Some Bayesian Contributions to Small Area Estimation

by

Juhyung Lee

(Under the Direction of Gauri Sankar Datta)

Abstract

In a sample survey, a subpopulation is referred to as a "small area" if its sample is not large enough to yield direct estimates of adequate precision. One main interest in small area estimation is estimation of small area means. The observed best prediction (OBP) is a model-based prediction procedure for small area means that has been shown to be more robust than the empirical best linear unbiased prediction (EBLUP) against model misspecifications. We derive a pseudo-Bayesian alternative to the OBP under the Fay-Herriot model by converting the OBP objective function to a likelihood function. Real data examples and simulation studies show that the pseudo-Bayesian estimator (PBE) competes favorably with the OBP. In terms of interval estimation, the PBE credible interval attains the nominal coverage probability, while the OBP confidence interval exhibits unsatisfactory coverage. In addition to the PBE, we propose two compromise pseudo-Bayesian estimators (CPBE) of small area means using regression weights that compromise between those of the EBLUP and OBP. Real data examples show that the CPBEs can outperform the EBLUP, OBP, and PBE in terms of both accuracy and stability of the estimates. Lastly, we consider a problem where direct estimates are available only at a higher level of aggregation instead of the desired lower level of small areas. We generalize the Fay-Herriot model and propose a hierarchical Bayesian version of the model to estimate the lower level small area means. We

decompose the posterior variance of small area mean and identify the source of increase in uncertainty caused by using aggregate information. A real data example is provided, where direct estimates at different levels of small areas are considered.

INDEX WORDS: Aggregate information, Compromise regression weights, Fay-Herriot model, Hierarchical Bayes, Model misspecification, Pseudo-Bayes, Robustness, Small area levels.

Some Bayesian Contributions to Small Area Estimation

by

Juhyung Lee

B.A., Yonsei University, Republic of Korea, 2013

M.A., Yonsei University, Republic of Korea, 2015

M.S., University of Georgia, 2020

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2021

Some Bayesian Contributions to Small Area Estimation

by

Juhyung Lee

Major Professor:	Gauri Sankar Datta
Committee:	Abhyuday Mandal Cheolwoo Park T.N. Sriram

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2021

Dedication

To my parents

Acknowledgments

I was able to finish this long journey with the help of many people.

I am very lucky to have Dr. Gauri S. Datta as my advisor who has always been insightful, understanding, and encouraging. My sincere gratitude goes to Dr. Cheolwoo Park who gave me the opportunity to study at UGA. Thanks to Dr. T.N. Sriram, I could gain my teaching experience in statistics, which eventually made me pursue a career in academia. I learned a lot from Dr. Daniel B. Hall, not only from his lectures, but also from working under his supervision at the UGA Statistical Consulting Center. I am grateful to Dr. Abhyuday Mandal for serving on my committee and providing valuable comments.

My special thanks extend to my academic brother Dr. Hee Cheol Chung who took care of me in terms of both living and learning even when he was struggling himself.

Needless to say, my six years in Athens were built on unconditional love and support from my parents, for which I owe them forever. I thank my younger sister Juhae for taking such good care of mother and father during my stay in Athens.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sample surveys are extensively used in practice to provide estimates for both the total population of interest and a variety of subpopulations. Subpopulations are also referred to as domains or areas. A domain can be defined by a geographic area, such as a state or a county, or by a socio-demographic group, such as a specific age-sex-race group within a large geographic area.

A domain estimator of a suitable characteristic of interest is called "direct" if it is based only on the domain-specific sample data on that characteristic. A direct estimator may depend on any known auxiliary variables that are believed to be related to the variable of interest. A domain is referred to as a "small area" if the domain-specific sample is not large enough to yield direct estimates of adequate precision. Small area estimation problems are actually very common in sample surveys. For example, oversampling is frequently employed in surveys to increase the sample sizes of certain domains, but this could make the sample sizes of other domains very small or even zero due to the limited budget. Even when a survey has large enough sample sizes for all domains, say states, to support direct estimates for the total state populations, the sample sizes may well not be large enough to support direct estimates for subgroups of the state populations, such as school-age children or persons in poverty in certain geography.

The demand for small area statistics has greatly increased globally over the last few decades. For example, small area statistics are widely used in the allocation of government funds to its various constituents. In both developed and developing countries, governmental policies increasingly require income and poverty estimates for small areas. In the United States, for example, the state-level median income estimates of four-person families are used by the U.S. Department of Health and Human Services to determine the eligibility for its energy assistance program to low income families. Also, the U.S. Department of Education allocates several billions of dollars annually to counties using the county estimates of poor school-age children. The allocated funds support compensatory education programs to meet the needs of educationally disadvantaged children.

To obtain reliable small area estimates, it is often necessary to use "indirect" estimators that "borrow strength" from direct estimates of other related areas and/or time periods. This is usually done by an implicit/explicit model that links the direct estimates using relevant auxiliary information, such as recent census counts and current administrative records. See Rao and Molina (2015) for various indirect methods that combine information using implicit models. An advantage of explicit small area model is that it can be spelled out so that one can understand how the data are generated and how information from different sources is combined. Moreover, an explicit model allows model selection and model diagnostics, and provides a measure of uncertainty for point estimators. Mixed effects models, or mixed models, are explicit models particularly suitable for small area estimation. A mixed model generally incorporates area-specific random effects that account for between area variation beyond that explained by the mean function of auxiliary variables (i.e., fixed effects) in the model. Complex data structures, such as spatial dependence and time series, can also be handled using mixed models.

The main topic of this dissertation is model-based small area estimation with a Bayesian approach. We discuss two small area estimation problems in the next three chapters. In the first problem, we note that any proposed model is subject to model misspecification

and focus on misspecification of the mean function. The observed best prediction (OBP) proposed by Jiang et al. (2011) is a new prediction procedure that has been shown to be more robust than the traditional empirical best linear unbiased prediction (EBLUP) with respect to misspecification of the mean function. In Chapter 2, we propose a pseudo-Bayesian alternative to the OBP. We apply our pseudo-Bayesian method to the same data that Jiang et al. (2011) used. We also present another real data example where the true small area means are available for evaluating the estimators. Then we carry out similar simulation studies conducted by Jiang et al. (2011) and compare the performance of our pseudo-Bayesian estimator with that of the OBP and EBLUP. Chapter 3 serves as an extension to Chapter 2. In Chapter 3, we propose two compromise pseudo-Bayesian methods using regression weights that compromise between the different weighting schemes adopted by the OBP and EBLUP. As before, our compromise pseudo-Bayesian methods are compared with other methods using real data and simulations.

The second problem considers the situation where surveys provide data on the outcome of interest only at a higher level of aggregation, while the estimates are needed at a lower level of small areas. In Chapter 4, we generalize the Fay-Herriot model (Fay and Herriot, 1979) to accommodate such situations using covariates available at the desired lower level. Then we propose a hierarchical Bayesian version of the generalized Fay-Herriot model to estimate the small area means at the desired lower level. We provide theoretical and empirical comparisons of the posterior variances of the small area means based on higher and lower level direct estimates.

Chapter 2

Pseudo-Bayesian Small Area Estimation

2.1 Introduction

In small area estimation, the problem of main interest is typically estimation of the small area means. Consider the Fay-Herriot model (Fay and Herriot, 1979) that is widely used in small area estimation. The model is a mixed effects model that can be expressed as

$$y_i = \theta_i + e_i,\ \theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + v_i,\ i = 1, \ldots, m, \tag{2.1}$$

where y_i is a direct estimator of the ith small area mean θ_i, $\mathbf{x}_i$ is a $p \times 1$ vector of known covariates, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients. Here, v_i's are area-specific random effects and e_i's are sampling errors. It is assumed that $v_i \overset{\text{iid}}{\sim} \mathrm{N}(0, A)$ independent of $e_i \overset{\text{ind}}{\sim} \mathrm{N}(0, D_i)$, where the variance A is unknown, but the sampling variances D_i's are treated as known. Under the Fay-Herriot model, estimation of the small area means becomes prediction of the mixed effects $\theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + v_i,\ i = 1, \ldots, m$.

Recall that for a predictor $\widetilde{\boldsymbol{\theta}} = (\widetilde{\theta}_i)_{1 \le i \le m}$ of $\boldsymbol{\theta} = (\theta_i)_{1 \le i \le m}$, the mean squared prediction

error (MSPE) is defined as

$$\mathrm{MSPE}(\widetilde{\boldsymbol{\theta}}) = \mathrm{E}(|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}|^2) = \sum_{i=1}^{m} \mathrm{E}(\widetilde{\theta}_i - \theta_i)^2, \tag{2.2}$$

under which the best predictor (BP) of $\boldsymbol{\theta}$ is its conditional expectation $\widetilde{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta}|\mathbf{y})$, where $\mathbf{y} = (y_i)_{1\le i\le m}$. Under the assumed model (2.1) and given the parameter $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, A)^T$, the BP can be expressed as

$$\widetilde{\boldsymbol{\theta}}(\boldsymbol{\psi}) = \mathrm{E}_{\mathrm{M},\boldsymbol{\psi}}(\boldsymbol{\theta}|\mathbf{y}) = \left[\mathbf{x}_i^T\boldsymbol{\beta} + \frac{A}{A+D_i}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\right]_{1\le i\le m} \tag{2.3}$$

or, componentwise, $\widetilde{\theta}_i(\boldsymbol{\psi}) = \mathbf{x}_i^T\boldsymbol{\beta} + B_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta})$, where $B_i = A/(A+D_i)$ and $\mathrm{E}_{\mathrm{M},\boldsymbol{\psi}}$ represents expectation under the assumed model with $\boldsymbol{\psi}$ being the true parameter. Note that E in (2.2) is with respect to the true underlying distribution, which is unknown but not model dependent, and hence $\mathrm{E}_{\mathrm{M},\boldsymbol{\psi}}$ in (2.3) is different from E unless the assumed model is correct and $\boldsymbol{\psi}$ is the true parameter. To obtain the best linear unbiased predictor (BLUP) of $\boldsymbol{\theta}$ under the model (2.1), one assumes A is known and replaces $\boldsymbol{\beta}$ in the BP (2.3) with its maximum likelihood estimator (MLE)

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \left(\sum_{i=1}^{m}\frac{\mathbf{x}_i\mathbf{x}_i^T}{A+D_i}\right)^{-1}\sum_{i=1}^{m}\frac{\mathbf{x}_i y_i}{A+D_i},$$

where the $m \times p$ design matrix $\mathbf{X} = (\mathbf{x}_i^T)_{1\le i\le m}$ is assumed to be of full column rank without loss of generality and $\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \mathrm{diag}(A + D_i,\ 1 \le i \le m)$. Then one replaces the unknown variance A in the BLUP with its MLE or residual MLE to obtain the empirical best linear unbiased predictor (EBLUP) of $\boldsymbol{\theta}$. The implication here is that the EBLUP may be regarded as a hybrid of optimal prediction (i.e., BP) and optimal estimation (e.g., MLE).

Since the EBLUP depends on the underlying mean function $\mathbf{x}_i^T\boldsymbol{\beta}$, its efficiency is affected under misspecification of the mean function. Jiang et al. (2011) considered the case where prediction is of main interest and derived a purely predictive procedure in which both the

predictor of $\boldsymbol{\theta}$ and estimators of model parameters are derived from predictive considerations. Suppose that the true underlying model can be expressed as

$$y_i = \theta_i + e_i,\ \theta_i = \mathrm{E}(y_i) + v_i,\ i = 1, \ldots, m, \tag{2.4}$$

where $\mathrm{E}(y_i)$ is the true, but unknown mean, and v_i's and e_i's are the same as in (2.1). Under the true underlying model (2.4),

$$\begin{aligned}
\mathrm{MSPE}[\widetilde{\boldsymbol{\theta}}(\boldsymbol{\psi})] &= \mathrm{E}[|\widetilde{\boldsymbol{\theta}}(\boldsymbol{\psi}) - \boldsymbol{\theta}|^2] \\
&= \sum_{i=1}^{m} \mathrm{E}[\{\widetilde{\theta}_i(\boldsymbol{\psi}) - \theta_i\}^2] \\
&= \sum_{i=1}^{m} \mathrm{E}[\{\mathbf{x}_i^T\boldsymbol{\beta} + B_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \theta_i\}^2] \\
&= \sum_{i=1}^{m} \mathrm{E}[\{y_i - \theta_i - (1 - B_i)(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\}^2] \\
&= \sum_{i=1}^{m} \mathrm{E}[\{(1 - B_i)(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - e_i\}^2] \\
&= \sum_{i=1}^{m} \{\mathrm{E}[(1 - B_i)^2(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2] - 2(1 - B_i)D_i + D_i\} \\
&= \sum_{i=1}^{m} \mathrm{E}[(1 - B_i)^2(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + 2B_iD_i - D_i] \\
&= \mathrm{E}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Gamma}^2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2A \cdot \mathrm{tr}(\boldsymbol{\Gamma}) - \mathrm{tr}(\mathbf{D})],
\end{aligned} \tag{2.5}$$

where $\boldsymbol{\Gamma} = \mathrm{diag}(1 - B_i = D_i/(A + D_i),\ 1 \le i \le m)$ and $\mathbf{D} = \mathrm{diag}(D_i,\ 1 \le i \le m)$. Jiang et al. (2011) proposed an estimator of $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, A)^T$ by minimizing the expression inside the expectation on the right-hand side of (2.5), which is equivalent to minimizing

$$Q(\boldsymbol{\psi}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Gamma}^2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2A \cdot \mathrm{tr}(\boldsymbol{\Gamma}). \tag{2.6}$$

The resulting estimator $\widetilde{\boldsymbol{\psi}} = (\widetilde{\boldsymbol{\beta}}^T, \widetilde{A})^T$ is referred to as the best predictive estimator (BPE)

of $\boldsymbol{\psi}$. Note that if A is known, the BPE of $\boldsymbol{\beta}$ can be obtained by minimizing the first term on the right-hand side of (2.6), which yields a closed-form solution

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Gamma}^2\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Gamma}^2\mathbf{y} = \left\{\sum_{i=1}^{m}\left(\frac{D_i}{A+D_i}\right)^2\mathbf{x}_i\mathbf{x}_i^T\right\}^{-1}\sum_{i=1}^{m}\left(\frac{D_i}{A+D_i}\right)^2\mathbf{x}_i y_i.$$

The BPE of A can be obtained by minimizing $Q(\boldsymbol{\psi})$ in (2.6) with $\boldsymbol{\beta}$ replaced by $\widetilde{\boldsymbol{\beta}}$ subject to $A > 0$. Then the BPE of $\boldsymbol{\beta}$ when A is unknown, which is usually the case, can be obtained by $\widetilde{\boldsymbol{\beta}}$ with A replaced by its BPE $\widetilde{A}$. Finally, Jiang et al. (2011) proposed a predictor of the mixed effects $\boldsymbol{\theta}$ obtained by replacing $\boldsymbol{\psi}$ in the BP (2.3) with its BPE $\widetilde{\boldsymbol{\psi}} = (\widetilde{\boldsymbol{\beta}}^T, \widetilde{A})^T$. Since the BPE minimizes the expression inside the expectation in MSPE$[\widetilde{\boldsymbol{\theta}}(\boldsymbol{\psi})]$, that is, the "observed" MSPE, the predictor is referred to as the observed best predictor (OBP).

Jiang et al. (2011) showed that the OBP can significantly outperform EBLUP in terms of the MSPE if the underlying model is misspecified. When the underlying model is correctly specified, on the other hand, Jiang et al. (2011) showed that the overall predictive performance of the OBP is very similar to that of the EBLUP if the number of small areas is large.

2.2 A Pseudo-Bayesian Alternative to OBP

Recall the objective function $Q(\boldsymbol{\psi})$ in (2.6) defined by Jiang et al. (2011). To be more explicit, we now denote it by $Q(\boldsymbol{\beta}, A, \mathbf{y})$. In this section, we derive a pseudo-Bayesian alternative to the OBP by converting $Q(\boldsymbol{\beta}, A, \mathbf{y})$ to a likelihood function and introducing suitable prior densities for $\boldsymbol{\beta}$ and A. In the following, with a slight abuse of notation, we use $\pi(\cdot)$ to denote relevant prior/posterior densities when there is no potential confusion. For simplicity, we first assume A is known, but $\boldsymbol{\beta}$ is unknown. Then the objective function becomes $Q(\boldsymbol{\beta}, \mathbf{y}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Gamma}^2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. With the popular noninformative uniform prior

$\pi(\boldsymbol{\beta}) \propto 1,\ \boldsymbol{\beta} \in \mathbb{R}^p$, a posterior density of $\boldsymbol{\beta}$ can be defined as

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto \exp[-\frac{1}{2}Q(\boldsymbol{\beta}, \mathbf{y})] \cdot \pi(\boldsymbol{\beta}) \propto \exp[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Gamma}^2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]. \tag{2.7}$$

Note that from the objective function, we have created a likelihood function that is proportional to $\exp[-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Gamma}^2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/2]$, which is the kernel of a $\mathrm{N}(\mathbf{X}\boldsymbol{\beta}, (\boldsymbol{\Gamma}^2)^{-1})$ distribution. The rationale behind this approach is that $Q(\boldsymbol{\beta}, \mathbf{y})$ is actually a loss function to be minimized and mathematically, minimizing the loss function is equivalent to maximizing $-Q(\boldsymbol{\beta}, \mathbf{y})$, where $\exp[-Q(\boldsymbol{\beta}, \mathbf{y})/2]$ is proportional to the likelihood function (Mallick et al., 2005). This duality between "likelihood" and "loss," particularly viewing the loss as the negative of the log-likelihood, is referred to in the Bayesian literature as a "logarithmic scoring rule" (Bernardo, 1979).

It can be shown by simple algebra that $\boldsymbol{\beta}|\mathbf{y} \sim \mathrm{N}\left(\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Gamma}^2\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Gamma}^2\mathbf{y}, (\mathbf{X}^T\boldsymbol{\Gamma}^2\mathbf{X})^{-1}\right)$. However, there is a scaling issue related to $Q(\boldsymbol{\beta}, \mathbf{y})$, which in turn affects the Bayesian decision for $\boldsymbol{\beta}$ based on the posterior density (2.7). Suppose we scale the data $\mathbf{y}$ to $\mathbf{y}^* = \mathbf{y}/c$ for some $c > 0$. For the $\mathbf{y}^*$ data, the relevant parameters are $\boldsymbol{\beta}^* = \boldsymbol{\beta}/c$ and $A^* = A/c^2$, and the sampling error variances become $D_i^* = D_i/c^2,\ i = 1, \ldots, m$. In this setting, similar to (2.7), we obtain

$$\begin{aligned}\pi(\boldsymbol{\beta}^*|\mathbf{y}^*) &\propto \exp[-\frac{1}{2}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}^*)^T\boldsymbol{\Gamma}^{*2}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}^*)] \\ &= \exp[-\frac{1}{2c^2}\{\mathbf{y} - \mathbf{X}(c\boldsymbol{\beta}^*)\}^T\boldsymbol{\Gamma}^2\{\mathbf{y} - \mathbf{X}(c\boldsymbol{\beta}^*)\}],\end{aligned}$$

where $\boldsymbol{\Gamma}^* = \mathrm{diag}(D_i^*/(A^* + D_i^*),\ 1 \le i \le m) = \mathrm{diag}(D_i/(A + D_i),\ 1 \le i \le m) = \boldsymbol{\Gamma}$. If we define $\boldsymbol{\beta}_c = c\boldsymbol{\beta}^*$, then the posterior density of $\boldsymbol{\beta}_c$ given $\mathbf{y}^*$ is given by

$$\pi(\boldsymbol{\beta}_c|\mathbf{y}^*) \propto \exp[-\frac{1}{2c^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_c)^T\boldsymbol{\Gamma}^2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_c)] = \exp[-\frac{1}{2c^2}Q(\boldsymbol{\beta}_c, \mathbf{y})]. \tag{2.8}$$

Comparing (2.7) and (2.8), $\pi(\boldsymbol{\beta}|\mathbf{y}) \neq \pi(\boldsymbol{\beta}_c|\mathbf{y}^*)$ and this implies the Bayesian decision for

$\boldsymbol{\beta}$ will not be scale-equivariant. To avoid the problem due to scaling, we suggest scaling the data at the outset. We define $\mathbf{y}^* = (y_i^*)_{1\le i\le m} = (y_i/\sqrt{D_{\max}})_{1\le i\le m}$, $\boldsymbol{\theta}^* = (\theta_i^*)_{1\le i\le m} = (\theta_i/\sqrt{D_{\max}})_{1\le i\le m}$, $\boldsymbol{\beta}^* = \boldsymbol{\beta}/\sqrt{D_{\max}}$, $A^* = A/D_{\max}$, and $\mathbf{D}^* = \text{diag}(D_i^*,\ 1 \le i \le m) = \text{diag}(D_i/D_{\max},\ 1 \le i \le m)$, where $D_{\max} = \max_{1\le i\le m} D_i$. The scaled data lead to the objective function

$$Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*) = (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}^*)^T \boldsymbol{\Gamma}^{*2} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}^*) + 2A^* \cdot \text{tr}(\boldsymbol{\Gamma}^*). \tag{2.9}$$

Suppose we define a posterior density for $\boldsymbol{\beta}^*$ and A^* using $Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)$ by

$$\pi(\boldsymbol{\beta}^*, A^*|\mathbf{y}^*) \propto \exp[-\frac{1}{2}Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)] \cdot \pi(\boldsymbol{\beta}^*, A^*) \propto \exp[-\frac{1}{2}Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)], \tag{2.10}$$

where the prior density $\pi(\boldsymbol{\beta}^*, A^*) \propto 1,\ \boldsymbol{\beta}^* \in \mathbb{R}^p,\ A^* > 0$. Then

$$\begin{aligned}
\pi(A^*|\mathbf{y}^*) &= \int_{\mathbb{R}^p} \pi(\boldsymbol{\beta}^*, A^*|\mathbf{y}^*) d\boldsymbol{\beta}^* \\
&\propto \int_{\mathbb{R}^p} \exp[-\frac{1}{2}Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)] d\boldsymbol{\beta}^* \\
&= \int_{\mathbb{R}^p} \exp[-\frac{1}{2}\{(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}^*)^T \boldsymbol{\Gamma}^{*2} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}^*) + 2A^* \cdot \text{tr}(\boldsymbol{\Gamma}^*)\}] d\boldsymbol{\beta}^* \\
&= \int_{\mathbb{R}^p} \exp[-\frac{1}{2}\{(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*)^T \boldsymbol{\Gamma}^{*2} (\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*) + (\mathbf{y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*)^T \boldsymbol{\Gamma}^{*2} (\mathbf{y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*) \\
&\quad + 2A^* \cdot \text{tr}(\boldsymbol{\Gamma}^*)\}] d\boldsymbol{\beta}^* \\
&= \int_{\mathbb{R}^p} \exp[-\frac{1}{2}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}^*)^T \mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X} (\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}^*)] d\boldsymbol{\beta}^* \\
&\quad \times \exp[-\frac{1}{2}\{(\mathbf{y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*)^T \boldsymbol{\Gamma}^{*2} (\mathbf{y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*) + 2A^* \cdot \text{tr}(\boldsymbol{\Gamma}^*)\}] \\
&\propto |\mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X}|^{-\frac{1}{2}} \exp[-\frac{1}{2}Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*)],
\end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}^* = (\mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{y}^*$. Note that as $A^* \to \infty$,

$$|\mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X}|^{-\frac{1}{2}} = \left| \sum_{i=1}^m \left(\frac{D_i^*}{A^* + D_i^*} \right)^2 \mathbf{x}_i \mathbf{x}_i^T \right|^{-\frac{1}{2}} \to \infty$$

and

$$
\begin{aligned}
Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*) &= (\mathbf{y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*)^T \boldsymbol{\Gamma}^{*2} (\mathbf{y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*) + 2A^* \cdot \mathrm{tr}(\boldsymbol{\Gamma}^*) \\
&= \sum_{i=1}^{m} \left(\frac{D_i^*}{A^* + D_i^*} \right)^2 (y_i^* - \mathbf{x}_i^T \widetilde{\boldsymbol{\beta}}^*)^2 + 2 \sum_{i=1}^{m} \frac{A^* D_i^*}{A^* + D_i^*} \\
&\to 2 \sum_{i=1}^{m} D_i^* < \infty.
\end{aligned}
$$

This implies $|\mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X}|^{-1/2} \exp[-Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*)/2] \to \infty$ as $A^* \to \infty$ and hence $\pi(A^*|\mathbf{y}^*)$ is non-integrable over the region (a, ∞) for any $a > 0$. As a result, the posterior density $\pi(\boldsymbol{\beta}^*, A^*|\mathbf{y}^*)$ defined in (2.10) would not be proper.

To achieve the propriety of posterior density obtained from the objective function of $\boldsymbol{\beta}^*$ and A^*, we define a modified objective function by adding a penalty term for estimation of A^* to $Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)$ in (2.9). The penalty term may be viewed as prior information on the variance parameter. By introducing a suitable tuning parameter $\lambda > 0$, a modified objective function is defined as

$$
Q_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*) = Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*) + \lambda \sum_{i=1}^{m} \log(A^* + D_i^*),
$$

from which a posterior density for $\boldsymbol{\beta}^*$ and A^* can be defined as

$$
\begin{aligned}
\pi_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*|\mathbf{y}^*) &\propto \exp[-\frac{1}{2} Q_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)] \\
&= \exp[-\frac{1}{2} Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)] \prod_{i=1}^{m} (A^* + D_i^*)^{-\frac{\lambda}{2}}. \qquad (2.11)
\end{aligned}
$$

The penalty term $\lambda \sum_{i=1}^{m} \log(A^* + D_i^*)$ can be interpreted as the prior density $\pi_{\mathrm{m}}(A^*)$ that is proportional to $\prod_{i=1}^{m} (A^* + D_i^*)^{-\lambda/2}$. Then it is implied in (2.11) that $\pi(\boldsymbol{\beta}^*) \propto 1,\ \boldsymbol{\beta}^* \in \mathbb{R}^p$, so that jointly $\pi_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*) = \pi(\boldsymbol{\beta}^*) \pi_{\mathrm{m}}(A^*) \propto \prod_{i=1}^{m} (A^* + D_i^*)^{-\lambda/2},\ \boldsymbol{\beta}^* \in \mathbb{R}^p,\ A^* > 0$. Clearly, it should be that $m\lambda/2 > 1 \Longleftrightarrow \lambda > 2/m$ for $\pi_{\mathrm{m}}(A^*)$ to be integrable over the region (a, ∞)

for any $a > 0$. Now,

$$\begin{aligned}
\pi_{\mathrm{m}}(A^*|\mathbf{y}^*) &= \int_{\mathbb{R}^p} \pi_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*|\mathbf{y}^*) d\boldsymbol{\beta}^* \\
&\propto \int_{\mathbb{R}^p} \exp[-\frac{1}{2} Q_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)] d\boldsymbol{\beta}^* \\
&= \int_{\mathbb{R}^p} \exp[-\frac{1}{2} Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)] d\boldsymbol{\beta}^* \prod_{i=1}^{m} (A^* + D_i^*)^{-\frac{\lambda}{2}} \\
&\propto |\mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X}|^{-\frac{1}{2}} \exp[-\frac{1}{2} Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*)] \prod_{i=1}^{m} (A^* + D_i^*)^{-\frac{\lambda}{2}} \\
&= \left| \sum_{i=1}^{m} \left(\frac{D_i^*}{A^* + D_i^*} \right)^2 \mathbf{x}_i \mathbf{x}_i^T \right|^{-\frac{1}{2}} \exp[-\frac{1}{2} Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*)] \prod_{i=1}^{m} (A^* + D_i^*)^{-\frac{\lambda}{2}}. \qquad (2.12)
\end{aligned}$$

The right-hand side of (2.12) will be integrable over the region (a, ∞) for any $a > 0$ if the function $g(A^*) = (A^*)^{p - m\lambda/2}$ is integrable over the same region. This results in a condition on the tuning parameter λ given by

$$p - \frac{m\lambda}{2} < -1 \Longleftrightarrow \lambda > \frac{2(p+1)}{m}.$$

To ensure the finite posterior mean of A^*, we need

$$p - \frac{m\lambda}{2} + 1 < -1 \Longleftrightarrow \lambda > \frac{2(p+2)}{m}.$$

Similarly, the condition

$$p - \frac{m\lambda}{2} + 2 < -1 \Longleftrightarrow \lambda > \frac{2(p+3)}{m}$$

ensures the finite posterior variance of A^*.

2.3 Pseudo-Bayesian Estimation of Small Area Means

Recall that our goal is estimation of the small area means, which, under the Fay-Herriot model (2.1), becomes prediction of the mixed effects $\theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + v_i,\ i = 1, \ldots, m$. Pseudo-Bayesian estimation of the small area means can be achieved by drawing $(\boldsymbol{\beta}^{*T}, A^*)^T$ from $\pi_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*|\mathbf{y}^*)$ in (2.11) with a suitably chosen value of the tuning parameter λ subject to $\lambda > 2(p+1)/m$, the condition on λ to make $\pi_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*|\mathbf{y}^*)$ proper. To this end, S values of A^* can be first drawn independently from its marginal posterior distribution $\pi_{\mathrm{m}}(A^*|\mathbf{y}^*)$ in (2.12) using one of the standard sampling techniques such as rejection sampling or Markov chain Monte Carlo (MCMC) methods. Then, given $A^* = A^{*(s)},\ s = 1, \ldots, S$, where the superscript s denotes the sth drawn value, $\boldsymbol{\beta}^{*(s)}$ can be drawn from the conditional posterior distribution

$$\begin{aligned}
\pi_{\mathrm{m}}(\boldsymbol{\beta}^*|A^*, \mathbf{y}^*) \propto \pi_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*|\mathbf{y}^*) &\propto \exp[-\frac{1}{2}Q_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)] \\
&\propto \exp[-\frac{1}{2}Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)] \\
&\propto \exp[-\frac{1}{2}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}^*)^T\boldsymbol{\Gamma}^{*2}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}^*)] \\
&\propto \exp[-\frac{1}{2}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}^*)^T\mathbf{X}^T\boldsymbol{\Gamma}^{*2}\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}^*)],
\end{aligned}$$

which is a $\mathrm{N}\left(\widetilde{\boldsymbol{\beta}}^* = (\mathbf{X}^T\boldsymbol{\Gamma}^{*2}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Gamma}^{*2}\mathbf{y}^*, (\mathbf{X}^T\boldsymbol{\Gamma}^{*2}\mathbf{X})^{-1}\right)$ distribution and the distribution depends on A^* via $\boldsymbol{\Gamma}^* = \mathrm{diag}(D_i^*/(A^* + D_i^*),\ 1 \le i \le m)$.

Now, note that the model (2.1) implies

$$\begin{pmatrix} y_i \\ \theta_i \end{pmatrix} \Bigg| \boldsymbol{\beta}, A \overset{\text{ind}}{\sim} \mathrm{N}\left(\begin{pmatrix} \mathbf{x}_i^T\boldsymbol{\beta} \\ \mathbf{x}_i^T\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} A + D_i & A \\ A & A \end{pmatrix}\right)$$

and hence

$$\theta_i|\boldsymbol{\beta}, A, \mathbf{y} \overset{\text{ind}}{\sim} \mathrm{N}\left(\mathbf{x}_i^T\boldsymbol{\beta} + \frac{A}{A+D_i}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}), \frac{AD_i}{A+D_i}\right).$$

Therefore, given $(\boldsymbol{\beta}^{*T}, A^*)^T = (\{\boldsymbol{\beta}^{*(s)}\}^T, A^{*(s)})^T$, the sth simulation of the ith (scaled) small area mean $\theta_i^{*(s)}$ can be drawn from a $\mathrm{N}\left(\mathbf{x}_i^T\boldsymbol{\beta}^* + A^*/(A^*+D_i^*)\cdot(y_i^* - \mathbf{x}_i^T\boldsymbol{\beta}^*), A^*D_i^*/(A^*+D_i^*)\right)$ distribution.

Finally, by rescaling and taking the average, a pseudo-Bayesian estimator (PBE) of θ_i can be obtained as

$$\widetilde{\theta}_i^{\mathrm{PB}} = \frac{1}{S}\sum_{s=1}^{S}\theta_i^{(s)} = \frac{1}{S}\sum_{s=1}^{S}\sqrt{D_{\max}}\cdot\theta_i^{*(s)},$$

where the posterior sample size S is suitably large and $\theta_i^{(s)} = \sqrt{D_{\max}}\cdot\theta_i^{*(s)}$.

2.4 Tuning Parameter Selection

In our pseudo-Bayesian alternative to the OBP, we convert the objective function $Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)$ to a likelihood function and obtain the posterior density $\pi_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*|\mathbf{y}^*)$ by introducing the prior density $\pi_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*) \propto \prod_{i=1}^{m}(A^*+D_i^*)^{-\lambda/2},\ \boldsymbol{\beta}^* \in \mathbb{R}^p,\ A^* > 0$. Clearly, we need a reasonable value for the tuning parameter λ. Below we propose a selection procedure for λ.

Recall that the BPE of $(\boldsymbol{\beta}^{*T}, A^*)^T$ minimizes $Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)$. In the pseudo-Bayesian framework, we seek a value of λ that minimizes $\mathrm{E}_{\boldsymbol{\beta}^*, A^*}[Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)|\mathbf{y}^*]$. Using iterated expectation,

$$\mathrm{E}_{\boldsymbol{\beta}^*, A^*}[Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)|\mathbf{y}^*] = \mathrm{E}_{A^*}[\mathrm{E}_{\boldsymbol{\beta}^*}\{Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)|A^*, \mathbf{y}^*\}|\mathbf{y}^*],$$

where the second expectation on the right-hand side

$$
\begin{aligned}
&\mathrm{E}_{\boldsymbol{\beta}^*}[Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)|A^*, \mathbf{y}^*] \\
=&\mathrm{E}_{\boldsymbol{\beta}^*}[(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}^*)^T \boldsymbol{\Gamma}^{*2}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}^*) + 2A^* \cdot \mathrm{tr}(\boldsymbol{\Gamma}^*)|A^*, \mathbf{y}^*] \\
=&\mathrm{E}_{\boldsymbol{\beta}^*}[(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*)^T \boldsymbol{\Gamma}^{*2}(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*)|A^*, \mathbf{y}^*] + (\mathbf{y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*)^T \boldsymbol{\Gamma}^{*2}(\mathbf{y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}^*) + 2A^* \cdot \mathrm{tr}(\boldsymbol{\Gamma}^*) \\
&\qquad\qquad (\widetilde{\boldsymbol{\beta}}^* = (\mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{y}^*) \\
=&\mathrm{E}_{\boldsymbol{\beta}^*}[(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}^*)^T \mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}^*)|A^*, \mathbf{y}^*] + Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*) \\
=&\mathrm{tr}[\mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X}(\mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X})^{-1}] + Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*) \qquad (\boldsymbol{\beta}^*|A^*, \mathbf{y}^* \sim \mathrm{N}(\widetilde{\boldsymbol{\beta}}^*, (\mathbf{X}^T \boldsymbol{\Gamma}^{*2} \mathbf{X})^{-1})) \\
=&\mathrm{tr}(\mathbf{I}_p) + Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*) \\
=&p + Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*).
\end{aligned}
$$

Therefore, $\mathrm{E}_{\boldsymbol{\beta}^*, A^*}[Q(\boldsymbol{\beta}^*, A^*, \mathbf{y}^*)|\mathbf{y}^*] = \mathrm{E}_{A^*}[p + Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*)|\mathbf{y}^*]$, minimizing which is equivalent to minimizing $\mathrm{E}_{A^*}[Q(\widetilde{\boldsymbol{\beta}}^*, A^*, \mathbf{y}^*)|\mathbf{y}^*]$. We propose the following procedure for selecting the tuning parameter λ:

(i) Set up a grid of λ values $\lambda_j,\ j = 1, \ldots, J$, where $2(p+1)/m < \lambda_1 < \cdots < \lambda_J$;

(ii) Given λ_j, draw $A_j^{*(s)},\ s = 1, \ldots, S$, independently from $\pi_{\mathrm{m}}(A^*|\mathbf{y}^*)$ with S suitably large;

(iii) Compute $\overline{Q}(\lambda_j) \triangleq S^{-1} \sum_{s=1}^{S} Q(\widetilde{\boldsymbol{\beta}}^*(A_j^{*(s)}), A_j^{*(s)}, \mathbf{y}^*)$, where $\widetilde{\boldsymbol{\beta}}^*(A_j^{*(s)})$ denotes $\widetilde{\boldsymbol{\beta}}^*$ with A^* replaced by $A_j^{*(s)}$;

(iv) Choose $\widetilde{\lambda} \in \{\lambda_j,\ j = 1, \ldots, J\}$ such that $\overline{Q}(\widetilde{\lambda}) = \min_{1 \le j \le J} \overline{Q}(\lambda_j)$ as the final tuning parameter.

2.5 Real Data Examples

In this section, we illustrate our pseudo-Bayesian alternative to the OBP with two real data examples.

2.5.1 Hospital Data

Morris and Christiansen (1996) presented a dataset involving 23 hospitals (out of a total of 219 hospitals) that had at least 50 kidney transplants during a 27-month period. Here, the $m = 23$ hospitals are treated as small areas. The responses $y_i,\ i = 1, \ldots, m$, are graft failure rates for kidney transplant operations, that is, y_i equals the number of graft failures at hospital i divided by the number of kidney transplants at that hospital. The variance for the graft failure rate, D_i, is approximated by $0.2 \times 0.8 = 0.16$ divided by the number of kidney transplants at hospital i, where 0.2 is the observed failure rate for all hospitals. In addition, a severity index x_i is available for each hospital, which is the average fraction of females, blacks, children, and extremely ill kidney recipients at hospital i. The hospital data are presented in Table 2.1 and a scatterplot of graft failure rate vs. severity index is shown in Figure 2.1.

Since the graft failure rates are binomial proportions of fairly large denominators (at least 50), a normal distribution for the y_i's is not unreasonable, at least from an approximation point of view, by the central limit theorem. Several Fay-Herriot models of the form (2.1) with different mean functions $\mathbf{x}_i^T\boldsymbol{\beta}$ have been proposed for the hospital data. Ganesh (2009) proposed a linear Fay-Herriot model with $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_i$. An inspection of the scatterplot in Figure 2.1, however, suggests that a quadratic model would fit the data well except for the point at the upper right corner. To accommodate this potential outlier, Jiang et al. (2010) proposed a cubic Fay-Herriot model with $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0+\beta_1 x_i+\beta_2 x_i^2+\beta_3 x_i^3$. See also Datta et al. (2011), where the same cubic mean function was considered but with the logit transformed y_i's, which were originally proportions. On the other hand, Jiang et al. (2011) proposed a quadratic-outlying (Q-O) Fay-Herriot model with $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + d \cdot I(x_i > 0.3)$, where $I(\cdot)$ is the indicator function. Note that the Q-O model assumes an abrupt "jump" in the mean response when x_i is greater than 0.3; otherwise, the mean response is a quadratic function of the covariate.

Table 2.1: The hospital data, PBE, OBP with measures of uncertainty. Post SD stands for posterior standard deviation.

Area	y_i	x_i	$\sqrt{D_i}$	PBE	Post SD	OBP	$\widetilde{\text{RMSPE}}^{\text{J}}$
1	0.302	0.112	0.055	0.234	0.025	0.239	0.060
2	0.140	0.206	0.053	0.183	0.024	0.181	0.019
3	0.203	0.104	0.052	0.219	0.024	0.220	0.017
4	0.333	0.168	0.052	0.243	0.025	0.249	0.085
5	0.347	0.337	0.047	0.347	0.056	0.347	0.047
6	0.216	0.169	0.046	0.234	0.023	0.234	0.016
7	0.156	0.211	0.046	0.174	0.026	0.172	0.020
8	0.143	0.195	0.046	0.200	0.021	0.197	0.045
9	0.220	0.221	0.044	0.160	0.034	0.162	0.058
10	0.205	0.077	0.044	0.178	0.031	0.180	0.017
11	0.209	0.195	0.042	0.206	0.020	0.206	0.015
12	0.266	0.185	0.041	0.224	0.021	0.228	0.031
13	0.240	0.202	0.041	0.198	0.022	0.201	0.031
14	0.262	0.108	0.036	0.229	0.023	0.234	0.024
15	0.144	0.204	0.036	0.184	0.022	0.180	0.032
16	0.116	0.072	0.035	0.158	0.032	0.154	0.040
17	0.201	0.142	0.033	0.239	0.024	0.236	0.032
18	0.212	0.136	0.032	0.239	0.023	0.238	0.020
19	0.189	0.172	0.031	0.226	0.021	0.223	0.031
20	0.212	0.202	0.029	0.196	0.020	0.199	0.014
21	0.166	0.087	0.029	0.189	0.023	0.187	0.017
22	0.173	0.177	0.027	0.218	0.020	0.212	0.039
23	0.165	0.072	0.025	0.164	0.029	0.165	0.017

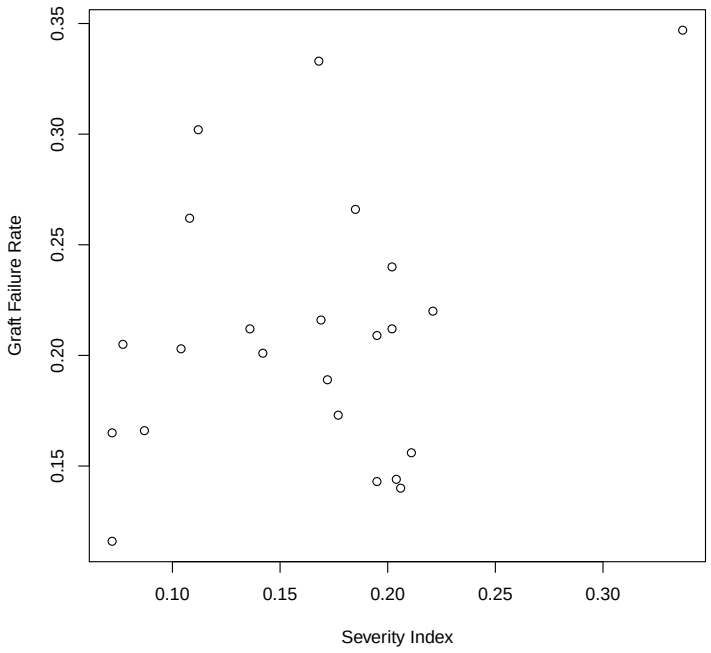

Figure 2.1: Scatterplot of graft failure rate vs. severity index (hospital data).

Following Jiang et al. (2011), we fit the Q-O model to the hospital data, but using our pseudo-Bayesian method. Recall that we need to scale the data first. For the hospital data, $D_{\max} = 0.055^2$ and the data are scaled accordingly. Next, we search for the optimal tuning parameter $\widetilde{\lambda}$ with the procedure proposed in Section 2.4. Since $m = 23$ for the hospital data and $p = 4$ for the Q-O model ($\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, d)^T$), we need the condition $\lambda > 2(p+1)/m = 0.435$ to guarantee the propriety of $\pi_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*|\mathbf{y}^*)$. For a grid of λ values with the lower bound of 0.435, the $\overline{Q}(\lambda)$ values are computed and shown in the left panel of Figure 2.2. The optimal λ is found to be $\widetilde{\lambda} = 1.1$, at which $\overline{Q}(\lambda)$ is minimized. With $\widetilde{\lambda} = 1.1$, $S = 5000$ posterior simulations of A are obtained and a histogram is shown in the right panel of Figure 2.2. The PBE of A, obtained by the posterior mean of A, is $\widetilde{A}^{\mathrm{PB}} = 2.1 \times 10^{-4}$, which is fairly smaller than the BPE $\widetilde{A}^{\mathrm{BPE}} = 3.4 \times 10^{-4}$ reported by Jiang et al. (2011).

Given the $S = 5000$ simulation draws of A, S posterior simulations of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, d)^T$

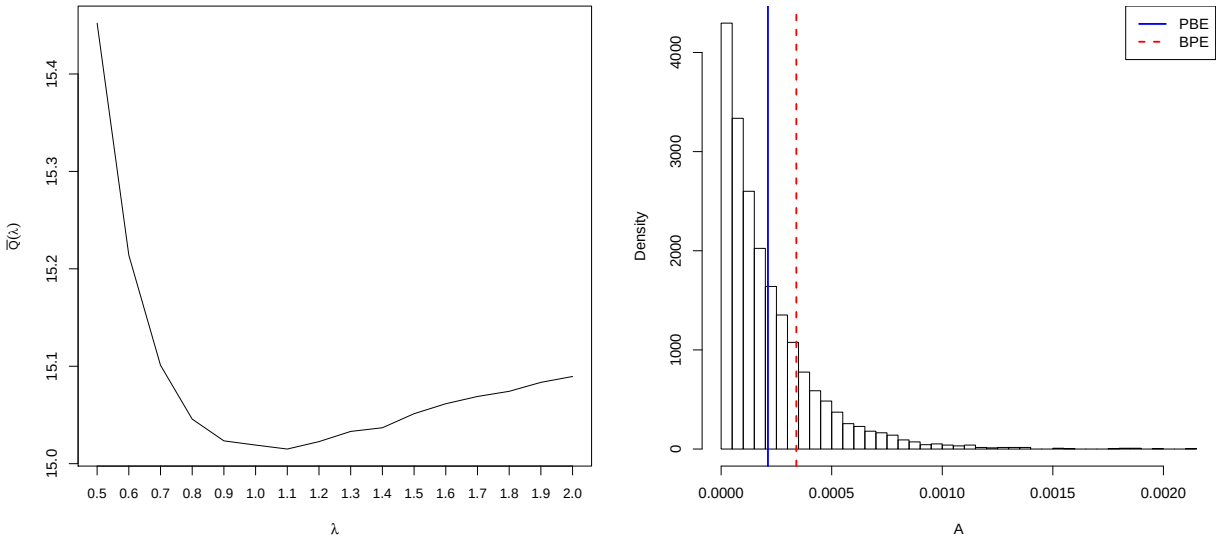

Figure 2.2: Optimal λ search (left) and histogram of posterior simulations of A (right) (hospital data).

are drawn independently from its conditional posterior distribution and histograms are shown in Figure 2.3. The PBE of $\boldsymbol{\beta}$ is computed to be $\widetilde{\boldsymbol{\beta}}^{\mathrm{PB}} = (\widetilde{\beta}_0^{\mathrm{PB}}, \widetilde{\beta}_1^{\mathrm{PB}}, \widetilde{\beta}_2^{\mathrm{PB}}, \widetilde{d}^{\mathrm{PB}})^T = (-0.077, 4.475, -15.518, 0.679)^T$, which is very close to the BPE $\widetilde{\boldsymbol{\beta}}^{\mathrm{BPE}} = (\widetilde{\beta}_0^{\mathrm{BPE}}, \widetilde{\beta}_1^{\mathrm{BPE}}, \widetilde{\beta}_2^{\mathrm{BPE}}, \widetilde{d}^{\mathrm{BPE}})^T = (-0.084, 4.614, -16.045, 0.698)^T$ reported by Jiang et al. (2011).

Now, given the $S = 5000$ simulation draws of $(\boldsymbol{\beta}^T, A)^T$, S posterior simulations of $\boldsymbol{\theta} = (\theta_i)_{1\le i\le m}$ are drawn independently from its posterior distribution. Based on the simulated values, we obtain $\widetilde{\theta}_i^{\mathrm{PB}}$, the PBE of θ_i, and a 95% credible interval for θ_i by directly computing the posterior mean and the 0.025 and 0.975 quantiles. The PBE and the posterior standard deviation of θ_i are reported in Table 2.1. The 95% credible intervals are shown in Figure 2.4.

On the other hand, Jiang et al. (2011) obtained $\widetilde{\theta}_i^{\mathrm{OBP}}$, the OBP of θ_i, and derived a second-order unbiased estimator of the MSPE of $\widetilde{\theta}_i^{\mathrm{OBP}}$ given by

$$\widetilde{\mathrm{MSPE}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}}) = \widetilde{\mathrm{MSPE}}_{\mathrm{n}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}}) + 2(1-\widetilde{B}_i)^2\widetilde{\mathbf{h}}_2^T\widetilde{\mathbf{f}}_i + 4D_i(1-\widetilde{B}_i)^3\mathrm{tr}(\widetilde{\mathbf{G}}_2^{-1}\widetilde{\mathbf{W}}_i), \qquad (2.13)$$

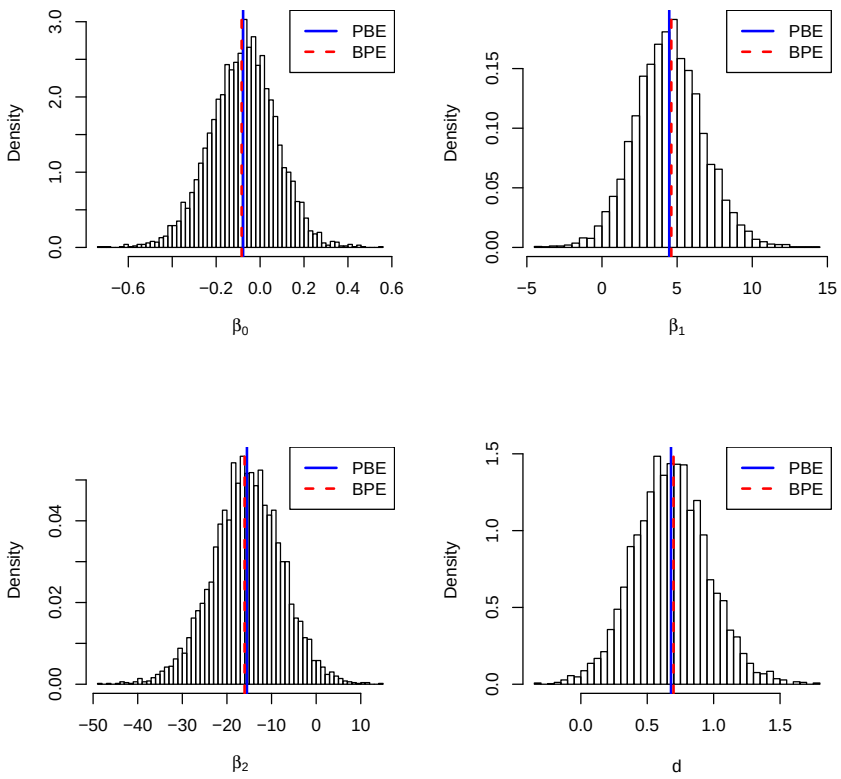

Figure 2.3: Histograms of posterior simulations of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, d)^T$ (hospital data).

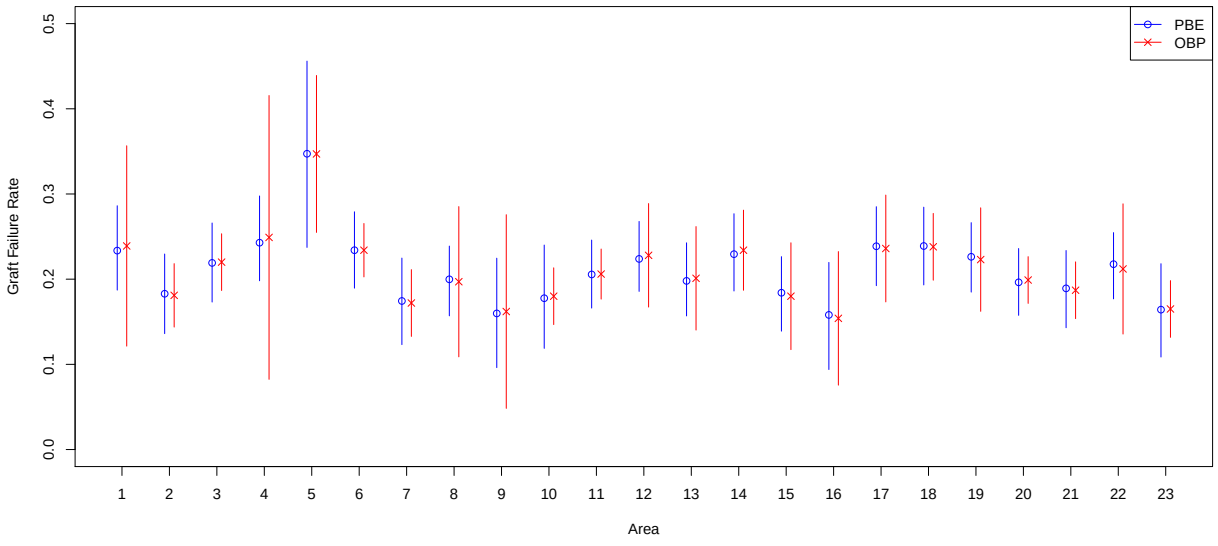

Figure 2.4: PBE and OBP of $\boldsymbol{\theta}$ with 95% credible/confidence intervals (hospital data).

where

$$\widetilde{\mathrm{MSPE}}_{\mathrm{n}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}}) = (\widetilde{\theta}_i^{\mathrm{OBP}} - y_i)^2 + D_i(2\widetilde{B}_i - 1)$$

is a naive estimator, $\widetilde{B}_i$ is B_i with A replaced by $\widetilde{A}^{\mathrm{BPE}}$, and the other quantities are given in detail in Section 6 of Jiang et al. (2011). Note that the naive estimator $\widetilde{\mathrm{MSPE}}_{\mathrm{n}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$ is obtained by plugging $(\boldsymbol{\beta}^T, A)^T = \left((\widetilde{\boldsymbol{\beta}}^{\mathrm{BPE}})^T, \widetilde{A}^{\mathrm{BPE}}\right)^T$ into the expression inside the penultimate expectation in (2.5), where the expectation is $\mathrm{MSPE}[\widetilde{\theta}_i(\boldsymbol{\beta}, A)]$, the MSPE of the BP $\widetilde{\theta}_i(\boldsymbol{\beta}, A)$ of θ_i under the model (2.1). While $\widetilde{\mathrm{MSPE}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$ has a bias of the order $o(m^{-1})$, $\widetilde{\mathrm{MSPE}}_{\mathrm{n}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$ has a bias of the order $O(m^{-1})$, that is, $\widetilde{\mathrm{MSPE}}_{\mathrm{n}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$ is only first-order unbiased. Although second-order unbiased, $\widetilde{\mathrm{MSPE}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$ is not guaranteed to be nonnegative with $\widetilde{\mathrm{MSPE}}_{\mathrm{n}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$ being more likely to take negative values. Alternatively, Jiang et al. (2011) proposed an MSPE estimator that is guaranteed to be nonnegative by using the following bootstrap method:

(i) Generate $\mathbf{y}_{(\ell)} = (y_{i(\ell)})_{1\le i\le m},\ \ell = 1, \ldots, L$, independently from $\mathrm{N}(\widetilde{\boldsymbol{\theta}}^{\mathrm{OBP}}, \mathbf{D})$, where $\widetilde{\boldsymbol{\theta}}^{\mathrm{OBP}} = (\widetilde{\theta}_i^{\mathrm{OBP}})_{1\le i\le m}$ and $\mathbf{D} = \mathrm{diag}(D_i,\ 1 \le i \le m)$;

(ii) Obtain $\widetilde{\theta}_{i(\ell)}^{\mathrm{OBP}}$, the OBP for θ_i based on $\mathbf{y}_{(\ell)},\ i = 1, \ldots, m,\ \ell = 1, \ldots, L$;

(iii) The bootstrap estimator of $\mathrm{MSPE}(\widetilde{\theta}_i^{\mathrm{OBP}})$ is given by

$$\widetilde{\mathrm{MSPE}}_{\mathrm{b}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}}) = \frac{1}{L}\sum_{\ell=1}^{L}\{\widetilde{\theta}_{i(\ell)}^{\mathrm{OBP}} - \widetilde{\theta}_i^{\mathrm{OBP}}\}^2. \tag{2.14}$$

Despite its nonnegative nature, the bootstrap MSPE estimator is only first-order unbiased. Jiang et al. (2011) recommend using the second-order unbiased MSPE estimator (2.13) if it is nonnegative; otherwise using the bootstrap MSPE estimator (2.14).

The OBP of θ_i for the hospital data and the square root of its MSPE estimate (RMSPE) reported by Jiang et al. (2011) are presented in Table 2.1. For the areas 3, 6, 7, 11, 20, and 23, the MSPE estimates (2.13) were negative and Jiang et al. (2011) replaced them with the

bootstrap MSPE estimates (2.14) with $L = 100$. Note that the posterior standard deviations in Table 2.1 are much more stable across the areas compared to the RMSPE estimates. Moreover, for 12 out of the 23 areas, the posterior standard deviations are smaller than the RMSPE estimates. The 6 areas for which the bootstrap MSPE estimates (2.14) were used due to the negative MSPE estimates (2.13) are all members of the other 11 areas.

Given the OBP $\widetilde{\theta}_i^{\mathrm{OBP}}$ and its RMSPE estimator $\widetilde{\mathrm{RMSPE}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$ (obtained by the square root of $\widetilde{\mathrm{MSPE}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$ if it is nonnegative; otherwise by the square root of $\widetilde{\mathrm{MSPE}}_{\mathrm{b}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$), an approximate $100(1-\alpha)\%$ confidence interval for θ_i can be constructed as

$$\left(\widetilde{\theta}_i^{\mathrm{OBP}} - z_{\alpha/2} \cdot \widetilde{\mathrm{RMSPE}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}}),\ \widetilde{\theta}_i^{\mathrm{OBP}} + z_{\alpha/2} \cdot \widetilde{\mathrm{RMSPE}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})\right), \tag{2.15}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathrm{N}(0,1)$ distribution. For the hospital data, OBP 95% confidence intervals of the form (2.15) are computed and shown in Figure 2.4. The PBE and OBP of $\boldsymbol{\theta}$ are very similar, but the 95% credible intervals seem to be overall shorter than the 95% confidence intervals. In fact, the average length of the confidence intervals is 0.124, whereas that of the credible intervals is 0.100, which is about 19% shorter than the former.

2.5.2 Median Income Data

For the second real data example, we consider the estimation of median income of four-person families for the 50 U.S. states and the District of Columbia (i.e., $m = 51$ small areas). The state-level estimates are needed by the U.S. Department of Health and Human Services to formulate its energy assistance program for low income families. The U.S. Census Bureau has been producing such estimates annually using the Current Population Survey (CPS). However, direct use of the CPS estimates is limited due to the smallness of the sample size, which causes substantial variability. In contrast, similar estimates from the decennial census for the year preceding the census year have negligible standard errors and are treated as

the true values for the four-person family median incomes. In this example, we estimate the 1989 four-person family median incomes using the 1990 CPS data and compare the estimates with those from the 1990 census, which are considered as truth.

Let $y_i,\ i = 1, \ldots, m$, be the direct estimate of the 1989 four-person family median income of the ith state from the CPS. As before, we fit a Fay-Herriot model of the form (2.1). For the covariates, we use the following two variables as suggested by Fay (1987):

(1) x_{i1}: 1979 four-person family median income of the ith state from the 1980 census;

(2) $x_{i2} = (\text{PCI}_{i,1989}/\text{PCI}_{i,1979}) \cdot x_{i1}$: 1989 adjusted census four-person family median income of the ith state,

where PCI is the per capita income from the U.S. Bureau of Economic Analysis. Then the mean function of our Fay-Herriot model becomes $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$. Note that x_{i2} attempts to adjust the base year (i.e., 1979) census median by the proportional growth in PCI to arrive at the current year (i.e., 1989) adjusted median. Any possible overstatement of the effect of change in PCI in estimating the current median incomes is believed to be adjusted for by the other covariate x_{i1}. Scatterplots of the response vs. covariates for the median income data are shown in Figure 2.5.

Now, we fit the model using our pseudo-Bayesian method. First, the optimal tuning parameter is found to be $\widetilde{\lambda} = 0.2$, where the lower bound for λ in this case is $2(p + 1)/m = 2(3 + 1)/51 = 0.157$. Next, $S = 5000$ posterior simulations of $(\boldsymbol{\beta}^T, A)^T = (\beta_0, \beta_1, \beta_2, A)^T$ are drawn independently from its posterior distribution and histograms are shown in Figure 2.6. The PBE of $(\boldsymbol{\beta}^T, A)^T$ is computed to be $\left((\widetilde{\boldsymbol{\beta}}^{\text{PB}})^T, \widetilde{A}^{\text{PB}}\right)^T = (\widetilde{\beta}_0^{\text{PB}}, \widetilde{\beta}_1^{\text{PB}}, \widetilde{\beta}_2^{\text{PB}}, \widetilde{A}^{\text{PB}})^T = (12485.9, -0.302, 0.777, 6017755)^T$. The BPE of $(\boldsymbol{\beta}^T, A)^T$ has also been computed and $\left((\widetilde{\boldsymbol{\beta}}^{\text{BPE}})^T, \widetilde{A}^{\text{BPE}}\right)^T = (\widetilde{\beta}_0^{\text{BPE}}, \widetilde{\beta}_1^{\text{BPE}}, \widetilde{\beta}_2^{\text{BPE}}, \widetilde{A}^{\text{BPE}})^T = (13018.2, -0.335, 0.783, 5645048)^T$, which is practically the same as the PBE $\left((\widetilde{\boldsymbol{\beta}}^{\text{PB}})^T, \widetilde{A}^{\text{PB}}\right)^T$ according to Figure 2.6.

Finally, given the $S = 5000$ simulation draws of $(\boldsymbol{\beta}^T, A)^T$, S posterior simulations of $\boldsymbol{\theta} =$

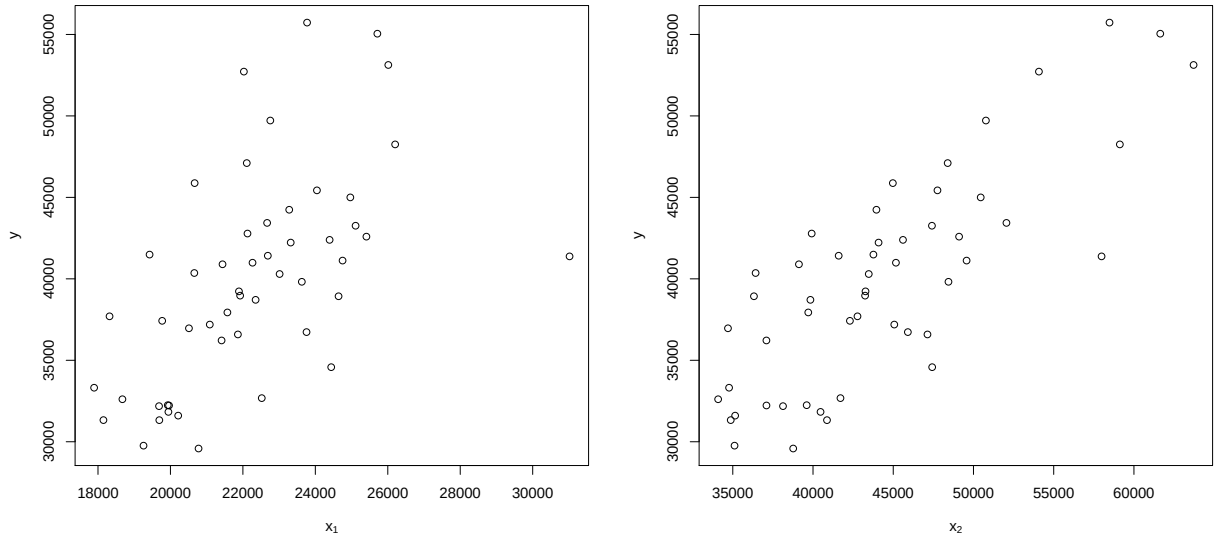

Figure 2.5: Scatterplots of the response vs. covariates (median income data).

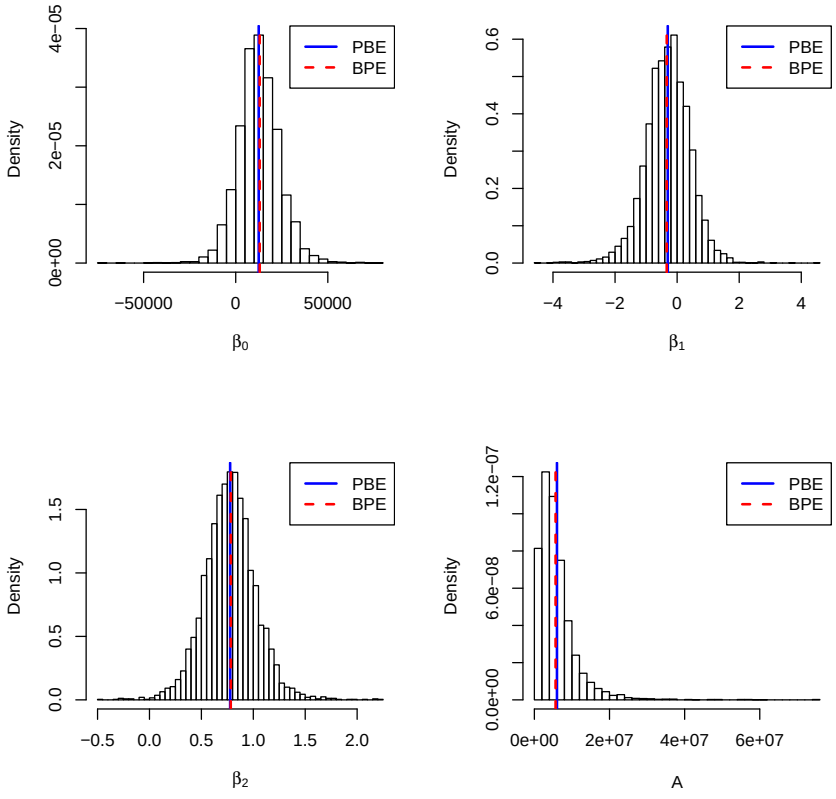

Figure 2.6: Histograms of posterior simulations of $(\boldsymbol{\beta}^T, A)^T = (\beta_0, \beta_1, \beta_2, A)^T$ (median income data).

$(\theta_i)_{1\le i\le m}$ are drawn independently from its posterior distribution. The PBE and OBP of θ_i as well as their measures of uncertainty (i.e., posterior standard deviation and $\widetilde{\mathrm{RMSPE}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$) are compared in Figure 2.7. Note that the PBE and OBP are consistent with each other, but their measures of uncertainty vary. Among the 51 states (including the District of Columbia), there are 19 states with posterior standard deviation smaller than the RMSPE estimate $\widetilde{\mathrm{RMSPE}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$. For the states of Arizona, Hawaii, Indiana, Louisiana, Maryland, Missouri, Rhode Island, Utah, and Vermont, the MSPE estimates (2.13) were negative and the bootstrap MSPE estimates (2.14) were used when computing the RMSPE estimates following the recommendation of Jiang et al. (2011). These 9 states are all members of the other 32 states as opposed to the forementioned 19 states.

In addition to the PBE and OBP, we compute the EBLUP of $\boldsymbol{\theta}$ for comparison. There are four widely used EBLUPs in small area estimation depending on the ways of estimating the variance A, namely, the maximum likelihood (ML), residual maximum likelihood (REML), Fay-Herriot (FH; Fay and Herriot, 1979), and Prasad-Rao (PR; Prasad and Rao, 1990) EBLUPs. We denote these EBLUPs by $\widehat{\boldsymbol{\theta}}^{\mathrm{ML}} = (\widehat{\theta}_i^{\mathrm{ML}})_{1\le i\le m}$, $\widehat{\boldsymbol{\theta}}^{\mathrm{REML}} = (\widehat{\theta}_i^{\mathrm{REML}})_{1\le i\le m}$, $\widehat{\boldsymbol{\theta}}^{\mathrm{FH}} = (\widehat{\theta}_i^{\mathrm{FH}})_{1\le i\le m}$, and $\widehat{\boldsymbol{\theta}}^{\mathrm{PR}} = (\widehat{\theta}_i^{\mathrm{PR}})_{1\le i\le m}$, respectively. See Prasad and Rao (1990), Datta and Lahiri (2000), and Datta et al. (2005) for second-order unbiased estimators of the MSPEs of these EBLUPs. We use $\widehat{\mathrm{MSPE}}(\widehat{\theta}_i)$ to denote those MSPE estimators, where $\widehat{\theta}_i$ is one of the four EBLUPs of θ_i. Given $\widehat{\theta}_i$ and $\widehat{\mathrm{MSPE}}(\widehat{\theta}_i)$, an approximate $100(1-\alpha)\%$ confidence interval for θ_i can be constructed as

$$\left(\widehat{\theta}_i - z_{\alpha/2}\cdot\{\widehat{\mathrm{MSPE}}(\widehat{\theta}_i)\}^{\frac{1}{2}},\ \widehat{\theta}_i + z_{\alpha/2}\cdot\{\widehat{\mathrm{MSPE}}(\widehat{\theta}_i)\}^{\frac{1}{2}}\right),$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathrm{N}(0,1)$ distribution.

Now, for the median income data, the four EBLUPs are computed and we compare them with the PBE and OBP. Recall that the true values of $\theta_i,\ i=1,\ldots,m$, are available from the 1990 census. For each estimator $\breve{\boldsymbol{\theta}} = (\breve{\theta}_i)_{1\le i\le m}$ of $\boldsymbol{\theta} = (\theta_i)_{1\le i\le m}$ we have considered (i.e.,

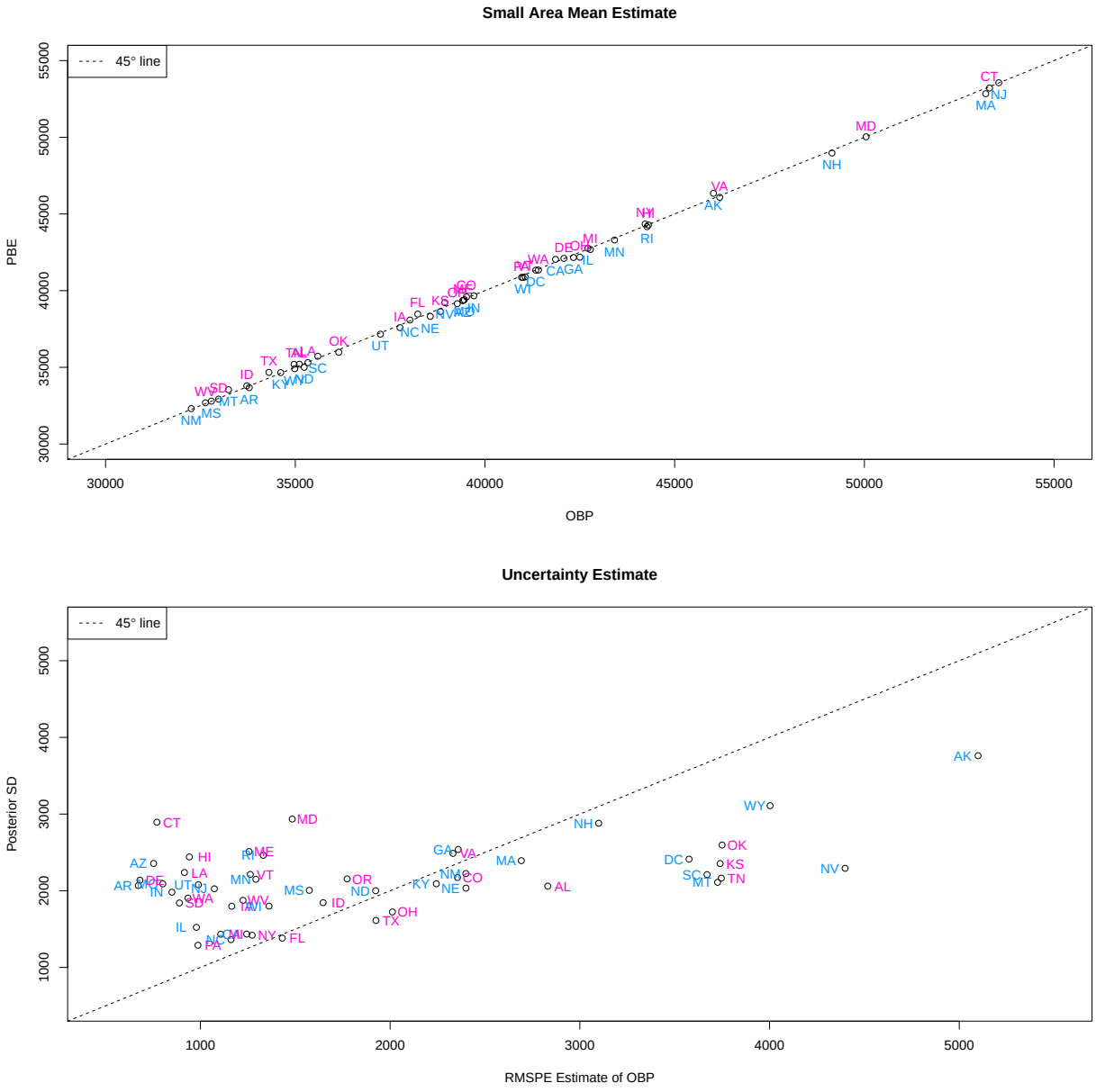

Figure 2.7: Comparisons of the PBE and OBP of $\boldsymbol{\theta}$ (top) and their measures of uncertainty (bottom) (median income data). Colors are for legibility purposes only.

Table 2.2: Comparison of estimators (median income data).

Estimate	AAD	ASD	AARD	ASRD	AL
Direct	2928.8	13.81×10^6	0.0735	0.0084	11424.4
ML	1394.9	3.19×10^6	0.0348	0.0019	7710.5
REML	1454.9	3.45×10^6	0.0363	0.0021	7716.3
FH	1464.4	3.49×10^6	0.0365	0.0021	7762.1
PR	1438.1	3.38×10^6	0.0359	0.0020	7824.6
OBP	1652.4	4.15×10^6	0.0420	0.0026	7540.7
PBE	1580.3	3.81×10^6	0.0403	0.0024	8421.5

OBP, PBE, and EBLUP), we compute four deviation measures: average absolute deviation $\text{AAD}(\check{\boldsymbol{\theta}}) = m^{-1}\sum_{i=1}^m |\check{\theta}_i - \theta_i|$, average squared deviation $\text{ASD}(\check{\boldsymbol{\theta}}) = m^{-1}\sum_{i=1}^m (\check{\theta}_i - \theta_i)^2$, average absolute relative deviation $\text{AARD}(\check{\boldsymbol{\theta}}) = m^{-1}\sum_{i=1}^m |(\check{\theta}_i - \theta_i)/\theta_i|$, and average squared relative deviation $\text{ASRD}(\check{\boldsymbol{\theta}}) = m^{-1}\sum_{i=1}^m \{(\check{\theta}_i - \theta_i)/\theta_i\}^2$. The deviations are reported in Table 2.2, where the row "Direct" corresponds to the direct estimates from the CPS. As expected, the performance of the direct estimator is very poor compared to the other estimators. The PBE is better than the OBP in terms of all four deviation measures, but it is outperformed by all the EBLUPs.

In Table 2.2, the column "AL" indicates the average length of 95% confidence/credible intervals for θ_i based on each estimator. Note that the OBP produces the smallest AL of 7540.7, which is substantially smaller than that of the PBE, 8421.5. This is mainly due to the use of the bootstrap MSPE estimates (2.14) for the nine states with negative MSPE estimates (2.13) (i.e., Arizona, Hawaii, Indiana, Louisiana, Maryland, Missouri, Rhode Island, Utah, and Vermont). For those 9 states, the OBP AL is 4063.5 and the PBE AL is 9249.8, whereas for the other 42 states, the OBP AL is 8285.8 and the PBE AL is 8243.9. In fact, it is shown in Section 2.8 that under the correctly specified Fay-Herriot model, a bootstrap MSPE estimator of the form (2.14) tends to underestimate the actual MSPE. While we do not know if the mean function $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ is correctly specified for the median

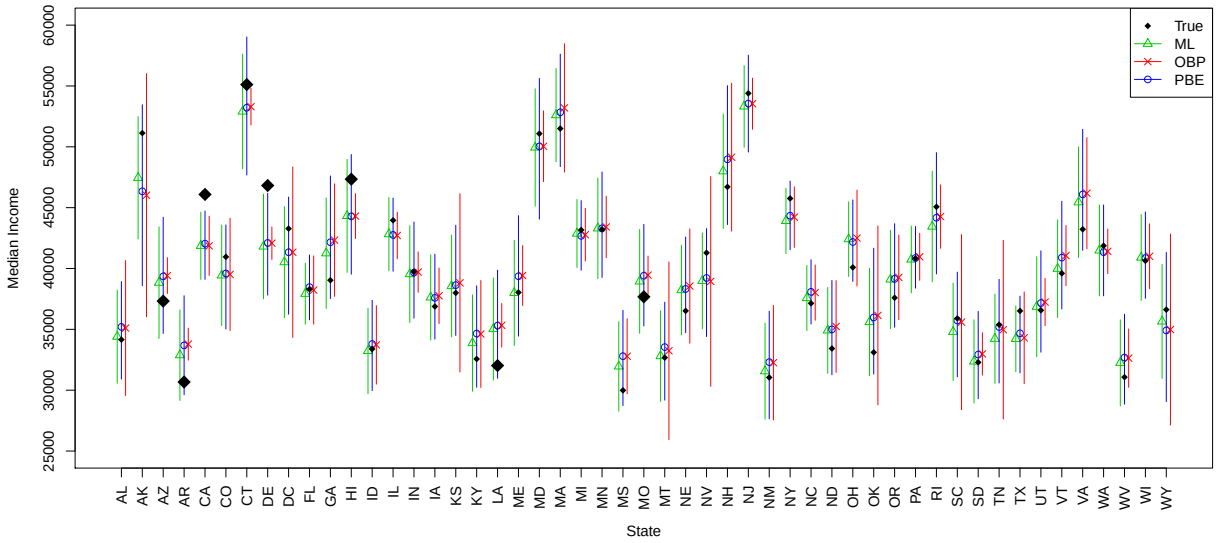

Figure 2.8: 95% confidence/credible intervals for θ_i based on the ML EBLUP, OBP, and PBE (median income data). The eight larger black diamonds indicate the true values the OBP confidence intervals fail to cover.

income data, the model has been used for the last few decades and is believed to be at least close to the truth. Of course, the argument in Section 2.8 is not entirely fair to the OBP since the OBP does not make the assumption that the underlying model is correctly specified, and the argument does not necessarily invalidate the use of (2.14).

We end this example by comparing 95% confidence/credible intervals for θ_i based on the ML EBLUP, OBP, and PBE. We choose the ML EBLUP since it performs the best among the EBLUPs according to Table 2.2. The confidence/credible intervals are shown in Figure 2.8. The black diamonds in Figure 2.8 indicate the true values and the eight larger ones are the ones the OBP confidence intervals fail to cover. They are the true values for the states of Arizona, Arkansas, California, Connecticut, Delaware, Hawaii, Louisiana, and Missouri. On the other hand, both the ML EBLUP confidence intervals and the PBE credible intervals miss only two states: California and Delaware, all due to underestimation.

2.6 Simulation Studies

A simulation study typically involves applying a proposed method repeatedly to different datasets. Since the number of repetitions is usually very high for accuracy, it is important to design the study as computationally efficient as possible. Recall the procedure proposed in Section 2.4 for selecting the tuning parameter λ in our pseudo-Bayesian method. While the procedure is sensible, it is really data-dependent and it can be computationally intensive when it comes to conducting a simulation study. Therefore, it is desirable to come up with a single reasonable value of λ for datasets under the same scenario in a simulation study. To this end, we investigate the effect of λ on the prior density $\pi_{\mathrm{m}}(\boldsymbol{\beta}^*, A^*) \propto \prod_{i=1}^{m}(A^*+D_i^*)^{-\lambda/2},\ \boldsymbol{\beta}^* \in \mathbb{R}^p,\ A^* > 0$, or, equivalently, on the prior density $\pi_{\mathrm{m}}(A^*) \propto \prod_{i=1}^{m}(A^* + D_i^*)^{-\lambda/2},\ A^* > 0$. Let $h(A^*;\lambda) = \prod_{i=1}^{m}(A^* + D_i^*)^{-\lambda/2},\ A^* > 0,\ \lambda > 0$, and note that

$$\frac{\partial}{\partial A^*}\log h(A^*;\lambda) = \frac{\partial}{\partial A^*}\left\{-\frac{\lambda}{2}\sum_{i=1}^{m}\log(A^* + D_i^*)\right\} = -\frac{\lambda}{2}\sum_{i=1}^{m}\frac{1}{A^* + D_i^*} < 0.$$

For $0 < \lambda_1 < \lambda_2$ and a given set of values of $D_i^*,\ i = 1, \ldots, m$,

$$\frac{\partial}{\partial A^*}\log h(A^*;\lambda_2) = -\frac{\lambda_2}{2}\sum_{i=1}^{m}\frac{1}{A^* + D_i^*} < -\frac{\lambda_1}{2}\sum_{i=1}^{m}\frac{1}{A^* + D_i^*} = \frac{\partial}{\partial A^*}\log h(A^*;\lambda_1) < 0$$

for all $A^* > 0$, which implies $\pi_{\mathrm{m}}(A^*)$ is decreasing and it decreases more rapidly as λ increases. In other words, as λ increases, $\pi_{\mathrm{m}}(A^*)$ assigns a higher probability to small values of A^* and hence becomes more informative. This is illustrated in Figure 2.9 using the hospital data and the median income data, where in each case $\log h(A^*;\lambda) = -\lambda/2 \cdot \sum_{i=1}^{m}\log(A^* + D_i^*)$ (i.e., $\log \pi_{\mathrm{m}}(A^*)$ apart from an additive constant) is plotted with three different values of λ.

Since we do not have any prior information on A^* in most cases, a large value of λ does not seem to be a reasonable choice unless the selection procedure in Section 2.4 identifies it as the optimal value. Therefore, for simulation purposes, we suggest taking $\lambda = 2(p+1.5)/m$ in each scenario, which is $1/m$ larger than its lower bound $2(p+1)/m$. In the following,

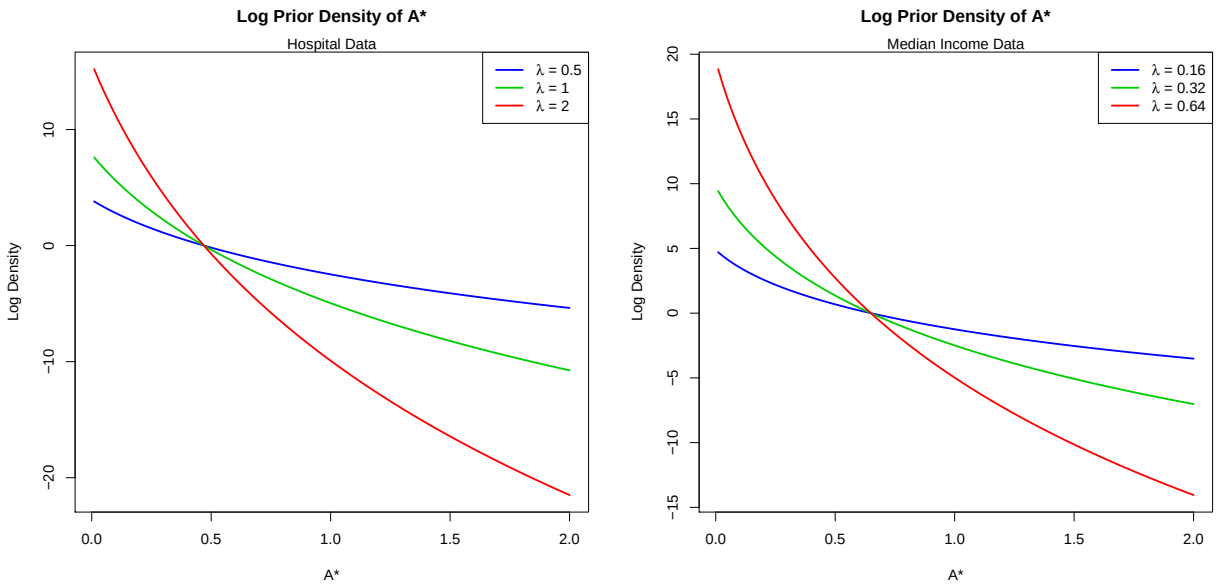

Figure 2.9: Log prior density of A^* for different values of λ (left: hospital data, right: median income data).

we consider two simulation studies similar to those conducted by Jiang et al. (2011) and apply our pseudo-Bayesian method to simulated data in addition to the existing methods (i.e., EBLUP and OBP). For both studies, we take $\lambda = 2(p+1.5)/m$ as the pseudo-Bayesian tuning parameter.

2.6.1 A Simple Example

In their first simulation study, Jiang et al. (2011) showed by a simple example that the gain of OBP over EBLUP can be substantial, if the underlying model is misspecified. Here, we conduct a similar simulation study to compare the performance of the PBE to that of the EBLUP and OBP under a model misspecification.

We consider a special case of the Fay-Herriot model (2.1). Suppose that the true under-

lying model is

$$y_i = \begin{cases} \mu_1 + v_i + e_i, & 1 \le i \le n \\ \mu_2 + v_i + e_i, & n+1 \le i \le m, \end{cases} \tag{2.16}$$

where $\mu_1 \neq \mu_2$, $m = 2n$, and the sampling variances are $D_i = \sigma_1^2,\ 1 \le i \le n$, and $D_i = \sigma_2^2,\ n+1 \le i \le m$. Further suppose that we actually fit the Fay-Herriot model (2.1) with $\mathbf{x}_i^T\boldsymbol{\beta} = \beta$, that is, we have a model misspecification by assuming $\mu_1 = \mu_2$. We set $\mu_1 = 0$, $\sigma_1^2 = 4$, $\sigma_2^2 = 1$, while $\mu_2 = 1$ or 5 so that we have a mild case and a severe case of model misspecification. We take $A = 0.2$ for the variance of v_i's and consider three different numbers of small areas: $m = 50$, 100, and 200. In each of the six scenarios (i.e., the six combinations of μ_2 and m), $K = 500$ simulation runs are carried out. Within each simulation run, a dataset is generated under the true underlying model (2.16), but a misspecified Fay-Herriot model with $\mathbf{x}_i^T\boldsymbol{\beta} = \beta$ is fitted using the EBLUP, OBP, and our pseudo-Bayesian method. Let $\boldsymbol{\theta}_{(k)} = (\theta_{i(k)})_{1\le i\le m}$, $k = 1, \ldots, K$, denote the true $\boldsymbol{\theta} = (\theta_i)_{1\le i\le m}$ from the kth simulation run, where $\theta_i = \mu_1 + v_i,\ 1 \le i \le n$, and $\theta_i = \mu_2 + v_i,\ n+1 \le i \le m$. Also, let $\breve{\boldsymbol{\theta}}_{(k)} = (\breve{\theta}_{i(k)})_{1\le i\le m}$ denote an estimator of $\boldsymbol{\theta}_{(k)}$. Then the empirical MSPE for that particular estimator is given by

$$\text{MSPE}^* = \frac{1}{K}\sum_{k=1}^{K} |\breve{\boldsymbol{\theta}}_{(k)} - \boldsymbol{\theta}_{(k)}|^2 = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} \{\breve{\theta}_{i(k)} - \theta_{i(k)}\}^2 = \sum_{i=1}^{m} \text{MSPE}_i^*,$$

where $\text{MSPE}_i^* = K^{-1}\sum_{k=1}^{K}\{\breve{\theta}_{i(k)} - \theta_{i(k)}\}^2$ is the empirical area-specific MSPE.

The empirical MSPEs are reported in Table 2.3, where the number in the parentheses is the percentage increase in MSPE by the EBLUP/OBP over the PBE, with negative percentage indicating decrease. Note that the EBLUPs perform very similarly with the FH EBLUP (resp. PR EBLUP) slightly better than the others when $\mu_2 = 1$ (resp. $\mu_2 = 5$). As demonstrated by Jiang et al. (2011), the OBP is substantially better than the EBLUPs, especially when $\mu_2 = 5$, that is, when there is a more serious model misspecification. The

Table 2.3: Empirical MSPEs with percentage increases over PBE (first simulation study).

μ_2	m	ML	REML	FH	PR	OBP	PBE
1	50	25.86	25.71	25.62	26.49	22.92	22.05
		(17%)	(17%)	(16%)	(20%)	(4.0%)	
1	100	47.95	47.72	47.16	48.20	40.24	39.23
		(22%)	(22%)	(20%)	(23%)	(2.6%)	
1	200	95.31	95.05	93.51	94.75	75.72	74.82
		(27%)	(27%)	(25%)	(27%)	(1.2%)	
5	50	97.20	96.38	94.75	94.27	67.68	68.81
		(41%)	(40%)	(38%)	(37%)	(−1.6%)	
5	100	186.87	186.21	183.76	182.96	131.56	132.47
		(41%)	(41%)	(39%)	(38%)	(−0.7%)	
5	200	378.94	378.24	372.90	371.00	261.43	262.42
		(44%)	(44%)	(42%)	(41%)	(−0.4%)	

same applies to the PBE with the percentage increase in MSPE by the EBLUPs ranging between 16% and 44%. The percentage increase in MSPE by the OBP, on the other hand, is as high as 4% when $\mu_2 = 1$ and it is negative when $\mu_2 = 5$, but becomes negligible as m increases.

Next, we compare the PBE, OBP, and EBLUPs in terms of area-specific MSPEs. Although the OBP is defined by minimizing the overall (observed) MSPE and the PBE is a pseudo-Bayesian alternative to it, there is no guarantee that their area-specific MSPEs are minimal as well. However, area-specific MSPEs are often of interest in small area estimation and it is important to compare those of the PBE, OBP, and EBLUP. Moreover, such comparisons may tell us the story behind the overall MSPEs reported in Table 2.3. The empirical area-specific MSPEs are summarized using boxplots and histograms in Figures 2.10 and 2.11. We have successfully reproduced the results of the OBP and EBLUPs presented in Jiang et al. (2011). For the PBE, the distribution of the empirical MSPEs is very similar to that of the OBP rather than those of the EBULPs. Specifically, regarding the boxplots,

not only do the PBE and OBP have smaller median empirical MSPEs in each case, but they also have much less variable empirical MSPEs compared to the EBLUPs.

The differences are more prominent from the histograms. While the histograms for the PBE and OBP are fairly close to normal, the histgrams for the EBLUPs are all bimodal with half of the empirical MSPEs slightly to moderately smaller than those of the PBE and OBP, but the other half much larger. As pointed out by Jiang et al. (2011), by fitting a misspecified model with one common mean when there are actually two different means evenly among the areas, the EBLUP sides with one mean and abandons the other, whereas the OBP balances between the two means. Our results here indicate that the PBE adopts the same strategy as the OBP and both of them are more robust to the model misspecification compared to the EBLUPs.

2.6.2 A Simulation Study Imitating the Hospital Data

The second simulation study conducted by Jiang et al. (2011) was an example imitating the hospital data. As before, we conduct a similar simulation study using our pseudo-Bayesian method.

Recall the potential outlier in the hospital data, which is the point at the upper right corner of the scatterplot in Figure 2.1. This outlying observation has led to the Q-O Fay-Herriot model

$$y_i = \theta_i + e_i,\ \theta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + d \cdot I(x_i > 0.3) + v_i,\ i = 1, \ldots, m, \tag{2.17}$$

for the hospital data in Section 2.5.1, where $v_i \overset{\text{iid}}{\sim} \mathrm{N}(0, A)$ independent of $e_i \overset{\text{ind}}{\sim} \mathrm{N}(0, D_i)$. We set up a simulation to investigate the outlying effect. We consider three different numbers of small areas: $m = 23$, 115, and 230. When $m = 23$, which is the case for the hospital data, $K = 5000$ simulation runs are carried out. Within each simulation run, a dataset is generated from the Q-O model (2.17) with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, d)^T = (-1.1, 20, -50, 0.9)^T$,

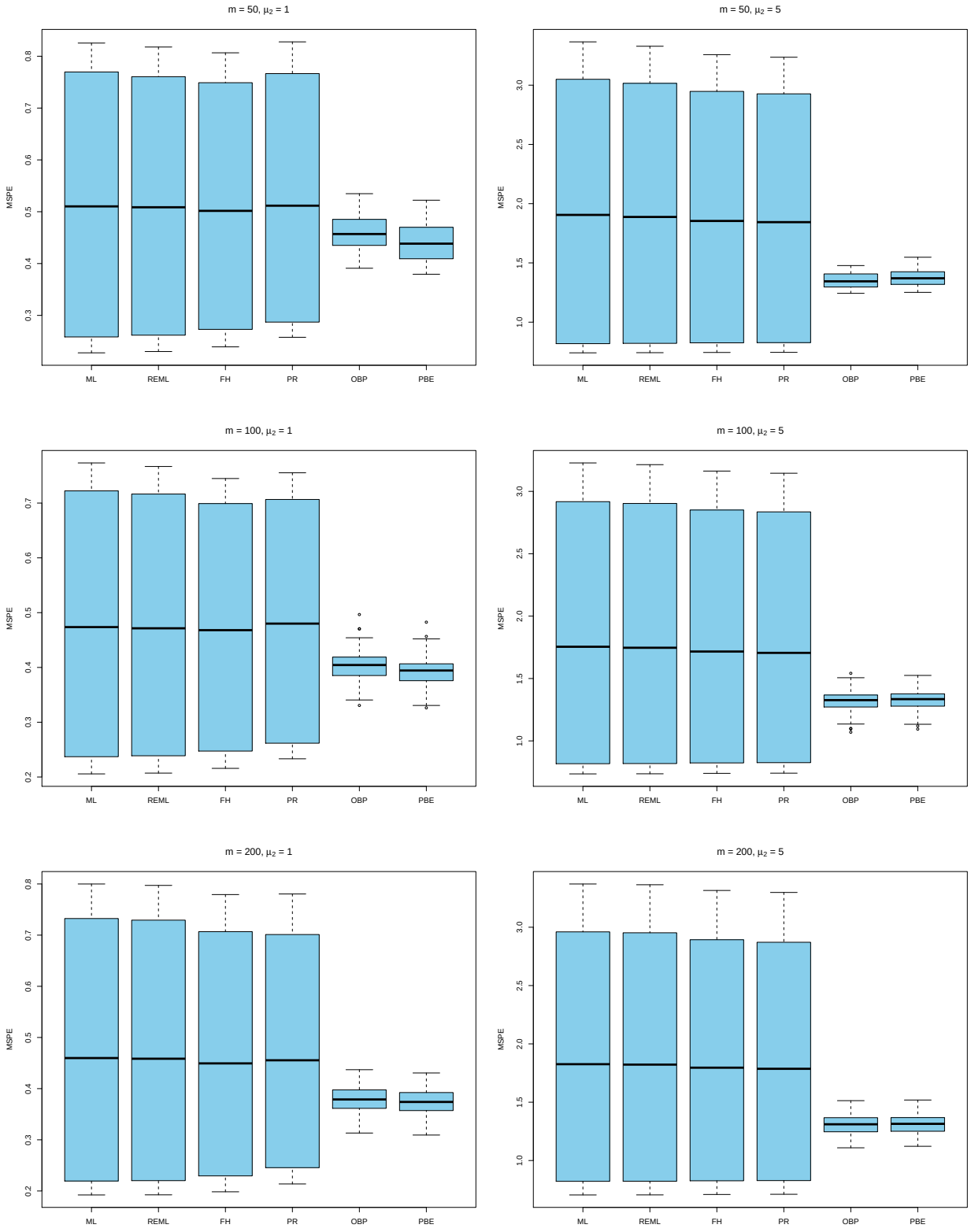

Figure 2.10: Boxplots of empirical area-specific MSPEs (first simulation study).

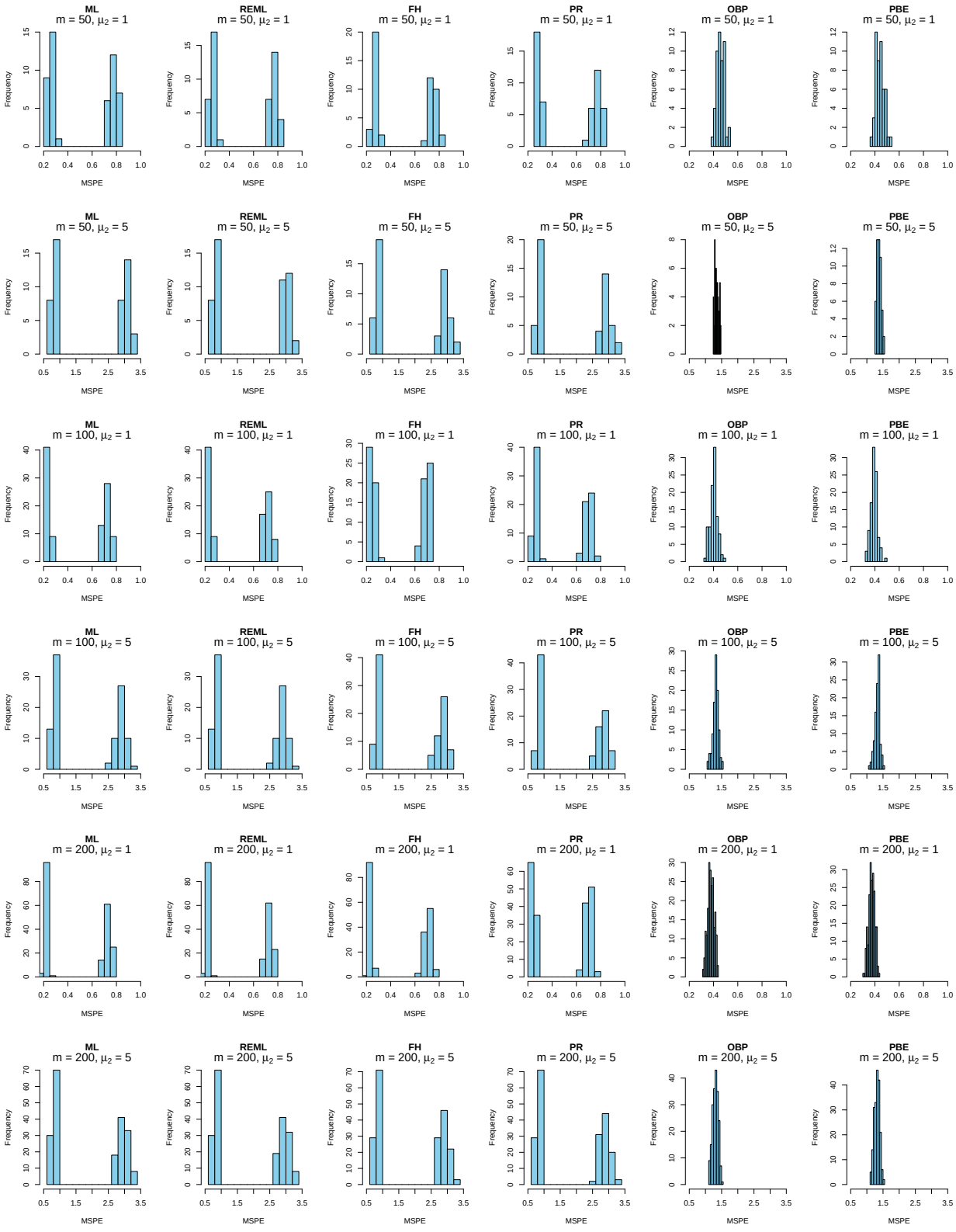

Figure 2.11: Histograms of empirical area-specific MSPEs (first simulation study).

$A = 0.0016$, and with x_i's and D_i's the same as those of the hospital data (see Table 2.1). The values for $(\boldsymbol{\beta}^T, A)^T$ were originally chosen by Jiang et al. (2011) so that, in a way, each simulated dataset mimics the hospital data. When $m = 115$ or 230, we replicate the design (i.e., the x_i's and D_i's) 5 or 10 times with $(\boldsymbol{\beta}^T, A)^T$ unchanged and carry out $K = 500$ simulation runs.

For the actual model to be fitted to the data, we consider two scenarios, namely, a misspecified case and a correctly specified case. In the misspecified case, a slightly misspecified model is fitted by omitting the term "$d \cdot I(x_i > 0.3)$" from the Q-O model (2.17), that is, a quadratic model is fitted. In the correctly specified case, the Q-O model (2.17) is fitted with no misspecification.

The empirical MSPEs (multiplied by 100) are reported in Table 2.4, where, as before, the number in the parentheses is the percentage increase in MSPE by the EBLUP/OBP over the PBE. When $m = 23$, the MSPE of the OBP is larger than those of the EBLUPs in both misspecified and correctly specified cases. This is not quite consistent with the results reported by Jiang et al. (2011), where the OBP outperformed the EBLUPs in the misspecified case. As for the PBE, when $m = 23$, its MSPE is smaller than that of the OBP in both misspecified and correctly specified cases, but it is still larger than those of the EBLUPs (except for the ML EBLUP in the correctly specified case). When $m = 115$ or 230, the OBP and PBE perform very similarly. They are equally better than the EBLUPs in the misspecified case and equally outperformed by the EBLUPs in the correctly specified case. Although this time the results are consistent with those reported by Jiang et al. (2011), the latter results are slightly more dramatic in favor of the OBP.

Next, we compute the empirical coverage probabilities of 95% confidence/credible intervals for θ_i's by averaging over the areas. The coverage probabilities are reported in Table 2.5, where the number in the parentheses is the average length of the intervals. While Jiang et al. (2011) have not reported such results, our results suggest the following. The OBP confidence intervals exhibit rather poor coverage probabilities, which are about 0.90 in the misspecified

Table 2.4: Empirical MSPEs (multiplied by 100) with percentage increases over PBE (second simulation study).

m	Model	ML	REML	FH	PR	OBP	PBE
23	Misspecified	2.925	2.870	2.868	2.876	2.981	2.954
		(−0.99%)	(−2.84%)	(−2.89%)	(−2.64%)	(0.93%)	
23	Correct	2.368	2.288	2.289	2.306	2.442	2.342
		(1.11%)	(−2.32%)	(−2.26%)	(−1.52%)	(4.28%)	
115	Misspecified	12.890	12.872	12.872	12.878	12.805	12.808
		(0.63%)	(0.49%)	(0.50%)	(0.55%)	(−0.03%)	
115	Correct	9.495	9.471	9.477	9.504	9.640	9.638
		(−1.48%)	(−1.73%)	(−1.67%)	(−1.39%)	(0.03%)	
230	Misspecified	25.544	25.531	25.533	25.540	25.277	25.287
		(1.01%)	(0.96%)	(0.97%)	(1.00%)	(−0.04%)	
230	Correct	18.602	18.592	18.607	18.642	18.774	18.771
		(−0.90%)	(−0.95%)	(−0.87%)	(−0.69%)	(0.01%)	

case and about 0.80 in the correctly specified case. Also, the OBP confidence intervals do not seem to be significantly shorter than the other intervals given the poor coverage. In contrast, the EBLUP and PBE intervals attain the 0.95 nominal coverage probability in both misspecified and correctly specified cases with the PBE credible intervals having slightly higher coverage probabilities and being slightly wider.

Overall, Tables 2.4 and 2.5 show that our pseudo-Bayesian method has an edge on the OBP. The PBE is very similar to or better than the OBP in terms of the MSPE and the interval estimator of the PBE remarkably outperforms that of the OBP. Also, recall that the second-order unbiased MSPE estimator $\widetilde{\text{MSPE}}^{\text{J}}(\widetilde{\theta}_i^{\text{OBP}})$ in (2.13) proposed by Jiang et al. (2011) can take a negative value and when it is negative, it is replaced by the bootstrap MSPE estimator $\widetilde{\text{MSPE}}_{\text{b}}^{\text{J}}(\widetilde{\theta}_i^{\text{OBP}})$ in (2.14), which is only first-order unbiased. For each fitted model in the simulation, the proportion of negative $\widetilde{\text{MSPE}}^{\text{J}}(\widetilde{\theta}_i^{\text{OBP}})$ values is calculated and boxplots of these proportions are shown in Figure 2.12. Among the K simulation runs in each scenario, where $K = 5000$ when $m = 23$ and $K = 500$ when $m = 115$ or 230, there are

Table 2.5: Empirical coverage probabilities of 95% confidence/credible intervals for θ_i's with average lengths of the intervals (second simulation study).

m	Model	ML	REML	FH	PR	OBP	PBE
23	Misspecified	0.9463	0.9470	0.9477	0.9481	0.8834	0.9606
		(0.1365)	(0.1356)	(0.1358)	(0.1362)	(0.1314)	(0.1486)
23	Correct	0.9401	0.9397	0.9409	0.9473	0.8343	0.9614
		(0.1242)	(0.1216)	(0.1220)	(0.1244)	(0.1177)	(0.1345)
115	Misspecified	0.9493	0.9494	0.9497	0.9497	0.8930	0.9526
		(0.1292)	(0.1291)	(0.1293)	(0.1294)	(0.1226)	(0.1320)
115	Correct	0.9466	0.9471	0.9472	0.9468	0.8049	0.9499
		(0.1114)	(0.1114)	(0.1115)	(0.1116)	(0.1007)	(0.1142)
230	Misspecified	0.9495	0.9495	0.9497	0.9496	0.8937	0.9511
		(0.1284)	(0.1284)	(0.1285)	(0.1286)	(0.1214)	(0.1297)
230	Correct	0.9486	0.9488	0.9485	0.9480	0.8056	0.9495
		(0.1103)	(0.1103)	(0.1102)	(0.1102)	(0.0990)	(0.1114)

many cases with very high proportions of negative $\widetilde{\mathrm{MSPE}}^{\mathrm{J}}(\widetilde{\theta}_i^{\mathrm{OBP}})$ values. In particular, in the correctly specified case, the median proportions are all about 0.2. On the other hand, the corresponding measure of uncertainty for the PBE of θ_i is simply the posterior variance, which is always nonnegative by definition.

Finally, for the area-specific MSPEs, we report the percentage of small areas for which the empirical area-specific MSPEs of the PBE are smaller than those of the EBLUPs and OBP. The percentages are presented in Table 2.6. When $m = 23$, the PBE does better than the OBP in most of the small areas in both misspecified and correctly specified cases; however, the PBE is outperformed by the EBLUPs, especially in the correctly specified case (except for the ML EBLUP). When $m = 115$ or 230, the PBE is slightly outperformed by the OBP except for the correctly specified case when $m = 115$; the PBE does better than the EBLUPs in more than half of the small areas in the misspecified case, but again, the EBLUPs dominate the PBE in the correctly specified case. Nevertheless, the overall advantage of the

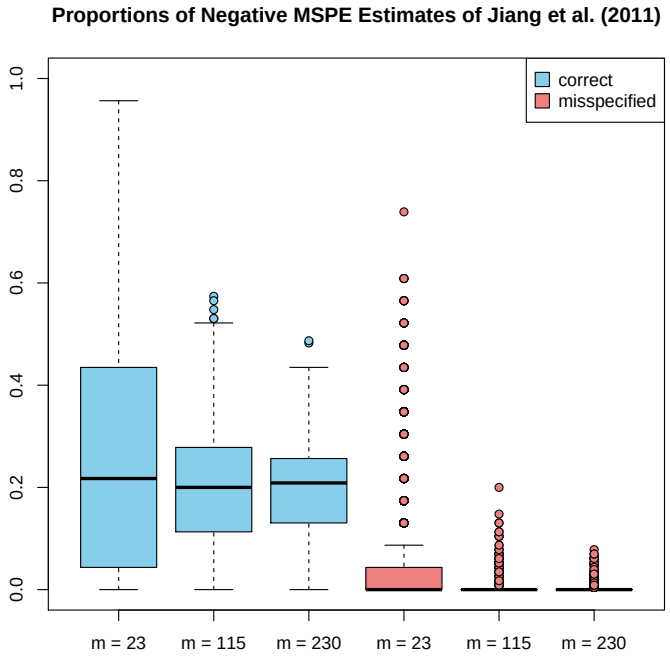

Figure 2.12: Boxplots of proportions of negative $\widetilde{\text{MSPE}}^{\text{J}}(\widetilde{\theta}_i^{\text{OBP}})$ values (second simulation study).

Table 2.6: Percentage of small areas for which the empirical area-specific MSPEs of the PBE are smaller than those of the EBLUPs and OBP (second simulation study).

m	Model	ML	REML	FH	PR	OBP
23	Misspecified	43%	48%	48%	48%	74%
	Correct	61%	22%	17%	30%	96%
115	Misspecified	50%	52%	52%	52%	37%
	Correct	10%	7%	6%	10%	57%
230	Misspecified	56%	53%	53%	52%	41%
	Correct	13%	12%	11%	16%	46%

EBLUP over PBE vanishes as m increases, as shown in Table 2.4.

2.7 Conclusions

The OBP is a new prediction procedure proposed by Jiang et al. (2011) that is different from the traditional EBLUP. Jiang et al. (2011) derived the OBP under the Fay-Herriot model by minimizing the observed MSPE and showed that the OBP is more robust to model misspecifications than the EBLUP.

In this chapter, we proposed a pseudo-Bayesian alternative to the OBP by converting the objective function, which the BPE minimizes to obtain the OBP, to a likelihood function. We introduced a prior density for the model parameters, where the prior density depends on a tuning parameter. Then we derived a condition, which is actually a lower bound, on the tuning parameter that guarantees the propriety of the resulting posterior density. Furthermore, we proposed a procedure for selecting the tuning parameter. However, the selection procedure is data-dependent and can be computationally intensive. We also note that a larger value of the tuning parameter amounts to more prior information on the variance component of the model parameters. Consequently, when the selection procedure is not feasible for any practical reasons, we suggest taking a small value that slightly exceeds the

lower bound as the tuning parameter.

Using real data examples and simulations, we demonstrated that the PBE is very similar to or better than the OBP in terms of the overall and area-specific MSPEs. One major advantage of the PBE over the OBP is that the PBE has a nonnegative measure of uncertainty (i.e., posterior variance), whereas the second-order unbiased MSPE estimator for the OBP proposed by Jiang et al. (2011) can take a negative value. For example, the MSPE estimates were negative for 9 out of 51 states in the median income data example. In the second simulation study, the PBE credible intervals attained the 0.95 nominal coverage probability, whereas the coverage probabilities of the OBP confidence intervals were all below 0.90 with only minor decreases in the lengths of the intervals. In summary, our pseudo-Bayesian method enjoys both the Bayesian and frequentist properties.

2.8 Appendix

We show that under the correctly specified Fay-Herriot model, a bootstrap MSPE estimator of the form (2.14) tends to underestimate the actual MSPE.

Proof. Suppose the Fay-Herriot model (2.1) is correctly specified. Then the BP of θ_i is $\widetilde{\theta}_i(\boldsymbol{\beta}, A) = \mathbf{x}_i^T\boldsymbol{\beta} + B_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta})$, where $B_i = A/(A+D_i)$. When A is known, but $\boldsymbol{\beta}$ is unknown, we replace $\boldsymbol{\beta}$ in $\widetilde{\theta}_i(\boldsymbol{\beta}, A)$ with $\check{\boldsymbol{\beta}}(A)$, the MLE or BPE of $\boldsymbol{\beta}$, and obtain $\widetilde{\theta}_i(\check{\boldsymbol{\beta}}(A), A)$. Note that $\widetilde{\theta}_i(\check{\boldsymbol{\beta}}(A), A)$ is the BLUP in case $\check{\boldsymbol{\beta}}(A)$ is the MLE of $\boldsymbol{\beta}$. When A is also unknown, we replace A in $\widetilde{\theta}_i(\check{\boldsymbol{\beta}}(A), A)$ with an estimator $\check{A}$ and obtain $\widetilde{\theta}_i(\check{\boldsymbol{\beta}}(\check{A}), \check{A})$. Note that $\widetilde{\theta}_i(\check{\boldsymbol{\beta}}(\check{A}), \check{A})$ becomes the EBLUP when $\check{A}$ is the ML, REML, FH, or PR estimator and $\check{\boldsymbol{\beta}}(\check{A})$ is the MLE of $\boldsymbol{\beta}$ given $\check{A}$. When $(\{\check{\boldsymbol{\beta}}(\check{A})\}^T, \check{A})^T$ is the BPE of $(\boldsymbol{\beta}^T, A)^T$, on the other hand, $\widetilde{\theta}_i(\check{\boldsymbol{\beta}}(\check{A}), \check{A})$ becomes the OBP. Now, recall that $\mathrm{Var}(\theta_i) = \mathrm{Var}(v_i) = A$ and the leading term of $\mathrm{MSPE}[\widetilde{\theta}_i(\check{\boldsymbol{\beta}}(\check{A}), \check{A})] = \mathrm{E}[\{\widetilde{\theta}_i(\check{\boldsymbol{\beta}}(\check{A}), \check{A}) - \theta_i\}^2]$ is $\mathrm{E}[\{\widetilde{\theta}_i(\boldsymbol{\beta}, A) - \theta_i\}^2] = AD_i/(A + D_i)$ (see,

e.g., Datta et al., 2005). Note that

$$\mathrm{Var}[\widetilde{\theta}_i(\boldsymbol{\beta}, A)] = \mathrm{Var}[\mathbf{x}_i^T\boldsymbol{\beta} + B_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta})] = \mathrm{E}[B_i^2(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2] = B_i^2(A + D_i) = B_iA,$$

$$\begin{aligned}
\mathrm{Var}[\widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A)] =& \mathrm{Var}[\widetilde{\theta}_i(\boldsymbol{\beta}, A) + \widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A) - \widetilde{\theta}_i(\boldsymbol{\beta}, A)] \\
=& \mathrm{Var}[\widetilde{\theta}_i(\boldsymbol{\beta}, A) + (1 - B_i)\mathbf{x}_i^T(\breve{\boldsymbol{\beta}}(A) - \boldsymbol{\beta})] \\
=& \mathrm{Var}[\widetilde{\theta}_i(\boldsymbol{\beta}, A)] + (1 - B_i)^2\mathbf{x}_i^T\mathrm{Var}[\breve{\boldsymbol{\beta}}(A)]\mathbf{x}_i \\
&+ 2(1 - B_i)\mathrm{Cov}[\widetilde{\theta}_i(\boldsymbol{\beta}, A), \mathbf{x}_i^T\breve{\boldsymbol{\beta}}(A)] \\
=& B_iA + (1 - B_i)^2\mathbf{x}_i^T\mathrm{Var}[\breve{\boldsymbol{\beta}}(A)]\mathbf{x}_i + 2B_i(1 - B_i)\mathrm{Cov}[y_i, \mathbf{x}_i^T\breve{\boldsymbol{\beta}}(A)],
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}[\widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(\breve{A}), \breve{A})] =& \mathrm{Var}[\widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A) + \widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(\breve{A}), \breve{A}) - \widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A)] \\
=& \mathrm{Var}[\widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A)] + \mathrm{Var}[\widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(\breve{A}), \breve{A}) - \widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A)] \\
&+ 2\mathrm{Cov}[\widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A), \widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(\breve{A}), \breve{A}) - \widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A)] \\
=& B_iA + (1 - B_i)^2\mathbf{x}_i^T\mathrm{Var}[\breve{\boldsymbol{\beta}}(A)]\mathbf{x}_i + 2B_i(1 - B_i)\mathrm{Cov}[y_i, \mathbf{x}_i^T\breve{\boldsymbol{\beta}}(A)] \\
&+ \mathrm{Var}[\widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(\breve{A}), \breve{A}) - \widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A)] \\
&+ 2\mathrm{Cov}[\widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A), \widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(\breve{A}), \breve{A}) - \widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(A), A)] \\
=& B_iA + O(m^{-1}).
\end{aligned}$$

This implies under the correctly specified Fay-Herriot model, the leading term of the expectation of a bootstrap MSPE estimator of the form (2.14) is $B_iAD_i/(B_iA + D_i)$, which is smaller than $AD_i/(A + D_i)$, the leading term of $\mathrm{MSPE}[\widetilde{\theta}_i(\breve{\boldsymbol{\beta}}(\breve{A}), \breve{A})]$. □

Chapter 3

Compromise Pseudo-Bayesian Small Area Estimation

3.1 Introduction

This chapter is an extension to Chapter 2, where we discussed the observed best predictor (OBP) proposed by Jiang et al. (2011) and proposed a pseudo-Bayesian alternative to the OBP. The alternative called the pseudo-Bayesian estimator (PBE) is obtained by converting the objective function, which the best predictive estimator (BPE) minimizes to obtain the OBP, to a likelihood function. For both the OBP and PBE, the assumed model is the Fay-Herriot model (Fay and Herriot, 1979):

$$y_i = \theta_i + e_i,\ \theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + v_i,\ i = 1, \ldots, m, \tag{3.1}$$

where y_i is a direct estimator of the ith small area mean θ_i, $\mathbf{x}_i$ is a $p \times 1$ vector of known covariates, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients. Here, v_i's are area-specific random effects and e_i's are sampling errors. It is assumed that $v_i \overset{\text{iid}}{\sim} \text{N}(0, A)$ independent of $e_i \overset{\text{ind}}{\sim} \text{N}(0, D_i)$, where the variance A is unknown, but the sampling variances D_i's are treated as known. Under the Fay-Herriot model (3.1), the likelihood function for the parameters $\boldsymbol{\beta}$

and A is

$$\begin{aligned} L_1(\boldsymbol{\beta}, A|\mathbf{y}) &\propto |\mathbf{V}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})] \\ &= |\mathbf{W}_1|^{\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_1(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})], \end{aligned} \tag{3.2}$$

where $\mathbf{y} = (y_i)_{1\le i\le m}$, $\mathbf{X} = (\mathbf{x}_i^T)_{1\le i\le m}$, $\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \mathrm{diag}(A + D_i,\ 1 \le i \le m)$, and $\mathbf{W}_1 = \mathbf{V}^{-1} = \mathrm{diag}((A + D_i)^{-1},\ 1 \le i \le m)$. Without loss of generality, the $m \times p$ design matrix $\mathbf{X}$ is assumed to be of full column rank. Given A, the likelihood function $L_1(\boldsymbol{\beta}, A|\mathbf{y})$ yields the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}_1\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_1\mathbf{y} = \left(\sum_{i=1}^m \frac{\mathbf{x}_i\mathbf{x}_i^T}{A+D_i}\right)^{-1} \sum_{i=1}^m \frac{\mathbf{x}_i y_i}{A+D_i}.$$

On the other hand, the OBP objective function to be minimized is

$$Q(\boldsymbol{\beta}, A, \mathbf{y}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Gamma}^2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + 2A \cdot \mathrm{tr}(\boldsymbol{\Gamma}),$$

which is obtained without assuming the mean function $\mathbf{x}_i^T\boldsymbol{\beta}$ of the Fay-Herriot model (3.1) is correctly specified, where $\boldsymbol{\Gamma} = \mathrm{diag}(D_i/(A + D_i),\ 1 \le i \le m)$. Given A, the objective function $Q(\boldsymbol{\beta}, A, \mathbf{y})$ is minimized by the BPE of $\boldsymbol{\beta}$

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Gamma}^2\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Gamma}^2\mathbf{y} = \left\{\sum_{i=1}^m \left(\frac{D_i}{A+D_i}\right)^2 \mathbf{x}_i\mathbf{x}_i^T\right\}^{-1} \sum_{i=1}^m \left(\frac{D_i}{A+D_i}\right)^2 \mathbf{x}_i y_i.$$

Note that both the MLE $\widehat{\boldsymbol{\beta}}$ and the BPE $\widetilde{\boldsymbol{\beta}}$ are in the class of weighted least squares (WLS) estimators of $\boldsymbol{\beta}$ that can be expressed as

$$\breve{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

for some positive definite weight matrix $\mathbf{W}$. Clearly, $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ adopt two different weighting schemes. For the MLE $\widehat{\boldsymbol{\beta}}$, $\mathbf{W} = \text{diag}(w_i^{\text{MLE}} \propto (A + D_i)^{-1},\ 1 \le i \le m)$, whereas for the BPE $\widetilde{\boldsymbol{\beta}}$, $\mathbf{W} = \text{diag}(w_i^{\text{BPE}} \propto \{D_i/(A + D_i)\}^2,\ 1 \le i \le m)$. Here, $w_i^{\text{MLE}} = (A + D_i)^{-1}/\sum_{h=1}^{m}(A + D_h)^{-1}$ and $w_i^{\text{BPE}} = \{D_i/(A + D_i)\}^2/\sum_{h=1}^{m}\{D_h/(A + D_h)\}^2$ are the MLE and BPE weights, respectively.

Given A, $\text{Var}(\widetilde{\boldsymbol{\beta}}) \ge \text{Var}(\widehat{\boldsymbol{\beta}})$ (i.e., $\text{Var}(\widetilde{\boldsymbol{\beta}}) - \text{Var}(\widehat{\boldsymbol{\beta}})$ is nonnegative definite) and this implies the BPE $\widetilde{\boldsymbol{\beta}}$ cannot be more efficient than the MLE $\widehat{\boldsymbol{\beta}}$ when the mean function $\mathbf{x}_i^T\boldsymbol{\beta}$ is correctly specified (Jiang et al., 2011). However, when the problem of interest is estimation of the small area means $\theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + v_i,\ i = 1, \ldots, m$, the BPE $\widetilde{\boldsymbol{\beta}}$ makes logical sense. The reason is the following. Recall that given A, one replaces $\boldsymbol{\beta}$ in the best predictor (BP) $\widetilde{\theta}_i(\boldsymbol{\beta}, A) = \mathbf{x}_i^T\boldsymbol{\beta} + B_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta})$ with its MLE or BPE to estimate θ_i, where $B_i = A/(A + D_i)$. The BPE $\widetilde{\boldsymbol{\beta}}$ assigns more weights to small areas with larger sampling variances D_i's so that those areas play a greater role in estimating $\boldsymbol{\beta}$, and it is those areas that are shrunken more toward the synthetic estimator $\mathbf{x}_i^T\boldsymbol{\beta}$. In fact, it can be shown that given A, the mean squared prediction error (MSPE) of $\widetilde{\boldsymbol{\theta}}(\check{\boldsymbol{\beta}}, A) = [\widetilde{\theta}_i(\check{\boldsymbol{\beta}}, A)]_{1 \le i \le m}$, where $\widetilde{\theta}_i(\check{\boldsymbol{\beta}}, A)$ is an empirical best predictor (EBP) of θ_i with a WLS estimator $\check{\boldsymbol{\beta}}$, can be decomposed into a model misspecification term and a variance term, which are minimized by the BPE weights and the MLE weights, respectively (Jiang et al., 2011).

This motivates us to consider weights that compromise between the MLE weights and the BPE weights. In the next section, we use convex combinations of weights that are proportional to the MLE and BPE weights to define two new likelihood functions. Then for each likelihood function, using a suitable prior density, we propose a compromise pseudo-Bayesian estimator (CPBE) of the small area means.

3.2 Compromise Pseudo-Bayesian Estimators

3.2.1 First Approach

Recall the likelihood function $L_1(\boldsymbol{\beta}, A|\mathbf{y})$ in (3.2) and the corresponding weight matrix $\mathbf{W}_1 = \text{diag}((A+D_i)^{-1},\ 1 \le i \le m)$, whose diagonal elements are proportional to the MLE weights $w_i^{\text{MLE}} = (A+D_i)^{-1}/\sum_{h=1}^{m}(A+D_h)^{-1}$. In addition to $L_1(\boldsymbol{\beta}, A|\mathbf{y})$, we define another likelihood function

$$L_2(\boldsymbol{\beta}, A|\mathbf{y}) \propto |\mathbf{W}_2|^{\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})],$$

where $\mathbf{W}_2 = \text{diag}(\{D_i/(A+D_i)\}^2/(A+\overline{D}),\ 1 \le i \le m)$ with $\overline{D} = m^{-1}\sum_{i=1}^{m} D_i$. Note that the diagonal elements of the weight matrix $\mathbf{W}_2$ are proportional to the BPE weights $w_i^{\text{BPE}} = \{D_i/(A+D_i)\}^2/\sum_{h=1}^{m}\{D_h/(A+D_h)\}^2$, where the common factor $(A+\overline{D})^{-1}$ in $\mathbf{W}_2$ is to avoid the scaling issue discussed in Section 2.2. As a result, the likelihood function $L_2(\boldsymbol{\beta}, A|\mathbf{y})$ is maximized at the BPE $\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_2\mathbf{y} = (\mathbf{X}^T\boldsymbol{\Gamma}^2\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Gamma}^2\mathbf{y}$ given A. Now, using $L_1(\boldsymbol{\beta}, A|\mathbf{y})$ and $L_2(\boldsymbol{\beta}, A|\mathbf{y})$, we define our first compromise likelihood function as

$$\begin{aligned}
L_{\text{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y}) =& L_1^{\alpha}(\boldsymbol{\beta}, A|\mathbf{y})L_2^{1-\alpha}(\boldsymbol{\beta}, A|\mathbf{y}) \\
\propto& \{|\mathbf{W}_1|^{\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_1(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]\}^{\alpha} \\
& \times \{|\mathbf{W}_2|^{\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]\}^{1-\alpha} \\
=& |\mathbf{W}_1|^{\frac{\alpha}{2}}|\mathbf{W}_2|^{\frac{1-\alpha}{2}} \exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\{\alpha\mathbf{W}_1 + (1-\alpha)\mathbf{W}_2\}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})] \\
=& |\mathbf{W}_1|^{\frac{\alpha}{2}}|\mathbf{W}_2|^{\frac{1-\alpha}{2}} \exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_{\text{c},1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})],
\end{aligned}$$

where $\alpha \in [0,1]$ is a mixing parameter and $\mathbf{W}_{\text{c},1} = \alpha\mathbf{W}_1 + (1-\alpha)\mathbf{W}_2$ is our first compromise weight matrix. Given A and α, the likelihood function $L_{\text{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$ is maximized at

$\widetilde{\boldsymbol{\beta}}_{\mathrm{c},1} = (\mathbf{X}^T\mathbf{W}_{\mathrm{c},1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_{\mathrm{c},1}\mathbf{y}$. Based on $L_{\mathrm{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$ and an improper prior density $\pi(\boldsymbol{\beta}, A, \alpha) \propto A^{-b},\ \boldsymbol{\beta} \in \mathbb{R}^p,\ A > 0,\ 0 \leq \alpha \leq 1$, we obtain the posterior density

$$\begin{aligned}\pi_{\mathrm{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y}) &\propto L_{\mathrm{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})\pi(\boldsymbol{\beta}, A, \alpha) \\ &\propto A^{-b}|\mathbf{W}_1|^{\frac{\alpha}{2}}|\mathbf{W}_2|^{\frac{1-\alpha}{2}} \exp[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_{\mathrm{c},1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]. \qquad (3.3)\end{aligned}$$

A suitable condition on the prior parameter b is needed to achieve the propriety of the posterior density $\pi_{\mathrm{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$. The condition is specified in the following theorem and the proof is given in Section 3.6.1.

Theorem 3.1. *The posterior density* $\pi_{\mathrm{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$ *in* (3.3) *is proper if* $1 - (m - p)/2 < b < 1$.

Now, recall that under the Fay-Herriot model (3.1),

$$\theta_i|\boldsymbol{\beta}, A, \mathbf{y} \overset{\text{ind}}{\sim} \mathrm{N}\left(\mathbf{x}_i^T\boldsymbol{\beta} + \frac{A}{A + D_i}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}), \frac{AD_i}{A + D_i}\right).$$

The CPBE of $\theta_i,\ i = 1, \ldots, m$, using our first approach is obtained by the conditional posterior mean of θ_i given $\boldsymbol{\beta}$, A, and α, whose posterior density is given by $\pi_{\mathrm{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$. We denote this CPBE as CPBE_1.

3.2.2 Second Approach

For our second approach, we define a new diagonal weight matrix such that each diagonal element is a convex combination of the MLE weight and the BPE weight. Specifically, for $\alpha \in [0, 1]$, we define our second compromise weight matrix as $\mathbf{W}_{\mathrm{c},2} = \mathrm{diag}(w_i^{\mathrm{c}},\ 1 \leq i \leq m)$, where

$$\begin{aligned}w_i^{\mathrm{c}} &= \alpha \cdot w_i^{\mathrm{MLE}} + (1 - \alpha) \cdot w_i^{\mathrm{BPE}} \\ &= \alpha \cdot \frac{(A + D_i)^{-1}}{\sum_{h=1}^m (A + D_h)^{-1}} + (1 - \alpha) \cdot \frac{\{D_i/(A + D_i)\}^2}{\sum_{h=1}^m \{D_h/(A + D_h)\}^2}.\end{aligned}$$

Henderson et al. (2020) used w_i^{c}'s and derived a frequentist estimator called the compromise best predictor (CBP) that compromises between the EBLUP and OBP. Note that $\mathbf{W}_{\mathrm{c},2}$ is different from our first compromise weight matrix $\mathbf{W}_{\mathrm{c},1} = \alpha\mathbf{W}_1 + (1-\alpha)\mathbf{W}_2$, which is also diagonal, but with the diagonal elements $\alpha(A+D_i)^{-1} + (1-\alpha)\{D_i/(A+D_i)\}^2/(A+\overline{D})$, $i = 1, \ldots, m$. Now, using $\mathbf{W}_{\mathrm{c},2}$, we define our second compromise likelihood function as

$$\begin{aligned} L_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y}) &\propto |\mathbf{V}_{\mathrm{c},2}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}_{\mathrm{c},2}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})] \\ &= \left|\frac{m}{A+\overline{D}}\mathbf{W}_{\mathrm{c},2}\right|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\left\{\frac{m}{A+\overline{D}}\mathbf{W}_{\mathrm{c},2}\right\}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] \\ &\propto (A+\overline{D})^{-\frac{m}{2}}|\mathbf{W}_{\mathrm{c},2}|^{\frac{1}{2}} \exp\left[-\frac{m}{2(A+\overline{D})}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_{\mathrm{c},2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right], \end{aligned}$$

where $\mathbf{V}_{\mathrm{c},2} = \{m/(A+\overline{D})\cdot\mathbf{W}_{\mathrm{c},2}\}^{-1}$. Here, the factor $m/(A+\overline{D})$ is multiplied to $\mathbf{W}_{\mathrm{c},2}$ because $\mathbf{V}_{\mathrm{c},2}$ should behave like an alternative to the covariance matrix $\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \mathrm{diag}(A + D_i,\ 1 \le i \le m)$ and w_i^{c}'s, the diagonal elements of $\mathbf{W}_{\mathrm{c},2}$, are $O(m^{-1})$. Given A and α, the likelihood function $L_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$ is maximized at $\widetilde{\boldsymbol{\beta}}_{\mathrm{c},2} = (\mathbf{X}^T\mathbf{W}_{\mathrm{c},2}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_{\mathrm{c},2}\mathbf{y}$. Based on $L_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$ and again the improper prior density $\pi(\boldsymbol{\beta}, A, \alpha) \propto A^{-b},\ \boldsymbol{\beta} \in \mathbb{R}^p,\ A > 0,\ 0 \le \alpha \le 1$, we obtain the posterior density

$$\begin{aligned} \pi_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y}) &\propto L_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})\pi(\boldsymbol{\beta}, A, \alpha) \\ &\propto A^{-b}(A+\overline{D})^{-\frac{m}{2}}|\mathbf{W}_{\mathrm{c},2}|^{\frac{1}{2}} \exp\left[-\frac{m}{2(A+\overline{D})}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_{\mathrm{c},2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]. \end{aligned} \tag{3.4}$$

The condition on the prior parameter b to make the posterior density $\pi_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$ proper is specified in the following theorem and the proof is given in Section 3.6.2.

Theorem 3.2. *The posterior density* $\pi_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$ *in* (3.4) *is proper if* $1 - (m-p)/2 < b < 1$.

As before, the CPBE of $\theta_i,\ i = 1, \ldots, m$, using our second approach is obtained by the conditional posterior mean of θ_i given $\boldsymbol{\beta}$, A, and α, but with the posterior density

$\pi_{c,2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$. We denote this CPBE as CPBE_2.

3.3 Real Data Examples

In Chapter 2, we applied our pseudo-Bayesian alternative to the OBP to two real datasets, namely, the hospital data and the median income data. In this section, we illustrate our compromise pseudo-Bayesian methods using the same datasets.

3.3.1 Hospital Data

The hospital data contain data collected from $m = 23$ hospitals, which are treated as small areas. The responses $y_i,\ i = 1, \ldots, m$, are the graft failure rates for kidney transplant operations and one available explanatory variable is the severity index x_i. See Table 2.1 for the entire data and Figure 2.1 for a scatterplot of the data.

Recall that in Chapter 2, we fit the quadratic-outlying (Q-O) model proposed by Jiang et al. (2011) to the hospital data. The Q-O model is a Fay-Herriot model with the mean function $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + d \cdot I(x_i > 0.3)$, where $I(\cdot)$ is the indicator function. Here, we fit the Q-O model to the hospital data using our compromise pseudo-Bayesian methods. Two popular choices for the prior parameter b are 0 and 0.5, where both of them satisfy the propriety condition $1 - (m - p)/2 < b < 1$ specified in Theorems 3.1 and 3.2 with $m = 23$ and $p = 4$ ($\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, d)^T$). For the hospital data, it turns out that $b = 0.5$ works better than $b = 0$ in terms of stability of the estimates in both cases of CPBE_1 and CPBE_2. Hence, we report the results based on $b = 0.5$.

Histograms of posterior simulations of the mixing parameter α for CPBE_1 and CPBE_2 are shown in the left and right panels of Figure 3.1, respectively. Clearly, both histograms are left skewed indicating both CPBE_1 and CPBE_2 tend to assign more weights to the MLE than the BPE in estimating $\boldsymbol{\beta}$. The posterior mean of α is 0.68 for CPBE_1 and is 0.66 for CPBE_2 as reported in Table 3.1. Other estimates of the model parameters based on different

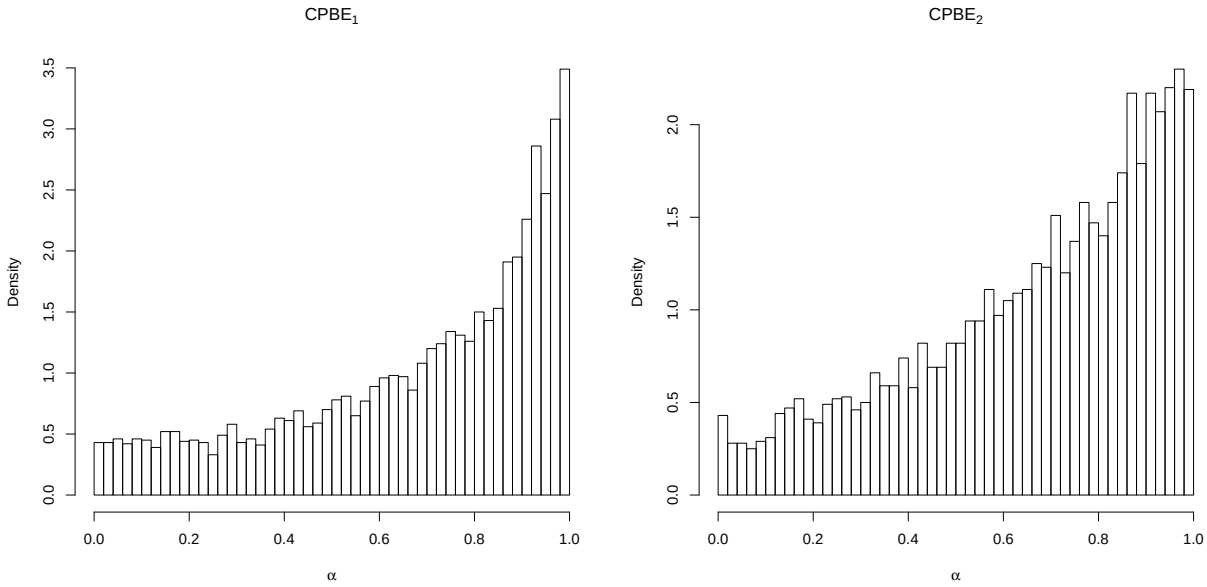

Figure 3.1: Histograms of posterior simulations of α for CPBE_1 (left) and CPBE_2 (right) (hospital data).

methods are also reported in Table 3.1, where "ML," "REML," "FH," and "PR" are the four EBLUPs discussed in Chapter 2. Note that the ML estimate of A is substantially smaller than the other estimates of A. Except that, all estimates seem roughly consistent with each other.

Now, we compare the small area mean estimates as well as their uncertainty estimates. For the frequentist estimates (i.e., EBLUPs and OBP), uncertainty is measured by the square roots of the MSPE estimates (RMSPE), which are discussed in Chapter 2. For the Bayesian estimates (i.e., PBE and CPBEs), on the other hand, uncertainty is measured by the posterior standard deviations. Among the four EBLUPs, we choose the ML EBLUP for comparison since it has the smallest average RMSPE estimate as reported in Table 3.2. Overall, however, CPBE_1 is the most stable with an average posterior standard deviation (APSD) of 0.023. The small area mean estimates and their uncertainty estimates are shown in Figure 3.2. The small area mean estimates are mostly consistent with each other, but

Table 3.1: Estimates of the model parameters (hospital data).

Estimate	α	A	β_0	β_1	β_2	d
ML	–	2.9×10^{-5}	-0.015	3.247	-11.014	0.519
REML	–	4.0×10^{-4}	-0.025	3.437	-11.701	0.543
FH	–	5.9×10^{-4}	-0.029	3.504	-11.944	0.551
PR	–	7.6×10^{-4}	-0.031	3.555	-12.130	0.558
OBP	–	3.4×10^{-4}	-0.084	4.614	-16.045	0.698
PBE	–	2.1×10^{-4}	-0.077	4.475	-15.518	0.679
CPBE_1	0.68	4.2×10^{-4}	-0.039	3.706	-12.691	0.577
CPBE_2	0.66	6.6×10^{-4}	-0.049	3.906	-13.425	0.603

Table 3.2: Averages of the uncertainty estimates of the small area mean estimates (hospital data). APSD stands for average posterior standard deviation.

Average RMSPE estimate					APSD		
ML	REML	FH	PR	OBP	PBE	CPBE_1	CPBE_2
0.026	0.026	0.028	0.030	0.032	0.025	0.023	0.025

the uncertainty estimates vary substantially. The OBP RMSPE estimates are very unstable across the areas. The ML RMSPE estimates are much more stable, but they are still larger than the posterior standard deviations of the Bayesian estimates on average as reported in Table 3.2. In particular, Figure 3.2 shows that the posterior standard deviations of CPBE_1 are smaller than the other uncertainty estimates in most areas.

3.3.2 Median Income Data

We revisit the median income data discussed in Chapter 2. Recall that our goal is to estimate the four-person family median incomes of the 50 U.S. states and the District of Columbia (i.e., $m = 51$ small areas) for the year 1989. The response $y_i,\ i = 1, \ldots, m$, is the direct estimate of the 1989 four-person family median income of the ith state from the Current

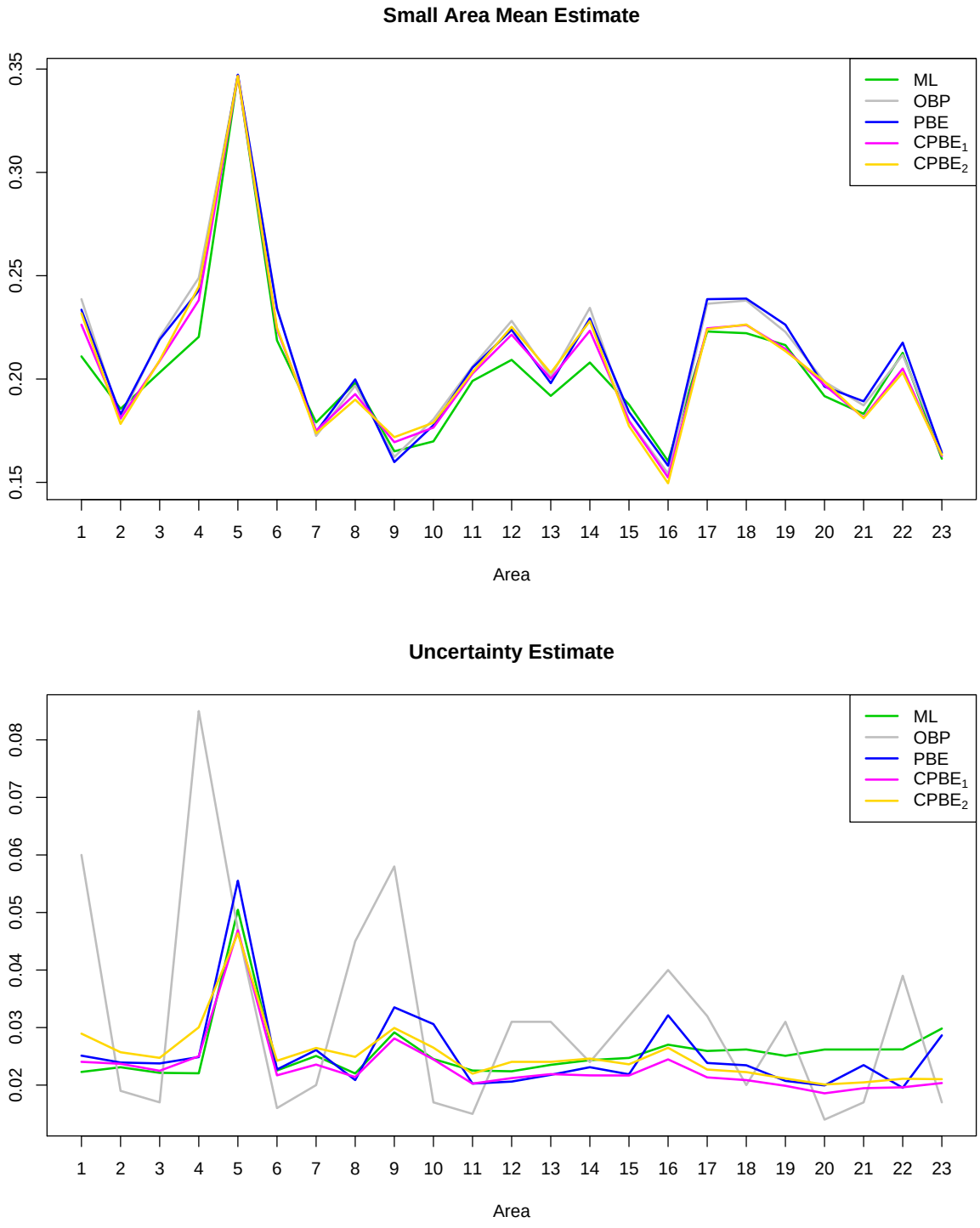

Figure 3.2: Small area mean estimates (top) and their uncertainty estimates (bottom) (hospital data).

Population Survey (CPS). Contrary to the noisy CPS estimates, the estimates from the 1990 census have negligible standard errors and we treat them as the true values. As before, we use the two covariates suggested by Fay (1987):

(1) x_{i1}: 1979 four-person family median income of the ith state from the 1980 census;

(2) $x_{i2} = (\text{PCI}_{i,1989}/\text{PCI}_{i,1979}) \cdot x_{i1}$: 1989 adjusted census four-person family median income of the ith state,

where PCI is the per capita income from the U.S. Bureau of Economic Analysis. Here, we fit a Fay-Herriot model with the mean function $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ to the median income data using our compromise pseudo-Bayesian methods. Regarding the prior parameter b, we use $b = 0.5$ as it works better than $b = 0$ in terms of both accuracy and stability of the estimates in both cases of CPBE_1 and CPBE_2.

Histograms of posterior simulations of the mixing parameter α for CPBE_1 and CPBE_2 are shown in the left and right panels of Figure 3.3, respectively. Both histograms are left skewed, but the first one is much more seriously skewed with mean 0.89, while the second one has mean 0.76 as reported in Table 3.3. According to Table 3.3, the estimates of A are all reasonably comparable, but the estimates of β_0 and β_1 from the OBP and PBE are totally different from others with the estimates of β_1 being negative. However, 95% credible intervals for β_0 and β_1 based on the PBE, CPBE_1, and CPBE_2 all contain 0 (not shown), implying β_0 and β_1 are not significant. On the other hand, the estimates of β_2 in Table 3.3 are all fairly close to each other and all 95% credible intervals for β_2 are strictly above 0 (not shown).

As in Chapter 2, we compare the small area mean estimates using four deviation measures, namely, average absolute deviation (AAD), average squared deviation (ASD), average absolute relative deviation (AARD), and average squared relative deviation (ASRD). See Chapter 2 for the definitions of the deviation measures. The deviation measures are computed using the true values from the 1990 census and reported in Table 3.4. Also, the small

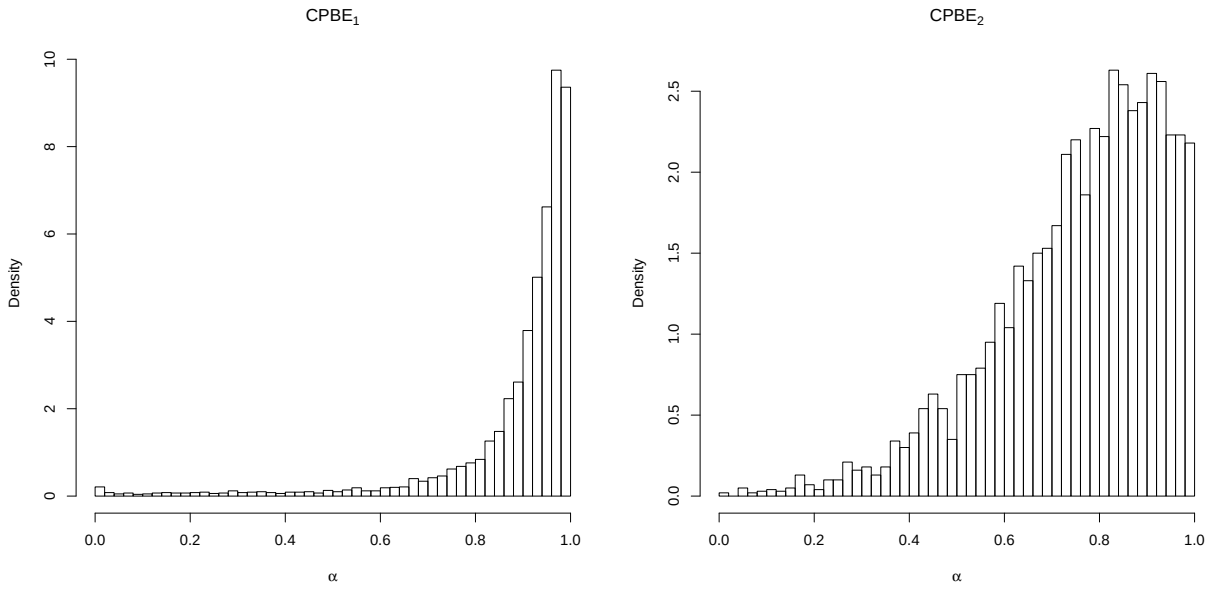

Figure 3.3: Histograms of posterior simulations of α for CPBE_1 (left) and CPBE_2 (right) (median income data).

Table 3.3: Estimates of the model parameters (median income data).

Estimate	α	A	β_0	β_1	β_2
ML	–	5.57×10^6	6362.2	0.090	0.705
REML	–	6.41×10^6	6489.0	0.080	0.707
FH	–	6.55×10^6	6507.8	0.079	0.707
PR	–	6.19×10^6	6456.0	0.082	0.707
OBP	–	5.65×10^6	13018.2	-0.335	0.783
PBE	–	6.02×10^6	12485.9	-0.302	0.777
CPBE_1	0.89	5.60×10^6	6712.2	0.071	0.707
CPBE_2	0.76	4.41×10^6	7514.8	0.008	0.722

Table 3.4: Comparison of estimators (median income data).

Estimate	AAD	ASD	AARD	ASRD
Direct	2928.8	13.81×10^6	0.0735	0.0084
ML	1394.9	3.19×10^6	0.0348	0.0019
REML	1454.9	3.45×10^6	0.0363	0.0021
FH	1464.4	3.49×10^6	0.0365	0.0021
PR	1438.1	3.38×10^6	0.0359	0.0020
OBP	1652.4	4.15×10^6	0.0420	0.0026
PBE	1580.3	3.81×10^6	0.0403	0.0024
CPBE_1	1366.5	3.07×10^6	0.0342	0.0018
CPBE_2	1296.9	2.77×10^6	0.0326	0.0017

area mean estimates and the true values are shown in the top panel of Figure 3.4, from which the estimates based on different methods seem very similar. According to Table 3.4, however, CPBE_2 performs the best, followed by CPBE_1. As we have already checked in Chapter 2, both the OBP and PBE are outperformed by the EBLUPs. But now with the compromise regression weights, the CPBEs have an edge on the EBLUPs.

The uncertainty estimates of the small area mean estimates are shown in the bottom panel of Figure 3.4 and their averages are reported in Table 3.5. As in the hospital data example, the OBP RMSPE estimates are very unstable across the areas. On the other hand, the posterior standard deviations from CPBE_1 and CPBE_2 are much more stable and smaller than the other uncertainty estimates in most areas. According to Table 3.5, the uncertainty estimates for CPBE_1 and CPBE_2 are smallest on average, where the APSDs are about 5.8% and 10.5% smaller, respectively, than the average RMSPE estimate of the ML EBLUP.

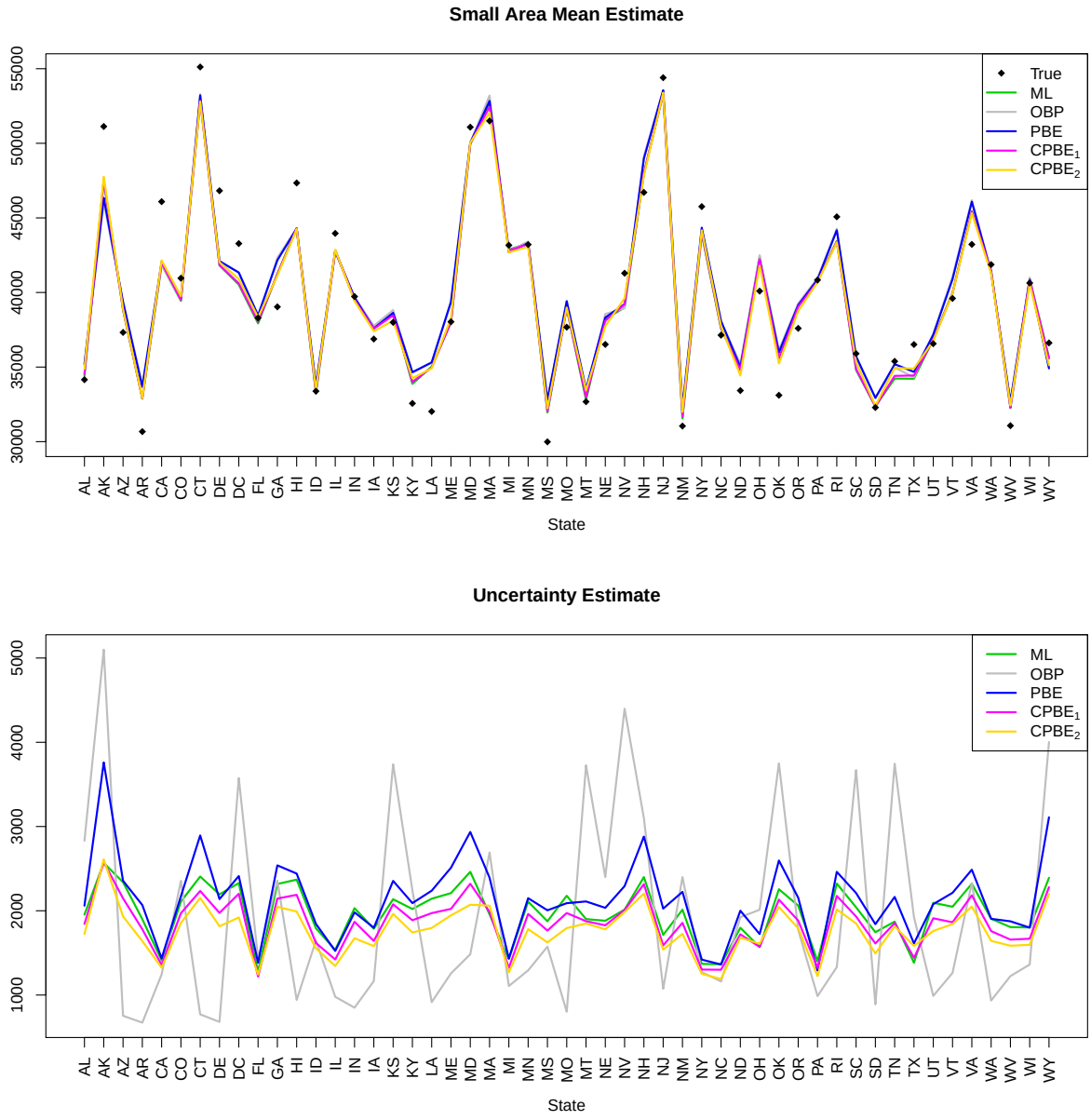

Figure 3.4: Small area mean estimates (top) and their uncertainty estimates (bottom) (median income data).

Table 3.5: Averages of the uncertainty estimates of the small area mean estimates (median income data).

Average RMSPE estimate					APSD		
ML	REML	FH	PR	OBP	PBE	$CPBE_1$	$CPBE_2$
1967.0	1968.5	1980.2	1996.1	1923.7	2135.0	1853.5	1760.1

3.4 A Simulation Study

We extend the simulation study conducted in Section 2.6.2 by adding CPBE_1 and CPBE_2. Recall the Q-O model, which is a Fay-Herriot model, fitted to the hospital data

$$y_i = \theta_i + e_i,\ \theta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + d \cdot I(x_i > 0.3) + v_i,\ i = 1, \ldots, m, \tag{3.5}$$

where $v_i \overset{\text{iid}}{\sim} \text{N}(0, A)$ independent of $e_i \overset{\text{ind}}{\sim} \text{N}(0, D_i)$. The simulation study imitates the hospital data by generating data from the Q-O model (3.5) with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, d)^T = (-1.1, 20, -50, 0.9)^T$, $A = 0.0016$, and with x_i's and D_i's the same as those of the hospital data. We consider three different numbers of small areas: $m = 23$, 115, and 230, where $m = 23$ corresponds to the hospital data. When $m = 115$ or 230, we replicate the design (i.e., the x_i's and D_i's) 5 or 10 times with $(\boldsymbol{\beta}^T, A)^T$ unchanged. We carry out $K = 5000$ simulation runs when $m = 23$ and $K = 500$ simulation runs when $m = 115$ or 230. For each dataset generated, we fit two different models corresponding to a misspecified case and a correctly specified case. In the misspecified case, we fit a quadratic model by omitting the term "$d \cdot I(x_i > 0.3)$" from the Q-O model (3.5). In the correctly specified case, the Q-O model (3.5) is fitted with no misspecification.

Let $\boldsymbol{\theta}_{(k)} = (\theta_{i(k)})_{1 \le i \le m}$, $k = 1, \ldots, K$, denote the true $\boldsymbol{\theta} = (\theta_i)_{1 \le i \le m}$ from the kth simulation run. For each estimator $\breve{\boldsymbol{\theta}}_{(k)} = (\breve{\theta}_{i(k)})_{1 \le i \le m}$ of $\boldsymbol{\theta}_{(k)}$, we compute the empirical MSPE

$$\text{MSPE}^* = \frac{1}{K} \sum_{k=1}^{K} |\breve{\boldsymbol{\theta}}_{(k)} - \boldsymbol{\theta}_{(k)}|^2 = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{m} \{\breve{\theta}_{i(k)} - \theta_{i(k)}\}^2.$$

The empirical MSPEs (multiplied by 100) are reported in Table 3.6. We choose the REML EBLUP since it performs the best among the EBLUPs according to Table 2.4 in Chapter 2. According to Table 3.6, the EBLUP is optimal when the model is correctly specified, but if the model is misspecified, the OBP and PBE can outperform the EBLUP. On the other

Table 3.6: Empirical MSPEs (multiplied by 100).

m	Model	REML	OBP	PBE	CPBE_1	CPBE_2
23	Misspecified	2.870	2.981	2.954	3.058	2.900
23	Correct	2.288	2.442	2.342	2.396	2.288
115	Misspecified	12.872	12.805	12.808	12.906	12.853
115	Correct	9.471	9.640	9.638	9.589	9.625
230	Misspecified	25.531	25.277	25.287	25.553	25.490
230	Correct	18.592	18.774	18.771	18.635	18.730

hand, the CPBEs seem to perform somewhere between the EBLUP and OBP/PBE except CPBE_1 is not very effective when the model is misspecified. This suggests CPBE_2 can be a safer choice that compromises between the EBLUP and OBP/PBE when there is not enough information to determine if the model is correctly specified.

Finally, we report the empirical coverage probabilities of 95% confidence/credible intervals for θ_i's in Table 3.7. The number in the parentheses is the average length of the intervals. As we have checked in Chapter 2, the coverage probabilities of the OBP are well below 0.95, while the EBLUP and PBE intervals attain the 0.95 nominal coverage probability. On the other hand, CPBE_1 and CPBE_2 yield coverage probabilities that are only slightly less than 0.95 with shoter intervals compared to the EBLUP and PBE.

3.5 Conclusions

In addition to the PBE proposed in Chapter 2, we proposed two more pseudo-Bayesian estimators that compromise between the EBLUP and OBP. Specifically, the compromise pseudo-Bayesian estimators CPBE_1 and CPBE_2 use convex combinations of weights that are proportional to the MLE and BPE weights from the EBLUP and OBP, respectively. The mixing parameter in the compromise weights is treated as an additional parameter. The real data examples showed that the gain of CPBEs over EBLUP, OBP, and PBE can

Table 3.7: Empirical coverage probabilities of 95% confidence/credible intervals for θ_i's with average lengths of the intervals.

m	Model	REML	OBP	PBE	CPBE_1	CPBE_2
23	Misspecified	0.9470	0.8834	0.9606	0.9229	0.9406
		(0.1356)	(0.1314)	(0.1486)	(0.1286)	(0.1329)
23	Correct	0.9397	0.8343	0.9614	0.9022	0.9346
		(0.1216)	(0.1177)	(0.1345)	(0.1078)	(0.1172)
115	Misspecified	0.9494	0.8930	0.9526	0.9472	0.9474
		(0.1291)	(0.1226)	(0.1320)	(0.1278)	(0.1280)
115	Correct	0.9471	0.8049	0.9499	0.9360	0.9346
		(0.1114)	(0.1007)	(0.1142)	(0.1073)	(0.1069)
230	Misspecified	0.9495	0.8937	0.9511	0.9477	0.9482
		(0.1284)	(0.1214)	(0.1297)	(0.1277)	(0.1275)
230	Correct	0.9488	0.8056	0.9495	0.9444	0.9394
		(0.1103)	(0.0990)	(0.1114)	(0.1085)	(0.1066)

be substantial in terms of both accuracy and stability of the small area mean estimates. The simulation study showed that CPBE_2 can be a safer alternative to the EBLUP, OBP, and PBE as its performance was somewhere between that of the EBLUP and OBP/PBE in both misspecified and correctly specified cases. However, the simulation study was originally designed to compare the performance of the OBP/PBE to that of the EBLUP, where only one misspecified case was considered as opposed to the correctly specified case. An in-depth simulation study that considers various degrees of model misspecification will help explore the performance of the CPBEs more thoroughly.

3.6 Proofs

Throughout this section, C denotes a generic positive constant that does not depend on the model parameters $\boldsymbol{\beta}$, A, and α.

3.6.1 Proof of Theorem 3.1

Proof. First, note that

$$\begin{aligned}
&\int_{\mathbb{R}^p} L_1(\boldsymbol{\beta}, A|\mathbf{y})d\boldsymbol{\beta} \\
=&C\int_{\mathbb{R}^p} |\mathbf{W}_1|^{\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_1(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]d\boldsymbol{\beta} \\
=&C|\mathbf{W}_1|^{\frac{1}{2}}\int_{\mathbb{R}^p} \exp[-\frac{1}{2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{W}_1\mathbf{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})]d\boldsymbol{\beta}\exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^T\mathbf{W}_1(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})] \\
&\qquad\qquad (\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}_1\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_1\mathbf{y}) \\
=&C|\mathbf{W}_1|^{\frac{1}{2}}|\mathbf{X}^T\mathbf{W}_1\mathbf{X}|^{-\frac{1}{2}}\exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^T\mathbf{W}_1(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})] \\
\leq&C|\mathbf{W}_1|^{\frac{1}{2}}|\mathbf{X}^T\mathbf{W}_1\mathbf{X}|^{-\frac{1}{2}}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\int_{\mathbb{R}^p} L_2(\boldsymbol{\beta}, A|\mathbf{y})d\boldsymbol{\beta} &= C\int_{\mathbb{R}^p} |\mathbf{W}_2|^{\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]d\boldsymbol{\beta} \\
&\leq C|\mathbf{W}_2|^{\frac{1}{2}}|\mathbf{X}^T\mathbf{W}_2\mathbf{X}|^{-\frac{1}{2}}.
\end{aligned}$$

Recall that $\mathbf{W}_1 = \text{diag}((A+D_i)^{-1},\ 1 \leq i \leq m)$ and $\mathbf{W}_2 = \text{diag}(\{D_i/(A+D_i)\}^2/(A+\overline{D}),\ 1 \leq i \leq m)$, and let $D_{(1)} = \min_{1\leq i\leq m} D_i$ and $D_{(m)} = \max_{1\leq i\leq m} D_i$. Then $|\mathbf{W}_1| \leq (A+D_{(1)})^{-m}$, $|\mathbf{X}^T\mathbf{W}_1\mathbf{X}| \geq (A+D_{(m)})^{-p}|\mathbf{X}^T\mathbf{X}|$, and

$$\begin{aligned}
|\mathbf{W}_1|^{\frac{1}{2}}|\mathbf{X}^T\mathbf{W}_1\mathbf{X}|^{-\frac{1}{2}} &\leq (A+D_{(1)})^{-\frac{m}{2}}(A+D_{(m)})^{\frac{p}{2}}|\mathbf{X}^T\mathbf{X}|^{-\frac{1}{2}} \\
&\leq \{\min(D_{(1)},1)\}^{-\frac{m}{2}}(A+1)^{-\frac{m}{2}}\{\max(D_{(m)},1)\}^{\frac{p}{2}}(A+1)^{\frac{p}{2}}|\mathbf{X}^T\mathbf{X}|^{-\frac{1}{2}} \\
&= C(A+1)^{-\frac{m-p}{2}};
\end{aligned}$$

$|\mathbf{W}_2| \leq \{D_{(m)}/(A+D_{(m)})\}^{2m}/(A+\overline{D})^m$, $|\mathbf{X}^T\mathbf{W}_2\mathbf{X}| \geq \{D_{(1)}/(A+D_{(1)})\}^{2p}/(A+\overline{D})^p \cdot |\mathbf{X}^T\mathbf{X}|$, and

$$
\begin{aligned}
|\mathbf{W}_2|^{\frac{1}{2}}|\mathbf{X}^T\mathbf{W}_2\mathbf{X}|^{-\frac{1}{2}} \leq & \left(\frac{D_{(m)}}{A+D_{(m)}}\right)^m (A+\overline{D})^{-\frac{m}{2}} \left(\frac{D_{(1)}}{A+D_{(1)}}\right)^{-p} (A+\overline{D})^{\frac{p}{2}}|\mathbf{X}^T\mathbf{X}|^{-\frac{1}{2}} \\
\leq & D_{(m)}^m\{\min(D_{(m)},1)\}^{-m}(A+1)^{-m}\{\min(\overline{D},1)\}^{-\frac{m}{2}}(A+1)^{-\frac{m}{2}} \\
& \times D_{(1)}^{-p}\{\max(D_{(1)},1)\}^p(A+1)^p\{\max(\overline{D},1)\}^{\frac{p}{2}}(A+1)^{\frac{p}{2}}|\mathbf{X}^T\mathbf{X}|^{-\frac{1}{2}} \\
= & C(A+1)^{-\frac{3(m-p)}{2}}.
\end{aligned}
$$

Now,

$$
\begin{aligned}
& \int_0^\infty \int_{\mathbb{R}^p} L_{\mathrm{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})\pi(\boldsymbol{\beta}, A, \alpha)d\boldsymbol{\beta}dA \\
= & C\int_0^\infty \int_{\mathbb{R}^p} L_1^\alpha(\boldsymbol{\beta}, A|\mathbf{y})L_2^{1-\alpha}(\boldsymbol{\beta}, A|\mathbf{y})A^{-b}d\boldsymbol{\beta}dA \\
= & C\int_0^\infty \int_{\mathbb{R}^p} \{A^{-b}L_1(\boldsymbol{\beta}, A|\mathbf{y})\}^\alpha\{A^{-b}L_2(\boldsymbol{\beta}, A|\mathbf{y})\}^{1-\alpha}d\boldsymbol{\beta}dA \\
\leq & C\left\{\int_0^\infty \int_{\mathbb{R}^p} A^{-b}L_1(\boldsymbol{\beta}, A|\mathbf{y})d\boldsymbol{\beta}dA\right\}^\alpha \left\{\int_0^\infty \int_{\mathbb{R}^p} A^{-b}L_2(\boldsymbol{\beta}, A|\mathbf{y})d\boldsymbol{\beta}dA\right\}^{1-\alpha} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Hölder's inequality)} \\
\leq & C\left\{\int_0^\infty A^{-b}|\mathbf{W}_1|^{\frac{1}{2}}|\mathbf{X}^T\mathbf{W}_1\mathbf{X}|^{-\frac{1}{2}}dA\right\}^\alpha \left\{\int_0^\infty A^{-b}|\mathbf{W}_2|^{\frac{1}{2}}|\mathbf{X}^T\mathbf{W}_2\mathbf{X}|^{-\frac{1}{2}}dA\right\}^{1-\alpha} \\
\leq & C\left\{\int_0^\infty A^{-b}(A+1)^{-\frac{m-p}{2}}dA\right\}^\alpha \left\{\int_0^\infty A^{-b}(A+1)^{-\frac{3(m-p)}{2}}dA\right\}^{1-\alpha} \\
< & C\left\{\int_0^\infty A^{-b}(A+1)^{-\frac{m-p}{2}}dA\right\}^\alpha \left\{\int_0^\infty A^{-b}(A+1)^{-\frac{m-p}{2}}dA\right\}^{1-\alpha} \\
= & C\int_0^\infty A^{-b}(A+1)^{-\frac{m-p}{2}}dA \\
= & C\cdot \mathrm{B}\left(1-b, \frac{m-p}{2}-(1-b)\right),
\end{aligned}
$$

where $\mathrm{B}(\cdot,\cdot)$ is the beta function and the last equality follows provided that

$$
1-b>0,\ \frac{m-p}{2}-(1-b)>0 \Longleftrightarrow 1-\frac{m-p}{2}<b<1.
$$

Finally,

$$
\begin{aligned}
&\int_0^1 \int_0^\infty \int_{\mathbb{R}^p} L_{\mathrm{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})\pi(\boldsymbol{\beta}, A, \alpha)d\boldsymbol{\beta}dAd\alpha \\
<&\int_0^1 C \cdot \mathrm{B}\left(1-b, \frac{m-p}{2} - (1-b)\right) d\alpha \\
=&C \cdot \mathrm{B}\left(1-b, \frac{m-p}{2} - (1-b)\right) \\
<&\infty,
\end{aligned}
$$

where $1-(m-p)/2 < b < 1$. Therefore, the posterior density $\pi_{\mathrm{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y}) \propto L_{\mathrm{c},1}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})$ $\times \pi(\boldsymbol{\beta}, A, \alpha)$ is proper if $1-(m-p)/2 < b < 1$. □

3.6.2 Proof of Theorem 3.2

Proof. First,

$$
\begin{aligned}
&\int_{\mathbb{R}^p} L_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})\pi(\boldsymbol{\beta}, A, \alpha)d\boldsymbol{\beta} \\
=&C \cdot A^{-b}(A+\overline{D})^{-\frac{m}{2}}|\mathbf{W}_{\mathrm{c},2}|^{\frac{1}{2}} \int_{\mathbb{R}^p} \exp\left[-\frac{m}{2(A+\overline{D})}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}_{\mathrm{c},2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] d\boldsymbol{\beta} \\
=&C \cdot A^{-b}(A+\overline{D})^{-\frac{m}{2}}|\mathbf{W}_{\mathrm{c},2}|^{\frac{1}{2}} \int_{\mathbb{R}^p} \exp\left[-\frac{m}{2(A+\overline{D})}(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}_{\mathrm{c},2})^T\mathbf{X}^T\mathbf{W}_{\mathrm{c},2}\mathbf{X}(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}_{\mathrm{c},2})\right] d\boldsymbol{\beta} \\
&\times \exp\left[-\frac{m}{2(A+\overline{D})}(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}_{\mathrm{c},2})^T\mathbf{W}_{\mathrm{c},2}(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}_{\mathrm{c},2})\right] \qquad (\widetilde{\boldsymbol{\beta}}_{\mathrm{c},2} = (\mathbf{X}^T\mathbf{W}_{\mathrm{c},2}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_{\mathrm{c},2}\mathbf{y}) \\
=&C \cdot A^{-b}(A+\overline{D})^{-\frac{m}{2}}|\mathbf{W}_{\mathrm{c},2}|^{\frac{1}{2}} \left|\frac{m}{A+\overline{D}}\mathbf{X}^T\mathbf{W}_{\mathrm{c},2}\mathbf{X}\right|^{-\frac{1}{2}} \\
&\times \exp\left[-\frac{m}{2(A+\overline{D})}(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}_{\mathrm{c},2})^T\mathbf{W}_{\mathrm{c},2}(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}_{\mathrm{c},2})\right] \\
=&C \cdot A^{-b}(A+\overline{D})^{-\frac{m-p}{2}}|\mathbf{W}_{\mathrm{c},2}|^{\frac{1}{2}}|\mathbf{X}^T\mathbf{W}_{\mathrm{c},2}\mathbf{X}|^{-\frac{1}{2}} \exp\left[-\frac{m}{2(A+\overline{D})}(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}_{\mathrm{c},2})^T\mathbf{W}_{\mathrm{c},2}(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}_{\mathrm{c},2})\right] \\
\leq&C \cdot A^{-b}(A+\overline{D})^{-\frac{m-p}{2}}|\mathbf{W}_{\mathrm{c},2}|^{\frac{1}{2}}|\mathbf{X}^T\mathbf{W}_{\mathrm{c},2}\mathbf{X}|^{-\frac{1}{2}} \\
\leq&C \cdot A^{-b}(A+\overline{D})^{-\frac{m-p}{2}}|\mathbf{X}^T\mathbf{W}_{\mathrm{c},2}\mathbf{X}|^{-\frac{1}{2}},
\end{aligned}
$$

where the last inequality follows by noting that

$$|\mathbf{W}_{\mathrm{c},2}| = \prod_{i=1}^{m} w_i^{\mathrm{c}} = \prod_{i=1}^{m} \{\alpha \cdot w_i^{\mathrm{MLE}} + (1-\alpha) \cdot w_i^{\mathrm{BPE}}\}$$
$$= \prod_{i=1}^{m} \left\{ \alpha \cdot \frac{(A+D_i)^{-1}}{\sum_{h=1}^{m}(A+D_h)^{-1}} + (1-\alpha) \cdot \frac{\{D_i/(A+D_i)\}^2}{\sum_{h=1}^{m}\{D_h/(A+D_h)\}^2} \right\} \leq 1$$

for all $A > 0$ and $\alpha \in [0,1]$. Also, note that as $A \to \infty$,

$$w_i^{\mathrm{MLE}} = \frac{(A+D_i)^{-1}}{\sum_{h=1}^{m}(A+D_h)^{-1}} \to \frac{1}{m}, \; w_i^{\mathrm{BPE}} = \frac{\{D_i/(A+D_i)\}^2}{\sum_{h=1}^{m}\{D_h/(A+D_h)\}^2} \to \frac{D_i^2}{\sum_{h=1}^{m} D_h^2},$$

and

$$w_i^{\mathrm{c}} = \alpha \cdot w_i^{\mathrm{MLE}} + (1-\alpha) \cdot w_i^{\mathrm{BPE}} \to \alpha \cdot \frac{1}{m} + (1-\alpha) \cdot \frac{D_i^2}{\sum_{h=1}^{m} D_h^2} \geq \min\left(\frac{1}{m}, \frac{D_{(1)}^2}{\sum_{h=1}^{m} D_h^2}\right),$$

where $D_{(1)} = \min_{1 \leq i \leq m} D_i$. This implies $|\mathbf{X}^T \mathbf{W}_{\mathrm{c},2} \mathbf{X}|^{-1/2} = |\sum_{i=1}^{m} w_i^{\mathrm{c}} \mathbf{x}_i \mathbf{x}_i^T|^{-1/2} < \infty$ for all $A > 0$ and $\alpha \in [0,1]$. Finally,

$$\begin{aligned}
&\int_0^1 \int_0^\infty \int_{\mathbb{R}^p} L_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha | \mathbf{y}) \pi(\boldsymbol{\beta}, A, \alpha) d\boldsymbol{\beta} dA d\alpha \\
\leq & C \int_0^1 \int_0^\infty A^{-b} (A + \overline{D})^{-\frac{m-p}{2}} |\mathbf{X}^T \mathbf{W}_{\mathrm{c},2} \mathbf{X}|^{-\frac{1}{2}} dA d\alpha \\
\leq & C \int_0^1 \int_0^\infty A^{-b} (A + \overline{D})^{-\frac{m-p}{2}} dA d\alpha \\
= & C \int_0^\infty A^{-b} (A + \overline{D})^{-\frac{m}{2}} (A + \overline{D})^{\frac{p}{2}} dA \\
\leq & C \int_0^\infty A^{-b} \min(\overline{D}, 1)(A+1)^{-\frac{m}{2}} \max(\overline{D}, 1)(A+1)^{\frac{p}{2}} dA \\
= & C \int_0^\infty A^{-b} (A+1)^{-\frac{m-p}{2}} dA \\
= & C \cdot \mathrm{B}\left(1-b, \frac{m-p}{2} - (1-b)\right) \\
< & \infty,
\end{aligned}$$

provided that $1-(m-p)/2 < b < 1$. Therefore, the posterior density $\pi_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y}) \propto L_{\mathrm{c},2}(\boldsymbol{\beta}, A, \alpha|\mathbf{y})\pi(\boldsymbol{\beta}, A, \alpha)$ is proper if $1-(m-p)/2 < b < 1$. □

Chapter 4

Small Area Estimation using Aggregate Information

4.1 Introduction

Suppose there are m small areas of interest and our goal is to estimate the small area mean $\boldsymbol{\theta} = (\theta_i)_{1\leq i\leq m}$. Further suppose for each of those m small areas, p covariates are available, forming an $m \times p$ design matrix $\mathbf{X} = (\mathbf{x}_i^T)_{1\leq i\leq m}$ of rank $p(< m)$. Ideally, we have a direct estimator $\mathbf{y} = (y_i)_{1\leq i\leq m}$ from a suitable survey that is design unbiased for $\boldsymbol{\theta}$ so that $\mathrm{E}(\mathbf{y}|\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $\mathrm{Var}(\mathbf{y}|\boldsymbol{\theta}) = \mathbf{D} = \mathrm{diag}(D_i,\ 1 \leq i \leq m)$, where the sampling variances D_i's are known. A widely used model in this case is the Fay-Herriot model (Fay and Herriot, 1979)

$$\begin{aligned} &\text{I. } \mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\boldsymbol{\theta}, \mathbf{D}); \\ &\text{II. } \boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_m), \end{aligned} \tag{4.1}$$

where the $p \times 1$ vector of regression coefficients $\boldsymbol{\beta}$ and the common area-specific random effect variance σ^2 are the model parameters. Under squared error loss, the best predictor

(BP) of $\boldsymbol{\theta}$ is the conditional expectation of $\boldsymbol{\theta}$ given $\boldsymbol{\beta}$, σ^2, and $\mathbf{y}$

$$\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{\beta},\sigma^2,\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \sigma^2(\mathbf{D}+\sigma^2\mathbf{I}_m)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \tag{4.2}$$

where the conditional variance is

$$\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{\beta},\sigma^2,\mathbf{y}) = \sigma^2\mathbf{I}_m - (\sigma^2)^2(\mathbf{D}+\sigma^2\mathbf{I}_m)^{-1} = \{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{D}^{-1}\}^{-1}. \tag{4.3}$$

Note that $\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{\beta},\sigma^2,\mathbf{y})$ is the inverse of the sum of the prior precision matrix $(\sigma^2)^{-1}\mathbf{I}_m$ for $\boldsymbol{\theta}$ and the sample precision matrix $\mathbf{D}^{-1}$ for $\boldsymbol{\theta}$ based on $\mathbf{y}$.

In some small area estimation problems, covariates are available at the desired level of small areas, but surveys provide direct estimates only at a higher level of aggregation. However, small area estimates are often needed at the same lower level as the covariates for policy making, research, etc. For example, survey estimates on childhood poverty may be available only at the county level when those estimates are needed at the school district level.

Suppose m small areas $A_i,\ i = 1,\ldots,m$, are grouped into $r(\leq m)$ regions $R_1 = \{A_1,\ldots,A_{n_1}\}, R_2 = \{A_{n_1+1},\ldots,A_{n_1+n_2}\},\ldots,R_r = \{A_{n_1+\cdots+n_{r-1}+1},\ldots,A_m\}$ so that there are n_j small areas in $R_j,\ j = 1,\ldots,r$, with $m = \sum_{j=1}^{r} n_j$. For example, if A_i's are the 50 U.S. states and the District of Columbia (i.e., $m = 51$), R_j's may be the $r = 9$ U.S. census divisions: (1) New England, (2) Middle Atlantic, (3) East North Central, (4) West North Central, (5) South Atlantic, (6) East South Central, (7) West South Central, (8) Mountain, and (9) Pacific. Suppose $\mathbf{w} = (w_j)_{1\leq j\leq r}$ is a vector of direct estimators such that w_j is design unbiased for some known linear combination of the small area means in $R_j,\ j = 1,\ldots,r$. That is, w_j is a design unbiased direct estimator of

$$\mathbf{c}_j^T\boldsymbol{\theta}_j = c_{j,1}\cdot\theta_{n_1+\cdots+n_{j-1}+1} + c_{j,2}\cdot\theta_{n_1+\cdots+n_{j-1}+2} + \cdots + c_{j,n_j}\cdot\theta_{n_1+\cdots+n_{j-1}+n_j}$$

for some known $\mathbf{c}_j = (c_{j,1}, c_{j,2}, \ldots, c_{j,n_j})^T$, where $\boldsymbol{\theta}_j = (\theta_{n_1+\cdots+n_{j-1}+1}, \theta_{n_1+\cdots+n_{j-1}+2}, \ldots, \theta_{n_1+\cdots+n_{j-1}+n_j})^T$ is the vector of small area means in R_j. This implies $\mathbf{w}$ is design unbiased for $\mathbf{C}\boldsymbol{\theta}$, where $\mathbf{C} = \text{diag}(\mathbf{c}_j^T,\ 1 \leq j \leq r)$ is a known $r \times m$ matrix with rank r. A simple example would be that w_j is a direct estimator of the simple average of the small area means in R_j, in which case $\mathbf{C} = \text{diag}(n_j^{-1}\mathbf{1}^T,\ 1 \leq j \leq r)$, where $\mathbf{1}$ is a vector of ones with appropriate dimension. Without loss of generality, we take $\mathbf{C} = \mathbf{I}_m$ when $r = m$, so that $r = m$ if and only if $\mathbf{C} = \mathbf{I}_m$. In that special case, $\mathbf{w}$ is design unbiased for $\boldsymbol{\theta}$ and can be treated as $\mathbf{y}$.

In this chapter, we consider a small area estimation problem where $\mathbf{y}$ is not available, but instead $\mathbf{w}$ and $\mathbf{C}$ are given. Moreover, we assume the relationship $\mathbf{w} = \mathbf{C}\mathbf{y}$ between $\mathbf{y}$ and $\mathbf{w}$, which makes logical sense since $\mathbf{y}$ and $\mathbf{w}$ are design unbiased for $\boldsymbol{\theta}$ and $\mathbf{C}\boldsymbol{\theta}$, respectively. Also, we treat $\text{Var}(\mathbf{w}|\boldsymbol{\theta}) = \text{Var}(\mathbf{C}\mathbf{y}|\boldsymbol{\theta}) = \mathbf{C}\mathbf{D}\mathbf{C}^T$ as known. It could be that $\mathbf{C}\mathbf{D}\mathbf{C}^T$ is given as a whole or $\mathbf{D}$ itself is known from a reliable source. We generalize the Fay-Herriot model (4.1) using direct estimators at a higher level of aggregation and propose a hierarchical Bayesian version of the model to estimate the small area means at the desired lower level. We provide theoretical comparisons of the posterior variances of the small area means based on higher and lower level direct estimators, and illustrate our findings with a real data example.

4.2 A Generalized Fay-Herriot Model

When $\mathbf{y}$ is not available, based on $\mathbf{w}$ and $\mathbf{C}$, the Fay-Herriot model (4.1) can be generalized as

$$\begin{aligned} &\text{I. } \mathbf{w}|\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 \sim \text{N}(\mathbf{C}\boldsymbol{\theta}, \mathbf{D}_{\mathbf{w}}); \\ &\text{II. } \boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2 \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_m), \end{aligned} \tag{4.4}$$

where $\mathbf{D_w} = \mathbf{CDC}^T$. Clearly, the Fay-Herriot model (4.1) is a special case of the generalized Fay-Herriot model (4.4) with $\mathbf{C} = \mathbf{I}_m$, in which case $\mathbf{w} = \mathbf{y}$. From

$$\begin{pmatrix} \mathbf{w} \\ \boldsymbol{\theta} \end{pmatrix} \Bigg| \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{Z}\boldsymbol{\beta} \\ \mathbf{X}\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \mathbf{D_w} + \sigma^2\mathbf{CC}^T & \sigma^2\mathbf{C} \\ \sigma^2\mathbf{C}^T & \sigma^2\mathbf{I}_m \end{pmatrix} \right),$$

where $\mathbf{Z} = \mathbf{CX}$, it is immediate that the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{\beta}$, σ^2, and $\mathbf{w}$ is normal with mean

$$\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) = \mathbf{X}\boldsymbol{\beta} + \sigma^2\mathbf{C}^T(\mathbf{D_w} + \sigma^2\mathbf{CC}^T)^{-1}(\mathbf{w} - \mathbf{Z}\boldsymbol{\beta}) \tag{4.5}$$

and variance

$$\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) = \sigma^2\mathbf{I}_m - (\sigma^2)^2\mathbf{C}^T(\mathbf{D_w} + \sigma^2\mathbf{CC}^T)^{-1}\mathbf{C}. \tag{4.6}$$

Under squared error loss, $\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w})$ is the BP of $\boldsymbol{\theta}$ based on $\mathbf{w}$. Note that the conditional mean and variance in (4.5) and (4.6) become those in (4.2) and (4.3) when $\mathbf{C} = \mathbf{I}_m$.

Alternatively, from a Bayesian perspective, the posterior density of $\boldsymbol{\theta}$ given $\boldsymbol{\beta}$, σ^2, and $\mathbf{w}$ is

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) &\propto \pi(\mathbf{w}|\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2) \\ &\propto \exp[-\frac{1}{2}(\mathbf{w} - \mathbf{C}\boldsymbol{\theta})^T\mathbf{D}_\mathbf{w}^{-1}(\mathbf{w} - \mathbf{C}\boldsymbol{\theta})] \exp[-\frac{1}{2\sigma^2}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})] \\ &\propto \exp[-\frac{1}{2}\boldsymbol{\theta}^T\{\mathbf{C}^T\mathbf{D}_\mathbf{w}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}\boldsymbol{\theta} + \boldsymbol{\theta}^T\{\mathbf{C}^T\mathbf{D}_\mathbf{w}^{-1}\mathbf{w} + (\sigma^2)^{-1}\mathbf{X}\boldsymbol{\beta}\}], \end{aligned}$$

which is the kernel of a normal distribution with mean

$$\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) = \{\mathbf{C}^T\mathbf{D}_\mathbf{w}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1}\{\mathbf{C}^T\mathbf{D}_\mathbf{w}^{-1}\mathbf{w} + (\sigma^2)^{-1}\mathbf{X}\boldsymbol{\beta}\} \tag{4.7}$$

and variance

$$\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{\beta},\sigma^2,\mathbf{w}) = \{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1}. \tag{4.8}$$

Similar to $\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{\beta},\sigma^2,\mathbf{y})$ in (4.3), the inverse of $\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{\beta},\sigma^2,\mathbf{w})$ can be decomposed into the prior precision matrix $(\sigma^2)^{-1}\mathbf{I}_m$ for $\boldsymbol{\theta}$ and the matrix $\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}$ that can be interpreted as the sample precision for $\boldsymbol{\theta}$ based on $\mathbf{w}$. The equivalence between the expectations in (4.5) and (4.7), and that between the variances in (4.6) and (4.8) are stated in the following proposition and the proof is given in Section 4.6.1.

Proposition 4.1. *The expectations in* (4.5) *and* (4.7) *are equivalent, and so are the variances in* (4.6) *and* (4.8).

Now, we consider a hierarchical Bayesian version of the generalized Fay-Herriot model (4.4). A classic improper prior density for the model parameters is $\pi(\boldsymbol{\beta},\sigma^2) \propto (\sigma^2)^{-\alpha},\ \boldsymbol{\beta} \in \mathbb{R}^p,\ \sigma^2 > 0$, where α is a prior parameter on which a suitable condition is needed to achieve the propriety of the resulting posterior density. Using this prior density, a hierarchical Bayesian version of the generalized Fay-Herriot model (4.4) can be expressed as

$$\begin{aligned} &\text{I. } \mathbf{w}|\boldsymbol{\theta},\boldsymbol{\beta},\sigma^2 \sim \mathrm{N}(\mathbf{C}\boldsymbol{\theta},\mathbf{D}_{\mathbf{w}});\\ &\text{II. } \boldsymbol{\theta}|\boldsymbol{\beta},\sigma^2 \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_m);\\ &\text{III. } \pi(\boldsymbol{\beta},\sigma^2) \propto (\sigma^2)^{-\alpha},\ \boldsymbol{\beta} \in \mathbb{R}^p,\ \sigma^2 > 0. \end{aligned} \tag{4.9}$$

The condition on α that guarantees the propriety of the posterior density of the hierarchical Bayesian generalized Fay-Herriot (HB-GFH) model (4.9) is specified in the following theorem and the proof is given in Section 4.6.2.

Theorem 4.1. *The posterior density of the HB-GFH model* (4.9) *is proper if* $1-(r-p)/2 < \alpha < 1$.

4.3 Posterior Variance of Small Area Mean

We now study the posterior variance $\text{Var}(\boldsymbol{\theta}|\mathbf{w})$ of the small area mean $\boldsymbol{\theta}$ based on the HB-GFH model (4.9). By comparing $\text{Var}(\boldsymbol{\theta}|\mathbf{w})$ with $\text{Var}(\boldsymbol{\theta}|\mathbf{y})$, which corresponds to the special case of $\mathbf{C} = \mathbf{I}_m$, we investigate and quantify the source of increase in uncertainty caused by using direct estimators at a higher level of aggregation instead of those at the desired lower level. We first note the following lemma, whose proof is given in Section 4.6.3, that will be useful for proving the theorem that follows.

Lemma 4.1. *Based on the HB-GFH model* (4.9)*,*

$$\boldsymbol{\beta}|\sigma^2, \mathbf{w} \sim \text{N}\left((\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{w}, (\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\right),$$

where $\boldsymbol{\Omega} = \text{Var}(\mathbf{w}|\boldsymbol{\beta}, \sigma^2) = \mathbf{D}_\mathbf{w} + \sigma^2\mathbf{C}\mathbf{C}^T$*.*

Now, the following theorem states that $\text{Var}(\boldsymbol{\theta}|\mathbf{w})$ can be decomposed into three different components and the proof is given in Section 4.6.4.

Theorem 4.2. *The posterior variance* $\text{Var}(\boldsymbol{\theta}|\mathbf{w})$ *of the small area mean* $\boldsymbol{\theta}$ *based on the HB-GFH model* (4.9) *can be decomposed as*

$$\begin{aligned}\text{Var}(\boldsymbol{\theta}|\mathbf{w}) &= \text{E}[\mathbf{G}_{1,\sigma^2}(\mathbf{C})|\mathbf{w}] + \text{E}[\mathbf{G}_{2,\sigma^2}(\mathbf{C})|\mathbf{w}] + \text{Var}[\mathbf{g}_{3,\sigma^2}(\mathbf{C})|\mathbf{w}] \\ &= \mathbf{G}_1(\mathbf{C}) + \mathbf{G}_2(\mathbf{C}) + \mathbf{G}_3(\mathbf{C}),\end{aligned}$$

where

$$\begin{aligned}\mathbf{G}_{1,\sigma^2}(\mathbf{C}) &= \{\mathbf{C}^T\mathbf{D}_\mathbf{w}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1}, \\ \mathbf{G}_{2,\sigma^2}(\mathbf{C}) &= (\sigma^2)^{-2}\{\mathbf{G}_{1,\sigma^2}(\mathbf{C})\}\mathbf{X}(\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{X}^T\{\mathbf{G}_{1,\sigma^2}(\mathbf{C})\}, \\ \mathbf{g}_{3,\sigma^2}(\mathbf{C}) &= \{\mathbf{G}_{1,\sigma^2}(\mathbf{C})\}\{\mathbf{C}^T\mathbf{D}_\mathbf{w}^{-1}\mathbf{w} + (\sigma^2)^{-1}\mathbf{X}(\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{w}\},\end{aligned}$$

and $\mathbf{G}_1(\mathbf{C}) = \mathrm{E}[\mathbf{G}_{1,\sigma^2}(\mathbf{C})|\mathbf{w}]$, $\mathbf{G}_2(\mathbf{C}) = \mathrm{E}[\mathbf{G}_{2,\sigma^2}(\mathbf{C})|\mathbf{w}]$, *and* $\mathbf{G}_3(\mathbf{C}) = \mathrm{Var}[\mathbf{g}_{3,\sigma^2}(\mathbf{C})|\mathbf{w}]$.

It can be easily checked from Theorem 4.2 (and its proof) that if both $\boldsymbol{\beta}$ and σ^2 are known (i.e., fixed), the posterior variance $\mathrm{Var}(\boldsymbol{\theta}|\mathbf{w})$ simplifies to $\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) = \mathbf{G}_{1,\sigma^2}(\mathbf{C})$; if σ^2 is known, but $\boldsymbol{\beta}$ is unknown (i.e., random), $\mathrm{Var}(\boldsymbol{\theta}|\mathbf{w})$ simplifies to $\mathrm{Var}(\boldsymbol{\theta}|\sigma^2, \mathbf{w}) = \mathbf{G}_{1,\sigma^2}(\mathbf{C}) + \mathbf{G}_{2,\sigma^2}(\mathbf{C})$. The terms $\mathbf{G}_{1,\sigma^2}(\mathbf{C})$ and $\mathbf{G}_{2,\sigma^2}(\mathbf{C})$, which are $m \times m$ matrices, can be explicitly compared with those when $\mathbf{C} = \mathbf{I}_m$, that is, when $\mathbf{y}$ is available. We note the following theorem, whose proof is given in Section 4.6.5.

Theorem 4.3. *For the $m \times m$ matrix terms $\mathbf{G}_{1,\sigma^2}(\mathbf{C})$ and $\mathbf{G}_{2,\sigma^2}(\mathbf{C})$ from the decomposition of the posterior variance $\mathrm{Var}(\boldsymbol{\theta}|\mathbf{w})$ in Theorem 4.2, the following inequalities hold:*

$$\mathbf{G}_{1,\sigma^2}(\mathbf{C}) \geq \mathbf{G}_{1,\sigma^2}(\mathbf{I}_m) \text{ and } \mathbf{G}_{2,\sigma^2}(\mathbf{C}) \geq \mathbf{G}_{2,\sigma^2}(\mathbf{I}_m),$$

where for two square matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \geq \mathbf{B}$ indicates that $\mathbf{A} - \mathbf{B}$ is nonnegative definite.

4.4 A Real Data Example: Median Income Data

In this section, we consider several levels of data aggregation using a real dataset and fit the HB-GFH model (4.9) to estimate the small area means at the desired lower level. Our dataset is called the median income data and our goal is to estimate the four-person family median incomes of the 48 contiguous U.S. states and the District of Columbia (i.e., $m = 49$ small areas) for the year 1989. The response $y_i,\ i = 1, \ldots, m$, is the direct estimate of the 1989 four-person family median income of the ith state from the Current Population Survey (CPS). The CPS estimates suffer from a small sample size problem, which causes substantial variability and limits their direct use. On the other hand, the estimates from the 1990 census have negligible standard errors and can be treated as the true values for the 1989 four-person family median incomes. For the covariates, we use the following two variables as suggested by Fay (1987):

(1) x_{i1}: 1979 four-person family median income of the ith state from the 1980 census;

(2) $x_{i2} = (\text{PCI}_{i,1989}/\text{PCI}_{i,1979}) \cdot x_{i1}$: 1989 adjusted census four-person family median income of the ith state,

where PCI is the per capita income from the U.S. Bureau of Economic Analysis. Then the design matrix and the regression coefficient for the HB-GFH model (4.9) are $\mathbf{X} = (\mathbf{x}_i^T)_{1 \le i \le m} = ((1, x_{i1}, x_{i2})^T)_{1 \le i \le m}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, respectively.

4.4.1 A Simple Case: Two States Combined

We first consider a simple case where 2 of the 49 states in the median income data are combined such that we have only one direct estimator for the 2 states that is unbiased for the average of their true median incomes. That is, if the first two states are the states to be combined, we are considering the situation where we have only $(y_1 + y_2)/2$, which is unbiased for $(\theta_1 + \theta_2)/2$, instead of the individual y_1 and y_2. In this case, $r = m - 1 = 48$ and the $\mathbf{C}$ matrix for the HB-GFH model (4.9) is

$$\mathbf{C}_{(m-1)\times m} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{m-2} \end{pmatrix}.$$

Then

$$\mathbf{D_w} = \mathbf{CDC}^T = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{m-2} \end{pmatrix} \begin{pmatrix} D_1 & 0 & \mathbf{0}^T \\ 0 & D_2 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{D}^* \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \mathbf{0}^T \\ \frac{1}{2} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{I}_{m-2} \end{pmatrix} = \begin{pmatrix} \frac{D_1+D_2}{4} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{D}^* \end{pmatrix},$$

where $\mathbf{D}^* = \text{diag}(D_i,\ 3 \le i \le m)$. Now,

$$\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} = \begin{pmatrix} \frac{1}{2} & \mathbf{0}^T \\ \frac{1}{2} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{I}_{m-2} \end{pmatrix} \begin{pmatrix} \left(\frac{D_1+D_2}{4}\right)^{-1} & \mathbf{0}^T \\ \mathbf{0} & (\mathbf{D}^*)^{-1} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{m-2} \end{pmatrix}$$
$$= \begin{pmatrix} (D_1+D_2)^{-1} & (D_1+D_2)^{-1} & \mathbf{0}^T \\ (D_1+D_2)^{-1} & (D_1+D_2)^{-1} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & (\mathbf{D}^*)^{-1} \end{pmatrix}$$

and

$$\begin{aligned} \text{Var}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) &= \mathbf{G}_{1,\sigma^2}(\mathbf{C}) \\ &= \{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1} \\ &= \begin{pmatrix} (D_1+D_2)^{-1} + (\sigma^2)^{-1} & (D_1+D_2)^{-1} & \mathbf{0}^T \\ (D_1+D_2)^{-1} & (D_1+D_2)^{-1} + (\sigma^2)^{-1} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & (\mathbf{D}^*)^{-1} + (\sigma^2)^{-1}\mathbf{I}_{m-2} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \frac{\sigma^2(\sigma^2+D_1+D_2)}{2\sigma^2+D_1+D_2} & -\frac{(\sigma^2)^2}{2\sigma^2+D_1+D_2} & \mathbf{0}^T \\ -\frac{(\sigma^2)^2}{2\sigma^2+D_1+D_2} & \frac{\sigma^2(\sigma^2+D_1+D_2)}{2\sigma^2+D_1+D_2} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & \{(\mathbf{D}^*)^{-1} + (\sigma^2)^{-1}\mathbf{I}_{m-2}\}^{-1} \end{pmatrix}. \end{aligned}$$

Clearly,

$$\text{Var}(\theta_1|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) = \text{Var}(\theta_2|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) = \frac{\sigma^2(\sigma^2 + D_1 + D_2)}{2\sigma^2 + D_1 + D_2}$$

is different from $\text{Var}(\theta_1|\boldsymbol{\beta}, \sigma^2, \mathbf{y}) = \sigma^2 D_1/(\sigma^2 + D_1)$ and $\text{Var}(\theta_2|\boldsymbol{\beta}, \sigma^2, \mathbf{y}) = \sigma^2 D_2/(\sigma^2 + D_2)$. In fact, when two small areas are combined as above, the posterior variance of θ_1 given $\boldsymbol{\beta}$

and σ^2 increases by a factor of

$$\begin{aligned}\frac{\text{Var}(\theta_1|\boldsymbol{\beta},\sigma^2,\mathbf{w})}{\text{Var}(\theta_1|\boldsymbol{\beta},\sigma^2,\mathbf{y})} &= \frac{\sigma^2(\sigma^2+D_1+D_2)}{2\sigma^2+D_1+D_2}\Big/\frac{\sigma^2 D_1}{\sigma^2+D_1} \\ &= \frac{(\sigma^2+D_1)(\sigma^2+D_1+D_2)}{D_1(2\sigma^2+D_1+D_2)} \\ &= \frac{D_1(2\sigma^2+D_1+D_2)+\sigma^2(\sigma^2+D_2)}{D_1(2\sigma^2+D_1+D_2)},\end{aligned}$$

which is greater than 1 as implied by Theorem 4.3. Given $\boldsymbol{\beta}$ and σ^2, the increase in the posterior variance will be substantial if D_2 is large.

For the median income data, each state has been classified as a direct-use state or an indirect-use state, depending on the available sample size for the state in the CPS sample. If the sample size available from a state is large, then it is classified as a direct-use state; otherwise, it is classified as an indirect-use state. There are 11 direct-use states in the median income data: California, Florida, Illinois, Massachusetts, Michigan, New Jersey, New York, North Carolina, Ohio, Pennsylvania, and Texas. The remaining states including the District of Columbia are considered indirect-use states. The sampling variances of the direct estimates in the median income data are shown in Figure 4.1. Note that the direct-use states tend to have smaller sampling variances, while there are a few indirect-use states with smaller sampling variances than Massachusetts.

For the two states to be combined, we choose (1) Georgia (GA) and South Carolina (SC); (2) North Carolina (NC) and Virginia (VA). In the first case, we have two neighboring indirect-use states (i.e., GA and SC), whereas in the second case, we have two neighboring direct-use and indirect-use states (i.e., NC and VA). The direct estimates, true values, and sampling variances of these four states are presented in Table 4.1. Clearly, the direct-use state NC has the smallest sampling variance, which is only 26% of that of SC and 12% of that of GA or VA.

Now, we fit the HB-GFH model (4.9). Two popular choices for the prior parameter α are 0 and 0.5, where both of them satisfy the propriety condition $1-(r-p)/2 < \alpha < 1$ specified

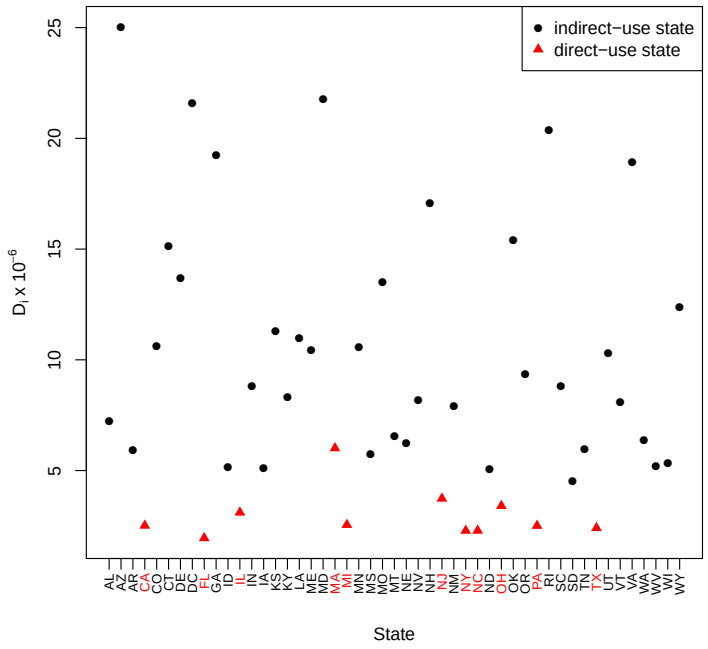

Figure 4.1: Sampling variances of the direct estimates.

Table 4.1: Direct estimates (y_i), true values (θ_i), and sampling variances (D_i) of Georgia, South Carolina, North Carolina, and Virginia.

State	y_i	θ_i	D_i
GA	45876	39036	4385^2
SC	31834	35901	2965^2
NC	37423	37139	1513^2
VA	49718	43223	4349^2

Table 4.2: Posterior means of the model parameters.

Case	β_0	β_1	β_2	σ^2
49-state	1408.3	0.395	0.667	6.43×10^6
(GA, SC)	1976.2	0.379	0.663	6.09×10^6
(NC, VA)	1975.9	0.363	0.671	6.86×10^6

in Theorem 4.1 with $r = 48$ and $p = 3$ ($\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$). According to Berger et al. (2020), however, $\alpha = 0.5$ is a better choice as $\alpha = 0$ does not produce optimal solution in terms of admissiblity. In fact, for the median income data, $\alpha = 0.5$ works better than $\alpha = 0$ in terms of several deviation measures we will introduce soon and our results below are based on $\alpha = 0.5$. In addition to the two combined cases, we also fit the model using the original 49 state-level direct estimates for comparison. Posterior means of the model parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ and σ^2 are reported in Table 4.2, where (State 1, State 2) indicates the two states are combined. In each of the three cases, only β_2 is significant among the regression coefficients as 95% credible intervals for β_0 and β_1 all contain 0 (not shown). Moreover, the posterior means of β_2 are practically the same in all three cases. The posterior means of σ^2 are also close to each other.

Next, we compare the small area mean estimates from the three cases. For each estimate (i.e., posterior mean) $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_i)_{1 \le i \le m}$ of $\boldsymbol{\theta} = (\theta_i)_{1 \le i \le m}$, where $\boldsymbol{\theta}$ is available from the 1990 census, we compute four deviation measures: average absolute deviation $\mathrm{AAD}(\widehat{\boldsymbol{\theta}}) = m^{-1} \sum_{i=1}^m |\widehat{\theta}_i - \theta_i|$, average squared deviation $\mathrm{ASD}(\widehat{\boldsymbol{\theta}}) = m^{-1} \sum_{i=1}^m (\widehat{\theta}_i - \theta_i)^2$, average absolute relative deviation $\mathrm{AARD}(\widehat{\boldsymbol{\theta}}) = m^{-1} \sum_{i=1}^m |(\widehat{\theta}_i - \theta_i)/\theta_i|$, and average squared relative deviation $\mathrm{ASRD}(\widehat{\boldsymbol{\theta}}) = m^{-1} \sum_{i=1}^m \{(\widehat{\theta}_i - \theta_i)/\theta_i\}^2$. In addition, we compute average posterior standard deviation (APSD) of $\theta_i,\ i = 1, \ldots, m$. The deviations are reported in Table 4.3, where the row "Direct" corresponds to the direct estimates from the CPS and APSD in that case is actually the average of the sampling standard errors $\sqrt{D_i},\ i = 1, \ldots, m$. The number in the parentheses is the percentage increase in the deviation by the combined case over the

Table 4.3: Comparison of small area mean estimates with percentage increases over the 49-state case.

Estimate	Case	AAD	ASD	AARD	ASRD	APSD
Direct	49-state	2801.6	12.32×10^6	0.0716	0.0079	2861.1
HB-GFH	49-state	1263.5	2.71×10^6	0.0322	0.0017	1890.2
HB-GFH	(GA, SC)	1208.8	2.54×10^6	0.0309	0.0016	1875.7
		(−4.32%)	(−6.38%)	(−4.28%)	(−6.39%)	(−0.77%)
HB-GFH	(NC, VA)	1311.6	2.84×10^6	0.0335	0.0018	1946.5
		(3.81%)	(4.82%)	(3.99%)	(5.71%)	(2.98%)

49-state case (based on the HB-GFH model), with negative percentage indicating decrease. In the (NC, VA) case, the increases are somewhere between 3.8% to 5.7% in terms of the four deviation measures and it is 3% for APSD. In the (GA, SC) case, on the other hand, the fit has actually improved compared to the 49-state case and APSD has even decreased slightly. While this is unexpected, it is not totally implausible given that we combined only 2 states out of 49 states. Another possibility is that if two states are highly correlated, which is likely if they are neighbors as in our example, it will not be surprising if we benefit from combining them.

Finally, we compare the posterior variances of the small area means from the combined cases with those from the 49-state case. The posterior variances are shown in Figure 4.2, where the left (resp. right) panel compares the (GA, SC) (resp. (NC, VA)) case with the 49-state case. In each case, two combined states have very similar posterior variances despite their substantially different sampling variances (see Table 4.1). As a result, the one with the smaller sampling variance, especially North Carolina, a direct-use state, has a very dramatic increase in the posterior variance after being combined with the other. In fact, we have already checked that $\mathrm{Var}(\theta_1|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) = \mathrm{Var}(\theta_2|\boldsymbol{\beta}, \sigma^2, \mathbf{w})$ when two small areas are combined by averaging their direct estimates, although they are not exactly the same as $\mathrm{Var}(\theta_1|\mathbf{w})$ or $\mathrm{Var}(\theta_2|\mathbf{w})$ presented here.

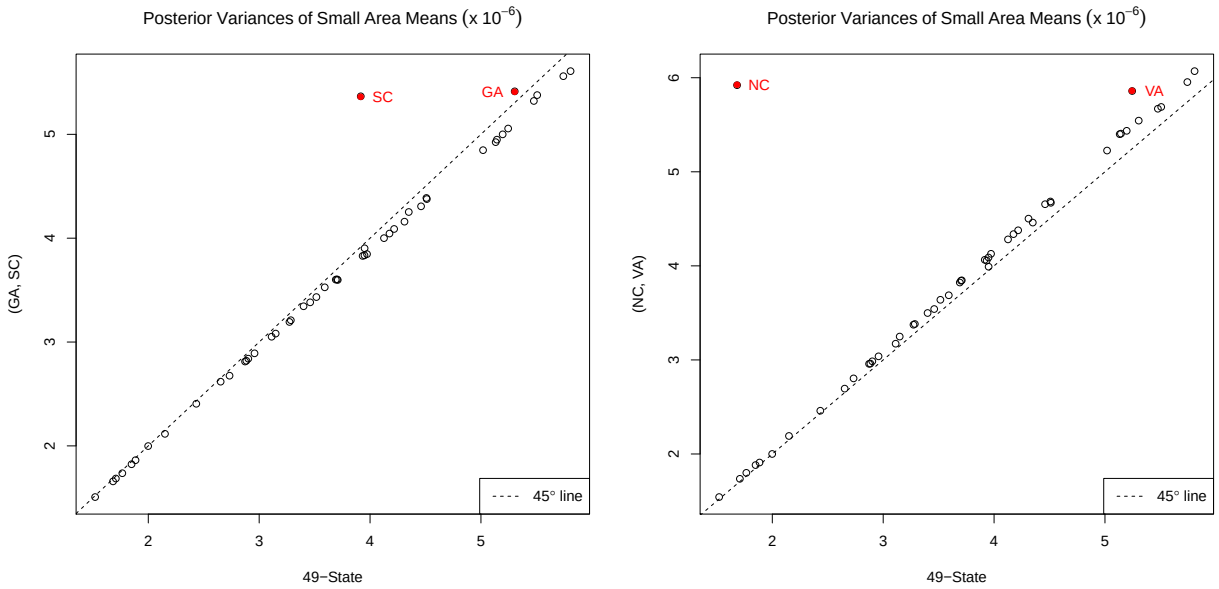

Figure 4.2: Comparisons of posterior variances of small area means.

To investigate the source of increase in the posterior variance, we decompose $\text{Var}(\boldsymbol{\theta}|\mathbf{w})$ as stated in Theorem 4.2. Let $g_{ki}(\mathbf{C}) = \{\mathbf{G}_k(\mathbf{C})\}_{ii},\ k = 1, 2, 3,\ i = 1, \ldots, m$, be the ith diagonal element of $\mathbf{G}_k(\mathbf{C})$, where $\text{Var}(\boldsymbol{\theta}|\mathbf{w}) = \mathbf{G}_1(\mathbf{C}) + \mathbf{G}_2(\mathbf{C}) + \mathbf{G}_3(\mathbf{C})$ according to Theorem 4.2. Then the posterior variance of the ith small area mean can be decomposed as

$$\text{Var}(\theta_i|\mathbf{w}) = g_{1i}(\mathbf{C}) + g_{2i}(\mathbf{C}) + g_{3i}(\mathbf{C}),\ i = 1, \ldots, m.$$

The g_{1i}, g_{2i}, and g_{3i} terms from the combined cases are compared with those from the 49-state case in Figure 4.3. Clearly, the leading term g_{1i} is the dominant term and the remaining terms g_{2i} and g_{3i} are very small or negligible compared to g_{1i}. This implies the increase in the posterior variance of the combined small area mean can be attributed to that in the g_{1i} term.

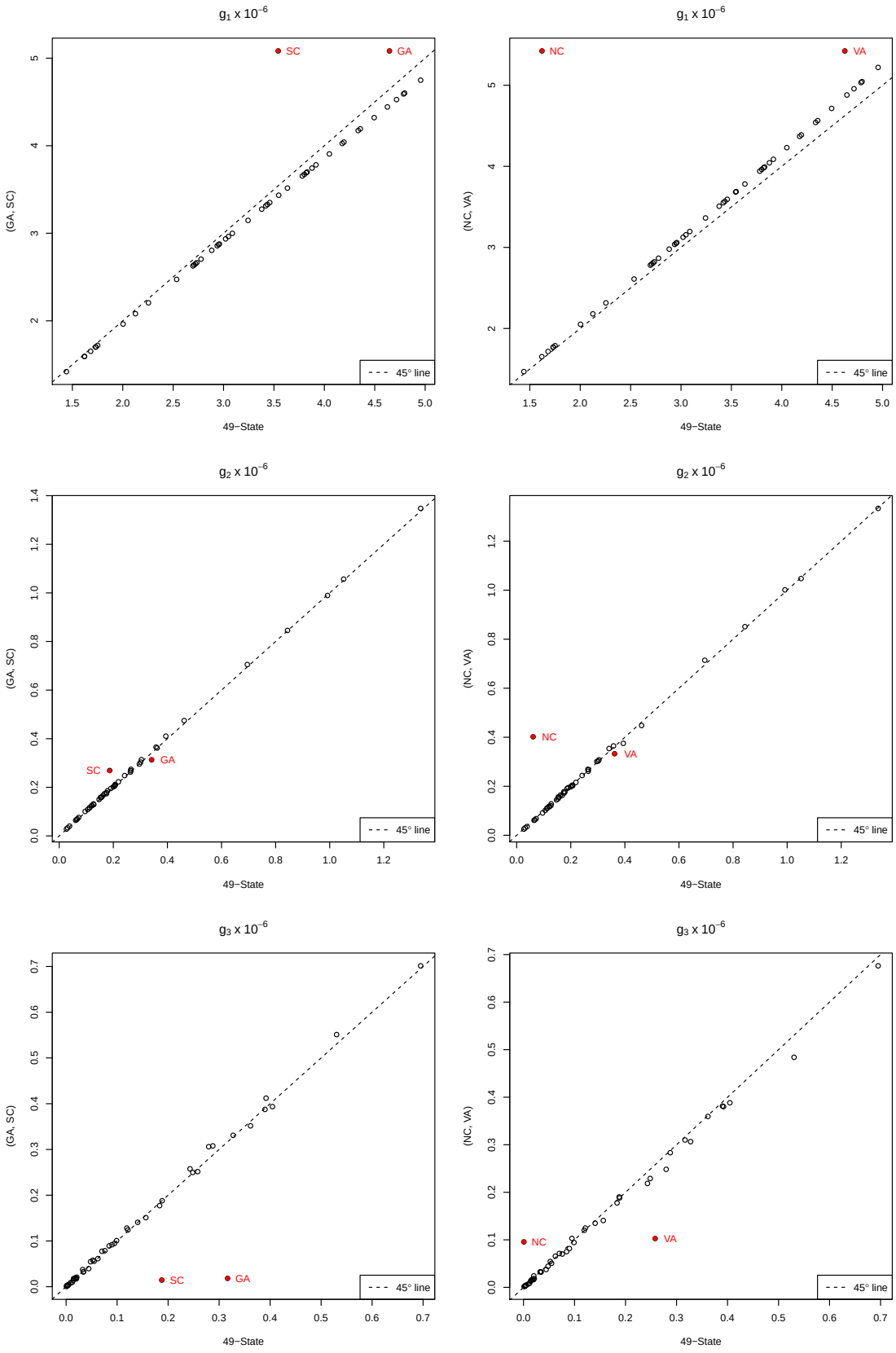

Figure 4.3: Componentwise comparisons of posterior variances of small area means.

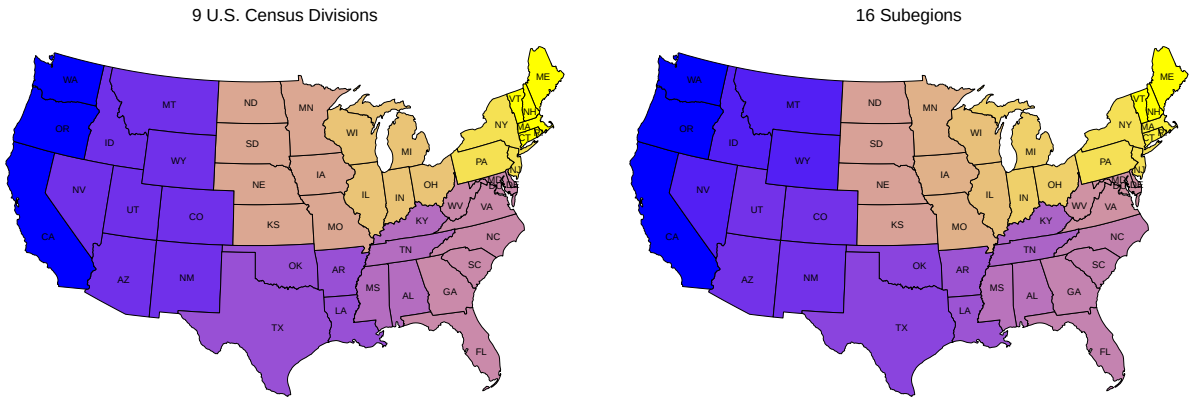

Figure 4.4: Nine U.S. census divisions and 16 subregions.

4.4.2 Higher Levels of Aggregation: 9 U.S. Census Divisions and 16 Subregions

We now consider higher levels of aggregation where each state in the median income data is combined with at least one of its neighboring states. The U.S. Census Bureau divides the United States into four census regions and nine census divisions. Using the nine U.S. census divisions, we consider the situation where we have only 9 division-level direct estimates of median incomes instead of the 49 state-level direct estimates. To obtain the division-level direct estimates, we take the simple average of the state-level direct estimates within the same division. Furthermore, we artificially split the 9 divisions into 16 subregions and calculate 16 subregion-level direct estimates, again by taking averages. The 9 U.S. census divisions and 16 subregions are summarized and shown in Table 4.4 and Figure 4.4. They will help us understand the general impact of using direct estimators at a higher level of aggregation on estimating lower level small area means.

As before, we fit the HB-GFH model (4.9) with $\alpha = 0.5$. Note that in either the 9-division or 16-subregion case (i.e., $r = 9$ or 16), $\alpha = 0.5$ satisfies the propriety condition $1-(r-p)/2 < \alpha < 1$, where $p = 3$. Posterior means of the model parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$

Table 4.4: Nine U.S. census divisions and 16 subregions.

Division	Subregion	State				
1. New England	1	ME	NH	VT		
	2	CT	MA	RI		
2. Middle Atlantic	3	NJ	NY	PA		
3. East North Central	4	IN	MI	OH		
	5	IL	WI			
4. West North Central	6	IA	MN	MO		
	7	KS	ND	NE	SD	
5. South Atlantic	8	DC	DE	MD	VA	WV
	9	FL	GA	NC	SC	
6. East South Central	10	AL	MS			
	11	KY	TN			
7. West South Central	12	AR	LA			
	13	OK	TX			
8. Mountain	14	AZ	CO	NM	UT	
	15	ID	MT	NV	WY	
9. Pacific	16	CA	OR	WA		

Table 4.5: Posterior means of the model parameters (9-division and 16-subregion cases added).

Case	β_0	β_1	β_2	σ^2
49-state	1408.3	0.395	0.667	6.43×10^6
(GA, SC)	1976.2	0.379	0.663	6.09×10^6
(NC, VA)	1975.9	0.363	0.671	6.86×10^6
9-division	-4545.2	0.474	0.771	55.77×10^6
16-subregion	796.3	0.277	0.747	15.15×10^6

and σ^2 are reported in Table 4.5. In the 9-division and 16-subregion cases, β_2 is again the only significant regression coefficient as 95% credible intervals for β_0 and β_1 all contain 0 (not shown). The posterior means of β_2 in these two cases are slightly larger than those in the other cases. On the other hand, there are striking increases in the posterior means of σ^2 in the 9-division and 16-subregion cases compared to the other cases, with the increase being much greater in the 9-division case.

The small area mean estimates (i.e., posterior means) are compared using the four deviation measures (AAD, ASD, AARD, and ASRD) and APSD in Table 4.6. As before, the number in the parentheses is the percentage increase in the deviation over the 49-state case (based on the HB-GFH model), with negative percentage indicating decrease. In terms of the four deviation measures, the increases are somewhere between 4.7% to 7.9% in the 9-division case, whereas in the 16-subregion case, there are decreases of at least 7%, implying the fit has actually improved. On the other hand, there are remarkable increases in APSD in both the 9-division and 16-subregion cases. This is not surprising given the posterior means of σ^2 in those cases, which are substantially larger than that in the 49-state case (see Table 4.5).

Now, recall the decomposition of the posterior variance $\text{Var}(\theta_i|\mathbf{w}) = g_{1i}(\mathbf{C}) + g_{2i}(\mathbf{C}) + g_{3i}(\mathbf{C})$, $i = 1, \ldots, m$. The g_{1i}, g_{2i}, and g_{3i} terms from the 49-state, 16-subregion, and 9-division cases are shown in Figure 4.5. As we have observed in the (GA, SC) and (NC, VA)

Table 4.6: Comparison of small area mean estimates with percentage increases over the 49-state case (9-division and 16-subregion cases added).

Estimate	Case	AAD	ASD	AARD	ASRD	APSD
Direct	49-state	2801.6	12.32×10^6	0.0716	0.0079	2861.1
HB-GFH	49-state	1263.5	2.71×10^6	0.0322	0.0017	1890.2
HB-GFH	(GA, SC)	1208.8	2.54×10^6	0.0309	0.0016	1875.7
		(−4.32%)	(−6.38%)	(−4.28%)	(−6.39%)	(−0.77%)
HB-GFH	(NC, VA)	1311.6	2.84×10^6	0.0335	0.0018	1946.5
		(3.81%)	(4.82%)	(3.99%)	(5.71%)	(2.98%)
HB-GFH	9-division	1357.0	2.84×10^6	0.0348	0.0019	7066.8
		(7.40%)	(4.68%)	(7.81%)	(7.86%)	(273.86%)
HB-GFH	16-subregion	1136.7	2.06×10^6	0.0300	0.0015	3556.0
		(−10.03%)	(−23.93%)	(−7.01%)	(−15.58%)	(88.12%)

cases, it again shows that g_{1i} is the dominant term among the three terms. Moreover, the dominance becomes stronger as we use direct estimates at a higher level of aggregation. Also, note that the states in the same group (i.e., division or subregion) have the same g_{1i} value. This is because we took the simple average when calculating the division/subregion-level direct estimates from the original state-level estimates.

As the last step of our analysis, we check the percentage increase in the posterior standard deviation of small area mean by the 9-division/16-subregion case over the 49-state case. First, suppose both $\boldsymbol{\beta}$ and σ^2 are known. Then the percentage increase is

$$\left(\sqrt{\frac{\mathrm{Var}(\theta_i|\boldsymbol{\beta},\sigma^2,\mathbf{w})}{\mathrm{Var}(\theta_i|\boldsymbol{\beta},\sigma^2,\mathbf{y})}}-1\right)\times 100\% = \left(\sqrt{\frac{g_{1i,\sigma^2}(\mathbf{C})}{g_{1i,\sigma^2}(\mathbf{I}_m)}}-1\right)\times 100\%,\ i=1,\ldots,m, \quad (4.10)$$

where $\mathbf{y}$ is the state-level direct estimate, $\mathbf{w} = \mathbf{C}\mathbf{y}$ is either the division-level or subregion-level direct estimate with a suitable $\mathbf{C}$ matrix, and $g_{1i,\sigma^2}(\mathbf{C}) = \{\mathbf{G}_{1,\sigma^2}(\mathbf{C})\}_{ii}$ is the ith diagonal element of $\mathbf{G}_{1,\sigma^2}(\mathbf{C})$. If we relax the condition and assume σ^2 is known, but $\boldsymbol{\beta}$ is

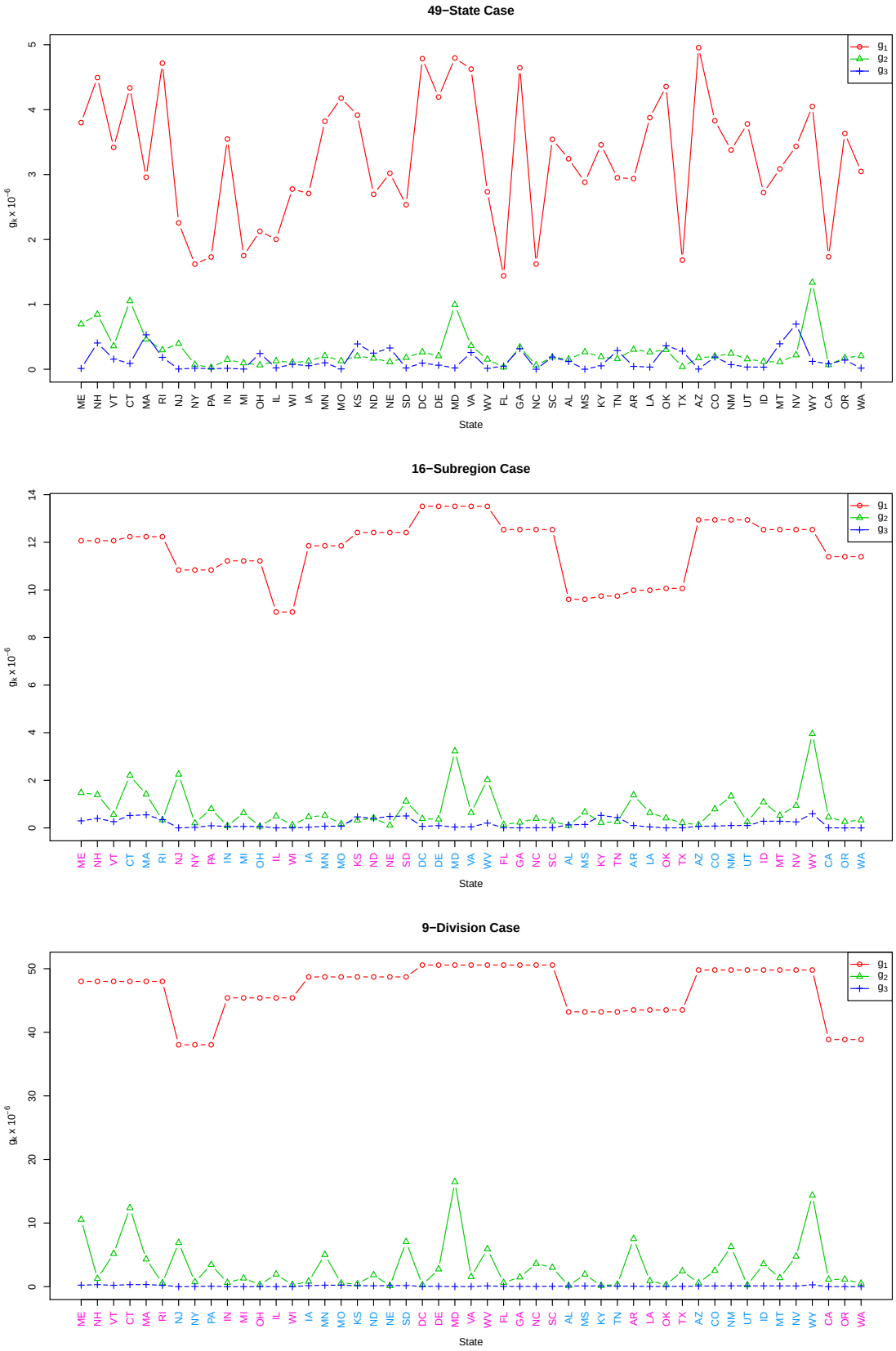

Figure 4.5: Decomposition of the posterior variance of small area mean. On the x-axis, states next to each other in the same color indicate they are in the same division or subregion.

unknown, the percentage increase becomes

$$\left(\sqrt{\frac{\mathrm{Var}(\theta_i|\sigma^2,\mathbf{w})}{\mathrm{Var}(\theta_i|\sigma^2,\mathbf{y})}}-1\right)\times 100\% = \left(\sqrt{\frac{g_{1i,\sigma^2}(\mathbf{C})+g_{2i,\sigma^2}(\mathbf{C})}{g_{1i,\sigma^2}(\mathbf{I}_m)+g_{2i,\sigma^2}(\mathbf{I}_m)}}-1\right)\times 100\%,\ i=1,\ldots,m, \tag{4.11}$$

where $g_{2i,\sigma^2}(\mathbf{C}) = \{\mathbf{G}_{2,\sigma^2}(\mathbf{C})\}_{ii}$ is the ith diagonal element of $\mathbf{G}_{2,\sigma^2}(\mathbf{C})$. Finally, if both $\boldsymbol{\beta}$ and σ^2 are unknown, the percentage increase becomes

$$\left(\sqrt{\frac{\mathrm{Var}(\theta_i|\mathbf{w})}{\mathrm{Var}(\theta_i|\mathbf{y})}}-1\right)\times 100\% = \left(\sqrt{\frac{g_{1i}(\mathbf{C})+g_{2i}(\mathbf{C})+g_{3i}(\mathbf{C})}{g_{1i}(\mathbf{I}_m)+g_{2i}(\mathbf{I}_m)+g_{3i}(\mathbf{I}_m)}}-1\right)\times 100\%,\ i=1,\ldots,m. \tag{4.12}$$

Here, note that $g_{ki}(\mathbf{C}) = \mathrm{E}[g_{ki,\sigma^2}(\mathbf{C})|\mathbf{w}]$ as $\mathbf{G}_k(\mathbf{C}) = \mathrm{E}[\mathbf{G}_{k,\sigma^2}(\mathbf{C})|\mathbf{w}],\ k=1,2$ (see Theorem 4.2). We compute the percentage increases (4.10)–(4.12). Since σ^2 is assumed known in (4.10) and (4.11), we need a fixed value of σ^2 to compute them. We take σ^2 to be $\widehat{\sigma}^2 = 6.43\times 10^6$, the posterior mean of σ^2 in the 49-state case (see Table 4.5). For (4.12), where σ^2 (and $\boldsymbol{\beta}$) is unknown, posterior simulations of σ^2 will be used. The percentage increases are plotted on maps in Figure 4.6 and five number summaries of them are reported in Table 4.7. The state in the parentheses in Table 4.7 corresponds to the relevant percentage increase. As the number of small areas reduces from 16 subregions to 9 divisions, the small area mean estimates become more unstable compared to the ideal 49-state case. The increase in the uncertainty is especially substantial when none of the parameters $\boldsymbol{\beta}$ and σ^2 is known. Nevertheless, all six maps in Figure 4.6 reveal similar patterns of the percentage increases. Throughout the six cases, Florida (FL) and North Carolina (NC) are the two states with the highest percentage increases, while Oklahoma (OK) is the state with the lowest or very low percentage increases.

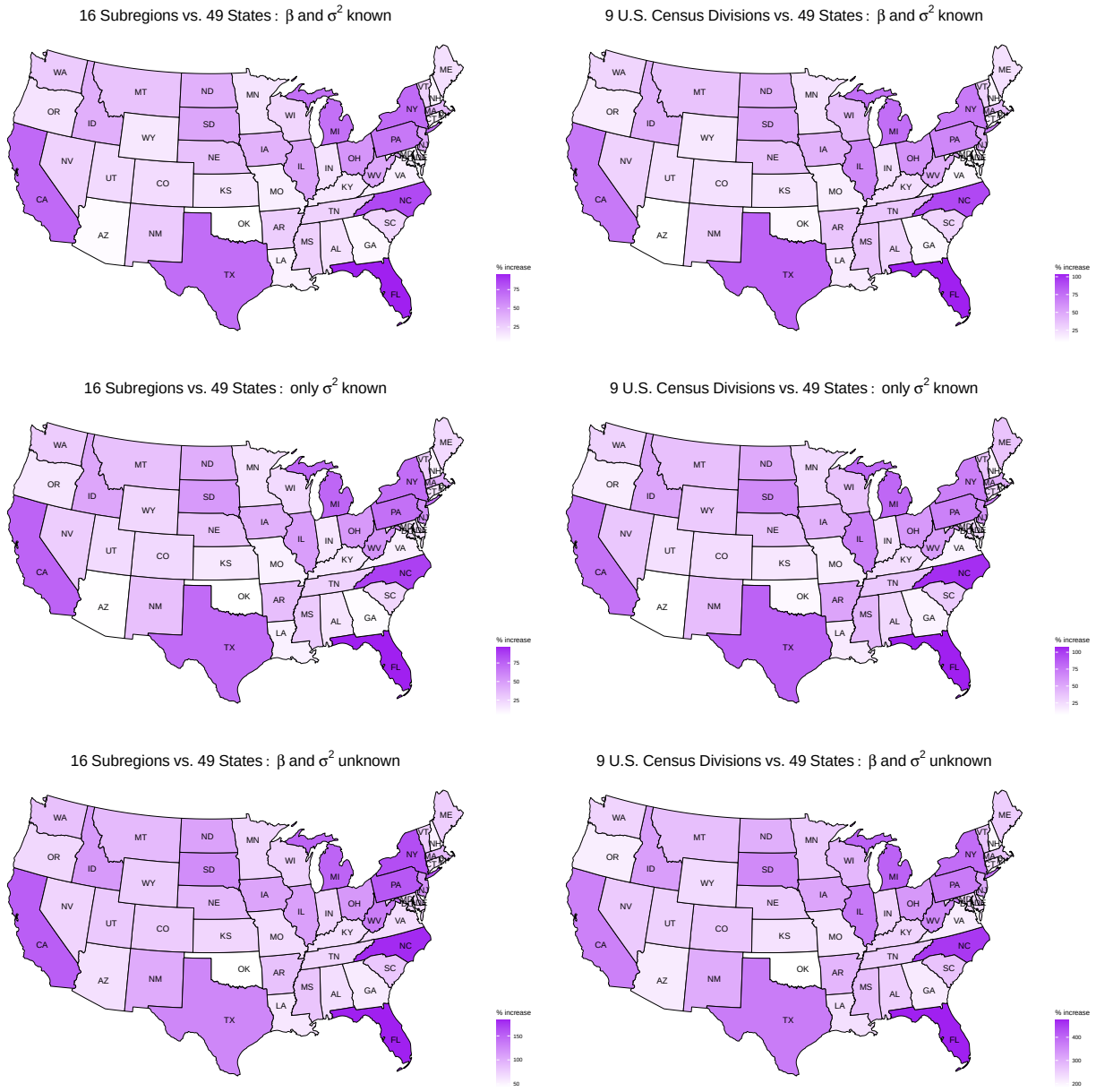

Figure 4.6: Percentage increase in the posterior standard deviation of small area mean by the 9-division/16-subregion case over the 49-state case.

Table 4.7: Five number summaries of the percentage increases (4.10)–(4.12) with corresponding states.

Case	Parameters known	Minimum	Q_1	Median	Q_3	Maximum
16-subregion	$\boldsymbol{\beta}$, σ^2	5.8%	16.3%	25.8%	41.4%	95.4%
		(OK)	(KY)	(VT)	(ID)	(FL)
16-subregion	σ^2	6.4%	17.7%	26.6%	46.3%	96.5%
		(OK)	(OR)	(TN)	(ID)	(FL)
16-subregion	none	44.3%	70.1%	79.7%	110.5%	183.8%
		(OK)	(MD)	(CO)	(IL)	(FL)
9-division	$\boldsymbol{\beta}$, σ^2	9.5%	19.0%	30.1%	45.8%	103.3%
		(AZ)	(LA)	(VT)	(ND)	(FL)
9-division	σ^2	8.8%	22.1%	36.6%	53.5%	107.7%
		(AZ)	(IN)	(NV)	(AR)	(FL)
9-division	none	187.0%	239.3%	262.8%	317.3%	475.8%
		(OK)	(WY)	(NV)	(IA)	(FL)

4.5 Conclusions and Future Directions

We proposed a model-based approach to estimate small area means at a lower level when covariates are available at that lower level, but direct estimates are available only at a higher level of aggregation. If there are direct estimates at the desired lower level, one may use the Fay-Herriot model originally proposed by Fay and Herriot (1979). We generalized the Fay-Herriot model to accommodate the situation where we have only direct estimates that are designed to estimate some known linear combination of the lower level small area means. Then we proposed a hierarchical Bayesian version of the generalized Fay-Herriot model with an improper prior density for the model parameters.

As a real data example, we used the median income data collected from the 48 contiguous U.S. states and the District of Columbia. We suitably combined the 49 state-level direct estimates from the median income data and obtained two sets of higher level direct estimates, namely, 9 division-level and 16 subregion-level direct estimates. Comparisons of small area mean estimates from the two combined cases with those from the 49-state case showed a loss of no more than 8% in accuracy of the estimates in the 9-division case, while there was actually an improvement of at least 7% in the 16-division case. The stability of the small area mean estimates, on the other hand, suffered significantly when states were combined. The percentage increase in APSD of the small area means over the 49-state case was 274% in the 9-division case and was 88% in the 16-subregion case. We also showed that the posterior variance of small area mean can be decomposed into three different components and identified the dominant term among them in the median income data example. In summary, while the increase in the uncertainty is inevitable when no direct estimates are available at the desired level of small areas, our approach enables one to estimate the original lower level small area means by retrieving information from direct estimates at a higher level of aggregation.

For a potential improvement on our approach, note that both the Fay-Herriot model and our hierarchical Bayesian version of the model take an independent covariance structure for

the area-specific random effects by assuming $\text{Var}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2) = \sigma^2 \mathbf{I}_m$. While the independent Fay-Herriot model works well in many applications, the model may benefit from a dependent covariance structure as population means from adjacent areas usually exhibit a spatial pattern. In their recent work, Chung and Datta (2021) considered several spatial mixed effects models by replacing the independent covariance structure in the Fay-Herriot model with some autocorrelation structure such as conditional autoregressive and simultaneous autoregressive structures. Chung and Datta (2021) showed by simulation and example that the independent Fay-Herriot model can be substantially outperformed by its spatial alternatives. Moreover, the improvement was especially remarkable when there was no efficient covariates to sufficiently account for the variation among the small area means. The approach of Chung and Datta (2021) can be directly applied to our problem. It will be a sensible next step of our research as our main challenge is to control the extra variability in the small area mean estimates caused by using direct estimates at a higher level of aggregation.

4.6 Proofs

4.6.1 Proof of Proposition 4.1

Proof. By simple matrix algebra, we rewrite the inverse matrix in (4.5) as

$$(\mathbf{D_w} + \sigma^2 \mathbf{C}\mathbf{C}^T)^{-1} = \mathbf{D_w}^{-1} - \mathbf{D_w}^{-1}\mathbf{C}\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D_w}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D_w}^{-1}$$

and note that

$$\begin{aligned}
&\mathbf{C}^T(\mathbf{D_w} + \sigma^2 \mathbf{C}\mathbf{C}^T)^{-1} \\
=&\mathbf{C}^T\mathbf{D_w}^{-1} - \mathbf{C}^T\mathbf{D_w}^{-1}\mathbf{C}\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D_w}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D_w}^{-1} \\
=&\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D_w}^{-1}\mathbf{C}\}\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D_w}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D_w}^{-1} \\
&- \mathbf{C}^T\mathbf{D_w}^{-1}\mathbf{C}\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D_w}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D_w}^{-1}
\end{aligned}$$

$$
\begin{aligned}
&=\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} - \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1} \\
&=(\sigma^2)^{-1}\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}.
\end{aligned}
$$

Then the expectation in (4.5) can be rewritten as

$$
\begin{aligned}
&\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) \\
=&\mathbf{X}\boldsymbol{\beta} + \sigma^2\mathbf{C}^T(\mathbf{D}_{\mathbf{w}} + \sigma^2\mathbf{C}\mathbf{C}^T)^{-1}(\mathbf{w} - \mathbf{Z}\boldsymbol{\beta}) \\
=&\{\mathbf{I}_m - \sigma^2\mathbf{C}^T(\mathbf{D}_{\mathbf{w}} + \sigma^2\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{C}\}\mathbf{X}\boldsymbol{\beta} + \sigma^2\mathbf{C}^T(\mathbf{D}_{\mathbf{w}} + \sigma^2\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{w} \qquad (\mathbf{Z} = \mathbf{C}\mathbf{X}) \\
=&[\mathbf{I}_m - \{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}]\mathbf{X}\boldsymbol{\beta} + \{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{w} \\
=&[\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\} \\
&- \{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}]\mathbf{X}\boldsymbol{\beta} + \{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{w} \\
=&\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} - \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}\mathbf{X}\boldsymbol{\beta} \\
&+ \{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{w} \\
=&\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}(\sigma^2)^{-1}\mathbf{X}\boldsymbol{\beta} + \{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{w} \\
=&\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\{(\sigma^2)^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{w}\},
\end{aligned}
$$

which is the expectation in (4.7). Similarly, the variance in (4.6) can be rewritten as

$$
\begin{aligned}
&\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w}) \\
=&\sigma^2\mathbf{I}_m - (\sigma^2)^2\mathbf{C}^T(\mathbf{D}_{\mathbf{w}} + \sigma^2\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{C} \\
=&\sigma^2\mathbf{I}_m - \sigma^2\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} \\
=&\sigma^2\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\} \\
&- \sigma^2\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} \\
=&\sigma^2\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1}\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} - \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\} \\
=&\{(\sigma^2)^{-1}\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C}\}^{-1},
\end{aligned}
$$

which is the variance in (4.8). □

4.6.2 Proof of Theorem 4.1

Proof. First, note that $\mathbf{w}|\boldsymbol{\beta},\sigma^2 \sim \mathrm{N}(\mathbf{Z}\boldsymbol{\beta},\boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \mathbf{D}_{\mathbf{w}} + \sigma^2\mathbf{C}\mathbf{C}^T = \mathbf{C}\mathbf{D}\mathbf{C}^T + \sigma^2\mathbf{C}\mathbf{C}^T = \mathbf{C}(\mathbf{D} + \sigma^2\mathbf{I}_m)\mathbf{C}^T$. Then the posterior density of $\boldsymbol{\beta}$ and σ^2 is

$$\pi(\boldsymbol{\beta},\sigma^2|\mathbf{w}) \propto \pi(\mathbf{w}|\boldsymbol{\beta},\sigma^2)\pi(\boldsymbol{\beta},\sigma^2) \propto |\boldsymbol{\Omega}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{w}-\mathbf{Z}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\mathbf{w}-\mathbf{Z}\boldsymbol{\beta})](\sigma^2)^{-\alpha}$$

and

$$\begin{aligned}
&\int_{\mathbb{R}^p} |\boldsymbol{\Omega}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{w}-\mathbf{Z}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\mathbf{w}-\mathbf{Z}\boldsymbol{\beta})](\sigma^2)^{-\alpha} d\boldsymbol{\beta} \\
=&(\sigma^2)^{-\alpha}|\boldsymbol{\Omega}|^{-\frac{1}{2}} \int_{\mathbb{R}^p} \exp[-\frac{1}{2}\{(\mathbf{Z}\boldsymbol{\beta}-\mathbf{Z}\widehat{\boldsymbol{\beta}})^T\boldsymbol{\Omega}^{-1}(\mathbf{Z}\boldsymbol{\beta}-\mathbf{Z}\widehat{\boldsymbol{\beta}}) + (\mathbf{w}-\mathbf{Z}\widehat{\boldsymbol{\beta}})^T\boldsymbol{\Omega}^{-1}(\mathbf{w}-\mathbf{Z}\widehat{\boldsymbol{\beta}})\}] d\boldsymbol{\beta} \\
=&(\sigma^2)^{-\alpha}|\boldsymbol{\Omega}|^{-\frac{1}{2}} \int_{\mathbb{R}^p} \exp[-\frac{1}{2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})] d\boldsymbol{\beta} \exp[-\frac{1}{2}(\mathbf{w}-\mathbf{Z}\widehat{\boldsymbol{\beta}})^T\boldsymbol{\Omega}^{-1}(\mathbf{w}-\mathbf{Z}\widehat{\boldsymbol{\beta}})] \\
\leq&C(\sigma^2)^{-\alpha}|\boldsymbol{\Omega}|^{-\frac{1}{2}}|\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z}|^{-\frac{1}{2}},
\end{aligned}$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{w}$ and C is a generic positive constant that does not depend on $\boldsymbol{\beta}$ and σ^2. Now, let $\lambda_1 \leq \cdots \leq \lambda_r$ be the r positive eigenvalues of $\mathbf{C}\mathbf{C}^T$ (recall that $\mathbf{C}$ is $r \times m$ with rank r). Also, let $D_{(1)} = \min_{1\leq i\leq m} D_i$ and $D_{(m)} = \max_{1\leq i\leq m} D_i$. Note that $(D_{(1)}+\sigma^2)\lambda_1\mathbf{I}_r \leq \boldsymbol{\Omega} \leq (D_{(m)}+\sigma^2)\lambda_r\mathbf{I}_r$. This implies

$$\begin{aligned}
&|\boldsymbol{\Omega}|^{-\frac{1}{2}} \leq \{(D_{(1)}+\sigma^2)\lambda_1\}^{-\frac{r}{2}} \leq \{\min(D_{(1)}\lambda_1,\lambda_1)\}^{-\frac{r}{2}}(1+\sigma^2)^{-\frac{r}{2}}, \\
&|\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z}|^{-\frac{1}{2}} \leq \{(D_{(m)}+\sigma^2)\lambda_r\}^{\frac{p}{2}}|\mathbf{Z}^T\mathbf{Z}|^{-\frac{1}{2}} \leq \{\max(D_{(m)}\lambda_r,\lambda_r)\}^{\frac{p}{2}}(1+\sigma^2)^{\frac{p}{2}}|\mathbf{Z}^T\mathbf{Z}|^{-\frac{1}{2}},
\end{aligned}$$

and

$$\begin{aligned}
|\boldsymbol{\Omega}|^{-\frac{1}{2}}|\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z}|^{-\frac{1}{2}} &\leq \{\min(D_{(1)}\lambda_1,\lambda_1)\}^{-\frac{r}{2}}(1+\sigma^2)^{-\frac{r}{2}}\{\max(D_{(m)}\lambda_r,\lambda_r)\}^{\frac{p}{2}}(1+\sigma^2)^{\frac{p}{2}}|\mathbf{Z}^T\mathbf{Z}|^{-\frac{1}{2}} \\
&= C(1+\sigma^2)^{-\frac{r-p}{2}}.
\end{aligned}$$

Finally,

$$\begin{aligned}
&\int_0^\infty \int_{\mathbb{R}^p} |\boldsymbol{\Omega}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{w}-\mathbf{Z}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\mathbf{w}-\mathbf{Z}\boldsymbol{\beta})](\sigma^2)^{-\alpha} d\boldsymbol{\beta} d\sigma^2 \\
\leq & C\int_0^\infty (\sigma^2)^{-\alpha}|\boldsymbol{\Omega}|^{-\frac{1}{2}}|\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z}|^{-\frac{1}{2}} d\sigma^2 \\
\leq & C\int_0^\infty (\sigma^2)^{-\alpha}(1+\sigma^2)^{-\frac{r-p}{2}} d\sigma^2 \\
= & C\cdot B\left(1-\alpha, \frac{r-p}{2}-(1-\alpha)\right) \\
< & \infty,
\end{aligned}$$

where $\mathrm{B}(\cdot,\cdot)$ is the beta function and the equality follows provided that

$$1-\alpha>0,\ \frac{r-p}{2}-(1-\alpha)>0 \Longleftrightarrow 1-\frac{r-p}{2}<\alpha<1.$$

Therefore, the posterior density of the HB-GFH model (4.9) is proper if $1-(r-p)/2<\alpha<1$. □

4.6.3 Proof of Lemma 4.1

Proof. Based on the HB-GFH model (4.9), the posterior density of $\boldsymbol{\beta}$ given σ^2 and $\mathbf{w}$ is

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\sigma^2,\mathbf{w}) &\propto \pi(\mathbf{w}|\boldsymbol{\beta},\sigma^2)\pi(\boldsymbol{\beta},\sigma^2) \\
&\propto |\boldsymbol{\Omega}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{w}-\mathbf{Z}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\mathbf{w}-\mathbf{Z}\boldsymbol{\beta})](\sigma^2)^{-\alpha} \\
&\propto \exp[-\frac{1}{2}\boldsymbol{\beta}^T\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{w}],
\end{aligned}$$

which is the kernel of a $\mathrm{N}\left((\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{w}, (\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\right)$ distribution. □

4.6.4 Proof of Theorem 4.2

Proof. The proof is a series of applications of the laws of iterated expectations and variances. First, note that

$$
\begin{aligned}
&\mathrm{Var}(\boldsymbol{\theta}|\sigma^2, \mathbf{w}) \\
=&\mathrm{E}[\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w})|\sigma^2, \mathbf{w}] + \mathrm{Var}[\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w})|\sigma^2, \mathbf{w}] \\
=&\mathrm{E}[\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1}|\sigma^2, \mathbf{w}] \\
&+ \mathrm{Var}[\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1}\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{w} + (\sigma^2)^{-1}\mathbf{X}\boldsymbol{\beta}\}|\sigma^2, \mathbf{w}] \quad \text{(by (4.7) and (4.8))} \\
=&\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1} \\
&+ (\sigma^2)^{-2}\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1}\mathbf{X} \cdot \mathrm{Var}(\boldsymbol{\beta}|\sigma^2, \mathbf{w}) \cdot \mathbf{X}^T\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1} \\
=&\mathbf{G}_{1,\sigma^2}(\mathbf{C}) + (\sigma^2)^{-2}\{\mathbf{G}_{1,\sigma^2}(\mathbf{C})\}\mathbf{X}(\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{X}^T\{\mathbf{G}_{1,\sigma^2}(\mathbf{C})\} \quad \text{(by Lemma 4.1)} \\
=&\mathbf{G}_{1,\sigma^2}(\mathbf{C}) + \mathbf{G}_{2,\sigma^2}(\mathbf{C}).
\end{aligned}
$$

Next, note that

$$
\begin{aligned}
&\mathrm{E}(\boldsymbol{\theta}|\sigma^2, \mathbf{w}) \\
=&\mathrm{E}[\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{w})|\sigma^2, \mathbf{w}] \\
=&\mathrm{E}[\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1}\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{w} + (\sigma^2)^{-1}\mathbf{X}\boldsymbol{\beta}\}|\sigma^2, \mathbf{w}] \quad \text{(by (4.7))} \\
=&\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1}\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{w} + (\sigma^2)^{-1}\mathbf{X} \cdot \mathrm{E}(\boldsymbol{\beta}|\sigma^2, \mathbf{w})\} \\
=&\{\mathbf{G}_{1,\sigma^2}(\mathbf{C})\}\{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{w} + (\sigma^2)^{-1}\mathbf{X}(\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{w}\} \quad \text{(by Lemma 4.1)} \\
=&\mathbf{g}_{3,\sigma^2}(\mathbf{C}).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\mathrm{Var}(\boldsymbol{\theta}|\mathbf{w}) &= \mathrm{E}[\mathrm{Var}(\boldsymbol{\theta}|\sigma^2, \mathbf{w})|\mathbf{w}] + \mathrm{Var}[\mathrm{E}(\boldsymbol{\theta}|\sigma^2, \mathbf{w})|\mathbf{w}] \\
&= \mathrm{E}[\mathbf{G}_{1,\sigma^2}(\mathbf{C}) + \mathbf{G}_{2,\sigma^2}(\mathbf{C})|\mathbf{w}] + \mathrm{Var}[\mathbf{g}_{3,\sigma^2}(\mathbf{C})|\mathbf{w}]
\end{aligned}
$$

$$= \mathrm{E}[\mathbf{G}_{1,\sigma^2}(\mathbf{C})|\mathbf{w}] + \mathrm{E}[\mathbf{G}_{2,\sigma^2}(\mathbf{C})|\mathbf{w}] + \mathrm{Var}[\mathbf{g}_{3,\sigma^2}(\mathbf{C})|\mathbf{w}]$$
$$= \mathbf{G}_1(\mathbf{C}) + \mathbf{G}_2(\mathbf{C}) + \mathbf{G}_3(\mathbf{C}).$$

□

4.6.5 Proof of Theorem 4.3

Proof. Recall that the matrix $\mathbf{C}$ is $r \times m$ with rank $r(\leq m)$, where $r = m$ if and only if $\mathbf{C} = \mathbf{I}_m$. For $r < m$, let $\mathbf{F}$ be an $(m-r) \times m$ matrix such that the $m \times m$ matrix $\begin{pmatrix}\mathbf{C}^T & \mathbf{F}^T\end{pmatrix}^T$ is nonsingular and $\mathbf{CDF}^T = \mathbf{0}$. Then

$$\begin{pmatrix}\mathbf{C}\\ \mathbf{F}\end{pmatrix} \mathbf{D} \begin{pmatrix}\mathbf{C}^T & \mathbf{F}^T\end{pmatrix} = \begin{pmatrix}\mathbf{CDC}^T & \mathbf{CDF}^T\\ \mathbf{FDC}^T & \mathbf{FDF}^T\end{pmatrix} = \begin{pmatrix}\mathbf{CDC}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{FDF}^T\end{pmatrix}$$

and

$$\begin{aligned}\mathbf{D}^{-1} &= \begin{pmatrix}\mathbf{C}^T & \mathbf{F}^T\end{pmatrix} \begin{pmatrix}(\mathbf{CDC}^T)^{-1} & \mathbf{0}\\ \mathbf{0} & (\mathbf{FDF}^T)^{-1}\end{pmatrix} \begin{pmatrix}\mathbf{C}\\ \mathbf{F}\end{pmatrix}\\ &= \mathbf{C}^T(\mathbf{CDC}^T)^{-1}\mathbf{C} + \mathbf{F}^T(\mathbf{FDF}^T)^{-1}\mathbf{F}\\ &= \mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + \mathbf{F}^T(\mathbf{FDF}^T)^{-1}\mathbf{F}.\end{aligned}$$

This implies

$$\mathbf{G}_{1,\sigma^2}(\mathbf{I}_m) = \{\mathbf{D}^{-1} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1} \leq \{\mathbf{C}^T\mathbf{D}_{\mathbf{w}}^{-1}\mathbf{C} + (\sigma^2)^{-1}\mathbf{I}_m\}^{-1} = \mathbf{G}_{1,\sigma^2}(\mathbf{C}),$$

establishing the first inequality.

To show the second inequality, we first note that when $\mathbf{C} = \mathbf{I}_m$, $\mathbf{Z} = \mathbf{CX}$ becomes $\mathbf{X}$ and $\mathbf{\Omega} = \mathbf{D}_{\mathbf{w}} + \sigma^2\mathbf{CC}^T = \mathbf{CDC}^T + \sigma^2\mathbf{CC}^T = \mathbf{C}(\mathbf{D} + \sigma^2\mathbf{I}_m)\mathbf{C}^T$ becomes $\mathbf{\Sigma} \triangleq \mathbf{D} + \sigma^2\mathbf{I}_m$. Moreover, there exists an $m \times m$ nonsingular matrix $\mathbf{L}$ such that $\mathbf{\Sigma}^{-1} = \mathbf{LL}^T$ since $\mathbf{\Sigma} > \mathbf{0}$

(i.e., positive definite). Now, note that

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1} - \mathbf{C}^T\boldsymbol{\Omega}^{-1}\mathbf{C} &= \boldsymbol{\Sigma}^{-1} - \mathbf{C}^T(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T)^{-1}\mathbf{C} \\
&= \mathbf{L}\mathbf{L}^T - \mathbf{C}^T\{\mathbf{C}(\mathbf{L}\mathbf{L}^T)^{-1}\mathbf{C}^T\}^{-1}\mathbf{C} \\
&= \mathbf{L}\mathbf{L}^T - \mathbf{C}^T\{(\mathbf{L}^{-1}\mathbf{C}^T)^T(\mathbf{L}^{-1}\mathbf{C}^T)\}^{-1}\mathbf{C} \\
&= \mathbf{L}[\mathbf{I}_m - (\mathbf{L}^{-1}\mathbf{C}^T)\{(\mathbf{L}^{-1}\mathbf{C}^T)^T(\mathbf{L}^{-1}\mathbf{C}^T)\}^{-1}(\mathbf{L}^{-1}\mathbf{C}^T)^T]\mathbf{L}^T \\
&= \mathbf{L}\{\mathbf{I}_m - \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\}\mathbf{L}^T \\
&= [\mathbf{L}\{\mathbf{I}_m - \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\}][\mathbf{L}\{\mathbf{I}_m - \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\}]^T \\
&\geq \mathbf{0},
\end{aligned}$$

where $\mathbf{K} = \mathbf{L}^{-1}\mathbf{C}^T$. This implies

$$(\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1} = (\mathbf{X}^T\mathbf{C}^T\boldsymbol{\Omega}^{-1}\mathbf{C}\mathbf{X})^{-1} \geq (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}$$

and

$$\begin{aligned}
\mathbf{G}_{2,\sigma^2}(\mathbf{C}) &= (\sigma^2)^{-2}\{\mathbf{G}_{1,\sigma^2}(\mathbf{C})\}\mathbf{X}(\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{X}^T\{\mathbf{G}_{1,\sigma^2}(\mathbf{C})\} \\
&\geq (\sigma^2)^{-2}\{\mathbf{G}_{1,\sigma^2}(\mathbf{I}_m)\}\mathbf{X}(\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{X}^T\{\mathbf{G}_{1,\sigma^2}(\mathbf{I}_m)\} \qquad (\mathbf{G}_{1,\sigma^2}(\mathbf{C}) \geq \mathbf{G}_{1,\sigma^2}(\mathbf{I}_m)) \\
&\geq (\sigma^2)^{-2}\{\mathbf{G}_{1,\sigma^2}(\mathbf{I}_m)\}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\{\mathbf{G}_{1,\sigma^2}(\mathbf{I}_m)\} \\
&= \mathbf{G}_{2,\sigma^2}(\mathbf{I}_m),
\end{aligned}$$

concluding the proof. □

Bibliography

[1] Berger, J. O., Sun, D., and Song, C. (2020), "An objective prior for hyperparameters in normal hierarchical models," *Journal of Multivariate Analysis*, 178, 104606.

[2] Bernardo, J. M. (1979), "Expected information as expected utility," *The Annals of Statistics*, 7, 686–690.

[3] Chung, H. C. and Datta, G. S. (2021), "Bayesian hierarchical spatial models for small area estimation," *Preprint*.

[4] Datta, G. S., Hall, P., and Mandal, A. (2011), "Model selection by testing for the presence of small-area effects, and application to area-level data," *Journal of the American Statistical Association*, 106, 362–374.

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

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

[7] Fay, R. E. (1987), "Application of multivariate regression to small domain estimation," *Small Area Statistics*, 91–102.

[8] Fay, R. E. and Herriot, R. A. (1979), "Estimates of income for small places: an appli-

cation of James-Stein procedures to census data," *Journal of the American Statistical Association*, 74, 269–277.

[9] Ganesh, N. (2009), "Simultaneous credible intervals for small area estimation problems," *Journal of Multivariate Analysis*, 100, 1610–1621.

[10] Henderson, N. C., Varadhan, R., and Louis, T. A. (2020), "Improved small area estimation via compromise regression weights," *arXiv preprint arXiv:2006.15638*.

[11] Jiang, J., Nguyen, T., and Rao, J. S. (2010), "Fence method for nonparametric small area estimation," *Survey Methodology*, 36, 3–11.

[12] — (2011), "Best predictive small area estimation," *Journal of the American Statistical Association*, 106, 732–745.

[13] Mallick, B. K., Ghosh, D., and Ghosh, M. (2005), "Bayesian classification of tumours by using gene expression data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 219–234.

[14] Morris, C. N. and Christiansen, C. L. (1996), "Hierarchical models for ranking and for identifying extremes, with applications," *Bayesian Statistics 5*, New York: Oxford University Press, 277–296.

[15] 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.

[16] Rao, J. N. K. and Molina, I. (2015), *Small area estimation*, Hoboken, NJ: John Wiley & Sons, Inc.