A two-component normal mixture alternative to the Fay-Herriot model

August 22, 2015

Adrijo Chakraborty [1], **Gauri Sankar Datta** [2] [3] [4] **and Abhyuday Mandal** [5]

Abstract: This article considers a robust hierarchical Bayesian approach to deal with random effects of small area means when some of these effects assume extreme values, resulting in outliers. In presence of outliers, the standard Fay-Herriot model, used for modeling area-level data, under normality assumptions of the random effects may overestimate random effects variance, thus provides less than ideal shrinkage towards the synthetic regression predictions and inhibits borrowing information. Even a small number of substantive outliers of random effects result in a large estimate of the random effects variance in the Fay-Herriot model, thereby achieving little shrinkage to the synthetic part of the model or little reduction in posterior variance associated with the regular Bayes estimator for any of the small areas. While a scale mixture of normal distributions with known mixing distribution for the random effects has been found to be effective in presence of outliers, the solution depends on the mixing distribution. As a possible alternative solution to the problem, a two-component normal mixture model has been proposed based on noninformative priors on the model variance parameters, regression coefficients and the mixing probability. Data analysis and simulation studies based on real, simulated and synthetic data show advantage of the proposed method over the standard Bayesian Fay-Herriot solution derived under normality of random effects.

[1]NORC at the University of Chicago, Bethesda, MD 20814, chakraborty-adrijo@norc.org

[2]Department of Statistics, University of Georgia, Athens, GA 30602, USA. email:gauri@stat.uga.edu

[3]Center for Statistical Research and Methodology, US Census Bureau, Washington, D.C. 20233

[4]Disclaimer: This report is released to inform interested parties of research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the US Census Bureau

[5]Department of Statistics, University of Georgia, Athens, GA 30602, USA. email:amandal@stat.uga.edu

1 Introduction

Small area estimation methods are getting increasingly popular among survey practitioners. Reliable small area estimates are often solicited by the policy makers from both government and private sectors for planning, marketing and decision making. In order to support growing demand of reliable small area estimates, researchers have developed methods that combine information from the small areas and other related variables. Ghosh and Rao (1994), Rao (2003), Jiang and Lahiri (2006), Datta (2009) and Pfeffermann (2013) provided a comprehensive review of the research in small area estimation.

The landmark paper by Fay and Herriot (1979) used the EB (empirical Bayes) approach (see, for example, Efron and Morris, 1973) and popularized model-based small area estimation methods. Denoting the design-based direct survey estimator of the ith small area by Y_i and its auxiliary variable by x_i, an $r \times 1$ vector, Fay and Herriot (1979) introduced the model

$$Y_i = \theta_i + e_i, \quad \theta_i = x_i^T \beta + v_i, \quad i = 1, \ldots, m. \tag{1.1}$$

Here θ_i is a summary measure of the characteristic to be estimated for the ith small area, e_i is the sampling error of the estimator Y_i, and the random effects v_i denotes the model error measuring the departure of θ_i from its linear regression on x_i. It is assumed that $e_1, \ldots, e_m$ are independent and normally distributed with $e_i \sim N(0, D_i)$, and are independent of $v_1, \ldots, v_m$, which are i.i.d. $N(0, A)$. The sampling variances D_i's are treated as known, but the model parameters β and A are unknown. Random effects v_i's are also known as small area effects.

In this paper we focus on hierarchical Bayes (HB) methods for area-level models. The classical area-level Fay-Herriot model was primarily developed as a frequentist model, which was later given a Bayesian formulation (Rao 2003; Datta et al. 2005). Estimators obtained from Fay-Herriot model are shrinkage estimators, i.e., an weighted average of the direct estimator and the model-based synthetic estimator, these weights depend on the model assumption. Datta and Ghosh (2012) gave an extensive review of shrinkage estimation in small area estimation context. Shrinkage estimators are primarily constructed to improve the standard estimators. Datta and Lahiri (1995) discussed how outliers can affect shrinkage estimators, even a single outlier may lead all the small area estimates to collapse to their corresponding direct estimates. This phenomenon was also mentioned in the context of estimation of multiple normal means under the assumption of an exchangeable normal prior (cf. Efron and Morris 1971, Stein 1981, and Angers and Berger 1991). One or more substantive outliers considerably inflate a standard estimator of model variance.

Overestimation of model variance due to one or more substantive outliers practically results in no shrinkage of any of the direct estimates of the small area means to a synthetic regression estimator. This would also limit the reduction of the posterior variances of the model-based estimates. To rectify this problem, following the work of Angers and Berger (1991), who used a Cauchy distribution for the small area means θ_i, Datta and Lahiri (1995) recommended a broader class of heavy-tailed distributions through scale mixture of normal distributions. They showed that under these assumptions, in presence of substantive outliers, estimators corresponding to the outlying areas converge to their corresponding direct estimators but leave the non-outlying areas less affected. One difficulty with the last method is that the mixing distribution for the scale parameter is considered to be known. For example, one can use t-distribution for random effects as in Xie et al. (2007). However, in absence of any information regarding the degrees of freedom one needs to specify a prior. Xie et al. (2007) assumed a gamma prior for the degrees of freedom. The hyperparameters involved in this gamma distribution need to be specified. Bell and Huang (2006) argued that under practical circumstances limited information is obtained from the data regarding degrees of freedom and instead they used several fixed values for the degrees of freedom.

In order to avoid specifying the mixing distribution, in this paper we propose a two-component normal mixture distribution for the random small area effects. This allows that outlying areas to come from a distribution with the bigger variance. It is a simple extension of the Fay-Herriot model with a contaminated random effects distribution with possibly a small proportion of areas having a larger model variance. Contaminated models have been extensively used in empirical evaluations of robust EBLUP approach of Sinha and Rao (2009). We consider an HB approach by assigning non-subjective priors to the parameters involved in the model. Some components of these priors are improper, hence we provide sufficient conditions for the posterior distribution to be proper.

In a recent article, Datta et al. (2011) demonstrated that in the presence of good covariates x_i, the variability of the small area means θ_i may be accounted well by x_i, and including a random effects v_i in the model (1.1) may be unnecessary. These authors test a null hypothesis of no random effects in the small area model and if it is not rejected, they propose more accurate synthetic estimators for the small area means. In a more recent article, Datta and Mandal (2015) argued that even if the null hypothesis is rejected in this case, it is reasonable to expect only a small fraction of small areas means will not be adequately explained by covariates, and only these areas would require a random component to the regression model.

In their HB approach, Datta and Mandal (2015) considered a “spike and slab” distribution for the random small area effects in order to propose a flexible balance between Fay and

Herriot (1979) and Datta et al. (2011) models. However, often it is difficult to find reliable covariates that would describe the response well, particularly, if the number of small areas is large. For such datasets, not only the test proposed by Datta et al. (2011) would suggest an inclusion of small area effects, the model proposed by Datta and Mandal (2015) would also estimate the probability of existence of random effects to be as very high. This effectively would suggest the Fay-Herriot model, but in reality, only a small proportion of small areas may not be adequately explained by a model with one single A. This would result in overestimation of A, thereby resulting in a poor fit, particularly when the number of small areas m is large. Even if most of the small areas would require a random effects term in the regression model, it is more likely that only a small proportion of small areas would need a bigger value of A, and a smaller value of the same is sufficient for the other areas. In this paper, we assume that $v_1, \ldots, v_m$ are independently distributed with mean 0 and a two-component mixture of normal distributions with variance either A_1 or $A_2 (> A_1)$. This model is potentially useful to handle large outliers in small area means.

Bell and Huang (2006) presented an insightful discussion about use of t-distribution with known d.f. to handle outliers in Fay-Herriot model. The theoretical regression residuals from (1.1) consist of the sum of the sampling error and model error, which are not individually observable. They argued that residual may be an outlier either due to the sampling error or the model error. They explained that the consequences of these two types of outliers are quite different. If the model error v_i is an outlier for some area, then the regression model (or synthetic estimation) is not good for the area. In that case, the direct estimator Y_i should be used as the small area estimator. Datta and Lahiri (1995) considered this case using a scale mixture of normal distribution. An alternative to this approach is proposed in the present article through a two-component normal mixture.

On the other hand, if the sampling error e_i is an outlier due to an underestimation of the variance D_i, then the direct estimator Y_i is not reliable; Bell and Huang (2006) argued that the "synthetic estimator" $x_i^T \beta$ may be used for prediction. To address this issue, they proposed a t-distribution for the sampling distribution. They noted that it is difficult to distinguish between the two scenarios of a sampling error outlier or a model error outlier since the data in fitting the model (1.1) cannot readily disentangle the two cases. For further discussion, we refer to this article.

This paper is organized as follows. In Section 2 we describe the proposed model and discuss some properties of our new shrinkage estimators. In Section 3 we illustrate our method to estimate U.S. poverty rates for 3141 counties based on 5-year estimates from the American

Community Survey. Performance of the model is discussed in comparison with the traditional Fay-Herriot model in Section 4. Section 6 provides a concluding discussion. A detailed proof of the propriety of the posterior distribution is relegated to the Appendix.

2 Two-component normal mixture model

Fay and Herriot (1979) proposed a model which has been extensively used in many small area estimation applications to provide reliable estimates of poverty and income measures. While for regular data the model successfully produces accurate shrinkage estimators of small area means, it breaks down in the presence of substantial outliers among the small area means. In order to account for the outliers, we consider a two-component normal mixture extension of Fay-Herriot model. This model is given by

$$y_i = \theta_i + e_i, \quad \theta_i = x_i^T \beta + (1-\delta_i)v_{1i} + \delta_i v_{2i}, \ i = 1, \dots, m, \tag{2.1}$$

where e_i, δ_i, v_{1i}, v_{2i} are independently distributed with $P(\delta_i = 1|p) = 1 - p$, $v_{1i} \sim N(0, A_1)$ and $v_{2i} \sim N(0, A_2)$. As in (1.1), β is an $r \times 1$ vector of regression parameters, and the sampling errors $e_1, \dots, e_m$ are independently normally distributed. To complete our HB structure, we consider the following class of priors,

$$\pi(\beta, A_1, A_2, p) = \pi^*(A_1, A_2) \propto A_1^{-\alpha_1} A_2^{-\alpha_2} I(0 < A_1 < A_2 < \infty). \tag{2.2}$$

We use a uniform prior on the regression parameter β and the mixing proportion p. For the prior on the variance parameters, we choose $\alpha_1 < 1 < \alpha_2$ suitably, and we discuss permissible choices of the values of α_1 and α_2 later. We impose the restriction $A_1 < A_2$, so that we do not have a label switching problem leading to non-identifiability. The area specific random effects corresponding to the outlying areas in the model are assumed to follow a normal distribution with a larger variance, which remains the motivation behind imposing such a restriction. While for the parameter β common in all components of mixture an improper uniform prior is reasonable, the prior for A_1 and A_2, which are not common in all components of the mixing distributions, is required to be at least *partially proper*. By partially proper we mean that while the marginals are improper, conditional priors for A_2 given A_1, and A_1 given A_2 are proper. For this to hold for our class of priors for A_1, A_2, it is necessary and sufficient that $\alpha_1 < 1 < \alpha_2$. A partially proper prior is required for the parameters that are not common to all components of a Bayesian mixture model (cf. Scott and Berger, 2006).

Since the Bayesian model involves improper priors, in Theorem 2.1 below we provide sufficient conditions that ensure the resulting posterior distribution from the proposed model will be proper. A detailed proof of Theorem 2.1 is given in Section 6.

Theorem 2.1 *The resulting posterior distribution from model (2.1) and the prior in (2.2) will be proper if (a)* $m > r + 2(2 - \alpha_1 - \alpha_2)$ *and (b)* $2 - \alpha_1 - \alpha_2 > 0$.

The sufficient conditions in Theorem 2.1 provide a set of permissible values for α_1 and α_2. In conjunction with the condition $2 - \alpha_1 - \alpha_2 > 0$, the condition $\alpha_2 > 1$ implies $\alpha_1 < 1$. We noted earlier that the last two conditions are necessary to elicit partially proper priors. The special case $\alpha_1 = 0$ is feasible, which corresponds to a uniform prior, provided $1 < \alpha_2 < 2$. However, it is not possible to assign a uniform prior on A_2. If $\alpha_1 = \frac{1}{2}$, then $1 < \alpha_2 < \frac{3}{2}$. Also, for mixture models, Jeffreys' prior has no closed-form expression to work with.

Our choice of prior for the mixing parameter p is Uniform(0,1). We can modify this prior if subjective information is available. If past experience in an application suggests any information regarding the proportion of outlying areas, that can be incorporated in the model by modifying the prior for p. Sufficient conditions for the propriety of the posterior density will remain unchanged. For instance, if the model is modified with the assumption that p follows a known *Beta* distribution, the sufficient conditions provided in Theorem 2.1 will remain intact.

It is well-known that even a single substantial outlier will collapse shrinkage estimators of all θ_i's based on the model (1.1) to the direct estimators y_i's (see Dey and Berger, 1983; Stein, 1981). As a result, model-based estimators will fail to borrow strength from the other small areas. To protect against this odd behavior, Angers and Berger (1991) and Datta and Lahiri (1995) suggested a robust shrinkage model. These authors used suitable scale mixture of normal distributions to model long-tail distribution of the θ's. These methods assume the knowledge of the scale mixing distribution, which may not be available. The purpose of our proposed mixture model in (2.1) is to provide an alternative solution that does not require the knowledge of the mixing distribution and to facilitate borrowing information among non-outlying observations in the presence of some substantive outliers.

We discuss below a heuristic comparison of the the shrinkage property of the Bayes estimators of θ_i under the Fay-Herriot model and our proposed model, in presence of substantial outliers. For Fay-Herriot model, given the values of the parameters β and A, estimator of θ_i is

$$\theta_i^{FH} = y_i - \frac{D_i}{D_i + A}(y_i - x_i^T \beta), \; i = 1, \ldots, m. \tag{2.3}$$

In presence of outliers, frequentist estimators of A will be large, and the posterior density of A will have a long right tail, which will also result in a large Bayesian estimator of A. Consequently, an estimate of the shrinkage coefficient $D_i/(D_i + A)$ will be rather small, and the Bayes or the EB estimator of θ_i will borrow little from its synthetic regression prediction and it will collapse to direct estimator y_i for all i.

We now argue that the proposed mixture model is more flexible to retain shrinkage of the non-outlying observations in presence of outliers. Again, for known model parameters β, A_1, A_2 and p, the conditional mean $E(\theta_i|\beta, A_1, A_2, p, y) = \theta_i^{Mix}$(say). Using iterated expectation $E(\theta_i|\beta, A_1, A_2, p, y) = E[E(\theta_i|\beta, A_1, A_2, \delta_i, p, y)|\beta, A_1, A_2, p, y]$, and noting that $E(\theta_i|\beta, A_1, A_2, \delta_i, p, y) = \dfrac{D_i x_i^T\beta + A_{1+\delta_i} y_i}{D_i + A_{1+\delta_i}}$, $\tilde{p}_i = P(\delta_i = 0|\beta, A_1, A_2, p, y_i)$, we get

$$\theta_i^{Mix} = y_i - \left[\left(\frac{D_i}{D_i + A_1}\right)\tilde{p}_i + \left(\frac{D_i}{D_i + A_2}\right)(1 - \tilde{p}_i)\right](y_i - x_i^T\beta), \tag{2.4}$$

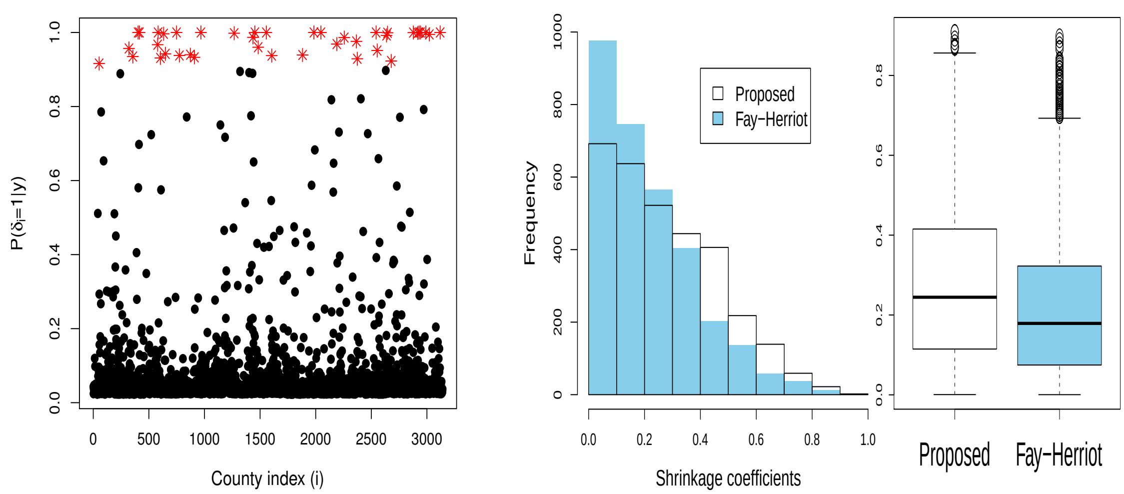
where

$$\tilde{p}_i = \frac{\frac{p}{(D_i+A_1)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\frac{(y_i - x_i^T\beta)^2}{(D_i+A_1)}\right\}}{\frac{p}{(D_i+A_1)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\frac{(y_i - x_i^T\beta)^2}{(D_i+A_1)}\right\} + \frac{(1-p)}{(D_i+A_2)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\frac{(y_i - x_i^T\beta)^2}{(D_i+A_2)}\right\}}, \tag{2.5}$$

for $i = 1, \ldots, m$. In presence of substantially large outliers, $(y_i - x_i^T\beta)^2$ and A_2 are expected to be high, hence $P(\delta_i = 0|\beta, A_1, A_2, p, y_i) \approx 0$. This will result in the second shrinkage term within square brackets in (2.4) to be dominant. However, since the posterior distribution of A_2 has long tail, the shrinkage coefficient associated with the second component will be small and $\theta_i^{Mix} \approx y_i$, i.e., if the i^{th} area is outlying then the small area estimator based on this model will be very close to its direct estimator. On the other hand, for any non-outlying areas $\tilde{p}_i$ will be away from 0, and their shrinkages will be less impacted by the outliers.

3 Data Analysis

We illustrate our proposed methodology by analyzing a real data obtained from the "American Fact Finder" website maintained by US Census Bureau. The data set contains 5-year ACS estimates of overall poverty rates for 3141 counties of United States along with their associated design-based standard errors. The county identifiers are not available due to confidentiality reasons. In order to improve direct design-based estimates, government agencies implement state-of-the-art small area estimation methods to produce model-based estimates using auxiliary data. For poverty estimation, domain level tax data are typically used as auxiliary information. However, tax data are not available for public use owing to legal restrictions. In our analysis we use foodstamp participation rate as our only auxiliary variable (correlation between foodstamp participation rate and overall poverty rate is 0.81). Initially we fit the Fay-Herriot model (1.1) with restricted maximum likelihood method (REML) as well as hierarchical Bayesian (HB) method assuming flat priors for regression and variance parameter. The REML and Bayes estimates of the model parameters are very close: $\hat{\beta}^{REML} = (0.056, 0.634)^T$, $\hat{A}^{REML} = 0.0009$ and $\hat{\beta}^{Bayes} = (0.051, 0.634)^T$, $\hat{A}^{Bayes} = 0.0009$.



Bayes estimates of $P(\delta_i = 1|y)$ Shrinkage coefficients

Figure 1: Analysis of the American Community Survey data

Table 1: HB estimates of model parameters (for the ACS county level poverty rates data)

Parameter	Posterior Mean	Posterior sd	Posterior Quantiles 2.5%	Median	97.5%
β_1	0.0465	0.0013	0.0440	0.0465	0.0491
β_2	0.6605	0.0075	0.6459	0.6607	0.6748
A_1	0.00054	0.00003	0.00049	0.00054	0.00059
A_2	0.00619	0.00103	0.00454	0.00609	0.00854
p	0.0725	0.0237	0.0470	0.0704	0.1037

We apply our proposed method to this data set and report the results in Table 1. Our choices of α_1 and α_2 are 0.3 and 1.3 respectively. We have also performed further analysis with other choices of α_1 and α_2 within the feasible range, but results were not considerably different. From Table 1, we see that the posterior mean of $A_2 (= 0.00619)$ is almost ten times larger than that of $A_1(= 0.00054)$. In addition, the estimate $\hat{p} = 0.07$ indicates that there are about 7% small areas which have much larger area specific variability compared to the most. The outlying areas can be identified by computing the Bayes estimates of posterior probabilities $P(\delta_i = 1|y)$. We plot the estimates of these probabilities for each area in Figure 1. It shows that although most areas have low probabilities of having high random effects, some of them have higher chances of having a large variability of the model error or random small area effects. According to our analysis, approximately 7% (221 out of 3141) small areas have posterior probability $P(\delta_i = 1|y) > 0.15$, and approximately 1.3% (40 out of 3141) small areas have posterior probability $P(\delta_i = 1|y) > 0.9$.

4 Exploration of the shrinkage coefficients

We compare the shrinkage coefficients resulting from our proposed method with those resulting from the standard Fay Herriot model. By simulations we demonstrate that our proposed method usually provides better shrinkage than the Fay-Herriot method in presence of outliers in the data. On the other hand, simulated data from a standard Fay-Herriot model yield shrinkage coefficients based on the proposed model that are very similar to those based on the Fay-Herriot model. These two simulations, presented in Figure 2 essentially show the robustness of the proposed method to outliers.

We mentioned in Section 2 that the proposed method is expected to provide better overall shrinkage than Fay-Herriot method in presence of outliers. In order to demonstrate this property of the model, we conduct the following simulations. We replace the direct estimates of first 10% small areas of the data by simulated values and retain the rest of the data set intact. The purpose is to artificially contaminate the data set. We generate the direct estimates of first 10% small areas from model (1.1). We use the sampling variances of these areas to generate the corresponding sampling errors. We use the estimated regression parameters $\beta = (0.06, 0.6)^T$ and model variance 0.0009 obtained from Fay-Herriot analysis of the original data using the Prasad-Rao method. We use these model parameter values and the values of the auxiliary variables from these 10% small areas to retain the mean structure and variability of the small area means which are nearly similar to the original population. We introduce outliers through use of heavy tail distribution or large model variance for random effects. Random small area effects are generated from (a) $v_i \sim t_1$, (b) $v_i \sim t_2$, (c) $v_i \sim t_3$, with proper scaling for each and (d) $v_i \sim N(0, 5^2 \times a^2)$. Note that t_1 distribution is the Cauchy distribution which does not have a variance (indeed it does not have a mean either). We rescale the draws from t_1, t_2 and t_3 by multiplying by the adjusting factor, $\dfrac{N_{0.75}}{T^{df}_{0.75}}a$, where $N_{0.75}$ and $T^{df}_{0.75}$ are the 75^{th} percentile of $N(0, 1^2)$ and t (for a specified df) respectively. By multiplying this adjusting factor, we intend to match the inter-quartile range of draws from the t-distribution to the inter-quartile range of a $N(0, a^2)$ distribution. Since the Prasad-Rao estimate of the random effects variance based on the original data is 0.0009, we choose $a^2 = 0.0009$ in order to maintain consistency.

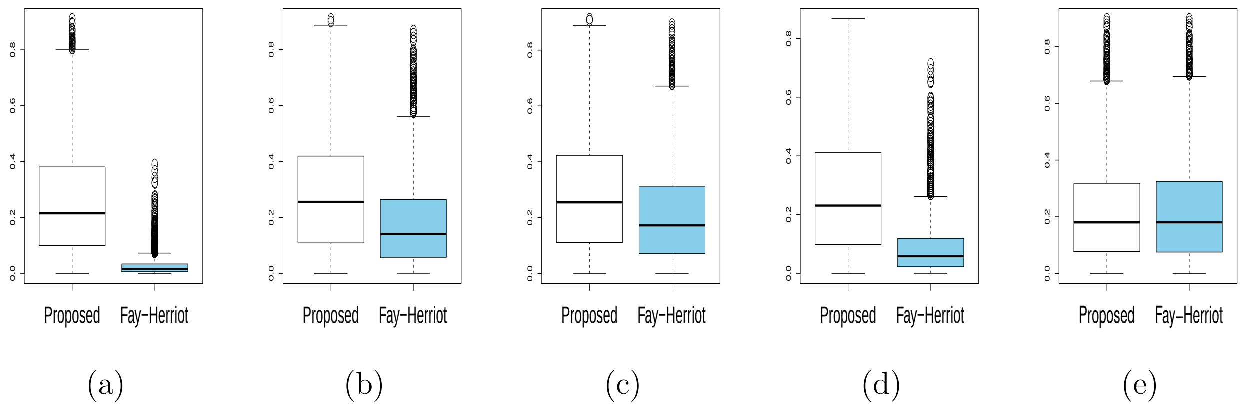


Figure 2: *Boxplots of estimated shrinkage coefficients for two methods. In plots (a)-(d), data are partially simulated for some small areas by drawing random effects from (a) t_1, (b) t_2, (c) t_3, (each of (a)-(c) scale adjusted) and (d) $N(0, 5^2 \times (0.03)^2)$. In the plot (e), we fully simulate the data for* **all** *areas by drawing random effects from $N(0, (0.03)^2)$.*

We apply the proposed method as well as the Fay-Herriot method and compare the estimates of shrinkage coefficients in Figures 2 and 3. We see from Figure 3 that when we partially contaminate the data set using (a) re-scaled t_1 (Cauchy) and (d) $\mathrm{N}(0, 5^2 \times (0.03)^2)$, the overall shrinkage obtained from the proposed model is considerably higher than the overall shrinkage obtained from the regular Fay-Herriot method. This result shows the flexibility of the proposed model in borrowing information from other areas when outliers in the random effects are present. Panels (b), (c) and (e) of Figure 2 show that the proposed method performs similarly as the Fay-Herriot method when the departure of the random effects distribution from the normal is moderate or none.

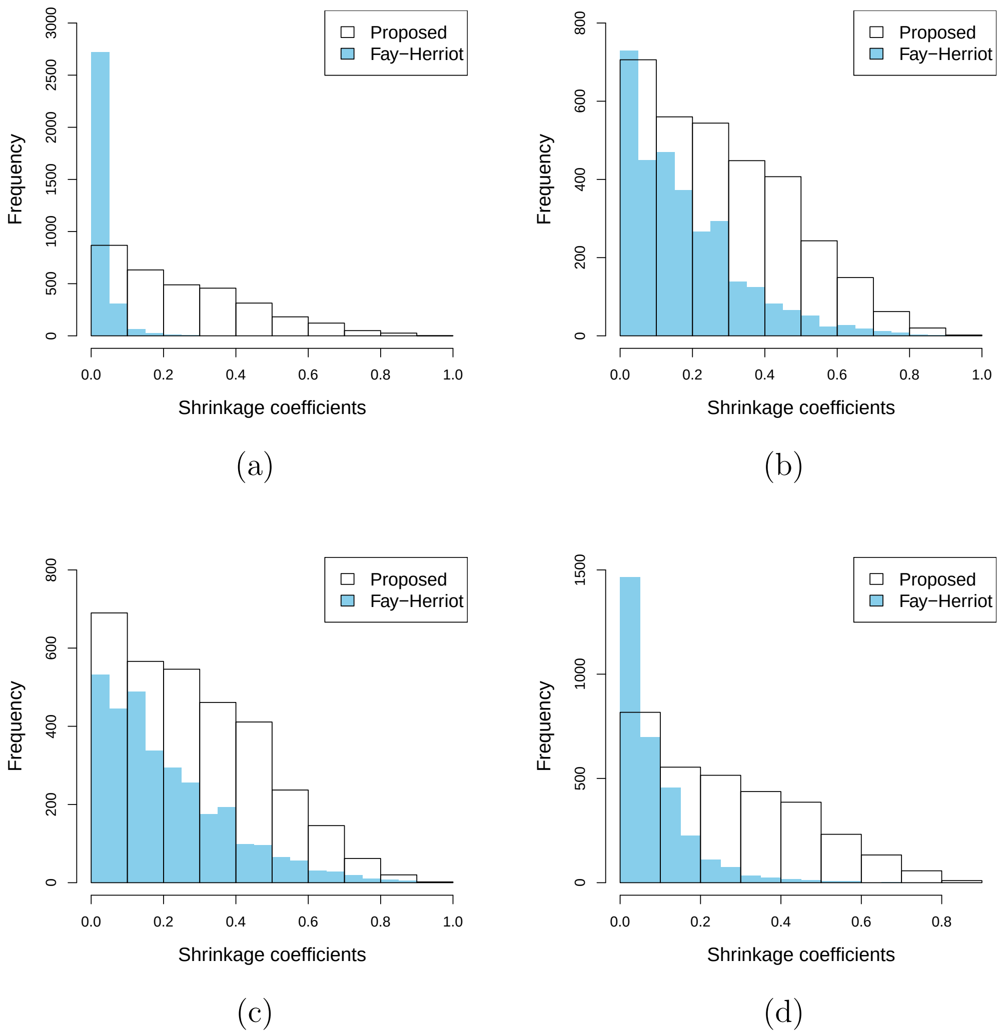


Figure 3: *Histograms of estimated shrinkage coefficients of two methods when the data are partially simulated by drawing random effects from (a)* t_1*, (b)* t_2*, (c)* t_3 *(each of (a)-(c) scale adjusted), and (d)* $N(0, 5^2 \times (0.03)^2)$

5 Performance of the proposed method

In order to evaluate the performance of the proposed model described in Section 2, we conduct a simulation study. This analysis is based on simulated data sets generated under different settings. For each $m = 100$, 500 and 1000, we generated 100 data sets. Here we set $r = 2$, $x = (1, x_1)^T$ and generate m copies of x_1 from $\mathrm{N}(10, (\sqrt{2})^2)$. For each choice of m, the set of covariates is generated exactly once and used the set for all 100 data sets. Our choice of β is $\beta = (20, 1)^T$. The sampling error e_i's are generated from $N(0, D_i)$, $i = 1, \ldots, m$, where D_i's are from the set $\{0.5,\ 1,\ 1.5,\ 2,\ 2.5,\ 3,\ 3.5,\ 4,\ 4.5,\ 5\}$, each value in the set is allocated to the same number of small areas. Random effects in model (1.1) are generated under three different settings,

$$v_i \sim N(0, 1^2), \tag{5.1}$$

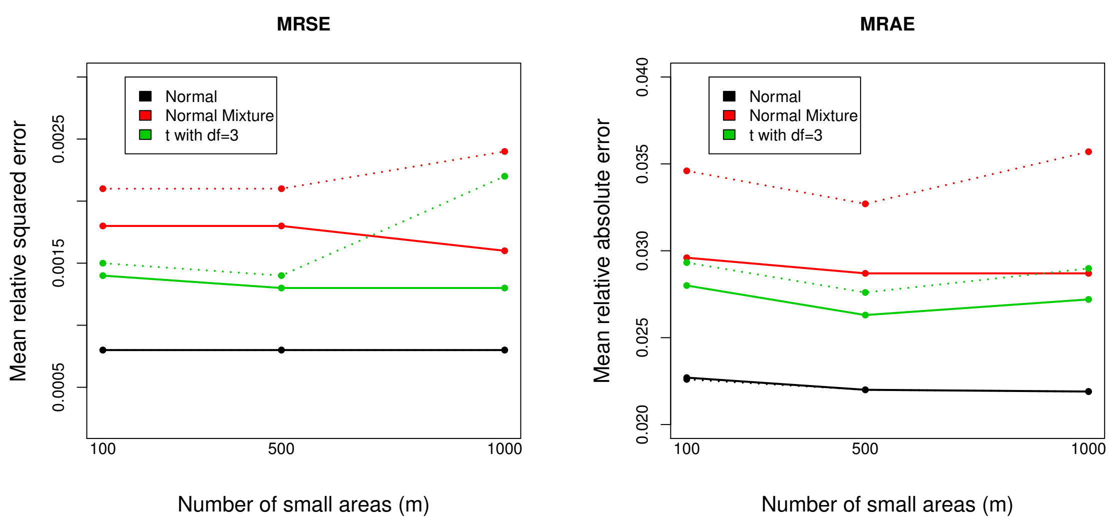
$$v_i \sim (1 - \delta_i)N(0, 1^2) + \delta_i N(0, 5^2), \quad \text{and} \tag{5.2}$$

$$v_i \sim t_3, \tag{5.3}$$

where $i = 1, \ldots, m$. For the normal-mixture setup (5.2), we set $\delta_i = 1$ for each i multiple of 5 and keep rest of the $\delta_i = 0$, the simulated data sets contain 20% observations from the normal distribution with variance 25. Based on a generated set of of v_i's , we compute both the θ_i's and y_i's by (1.1). For each of 100 simulated data sets for each setting, we predict θ_i's based on Fay-Herriot model and the proposed area-level normal mixture model. We measure the performance of each prediction method by computing (empirical) mean squared error (MSE)= $\frac{1}{m}\sum_{i=1}^{m}(\theta_i - \hat{\theta}_i)^2$, mean absolute error (MAE)=$\frac{1}{m}\sum_{i=1}^{m}|\theta_i - \hat{\theta}_i|$, mean relative squared error (MRSE)= $\frac{1}{m}\sum_{i=1}^{m}\frac{(\theta_i - \hat{\theta}_i)^2}{\theta_i^2}$ and mean relative absolute error (MRAE)= $\frac{1}{m}\sum_{i=1}^{m}\frac{|\theta_i - \hat{\theta}_i|}{\theta_i}$, where θ_i's are true and $\hat{\theta}_i$'s are estimated small area means (for our simulation setup, all the θ_i's are positive). These empirical deviation measures are typically used in small area estimation literature to compare the accuracy of various estimation methods (Rao, 2003). For each simulated dataset, we compute MSE, MAE, MRAE and MRSE for two different methods and report the average values based on all simulated data sets. Results of the simulation study are presented in Tables 2 and 3. Table 3 shows a more detailed result when the v_i's are drawn according to equation (5.2). From Table 3 we can compare performance of two prediction methods for outlying areas (random effects drawn from $N(0, 5^2)$) and non-outlying areas (random effects drawn from $N(0, 1^2)$), separately. Simulation results indicate that the proposed method tends to perform better than the Fay-Herriot method when the possibility of presence of outliers is high, and performs similarly otherwise.

Table 2: Comparison of the methods based on simulated MSE and MAE of prediction. Results are based on 100 simulated data sets

Scenario		m=100		m=500		m=1000	
		Proposed	FH	Proposed	FH	Proposed	FH
(5.1) Normal	MSE	0.72	0.71	0.69	0.69	0.68	0.68
	MAE	0.67	0.67	0.66	0.66	0.66	0.65
(5.2) Mixture	MSE	1.48	1.75	1.49	1.81	1.30	1.87
	MAE	0.86	1.01	0.85	0.98	0.84	1.04
(5.3) t_3	MSE	1.14	1.27	1.01	1.20	1.14	1.30
	MAE	0.83	0.84	0.79	0.81	0.80	0.84



(a) (b)

Figure 4: *(a) Mean relative squared error (MRSE) and (b) mean relative absolute error (MRAE) based on* 100 *simulated data sets; Dotted line for Fay-Herriot method and solid line for the proposed method*

Table 3: Comparison of the methods based on simulated MSE, MAE, MRSE and MRAE of prediction. Results are based on 100 simulated data sets. Performance of the methods are compared separately for outlying and non-outlying areas based on the simulation design.

		Scenario (5.2) Mixture					
		m=100		m=500		m=1000	
		Proposed	FH	Proposed	FH	Proposed	FH
MSE	$A_1 = 1^2$	0.90	1.26	0.80	1.06	0.80	1.32
	$A_2 = 5^2$	3.39	3.69	4.25	4.80	3.28	4.03
MAE	$A_1 = 1^2$	0.73	0.88	0.69	0.82	0.70	0.91
	$A_2 = 5^2$	1.43	1.47	1.49	1.61	1.39	1.59
100×MRSE	$A_1 = 1^2$	0.10	0.14	0.09	0.12	0.09	0.15
	$A_2 = 5^2$	0.43	0.50	0.53	0.56	0.44	0.61
10×MRAE	$A_1 = 1^2$	0.25	0.30	0.23	0.27	0.24	0.30
	$A_2 = 5^2$	0.50	0.52	0.51	0.54	0.49	0.57

6 Discussion

In this paper, we propose a robust alternative to Fay-Herriot model. The proposed hierarchical Bayesian estimation procedure is straightforward. Other robust alternative is a t-distribution for the random effects, which requires information regarding the degrees of freedom. Xie et al. (2007) proposed a method to estimate degrees of freedom, however Bell and Huang (2006) pointed out some issues associated with specifying this prior. We propose a method based on noninformative priors for the hyperparameters. We provide sufficient conditions for the propriety of the resulting posterior distributions.

Model-based small area estimates depend on the accuracy of the underlying model assumptions. Larger values of area specific random effects may be caused by poor choice of the linking model or lack of predictive quality of the auxiliary variables. If the model-based estimates of area specific random effects are significantly larger for some areas compared to the other areas, it is probably meaningful to retain the direct estimates instead of model-based estimates for those areas to avoid possible inaccuracy. Although we should be cautious in this recommendation if there is any indication that the sampling variance is underestimated.

Datta and Lahiri (1995) recommended heavy-tailed priors for random effects by emphasizing the fact that estimators obtained by using these priors are similar to direct estimators for the areas with extreme observations. However, the estimators for non-outlying areas should shrink direct estimators more toward synthetic estimators and the magnitude of this shrinkage may depend on the quality of the auxiliary information. While for an outlying observation our model limits shrinkage of Bayes predictor to the synthetic estimator, for non-outlying observations it enables the Bayes predictors to retain shrinkage to the synthetic estimator when the regression model provides a good fit.

Acknowledgment

Authors are thankful to Dr. Jerry Maples of Census Bureau for providing and explaining the poverty data in our application. They are also grateful to him and to Dr. William R. Bell for an internal review of an earlier version of the manuscript. Their many valuable comments, particularly some substantive comments of Dr. Bell, led to a significantly improved manuscript. Research of Mandal is partially supported by the NSA Grant H98230-13-1-0251.

References

1. Angers, J. F., and Berger, J. O. (1991). Robust hierarchical Bayes estimation of exchangeable means. *Canadian Journal of Statistics*, **19**, 39-56.

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

3. Berger, J. O. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness of Bayesian Analysis*, Ed.: J. Kadane, North-Holland, Amsterdam.

4. Datta, G. S. (2009). Model-based approach to small area estimation. *Handbook of Statistics: Sample Surveys: Inference and Analysis*, *29B*, Eds.: D. Pfeffermann and C.R. Rao, The Netherlands: North-Holland. 251-288.

5. Datta, G. S., and Ghosh, M. (2012). Small area shrinkage estimation. *Statistical Science*, **27**, 95-114.

6. Datta, G. S., Hall, P. G., and Mandal, A. (2011). Model selection by testing for the presence of small-area effects in area-level data. *Journal of the American Statistical Association*, **106**, 362-374.

7. 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.

8. Datta, G. S., and Mandal, A. (2015). Small area estimation with uncertain random effects. To appear in *Journal of the American Statistical Association*, **110**, DOI: 10.1080/01621459.2015.1016526.

9. Datta, G. S., Rao, J. N. K., and Smith, D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, **92,** 183-196.

10. Dey, D. K., and Berger, J. O. (1983). On truncation of shrinkage estimators in simultaneous estimation of normal means. *Journal of the American Statistical Association*, **78**, 865-869.

11. Efron, B., and Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators, Part I: The Bayes Case. *Journal of the American Statistical Association*, **67**, 130-139.

12. 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.

13. Ghosh, M., and Rao, J. N. K. (1994). Small are estimation: an appraisal. *Statistical Science*, **9**, 55-93.

14. Jiang, J., and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, **15**, 1-96.

15. Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40-68.

16. Prasad, N. G. N., and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163-171.

17. Rao, J. N. K. (2003). *Small Area Estimation*. Wiley-Interscience, Hoboken, NJ.

18. Rao, J. N. K. (2011). Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal. *Statistical Science*, **26**, 240-256.

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

20. Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1954−1955, **I**, 197-206. Univ. California Press, Berkeley.

21. Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135-1151.

22. Xie, D., Raghunathan, T. E., and Lepkowski, J. M. (2007). Estimation of the proportion of overweight individuals in small areas–a robust extension of the Fay-Herriot model. *Statistics in Medicine*, **26**, 2699-2715.

Appendix

Gibbs sampling for the proposed model

In order to apply our model, we use Gibbs sampling. We derive the set of full conditional distributions from the posterior joint density of $\theta = (\theta_1, \ldots, \theta_m)^T$, $\beta = (\beta_1, \ldots, \beta_r)^T$, $\delta = (\delta_1, \ldots, \delta_m)^T$, A_1, A_2 and p, which is given by

$$\begin{aligned}\pi(\theta, \beta, A_1, A_2, \delta, p|y) \propto & \left\{\prod_{i=1}^m \exp\left\{-\frac{(y_i-\theta_i)^2}{2D_i}\right\}\right\} \prod_{i=1}^m \Bigg[\left\{\frac{1}{\sqrt{A_1}} \times \exp\left\{-\frac{(\theta_i - x_i^T\beta)^2}{2A_1}\right\}\right\}^{\delta_i} \\ & \times \left\{\frac{1}{\sqrt{A_2}} \times \exp\left\{-\frac{(\theta_i - x_i^T\beta)^2}{2A_2}\right\}\right\}^{1-\delta_i} p^{\delta_i}(1-p)^{1-\delta_i}\Bigg] \\ & \times A_1^{-\alpha_1} A_2^{-\alpha_2} \times I(0 < A_1 < A_2). \end{aligned} \tag{6.1}$$

From (6.1), we get the following full conditional distributions:

(I) $\theta_i|\beta, A_1, A_2, \delta, p, y \overset{\text{ind}}{\sim} N\Big(\dfrac{D_i x_i^T\beta + A_{2-\delta_i} y_i}{D_i + A_{2-\delta_i}} \ , \ \dfrac{D_i A_{2-\delta_i}}{D_i + A_{2-\delta_i}}\Big)$, $i = 1, \ldots, m$;

(II) $\beta|\theta, A_1, A_2, \delta, p, y \sim N\Bigg(\left[\sum_{i=1}^m A_{2-\delta_i}^{-1} x_i x_i^T\right]^{-1} \left[\sum_{i=1}^m A_{2-\delta_i}^{-1} x_i \theta_i\right], \left[\sum_{i=1}^m A_{2-\delta_i}^{-1} x_i x_i^T\right]^{-1}\Bigg)$;

(III) $p|\theta, \beta, A_1, A_2, \delta, y \sim Beta\left(\sum_{i=1}^m \delta_i + 1, m - \sum_{i=1}^m \delta_i + 1\right)$;

(IV) $A_1|A_2, \theta, \beta, \delta, p, y$ has the pdf $f_1(A_1)$, where,

$$f_1(A_1) \propto A_1^{-(\alpha_1 + \sum_{i=1}^m \frac{\delta_i}{2})} \exp\left\{-\sum_{i=1}^m \frac{\delta_i(\theta_i - x_i^T\beta)^2}{2A_1}\right\} I(A_1 < A_2),$$

(V) $A_2|A_1, \theta, \beta, \delta, p, y$ has the pdf $f_2(A_2)$, where,

$$f_2(A_2) \propto A_2^{-(\alpha_2 + \sum_{i=1}^m \frac{(1-\delta_i)}{2})} \exp\left\{-\sum_{i=1}^m \frac{(1-\delta_i)(\theta_i - x_i^T\beta)^2}{2A_2}\right\} I(A_1 < A_2),$$

(VI) For $i = 1, \ldots, m,\ \delta_i|\theta, \beta, A_1, A_2, p, y$ are independent with

$$P(\delta_i = 1|\theta, \beta, p, y) = \frac{p \times \exp\left\{-\frac{(\theta_i - x_i^T\beta)^2}{2A_1}\right\} A_1^{-\frac{1}{2}}}{p \times \exp\left\{-\frac{(\theta_i - x_i^T\beta)^2}{2A_1}\right\} A_1^{-\frac{1}{2}} + (1-p) \times \exp\left\{-\frac{(\theta_i - x_i^T\beta)^2}{2A_2}\right\} A_2^{-\frac{1}{2}}}.$$

Our goal is to estimate θ_i, i.e., small area mean for the i^{th} area, $i = 1, \ldots, m$. We implement Gibbs sampling using the conditional distributions (I)$-$(VI) in order to find posterior means and standard deviations of θ_i's. Conditional distribution (IV) and (V) may not have always admit a closed form expression.

Proof of Theorem 2.1

Note that under the proposed mixture model, the likelihood function of the model parameter β, A_1, A_2 and p based on the marginal distribution of $y_1, \ldots, y_m$ is given by

$$L(\beta, A_1, A_2, p) = C \times \prod_{i=1}^{m} \left[\frac{p}{(A_1 + D_i)^{\frac{1}{2}}} e^{-\frac{(y_i - x_i^T\beta)^2}{2(A_1 + D_i)}} + \frac{(1-p)}{(A_2 + D_i)^{\frac{1}{2}}} e^{-\frac{(y_i - x_i^T\beta)^2}{2(A_2 + D_i)}}\right], \quad (6.2)$$

where C is a generic positive constant not depending on the model parameters. Suppose for $0 < a < b < \infty$ we have $a \le D_i \le b$, $i = 1, \ldots, m$. Since $(A_1 + b) \ge (A_1 + D_i) \ge (a/b)(A_1 + b)$, $(A_2 + b) \ge (A_2 + D_i) \ge (a/b)(A_2 + b)$, from (6.2)

$$L(\beta, A_1, A_2, p) \le C \times \prod_{i=1}^{m} \left[\frac{p}{(A_1 + b)^{\frac{1}{2}}} e^{-\frac{(y_i - x_i^T\beta)^2}{2(A_1 + b)}} + \frac{(1-p)}{(A_2 + b)^{\frac{1}{2}}} e^{-\frac{(y_i - x_i^T\beta)^2}{2(A_2 + b)}}\right]. \quad (6.3)$$

For $k = 0, 1, \cdots, m$, let $P_k = \{S_1^{(k)}, S_2^{(k)}\}$ be an arbitrary partition of $\{1, 2, \cdots, m\}$, where $S_1^{(k)}$ has k elements and $S_2^{(k)}$ has $m - k = l$(say) elements. Let $\mathcal{P}_k$ denote all $\binom{m}{k}$ collections of $\{S_1^{(k)}, S_2^{(k)}\}$. Then, expanding the product of the right hand side of (6.3), we get

$$L(\beta, A_1, A_2, p) \le C \times \sum_{k=0}^{m} \sum_{P_k \in \mathcal{P}_k} \frac{p^k (1-p)^{m-k}}{(A_1 + b)^{\frac{k}{2}} (A_2 + b)^{\frac{m-k}{2}}} e^{-\sum_{i \in S_1^{(k)}} \frac{(y_i - x_i^T\beta)^2}{2(A_1 + b)} - \sum_{i \in S_2^{(k)}} \frac{(y_i - x_i^T\beta)^2}{2(A_2 + b)}}. \quad (6.4)$$

To establish propriety of the posterior density, we show integrability of each of the 2^m summands on the right hand side of (6.4) with respect to $\pi(\beta, A_1, A_2, p)$ given in (2.2).

We first consider the case $k = 0$. Here $\mathcal{P}_0$ has one element and $S^{(0)}$ is a null set. Let

$Q(y) = y^T[I - X(X^TX)^{-1}X^T]y$. In this case, the integral $I^{(0)}$ of the term is

$$I^{(0)} = C\int_0^\infty \int_{R^r} \int_0^{A_2} \int_0^1 (1-p)^m dp A_1^{-\alpha_1} dA_1 A_2^{-\alpha_2}(A_2+b)^{-\frac{m}{2}} e^{-\sum_{i=1}^m \frac{(y_i - x_i^T\beta)^2}{2(A_2+b)}} d\beta dA_2$$

$$\begin{aligned} &= C\int_0^\infty \int_{R^r} A_2^{1-\alpha_1-\alpha_2}(A_2+b)^{-\frac{m}{2}} e^{-\sum_{i=1}^m \frac{(y_i - x_i^T\beta)^2}{2(A_2+b)}} d\beta dA_2 \qquad (\text{since } \alpha_1 < 1) \\ &= C\int_0^\infty A_2^{1-\alpha_1-\alpha_2}(A_2+b)^{-\frac{m-r}{2}} e^{-\frac{1}{2}\frac{Q(y)}{A_2+b}} dA_2 \le C\int_0^\infty A_2^{1-\alpha_1-\alpha_2}(A_2+b)^{-\frac{m-r}{2}} dA_2 \\ &< \infty, \qquad \text{if and only if } 2-\alpha_1-\alpha_2 > 0 \text{ and } 1-\alpha_1-\alpha_2-\frac{m-r}{2} < -1, \end{aligned} \tag{6.5}$$

which are equivalent to the conditions outlined in Theorem 2.1.

For the case $k = m$, again there is one term in $\mathcal{P}_m$ and the resulting integral, proceeding as in $I^{(0)}$, is bounded above by

$$\begin{aligned} & C\int_0^\infty A_1^{-\alpha_1}(A_1+b)^{-\frac{m-r}{2}} \int_{A_1}^\infty A_2^{-\alpha_2} dA_2 dA_1 \\ &= C\int_0^\infty A_1^{1-\alpha_1-\alpha_2}(A_1+b)^{-\frac{m-r}{2}} dA_1 \quad (\text{since } \alpha_2 > 1) \\ &< \infty, \quad \text{under the conditions of the theorem.} \end{aligned} \tag{6.6}$$

Now consider a case where $1 \le k \le m-1$. Let $S_1^{(k)}$ be a set of indices $\{i_1, \ldots, i_k\}$ and let $S_2^{(k)} = \{j_1, \ldots, j_l\} = \{1, 2, \cdots, m\} \setminus S_1^{(k)}$. Let us define, $M_1 = (x_{i_1}, \ldots, x_{i_k})^T$ and $M_2 = (x_{j_1}, \ldots, x_{j_l})^T$. Suppose $g = \text{rank}(M_1)$. If $g > 0$, suppose $B \equiv \{\alpha_1, \ldots, \alpha_g\} \subset \{i_1, \ldots, i_k\}$, so that $\{x_{\alpha_1}, \ldots, x_{\alpha_g}\}$ is linearly independent. If $g = 0$, the set B is always empty. Suppose $\{\gamma_1, \ldots, \gamma_{r-g}\} \subset \{j_1, \ldots, j_l\}$ such that $\{x_{\alpha_1}, \ldots, x_{\alpha_g}, x_{\gamma_1}, \ldots, x_{\gamma_{r-g}}\}$ is linearly independent. Let us define the $r \times r$ matrix $F = (x_{\alpha_1}, \ldots, x_{\alpha_g}, x_{\gamma_1}, \ldots, x_{\gamma_{r-g}})^T$, which is non-singular. Consider the non-singular linear transformation of β by $\phi = F\beta$. With these developments, the integral of the term identified by $\{S_1^{(k)}, S_2^{(k)}\}$ in the right hand side of (6.4) with respect to the prior $\pi(\beta, A_1, A_2, p)$ is bounded above by

$$\begin{aligned}
& C\int_0^\infty \int_{A_1}^\infty \int_{R^r} \frac{A_1^{-\alpha_1} A_2^{-\alpha_2}}{(A_1+b)^{\frac{k}{2}}(A_2+b)^{\frac{l}{2}}} e^{-\sum_{i\in S_1^{(k)}} \frac{(y_i - x_i^T\beta)^2}{2(A_1+b)} - \sum_{i\in S_2^{(k)}} \frac{(y_i - x_i^T\beta)^2}{2(A_2+b)}} d\beta dA_2 dA_1 \\
\leq\ & C\int_0^\infty \int_{A_1}^\infty \int_{R^r} \frac{A_1^{-\alpha_1} A_2^{-\alpha_2}}{(A_1+b)^{\frac{k}{2}}(A_2+b)^{\frac{l}{2}}} e^{-\sum_{u=1}^{g} \frac{(y_{\alpha_u} - x_{\alpha_u}^T\beta)^2}{2(A_1+b)} - \sum_{t=1}^{r-g} \frac{(y_{\gamma_t} - x_{\gamma_t}^T\beta)^2}{2(A_2+b)}} d\beta dA_2 dA_1 \\
=\ & C\int_0^\infty \int_{A_1}^\infty \int_{R^r} \frac{A_1^{-\alpha_1} A_2^{-\alpha_2}}{(A_1+b)^{\frac{k}{2}}(A_2+b)^{\frac{l}{2}}} e^{-\sum_{u=1}^{g} \frac{(y_{\alpha_u} - \phi_u)^2}{2(A_1+b)} - \sum_{t=1}^{r-g} \frac{(y_{\gamma_t} - \phi_{g+t})^2}{2(A_2+b)}} d\phi dA_2 dA_1
\end{aligned}$$

$$\begin{aligned}
=\ & C\int_0^\infty \int_{A_1}^\infty \frac{A_1^{-\alpha_1} A_2^{-\alpha_2}}{(A_1+b)^{\frac{k-g}{2}}(A_2+b)^{\frac{l-r+g}{2}}} dA_1 dA_2 \\
\leq\ & C\int_0^\infty \int_{A_1}^\infty \frac{A_1^{-\alpha_1} A_2^{-\alpha_2}}{(A_1+b)^{\frac{k-g}{2}}(A_1+b)^{\frac{l-r+g}{2}}} dA_2 dA_1 \\
=\ & C\int_0^\infty \frac{A_1^{1-\alpha_1-\alpha_2}}{(A_1+b)^{\frac{k-g}{2}}(A_1+b)^{\frac{l-r+g}{2}}} dA_1 \\
=\ & C\int_0^\infty \frac{A_1^{1-\alpha_1-\alpha_2}}{(A_1+b)^{\frac{m-r}{2}}} dA_1 < \infty, \qquad (6.7)
\end{aligned}$$

by the conditions of the theorem. Since the integrability conditions do not depend k or on the indices $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_l\}$ and on the values k and l, the conditions $2 - \alpha_1 - \alpha_2 > 0$ and $m > r + 2(2 - \alpha_1 - \alpha_2)$ will be sufficient to ensure the propriety of the posterior. □