayesian methods, are highly influenced by major outliers in the data, or break down if the outliers grossly violate distributional assumptions.

Sinha and Rao (2009) investigated robustness, or lack thereof, of the EBLUPs from the usual normal NER model in presence of “representative outliers”. According to Chambers (1986), a representative outlier is a “sample element with a value that has been correctly recorded and cannot be regarded as unique. In particular, there is no reason to assume that there are no more similar outliers in the nonsampled part of the population.” Sinha and Rao (2009) showed via simulations for the NER model that while the EBLUPs are efficient under normality, they are very sensitive to outliers that deviate from the assumed model.
To address the non-robustness issue of EBLUPs, Sinha and Rao (2009) used $\psi$-function, Huber’s Proposal 2 influence function in M-estimation, to downweight contribution of outliers in the BLUPs and the estimators of the model parameters, both regression coefficients and variance components. Using M-estimation for robust maximum likelihood estimators of model parameters and robust predictors of random effects, Sinha and Rao (2009) for mixed linear models proposed a robust EBLUP (REBLUP) of mixed effects, which they used to estimate small area means for the NER model. By using a parametric bootstrap procedure they have also developed estimators of the MSEs of the REBLUPs. We refer to Sinha and Rao (2009) for details of this method. Their simulations show that when the normality assumptions hold, the proposed REBLUPs perform similar to the EBLUPs in terms of empirical bias and empirical MSE. But, in presence of outliers in the unit-level errors, while both EBLUPs and REBLUPs remain approximately unbiased, the empirical MSEs of the EBLUPs are significantly larger than those of the REBLUPs.

Datta and Ghosh (1991) proposed a noninformative HB model to predict finite population small area means. In this article we follow the approach to finite population sampling which was also followed by Datta and Ghosh (1991). Our suggested model includes the treatment of NER model by Datta and Ghosh (1991) as a special case. Our model facilitates accommodating outliers in the population and in the sample values. We replace the normality of the unit-level error terms by a two-component mixture of normal distributions, each component centered at zero. As in Datta and Ghosh (1991), we assume normality of the small area effects.

Simulation results of Sinha and Rao (2009) indicated that there was not enough improvement in performance of the REBLUP procedures over the EBLUPs when they considered outliers in both the unit-level error and the model error terms. To keep both analytical and computational challenges for our noninformative HB analysis manageable, we use a realistic framework and we restrict ourselves to the normality assumption for the random effects. Moreover, the assumption of zero means for the unit-level error terms is similar to the assumption made by Sinha and Rao (2009). While allowing the component of the
unit-level error terms with the bigger variance to also have non-zero means to accommodate outliers might appear attractive, we note it later that it is not possible to conduct a noninformative Bayesian analysis with an improper prior on the new parameter.

We focus only unit-level model robust small area estimation in this article. There is a substantial literature on small area estimation based on area-level data using Fay-Herriot model (see Fay and Herriot, 1979; Prasad and Rao, 1990). The paper by Sinha and Rao (2009) also discussed robust small area estimation for area-level model. In another paper, Lahiri and Rao (1995) discussed EBLUP and estimation of MSE under non-normality assumption for the random effects. An early robust Bayesian approach for area-level model is due to Datta and Lahiri (1995), where they used scale mixture of normal distributions for the random effects. It is worth mentioning that the $t$-distributions are special cases of scale mixture of normal distributions. While Datta and Lahiri (1995) assumed long-tailed distributions for the random effects, Bell and Huang (2006) used HB method based on $t$ distribution, either only for the unit-level errors or only for the model errors. Bell and Huang (2006) assumed that outliers can arise either in model errors or in unit-level errors.

The scale mixture of normal distributions requires specification of the mixing distribution, or in the specific case for $t$ distributions, it requires the degrees of freedom. In an attempt to avoid this specification, in a recent article Chakraborty et al. (2016) proposed a simple alternative via a two-component mixture of normal distributions in terms of the variance components for the model errors.

2 Unit-Level HB Models for Small Area Estimation

The model-based approach to finite population sampling is very useful to model unit-level data in small area estimation. The NER model of Battese et al. (1988) is a popular model for unit-level data. Suppose a finite population is partitioned into $m$ small areas, with $i$th area having $N_i$ units. The NER model relates $Y_{ij}$, the value of a response variable $Y$ for the $j$th unit in the $i$th small area, with $x_{ij} = (x_{ij1}, \cdots , x_{ijp})^T$, the value of
a $p$-component covariate vector associated with that unit, through a mixed linear model given by
\[ Y_{ij} = x_{ij}^T \beta + v_i + e_{ij}, \quad j = 1, \cdots, N_i, \quad i = 1, \cdots, m, \] (2.1)
where all the random variables $v_i$'s and $e_{ij}$'s are assumed independent. Distributions of these variables are specified by assuming that random effects $v_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_v^2)$ and unit-level errors $e_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$. Here $\beta = (\beta_1, \cdots, \beta_p)^T$ is the regression coefficient vector. We want to predict the $i$th small area finite population mean $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}, \ i = 1, \cdots, m$.

Battese et al. (1988), Prasad and Rao (1990), among others, considered noninformative sampling, where a simple random sample of size $n_i$ is selected from the $i$th small area. For notational simplicity we denote the sample by $Y_{ij}, j = 1, \cdots, n_i, i = 1, \cdots, m$. To develop predictors of small area means $\bar{Y}_i, i = 1, \cdots, m$, these authors first derived, for known model parameters, the conditional distribution of the unsampled values, $Y_{ij}, j = n_i + 1, \cdots, N_i, i = 1, \cdots, m$, given the sampled values $Y_{ij}, j = 1, \cdots, n_i, i = 1, \cdots, m$. Under squared error loss, the best predictor of $\bar{Y}_i$ is its mean with respect to this conditional distribution, also known as predictive distribution. In the frequentist approach, Battese et al. (1988), Prasad and Rao (1990) obtained the EBLUP of $\bar{Y}_i$ by replacing in the conditional mean the unknown model parameters $(\beta^T, \sigma_e^2, \sigma_v^2)^T$ by their estimators using $Y_{ij}, j = 1, \cdots, n_i, i = 1, \cdots, m$. In the Bayesian approach, on the other hand, Datta and Ghosh (1991) developed HB predictors of $\bar{Y}_i$ by integrating out these parameters in the conditional mean of $\bar{Y}_i$ with respect to their posterior density, which is derived based on a prior distribution on the parameters and the distribution of the sample $Y_{ij}, j = 1, \cdots, n_i, i = 1, \cdots, m$, derived under the model (2.1).

While the frequentist approach for the NER model under distributional assumptions in (2.1) continues with accurate approximation and estimation of the MSEs of the EBLUPs, the Bayesian approach typically proceeds under some noninformative priors, and computes numerically, usually by MCMC method, the exact posterior means and variances of $\bar{Y}_i$’s. Among various noninformative priors for $\beta, \sigma_e^2, \sigma_v^2$, a popular choice is
\[ \pi_P(\beta, \sigma_e^2, \sigma_v^2) = \frac{1}{\sigma_e^2}, \] (2.2)
The standard NER model in (2.1) is unable to explain outlier behavior of unit-level error terms. To avoid breakdown of EBLUPs and their MSEs in presence of outliers, Sinha and Rao (2009) modified all estimating equations for the model parameters and random effects terms by robustifying various “standardized residuals” that appear in estimating equations by using Huber’s \( \psi \)-function, which truncates large absolute values to a certain threshold. They did not replace the working NER model in (2.1) to accommodate outliers, but they accounted for their potential impacts on the EBLUPs and estimated MSEs by downweighting large standardized residuals that appear in various estimating equations through Huber’s \( \psi \)-function. Their approach, in the terminology of Chambers et al. (2014), may be termed \textit{robust projective}, where they estimated the working model in a robust fashion and used that to project sample non-outlier behavior to the unsampled part of the model.

To investigate the effectiveness of their proposal, Sinha and Rao (2009) conducted simulations based on various long-tailed distributions for the random effects and/or the unit-level error terms. In one of their simulation scenarios which is reasonably simple but useful, they used a two-component mixture of normal distributions for the unit-level error terms, both components centered at zero but with unequal variances, and the component with the larger variance appears with a small probability. This reflects the regular setup of the NER model with the possibility of outliers arising as a small fraction of contamination caused by the error corresponding to the larger variance component. In this article, we incorporate this mixture distribution to modify the model in (2.1) to develop new Bayesian methods that would be robust to outliers. Our proposed population level HB model is given by

**NM HB Model:**

1. Conditional on \( \beta = (\beta_1, \ldots, \beta_p)^T, v_1, \ldots, v_m, z_{ij}, j = 1, \ldots, N_i, i = 1, \ldots, m, p_e, \sigma^2_1, \sigma^2_2 \) and \( \sigma^2_v \),

\[
Y_{ij} \overset{ind}{\sim} z_{ij}N(x_{ij}^T\beta + v_i, \sigma^2_1) + (1 - z_{ij})N(x_{ij}^T\beta + v_i, \sigma^2_2), \ j = 1, \ldots, N_i, i = 1, \ldots, m.
\]
(II) The indicator variables \( z_{ij} \)'s are iid with \( P(z_{ij} = 1|p_e) = p_e, \ j = 1, \cdots, N_i, i = 1, \cdots, m, \) and are independent of \( \beta = (\beta_1, \cdots, \beta_p)^T, v_1, \cdots, v_m, \sigma^2_1, \sigma^2_2 \) and \( \sigma^2_v \).

(III) Conditional on \( \beta, z = (z_{11}, \cdots, z_{1N_1}, \cdots, z_{m1}, \cdots, z_{mN_m})^T, p_e, \sigma^2_1, \sigma^2_2 \) and \( \sigma^2_v \), random small area effects \( v_i \sim iid \sim N(0, \sigma^2_v) \) for \( i = 1, \cdots, m \).

For simplicity, we assume the contamination probability \( p_e \) to remain the same for all units in all small areas. Gershunskaya (2010) proposed this mixture model for empirical Bayes point estimation of small area means. We assume independent simple random samples of size \( n_1, \cdots, n_m \) from the \( m \) small areas. For simplicity of notation, we denote the responses for the sampled units from the \( i \)th small area by \( Y_{i1}, \cdots, Y_{in_i}, i = 1, \cdots, m \).

The SRS results in a noninformative sample and that the joint distribution of responses of the sampled units can be obtained from the NM HB model above by replacing \( N_i \) by \( n_i \). This marginal distribution in combination with the prior distribution provided below will yield the posterior distribution of \( v_i \)'s, and all the parameters in the model. For the informative sampling development in small area estimation we refer to Pfeffermann and Sverchkov (2007) and Verret et al. (2015).

Two components of the normal mixture distribution differ only by their variances. We will assume the variance component \( \sigma^2_2 \) is larger than \( \sigma^2_1 \) and is intended to explain any outliers in a data set. However, if a data set does not include any outliers, the two component variances \( \sigma^2_1, \sigma^2_2 \) may only minimally differ. In such situation, the likelihood based on the sample will include limited information to distinguish between these variance parameters, and consequently, the likelihood will also have little information about the mixing proportion \( p_e \). We notice this behavior in our application to a subset of the corn data in Section 5.

In this article, we carry out an objective Bayesian analysis by assigning a noninformative prior to the model parameters. In particular, we propose a noninformative prior

\[
\pi(\beta, \sigma^2_1, \sigma^2_2, \sigma^2_v, p_e) = \frac{I(0 < \sigma^2_1 < \sigma^2_2 < \infty)}{(\sigma^2_2)^2},
\]

(2.3)
where we have assigned an improper prior on $\beta, \sigma^2_v, \sigma^2_1, \sigma^2_2$ and a proper uniform prior on the mixing proportion $p_e$. However, subjective priors could also be assigned when such subjective information is available. Notably, it is possible to use some other proper prior on $p_e$ that may elicit the extent of contamination to the basic model to reflect prevalence of outliers. While many such subjective prior can be reasonably modeled by a beta distribution, we use a uniform distribution from this class to reflect noninformativeness or little information about this parameter. We also use traditional uniform prior on $\beta$ and $\sigma^2_v$. In the Supplementary materials, we explore the propriety of the posterior distribution corresponding to the improper priors in (2.3).

The improper prior distribution on the two variances for the mixture distribution has been carefully chosen so that the prior will yield conditionally proper distribution for each parameter given the other. This conditional propriety is necessary for parameters appearing in the mixture distribution in order to ensure under suitable conditions the propriety of the posterior density resulting from the HB model. The specific prior distribution that we propose above is such that the resulting marginal densities for $\sigma^2_1$ and $\sigma^2_2$ respectively, are $\pi_{\sigma^2_1}(\sigma^2_1) = (\sigma^2_1)^{-1}$ and $\pi_{\sigma^2_2}(\sigma^2_2) = (\sigma^2_2)^{-1}$. These two densities are of the same form as that of $\sigma^2_e$ in the regular model in (2.2) introduced earlier. Indeed by setting $p_e = 0$ or 1 in our analysis, we can reproduce the HB analysis of the regular model given by (2.1) and (2.2).

We use the NM HB Model under noninformative sampling and the noninformative priors given by (2.3) to derive the posterior predictive distribution of $\bar{Y}_i, i = 1, \cdots, m$. The NM HB model and noninformative sampling that we propose here facilitate building model for representative outliers (Chambers, 1986). According to Chambers, a representative outlier is a value of a sampled unit which is not regarded as unique in the population, and one can expect existence of similar values in the non-sampled part of the population which will influence the value of the finite population means $\bar{Y}_i$’s or the other parameters involved in the superpopulation model.

Following the practice of Battese et al. (1988) and Prasad and Rao (1990), we approx-
imated the predictand $\bar{Y}_i$ by $\theta_i = X_i^T \beta + v_i$ to draw inference on the finite population small area means. Here $X_i = N_i^{-1} \sum_{j=1}^{N_i} x_{ij}$ is assumed known. This approximation works well for small sampling fractions $n_i/N_i$ and large $N_i$'s. It has been noted by these authors, and by Sinha and Rao (2009), that even for the case of outliers in the sample the difference between the inference results for $\bar{Y}_i$ and $\theta_i$ is negligible. Our own simulations for our model also confirm that. Once MCMC samples from the posterior distribution of $\beta, v_i$'s and $\sigma^2_0, \sigma^2_1, \sigma^2_2, p_e$ have been generated, using the NM HB Model the MCMC samples of $Y_{ij}, j = n_i + 1, \cdots, N_i, i = 1, \cdots, m$ from their posterior predictive distributions can be easily generated. Finally, using the relation $\bar{Y}_i = N_i^{-1} \left[ \sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} Y_{ij} \right]$, (posterior predictive) MCMC samples for $\bar{Y}_i$'s can be easily generated for inference on these quantities. In our own data analysis, where the sampling fractions are negligible, we do inference for the approximated predictands $\theta_i$'s.

Chambers and Tzavidis (2006) took a new frequentist approach to small area estimation that is different from the mixed model prediction used in EBLUP. Instead of using a mixed model for the response, they suggested a method based on quantile regression. We briefly review their M-quantile small area estimation method in Section 3. They also proposed an estimator of MSE of their point estimators.

Our Bayesian proposal has two advantages over the REBLUP of Sinha and Rao (2009). First, instead of a working model for the non-outliers, we use an explicit mixture model to specify the joint distribution of responses of all the units in the population, and not only the non-outliers part of the population. It enables us to use all the sampled observations to predict the entire non-sampled part, consisting of outliers and non-outliers, of the population. Our method is robust predictive and the noninformative HB predictors are less susceptible to bias. Second, the main thrust of the EBLUP approach in small area estimation is to develop accurate approximations and estimation of MSEs of EBLUPs (cf. Prasad and Rao, 1990). Datta and Lahiri (2000) and Datta et al. (2005) termed this approximation as second-order accurate approximation, which neglects terms lower order than $m^{-1}$ in the approximation. Second-order accurate approximation results
for REBLUPs have not been obtained by Sinha and Rao (2009). Also, their bootstrap proposal for estimation of the MSE under the working model has not been shown to be second-order accurate. Our HB proposal does not rely on any asymptotic approximations. Analysis of the corn data set and simulation study show better stability and less uncertainty of our method compared to the M-quantile method.

3 M-quantile Small Area Estimation

Small area estimation is dominated by linear mixed effects models where the conditional mean of $Y_{ij}$, the response of the $j$th unit in the $i$th small area, is expressed as $E(Y_{ij}|x_{ij}, v_i) = x_{ij}^T \beta + z_{ij}^T v_i$, where $x_{ij}$ and $z_{ij}$ are suitable known covariates, and $v_i$ is a random effects vector and $\beta$ is a common regression coefficient vector. This assumption is the building block for EBLUPs of small area means, based on suitable additional assumptions for this conditional distribution and the distribution of the random effects. Also with suitable prior distribution on the model parameters, HB methodology for prediction of small area means is developed.

As an alternative to linear regression which models $E(Y|x)$, the mean of the conditional distribution of $Y$ given covariates $x$, the quantile regression has been developed by modeling suitable quantiles of the conditional distribution of $Y$ given $x$. In particular in quantile linear regression, for $0 < q < 1$, the $q$th quantile $Q_q(y, x)$ of this distribution is modeled as $Q_q(y, x) = x^T \beta_q$, where $\beta_q$ is a suitable parameter modeling the linear quantile function. For a given quantile regression function, the quantile coefficient $q_i \in (0, 1)$ of an observation $y_i$ satisfies $Q_{q_i}(y, x_i) = y_i$. For linear quantile function, $q_i$ satisfies $x_i^T \beta_{q_i} = y_i$.

While in linear regression setup the regression coefficient $\beta$ is estimated from a set of data $\{y_i, x_i : i = 1, \cdots, n\}$ by minimizing the sum of squared errors $\sum_{i=1}^n (y_i - x_i^T \beta)^2$ with respect to $\beta$, the quantile regression coefficient $\beta_q$ for a fixed $q \in (0, 1)$ is obtained by minimizing the loss function $\sum_{i=1}^n |y_i - x_i^T b| \{(1 - q)I(y_i - x_i^T b \leq 0) + qI(y_i - x_i^T b > 0)\}$ with respect to $b$. Here $I(\cdot)$ is a usual indicator function.
Following the idea of M-estimation in robust linear regression, Breckling and Chambers (1988) generalized quantile regression by minimizing an objective function \[ \sum_{i=1}^{n} d(|y_i - x_i^T b|) \{(1 - q)I(y_i - x_i^T b \leq 0) + qI(y_i - x_i^T b > 0)\} \] with respect to \( b \) for some given loss function \( d(\cdot) \). [Linear regression is a special case for \( q = .5 \) and \( d(u) = u^2 \).] Estimator of \( \beta_q \) is obtained by solving the equation

\[
\sum_{i=1}^{n} \psi_q(r_{iq}) x_i = 0,
\]

where \( r_{iq} = y_i - x_i^T \beta_q \), \( \psi_q(r_{iq}) = \psi(s^{-1}r_{iq}) \{(1 - q)I(r_{iq} \leq 0) + qI(r_{iq} > 0)\} \), the function \( \psi(\cdot) \), known as influence function in M-estimation, is determined by \( d(\cdot) \) (actually, \( \psi(u) \) is related to the derivative of \( d(u) \), assuming it is differentiable). The quantity \( s \) is a suitable scale factor determined from data (cf. Chambers and Tzavidis, 2006). In M-quantile regression, these authors suggested using \( \psi(\cdot) \) as the Huber Proposal 2 influence function \( \psi(u) = uI(|u| \leq c) + c \text{sgn}(u)I(|u| > c) \), where \( c \) is a given positive number bounded away from 0.

To apply M-quantile method in small area estimation for a set of data \( \{y_{ij}, x_{ij}, \hat{j} = 1, \cdots, n_i, i = 1, \cdots, m\} \), Fabrizi et al. (2012) followed Chambers and Tzavidis (2006) and suggested determining a set of \( \hat{\beta}_q \) in a fine grid for \( q \in (0, 1) \) by solving

\[
\sum_{i=1}^{m} \sum_{j=1}^{n_i} \psi_q(r_{ijq}) x_{ij} = 0,
\]

where \( r_{ijq} = y_{ij} - x_{ij}^T \hat{\beta}_q \). Fabrizi et al. (2012) defined M-quantile estimator of \( \bar{Y}_i \) by

\[
\hat{Y}_{i,MQ} = \frac{1}{N_i} \left[ \sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} x_{ij}^T \hat{\beta}_{qi} + (N_i - n_i)(\bar{y}_i - x_i^T \hat{\beta}_{qi}) \right], \tag{3.1}
\]

where \( (\bar{y}_i, \bar{x}_i) \) is the sample mean of \( \{(y_{ij}, x_{ij}), \hat{j} = 1, \cdots, n_i\} \). Here \( \bar{q}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} q_{ij} \) is the average estimated quantile coefficient of the \( i \)th small area, where \( q_{ij} \) is obtained by solving \( x_{ij}^T \hat{\beta}_q = y_{ij} \), based on the set \( \{\hat{\beta}_q\} \) described above (if necessary, interpolation for \( q \) is made to solve \( x_{ij}^T \hat{\beta}_q = y_{ij} \) accurately). Here we suppress the dependence of \( \hat{\beta}_q \) and \( q_{ij} \) on the influence function \( \psi(\cdot) \). For details on M-quantile small area estimators and associated estimators of MSE based on a pseudo-linearization method, we refer to Chambers et al. (2014).
4 Robust Empirical Best Linear Unbiased Prediction

Empirical best linear unbiased predictors (EBLUPs) of small area means, developed under normality assumptions for the random effects and the unit-level errors, play a very useful role in production of reliable model-based estimation methods. While the EBLUPs are efficient under the normality assumptions, they may be highly influenced by outliers in the data. Sinha and Rao (2009) investigated the robustness of the classical EBLUPs to the departure from normality assumptions and proposed a new class of predictors which are resistant to outliers. Their proposed robust modification of EBLUPs of small area means, which they termed robust EBLUP (REBLUP), downweight any influential observations in the data in estimating the model parameters and the random effects.

Sinha and Rao (2009) considered a general linear mixed effects model with a block-diagonal variance-covariance matrix. Their model, which is sufficiently general to include the popular Fay-Herriot model and the nested error regression model as special cases, is given by

\[ y_i = X_i \beta + Z_i v_i + e_i, \quad i = 1, \ldots, m, \]  

(4.1)

for specified design matrices \(X_i, Z_i\), random effects vector \(v_i\) and unit-level error vector \(e_i\) associated with the data \(y_i\) from the \(i\)th small area. They assumed normality and independence of the random vectors \(v_1, \ldots, v_m, e_1, \ldots, e_m\), where \(v_i \sim N(0, G_i(\delta))\) and \(e_i \sim N(0, R_i(\delta))\). Here \(\delta\) includes the variance parameters associated with the model (4.1).

To develop a robust predictor of a mixed effect \(\mu_i = h_i^T \beta + k_i^T v_i\), Sinha and Rao (2009) started with well-known mixed model equations given by

\[
\sum_{i=1}^{m} X_i^T R_i^{-1}(y_i - X_i \beta - Z_i v_i) = 0, \quad Z_i^T R_i^{-1}(y_i - X_i \beta - Z_i v_i) - G_i^{-1} v_i = 0, \quad i = 1, \ldots, m, \]  

(4.2)

which are derived as estimating equations by differentiating the joint density of \(y_1, \ldots, y_m, v_1, \ldots, v_m\) with respect to \(\beta\), and \(v_1, \ldots, v_m\) to obtain “maximum likelihood” estimators of \(\beta, v_1, \ldots, v_m\) for known \(\delta\). Unique solution \(\tilde{\beta}(\delta), \tilde{v}_1(\delta), \ldots, \tilde{v}_m(\delta)\) to these
equations lead to the BLUP $h_i^T \hat{\beta} + k_i^T \hat{v}_i$ of $\mu_i$. To estimate the variance parameters $\delta$, Sinha and Rao (2009) maximized the profile likelihood of $\delta$, which is the value of likelihood of $\beta, \delta$ based on the joint distribution of the data $y_1, \cdots, y_m$ at $\beta = \bar{\beta}(\delta)$.

To mitigate the impact of outliers on the estimators of variance parameters, the regression coefficients and the random effects, Sinha and Rao (2009) extended the work of Fellner (1986) to robustify all the “estimating equations” by using Huber’s $\psi$–function in M-estimation. Based on the robustified estimating equations, Sinha and Rao (2009) denoted the robust estimators of $\beta, \delta$ and $v_i, i = 1, \cdots, m$ by $\hat{\beta}_M, \hat{\delta}_M$ and $\hat{v}_{iM}$. These estimators lead to REBLUP of $\mu_i$ given by $h_i^T \hat{\beta}_M + k_i^T \hat{v}_{iM}$. For details of the REBLUP and the associated parametric bootstrap estimators of the MSE of the REBLUPs of $\mu_i$, we refer the readers to the paper by Sinha and Rao (2009).

5 Data Analysis

We illustrate our method by analyzing the crop areas data published by Battese et al. (1988) who considered EBLUP prediction of county crop areas for 12 counties in Iowa. Based on U.S. farm survey data in conjunction with LANDSAT satellite data they developed predictors of county means of hectares of corn and soybeans. Battese et al. (1988) were the first to put forward the nested error regression model for the prediction of the county crop areas. Datta and Ghosh (1991) later used HB prediction approach on this data to illustrate the nested error regression model. In the USDA farm survey data on 37 sampled segments from these 12 counties, Battese et al. (1988) determined in their reported data that the second observation for corn in Hardin county was an outlier. In order that this outlier does not unduly affect the model-based estimates of the small area means, Battese et al. (1988) initially recommended, and Datta and Ghosh (1991) subsequently followed, to remove this suspected outlier observation from their analyses. Discarding this observation results in a better fit for the nested error regression model. However, removing any data which may not be a non-representative outlier from analysis will result in loss of valuable information about a part of the non-sampled units of the
population which may contain outliers.

Table 1: Various point estimates and standard errors of county hectares of corn (Full)

<table>
<thead>
<tr>
<th>SA</th>
<th>n_i</th>
<th>DG HB Mean</th>
<th>DG HB SD</th>
<th>NM HB Mean</th>
<th>NM HB SD</th>
<th>SR Mean</th>
<th>SR SD</th>
<th>MQ Mean</th>
<th>MQ SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>123.8 11.7</td>
<td>123.4 9.8</td>
<td>123.7 9.9</td>
<td>130.0 5.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>124.9 11.4</td>
<td>126.6 10.3</td>
<td>125.3 9.7</td>
<td>134.2 8.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>110.0 12.3</td>
<td>108.0 11.3</td>
<td>110.3 9.4</td>
<td>86.0 18.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>114.2 10.7</td>
<td>112.3 10.2</td>
<td>114.1 8.8</td>
<td>114.4 3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>140.3 10.8</td>
<td>142.1 8.1</td>
<td>140.8 7.8</td>
<td>144.2 11.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>110.0 9.6</td>
<td>111.4 7.6</td>
<td>110.8 7.6</td>
<td>108.6 3.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>116.0 9.7</td>
<td>114.3 7.6</td>
<td>115.2 7.3</td>
<td>116.3 4.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>123.2 9.5</td>
<td>122.7 7.9</td>
<td>122.7 7.5</td>
<td>122.5 3.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>112.6 9.9</td>
<td>113.9 6.9</td>
<td>113.5 6.5</td>
<td>115.3 5.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>124.4 8.9</td>
<td>123.5 6.1</td>
<td>124.1 6.3</td>
<td>121.6 4.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>111.3 8.9</td>
<td>108.2 6.8</td>
<td>109.5 6.2</td>
<td>106.9 10.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>130.7 8.3</td>
<td>135.3 7.5</td>
<td>136.9 6.0</td>
<td>135.8 4.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We reanalyze the full data set for corn using our proposed HB method as well as the other methods we reviewed above. In Tables 1 and 2 we reported various point estimates and standard error estimates. We compare our proposed robust HB prediction method with the standard HB method of Datta and Ghosh (1991), and two other robust frequentist methods, the REBLUP method of Sinha and Rao (2009) and the MQ method of Fabrizi et al. (2012) and Chambers et al. (2014). We listed in the table various estimates of county hectares of corn, along with their estimated standard errors or posterior standard deviations. Our analysis of the full data set including the potential outlier from the last small area shows that for the first 11 small areas there is a close agreement among the three sets of point estimates by Datta and Ghosh (1991), Sinha and Rao (2009) and the proposed normal mixture HB method. The Datta and Ghosh method, which was not developed to handle outliers, yields point estimate for the 12th small area that is much different from the point estimates from Sinha-Rao or the proposed NM HB method. The latter two robust estimates are very similar in terms of point estimates for all the small areas. But when we compare these two sets of robust estimates with those from
Table 2: Various point estimates and standard errors of county hectares of corn (Reduced)

<table>
<thead>
<tr>
<th>SA</th>
<th>n_i</th>
<th>DG HB</th>
<th>Mean SD</th>
<th>NM HB</th>
<th>Mean SD</th>
<th>SR</th>
<th>Mean SD</th>
<th>MQ</th>
<th>Mean SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>122.0 11.6</td>
<td>121.7 9.7</td>
<td>122.2 9.9</td>
<td>128.0 3.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>126.4 10.9</td>
<td>127.2 9.7</td>
<td>126.5 9.5</td>
<td>133.4 6.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>107.6 12.4</td>
<td>105.6 10.1</td>
<td>106.7 9.5</td>
<td>94.6 14.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>108.9 10.5</td>
<td>108.2 8.7</td>
<td>111.0 8.3</td>
<td>113.3 3.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>143.6 9.7</td>
<td>144.1 7.0</td>
<td>143.3 7.1</td>
<td>144.2 9.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>112.3 9.7</td>
<td>112.5 6.5</td>
<td>112.3 7.1</td>
<td>114.5 5.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>113.4 9.1</td>
<td>112.5 6.8</td>
<td>112.9 7.1</td>
<td>115.4 3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>121.9 8.8</td>
<td>121.9 6.6</td>
<td>121.9 7.1</td>
<td>122.7 4.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>115.5 9.2</td>
<td>115.7 5.7</td>
<td>115.3 6.4</td>
<td>115.7 4.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>124.8 8.4</td>
<td>124.4 5.4</td>
<td>124.5 5.3</td>
<td>123.1 4.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>107.7 8.5</td>
<td>106.3 5.7</td>
<td>106.8 5.4</td>
<td>105.5 7.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>142.6 9.0</td>
<td>143.5 5.9</td>
<td>143.1 5.8</td>
<td>140.6 4.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Another robust method, namely, the MQ estimates, we find that the MQ estimates for the first three small areas are widely different from those for the other two methods. These numbers possibly indicate a potential bias of the MQ estimates.

To compare performance of all these methods in the absence of any potential outliers, we reanalyzed the corn data by removing a suspected outlier (Our robust HB analysis confirmed the outlier status of this observation, cf. Figure 1 below). When we compare the MQ estimates with four other sets of estimates, the DG HB, the SR, the NM, which are reported in Table 2, and the EB estimates from Table 3 of Fabrizi et al. (2012), we notice a great divide between MQ estimates and the other estimates. Out of the twelve small areas, the estimates for areas 1, 2, 3, 5, and 6 from the MQ method differ substantially from the estimates from the other four methods. On the other hand, the close agreement among the last four sets of estimates also shows in general the usefulness of the robust predictors, the proposed HB predictors and the Sinha-Rao robust EBLUP predictors.

In a recent article Chambers et al. (2014) suggested a bias-corrected version of REBLUP by Sinha-Rao for estimation of small area means for unit-level data in presence of outliers.
However, the associated estimator of MSE (see equation (29) of Chambers et al. (2014)) cannot be applied since the method requires $n_i > p + q = 3 + 1 = 4$ for all small areas, which is not true for the corn data set.

To examine the influence of the outlier on the estimates we compare changes in the estimates from the full data and the reduced data. From Table 2 we find that the largest change occurs, not surprisingly, for the DG HB method for the small area suspected of the outlier. Such a large change occurred since the DG method cannot suitably downweight an outlier. The method is not meant to handle outliers, consequently, it treated the outlier value of 88.59 in the same manner as it treated any other non-outlier observation. As a result, the predictor substantially underestimated the true mean $\bar{Y}_i$ for the Hardin county. The next largest difference occurred for the MQ method for small area 3 which is not known to include any outlier. Such a large change is contrary to behavior of a robust method.

The changes in point estimates for the robust HB and the REBLUP methods are moderate for the areas not known to include any outliers, and the changes seem proportionate for the small area suspected of an outlier. The corresponding changes in the estimates from the MQ method for some of the areas not including any outlier seem dispropor-
tionately large, and change in the estimate for the area suspected of an outlier is not as large. This behavior to some extent indicates a lack of robustness of the MQ method to outliers.

An inspection of the posterior standard deviations of the two Bayesian methods reveals some interesting points. First, the posterior SDs of small area means for the proposed mixture model appear to be substantially smaller than the posterior SDs associated with the Datta-Ghosh HB estimators. Smaller posterior SDs suggest posterior distribution of the small area means under the mixture model are more concentrated than those under the Datta-Ghosh model. It has been confirmed by simulation study, reported in the next section.

Next, when we compare the posterior SDs of small area means for our proposed method based on the full data and the reduced data, all posterior SDs increase for the full data (which likely contain an outlier). In the presence of outliers, unit-level variance is expected to be large. Even though the posterior SDs of small area means do not depend entirely only on the unit-level error variance, it is expected to increase with this variance. This monotonic increasingness appears reasonable due to the suspected outlier. While this intuitive property holds for our proposed method, it does not hold for the standard Datta-Ghosh method.

For further demonstration of effectiveness of our proposed robust HB method, we computed model parameter estimates for both the reduced and the full data sets. These estimates are displayed in Table 3. The HB estimate of the larger variance component (976, based on mean) of the mixture is much larger than the estimate of the smaller component (182) for the full data, indicating a necessity of the mixture model. On the other hand, for the reduced data the estimates of variances for the two mixing components, which, respectively, are 231 and 121, are very similar and can be argued identical within errors in estimation, indicating limited need of the mixture distribution. A comparison of the estimates of $p_e$ for the two cases also reveals appropriateness of the mixture model for the full data, and its absence for the reduced data.
Table 3: Parameter estimates for various models with and without the suspected outlier

<table>
<thead>
<tr>
<th>Estimates</th>
<th>Datta-Ghosh HB</th>
<th>Datta-Ghosh HB</th>
<th>Proposed Mixture HB</th>
<th>Proposed Mixture HB</th>
<th>Sinha-Rao</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MEAN</td>
<td>MEDIAN</td>
<td>MEAN</td>
<td>MEDIAN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Full Data</td>
<td>Reduced Data</td>
<td>Full Data</td>
<td>Reduced Data</td>
<td>Full Data</td>
</tr>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>17.29</td>
<td>50.35</td>
<td>16.17</td>
<td>50.92</td>
<td>30.89</td>
</tr>
<tr>
<td></td>
<td>31.46</td>
<td>50.78</td>
<td>29.14</td>
<td>48.20</td>
<td></td>
</tr>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>0.37</td>
<td>0.33</td>
<td>0.37</td>
<td>0.33</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.33</td>
<td>0.35</td>
<td>0.33</td>
<td>0.36</td>
</tr>
<tr>
<td>( \hat{\beta}_2 )</td>
<td>-0.03</td>
<td>-0.13</td>
<td>-0.03</td>
<td>-0.13</td>
<td>-0.07</td>
</tr>
<tr>
<td></td>
<td>-0.07</td>
<td>-0.13</td>
<td>-0.07</td>
<td>-0.13</td>
<td>-0.07</td>
</tr>
<tr>
<td>( \hat{\sigma}_1^2 )</td>
<td>175.68</td>
<td>231.87</td>
<td>127.68</td>
<td>186.07</td>
<td>205.01</td>
</tr>
<tr>
<td></td>
<td>160.22</td>
<td>203.55</td>
<td>102.74</td>
<td>155.15</td>
<td></td>
</tr>
<tr>
<td>( \hat{\sigma}_2^2 )</td>
<td>370.00</td>
<td>216.00</td>
<td>341.00</td>
<td>192.00</td>
<td>976.00</td>
</tr>
<tr>
<td></td>
<td>483.00</td>
<td>188.00</td>
<td>225.60</td>
<td>161.50</td>
<td></td>
</tr>
</tbody>
</table>

The posterior density in a reasonable noninformative Bayesian analysis is usually dominated by the likelihood of the parameters generated by the data. In case the data do not provide much information about some parameters to the likelihood, posterior densities of such parameters will be dominated by their prior information. Consequently, posterior distribution for some of them may be very similar to the prior distribution. An overparameterized likelihood usually carries little information for some parameters responsible for overparameterization. In particular, if our mixture model is overparameterized in the sense that variances of mixture components are similar, then the integrated likelihood may be flat on the mixing proportion. We observe this scenario in our data analysis when we removed the suspected outlier observation from analysis based on our model. Since our mixture model is meant to accommodate outliers based on unequal variances for the mixing components, in the absence of any outliers the mixture of two normal distributions may not be required. In particular, we noticed earlier that with the suspected outlier removed the estimates of the two variance components \( \sigma_1^2 \) and \( \sigma_2^2 \) are very similar. Also, the posterior histogram of the mixing proportion \( p_e \), not presented here, resembles a uniform distribution, the prior distribution assigned in our Bayesian analysis. In fact, the posterior mean of this parameter for the reduced data is the same as the prior mean 0.5. This essentially says that the likelihood is devoid of any information about \( p_e \) to
update the prior distribution.

One advantage of our mixture model is that it explicitly models any representative outlier through the latent indicator variable \( Z_{ij} \). By computing posterior probability of \( Z_{ij} = 0 \) we can compute the posterior probability that an observed \( y_{ij} \) is an outlier. While the REBLUP method does not give similar measure for an observation, one can determine the outlier status by computing standardized residual associated with an observation.

To show the effectiveness of our method, in Figure 1, we plotted the posterior probabilities of an individual observation being an outlier against the observation’s standardized residual. In the left panel, we showed the plot of these posterior probabilities for the full data, and in the right panel we included the same by removing the suspected outlier. These two figures are in sharp contrast; the left panel clearly showed that there is a high probability (0.86) that the second observation in the Hardin county is an outlier. The associated large negative standardized residual of this observation also confirmed that, and from this plot an approximate monotonicity of these posterior probabilities with respect to the absolute values of the standardized residuals may also be discerned. However, the right panel shows that for the reduced data excluding the suspected outlier, the standardized residuals for the remaining observations are between \(-3\) and \(3\), with the associated posterior probabilities of being outlier observations are all between 0.44 and 0.64. None of these probabilities is particularly larger than prior probability 0.5 to indicate outlier status of that corresponding observation. This little change of the outlier prior probabilities in the posterior distribution for the reduced data essentially confirms that a discrete scale mixture of normal distributions is not supported by the data, or in other words, the scale mixture model is not required to explain the data, which is the same as that there are possibly no outliers in the data set.

## 6 A Simulation Study

In our extensive simulation study, we followed the simulation setup used by Sinha and Rao (2009). Corresponding to the model in (2.1), we use a single auxiliary variable \( x \), which we
generated independently from a normal distribution with mean 1 and variance 1. In our simulations we use \( m = 40 \). We generated 40 sets of 200 (= \( N_i \)) values of \( x \) to create the finite population of covariates for the 40 small areas. Based on these simulated values we computed \( \bar{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij} \). Throughout our simulations we keep the generated \( x \) values fixed. We used these generated \( x_{ij} \) values and generated \( v_i, i = 1, \cdots, m \) independently from \( N(0, \sigma_v^2) \) with \( \sigma_v^2 = 1 \). We generated \( e_{ij}, j = 1, \cdots, N_i, i = 1, \cdots, m \) as iid from one of three possible distributions: (i) the case of no outliers where \( e_{ij} \) are generated from \( N(0, 1) \) distribution; (ii) a mixture of normal distributions, with 10% outliers from \( N(0, 25) \) distribution and the remaining 90% from the \( N(0, 1) \) distribution; and (iii) \( e_{ij} \)'s are iid from a \( t \)-distribution with 4 degrees of freedom. We also took \( \beta_0 = 1 \) and \( \beta_1 = 1 \) as in Sinha and Rao (2009), and generated \( m \) small area finite populations based on the generated \( x_{ij} \)'s, \( v_i \)'s and \( e_{ij} \)'s by computing \( Y_{ij} = \beta_0 + \beta_1 x_{ij} + v_i + e_{ij} \) based on the NER model in (2.1). Our goal is prediction of finite population small area means \( \bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}, i = 1, \cdots, m \). After examining no significant difference between \( \bar{Y}_i \) and \( \beta_0 + \beta_1 \bar{X}_i + v_i = \theta_i \) (say) in the simulated populations, as in Sinha and Rao (2009), we also consider prediction of \( \theta_i \).

From each simulated small area finite population we selected a simple random sample of size \( n_i = 4 \) for each small area. Based on the selected samples we derived the HB predictors of Datta and Ghosh (1991) (referred as DG), the REBLUPs of Sinha and Rao (2009) (referred as SR), MQ predictors of Chambers et al. (2014) (referred as CCST-MQ, based on their equation (38)) and our proposed robust HB predictors (referred as NM). In addition to the point predictors we also obtained the posterior variances of both the HB predictors and the estimates of the MSE of the REBLUPs based on the bootstrap method proposed by Sinha and Rao (2009), and the estimates of MSE of MQ predictors, obtained by using pseudo-linearization in equation (39) by Chambers et al. (2014).

For each simulation setup, we have simulated \( S = 100 \) populations. For the \( s \)th created population, \( s = 1, \cdots, S \), we computed the values of \( \theta_i^{(s)} \), which will be treated as the true values. We denote the \( s \)th simulation sample by \( d^{(s)} \), and based on this data we
calculate the REBLUP predictors $\hat{\theta}^{(s)}_{i,SR}$ and their estimated MSE, $mse(\hat{\theta}^{(s)}_{i,SR})$ using the procedure proposed by Sinha and Rao (2009). To assess the accuracy of the point predictors we computed empirical bias $eB_{i,SR} = \frac{1}{S} \sum_{s=1}^{S}(\hat{\theta}^{(s)}_{i,SR} - \theta^{(s)}_{i})$ and empirical MSE $eM_{i,SR} = \frac{1}{S} \sum_{s=1}^{S}(\hat{\theta}^{(s)}_{i,SR} - \theta^{(s)}_{i})^2$. Treating $eM_{i,SR}$ as the “true” measure of variability of $\hat{\theta}_{i,SR}$, we also evaluate the accuracy of the MSE estimator $mse(\hat{\theta}_{i,SR})$, suggested by Sinha and Rao (2009). Accuracy of the mse estimator is evaluated by the percent difference between the empirical MSE and the average (over simulations) estimated mse, given by $RE_{mse,SR,i} = 100\{(1/S) \sum_{s=1}^{S} mse(\hat{\theta}^{(s)}_{i,SR}) - eM_{i,SR}\}/eM_{i,SR}$. Similarly, we obtained the predictors $\hat{\theta}^{(s)}_{i,CCST}$, estimated MSEs $mse(\hat{\theta}^{(s)}_{i,CCST})$ of Chambers et al. (2014), empirical biases and empirical MSEs of point estimators and relative biases of the estimated MSEs. Using the point estimates and MSE estimates we created approximate 90% prediction intervals $I_{i,SR,90}^{(s)} = [\hat{\theta}^{(s)}_{i,SR} - 1.645 \sqrt{mse(\hat{\theta}^{(s)}_{i,SR})}, \hat{\theta}^{(s)}_{i,SR} + 1.645 \sqrt{mse(\hat{\theta}^{(s)}_{i,SR})}]$ and 95% prediction intervals $I_{i,SR,95}^{(s)} = [\hat{\theta}^{(s)}_{i,SR} - 1.96 \sqrt{mse(\hat{\theta}^{(s)}_{i,SR})}, \hat{\theta}^{(s)}_{i,SR} + 1.96 \sqrt{mse(\hat{\theta}^{(s)}_{i,SR})}]$. We also obtained similar intervals for the MQ method of Chambers et al. (2014). We evaluated empirical biases, empirical MSEs, relative biases of estimated MSEs, and empirical coverage probabilities of prediction intervals for all four methods. These quantities for all 40 small areas are plotted in Figures 2, 3 and 4.
Figure 2: Plot of empirical biases and empirical MSEs of $\hat{\theta}$s
We plotted the empirical biases on the left panel and the empirical MSEs on the right panel of Figure 2. These estimators do not show any systematic bias. In terms of eM, the REBLUP and the proposed NM HB predictor appear to be most accurate and perform similarly (in fact, based on all evaluation criteria considered here, the proposed NM HB and the REBLUP methods have equivalent performance). In terms of eM, the MQ predictor has maximum variability and the standard DG HB predictor is in third place. In the case of no outliers, while the other three predictors have the same eM, the MQ predictor is slightly more variable. Moreover, we examined how closely the posterior variances of the Bayesian predictors and the MSE estimators of the frequentist robust predictors track their respective eM of prediction (see Figure 3). The posterior variance of the proposed NM HB predictor and the estimated MSE of REBLUP appear to track the eM the best without any evidence of bias. The posterior variance of the standard HB predictor appears to overestimate the eM and the estimated MSE of the MQ predictor appears to underestimate. An undesirable consequence of this negative bias of the MSE estimator of the MQ method is that the related prediction intervals often fail to cover the true small area means (see the plots in Figure 4).

Our sampling-based Bayesian approach allowed us to create credible intervals for the small area means at the nominal levels of 0.90 and 0.95 based on sample quantiles of the Gibbs samples of the $\theta_i$’s. For the Sinha-Rao and the Chambers et al. methods we used their respective estimated root MSE of the REBLUPs or MQ-predictors to create symmetric approximate 90% and 95% prediction intervals of the small area means.

Next, to assess the coverage rate of these prediction intervals we computed empirical coverage probabilities $eC_{i,SR,90} = \frac{1}{S} \sum_{s=1}^{S} I[\theta_i^{(s)} \in \hat{I}_{i,SR,90}]$ and $eC_{i,SR,95} = \frac{1}{S} \sum_{s=1}^{S} I[\theta_i^{(s)} \in \hat{I}_{i,SR,95}]$, where $I[x \in A]$ is the usual indicator function that is one for $x \in A$ and 0 otherwise.

Based on the same setup and same set of simulated data we also evaluated the two HB procedures. In the Bayesian approach, the point predictor, the posterior variance and the credible intervals for $\theta_i^{(s)}$ in the $s$th simulation were computed based on the
MCMC samples of $\theta_i^{(s)}$ from its posterior distribution, generated by Gibbs sampling. The posterior mean and posterior variance are computed by the sample mean and the sample variance of the MCMC samples. An equi-tailed $100(1 - 2\alpha)\%$ credible interval for $\theta_i^{(s)}$ is created, where the lower limit is the $100\alpha$th sample percentile and the upper limit is the $100(1 - \alpha)$th sample percentile of the MCMC samples of $\theta_i^{(s)}$ from the $s$th simulation.

Suppose in the $s$th simulation $\hat{\theta}_{i,DG}^{(s)}$ denotes the Datta-Ghosh HB predictor of $\theta_i$ and $V_{i,DG}^{(s)}$ denotes the posterior variance. The empirical bias of the Datta-Ghosh predictor of $\theta_i$ is defined by $eB_{i,DG} = \frac{1}{S} \sum_{s=1}^{S} (\hat{\theta}_{i,DG}^{(s)} - \theta_i^{(s)})$ and empirical MSE by $eM_{i,DG} = \frac{1}{S} \sum_{s=1}^{S} (\hat{\theta}_{i,DG}^{(s)} - \theta_i^{(s)})^2$. To investigate the extent $V_{i,DG}^{(s)}$ may be interpreted as an estimated mse of the predictor $\hat{\theta}_{i,DG}$, we compute the percent difference between the empirical MSE and the average (over simulations) posterior variance, given by $RE_{V,DG,i} = 100\left\{ \frac{1}{S} \sum_{s=1}^{S} V_{i,DG}^{(s)} - eM_{i,DG} \right\} / eM_{i,DG}$. These quantities for all 40 small areas are plotted in Figure 3.

Based on the MCMC samples of $\theta_i$’s for the $s$th simulated data set, let $I_{i,DG,90}^{(s)}$ be the 90% credible interval for $\theta_i$. To evaluate the frequentist coverage probability of the credible interval for $\theta_i$, we computed empirical coverage probabilities $eC_{i,DG,90} = \frac{1}{S} \sum_{s=1}^{S} I[\theta_i^{(s)} \in I_{i,DG,90}^{(s)}]$. Corresponding to a credible interval $I_{i,DG,90}^{(s)}$, we use $L_{i,DG,90}^{(s)}$ to denote its length, and computed empirical average length of a 90% credible interval for $\theta_i$ based on Datta-Ghosh approach by $\bar{L}_{i,DG,90} = \frac{1}{S} \sum_{s=1}^{S} L_{i,DG,90}^{(s)}$. Similarly, we computed $eC_{i,DG,95}$ and $\bar{L}_{i,DG,95}$ for the 95% credible intervals for $\theta_i$.

Finally, as we did for the Datta-Ghosh HB predictor, we computed similar quantities for our new robust HB predictor. Specifically, suppose $\hat{\theta}_{i,NM}^{(s)}$ is the newly proposed NM HB predictor of $\theta_i^{(s)}$ and $V_{i,NM}^{(s)}$ is the posterior variance. For the new predictor we define the empirical bias by $eB_{i,NM} = \frac{1}{S} \sum_{s=1}^{S} (\hat{\theta}_{i,NM}^{(s)} - \theta_i^{(s)})$ and empirical MSE by $eM_{i,NM} = \frac{1}{S} \sum_{s=1}^{S} (\hat{\theta}_{i,NM}^{(s)} - \theta_i^{(s)})^2$. Again, to investigate the extent $V_{i,NM}^{(s)}$ may be viewed as an estimated mse of the predictor $\hat{\theta}_{i,NM}$, we computed the percent difference between the empirical MSE and the average (over simulations) posterior variance, given by
\[ RE_{V-NM,i} = 100\left\{(1/S) \sum_{s=1}^{S} V_{i,NM}^{(s)} - eM_{i,NM}\right\}/eM_{i,NM}. \] These quantities for all 40 small areas are plotted in Figure 3. Based on the MCMC samples of \( \theta_i \)'s for the \( s \)th simulated data set, let \( I_{i,NM,90}^{(s)} \) be the 90\% credible interval for \( \theta_i \). To evaluate the frequentist coverage probability of the credible interval for \( \theta_i \) we computed empirical coverage probabilities \( eC_{i,NM,90} = \frac{1}{S} \sum_{s=1}^{S} I_\left[ \theta_i^{(s)} \in I_{i,NM,90}^{(s)} \right] \). Corresponding to a credible interval \( I_{i,NM,90}^{(s)} \), we use \( L_{i,NM,90}^{(s)} \) to denote its length, and computed empirical average length of a 90\% credible interval for \( \theta_i \) based on new approach by \( L_{i,NM,90} = \frac{1}{S} \sum_{s=1}^{S} L_{i,NM,90}^{(s)} \). Similarly, we computed \( eC_{i,NM,95} \) and \( L_{i,NM,95} \) for the 95\% credible intervals for \( \theta_i \).

We plotted the empirical coverage probabilities for the four methods that we considered in this article. The plot reveals significant undercoverage of the approximate prediction intervals created by using the estimated prediction MSE proposed by Chambers et al. (2014). This undercoverage is not surprising since their estimated MSE mostly underestimates the true MSE (measured by the eM) (see Figure 3). Coverage probabilities of the Sinha-Rao prediction intervals and the two Bayesian credible intervals are remarkably accurate. This lends dual interpretation of our proposed credible intervals, Bayesian by construction, and frequentist by simulation validation. This property is highly desirable to practitioners, who often do not care for a paradigm or a philosophy. In the same plot, we also plotted the ratio of the average lengths of the DG credible intervals to the newly proposed robust HB credible intervals. These plots show the superiority of the proposed method, yielding intervals which meet coverage accurately with average lengths about 25-30\% shorter compared to the DG method for normal mixture model with 10\% contamination. Again these two intervals meet the coverage accurately when the unit-level errors are generated from normal (no outliers) or a moderately heavy-tail distribution (\( t_4 \)). In these cases, the reduction in length of the intervals is less, which is about 10\%. This shorter prediction intervals from the new method even for normal distribution for the unit-level error is interesting; it shows that the proposed method does not lose any efficiency in comparison with the Datta-Ghosh method even when the normality of the unit-level errors holds.
The comparison of NM HB prediction intervals and the Sinha-Rao prediction intervals yields a mixed picture. In the mixture setup, the NM HB prediction intervals attained coverage probability more accurately than the Sinha-Rao intervals, which undercover by 1%, and on an average the Bayesian prediction intervals are about 2% shorter than the frequentist intervals. When the data are simulated from a $t_4$ distribution, the coverage probabilities of the Sinha-Rao prediction intervals are about 1% below the target, but these intervals are about 3% shorter than the NM HB prediction intervals, which attained the nominal coverage. Finally, when the population does not include any outlier, these two methods perform the same, both attained the nominal coverage and yield the same average length.

7 Conclusion

The NER model by Battese et al. (1988) plays an important role in small area estimation for unit-level data. While Battese et al. (1988), Prasad and Rao (1990) and Datta and Lahiri (2000) investigated EBLUPs of small area means, Datta and Ghosh (1991) proposed an HB approach for this model. Sinha and Rao (2009) investigated robustness of the MSE estimates of EBLUPs in Prasad and Rao (1990) for outliers in the response. They showed in presence of outliers robustness of their REBLUPs and lack of robustness of the EBLUPs.

In this article we showed that non-robustness also persists for the HB predictors by Datta and Ghosh (1991). To deal with this undesirable issue we proposed an alternative to the HB predictors by using a mixture of normal distributions for the unit-level error part of the NER model. An illustrative application and simulation study show the superiority of our proposed method over the existing HB, EBLUP and M-quantile solutions. Indeed simulation results show superiority of our method over the Datta and Ghosh (1991) HB predictors and the M-quantile small area estimators of Chambers et al. (2014). Performance of our proposed NM HB method is found to be as good as the frequentist solution of Sinha and Rao (2009). Our proposed Bayesian intervals also achieve the
corresponding frequentist coverage. Thus, unlike the frequentist solutions, our proposed
HB solution enjoys dual interpretation, Bayesian by construction, and frequentist via
simulation, a feature attractive to practitioners. Moreover, suggested credible intervals
are shorter in length in comparison with the other nominal prediction intervals. In fact,
the application and simulations show the proposed NM HB method is the best among
the four methods in presence of outliers. Our proposed method is as good as the HB
method of Datta and Ghosh (1991), even in absence of outliers. Thus there will be no
loss in using proposed HB method for all data sets.

References

prediction of county crop areas using survey and satellite data, *Journal of the American

Bell, W. R. and Huang, E. T. (2006), Using t-distribution to deal with outliers in small
area estimation, *Proceedings of Statistics Canada Symposium 2006 Methodological is-
suces in measuring population health*.

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 and Survey

Chambers, R. L. (1986), Outlier robust finite population estimation, *Journal of the Amer-
ican Statistical Association*, 81, 1063–1069.

Datta, G. and Ghosh, M. (1991), Bayesian prediction in linear models: Applications to

Datta, G. S. and Lahiri, P. (1995), Robust hierarchical Bayesian estimation of small area
characteristics in presence of covariates and outliers, *Journal of Multivariate Analysis*,
54, 310–328.


Figure 3: Plot of posterior variances and MSE estimates and their empirical relative biases
Figure 4: Plot of lengths and coverages of credible and prediction intervals.
Robust Hierarchical Bayes Small Area Estimation for Nested Error Regression Model

8 Supplementary Materials

8.1 Exploration of the Propriety of the Posterior Density

Since improper prior distribution has been used in the HB model proposed in this article, it is important to explore the propriety of the resulting posterior distribution in order to avoid misleading results based on improper posteriors (cf. Hobert and Casella, 1996). In the following results, we first provide sufficient conditions for the propriety of the resulting posterior distribution based on the proposed model. We relax the condition $n_i \geq 2$ for all areas in the corollary below.

**Theorem 8.1** Let $\sum_{i=1}^{m} n_i = n$. The following conditions are sufficient for the propriety of the posterior distribution under the proposed model:

(a) $n_i \geq 2$ for $i = 1, \ldots, m$,
(b) $n \geq 2m + 2p - 1$,
(c) $m \geq p + 6$.

A detailed proof of Theorem 8.1 is provided in Section 8.2 of the Supplementary Materials. While Theorem 8.1 appears to be restrictive, the following corollary and lemma show that it is not the case.

**Corollary 8.2** If there exists a set $S$ of $m'$ ($m' \leq m$) small areas such that

(a) $n_i \geq 2$, $n_i$ being the number of sampled units from the $i^{th}$ small area, $i \in S$,
(b) $\sum_{i \in S} n_i \geq 2m' + 2p - 1$,
(c) $m' \geq p + 6$,

then the posterior distribution under the proposed model will be proper.

**Proof of Corollary 8.2**: Proof follows from an Application of the lemma below.
Lemma 8.3 Let \( \theta \sim \pi(\theta) \) and \( d|\theta \sim f(d|\theta) \). We partition \( d \) as \( d = (d^{(1)T}, d^{(2)T})^T \). If the posterior distribution \( \theta|d^{(1)} \) is proper, then the posterior distribution \( \theta|d \) is also proper.

Suppose there exists \( m' (\leq m) \) small areas which satisfy conditions (a), (b) and (c) of Theorem 8.1. Let \( S_{m'} \) be the set of small areas with at least two sampled units and \( S_{m'}^c \) contain rest of the small areas. Let us partition the responses for the sampled units as follows:

\[
d^{(1)} = \{ y_{ij} : i \in S_{m'}; j = 1, \ldots, n_i \} \quad \text{and} \quad d^{(2)} = \{ y_{ij} : i \in S_{m'}^c; j = 1, \ldots, n_i \}.
\]

Let \( \theta \) be the set of model parameters. By Theorem 8.1, \( f(\theta|d^{(1)}) \) is proper. Now, applying Lemma 8.3, we can say \( f(\theta|d) = f(\theta|d^{(1)}, d^{(2)}) \) is proper. This proves Corollary 8.2.

8.2 Proof of the Theorem

Proof of Theorem 8.1: We assume that there are at least two sampled units for each small area, i.e. \( n_i \geq 2, i = 1, \ldots, m \); and \( n \geq 2m + 2p - 1 \), where \( n = \sum_{i=1}^{m} n_i \). At first, we consider the case when \( n = 2m + 2p - 1 \), the argument can be extended to the case \( n > 2m + 2p - 1 \) by applying Lemma 8.3. Under the proposed model, the joint pdf of \( y_{ij} \)'s, \( j = 1, \ldots, n_i, i = 1, \ldots, m; v (m \times 1), \beta (p \times 1), \sigma^2_1, \sigma^2_2, \sigma^2_v \) and \( p_e \) is given by

\[
f(y, v, \beta, \sigma^2_1, \sigma^2_2, \sigma^2_v, p_e) \propto \sum_{\Omega} \left[ \prod_{i=1}^{m} \left( \prod_{k=1}^{n_{i1}} \frac{p_e}{\sqrt{\sigma^2_1}} \exp \left( -\frac{1}{2} \frac{(y_{ijk} - x_{ijk}^T \beta - v_i)^2}{\sigma^2_1} \right) \right) \right] \times \left[ \prod_{k=n_{i1}+1}^{n_i} (1-p_e) \exp \left( -\frac{1}{2} \frac{(y_{ijk} - x_{ijk}^T \beta - v_i)^2}{\sigma^2_2} \right) \right] \times \frac{1}{(\sigma^2_v)^m} \exp \left( -\frac{1}{2} \sum_{i=1}^{m} \frac{v_i^2}{\sigma^2_v} \right) \times \frac{1}{(\sigma^2_2)^2} I(\sigma^2_1 < \sigma^2_2) \tag{8.1}
\]

The summation \( \sum_{\Omega} \) and the quantities \( n_{i1}, n_{i2}, i = 1, \ldots, m \) are explained below. Let \( z_{ij} = 1 \), if the \( j^{th} \) sampled unit of the \( i^{th} \) small area corresponds to the mixture component \( \sigma^2_1 \) and \( z_{ij} = 0 \) otherwise. The set \( \Omega \) contains all possible choices of \( z = (z_{11}, \ldots, z_{mm}) \) vector. Hence the cardinality of \( \Omega \) is \( 2^n \). For a given \( z \), let \( n_{i1} = \sum_{j=1}^{n_i} z_{ij} \) and \( n_{i2} = n_i - n_{i1} \) for \( i = 1, \ldots, m \). Then \( n_{i1} \) is the number of units from the \( i^{th} \) small area whose unit-level
variance corresponds to the mixture component $\sigma_i^2$. The remaining $n_{i2}$ units from the $i^{th}$ small area corresponds to the mixture component $\sigma_i^2$.

Define, $S_1 = \{i : n_{i1} > 0\}$ and $S_2 = \{i : n_{i2} > 0\}$. Clearly, $S_1 \cup S_2 = \{1, \ldots, m\}$ and $S_1 \cap S_2$ may not be an empty set. Let $m_i$ be the cardinality of $S_i$, $i = 1, 2$, then $m \leq m_1 + m_2$. Note that $n_{i1}$ or $n_{i2}$ can be zero for some $i$. Indeed, if $i \notin S_1, n_{i1} = 0$ and if $i \notin S_2, n_{i2} = 0$. Define, $n_1^* = \sum_{i \in S_1} n_{i1}$ and $n_2^* = \sum_{i \in S_2} n_{i2}$.

From (8.1), a typical term under the sum over $\Omega$ is,

$$\varphi(y, v, \beta, \sigma_1^2, \sigma_2^2, \sigma_v^2, p_e) \quad = C \times p_e^{n_1^*} (1 - p_e)^{n_2^*} \times \frac{1}{(\sigma_1^2)^{\frac{n_1^*}{2}}} \times \exp \left( -\frac{1}{2\sigma_1^2} \sum_{i \in S_1} \sum_{k=1}^{n_i} (y_{ijk} - x_{ijk}^T \beta - v_i)^2 \right) \times \frac{1}{(\sigma_2^2)^{\frac{n_2^*}{2}}} \times \exp \left( -\frac{1}{2\sigma_2^2} \sum_{i \in S_2} \sum_{k=n_{i1}+1}^{n_i} (y_{ijk} - x_{ijk}^T \beta - v_i)^2 \right) \times \frac{1}{(\sigma_v^2)^{\frac{m}{2}}} \exp \left( -\frac{1}{2} \sum_{i=1}^{m} \frac{v_i^2}{\sigma_v^2} \right) \times \frac{I(\sigma_i^2 < \sigma_2^2)}{(\sigma_2^2)^2},$$

where $C$ is a generic, positive constant. In order to check the integrability of $f(y, v, \beta, \sigma_1^2, \sigma_2^2, \sigma_v^2, p_e)$ with respect to $\beta, v, \sigma_1^2, \sigma_2^2, \sigma_v^2, p_e$ in (8.1), we need to check the integrability of each typical term in (8.1) with respect to $\beta, v, \sigma_1^2, \sigma_2^2, \sigma_v^2, p_e$.

We introduce the following notation: $y_1 = \text{col}_{i \in S_1} \text{col}_{1 \leq k \leq n_{i1}} y_{ijk}; \quad X_1 = \text{col}_{i \in S_1} \text{col}_{1 \leq k \leq n_{i1}} x_{ijk}^T$ and $y_2 = \text{col}_{i \in S_2} \text{col}_{n_{i1}+1 \leq k \leq n_i} y_{ijk}; \quad X_2 = \text{col}_{i \in S_2} \text{col}_{n_{i1}+1 \leq k \leq n_i} x_{ijk}^T$, $Z_1 = \bigoplus_{i=1}^{m} 1_{n_{i1}}$ and $Z_2 = \bigoplus_{i=1}^{m} 1_{n_{i2}}$.

Note that, there are $m_1$ and $m_2$ components of $v$ are involved in $Z_1 v$ and $Z_2 v$ respectively.

Let the rank of $X_1 (n_1^* \times p)$ and $X_2 (n_2^* \times p)$ be $p_1$ and $p_2$ respectively, where $p_1 + p_2 \geq p$.

We now state the lemma below.

**Lemma 8.4** If $n = 2m + 2p - 1$ and $m \geq p + 6$, then one of the following conditions must hold. (a) $n_1^* \geq m_1 + p_1$, $m_1 > 3$ or (b) $n_2^* \geq m_2 + p_2$, $m_2 > 3$.

The proof of Lemma 8.4 is provided in Section 8.3. Without loss of generality, for the rest of the proof, we assume that $n_1^* \geq m_1 + p_1$ and $m_1 > 3$. Had we assumed $n_2^* > m_2 + p_2$, $m_2 > 3$, it will lead us to establish the same results. Note that we do not have to make
separate assumptions for $n^*_1$ and $m_1$, they come from the assumptions $n \geq 2m + 2p - 1$ and $m \geq p + 6$.

Without loss of generality, we assume that the rows of $X_1$ are arranged such that the first $p_1$ rows are linearly independent. These rows constitute a submatrix $X_{11}(p_1 \times p)$, the rest of the rows of $X_1$ can be expressed as $A_{21}X_{11}$ for some matrix $A_{21}((n^*_1 - p_1) \times p_1)$. Similarly, we assume that first $p_2$ rows of $X_2$ are so arranged that they are linearly independent. Since, $p_2 \geq p - p_1$, we further assume that the first $(p - p_1)$ of these $p_2$ rows are linearly independent of the rows of $X_{11}$, we denote this portion of $X_2$ as the submatrix $X_{211}((p - p_1) \times p_1)$.

Let, $X_{212}$ consists next $p_2 - (p - p_1)$ linearly independent rows of $X_2$, and $X_{22}$ contains the remaining $(n^*_2 - p_2)$ rows. Hence,

$$X_1 = \begin{pmatrix} X_{11} \\ A_{21}X_{11} \end{pmatrix} \quad \text{and} \quad X_2 = \begin{pmatrix} X_{211} \\ X_{212} \\ X_{22} \end{pmatrix}$$

where, $\text{rank}(X_{11}) = p_1$.

According to the construction of the matrices, $X_{212} = H \begin{pmatrix} X_{11} \\ X_{211} \end{pmatrix}$ for some $H = \begin{pmatrix} H_{21} & H_{22} \end{pmatrix}$, note that $H_{21} \neq 0$. we can write, $X_{22} = A_{22} \begin{pmatrix} X_{211} \\ X_{212} \end{pmatrix}$ for some $A_{22}((n^*_2 - p_2) \times p_2)$.

We consider the transformation: $\rho_1 = X_{11}\beta$ and $\rho_2 = X_{211}\beta; \rho = (\rho_1^T, \rho_2^T)^T$.

Now, $X_{11}\beta = \begin{pmatrix} X_{11} \\ A_{21}X_{11} \end{pmatrix} \beta = \begin{pmatrix} I_{p_1} \\ A_{21} \end{pmatrix} X_{11}\beta = M_1\rho_1$, where $M_1 = \begin{pmatrix} I_{p_1} \\ A_{21} \end{pmatrix}$, $\text{rank}(M_1) = p_1$.

Similarly,

$$X_{212}\beta = H \begin{pmatrix} X_{11}\beta \\ X_{211}\beta \end{pmatrix} = H \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} = H\rho \quad \text{and} \quad X_{22}\beta = A_{22} \begin{pmatrix} X_{211} \\ X_{212} \end{pmatrix} \beta = A_{22} \begin{pmatrix} \rho_2 \\ H\rho \end{pmatrix} = A_{22} \begin{pmatrix} 0 & I \\ H_{21} & H_{22} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} = A_{22}^*\rho,$$
hence, $X_2\beta = \begin{pmatrix} X_{211} \\ X_{212} \\ X_{22} \end{pmatrix} \beta = \begin{pmatrix} \rho_2 \\ H \rho_2 \\ A_{22}^* \rho_2 \end{pmatrix} = \begin{pmatrix} \rho_2 \\ G_\rho \\ A_{22}^* G_\rho \end{pmatrix}$, where $G = \begin{pmatrix} H \\ A_{22}^* \end{pmatrix}$, we partition $y_2$ and $Z_2$

according to the partitioned rows of $X_2$, i.e., $y_2 = \begin{pmatrix} y_{211} \\ y_{212} \\ y_{22} \end{pmatrix} = \begin{pmatrix} y_{211} \\ y_{22} \end{pmatrix}$, where $y_2^* = \begin{pmatrix} y_{212} \\ y_{22} \end{pmatrix}$

and $Z_2 = \begin{pmatrix} Z_{211} \\ Z_{212} \\ Z_{22} \end{pmatrix} = \begin{pmatrix} Z_{211} \\ Z_{22}^* \end{pmatrix}$, where $Z_{22}^* = \begin{pmatrix} Z_{22} \end{pmatrix}$.

After these transformations, we can rewrite the right hand side of (8.2) as

$$\int \tilde{\varphi}(y_1, y_{211}, y_{212}, y_{22}, v, \rho_1, \rho_2, \sigma_1^2, \sigma_2^2, \sigma_v^2, \rho_e) \, dy_2^* \, d\rho_2 \, d\sigma_2^2$$

We integrate with respect to $y_2^*$, $\rho_2$ and $\sigma_2^2$ respectively, to obtain

$$\int \tilde{\varphi}(y_1, y_{211}, y_{212}, y_{22}, v, \rho_1, \rho_2, \sigma_1^2, \sigma_2^2, \sigma_v^2, \rho_e) \, dy_2^* \, d\rho_2 \, d\sigma_2^2 = C \times \frac{1}{(\sigma_1^2)^{\frac{m_1}{2}}} \exp \left( -\frac{1}{2\sigma_1^2} (y_1 - M_1 \rho_1 - Z_1 v)^T (y_1 - M_1 \rho_1 - Z_1 v) \right)$$

After these transformations, we can rewrite the right hand side of (8.2) as

$$\int \tilde{\varphi}(y_1, y_{211}, y_{212}, y_{22}, v, \rho_1, \rho_2, \sigma_1^2, \sigma_2^2, \sigma_v^2, \rho_e) \, dy_2^* \, d\rho_2 \, d\sigma_2^2 = \tilde{\varphi}(y_1, y_{211}, y_{212}, y_{22}, v^*, \rho_1, \rho_2, \sigma_1^2, \sigma_2^2, \sigma_v^2, \rho_e)$$

We integrate with respect to $y_2^*$, $\rho_2$ and $\sigma_2^2$ respectively, to obtain

$$\int \tilde{\varphi}(y_1, y_{211}, y_{212}, y_{22}, v, \rho_1, \rho_2, \sigma_1^2, \sigma_2^2, \sigma_v^2, \rho_e) \, dy_2^* \, d\rho_2 \, d\sigma_2^2 = C \times \frac{1}{(\sigma_1^2)^{\frac{m_1}{2}}} \exp \left( -\frac{1}{2\sigma_1^2} (y_1 - M_1 \rho_1 - Z_1 v)^T (y_1 - M_1 \rho_1 - Z_1 v) \right)$$

As we mentioned earlier, there are $m_1$ components of $v$ involved in $Z_1 v$. We write those $m_1$ components as $v^{(1)} = (v_{i_1}, \ldots, v_{i_{m_1}})^T$. Then, $Z_1 v$ reduces to $Z_1^{(1)} v^{(1)}$, where $Z_1^{(1)} = \bigoplus_{j=1}^{m_1} 1_{n_{i,j}}$. Clearly, rank($Z_1^{(1)}$) = $m_1$ and $n_1^* = \sum_{j=1}^{m_1} n_{i,j}$. We integrate out $v^{(2)} = \{v_l :
Proof of Lemma 8.5 is discussed in Section 8.3. We have, rank
\[ R = \text{rank}(I - M_1^T M_1)^{-1} M_1^T. \]

Before we proceed, let us state the following lemma.

**Lemma 8.5** The following results hold: (a) rank \([R_1 Z_1^{(1)}] = \text{rank}(M_1 Z_1^{(1)}) - \text{rank}(M_1),\)
(b) rank \((R_2) = n_1^* - \text{rank}(M_1 Z_1^{(1)}), \)
where \(R_2 = I - \text{Proj} \left( M_1 Z_1^{(1)} \right).\)

Proof of Lemma 8.5 is discussed in Section 8.3. We have, rank \(M_1 Z_1^{(1)} = \text{rank}(M_1) + \text{rank}(Z_1^{(1)}) - 1 = p_1 + m_1 - 1.\)
Hence, we have rank \((R_2) = n_1^* - \text{rank}(M_1 Z_1^{(1)}) = n_1^* - (p_1 + m_1 - 1) = (n_1^* - p_1 - m_1) + 1 \geq 1\) (by (a) of Lemma 8.4). Thus \(R_2\) is positive-semidefinite and \(y_1^TR_2y_1 > 0\) with probability 1.

Let \(Q_1 = Z_1^{(1)T} R_1 Z_1^{(1)}.\) Since \(R_1\) is symmetric and idempotent, rank \((Q_1) = \text{rank}[R_1 Z_1^{(1)}] = \text{rank}(M_1 Z_1^{(1)}) - \text{rank}(M_1) = m_1 + p_1 - 1 - p_1 = m_1 - 1 = t_1\) (say). Let \(P_1\) be
an orthogonal matrix such that $P_1^TQ_1P_1 = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{t_1}, 0, \ldots, 0)$, where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_{t_1} > 0$ are the positive eigenvalues of $Q_1$.

Let $\hat{v}^{(1)}$ denote a minimizer of $(y_1 - Z_1^{(1)}v^{(1)})^T R_1 (y_1 - Z_1^{(1)}v^{(1)})$ wrt $v^{(1)}$. We use the transformation $w = P_1 v^{(1)}$ in (8.6).

\[
\int \tilde{\varphi}(y_1, y_{211}, y_2^*, v, \rho_1, \rho_2, \sigma_1^2, \sigma_2^2, v_0) \, dy_2^* \, d\rho_2 \, d\sigma_2^2 \, d\nu^{(2)} \, d\rho_1 \\
= C \times \int \frac{1}{\sigma_1^2} \exp \left\{ -\frac{w^T w}{2\sigma_1^2} \right\} \, dy_2^* \, d\rho_2 \, d\sigma_2^2 \, d\nu^{(2)} \, d\rho_1 \times \exp \left\{ -\frac{\sum_{j=1}^{t_1} \lambda_j (w_j - \hat{w}_j)^2}{2\sigma_1^2} \right\},
\]
(8.7)

where $P_1 \hat{v}^{(1)} = \hat{w}$. We integrate out $w_{t_1+1}, \ldots, w_{m_1}$:

\[
\int \tilde{\varphi}(y_1, y_{211}, y_2^*, v, \rho_1, \rho_2, \sigma_1^2, \sigma_2^2, v_0) \, dy_2^* \, d\rho_2 \, d\sigma_2^2 \, d\nu^{(2)} \, d\rho_1 \prod_{k=t_1+1}^{m_1} d\sigma_k \\
= C \times \int \frac{1}{\sigma_1^2} \exp \left\{ -\frac{y_1^T R_2 y_1}{2\sigma_1^2} \right\} \, dy_2^* \, d\rho_2 \, d\sigma_2^2 \, d\nu^{(2)} \, d\rho_1 \prod_{k=t_1+1}^{m_1} d\sigma_k \times \exp \left\{ -\frac{\sum_{j=1}^{t_1} w_j^2}{2\sigma_v^2} \right\}.
\]

We integrate the last equation with respect to $\sigma_1^2$ and $\sigma_v^2$ using inverse gamma density integration result. By the conditions $n^*_1 > p_1 + m_1$ and $m_1 > 3$, the shape parameters will be positive. Hence after substantial simplifications,

\[
\int \tilde{\varphi}(y_1, y_{211}, y_2^*, v, \rho_1, \rho_2, \sigma_1^2, \sigma_2^2, v_0) \, dy_2^* \, d\rho_2 \, d\sigma_2^2 \, d\nu^{(2)} \, d\rho_1 \prod_{k=t_1+1}^{m_1} d\sigma_k \times \frac{1}{\left( \sum_{j=1}^{t_1} w_j^2 \right)^{\frac{t_1-2}{2}}}.
\]
(8.8)
Let us denote \( \sum_{j=1}^{t_1} w_j^2 \) by \( d^2 \), then for any \( \epsilon > 0 \), and positive constants \( C_1 \) and \( C_2 \),

\[
\int \tilde{\varphi}(y_1, y_{211}, y_{22}^*, v, \rho_1, \rho_2, \sigma_1^2, \sigma_2^2, \sigma_v^2, p_e) \, dy_2^* \, d\rho_2 \, d\sigma_2^2 \, dv^{(2)} \, d\rho_1 \prod_{k=t_1+1}^{m_1} dw_k \, d\sigma_1^2 \, d\sigma_v^2 \\
\leq C_1 \times \left\{ \frac{1}{\{y_1^T R_2 y_1\}^{e_1+e_2}} \times \frac{1}{(\sum_{j=1}^{t_1} w_j^2)^{e_1+e_2}} \right\} \times \frac{I\left(\sum_{j=1}^{t_1} w_j^2 \leq 2d^2 + \epsilon\right)}{I\left(\sum_{j=1}^{t_1} w_j^2 \leq \lambda_{t_1}^2 \left[ \frac{1}{2} \sum_{j=1}^{t_1} w_j^2 - d^2 \right] \right)}
\]

\[
+ C_2 \times \frac{1}{\lambda_{t_1}^{(e_1+e_2)}} \left\{ \frac{1}{\left[ \frac{1}{2} \sum_{j=1}^{t_1} w_j^2 - d^2 \right]^{e_1+e_2}} \right\} \times \frac{I\left(\sum_{j=1}^{t_1} w_j^2 > 2d^2 + \epsilon\right)}{I\left(\sum_{j=1}^{t_1} w_j^2 \right)}.
\] (8.9)

Using the polar transformation for \( w_1, \ldots, w_{t_1} \), it follows after substantial simplifications, the integrability of the rhs of (8.9) follows.

So far we have proved that any arbitrary typical term in (8.1) satisfying conditions (a), (b) and (c) is integrable. Hence, we can conclude, \( f(y, v, \beta, \sigma_1^2, \sigma_2^2, \sigma_v^2, p_e) \) (in (8.1)) is integrable with respect to \( v, \beta, \sigma_1^2, \sigma_2^2, \sigma_v^2, p_e \) if condition (a), (b) and (c) are satisfied. \( \square \)

### 8.3 Proof of the Lemmas

**Proof of Lemma 8.4:**

**Proof** At first we note that at least one of these two conditions \( n_1^* \geq m_1 + p_1 \) and \( n_2^* \geq m_2 + p_2 \) holds. In order to establish that, let us assume, \( n_1^* \leq m_1 + p_1 - 1 \) and \( n_2^* \leq m_2 + p_2 - 1 \), i.e., \( n = n_1^* + n_2^* \leq m_1 + p_1 + m_2 + p_2 - 2 < 2m + 2p - 1 \), which contradicts to our assumption that \( n = 2m + 2p - 1 \).

Note that, \( n_1^* = \sum_{i \in S_1} n_{i} \leq 2m_1 + 2p - 1 \). If possible, let \( n_1^* > 2m_1 + 2p - 1 \), that is, \( m_1 \) small areas have more than \( 2m_1 + 2p - 1 \) observations. Since we previously assumed that (in Theorem 8.1) \( n_i \geq 2 \), for all \( i \). Hence the remaining \( (m - m_1) \) small areas have at least \( 2(m - m_1) \) observations overall. Therefore, \( n = n_1^* + n_2^* > 2m_1 + 2p - 1 + 2(m - m_1) = 2m + 2p - 1 \), which is a contradiction to the previous assumption that \( n = 2m + 2p - 1 \).

With similar arguments we can establish \( n_2^* \leq 2m_2 + 2p - 1 \).

Now we consider various scenarios and prove that either (a) or (b) holds.

**Case - I:** \( n_1^* \geq m_1 + p_1 \) and \( m_1 \leq 3 \).
We know $n_1^* \leq 2m_1 + 2p - 1$. Hence, $n_1^* \leq 6 + 2p - 1 = 2p + 5$. Also, $n_2^* = n - n_1^* \geq n - (2p + 5) = 2m - 6 \geq m + p \geq m_2 + p_2$.

Now, let us assume, $m_2 \leq 3$, $n_2^* \leq 2m_2 + 2p - 1 \Rightarrow n_2^* \leq 2p + 6 \leq p - 1 + m < m + p$, which contradicts to our earlier assertion $n_2^* \geq m + p$. Hence $m_2 > 3$. Therefore, $n_2^* \geq m_2 + p_2$, $m_2 > 3$; i.e. condition (b) holds.

Case - II: $n_2^* \geq m_2 + p_2$ and $m_2 \leq 3$. With the similar arguments, as in Case - I, it can be shown that condition (a) holds in this case.

Case - III: $n_2^* < m_2 + p_2$, $m_2 > 3$ or $m_2 \leq 3$. In this case, $n_1^* = n - n_2^* > 2m + 2p - 1 - (m_2 + p_2) \geq 2m + 2p - m - p - 1 \geq m_1 + p_1 - 1$. Hence, $n_2^* < m_2 + p_2 \Rightarrow n_1^* \geq m_1 + p_1$. Again, let us assume, $m_1 \leq 3$. Now, $n_1^* \leq 2m_1 + 2p - 1 \leq 2p + 5 \Rightarrow n = n_1^* + n_2^* \leq (2p + 5) + (m_2 + p_2 - 1) \leq 2p + 5 + m + p - 1 = 3p + m + 4 \Rightarrow n = 2m + 2p - 1 \leq 3p + m + 4 \iff m \leq p + 5$, which contradicts to our earlier assumption that $m \geq p + 6$. Therefore $m_1 \geq 3$ in this case, i.e. $m_1 > 3$. Hence, $n_2^* < m_2 + p_2 \Rightarrow n_1^* \geq m_1 + p_1$ and $m_1 > 3$, i.e., condition (a) holds.

Case - IV: $n_1^* < m_1 + p_1$, $m_1 > 3$ or $m_1 \leq 3$. It can be proved that condition (2) will hold in this scenario. Hence, under the proposed model at least one of the conditions stated will hold.

\[ \square \]

**Proof of Lemma 8.5**

**Proof** (a) Here, $R_1 = I - M_1(M_1^T M_1)^{-1} M_1^T \Rightarrow R_1 Z^{(1)} = Z^{(1)} - M_1(M_1^T M_1)^{-1} M_1^T Z^{(1)} \Rightarrow M_1^T (R_1 Z^{(1)}) = 0 \Rightarrow$ columns of $M_1$ are orthogonal to the columns of $R_1 Z^{(1)}$. Therefore,

\[ \text{rank}(R_1 Z^{(1)} M_1) = \text{rank}(R_1 Z^{(1)}) + \text{rank}(M_1). \]

Now,

\[ \begin{pmatrix} R_1 Z^{(1)} & M_1 \end{pmatrix} = \begin{pmatrix} \text{I} \\ -(M_1^T M_1)^{-1} M_1^T Z^{(1)} \text{I} \end{pmatrix}. \]

\[ \Rightarrow \text{rank}(R_1 Z^{(1)} M_1) = \text{rank}(Z^{(1)} M_1) \text{ (since } \begin{pmatrix} \text{I} \\ -(M_1^T M_1)^{-1} M_1^T Z^{(1)} \text{I} \end{pmatrix} \text{ is non singular) } \Rightarrow \text{rank}(R_1 Z^{(1)}) = \text{rank}(Z^{(1)} M_1) - \text{rank}(M_1) = \text{rank}(M_1 Z^{(1)}) - \text{rank}(M_1). \]
(b) $R_2 = R_1 - R_1Z_1^{(1)}(Z_1^{(1)T}R_1Z_1^{(1)}) - Z_1^{(1)T}R_1$, $R_2$ is idempotent.

Therefore, $\text{rank}(R_2) = \text{rank}(R_1) - \text{rank}(R_1Z_1^{(1)}) = (n_1^* - \text{rank}(M_1)) - \text{rank}\begin{pmatrix} M_1 & Z_1^{(1)} \end{pmatrix}$

$- \text{rank}(M_1) = n_1^* - \text{rank}\begin{pmatrix} Z_1^{(1)} & M_1 \end{pmatrix} = n_1^* - \text{rank}\begin{pmatrix} M_1 & Z_1^{(1)} \end{pmatrix}$. \hfill \Box