ABSTRACT

Title of dissertation: THE BAYESIAN AND APPROXIMATE BAYESIAN METHODS IN SMALL AREA ESTIMATION

Santanu Pramanik, Doctor of Philosophy, 2008

Dissertation directed by: Professor Partha Lahiri
Joint Program in Survey Methodology

For small area estimation, model based methods are preferred to the traditional design based methods because of their ability to borrow strength from related sources. The indirect estimates, obtained using mixed models, are usually more reliable than the direct survey estimates. To draw inferences from mixed models, one can use Bayesian or frequentist approach. We consider the Bayesian approach in this dissertation. The Bayesian approach is straightforward. The prior and likelihood produce the posterior, which is used for all inferential purposes. It overcomes some of the shortcomings of the empirical Bayes approach. For example, the posterior variance automatically captures all sources of uncertainties in estimating small area parameters. But this approach requires the specification of a subjective prior on the model parameters. Moreover, in almost all situation, the posterior moments involve multi-dimensional integration and consequently closed form expressions cannot be obtained. To overcome the computational difficulties one needs to apply computer intensive MCMC methods.

We apply linear mixed normal models (area level and unit level) to draw inferences for small areas when the variable of interest is continuous. We propose and evaluate a new prior distribution for the variance component. We use Laplace approximation to obtain accurate approximations to the posterior moments. The approximations present the Bayesian methodology in a transparent way, which facilitates the interpretation of the methodology to the data users. Our simulation study shows that the proposed prior yields good frequentist properties for the Bayes estimators relative to some other popular choices. This frequentist validation brings in an objective flavor to the so-called subjective Bayesian approach.

The linear mixed models are, usually, not suitable for handling binary or count data, which are often encountered in surveys. To estimate the small area proportions, we propose a binomial-beta hierarchical model. Our formulation allows a regression specification and hence extends the usual exchangeable assumption at the second level. We carefully choose a prior for the shape parameter of the beta density. This new prior helps to avoid the extreme skewness present in the posterior distribution of the model parameters so that the Laplace approximation performs well.

THE BAYESIAN AND APPROXIMATE BAYESIAN METHODS IN SMALL AREA ESTIMATION

by

Santanu Pramanik

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:
Professor Partha Lahiri, Chair/Advisor
Dr. William R. Bell
Dr. Barry Graubard
Professor Paul Smith
Professor Roger Tourangeau

Dedication

To my mother

Acknowledgments

I find this as a perfect opportunity to acknowledge the contribution of all the people who made this thesis possible. Without their help, it was never possible to achieve this level of success. This may appear to be a cliché, but I know how true it is.

First and foremost, I owe sincere gratitude to my advisor Prof. Partha Lahiri. It was a nice learning experience to work with him closely during these four years. It was my privilege to be involved with a researcher of his caliber and competence. The way he interprets the results, thinks about the implicit reasoning behind an analytical expression or the so-called naïve numbers from the computer output, plans about the next step after looking at discouraging results, helped me to grow as a researcher. I would definitely miss the time we spent in discussing research ideas.

I am thankful to other committee members, Prof. Roger Tourangeau, Prof. Paul Smith, Dr. Barry Graubard, and Dr. William, R. Bell. They all were very helpful with their useful comments and suggestions. Their constructive comments substantially improved the final version of the thesis. In particular, I would like to mention the contribution of Dr. William, R. Bell towards the progress of this dissertation. His insightful comments and careful examination of the manuscript identified some mistakes and lead to a correct presentation.

I am fortunate to get the opportunity of doing PhD in the Joint Program in Survey Methodology (JPSM). It is definitely one of the best places for survey re-

search in different dimensions of Survey Methodology. The courses I took in JPSM, from the award-winning faculty, were extremely helpful in clarifying several concepts, many of them were new to me at the beginning. In this regard, I should mention two of my teachers at JPSM, Prof. Richard Valliant (Rick) and Prof. Trivellore Raghunathan (Raghu). Whatever I learnt about the complex survey designs, prediction approach in finite population sampling, Bayesian approaches in surveys, weighting and imputation, can be attributed to the courses I took from Rick and Raghu. They are great teachers, perhaps the best teachers I have ever seen in my life. They also helped me to learn the statistical software `R`.

I would like to thank the support staff in the JPSM. The IT-coordinators Adam Kelley and Duane Gilbert helped me with lot of patience to make me computer-literate. The administrative staff Pam Ainsworth, Rupa Jethwa Eapen, and Sarah Gebremicael always extended their cordial help, whenever I needed them. I want to thank my fellow JPSM mates, Jill (first PhD in my cohort), Michael, Benmei, Aaron, Carolina, Sylvia, Dan, Susan, and Stephanie, to name a few. They helped me to get used to the new environment and enriched my graduate life in many academic and non-academic ways.

This acknowledgement would suffer from a huge incompleteness, if I do not write about my wife Arpita. During these four years she was equally involved in the process in spite of staying five hours apart. When things did not go well, she was more worried than me and put the most sincere effort to instill the ray of optimism in me. For the good part, she was more enthusiastic than me. It is really difficult to describe her role in words for me, who always has to look for the appropriate

words and phrases. I am thankful to the God to have a wife like her endowed with so many good, rare and humane qualities. Kudos to her parents (whom I consider my parents as well), who raised their kid in such a way.

I dedicated this dissertation to my mother. What a nice mom she is! Even staying thousands of miles away, she thought about me everyday, got worried about me for every single minor reason (actually to be worried is one of her favorite pastimes!). Starting from mid-term exams to the job interview, final defense; she prayed to the God for better outcome of every incident in my life. Throughout my life she gave me license and at the same time her love is fathomless. When license is backed by true love, it is difficult to abuse that. I would also like to mention about my dearest sister. Without her nagging effort for mother's consent, it would never been possible to come to the U.S. for a PhD.

At the end, I pray to the Almighty for strength and courage, as He always bestowed these handfuls. Whenever I prayed to Him, I got a bountiful support.

Table of Contents

List of Tables viii

List of Figures ix

1 Literature Review 1
1.1 Small Area Estimation . . . 1
1.2 Linear Mixed Models in Small Area Estimation . . . 5
1.2.1 Area Level Model . . . 6
1.2.2 Unit Level Model . . . 8
1.3 Generalized Linear Mixed Models in Small Area Estimation . . . 11
1.4 Inferential Procedures for Small Area Problems . . . 13
1.4.1 Empirical Bayes (EB) Approach . . . 14
1.4.2 Variance Component Estimation . . . 17
1.4.3 Hierarchical Bayesian Approach . . . 21
1.4.4 Approximations in Hierarchical Bayesian Approach . . . 22
1.4.4.1 Adjusted density method . . . 24
1.4.4.2 Laplace Approximation . . . 27
1.5 Discussion and Outline of the Dissertation . . . 28

2 On the Prior Selection and Approximations in the Fay-Herriot Model 30
2.1 Introduction . . . 30
2.2 Selection of Prior for the Hyperparameters . . . 34
2.3 Approximate Hierarchical Bayes Method . . . 39
2.3.1 Adjusted Density Method . . . 40
2.3.2 Laplace Approximation . . . 41
2.4 Simulation Study . . . 45
2.4.1 Design of the Simulation Study . . . 46
2.4.2 Comparison of Different Estimators of A . . . 47
2.4.3 Comparison of Different Estimators of B_i . . . 48
2.4.4 Comparison of Different Estimators of θ_i . . . 51
2.5 SAIPE Data Analysis . . . 57
2.5.1 Evaluation of the Adjusted Density Method . . . 57
2.5.2 Comparison of Different Estimators of A . . . 59
2.5.3 Results Comparing Different Estimators of B_i . . . 60
2.5.4 Results Comparing Different Estimators of θ_i . . . 61
2.5.5 Evaluation of Laplace Approximation . . . 66
2.6 Concluding Remarks . . . 69
2.7 Appendix A: Verification of the propriety of the posterior distribution $f(A|y)$ of A . . . 70
2.7.1 Propriety of the Posterior Corresponding to $\pi(A) \propto \frac{A}{(A+d_0)^{p/2}}$ . . . 72
2.7.2 Propriety of the Posterior Corresponding to $\pi(A) \propto A$ . . . 73
2.8 Appendix B: Adjusted Density Method . . . 73

2.9 Appendix C: BRugs model to implement the new prior in Fay-Herriot Model . . . 75
2.10 Appendix D: A note on WinBUGS convergence criteria . . . 76

3 The Prior Selection and Approximations for the Nested Error Regression Model: Estimation of Finite Population Mean for Small Areas 79
3.1 Introduction . . . 79
3.2 Estimation of Finite Population Means using Unit level Models: A Review . . . 82
3.3 Hierarchical Bayes Estimation of Finite Population Means . . . 86
3.3.1 Posterior moments of finite population mean when λ known . 90
3.3.2 Choice of Prior on λ . . . 92
3.3.3 Propriety of Posterior . . . 94
3.3.4 Laplace Approximation . . . 94
3.4 Simulation Study . . . 97
3.5 Data Analysis . . . 102
3.6 Concluding Remarks . . . 108
3.7 Appendix . . . 108

4 Hierarchical Bayes Estimation of Binary Data for Small Areas 110
4.1 Introduction . . . 110
4.2 Model and Methodology . . . 112
4.2.1 Hyperparameters Known . . . 114
4.2.2 Hyperparameters are not Known . . . 117
4.2.3 Choice of Prior on the Hyperparameters . . . 119
4.3 Laplace Approximation . . . 123
4.4 Data Analysis . . . 126
4.4.1 Missouri Turkey Hunting Survey Data . . . 126
4.4.2 Baseball Data . . . 130
4.5 Concluding Remarks . . . 137
4.6 Appendix . . . 138
4.6.1 Baseball Data . . . 138
4.6.2 BRugs model specification without covariates . . . 138
4.6.3 BRugs model specification including covariates . . . 139

5 Concluding Remarks and Future Research 140

List of Tables

2.1 Percentage of Zero Estimates for Different Estimators of the Variance Component A 48

2.2 Bias and MSE (in percent) of Different Estimators of the variance component A 48

2.3 Bias (and MSE) in percent of Different Estimators of the shrinkage factor B_i 49

2.4 Comparison of Different Estimators of Small Area Mean θ_i 55

2.5 Relative contribution (%) of the three terms to the posterior variance of θ_i using the prior $\pi_{LL}(A)$ on A: average over 10,000 simulations and 3 small areas within each group 56

2.6 Different variance estimates of A for two years of SAIPE data 61

2.7 Approximations to the small area mean θ_i and their measure of uncertainty using Fay-Herriot model with new prior: SAIPE 1997 data . 63

3.1 Percentage of trials that posterior mode of $\lambda = 0$ 98

3.2 Bias (and MSE) in percent of different estimators of the shrinkage factor B for different choices of λ 100

3.3 MSE of different estimators of the small area mean for different choices of λ 100

3.4 Coverage properties of different estimators of the small area mean for different choices of λ 101

3.5 Different point estimates of mean hectares of corn 105

3.6 Measure of uncertainty of the estimators of mean hectares of corn . . 106

4.1 Results from the Baseball Data Analysis: without using Covariate . . 136

4.2 Results from the Baseball Data Analysis: using 1969 Batting Average as Covariate 137

List of Figures

2.1 Plot of the proposed prior (2.10) for different p (number of covariates) and fixed central tendency measure d_0 (=10) of the sampling variances 39

2.2 Plot of average values of the different estimates of the shrinkage factor B_i along with true values: average being taken over 10,000 simulations 50

2.3 Plot of measure of uncertainty ratios for different estimators of the measure of uncertainty of the small area mean θ_i for 15 small areas . 53

2.4 Evaluation of the adjusted density method proposed by Morris: Approximating the posterior density of the shrinkage factor B_i by a beta density 58

2.5 Plot of the posterior density of A using two years of SAIPE data: a comparison between uniform prior and proposed prior under Fay-Herriot model 60

2.6 Plot of different shrinkage factor estimates using SAIPE 1997 data: Comparison between MEL, REML and our method considering MCMC and numerical integration (exact) as gold standard 62

2.7 Point estimates of small area means θ_i: Comparison between Bell (1999)≡MEL and our≡HB_LL(o2) method (with MCMC as the gold standard) using SAIPE 1997 data (By increasing D_i) 64

2.8 Plot of measure of uncertainty of the estimators of small area means: Comparison between Bell (1999)≡MEL and our≡HB_LL(o2) method (with MCMC as the gold standard) using SAIPE 1997 data (By increasing D_i) 65

2.9 Evaluation of Laplace approximation to the posterior moments of θ_i using SAIPE 1997 data and the new prior: percent difference from the exact as a summary measure 67

2.10 Evaluation of Laplace approximation to the posterior moments of θ_i using SAIPE 1997 data and the uniform prior: percent difference from the exact as a summary measure 68

3.1 Plot of the posterior density of λ using Battese et al. (1988) data: a comparison between uniform prior and proposed prior for λ under nested error regression model 104

3.2 Evaluation of Laplace approximation using new prior for λ: percent difference from EXACT (numerical integration) as a summary measure 107

4.1 Different Point Estimates of Hunting Success Rates 127

4.2 Standard Errors Associated with Different Point Estimates: Turkey Hunting Data 128

4.3 Evaluation of Laplace Approximation using Turkey Hunting Data: Percent Difference from MCMC as a Summary Measure 130

4.4 Comparison of Different Point Estimates: Baseball Data 132

4.5 Comparison of Point Estimates: Laplace Approximation vs MCMC . 133

4.6 Evaluation of Laplace Approximation without using Covariate: Percent Difference from MCMC as a Summary Measure 134

4.7 Evaluation of Laplace Approximation using Covariate: Percent Difference from MCMC as a Summary Measure 135

Chapter 1

Literature Review

1.1 Small Area Estimation

Small area estimation has received considerable attention over the last three decades. This attention reflects the demand for reliable small area estimates for allocating federal funds to local jurisdictions and for regional planning. Small areas can be a geographical region (e.g. state, county, municipality etc.) of a country, a demographic group (a particular sex, race or age group) or a demographic group within a geographical area. In the absence of adequate direct information in small areas, small area estimation technique borrows strength from related sources to produce precise small area estimates.

There are various reasons for the scarcity of direct reliable data on the variables of interest for small areas. In the context of sample survey, national surveys are usually designed to represent the whole nation (large area) and hence cannot guarantee reasonable representation to all the small areas within that large area. Over-sampling is often employed in surveys in order to increase sample sizes for some domains, but that leaves other domains with few sample cases or even no sample cases, since the total sample size is usually fixed by the survey budget. For example, in the National Health and Nutrition Examination Survey (NHANES) III, certain minority groups residing predominantly in certain states (e.g., California

and Texas) were oversampled. This design strategy resulted in small samples for the states that do not have large populations for these minority groups. There are even instances where surveys have no sample in some small areas. For example, until recently, in estimating the number of poor school age (5-17) children in counties, the U.S. Census Bureau used the March Income Supplement of the Current Population Survey (CPS) where more than half of the 3141 counties do not have any CPS sample (release of 2005 estimates mention the change to American Community Survey). This problem of small or zero sample size in small areas prevents the use of direct survey estimates for small area parameters since the estimates are likely to be highly unreliable (i.e. the estimators will have unacceptably large standard errors) or unavailable.

Small area statistics are needed for the planning of reforms, welfare and administration in many fields, including health programs, agriculture, poverty reduction programs. Thus, the importance of producing reliable small area statistics cannot be over-emphasized. For example, health planning often takes place at the small area level (e.g., state, county) since health characteristics are known to vary across geography. The U.S. National Health Planning Resource Development Act of 1974 mandates Health System Agencies to collect and analyze data related to the health status of the residents and the health delivery systems in their health service areas (Nandram, 1999). The U.S. National Center for Health Statictics (NCHS) pioneered the use of synthetic estimation, based on implicit models, to develope state estimates of disability and other characteristics for different groups from the National Health Interview Survey (NHIS) (Rao, 2003). Maps of regional morbid-

ity and mortality rates play an important role in assessing environmental equity (Marshall, 1991). They provide central tools for identifying areas with potentially elevated risk. Hence, mapping the incidence of a disease over different small areas is useful in allocation of government resources to various geographical areas and also to identify factors potentially causing a disease.

The U.S. National Agricultural Service (NASS) publishes crop acerage estimates at the county level using remote sensing satellite data as auxiliary information (Rao, 2003). County estimates assist the agricultural authorities in local agricultural decision making. Also, county crop yield estimates are used to administer federal programs involving payments to farmers if crop yields fall below certain threshold. The U.S. Substance Abuse and Mental Health Administration (SAMHSA) uses National Household Survey on Drug Use and Health (NSDUH) to produce state level and sub-state level (groups of counties or census tracts) small area estimates for more than 20 binary outcomes related to substance use, treatment, and mental health. These estimates are being used for treatment planning purposes by the states.

Small area estimation of variables studied in social surveys is a growing need for government. The Statistical Methodology Division of the Office of National Statistics (ONS) of U.K. established the Small Area Estimation Project (SAEP) in April 1998. The aim of this project was to derive estimates for variables contained in social surveys at the level of political wards (roughly 2000 households). The variables considered in this project are gross weekly household income, average weekly gross household income, number of people to help in a crisis with data from General Household Survey (GHS), Family Resources Survey (FRS)(Heady & Clarke, 2003).

In response to the growing need for precise income and poverty statistics for small areas, the US Census Bureau formed a committee on Small Area Income and Poverty Estimates (SAIPE) in the early 1990s. This committee was created with the goal of providing more timely and precise estimates for subnational areas such as states, counties, school districts, etc., these estimates were needed to allocate government funds. Improving America's Schools Act of 1994 called for the use of updated SAIPE estimates of poor school-age children (aged 5-17) for counties and school districts to allocate more than $7 billion (now over $12 billion) of funds annually for educationally disadvantaged children under Title I of the Elementary and Secondary Education Act (Citro & Kalton, 2000). Thus, small area estimation has wide applicability for survey sampling, disease mapping, poverty mapping, and mapping of health characteristics. For more use and application of small area statistics see Rao (2003), and Jiang & Lahiri (2006b).

In the absence of adequate direct information for small areas, it is customary to borrow strength from related sources to form indirect estimators that increase the effective sample size and hence reduce the sampling errors of the estimators. Such indirect estimators are usually based on implicit or explicit models which combine information from the sample survey, various administrative/census records, or previous surveys. In the disease mapping problem, where the main concern is the small population size, one can consider borrowing strength by exploiting the possible correlation among the neighboring areas and/or past disease incidence information for the small area under consideration (Jiang & Lahiri, 2006b).

Once relevant sources of information are identified for a particular small area

of interest, a model is developed. Various indirect methods that combine information using implicit models have been discussed in Ghosh & Rao (1994) and Rao (2003). Formal evaluation for indirect methods, such as synthetic and composite estimation, that use implicit models is problematic, because the model involved is not spelled out. A synthetic estimator for small areas involves a reliable direct estimator for a large area, covering several small areas. For example, a regression synthetic estimator for a particular area uses data from all the areas to estimate the regression coefficient. Synthetic estimator is derived under the assumption that the small areas have the same characteristics as the large area. This strong assumption often leads to the synthetic estimator having large bias. An intuitive way to balance the potential bias of a synthetic estimator against the instability of a direct estimator is to take a weighted average of the two estimators (Ghosh & Rao, 1994). The estimator obtained obtained in such a way is known as composite estimator. An explicit model is useful in small area estimation, this gives the users an idea of how different information sources are combined. These methods permit formal model building process, including model selection and model diagnostics, and provide a good measure of uncertainty of the point estimator or predictor under a reasonable working model.

1.2 Linear Mixed Models in Small Area Estimation

Linear mixed models are often used in small area estimation because of their flexibility in combining information from different sources and taking different sources

of errors into account. These models may be classified into two broad classes, area level and unit level models, based on the availability of data for the variable to be modeled.

1.2.1 Area Level Model

Area level models are used to model survey-weighted direct estimates of the true small area parameters. For this class of models, relevant auxiliary information can be used at the small area level. An area level model borrows strength from relevant sources and also captures the differences in the small areas not only through the available auxiliary variables, but also through area specific random effects. In contrast, an implicit regression model motivating a synthetic estimation method assumes no between area variations other than those explained by the area-specific auxiliary variables (Jiang & Lahiri, 2006b).

Area level models are particularly helpful when it is difficult to obtain detailed data for the sampled units for various reasons, including issues related to the confidentiality of the data at the sampled unit level. Since the survey estimates are directly modeled, area level models usually result in design-consistent small area estimators. Also, the use of area level model avoids the need to deal explicitly with the survey design; this is accounted for in the sampling variance estimates. However, the estimation of the sampling variances of the survey-weighted estimates, which are needed in a typical area level model, is a challenging problem, due to the small sample sizes in the small areas.

To deal with this problem, researchers have considered appropriate models to improve the stability of the sampling variances. Fay & Herriot (1979), Otto & Bell (1995), Hinrichs (2003), Gershunskaya & Lahiri (2005), and Wolter (1985) discuss the problem of smoothing variances. The Generalized variance function (GVF) (Wolter, 1985, p. 201-217) method is found to be useful in stabilizing the sampling variances in the context of small area estimation. In GVF method, a mathematical model is used to describe the relationship between the variance of a survey estimate and its expectation. The choice of the function is based on the premise that the relative variance is a decreasing function of the magnitude of the expectation (Wolter, 1985, p. 203). The parameters of the model are estimated using past data. In practice, both the expectation and the variance of the survey estimate are unknown, we need to use the survey estimate and an estimate of the variance. An estimate of the design-based sampling variance of the direct survey estimate is obtained using one of the standard variance estimating techniques (linearization, jackknife, balanced repeated replication etc.). In the context of SAIPE project, Otto & Bell (1995) used GVF method for smoothing the sampling variances at the state level. In their model they also accounted for the differences in the variances by state (through random state effects), dependence of variances on sample size, sampling variance correlations over time (through an autoregressive-moving-average time series model). Once the relatively stable sampling variances are obtained, they are assumed to be known in a typical area level model. Any procedure which treats the sampling variances as known will not take into account the variability in estimating the sampling variances. To incorporate this additional variability,

researchers have considered alternate modeling. See Arora & Lahiri (1997), Bell & Otto (1992), Kleffe & Rao (1992).

In order to estimate the per-capita income of small places (population less than 1000), Fay & Herriot (1979) used an area level model to combine survey data with relevant administrative and census records. The Fay-Herriot model and its different extensions have been found to be effective in many small area applications. Particular cases of the Fay-Herriot model, some even before their paper, can be found in the literature in the context of a wide variety of applications. This includes the estimation of false alarm probabilities in the New York city (Carter & Rolph, 1974), batting averages of major league baseball players (Efron & Morris, 1975), ranking of the 23 kidney transplant hospitals (Morris & Christiansen, 1996), estimation of the poverty ratios for U.S. states and counties (Citro & Kalton, 2000) in the SAIPE program.

1.2.2 Unit Level Model

A unit level model can be used to model study variables available at the sampled unit level. This class of models has the potential to use auxiliary information at both the unit and area level (Moura & Holt, 1999). In unit level mixed models, area-specific random effect terms capture the correlation possibly present among the sample units within a small area. One advantage of area level modeling, discussed in the previous section, is that it usually leads to a design-consistent small area estimator. But if the sampling weights are ignored in unit level modeling, this leads to an

estimator that is not design-consistent (unless the sampling design is self-weighting within the small area). Survey practitioners prefer to use design consistent model based estimators because such estimators provide protection against model failures as the small area sample size increases (Rao, 2003, p. 148).

A unit level model may be a good choice when all the relevant unit level data, including all the design information such as stratification and clustering, are available for the sampled units. To produce design-consistent small area estimators using a unit level model, it is necessary to incorporate all the design information while building the model. However, this modeling is more challenging than an area level model. The absence of relevant design information poses a problem in unit level modeling. Failure to incorporate some design variables which govern unequal selection probability of units may result in an estimator that is not design-consistent. For example, in the context of the SAIPE project, to obtain model based estimates of poverty ratios for counties using CPS data, if one wants to consider a unit level model, the prior decennial census population count for the counties must be included as a covariate or the resulting estimator may not be design-consistent. The CPS primary sampling units (counties or a group of counties) are drawn with probability proportional to prior census population count in the CPS design.

In the absence of detailed design information, one needs to use the survey weights following the approaches outlined in Kott (1989), Prasad & Rao (1999), You & Rao (2002), Rao (2003), Jiang & Lahiri (2006a), under a frequentist paradigm. The basic idea is to obtain a survey-weighted aggregated area level model from the unit level model by taking a weighted average with weights as normalized survey

weights. Then the model based small area estimators obtained from the aggregated model satisfy the design consistency property. These estimators also satisfy the benchmarking property without any adjustment in the sense that they add up to the direct survey regression estimator when aggregated over areas (You & Rao, 2002). This approach, perhaps, is possible to implement only for a linear unit level model. For other types of unit level small area models (e.g., binomial with logistic regression) this is essentially an unsolved problem– there is no generally applicable accepted solution. Some researchers have addressed this issue for specific applications. Jiang & Lahiri (2006a) proposed a general model assisted approach encompassing continuous and binary reposnse variable. Their model assisted estimator converges in probability to the customary design-consistent estimator as the domain and sample sizes increase. For use of survey weights in a hierarchical Bayesian set up, the readers are encouraged to see You & Rao (2003), Rao (2003), and Lahiri & Mukherjee (2007).

A good application of unit level model in small area estimation can be found in the paper by Battese et al. (1988). They used a nested error regression model to estimate mean area under different crops (corn and soybeans) for twelve counties (small areas) of north central Iowa. To consider the correlation structures, where reported crop hectares for geographically closer segments have stronger correlation than those farther apart, they included a random county effect in the model. The nested error regression model, considered by Battese et al. (1988) and Prasad & Rao (1990), is also termed as random intercept model. This can be viewed as a particular case of multilevel models which allow small area slopes as well as the intercept to

be random and lead to improved small area estimates with the potential to use area level covariates (Moura & Holt, 1999). Multilevel models considered by Moura & Holt (1999) are also known as random regression coefficient models (Dempster et al., 1981) in the literature.

Besides small area estimation, nested error regression models and their various extensions have been found to be very useful in other fields as well, including longitudinal studies and animal breeding. For example, to study the effect of a drug, Propranolol, on hypertension, blood pressure measurements were taken on several persons after administration of the drug and a placebo both in the upright and reclining positions (McCulloch, 2003). To predict the mean blood pressure at the person level, we need to consider a unit level model on the blood pressure measurement of a particular person, at a particular position, and with a particular drug condition. In the model we need to treat the person effect as random, as contrasted with treating persons as fixed, like the effects for position or drug, to capture the correlation between measurements taken on the same person.

1.3 Generalized Linear Mixed Models in Small Area Estimation

The linear mixed models discussed in Section 1.2 are designed for continuous dependent variables, but they are not suitable for handling binary or count data. Generalized Linear Models (GLMs) are an extension of linear models that allow the usual regression methodology to apply to discrete data such as counts, binary responses, survival times which are very frequently observed in real life. A nice

account of GLMs can be found in McCullagh & Nelder (1989). But inferences from GLMs are based on the assumption that the responses are independent. In many real life problems, the observations are correlated. For example, in longitudinal studies, measurements obtained from the same individual over different times are likely to be dependent. In multi-stage cluster sampling, the responses of the individuals within the same cluster are possibly correlated. To analyze such data, we apply another extension of linear models, known as mixed models or variance components models (Searle et al., 1992), discussed in Section 1.2. However, these mixed models may not be appropriate for discrete data. This leads to the development of Generalized Linear Mixed Models (GLMMs)(McCulloch, 2003; Jiang & Lahiri, 2006b). To estimate the number of poor school-age children at the county and state level, the SAIPE project of the U.S. Census Bureau uses an area level normal linear mixed model, though the unit level data is binary. To estimate the unemployment rate at the state level, Datta et al. (1999) proposed a time series generalization of the Fay-Herriot model, though the corresponding response variable (the employment status of an individual obtained from CPS) is binary. These modeling assumptions may not be appropriate in all situation. For small sample size within the small area, the validity of the assumption is questionable.

The idea behind GLMMs is conceptually simple: incorporate random effects into the linear predictor portion of a GLM. This simple change allows us to accommodate correlation in the context of a broad class of models for non-normally distributed data. In other words, it is a convenient way to build the multivariate distributions for non-normal data that can accommodate dependence among the

observations. Most survey data are binary or categorical in nature, hence the problem of estimating rates and proportions for small areas using GLMMs had received considerable attention in the recent past. For some particular uses of GLMMs in small area estimation, see Chapter 4.

1.4 Inferential Procedures for Small Area Problems

Once relevant sources of information are identified for a particular small area of interest, a model needs to be postulated. The inferential procedure can be Bayesian or frequentist (classical). In a classical approach, the Empirical Best Linear Unbiased Predictors (EBLUPs) are used to estimate the true small area quantities (see Jiang & Lahiri (2006b), and the references therein). Whether we are in Bayesian or frequentist paradigm depends on whether we assume a prior on the hyperparameters (see Section 1.4.1). That is why the empirical Bayes approach is also considered as a frequentist approach by many researchers (Jiang & Lahiri 2006b, p. 4-5, Rao 2003, p. 179, Datta & Ghosh 1991, p. 1748), since it does not consider any prior distribution for the hyperparameters. Instead, the hyperparameters are estimated using some classical method. As an inferential approach, empirical Bayes (EB) and hierarchical Bayes methods have wide applicability in small area estimation in the sense of handling models for binary and count data as well as normal linear mixed models (Rao, 2003). In the latter case, EB and EBLUP are identical as discussed in the following section.

1.4.1 Empirical Bayes (EB) Approach

It is convenient to explain the empirical Bayes or EBLUP approach in the context of a basic area level model (such as the Fay-Herriot model). Let y_i be the survey-weighted direct estimate of some true small area quantity (such as a mean) θ_i for the ith small area, $i = 1, \ldots, m$. The Fay-Herriot model, widely used in the small area estimation literature, consists of two levels. In level 1, a sampling model, $y_i|\theta_i \overset{ind}{\sim} N(\theta_i, D_i)$, is used to capture the sampling variability of the regular survey estimate y_i. In Level 2, a linking model, $\theta_i|\beta, A \overset{ind}{\sim} N(x_i'\beta, A)$, relates the true small area quantity θ_i to a $p \times 1$ vector of known covariates x_i. In this model, β is a $p \times 1$ vector of unknown regression coefficients and A is an unknown variance component. The sampling variances, D_i's, are assumed to be known, though in practice they are estimated by some suitable method, see Section 1.2.1. To estimate the number of poor school-age (5-17) children for the U.S. states, up through 1995, the Census Bureau employed an empirical Bayes methodology using the Fay-Herriot model that combines the direct survey (CPS) estimates of poverty ratio with the auxiliary information obtained from Internal Revenue Service individual income tax returns, food stamp administrative records, population estimates from the Census Bureau's demographic estimates program, and the previous census (Citro & Kalton, 2000).

In small area estimation, the main objective is usually to draw inferences about the high-dimensional parameters θ_i. However, as an intermediate step, estimation of the low-dimensional parameters β and A, usually referred to as hyperparame-

ters, is also important. Empirical Bayes methodology assumes the hyperparameters involved at level 2 to be known. When hyperparameters are known, the Bayes estimator of θ_i is in the form of a shrinkage estimator, $(1 - B_i)y_i + B_i x_i'\beta$, where $B_i = D_i/(A + D_i)$, is the shrinkage factor which shrinks the direct estimates to the regression synthetic estimates. In practice, β and A are unknown and estimated from the data.

When A is known but β is unknown, use of weighted least square estimate $\hat{\beta}_A = (X'W_A X)^{-1}(X'W_A y)$, of β in the Bayes estimator of θ_i is a standard practice, where $W_A = diag\,(1/(A + D_i))$, X is the $m \times p$ matrix of covariates, and y is the $m \times 1$ vector of small area direct estimates. Note that $\hat{\beta}_A$ is also the maximum likelihood estimator of β, under the Fay-Herriot model. When A is known, the empirical Bayes estimator of the ith small area mean θ_i and its measure of uncertainty are identical to the best linear unbiased predictor (BLUP) of θ_i ($\theta_i = x_i'\beta + v_i$, where v_i is the area level random effect term) and its mean squared error, respectively, under the linear mixed model $y_i = x_i'\beta + v_i + e_i$, where $\{v_i\}$ and $\{e_i\}$ are independently distributed with $v_i \sim$ iid $N(0, A)$ and $e_i \sim$ ind $N(0, D_i)$. Note that the weighted least square estimate of β involves the variance component A which is also unknown, in practice. The unknown variance component needs to be estimated using some suitable method (see Section 1.4.2). Then the estimate of A is plugged in the Bayes estimator given above along with $\hat{\beta}_A$ and we obtain the empirical Bayes estimator (equivalently, empirical best linear unbiased predictor) of the small area mean θ_i.

It is customary to judge any estimator by its corresponding measure of uncertainty. Estimation of the mean squared error (MSE) of the empirical Bayes (or

EBLUP) estimator of θ_i is quite complicated. It requires rigorous asymptotics to find a closed form expression of the measure of uncertainty under certain regularity conditions. The measure of uncertainty associated with an empirical Bayes estimator of the true small area mean can be decomposed into three parts. The first term measures the uncertainty in the model for estimating θ_i, and the second term measures the uncertainty in the estimation of β. Generally, there is no closed form expression available for the third term, which measures the uncertainty in estimating the variance component A. The third term depends on the method used in estimating the variance component, while the first two terms remain the same for any method used to estimate the variance component. Asymptotic expressions of the third term can be found in literature. See Prasad & Rao (1990) when the ANOVA method is used to estimate the variance component; Datta & Lahiri (2000) for the ML and REML estimators of A; Smith (2001) and Datta et al. (2005) for the Fay-Herriot method-of-moment estimator of A. Also for a detailed derivation of the MSE estimator (mse), readers are encouraged to see the above mentioned papers.

A naïve approach of obtaining the measure of uncertainty of the empirical Bayes estimator (equivalently the MSE of BLUP when A is known) is to plug-in the variance estimator of A in the posterior variance of θ_i when A is known. When A is known, the MSE of the BLUP of θ_i is given by $D_i\,(1 - B_i) + B_i^2 x_i'\,\Sigma_A\,x_i$, where $\Sigma_A = (X'W_AX)^{-1}$. In this expression, an estimate of the variance component is plugged in to obtain the measure of uncertainty of EBLUP of θ_i. This naïve approach does not take into account the uncertainty associated with the estimation of the variance component and only considers the first two terms as mentioned above.

The corresponding 95% interval estimation is based on the standard $EB \pm 1.96\sqrt{mse}$ type confidence interval. The simplicity of this method is quite appealing. However, the current research in small area estimation (see Hall & Maiti (2006), Chatterjee et al. (2008), and Li (2007)) suggests that this kind of confidence interval typically suffers from an undercoverage problem. When n_i, the number of sampled units in the ith small area, is small and m is large, the first two terms will dominate and one can ignore the third term. However, when m is moderate in size or n_i is not small, the third term can be substantial and should not be ignored. Also when the synthetic estimator is far apart from the direct estimator, the contribution of the measure of uncertainty from the third term will be large. See Kass and Steffey (1989) and Bell (1999) for further discussion.

1.4.2 Variance Component Estimation

In an empirical best linear unbiased prediction (EBLUP) or empirical Bayes approach, no prior distribution is assumed on the variance component A. A is estimated using some frequentist method. Three methods of variance component estimation are widely used in small area estimation. These are the analysis of variance (ANOVA; Prasad & Rao (1990)), the method of moments (Fay & Herriot (1979); Pfeffermann & Nathan (1981); Datta et al. (2005)) and the likelihood-based methods. All of these methods can produce unreasonable estimate of A, i.e., they can produce a negative or zero estimate of A. The likelihood-based method essentially maximizes certain likelihood function of A. The profile likelihood and residual

likelihood are well known in the literature (see Rao (2003); Jiang (2007)).

The maximum likelihood (ML) method, which maximizes the profile likelihood function, was introduced to variance component estimation by Hartley & Rao (1967). Under some regularity conditions, ML estimators have some good large sample properties. For example, they are consistent, efficient, and normally distributed. Another attractive feature of ML estimation is that the asymptotic dispersion matrix of the estimators is always available, except perhaps when ML estimate occurs at the boundary point. It is the inverse of the information matrix. Note that the weighted least square estimate ($\hat{\beta}_A$) of β is also ML if we plug-in the ML estimate of A.

There are certain situations met in practice where the ML estimator is inconsistent. Neyman & Scott (1948) showed that for a partially consistent series of observations (units in different small areas have different means but the same variance), the ML estimator of variance component has got a downward bias which is usually the case as the maximum likelihood method does not take into account the effect of estimating the fixed effects. Moreover, it is not a consistent estimator. In another example, again for a partially consistent series of observations (this time with same mean but different variances) Neyman and Scott showed that the ML estimator of the mean (μ), although consistent, does not have the property of asymptotic efficiency. That means it is possible to find an estimator of μ other than the ML estimator whose mean squared error is less than that of the maximum-likelihood estimator. Also for the p-variate normal distribution, the maximum-likelihood estimator of the mean vector is unbiased, consistent, normally distributed but inadmis-

sible under squared error loss for $p \geq 3$. An estimator (δ) of a parameter is said to be inadmissible if the average loss incurred by δ is greater than or equal to that of another estimator (δ'), for all values of the parameter space and greater than that of δ' for at least one parameter value. James-Stein's (1961) shrinkage estimator, although biased, has smaller average squared error loss.

REML maximizes the residual likelihood function. In classical inference, the residual maximum likelihood method is preferred to maximum likelihood as a variance component estimation method. REML is intended to reduce the downward bias of ML. REML separates the part of the data used for the estimation of variance components from that used for the estimation of fixed effects to eliminate the fixed effects from the likelihood. This is a deficiency of ML estimators of variance components, which take no account of the loss of degrees of freedom resulting from the estimation of the model's fixed effects. REML determines a linear transformation, $z = A'y$, of data y that is free of fixed effects. This can be done by considering the error contrasts. REML estimators of variance components maximize the likelihood based on the transformed data rather than the original data. For further details see Patterson & Thompson (1971) and Harville (1974). Under standard regularity conditions, the residual maximum likelihood estimator is consistent estimator of A, for large m (Jiang, 1996).

But in many practical applications, both maximum likelihood (ML) and residual maximum likelihood (REML) estimates of variance components occur at the boundary point. For example, in a two-level Poisson-gamma model, the ML estimate of the variance component can be infinity (Christiansen & Morris, 1997);

for the Fay-Herriot model, the ML or REML estimate of variance component can be zero (Bell, 1999). When that happens, we come up with several unreasonable implications on the estimators and its measure of uncertainty. For example, in the context of the Fay-Herriot model, when $\hat{A} = 0$, the empirical Bayes estimate of θ_i gives zero weight to the direct estimate, which is unreasonable for large areas. To overcome this problem i.e., to change the curvature of the likelihood function in order to have an estimate falling within its admissible range, an adjustment term is multiplied with the likelihood function (Christiansen & Morris (1997), Tang (2002), Morris (2006)) and then the adjusted likelihood function is maximized to obtain an estimate of the variance component. This method is termed as adjustment for density maximization (ADM) by Morris and his collaborators. Morris (2006, p. 72-76) suggests an adjustment term A that needs to be multiplied with the residual likelihood function. Recently, Li & Lahiri (2008) demonstrated the superiority of the ADM method that maximizes the adjusted profile likelihood over the one that maximizes the adjusted residual likelihood. They also proved analytically that the ADM estimators of A are strictly positive and consistent under the same regularity conditions used for the asymptotic properties of REML.

Bell (1999), in the context of SAIPE program, considered a relatively less familiar mean likelihood (MEL) approach to find an estimate of the model variance A, while producing state level estimates of poverty ratios among school-age children. The MEL estimate is the posterior mean of A obtained from the marginal posterior density of A, assuming uniform prior on A. Naturally, the MEL always produces positive estimates of A. In this paper he compared the empirical Bayes approach

with the plug-in MEL estimate of A to the fully Bayesian approach.

1.4.3 Hierarchical Bayesian Approach

The hierarchical Bayesian approach to inference considers the hyperparameters to be unknown all through the inferential procedure and requires some prior distribution on the hyperparameters. The beauty of this approach is its ability to structure complicated models, inferential goals, and analysis. The prior and likelihood produce the full joint posterior distribution, which is used for all inferential purposes. In that sense, this approach is very straightforward. It overcomes some of the shortcomings of the empirical Bayes approach. The posterior variance of θ_i, which measures the precision of the estimator of θ_i, automatically takes into account all sources of uncertainty. But, unlike the empirical Bayes approach, the introduction of a third level of prior specification often leads to a nonstandard posterior; one needs to apply the computer-intensive Markov Chain Monte Carlo (MCMC) technique to estimate the parameters, even with uniform prior. In almost all situations, closed form expressions for the posterior means and variances cannot be obtained and hence it becomes difficult to interpret the formulae. This makes the hierarchical Bayesian methodology less appealing to users. For specific implementation of the hierarchical Bayesian approach for small area estimation see the introduction sections of the Chapters 2, 3, and 4.

The prior distribution plays a major role in Bayesian analysis. Any subjective prior information, if available, should be used in Bayesian analysis. However, such

a prior distribution may not be available in many applications. In the absence of a subjective prior information, statisticians often use various noninformative priors that have been proposed in the literature to carry out a noninformative Bayesian analysis. Noninformative priors are usually improper. In a hierarchical Bayesian approach, it is important to check for the propriety of the posterior distributions involved, in case improper priors are used for the hyperparameters. Improper priors may lead to a posterior distribution (the basis of inference) which is not a proper density. The popular Bayesian software BUGS (Spiegelhalter et al., 1997), based on Gibbs sampling (Gelfand & Smith, 1990) or, more generally, the MCMC technique, cannot inform the users that the posterior is improper. Gibbs conditionals corresponding to an improper posterior may appear perfectly reasonable. For an example of this phenomenon, see Hobert & Casella (1996). That is why one should demonstrate the propriety of posterior before a MCMC technique is used.

1.4.4 Approximations in Hierarchical Bayesian Approach

The Monte Carlo Markov Chain (MCMC) technique can be used to implement complex hierarchical Bayesian models. In general, it is not feasible to draw independent samples from the joint posterior distribution of the parameters η (includes the small area parameters and the hyperparameters), because of the intractable form of the posterior. MCMC avoids this difficulty by constructing a Markov chain (a time dependent sequence of events) of the parameters such that the distribution of the chain converges to a unique stationary distribution, under certain conditions, which

is equivalent to the posterior distribution of the parameters. Then, the posterior mean of η can be approximated by the average of the sequences of the Markov chain, after ignoring a sufficiently large burn-in. These properties follow from the ergodic theorem of stochastic process, which can be viewed as law of large numbers for a dependent sequence. For further details on MCMC methods see Rao (2003) and Robert & Casella (2004).

Although the MCMC method is justified by the ergodic theorem, in practice results from a MCMC run can depend heavily on several factors. This includes the choice of the initial values for the parameters, the burn-in length, the number of replicates after discarding the burn-in samples, the number of chains, the time series plot for the chains corresponding to each parameter (to see how well the chain mixes and whether the chain has converged), and the autocorrelation plot (ideally there should be low autocorrelation for samples further apart). All these factors are carefully examined in a Bayesian analysis. If the Bayesian methodology is to be carried out routinely by someone with minimal knowledge of sophisticated MCMC methods, then the convergence of the MCMC technique may not be checked properly, which may lead to unreasonable conclusions. Also the slow computation speed of MCMC does not permit its evaluation by repeated use in simulation. Approximation to the complex posterior distribution and posterior moments can be useful in such situation. The approximations are designed to present the Bayesian methodology in a transparent way, which facilitates the interpretation of the methodology to the data users. Simple approximation of a complex posterior distribution and its moments has been discussed by many researchers, including Tierney et al. (1989); Kass &

Steffey (1989); Morris (1988, 2006); Christiansen & Morris (1997); Tang (2002).

1.4.4.1 Adjusted density method

The adjusted density method (Morris, 1988) approximates a complex univariate density (possibly a posterior density) with a Pearson density. To do so, this method first multiplies the given univariate density by the appropriate Pearson quadratic function, after which fitting is done by matching the first two derivatives of the adjusted density to that of a two-parameter Pearson family. The choice of a Pearson density depends, in most cases, on the support of the given density. For example, Christiansen & Morris (1997) examined beta approximation to the posterior density of shrinkage factors in the context of hierarchical Poisson regression modeling. Before going into the detail of the methodology, we describe the Pearson density.

In general, the Pearson family, with respect to the quadratic function, $Q(x) = q_2x^2 + q_1x + q_0 > 0$, has density $p(x) = K_Q(m, \mu_0) \cdot \exp\left\{-m \int \frac{(x-\mu_0)dx}{Q(x)}\right\} \frac{1}{Q(x)}$. For fixed Q, the Pearson family can be viewed as a two-parameter distribution, denoted by

$$\text{Pearson}(m, \mu_0; Q) = \text{Pearson}\left[\text{mean } = \mu_0, \ \text{ variance } = Q(\mu_0)/(m - q_2)\right].$$

Examples of the most familiar Pearson families include the normal, gamma, inverse gamma, beta, and F distributions.

Suppose the density $f(x)$ of some random variable X is to be approximated by a Pearson(m,μ_0;Q) $= p(x)$ density specified by Q, perhaps chosen because its

range agrees with that of f. Define $l(x) = \log(f(x)Q(x))$. Then, with respect to the Pearson measure $\frac{dx}{Q(x)}$, $f(x)Q(x)$ is a density. Express $f(x)Q(x)$ in the form of $\exp\left\{-m\int\frac{(x-\mu_0)dx}{Q(x)}\right\}$ by matching two derivatives of the logarithms of the adjusted density and the Person density at the modal value. Letting $l'(x) = 0$, with x_0 be the root of this derivative, then $x_0 = \mu_0$ and $-l''(x)|_{x=x_0} = m/Q(x_0)$.

An example: Suppose we need to calculate the moments (or approximate the posterior density $p(B)$) of shrinkage factor, B, of the type $m/(m+n)$; $m > 0,\ n > 0$. We come across with this kind of shrinkage factor frequently in small area estimation. Since $0 < B < 1$, we can choose the beta distribution as the approximating distribution from the Pearson family. The beta is a rich family of distributions in the sense that it exhibits a fairly wide variety of shapes on the unit interval (0, 1). It encompasses right skewed, symmetric, negatively-skewed distribution. Moreover, the uniform (0, 1) is a special case of Beta distribution.

Define $l(B) = \log\{p(B).B(1-B)\}$, where $B(1-B)$ is the adjustment to $p(B)$ required for this particular example. Let $\hat{B}$ maximize $l(B)$, so that $l'\left(\hat{B}\right) = 0$ and define $-l''\left(\hat{B}\right) = i_0$, the Pearson information. If the approximating Pearsonian density is $Beta(a_1, a_2)$, then we can write it as

$$K\exp\left\{-(a_1+a_2)\int\frac{B - a_1/(a_1+a_2)}{B(1-B)}\,dB\right\}\frac{1}{B(1-B)}.$$

Comparing it with the Pearsonian form (mentioned above), we get $\mu_0 = \frac{a_1}{a_1+a_2}$, $m = a_1 + a_2$, and $Q(B) = B(1-B)$. Thus, we obtain $\hat{B} = \frac{a_1}{a_1+a_2}$ and $i_0 = \frac{a_1+a_2}{\hat{B}\left(1-\hat{B}\right)}$. Simple calculation leads to $a_1 = \hat{B}^2\left(1-\hat{B}\right)i_0$ and $a_2 = \hat{B}\left(1-\hat{B}\right)^2 i_0$, and so $B \sim Beta\left(\hat{B}^2(1-\hat{B})i_0, \hat{B}(1-\hat{B})^2 i_0\right)$ is the recommended approximation to the

density of B.

This method allows approximation of univariate densities by distributions other than the normal, and hence there is a possibility of more accurate approximation for small sample sizes. This approximation makes the calculation of the posterior moments easy as the mean and variance of the Pearson density are known beforehand. For further details, see Morris (1988). The advantage of this approximation is that an entire distribution is being fitted by a Pearson family, not just the moments. So any kind of inference can be drawn conveniently. For example, to find the quantiles of the complex distribution, standard t, chi-square, F tables or the readily available softwares can be used. This method is useful for noninformative Bayesian analysis–that is, when noninformative priors, say uniform, are used at the third level. When we have a subjective prior in hand, chosen by following some objective criteria or on the basis of some prior knowledge, application of this method changes the prior ultimately because of the additional adjustment term and restricts us to use the subjective prior in the analysis. For an illustration of this phenomenon using SAIPE data, see Chapter 2. Another shortcoming of this approximation is that it does not generalize in an obvious way to the multidimensional case, whereas Laplace's method can be used to approximate the posterior moments easily for the multidimensional case.

1.4.4.2 Laplace Approximation

Laplace's method is a technique of classical applied mathematics and very useful for asymptotic evaluation of integrals. This remarkable method provides accurate approximations to the posterior means and variances of any real function of parameter vector θ in Bayesian analysis. Posterior moments can be expressed as ratio of integrals and the application of Laplace's method to ratio of integrals leads to accurate approximation for the posterior moments. This method has been applied by many authors in the context of Bayesian analysis. See Tierney & Kadane (1986); Tierney et al. (1989); Kass & Steffey (1989); Butar & Lahiri (2002); Datta et al. (2005).

The basic method can be described as follows. Consider an integral

$$\int b(\theta) \exp(nL(\theta))\, d\theta.$$

Laplace's method provides an approximation of the above integral when n (usually the sample size in statistical applications, but here the number of small areas) is large. The idea is that if L has a unique maximum at $\hat{\theta}$, then for large n the value of the integral depends only on the behavior of the function L near its maximum (Tierney & Kadane, 1986). In other words, the approximation is based on Taylor series approximation to $L(.)$ and $b(.)$ about $\hat{\theta}$. For the convenience of exposition, we assume that the parameter θ is one-dimensional and $b(\theta) = 1$. Then the second

order approximation of the above integral is given by:

$$\begin{aligned}\int \exp\left[nL(\theta)\right]\, d\theta &\approx \exp\left[nL(\hat{\theta})\right] \int \exp\left[-\frac{n}{2\sigma^2}\left(\theta-\hat{\theta}\right)^2\right]\, d\theta \\ &= \sqrt{2\pi}.\frac{\sigma}{\sqrt{n}}.\exp\left[nL(\hat{\theta})\right],\end{aligned}$$

where $\sigma^2 = -1/L''(\hat{\theta})$. The last step follows by comparing the integrand with a normal density. Higher order approximations may be derived by retaining higher order terms in the expansion of $L()$ and $b()$. When $b(\theta)$ is not equal to 1, i.e. $b(\theta)$ is any general function of θ, which is usually the case, then we need to calculate higher order derivatives of $L()$ and $b()$. But when we consider the ratio of integrals, the third derivatives of L disappear under certain form of Laplace approximation. See Kass et al. (1988) for several forms of Laplace approximation to ratio of integrals. In the context of the Fay-Herriot model, if we consider $\theta = A$, the variance component, then we can take $L()$ and $b()$ as log-posterior density of A and shrinkage factor respectively in order to approximate the posterior moments of the shrinkage factor.

1.5 Discussion and Outline of the Dissertation

In this chapter, we have given a broad overview of small area estimation, its usefulness and application in a wide variety of settings. We discussed model based approaches in drawing inferences for small areas. We discussed several methods for the estimation of the variance component, which plays an important role in obtaining reliable small area estimates and the associated measure of uncertainty. In a hierarchical Bayesian set up, for the ease of implementation and evaluation of the hierarchical Bayesian procedure, we discuss two approximate methods. However, in

this chapter no attempt is made to identify the research gap and how this dissertation is going to fill that. That is done in the next three chapters separately. See the introduction and outline section of each of the following chapters to find out the novel contribution of this dissertation to address some of the research gaps in small area estimation.

In Chapters 2 and 3, we apply linear mixed normal models (area level and unit level) to draw inferences for small areas when the variable of interest is continuous. We propose a new prior distribution for the variance component. We also use Laplace approximation to obtain accurate approximations to the posterior moments of interest. The approximations present the Bayesian methodology in a transparent way, that makes it easier for data users to interpret the methodology.

The linear mixed models used in Chapter 2 and 3 are not suitable for handling binary or count data. In Chapter 4, we consider hierarchical Bayes estimation of small area proportions. The binomial-beta hierarchical model we use is different from the usual mixed logistic model suitable for the proportion estimation problem. Our formulation allows a regression specification and hence extends the usual exchangeable assumption at the second level. Also, we carefully choose a prior for the shape parameter of the beta density. This new prior helps to avoid the extreme skewness present in the posterior distribution of the model parameters so that the Laplace approximation performs well.

In Chapters 2, 3, and 4 we carry out some empirical applications to demonstrate the utility and accuracy of our approach in solving real-life small-area estimation problem relative to existing methods.

Chapter 2

On the Prior Selection and Approximations in the Fay-Herriot Model

2.1 Introduction

In order to estimate the per-capita income of small places (population less than 1000), Fay & Herriot (1979) used an area level model to combine survey data with relevant administrative and census records. The Fay-Herriot model, extensively used in the small area estimation literature, consists of two levels. In Level 1, a sampling model captures the sampling variability of the regular survey estimates y_i of true small area means θ_i,

$$y_i|\theta_i \overset{ind}{\sim} N(\theta_i, D_i),\ i = 1, \ldots, m. \tag{2.1}$$

In Level 2, a linking model relates the true small area means θ_i to a $p \times 1$ vector of known covariates x_i,

$$\theta_i|\beta, A \overset{ind}{\sim} N(x_i'\beta, A),\ i = 1, \ldots, m. \tag{2.2}$$

In the above model, β is a $p \times 1$ vector of unknown regression coefficients and A is an unknown variance component. The sampling variances, D_i's are assumed to be known, though in practice they are estimated by some suitable method, see Chapter 1 (Section 1.2.1), for details about the estimation of D_i.

In small area estimation, usually the main objective is to draw inferences about the high-dimensional parameters, i.e. θ_i. However, as an intermediate step,

estimation of the low-dimensional parameters β and A, usually referred to as hyper-parameters, is also of importance. In the hierarchical Bayes implementation of the Fay-Herriot model, a prior distribution, often a vague or noninformative prior, is assumed on the hyperparameters. For example, Morris & Christiansen (1996) used the following prior distribution for the problem of ranking and identifying the best or worst of several individuals:

$$p(\beta, A) \propto 1;\ (\beta, A) \in R^p \times [0, \infty] \tag{2.3}$$

The prior distribution (2.3) for the hyperparameters is simple to interpret and is often recommended. The uniform prior for A is noninformative and yields a posterior distribution of A for which the mode is identical to the residual maximum likelihood (REML) estimator of A (Harville 1977; Berger 1985, p. 192). Thus, the posterior mode of A, under prior (2.3), enjoys good frequentist properties, which follows from the general theory on REML; see Jiang (1996).

In spite of good asymptotic properties of the posterior mode, under prior (2.3), it could produce an undesirable zero estimate of A for a given data set. For example, in estimating annual poverty ratios of school-age (5-17) children for the U.S. states using CPS data in the SAIPE program, the REML estimate of A is zero for the years 1989-1992 (Bell, 1999) and 1997 (see Section 2.5). When $\hat{A} = 0$, we come up with several unreasonable implications on the estimators and its measure of uncertainty (see Bell, 1999, for details). In discussing Jiang & Lahiri (2006b), Morris noted that the likelihood function of A is invariably right-skewed so that its mode will be smaller than most of its distribution, as we know for a right-skewed distribution

mean > median > mode. Hence, the maximum likelihood (ML) estimate of A is biased toward zero – substantially so if A is small. The same comments apply to REML and hence to the posterior mode of A, under prior (2.3). Morris (2006) proposed to multiply the residual likelihood by an adjustment term (A) to change the curvature of the likelihood function and then to maximize the adjusted likelihood so that the resulting estimator of A approximates the mean of the distribution, not the mode. This method is termed as adjustment for density maximization (ADM) by Morris and his collaborators (Christiansen & Morris, 1997; Morris, 2006; Tang, 2002).

Recently, Li & Lahiri (2008) proved that Morris' ADM estimator of A is strictly positive and is consistent, for large m, under standard regularity conditions. They also noted that frequentist properties of estimators of both A and the shrinkage factor B_i [see (2.4)] improve when the residual likelihood function of A is replaced by the profile likelihood function. A natural question, not addressed till date, is: For which prior on A, are the posterior mode and the ADM estimator of Morris 2006 (or Li & Lahiri, 2008) identical? By addressing this question in this chapter, we specify the full hierarchical Bayes model, which is useful in solving wide range of problems.

The Monte Carlo Markov Chain (MCMC) technique can be used to implement the proposed hierarchical Bayesian model. Although the MCMC method is justified by the ergodic theorem, in practice results from a MCMC run can depend heavily on several factors (see Chapter 1, Section 1.4.4). All these factors are carefully examined in a Bayesian analysis. If the Bayesian methodology is to

be carried out routinely by someone with minimal knowledge of the sophisticated MCMC method, then the convergence of the MCMC technique may not be checked properly, which may lead to unreasonable conclusions. Also its slow computation speed does not permit its evaluation by repeated use in simulation. That is why we intend to implement our proposed hierarchical Bayes procedure, which involves two new priors on A, by simple approximations, motivated from the application of Laplace's method to ratio of integrals. Simple approximation to a complex posterior distribution and its moments has been discussed by many researchers, including Tierney et al. (1989); Kass & Steffey (1989); Morris (1988, 2006); Christiansen & Morris (1997); Tang (2002). The approximations offer simple interpretations of the Bayesian methodology.

In Section 2.2, we review the empirical and hierarchical Bayes procedure for the Fay-Herriot model and discuss the ADM method in more detail. In this section we also introduce two new prior distributions on A and provide conditions for the propriety of the resulting posterior distributions. In Section 2.3, we apply the Laplace method to approximate the posterior moments of the parameters involved, using the new priors. We present results from a Monte Carlo simulation study in Section 2.4 and establish the superiority of the approximate hierarchical Bayes method resulting from one of our proposed priors relative to some other existing methods. Here we study the frequentist bias and variance of different estimators of A and the shrinkage factors B_i. In addition, we also examine the frequentist mean squared error and coverage properties of the resulting hierarchical Bayes estimators of θ_i. In Section 2.5, we apply our hierarchical Bayes procedure on SAIPE data and

present the results in detail. With the help of this data analysis we demonstrate the utility and accuracy of our methods relative to some other existing methods. A brief summary of the chapter is given in Section 2.6, along with potential future research ideas. We present some technical details, computer programs and convergence criteria of MCMC in the Appendices.

2.2 Selection of Prior for the Hyperparameters

When the hyperparameters β and A are known, the posterior distribution of θ_i is normal with mean and variance given by:

$$E(\theta_i|y, A, \beta) = (1 - B_i)y_i + B_i x_i'\beta \tag{2.4}$$

$$V(\theta_i|y, A, \beta) = D_i(1 - B_i), \tag{2.5}$$

where $B_i = D_i/(A + D_i)$, is the shrinkage factor which shrinks the direct estimates to a regression surface. Note that the right hand side of (2.4) is essentially the best predictor (BP) of θ_i, being the conditional mean of θ_i, given data, assuming known hyperparameters. Under the hierarchical Bayesian approach, to estimate θ_i along with a reliable measure of precision, we need to obtain $E(\theta_i|y)$ and $V(\theta_i|y)$. To this end, we first find the conditional posterior distribution of β, given A, and then the posterior distribution of A.

When A is known, the uniform prior on β in R^p, the p-dimensional real space, yields the following posterior distributions for β and θ_i:

$$\beta|y, A \sim N\left[\hat{\beta}_A, \Sigma_A\right], \tag{2.6}$$

$$\theta_i|y, A \sim N\left[(1-B_i)y_i + B_i x_i'\hat{\beta}_A,\ D_i(1-B_i) + B_i^2 x_i'\Sigma_A x_i\right], \tag{2.7}$$

where $\hat{\beta}_A = (X'W_AX)^{-1}(X'W_Ay)$, $\Sigma_A = (X'W_AX)^{-1}$ with $W_A = diag\,(1/(A+D_i))$, X is the $m \times p$ matrix of area-level covariates, and y is the $m \times 1$ vector of small area direct estimates. The subscript A in both $\hat{\beta}_A$ and Σ_A indicates the dependence of the terms on A. Note that the posterior mean and the posterior variance of β are identical to the maximum likelihood estimator of β and its standard variance estimator, respectively, under the marginal distribution of y. Also, the posterior mean, $E(\theta_i|y, A)$, and the posterior variance, $V(\theta_i|y, A)$, of θ_i [see (2.7)] are identical to the best linear unbiased predictor (BLUP) of $\theta_i = x_i'\beta + v_i$ and its mean squared error, respectively, under the linear mixed model $y_i = \theta_i + e_i = x_i'\beta + v_i + e_i$, where $\{v_i\}$ and $\{e_i\}$ are independently distributed with $v_i \sim$ iid $N(0, A)$ and $e_i \sim$ ind $N(0, D_i)$; $i = 1, \ldots, m$. Note that the BLUP does not require any specification of prior for β; however, the commonly used uniform prior for β on R^p yields frequentist inference for both β and θ_i.

In practice, the variance component A is unknown. In an empirical best linear unbiased prediction (EBLUP) or empirical Bayes approach, no prior distribution is assumed on the variance component A, which is estimated using some suitable classical method. Three most common methods of variance component estimation are widely used in the small area estimation. These are analysis of variance (ANOVA; Prasad & Rao (1990)), the method of moments (Fay & Herriot (1979); Pfeffermann & Nathan (1981); Datta et al. (2005)) and likelihood-based method. All of these methods can produce unreasonable estimate of A i.e. they can produce negative

or zero estimate of A. Bell (1999), in the context of SAIPE program, considered a relatively less familiar mean likelihood (MEL) approach to find an estimate of the model variance A, while producing state level estimates of poverty ratios among the school-age children. The MEL estimate is the posterior mean of A obtained from the marginal posterior density of A assuming uniform prior on A. Naturally, MEL always produces positive estimates of A. However, the frequentist properties such as the bias and variance of the mean likelihood estimator are not known. To avoid the boundary estimation problem in a non-Bayesian framework, Lahiri (2003) also suggested to consider alternate measures of central tendency (e.g., median, mean, etc.) of the probability distribution, instead of considering mode, obtained by standardizing the marginal likelihood function.

A likelihood-based method essentially maximizes a likelihood function of A. The profile likelihood (maximum likelihood) and residual likelihood of A are well-known in the literature (see Rao 2003, Jiang 2007). Certain adjustments to the residual and profile likelihood could significantly improve the estimation of both A and B_i. We define the following general likelihood function of A as

$$L(A) = \pi(A) L_R(A), \tag{2.8}$$

where

$$L_R(A) \propto |W_A|^{1/2} \, |X' W_A X|^{-1/2} \exp\left[-\frac{1}{2} (y - X\hat{\beta}_A)' W_A (y - X\hat{\beta}_A) \right] \tag{2.9}$$

is the residual likelihood function of A. The factor $\pi(A)$ can be viewed as an adjustment factor to the residual likelihood, which can produce different likelihood

functions considered in the literature. The following choices of $\pi(A)$ deserve special mention:

(i) $\pi(A) = 1$ or no adjustment yields the residual likelihood;

(ii) $\pi(A) = |X'W_AX|^{1/2}$ yields the profile likelihood;

(iii) $\pi(A) = A$ yields the adjusted residual likelihood proposed by Tang (2002) and Morris (2006);

(iv) $\pi(A) = A\,|X'W_AX|^{1/2}$ yields the adjusted profile likelihood considered by Li & Lahiri (2008).

Under standard regularity conditions, the maximum likelihood estimator and residual maximum likelihood estimator, obtained by maximizing (i) and (ii), are consistent estimators of A, for large m (Jiang, 2007). Recently, Li & Lahiri (2008) proved that the maximizations of (iii) and (iv) yield strictly positive estimators of A, which are also consistent under the same regularity conditions, for large m. The estimators of A, obtained by maximization of (iii) and (iv), are known as the ADM estimators. Li & Lahiri (2008) demonstrated the superiority of the ADM that maximizes (iv) over the one that maximizes (iii).

From a Bayesian perspective, the marginal posterior distribution $f(A|y)$ of A, under the prior $p(\beta, A) \propto \pi(A)$, is proportional to $\pi(A)L_R(A)$ - thus the posterior mode of A is identical to the frequentist estimator that maximizes the general likelihood $L(A)$ given by (2.8) and (2.9). We can match the posterior mode of A to the ML, REML, and ADM estimators of A by choosing the prior $\pi(A)$ suitably.

The priors corresponding to REML and ML are denoted by $\pi_{RL}(A)$ and $\pi_{PL}(A)$ respectively. However, because of the superiority of ADM estimators over the ML and REML we investigate only the choices (iii) and (iv) above. We denote the prior corresponding to the choice (iii) by $\pi_M(A)$ i.e. $\pi_M(A) \propto A$. The subscript M is used to acknowledge that the prior is derived from the adjustment term suggested by Morris (2006), though he did not propose the prior. The prior $\pi_M(A)$ leads to a proper posterior if $m > p + 4$ (see Appendix A of this chapter for a proof).

In this chapter, we propose a simpler prior on A starting from the choice (iv). To this end, note that for the balanced case i.e. when the sampling variances are equal ($D_i = D\ \forall i$), we have $\pi(A) \propto \frac{A}{(A+D)^{p/2}},\ A \in [0, \infty]$. Our choice of prior for the unequal variance case is heuristically motivated by the equal sampling variance case. If we replace $D_i,\ i = 1, \ldots, m$ by a particular representative value d_0 (say, the median of D_i) in W_A, then $\pi(A)$ becomes $\frac{|X'X|^{1/2}}{(A+d_0)^{p/2}}$. Leaving the proportionality constant aside, we obtain

$$\pi_{LL}(A) \propto \frac{A}{(A + d_0)^{p/2}},\ A \in [0, \infty] \tag{2.10}$$

Note that the prior proposed in (2.10) does not depend on the individual sampling variance, D_i, unlike Datta et al. (2005). The posterior $f(A|y)$ under the prior (2.10) is proper, provided $m > 4$ (see Appendix A for a proof). We use the notation $\pi_{LL}(A)$ for this prior, since it is motivated from the adjustment term given by Li & Lahiri (2008). Note that, for $p = 0$, i.e. for common mean model with known mean, $\pi_{LL}(A)$ is equivalent to $\pi_M(A)$ and consequently, the conditions for propriety are the same.

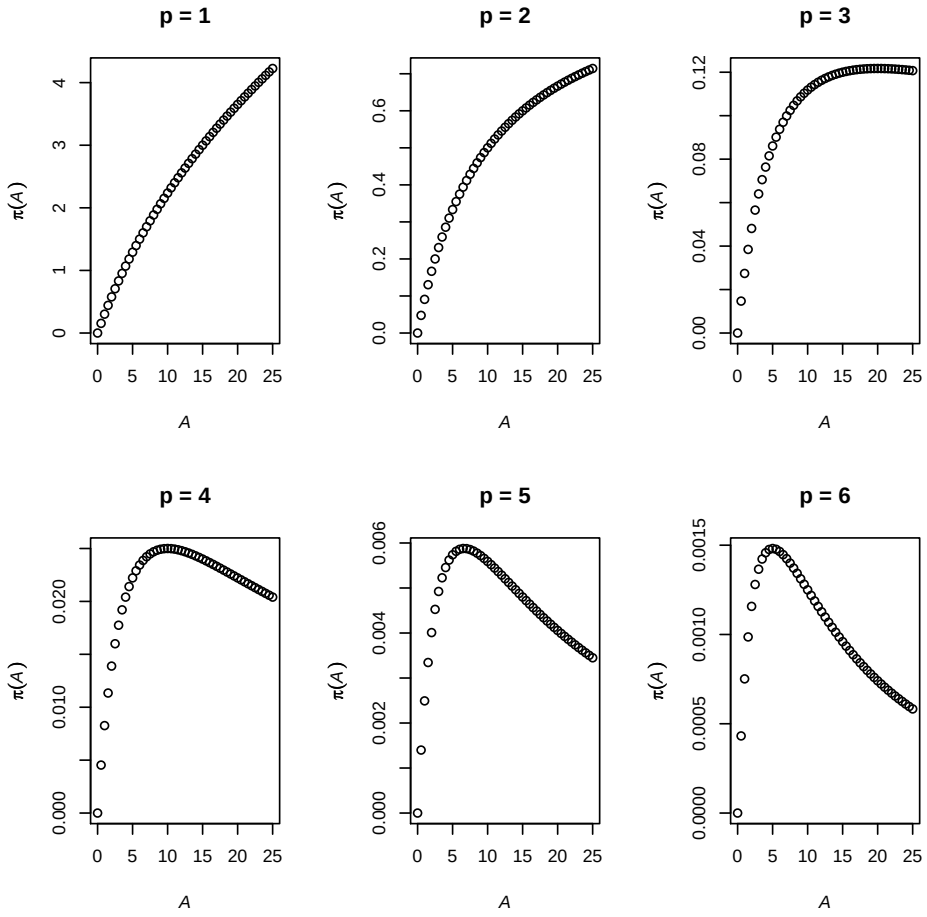

Figure 2.1: Plot of the proposed prior (2.10) for different p (number of covariates) and fixed central tendency measure d_0 (=10) of the sampling variances

2.3 Approximate Hierarchical Bayes Method

The posterior moments of B_i and θ_i, under the priors considered in Section 2.2, are not in a closed-form, but can be obtained either by numerical integration or by Monte Carlo Markov Chain (MCMC) method. We have written a program, using the `BRugs` package (Thomas et al., 2006) in `R` (R Development Core Team, 2008), that allows the hierarchical Bayes analysis for our new priors using the MCMC method. But its slow computation speed does not permit its evaluation by repeated use in simulation. Thus, for convenient implementation and evaluation of our hierarchical

Bayes method, we approximate the posterior moments of B_i and θ_i using Laplace's method. Before going into the detailed implementation of Laplace's method under the Fay-Herriot model, we would like to address another method of approximating nonstandard complex densities viz., the adjusted density method along with its shortcomings.

2.3.1 Adjusted Density Method

We discussed the adjusted density method (Morris, 1988) with some details in Section 1.4.4.1. This method is useful for noninformative Bayesian analysis i.e. when noninformative priors, say uniform, are used at the third level. When we have a subjective prior in hand, chosen by following some objective criteria as in Section 2.2 or on the basis of some prior knowledge, application of this method changes the prior ultimately because of the additional adjustment term and restricts us to use the subjective prior in the analysis. This fact can be illustrated as follows.

Following the adjusted density method of Morris (1988), if we want to approximate the posterior density of B_i by a beta density, say Beta(a_i, b_i), first we write the adjusted log-posterior density $l\,(B_i)$ of B_i.

$$\begin{aligned} l(B_i) &= \log\{f(B_i|y)B_i(1-B_i)\} \\ &= \log\{f(A|y)\ |J|\ B_i(1-B_i)\} \\ &= \log\left\{L_R(A)\frac{A^2}{(A+d_0)^{p/2}}\right\} \end{aligned} \tag{2.11}$$

where $|J|$ is absolute value of the jacobian of transformation $A \to B_i$. Note that, in this example $B_i(1-B_i)$ is the required adjustment term, as suggested by Morris

(1988). Now, if we match the first derivative of (2.11) to the beta mean, we are basically considering the mode of a different posterior density which uses $\pi(A) \propto \frac{A^2}{(A+d_0)^{p/2}}$ instead of the prior mentioned in (2.10). Although, the approximation performs quite accurately as illustrated by the SAIPE data analysis (see Section 2.5; Figure 2.4). For more on the approximation i.e. how to obtain the parameters of the beta density in order to implement the adjusted density method in practice see Appendix B of this Chapter. Another shortcoming of this method is that it does not generalize in an obvious way to the multidimensional case whereas Laplace's method can be used to approximate the posterior moments easily for the multidimensional case, albeit ours is a univariate case.

2.3.2 Laplace Approximation

Let, $\hat{A}$ be the posterior mode, which is obtained by maximizing the posterior distribution $f(A|y)$, with $\pi(A) = \pi_{LL}(A)$, and let i_0 be the negative second derivative (Hessian) of the log posterior density evaluated at $\hat{A}$. Also, we assume that the Hessian should be bounded away from zero at the mode. Now, following (Kass & Steffey, 1989), the first order approximation to the posterior mean and variance of any real-valued smooth function, $g(A)$, of A is given by

$$E\{g(A)|y\} = g\left(\hat{A}\right) + O\left(m^{-1}\right) \tag{2.12}$$

$$V\{g(A)|y\} = \left\{g'\left(\hat{A}\right)\right\}^2 \frac{1}{i_0} + O\left(m^{-2}\right), \tag{2.13}$$

where $g'\left(\hat{A}\right)$ is the first derivative of $g(A)$ at $\hat{A}$. This follows from the standard form of Laplace approximation to the ratio of integrals (Tierney et al., 1989). Both

the approximations in (2.12) and (2.13) have relative error $O\left(m^{-1}\right)$ and are termed as first order approximation. By relative error we mean the error term corresponding to the quantity $\frac{\text{approximation}}{\text{true}}$ and the true variance is possibly of order $O\left(m^{-1}\right)$.

We now discuss this method in the context of obtaining the first order approximation for the posterior moments of θ_i, under the model described in Section 2.2. From (2.7), we define

$$g_i(A) = E(\theta_i|y, A) = (1 - B_i)y_i + B_i x_i'\hat{\beta}_A$$

and

$$h_i(A) = V(\theta_i|y, A) = D_i(1 - B_i) + B_i^2 x_i'\Sigma_A x_i$$

Using iterative expectation technique on $E(\theta_i|y)$, we write the first order approximation to the posterior mean of θ_i as

$$E(\theta_i|y) = E\left\{g_i(A)|y\right\} = E(\theta_i|y, \hat{A}) + O(m^{-1}), \tag{2.14}$$

follows directly from (2.12). Using similar iterative expectation and variance technique on $V(\theta_i|y, A)$, the first order approximation to the posterior variance can be written as

$$\begin{aligned} V(\theta_i|y) &= E\left\{h_i(A)|y\right\} + V\left\{g_i(A)|y\right\} \\ &= \left[V(\theta_i|y, \hat{A}) + O(m^{-1})\right] + \left[\left\{E'(\theta_i|y, A)\right\}^2_{A=\hat{A}} \ \frac{1}{i_0} + O(m^{-2})\right] \quad (2.15) \\ &= D_i(1 - B_i) + \hat{B}_i^2 \ x_i'\Sigma_{\hat{A}}x_i + \\ &\quad \left\{y_i - x_i'\hat{\beta}_A + (A + D_i)x_i'u\right\}^2_{A=\hat{A}} \ V(B_i|y) + O(m^{-1}), \quad (2.16) \end{aligned}$$

where $\hat{B}_i$ and $\Sigma_{\hat{A}}$ are B_i and Σ_A evaluated at $\hat{A}$; $u = \frac{\partial\hat{\beta}_A}{\partial A}$. The first and the second

part of the equation (2.15) follows from (2.12) and (2.13) respectively. The term $V(B_i|y)$ can be obtained by applying (2.13) with $g_i(A) = D_i/(A + D_i)$.

The first order approximation of the posterior variance given by (2.16) has three clearly defined terms. The first term, $T_1 = D_i(1 - \hat{B}_i)$, measures the uncertainty in the model for estimating θ_i, the second term, $T_2 = \hat{B}_i^2 x_i' \, \Sigma_{\hat{A}} \, x_i$, measures the uncertainty in the estimation of β and the third term, T_3 (remaining portion), accounts for the uncertainty in estimating A. In many applications, the third term is quite small and often ignored (Bell, 1999). Kass & Steffey (1989) emphasized the importance of the third term in their data analysis. They considered the standard form for the first order Laplace approximation under a similar kind of model as Fay-Herriot, but with uniform prior for the variance component, under the label conditionally independent hierarchical models (CIHM).

To obtain a second order approximation of the posterior mean in standard form, we need to evaluate the third order derivatives of the log-likelihood. More complexity arises if we need a second order variance approximation, which demands fourth and fifth derivatives of the log-likelihood (Kass et al., 1988, p. 265). To avoid this complication, Tierney & Kadane (1986) proposed fully exponential form of Laplace approximation. The fully exponential form is applicable only for functions g bounded away from zero. We describe the method below in brief.

$$E\{g(A)|y\} = \frac{\sigma^*}{\sigma} \exp\left\{ m \left(L^*(\hat{A}^*) - L(\hat{A}) \right) \right\}, \tag{2.17}$$

where $L = \frac{1}{m} \log f\,(A\,|y)$, $L^* = \frac{1}{m} \{\log f\,(A\,|y) + \log g\,(A)\}$, and $\hat{A}^*$ is the maximizer of L^*. σ^2 and σ^{*2} be the inverse of the negative Hessian of L and L^* evaluated at

$\hat{A}$ and $\hat{A}^*$ respectively. To find the second order approximation of $V\{g(A)|y\}$, we write

$$V\{g(A)|y\} = E\left\{g^2(A)|y\right\} - E^2\{g(A)|y\}. \tag{2.18}$$

Then, apply (2.17) to the first term on the right hand side of (2.18) with $L^* = \frac{1}{m}\{\log f + \log g^2\}$ and subtract the square of (2.17). A similar approximation applies in the multiparameter case where the σ's are replaced by the determinant of the negative Hessian matrix. In this specific implementation of Laplace approximation, the approximations are more accurate than the conventional Laplace approximation. For this approximation the errors are $O(m^{-2})$, where as in conventional Laplace approximation the errors are $O(m^{-1})$. See Tierney & Kadane (1986) for the derivation of the error terms. Datta et al. (2005) considered standard form of second order Laplace approximation for approximating posterior mean of θ_i under Fay-Herriot model, which requires evaluation of the third derivatives of the log-likelihood. But their approximation to the posterior variance has relative error $O(m^{-1})$.

Tierney & Kadane (1986) warned against using the fully exponential method for nonpositive function. According to them, the positivity assumption is important to ensure that the numerator and denominator integrands are similar in shape so that application of the Laplace approximation to the numerator and denominator leads to similar error terms, and in taking the ratios, the order $O(m^{-1})$ terms cancel in both numerator and denominator. Tierney et al. (1989) extended the fully exponential approach for approximating moments of functions taking both negative and positive values, such as regression coefficients. They used the term moment

generating function (MGF) method, which finds the second order approximations of posterior moments of nonpositive functions without requiring the evaluation of third derivative of likelihood function. It is instructive to note that the posterior mode of A resulting from our prior always lies in the interior of the parameter space, and hence application of Laplace method is straightforward. For asymptotic expansions of posterior expectations when the mode is at the boundary, see Erkanli (1994).

2.4 Simulation Study

In this section, we compare two frequentist approaches with our hierarchical Bayes method. The conceptual difference between a frequentist and a Bayesian analysis are avoided by studying the frequentist properties of the estimators under two approaches. The frequentist properties of estimators are derived from the distribution of the estimators under repeated simulations of observations following the statistical model with known hyperparameters. We investigate the small sample ($m = 15$) frequentist properties of our proposed priors $\pi_M(A)$ and $\pi_{LL}(A)$ in estimating A, B_i, and θ_i, $i = 1, \ldots, m$, using a Monte Carlo simulation study. In the tables, we denote the hierarchical Bayes methods resulting from $\pi_M(A)$ by $\mathrm{HB_M}$ and that from $\pi_{LL}(A)$ by $\mathrm{HB_{LL}}$. This is generic and used for the estimation of any parameter. The hierarchical Bayes methods are approximated, using the approximations discussed in Section 2.3.2. For the estimation of A, we compare the posterior mode using new priors with the residual maximum likelihood (REML)

and mean likelihood (MEL) estimators, considered by Bell (1999). For the estimation of the shrinkage parameters, we consider the Laplace approximation method for our two priors along with the usual plugged-in REML and MEL estimators in $B_i = D_i/(A + D_i)$. For inference about the small area means θ_i, we compare our hierarchical Bayes estimators with the empirical best linear unbiased predictors (same as empirical Bayes), using the REML and MEL estimators of A.

2.4.1 Design of the Simulation Study

For our simulation experiment, we use the Fay-Herriot model with $m = 15$, and $A = 1$. We consider one covariate, which is generated from a gamma distribution with mean 10 and variance 50, and an intercept term i.e., we considered $p = 2$. We assume $\beta = (-2, 0.5)$. Following Datta et al. (2005), we divide the 15 small areas into 5 groups: G1, G2, G3, G4, and G5, each group containing 3 small areas with identical sampling variance. We consider the following sampling variance pattern for the five groups: (4.0, 0.6, 0.5, 0.4, 0.1). That is, $D_i = 4.0$, for the three small areas in group G1, $D_i = 0.6$, for the three small areas in group G2, and so on. This sampling variance pattern is the most variable pattern among the three patterns considered by Datta et al. (2005) and corresponds to their Type 3 pattern. In our simulation we multiply the sampling variances by 3 to increase the values of the variances. The median of D_i's is 1.5 i.e., we considered $d_0 = 1.5$ for our proposed prior. We generate $N = 10,000$ replicates of $\{(y_i, \theta_i),\ i = 1, \ldots, m = 15\}$ using the Fay-Herriot model in order to study various frequentist properties of our approximate

hierarchical Bayesian methods.

2.4.2 Comparison of Different Estimators of A

To compare the different estimators of A, we use the following criteria. Let $\hat{A}^{(j)}$ be an estimate of A in the jth simulation run, $j = 1, \ldots, N(= 10,000)$. We define,

- Bias$(A) = \bar{\hat{A}} - A$
- Variance$(A) = \frac{1}{N-1} \sum_{j=1}^{N} (\hat{A}^{(j)} - \bar{\hat{A}})^2$, where $\bar{\hat{A}} = \frac{1}{N} \sum_{j=1}^{N} \hat{A}^{(j)}$
- MSE = Variance + Bias^2

We define the bias, variance, and MSE for estimators of B_i and θ_i in a similar way. Since the estimators are biased, as a measure of uncertainty, we present the MSE's in the tables instead of variance. Before comparing different estimators of A, we spell out the proportion of zero estimates obtained over the simulation runs for all the four estimators. Only in this situation, we explore several choices of A. For other results we have used $A = 1$. Table 2.1 shows that the percentage of zero estimates for REML can be substantial for small values of A. Here it should be noted that the cases where REML turned out to be negative, it is truncated at zero. As expected, all other estimators always produce positive estimates.

Table 2.2 shows that the bias of REML is negligible (1.59 %) and is the least biased among all the methods considered. This is consistent with the theory, since the bias of REML is of order $o(m^{-1})$, under standard regularity conditions (Jiang,

Table 2.1: Percentage of Zero Estimates for Different Estimators of the Variance Component A

A	REML	MEL	HB_{LL}	HB_{M}
0.5	27	0	0	0
1	12.6	0	0	0
2	4.2	0	0	0
5	0.4	0	0	0

Table 2.2: Bias and MSE (in percent) of Different Estimators of the variance component A

	REML	MEL	HB_{LL}	HB_{M}
Bias(%)	1.59	141.84	45.46	93.24
MSE(%)	76.69	391.01	91.13	206.98

1996). However, the REML is subject to zero estimates. The MEL always produces positive estimates of A, but is subject to an overestimation problem. The posterior mode HB_{LL}, obtained using the prior $\pi_{LL}(A)$, performs the best among the three estimators, which never yield zero estimates of A. It is also better than the other three estimators, except REML, in terms of the MSE.

2.4.3 Comparison of Different Estimators of B_i

Table 2.3 reports the average percent bias and MSE of different estimators of B_i for each of the five groups, where the average is taken over all the three small areas in a group. Such an average makes sense, since B_i's are identical within group. The bias of the plugged-in REML estimator of B_i is worse than the bias of the REML estimator of A, especially for groups with small D_i's. This follows from

Table 2.3: Bias (and MSE) in percent of Different Estimators of the shrinkage factor B_i

Group	REML	MEL	HB_{LL}	HB_M
G1	0.25 (0.34)	-8.45 (1.24)	-2.78 (0.35)	-5.64 (0.72)
G2	5.04 (3.79)	-17.44 (4.86)	-5.73 (2.06)	-12.26 (3.31)
G3	6.03 (4.43)	-17.33 (4.78)	-5.60 (2.13)	-12.19 (3.31)
G4	7.33 (5.27)	-16.85 (4.51)	-5.30 (2.14)	-11.86 (3.19)
G5	13.91 (9.85)	-8.99 (1.32)	-2.02 (0.97)	-6.17 (1.07)

the Jenson's inequality, since B_i is a convex function of A. Even if we plug in an unbiased estimator of A, The estimator of B_i would have a positive bias. The REML is subject to an overestimation problem whereas the other estimators are subject to underestimation problem. This is an interesting observation, since this implies, in estimating θ_i, REML tends to put less weight on the direct estimator than the hierarchical Bayes do. Also, unlike the other estimators, REML could estimate B_i by 1, in which case the REML-based empirical best linear unbiased predictor of θ_i is identical to the regression synthetic estimator – this is surely an undesirable feature of REML. Among the other three methods, plugged-in MEL appears to yield the most negatively biased estimator of B_i. The hierarchical Bayes method HB_{LL} appears to be a good compromise. Here, HB_{LL} is the first order Laplace approximation to the posterior mean of B_i, using prior $\pi_{LL}(A)$. In this case, HB_{LL} is essentially the plugged-in estimator of B_i with new posterior mode of A. In terms of MSE, the performance of HB_{LL} is the best.

Figure 2.2 plots different average B_i estimates, the averages being taken over all the 10000 simulation runs, against small areas, arranged in decreasing order of

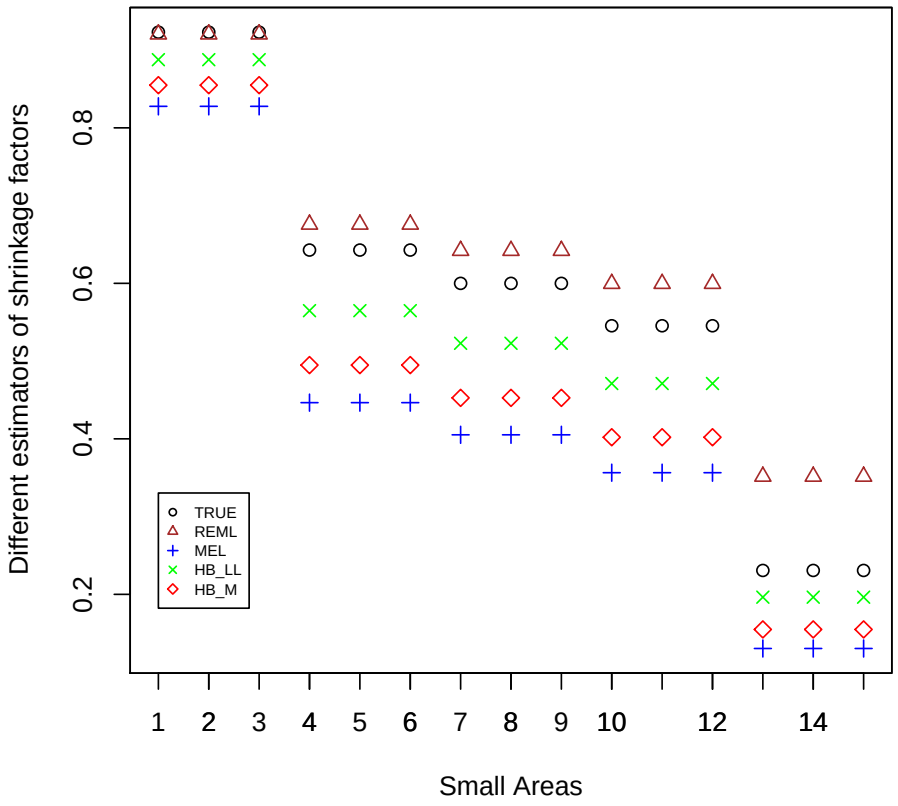

Figure 2.2: Plot of average values of the different estimates of the shrinkage factor B_i along with true values: average being taken over 10,000 simulations

the sampling variances, D_i. For all the estimators, the weight given to the direct estimator increases with the decrease in the sampling variance. The REML estimator puts the lowest weight to the direct estimator whereas both MEL and $\mathrm{HB_M}$ put more weight to the direct estimator compared to $\mathrm{HB_{LL}}$. In a sense the estimator $\mathrm{HB_{LL}}$ strikes a nice balance in putting weight to the direct and synthetic estimator compared to other three estimators.

2.4.4 Comparison of Different Estimators of θ_i

In small area estimation the objective of interest is usually to draw inferences about the small area mean θ_i. In this section, we compare four estimators of θ_i. REML and MEL are the plugged-in EBLUP estimator, two hierarchical Bayes estimators, HB_{LL} and HB_{M}, are the first order Laplace approximation to the posterior mean of θ_i with the corresponding prior. To compare different estimators of θ_i, we introduce two other summary statistics, besides bias, variance and MSE defined in Section 2.4.2. We study the frequentist coverage properties of the 95% confidence interval (CI) of the type: Estimator $\pm$ 1.96 $\times$ MOU, constructed using the four estimators based on the normal approximation. By MOU we mean the measure of uncertainty, e.g., the posterior standard deviation for a Bayes estimator. Since we have both Bayesian and frequentist estimators, we prefer to use this term for all the estimators, instead of using $\sqrt{MSE}$. For the comparison purposes, we define the following two summary statistics:

- Coverage probability of 95% CI = [number of times true θ_i is included in the interval]$/N$
- Average length of CI = $\frac{1}{N}\sum_{j=1}^{N}(UCL_j - LCL_j)$, where LCL_j and UCL_j are the lower and upper confidence limit of the corresponding CI.

Table 2.4 presents the percent MSE, coverage probability for a 95% confidence interval, and average length of the confidence interval for each of the four methods. Here, instead of considering summary statistics for 15 small areas we report the average summary statistics for each of the five groups, where the average is taken

over all the three small areas in a group. Although the areas within a group differ due to the differences in the covariate, unlike Datta et al. (2005) (they considered common mean model in the simulation), we adapt this approach to save space. On the basis of MSE, it is clear that $\mathrm{HB}_{\mathrm{LL}}$ performs better than MEL and HB_{M} for all the small areas, although the differences in the MSE tend to decrease as the sampling variance decreases. In terms of MSE, the performance of REML and $\mathrm{HB}_{\mathrm{LL}}$ is comparable, $\mathrm{HB}_{\mathrm{LL}}$ performs slightly better except for the group with the highest sampling variance.

To construct the confidence interval for REML, as a measure of uncertainty we use square root of both: (1) the naïve MSE estimator that does not take into account the uncertainty in estimating A, and (2) the approximately unbiased MSE estimator of the MSE of EBLUP, which takes into account all sources of uncertainties. For an explicit expression of the MSE estimator of the EBLUP of θ_i, under Fay-Herriot model, see Datta & Lahiri (2000). For MEL estimator of θ_i, we consider the naïve MSE estimator of the EBLUP of θ_i with plugged-in MEL estimate of A. Bell (1999) showed that this naïve MSE estimator with plugged-in MEL estimate of A estimates the measure of uncertainty quite accurately. For two Bayesian estimators we used the first order Laplace approximation to the posterior variance of θ_i. Before comparing the confidence intervals, we need to study the performance of the estimators of the measure of uncertainty which we use to construct the confidence intervals. To this end, we look at the ratio of average MOU to the true MOU (simulated MSE of different estimators of θ_i). We consider the simulated MSE as the true MOU as the estimators of MOU are biased. It is desirable that the ratio

is close to 1. Figure 2.3 shows that, for REML with MOU as in Datta & Lahiri (2000), these ratios are close to 1 for all the areas. For HB_LL, this ratio is close to 1 for all the areas, except for the areas with highest sampling variance.

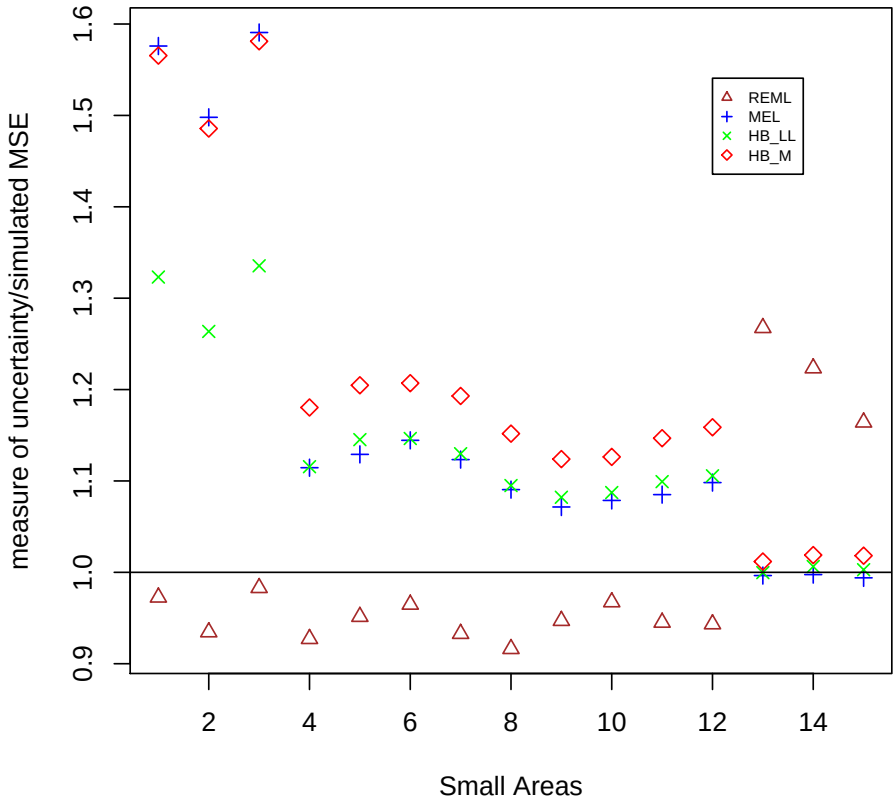

Figure 2.3: Plot of measure of uncertainty ratios for different estimators of the measure of uncertainty of the small area mean θ_i for 15 small areas

Regarding the performance of the confidence intervals, we can say that the REML has got the lowest coverage rate, as low as 0.84 for some areas, compare to the nominal rate of 0.95. The coverage property improves when we consider the measure of uncertainty in estimating A, as opposed to ignoring it. Still REML suffers from the undercoverage problem because of the zero estimation problem of

A. In spite of good MSE properties of the REML estimator of θ_i, its lower coverage rate restricts its use in real life data analysis. All other methods provide confidence intervals having at least nominal coverage rate. So, in terms of coverage, other three methods are comparable, although $\mathrm{HB}_{\mathrm{LL}}$ performs the best (smallest average length) in terms of the average length.

In this subsection, we also study the relative contributions of the three terms to the posterior variance of θ_i obtained using the prior $\pi_{LL}(A)$. In Table 2.5, the columns T1, T2, and T3 exhibit the relative contributions of term 1, term 2, and term 3 respectively. For the explicit expressions of the terms see the equation (2.16). From the values in Table 2.5, we conclude the contribution of the term which accounts for the uncertainty in estimating the variance component is substantially small relative to the first term which accounts for the uncertainty in the model in estimating the small area mean.

Table 2.4: Comparison of Different Estimators of Small Area Mean θ_i

	MSE (%)				Coverage probability of 95% CI					Average length of CI				
Group	REML	MEL	HB_{LL}	HB_M	REML[1]	REML[2]	MEL	HB_{LL}	HB_M	REML[1]	REML[2]	MEL	HB_{LL}	HB_M
G1	110.78	128.49	111.36	118.46	0.84	0.86	0.98	0.96	0.97	3.65	4.16	5.55	4.65	5.25
G2	108.84	113.23	105.92	109.48	0.88	0.89	0.96	0.95	0.96	3.50	3.43	4.29	4.11	4.28
G3	75.28	80.41	72.12	76.20	0.85	0.89	0.96	0.96	0.96	2.83	3.34	3.71	3.50	3.70
G4	71.50	74.96	68.33	71.59	0.86	0.91	0.96	0.95	0.96	2.77	3.19	3.50	3.34	3.50
G5	26.59	26.77	25.94	26.17	0.92	0.96	0.95	0.95	0.95	1.87	2.25	2.04	2.01	2.03

[1] naïve MSE estimate is used, does not consider the uncertainty in estimating A

[2] estimated MSE as documented in Datta & Lahiri (2000) is used

Table 2.5: Relative contribution (%) of the three terms to the posterior variance of θ_i using the prior $\pi_{LL}(A)$ on A: average over 10,000 simulations and 3 small areas within each group

Group	T1[1]	T2[2]	T3[3]
G1	83.70	11.86	4.44
G2	69.28	24.85	5.87
G3	84.13	7.84	8.02
G4	82.03	10.45	7.52
G5	91.61	4.29	4.10

[1] Uncertainty in the model (first two levels of FH model, assuming β and A known) in estimating θ_i

[2] Uncertainty in estimating the regression coefficient β

[3] Uncertainty in estimating the variance component A

2.5 SAIPE Data Analysis

In this section, we analyze the SAIPE state level data for the year 1997 and 1993 using the Fay-Herriot model. We compare our approximate hierarchical Bayes method with the method used by the U.S. Census Bureau as documented in Bell (1999). In our notation, y_i denotes the direct survey (CPS) estimate (expressed as percentage) of the true poverty ratio θ_i of school-age children in state $i,\ i = 1, \ldots, 51$ and x_i is a 5×1 vector consisting of an intercept term and four auxiliary variables obtained from administrative records as mentioned above. The sampling variances D_i's were obtained from the sampling error model of Otto & Bell (1995); they fitted a generalized variance function (GVF) to five years of direct variance and covariance estimates for each state produced by Fay & Train (1995). The sampling variances are assumed to be known throughout the inferential procedure i.e., the uncertainty about the sampling error is not considered in the analysis. For further details on the auxiliary variables, see Bell (1999) and Citro & Kalton (2000). In the tables and figures, we denote the hierarchical Bayes methods resulting from $\pi_{LL}(A)$ by $\mathrm{HB_{LL}}$ or HB_LL. This is generic and used for the estimation of any parameter. The hierarchical Bayes method is approximated, using the Laplace approximations proposed in Section 2.3.2.

2.5.1 Evaluation of the Adjusted Density Method

In this subsection, we evaluate the performance of the adjusted density method discussed in Section 2.3.1, in the context of the SAIPE 1997 data. For illustration,

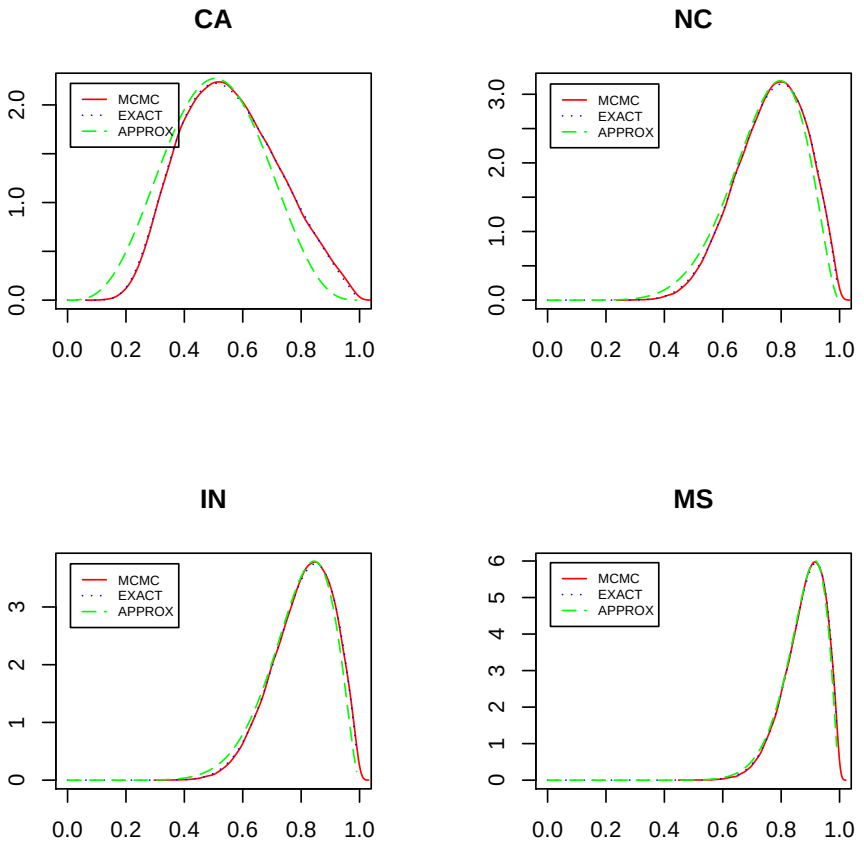

Figure 2.4: Evaluation of the adjusted density method proposed by Morris: Approximating the posterior density of the shrinkage factor B_i by a beta density

we consider the posterior distributions of B_i under prior $\pi_{LL}(A)$ for four states: California (CA), North Carolina (NC), Indiana (IN), and Mississippi (MS), using 1997 SAIPE data. Bell (1999) examined these four states for his data analysis for the years 1989-1993. We believe this is a representative sample from the 51 states to cover the range of sampling variances and sample sizes within states. Since the posterior distribution of B_i is in the form of a one-dimensional integral, it is possible to obtain a highly accurate approximation to the posterior density using numerical integration. This gives us an opportunity to compare the Beta approximation,

suggested by Morris, with the MCMC method, treating the numerical integration as an exact method. In Figure 2.4, we plot the posterior densities of B_i using numerical integration (dotted line) with prior $\pi_{LL}(A)$, densities obtained from applying MCMC method (solid line) with prior $\pi_{LL}(A)$ using the `BRugs` package (Thomas et al., 2006) in `R` (R Development Core Team, 2008) and approximate beta densities (long-dashed line) of B_i. Comparing the densities we can readily conclude that the approximation, as obtained using the adjusted density method, performs quite accurately for all the states (slight deviation for CA) but the approximation uses a different prior on A rather than the prior $\pi_{LL}(A)$ [see equation (2.11)].

2.5.2 Comparison of Different Estimators of A

In Figure 2.5, we present different posterior densities of A using two different priors on A: uniform prior and $\pi_{LL}(A)$. We can readily see, for SAIPE 1997 data, the posterior mode under uniform prior is zero, an undesirable estimate of the variance component A, which is by definition positive. The use of prior $\pi_{LL}(A)$ pushes the modes of the posterior densities, for both 1993 and 1997, to the right yielding the posterior mode HB_{LL} that is always positive, like the MEL estimate. Table 2.6 compares three different estimates of A: REML (posterior mode using uniform prior), MEL (posterior mean using uniform prior as in Bell 1999), and HB_{LL} (posterior mode using prior $\pi_{LL}(A)$). It is interesting to note that, for both the years considered, the difference between HB_{LL} estimate and REML is much less than that of MEL and REML.

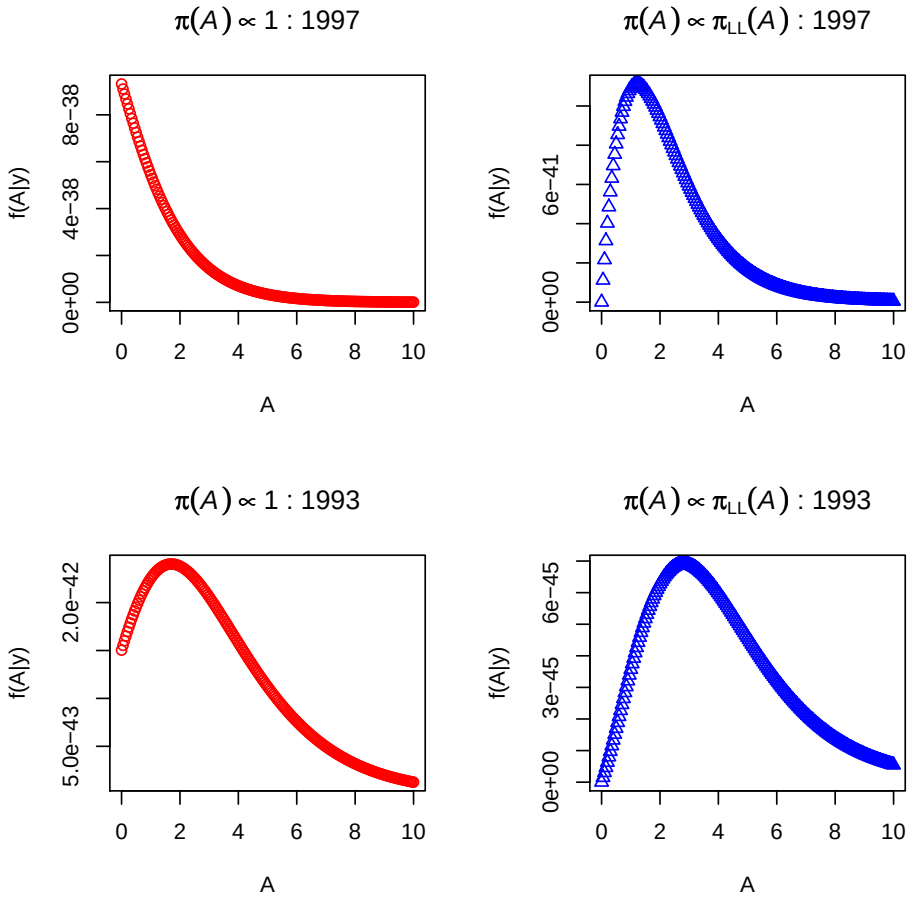

Figure 2.5: Plot of the posterior density of A using two years of SAIPE data: a comparison between uniform prior and proposed prior under Fay-Herriot model

2.5.3 Results Comparing Different Estimators of B_i

Since we have 51 small areas, i.e. 50 states and the District of Columbia, we compare the estimates of B_i using a graph. Figure 2.6 plots five estimates of B_i: REML, MCMC, HB_LL, EXACT and MEL, against states arranged in increasing order of the sampling variances D_i. Note that MEL and REML are regular plug-in estimates of B_i, whereas HB_{LL} is obtained using the second order Laplace approximation to the posterior mean of B_i. In other words, HB_{LL} is obtained by applying the fully exponential form of Laplace approximation. The estimator MCMC is the

Table 2.6: Different variance estimates of A for two years of SAIPE data

Year	REML	MEL	HB_{LL}
1993	1.70	3.37	2.80
1997	0	1.50	1.22

posterior mean of B_i obtained by applying MCMC method with prior $\pi_{LL}(A)$ using the `BRugs` package in `R`. The EXACT estimator of B_i is computed by performing numerical integration over 100 equal subintervals of A. Since REML estimate of A is zero for 1997, $B_i = D_i/(A + D_i) = 1$ for all the states. Thus, for any given state, irrespective of its size, the REML puts all the weight to the regression synthetic estimate and none to the direct estimate. This is not natural since, at least for large states, like California (CA), we would expect a reasonable procedure to put large weight to the direct estimate. From the Figure 2.6, it is evident that the Laplace approximation performs very well as it coincides with EXACT and MCMC for almost all the states. It is worth mentioning that our estimate would put more weight on the direct estimate of θ_i, compared to the MEL estimate in an empirical Bayes set up.

2.5.4 Results Comparing Different Estimators of θ_i

Table 2.7 presents some estimates of θ_i and the associated measure of uncertainty while considering the approximate hierarchical Bayesian approach using the Fay-Herriot model with our new prior. Also shown are the CPS sample sizes n_i (number of households in March CPS sample), CPS direct poverty ratio estimates

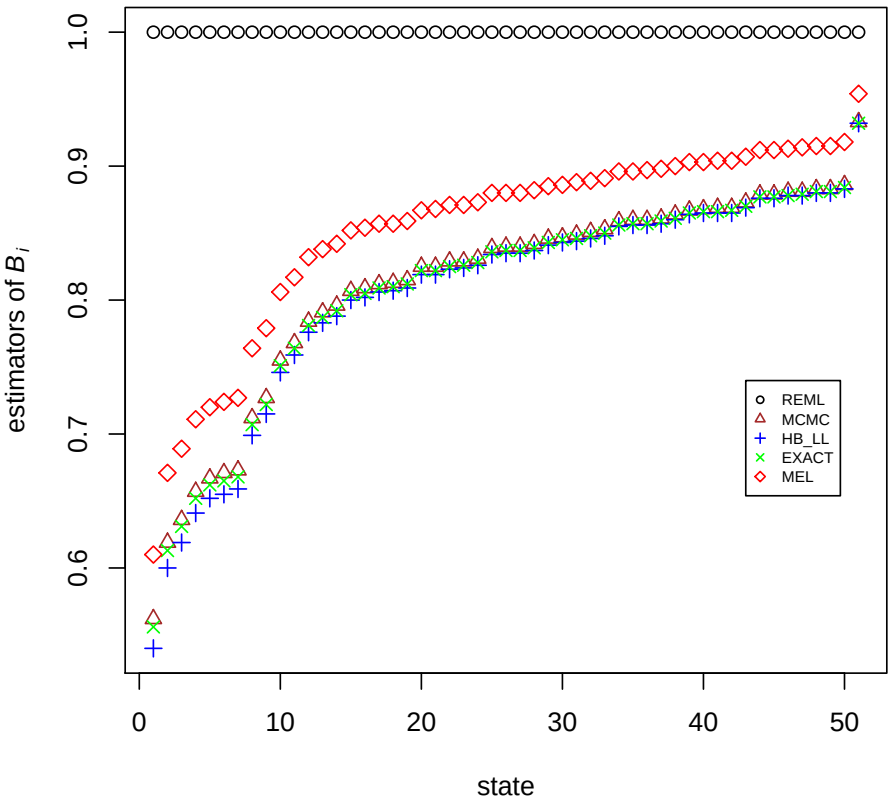

Figure 2.6: Plot of different shrinkage factor estimates using SAIPE 1997 data: Comparison between MEL, REML and our method considering MCMC and numerical integration (exact) as gold standard

y_i, and direct sampling variances D_i. Results are presented for 10 selected states in increasing order of D_i. In addition to the four states (CA, NC, IN, MS), as in Bell (1999), we include six other states: two small states (DC, MD), two large states (FL, NY) and two moderately large states (MA, VA). Size of the states is categorized on the basis of the CPS sample size. We believe that these ten states are a good representation of the 51 states as far as the discrepancies in sampling variance and sample size within states are concerned. Table 2.7 includes the synthetic estimate of θ_i (obtained by plugging in the posterior mode of A in $\hat{\beta}$), the first and second

Table 2.7: Approximations to the small area mean θ_i and their measure of uncertainty using Fay-Herriot model with new prior: SAIPE 1997 data

State	D_i	n_i	y_i	$x_i'\hat{\beta}$	Lo1	Lo2	Sdo1	V1	V2	V3
CA	2.34	4465	23.55	22.71	22.99	23.07	1.24	51.77	47.37	0.85
NY	3.05	3350	23.76	23.16	23.33	23.39	1.27	53.96	45.60	0.42
FL	3.93	2613	18.34	19.82	19.47	19.29	1.22	62.08	33.15	4.76
MA	6.22	1187	19.66	15.04	15.79	16.22	1.47	46.55	39.14	14.30
NC	6.70	1238	13.89	16.98	16.50	16.22	1.20	71.13	19.50	9.36
IN	8.74	684	11.07	13.082	12.83	12.67	1.20	74.01	23.30	2.68
VA	10.94	736	16.40	15.38	15.48	15.54	1.15	82.38	17.28	0.32
MD	12.21	564	9.89	13.56	13.22	12.98	1.27	67.82	26.57	5.60
MS	15.88	591	20.58	22.18	22.06	22.02	1.67	40.44	59.48	0.06
DC	30.81	548	35.85	32.42	32.55	32.67	2.47	19.07	80.60	0.32

order Laplace approximation (Lo1 and Lo2 column) to the posterior mean of θ_i.

Note that the first order approximation is basically the empirical Bayes estimator, with $\mathrm{HB}_{\mathrm{LL}}(\hat{A})$ plugged-in the Bayes estimator of θ_i when A is known. The next column (Sdo1) shows the square root of the first order approximation of the posterior variance of θ_i. In order to have an analytical expression of the posterior variance for illustration purposes, we consider the standard form of first order Laplace approximation that makes it possible to present three clearly defined components in the variance [see (2.16)]. We wish to explore the importance of the third component which is usually small compared to the other two terms and is ignored. The relative contributions (%) of the three components towards the posterior variance are given in the last three columns (V1, V2, V3) of Table 2.7. For the state MA, the third component is substantial (14%) compared to that of other states because of the substantial difference between the direct estimate (19.66) and synthetic estimate (15.04). The maximum contribution (19%) of the third component towards

the variance is observed for the state MI (not presented in the table). It can be concluded that for this particular dataset, if we consider first order approximation of the posterior variance, then the third term is small compared to other two terms. If we compare Lo1 and Lo2 column of Table 2.7, we can say that there is not much difference between the first and second order approximation of posterior mean of θ_i. For a more extensive evaluation of the approximation see Section 2.5.5.

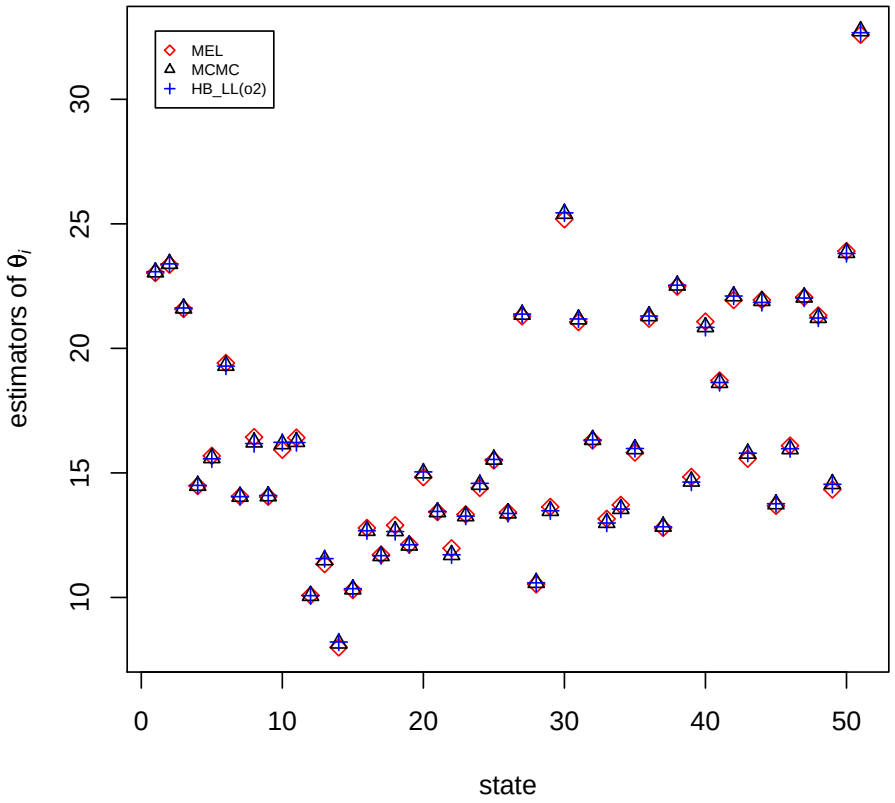

Figure 2.7: Point estimates of small area means θ_i: Comparison between Bell (1999)≡MEL and our≡HB_LL(o2) method (with MCMC as the gold standard) using SAIPE 1997 data (By increasing D_i)

Figure 2.7 plots three point estimates of θ_i, MEL, MCMC, HB_{LL}. The MEL estimate is obtained by plugging in the MEL estimate of A in $E(\theta_i|y, A)$. The

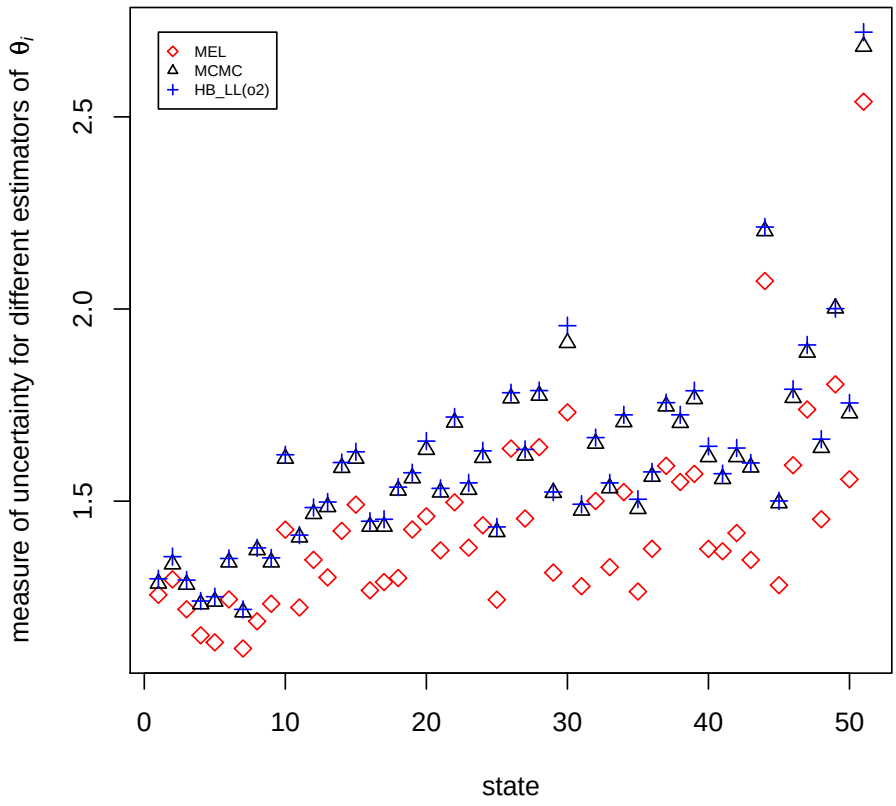

Figure 2.8: Plot of measure of uncertainty of the estimators of small area means: Comparison between Bell (1999)≡MEL and our≡HB_LL(o2) method (with MCMC as the gold standard) using SAIPE 1997 data (By increasing D_i)

MCMC estimator is the posterior mean of θ_i under Fay-Herriot model using the new prior. HB_LL(o2) is the second order Laplace approximation to the posterior mean of θ_i obtained by applying (2.17) with $g_i(A) = (1 - B_i)\, y_i + B_i x_i' \hat{\beta}_A$. Although, the positivity of this function cannot be guaranteed, in general, but in our data analysis the function $g_i(A)$ is always positive. That's why we think the application of (2.17) is valid. No significant disparity is observed among the three estimators. Figure 2.8 demonstrates the difference between the measures of uncertainty of the three estimators considered in Figure 2.7. The HB_LL estimator (second order

Laplace approximation to the posterior variance of θ_i) is obtained by applying (2.17) and (2.18) with $g(A) = D_i(1 - B_i) + B_i^2 x_i' \Sigma_A x_i$. The overlapping of the MCMC (posterior variance of θ_i using new prior) and HB_LL estimator justifies the efficiency of the approximation. In general, the measure of uncertainties associated with the three estimators increase with the increase of the sampling variance D_i. The measure of uncertainty of MEL discussed in Bell (1999) seems to suffer from an underestimation problem because it ignores the uncertainty in estimating of A.

2.5.5 Evaluation of Laplace Approximation

In this section, with the help of Figure 2.9 and Figure 2.10, we evaluate the precision of the Laplace approximation to the posterior moments of θ_i using SAIPE 1997 data. To measure the accuracy, we consider the percentage difference as the summary statistics. Mathematically, this can be defined as $\{(\text{exact} - \text{approximate})/\text{exact}\} \times 100$. We obtain exact posterior moments by performing numerical integration over 100 equal subintervals of A. Figure 2.9 shows that both the first and second order approximation values of mean are quite close to the exact value, although, the second order values are more accurate as the percentage difference values lie on the zero line for almost all areas. Evidently, the first order variance approximation underestimates the uncertainty, the percentage difference being more than 20% for most states with highest being 35% for the state Delaware (DE). The second order variance approximation slightly overestimates (percentage difference being negative) the uncertainty for almost all areas but the absolute dif-

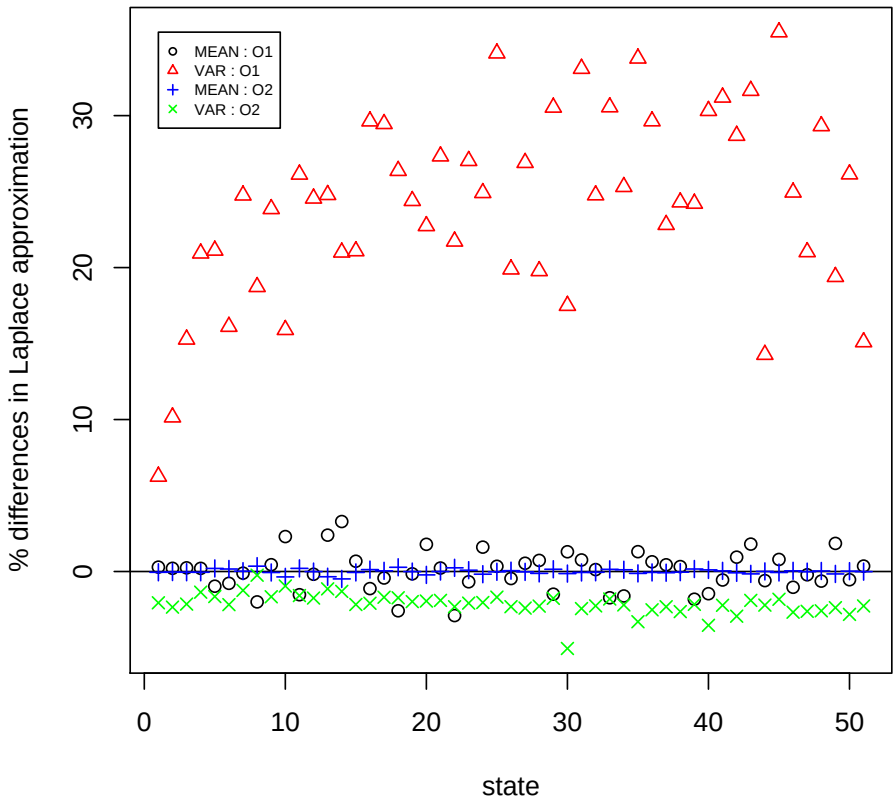

Figure 2.9: Evaluation of Laplace approximation to the posterior moments of θ_i using SAIPE 1997 data and the new prior: percent difference from the exact as a summary measure

ference is less than 4% for all areas except for the state New Mexico (NM) with 5.06%.

Figure 2.10 demonstrates the fact that if the posterior mode lies on the boundary point, the Laplace approximation may end up with misleading results. For the SAIPE 1997 data, use of uniform prior on A leads to the posterior mode to be zero. Although this does not affect the accuracy of the mean approximation much, it reduces the precision of the variance approximation to a large extent. Our proposed prior excludes the possibility of zero estimates of posterior mode. For the SAIPE

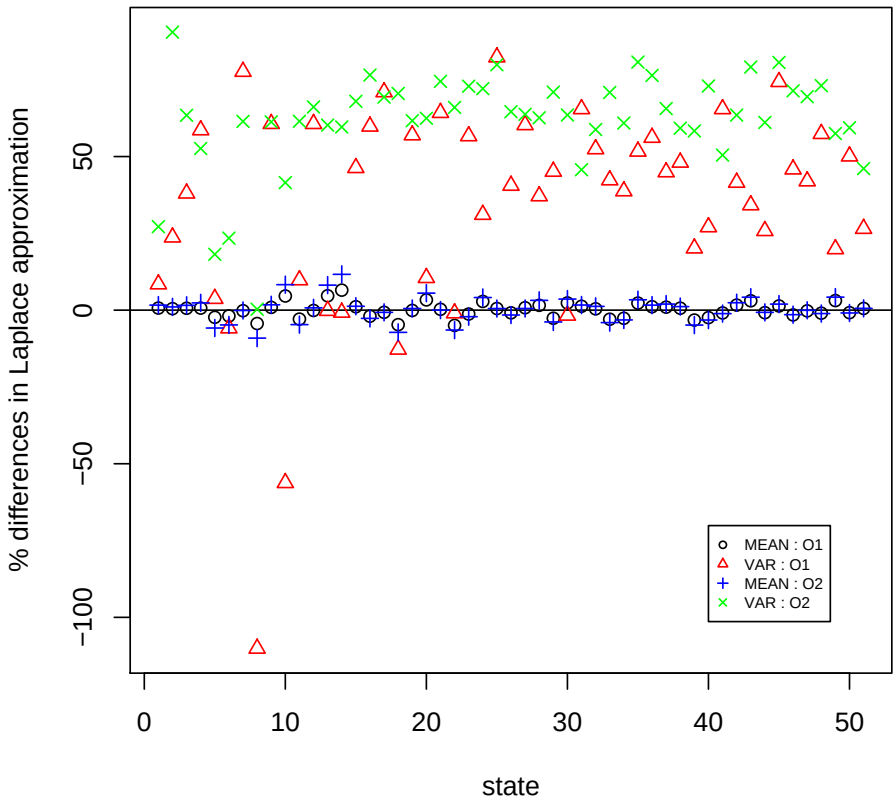

Figure 2.10: Evaluation of Laplace approximation to the posterior moments of θ_i using SAIPE 1997 data and the uniform prior: percent difference from the exact as a summary measure

1993 data, where the uniform prior on A does not lead to zero posterior mode (see Figure 2.5), the approximation works much better than what is depicted in Figure 2.10. But for 1993 also, our prior performs much better than that of uniform prior (not presented here). For example, for the second order variance approximation, the mean absolute difference is 5% for the uniform prior compared to 0.5% for our new prior.

2.6 Concluding Remarks

In this chapter, we examine the Fay-Herriot model which is extensively used in small area estimation. As an inferential procedure under the Fay-Herriot model we prefer the hierarchical Bayesian approach because of its ability to solve complex inferential problem in a straightforward way. Once we have the relevant posterior density, we can use it for any inferential problems, including the measure of uncertainty and interval estimation. Implementation of hierarchical Bayes procedure needs specification of priors on hyperparameters. We use some objective criteria to propose a prior on the variance component A. To select the prior on A we match the posterior distribution of A to the adjusted profile likelihood function of A. We recommend to use the prior $\pi(A) \propto \frac{A}{(A+d_0)^{p/2}}$ for the variance component. Besides being simple, our approach has two main advantages. It removes the possibility of yielding zero estimate for the variance component; the popular choice of uniform prior on A suffers from this drawback if posterior mode is considered as an estimator. Our prior also enjoys good small sample frequentist properties; simulation results justify that conclusion. Also, in order to have closed form expressions of the posterior mean and variance of the true small area mean, we use the Laplace approximation to ratio of integrals, following Kass & Steffey (1989). In addition to the computational simplicity, such approximations lend themselves to easy interpretations of the formulae involved.

In the context of the Small Area Income and Poverty Estimation (SAIPE) project, the REML estimator of the variance component A comes out to be zero

quite frequently. When that happens we come up with several unreasonable implications on the estimator and its measure of uncertainty. For example, the empirical Bayes estimator of the small area mean gives full weight to the synthetic estimator and none to the direct. To overcome this issue, we use the new prior on A along with the Fay-Herriot model. We implement our hierarchical Bayes procedure with second order Laplace approximation. By comparing the Laplace method to the numerical integration and MCMC method, using SAIPE data, we conclude that the second order fully exponential form of Laplace approximation works very accurately. If the posterior mode occurs at the boundary point, the approximation to the posterior variance is quite poor even for second order approximation. This suggests that the posterior mode needs to be sufficiently away from zero for the Laplace approximation to work well. We contrast our approximate hierarchical Bayes approach to the MEL empirical Bayes approach as documented in Bell (1999). Although, the point estimates of the small area means are quite close, there is an indication of underestimation in the measure of uncertainty in the Bell (1999) methodology relative to ours.

2.7 Appendix A: Verification of the propriety of the posterior distribution $f(A|y)$ of A

The proof follows the technique used by Datta & Smith (2003) and Smith (2001). In the proof, any constant term (i.e. the term does not involve A) is

denoted by C, even if they are different from step to step.

$$f(A|y) = C|W_A|^{1/2}|X'W_AX|^{-1/2}\exp\left[-\frac{1}{2}(y - X\hat{\beta}_A)'W_A(y - X\hat{\beta}_A)\right]\pi(A) \quad \text{(A.1)}$$

We need to show that,

$$\int_0^\infty f(A|y)dA < \infty \quad \text{(A.2)}$$

Let's denote $T_1 = |W_A|^{1/2}$, $T_2 = |X'W_AX|^{-1/2}$, $T_3 = \exp\left[-\frac{1}{2}(y - X\hat{\beta}_A)'W_A(y - X\hat{\beta}_A)\right]$.

Step 1: An upper bound for T_1

$|W_A| = \prod_{i=1}^m \frac{1}{A+D_i}$. Define $d = \min_i\{D_i\}(> 0)$.Thus,

$$A + D_i \geq A + d \; \forall i \quad \text{(A.3)}$$

When $d > 1$, $A + d \geq A + 1$. Thus, (A.3)$\Rightarrow$

$$A + D_i \geq A + 1 \Rightarrow T_1 \leq (A + 1)^{-m/2} \quad \text{(A.4)}$$

When $d \leq 1$, $A + d \geq Ad + d$. Thus, (A.3)$\Rightarrow$

$$A + D_i \geq C(A + 1) \Rightarrow T_1 \leq C(A + 1)^{-m/2} \quad \text{(A.5)}$$

Combining (A.4) and (A.5), we have

$$T_1 \leq C(A + 1)^{-m/2} \quad \text{(A.6)}$$

Step 2: An upper bound for T_2

Define $f = \max_i\{1, D_i\}$. Then

$$A + D_i \leq A + f = f(1 + A/f) \leq f(1 + A) \; \forall i \quad \text{(A.7)}$$

$$
\begin{aligned}
(A.7) \quad &\Rightarrow \quad \frac{1}{A+D_i} \geq \frac{1}{f(1+A)} \\
&\Rightarrow \quad W_A \geq \frac{1}{f(1+A)} I_m \\
&\Rightarrow \quad X'W_AX \geq \frac{1}{f(1+A)} \qquad (A.8)
\end{aligned}
$$

We restate the following result from Rao (1973, p. 70): Let A and B be non negative definite (nnd) matrices with $A - B$ nnd. Then $|A| \geq |B|$. Using this result we can say

$$
\begin{aligned}
|X'W_AX| &\geq C(1+A)^{-p/2}|X'X| \\
\Rightarrow T_2 &\leq C(1+A)^{p/2} \qquad (A.9)
\end{aligned}
$$

Step 3: An upper bound for T_3

$$
(y - X\hat{\beta}_A)'W_A(y - X\hat{\beta}_A) \geq 0 \Rightarrow T_3 \leq 1 \qquad (A.10)
$$

Combining (A.6), (A.9), and (A.10) in (A.1), we have,

$$
f(A|y) \leq C(1+A)^{-(m-p)/2}\pi(A) \qquad (A.11)
$$

2.7.1 Propriety of the Posterior Corresponding to $\pi(A) \propto \frac{A}{(A+d_0)^{p/2}}$

When $d_0 < 1, (A+d_0)^{p/2} \geq (Ad_0 + d_0)^{p/2} = C(1+A)^{p/2}$, for any $p > 0$. Thus,

$$
\frac{1}{(A+d_0)^{p/2}} \leq C\frac{1}{(1+A)^{p/2}} \qquad (A.12)
$$

When $d_0 \geq 1, (A+d_0)^{p/2} \geq (1+A)^{p/2}$, for any $p > 0$. Thus,

$$
\frac{1}{(A+d_0)^{p/2}} \leq \frac{1}{(1+A)^{p/2}} \qquad (A.13)
$$

Combining (A.12) and (A.13), we can say, for any d_0,

$$\frac{1}{(A+d_0)^{p/2}} \leq C\frac{1}{(1+A)^{p/2}} \tag{A.14}$$

Putting (A.14) in (A.11), we have,

$$f(A|y) \leq C\frac{A}{(A+1)^{m/2}} \tag{A.15}$$

The right hand side of (A.15) is integrable if $m/2 > 2 \Rightarrow m > 4$. Thus, the posterior $f(A|y)$ of A corresponding to the prior $\pi_(A) \propto \frac{A}{(A+d_0)^{p/2}}$ will be proper provided $m > 4$.

2.7.2 Propriety of the Posterior Corresponding to $\pi(A) \propto A$

(A.11)$\Rightarrow$

$$f(A|y) \leq C\frac{A}{(A+1)^{(m-p)/2}} \tag{A.16}$$

The righthand side of (A.16) is integrable if $(m-p)/2 > 2 \Rightarrow m > p+4$.

2.8 Appendix B: Adjusted Density Method

Suppose we want to approximate the posterior density of B_i by a beta density, say Beta(a_i, b_i). First we write the adjusted log-posterior density $l\,(B_i)$ of B_i.

$$\begin{aligned} l(B_i) &= \log\{f(B_i|y)B_i(1-B_i)\} \\ &= \log\{f(A|y)\ |J|\ B_i(1-B_i)\} \\ &= \log\left\{L_R(A)\frac{A^2}{(A+d_0)^{p/2}}\right\} \end{aligned} \tag{B.1}$$

The parameters of the beta density are obtained by matching the first two derivatives of $l(B_i)$ and log adjusted beta density. After some simplifications:

$$a_i = i_0 \hat{B}_i^2 (1 - \hat{B}_i) \tag{B.2}$$

$$b_i = i_0 \hat{B}_i (1 - \hat{B}_i)^2, \tag{B.3}$$

where $\hat{B}_i$ is obtained by equating $l'(B_i) = 0$ and i_0 is the negative second order derivative of the log posterior density of B_i, evaluated at $\hat{B}_i$, and is given by

$$i_0 = -l''(B_i)\,\big|_{B_i=\hat{B}_i} = -\left(\frac{D_i}{B_i^2}\right)^2 \frac{\partial^2 l(B_i)}{\partial A^2}\,\big|_{B_i=\hat{B}_i} - \left(\frac{2D_i}{B_i^3}\right) \frac{\partial l(B_i)}{\partial A}\,\big|_{B_i=\hat{B}_i} \tag{B.4}$$

The analytical expressions of first and second order derivative of l with respect to A, where $l = \log L_R(A)$ are given below.

$$L_R(A) = C|W_A|^{1/2}|X'W_AX|^{-1/2} \exp\left[-\frac{1}{2}(y - X\hat{\beta}_A)'W_A(y - X\hat{\beta}_A)\right] \tag{B.5}$$

The derivatives can be calculated numerically using the `numDeriv` package (Gilbert, 2008) of `R` (R Development Core Team, 2008), the expressions can be used as a check to the results produced by the `numDeriv` package. We need the following results from Searle et al. (1992, Appendix M) to calculate the derivatives. If the elements of a nonsingular matrix P are functions of a scalar t, then

1. $\frac{\partial P^{-1}}{\partial t} = -P^{-1}\frac{\partial P}{\partial t}P^{-1}$

2. $\frac{\partial \log|P|}{\partial t} = \text{trace}\left(P^{-1}\frac{\partial P}{\partial t}\right)$

Define, $e = y - X\hat{\beta}_A$. Also define $u = \frac{\partial \hat{\beta}_A}{\partial A} = -\Sigma_A X'W_A^2 e$, where $\Sigma_A = (X'W_AX)^{-1}$. Then $\frac{\partial u}{\partial A} = 2\Sigma_A X'W_A^3 e - 2MX'W_A^2 e$, where $M = \Sigma_A(X'W_A^2X)\Sigma_A$. Now, we write

the analytical expressions of the first and second order derivative as

$$\frac{\partial l}{\partial A} = -\frac{1}{2}\text{trace}(W_A) + \frac{1}{2}\text{trace}\left(\Sigma_A X' W_A^2 X\right) + \frac{1}{2}e' W_A^2 e + e' W_A X u \tag{B.6}$$

$$\frac{\partial^2 l}{\partial A^2} = \frac{1}{2}\text{trace}(W_A^2) + \frac{1}{2}\text{trace}\left(\Sigma_A X' W_A^2 X\right)^2 - \frac{1}{2}\text{trace}\left(\Sigma_A X' W_A^3 X\right) - e' W_A^3 e - 2e' W_A^2 X u - (Xu)' W_A X u + e' W_A X \frac{\partial u}{\partial A} \tag{B.7}$$

We have used $\frac{\partial W_A}{\partial A} = -W_A^2$, $\frac{\partial^2 W_A}{\partial A^2} = 2W_A^3$ to obtain the above expressions.

2.9 Appendix C: BRugs model to implement the new prior in Fay-Herriot Model

```
model
{
for (i in 1:m)
{
y[i] ~ dnorm(theta[i], tau[i])
tau[i] <- 1/d[i]
B[i] <- d[i]/(A + d[i])
theta[i] <- inprod(X[i,], beta[]) + v[i]
v[i] ~ dnorm(0, A.inv)
}
A.inv <- 1/A
beta[1] ~ dflat()
beta[2] ~ dflat()
beta[3] ~ dflat()
beta[4] ~ dflat()
beta[5] ~ dflat()
# to incorporate new prior
dummy <- 0
dummy ~ dgeneric(ll)
ll <- log(A) - (p/2) * log(A + d0)
A ~ dflat()T(0,)
}
```

2.10 Appendix D: A note on WinBUGS convergence criteria

Deviance Information Criterion (DIC; Spiegelhalter et al. 2002) and related statistics are used to assess model complexity and compare different models. It can be considered as a Bayesian measure of fit or adequacy. DIC returns the following summary statistics in BRugs (Thomas et al., 2006). BRugs is basically the R interface of OpenBUGS (Spiegelhalter et al., 2007).

Dbar: The posterior mean of the deviance, which is exactly the same if the node 'deviance' had been monitored in the samplesStats function of BRugs. This deviance is defined as -2 * log(likelihood): 'likelihood' is defined as p(y |theta), where y comprises of all data (i.e., all stochastic nodes for which values are given), and theta comprises the stochastic parents of y - 'stochastic parents' are the stochastic nodes upon which the distribution of y depends, when collapsing over all logical relationships.

Dhat: A point estimate of the deviance (-2 * log(likelihood)), obtained by substituting in the posterior means theta.bar of theta: thus, Dhat = -2 * log(p(y |theta.bar)). In other words, Dhat is the deviance at the posterior means of theta.

pD: The effective number of parameters in a model as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest i.e., pD = Dbar - Dhat. This is a measure of model complexity.

DIC: The Deviance Information Criterion is given by DIC = Dbar + pD = Dhat + 2 * pD.

Dbar (the posterior mean of the deviance) has often been used to compare models in

the literature, but such measure does not protect against the complexity of a model. `DIC` considers an additional complexity term `pD`. The models with negligible prior information, `DIC` will be approximately equivalent to the classical Akaike's criterion. The model with the smallest `DIC` is estimated to be the model that would best predict a replicate dataset of the same structure as that currently observed.

Caution: It is important to note that `DIC` assumes the posterior mean to be a good estimate of the stochastic parameters. If this is not so, say because of extreme skewness or even bimodality, then `DIC` may not be appropriate. There are also circumstances, such as with mixture models, in which `OpenBUGS` (Spiegelhalter et al., 2007) will not permit the calculation of `DIC`.

BGR (Brooks-Gelman-Rubin) convergence statistic

In `BRugs` (Thomas et al., 2006), `samplesBgr` function calculates (if plot = FALSE then returns the value only) and plots the Gelman-Rubin (1992) convergence statistic, as modified by Brooks & Gelman (1998).

The method assumes that m chains have been simulated in parallel, each with different starting points (preferably overdispersed). Having obtained suitable starting points, the chains are then run for $2n$ iterations, of which the first n are discarded to avoid the burn-in period. Given any individual sequence, the inferences about any parameter of interest are made by computing the sample mean and variance from the simulated draws. Thus, the m chains yield m possible inferences. Gelman and Rubin suggested comparing these to the inference made by mixing together the mn draws from all the sequences. For a particular parameter (node), calculate the between-sequence variance (B) and the within-sequence variance (W). An estimate

of these variance ratio (R = B/W) is called potential scale reduction factor (PSRF). If PSRF is close to 1, we can conclude that each of the m sets of n simulated observations is close to the target distribution.

Brooks & Gelman (1998) made some minor correction to the PSRF and obtained $\hat{R}_c$. $\hat{R}_c$ may ignore some information in the simulations. According to them the following three conditions should hold:

1. The mixtures-of-sequences variance should stabilize as function of n. Monitor the green line in the plot.

2. The within-sequence variance should stabilize as function of n. Monitor the blue line in the plot.

3. $\hat{R}_c$ should approach 1 i.e. the red line should be very close to 1 on the right-hand side of the plot.

Monitoring $\hat{R}_c$ alone considers only the third of these conditions. Brooks and Gelman emphasize that one should also be concerned with individual stability of within (blue) and pooled (green) variance estimates.

Chapter 3

The Prior Selection and Approximations for the Nested Error Regression Model: Estimation of Finite Population Mean for Small Areas

3.1 Introduction

In a typical sample survey, the goal is to draw inferences about various characteristics of a finite population, such as the mean or the total for a study variable y. The traditional approach towards this inferential problem is design based (Kish, 1965). The random variables according to this approach, are the sampling indicators (whether a population unit is included in the sample), and their probability distribution generates the sampling weights. The values of y are considered to be fixed in the design based approach. The sampling weights along with the observed values of y are used to obtain an estimate of the finite population quantity and the associated measure of uncertainty. The alternative view is the model based approach (Ericson, 1969; Royall, 1971; Valliant et al., 2000), which views the finite population as a sample from a hypothetical superpopulation, that is characterized by a model for y. According to this approach, y is random. Model based methods are useful in the context of small area estimation because this borrows strength from neighboring areas and other related sources. Small area estimation is well suited to

settings that involve several areas or strata with a small number of sample observations available from each individual stratum. Both the design and model based approaches can be frequentist, where such procedures do not make an explicit use of priors for the finite population or the superpopulation parameters (Ghosh, 2008). In a Bayesian paradigm, prior distributions are used for the hyperparameters involved in the model which describes the superpopulation. The traditional survey sampling approach may use prior information, in the form of auxiliary variables, but that is quite different from the Bayesian paradigm.

The mixed models are often used in small area estimation because of their flexibility in combining information from different sources and taking different sources of errors into account. These models may be classified into two broad classes, area level and unit level models, based on the availability of data for the variable to be modeled. A good application of a unit level model can be found in a paper by Battese et al. (1988). They used a nested error regression model to estimate the mean area under different crops (corn and soybeans) for twelve counties (small areas) of north central Iowa. To allow for the correlation structures (in which reported crop hectares for geographically closer segments have stronger correlations than those farther apart), they included a random county effect in the model. The nested error regression models in Battese et al. (1988) and Prasad & Rao (1990), are also termed random intercept models. This can be viewed as a particular type of multilevel model that allows small area slopes as well as the intercept to be random and lead to improved small area estimates with the potential to use area level covariates (Moura & Holt, 1999). The multilevel models proposed by Moura & Holt (1999)

are also known as random regression coefficient models in the literature (Dempster et al., 1981).

The objective of this chapter is to examine the hierarchical Bayesian (HB) approach, under a nested error regression model for estimating finite population means for small areas. We propose a new prior for the variance component and apply Laplace methods to approximate the various posterior moments involved in the hierarchical Bayesian analysis. The chapter is organized as follows. In Section 3.2, we present a review of classical and Bayesian approaches using unit level models for small area estimation. In Section 3.3, we describe the HB procedure in detail, justifying the choice of priors on the variance component. We propose a prior distribution for the variance component that results in an estimator for the shrinkage factor and small area means that have good frequentist properties. In this section, we also give an outline of how to use the Laplace approximation to approximate various posterior moments. This is followed by a section on simulation study, Section 3.4, to justify the choice of priors in estimating small area means. In Section 3.5, we implement our methodology in predicting areas under corn and soybeans for 12 counties in north central Iowa. This problem was originally explored by Battese et al. (1988) in a classical context and revisited by Datta & Ghosh (1991) from a Bayesian point of view with a diffuse prior for the variance components.

3.2 Estimation of Finite Population Means using Unit level Models: A Review

Ericson (1969) was the first to put forward a subjective Bayesian approach for estimating finite population means, using a unit level model. Following Ericson (1969), there is a series of papers that use unit level models in the context of finite population sampling for small area estimation. Ghosh & Meeden (1986); Ghosh & Lahiri (1987); Battese et al. (1988); Prasad & Rao (1990); Nandram & Sedransk (1993); Arora et al. (1997); Datta & Lahiri (2000), among others, examined empirical Bayes and empirical best linear unbiased prediction (EBLUP) approaches to small area estimation, using unit level models. Ghosh & Lahiri (1989) proposed a hierarchical Bayes procedure as an alternative to the EBLUP and the empirical Bayes approaches, in order to obtain a measure of uncertainty of the small area estimator in a straightforward way. The model considered by Ghosh & Lahiri (1989) was a special case of the nested error regression model considered by Battese et al. (1988). Datta & Ghosh (1991) discussed a unified hierarchical Bayes inference approach using a general linear mixed model to estimate the finite population means for small areas. They used inverse gamma distribution as the prior for the variance components in the general linear mixed model. To implement their hierarchical Bayes method, in the absence of closed form expressions for the posterior mean and variance, Datta & Ghosh (1991) relied on numerical integration and MCMC methods.

In the hierarchical Bayesian approach, it is important to check for the propriety

of the posterior distributions involved, in case improper priors are used for the hyperparameters. Smith (2001) presented a very useful sufficient condition for the propriety of the posterior distribution when Jeffreys' prior is used for the variance components in a nested error regression model. They calculated the Jeffreys' prior (square root of the determinant of the Fisher information matrix) from the residual likelihood of the variance component as well as from the full likelihood. In the later case, they obtained the Jeffrey's prior for the variance component and the regression coefficient from the joint likelihood of regression coefficient and variance component. Although, Datta & Smith (2003) recommended to use the Jeffreys' prior based on the residual likelihood to reduce the effects of the nuisance parameter on the inference of the parameter of interest.

In this chapter, we examine the estimation of the finite populations means $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij},\ i = 1, \ldots, m$, using the following nested error regression model:

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij};\ j = 1, \ldots, N_i,\ i = 1, \ldots, m, \tag{3.1}$$

where y_{ij} is the value of the study variable for the jth unit belonging to the ith area, x_{ij} is the unit level $p \times 1$ vector of known covariates, β is a $p \times 1$ vector of unknown regression coefficients, v_i is the random small area effect, and e_{ij} is the error term. We assume $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$. Our objective is to draw inferences about the finite population mean $\bar{Y}_i$ for the ith small area. Battese et al. (1988) applied the above model with $p = 3$. In their application, y_{ij} is the number of hectares of corn (or soybeans) in the jth segment of the ith county, as obtained from a survey; x_{1ij} and x_{2ij} are the number of pixels classified as corn and soybeans,

respectively, in the jth segment of the ith county, as obtained from satellite data.

Suppose a sample s_i of size n_i is drawn from the N_i units of the ith area following some specified sampling design; $i = 1, \ldots, m$ and $\sum_{i=1}^{m} n_i = n$. Following Battese et al. (1988) and Rao (2003, p 78), we assume that the sample values also follow the model (3.1), that is, we do not address the situation involving informative sampling. The assumption that the sample values also obey the model (3.1) holds true for simple random sampling from each area. According to Rao (2003), this assumption is satisfied more generally for sampling designs that use the auxiliary information x_{ij} in selecting the sample, including stratified simple random sampling, probability proportional to size designs. For a proof of this absence of selection bias, see Rao (2003, p 79). If there is clustering within the small area, we need to consider more complex models involving more variance components to incorporate the intra-cluster correlation. This dissertation will not discuss such complex models. Scott & Smith (1969), Malec & Sedransk (1985), Ghosh (2008), and others presented Bayesian analysis for multistage cluster sampling.

Battese et al. (1988) expressed the finite population mean in terms of model parameters and considered the prediction of a mixed effect term (combination of both fixed and random effect), i.e., they approximated $\bar{Y}_i$ as

$$\bar{Y}_i \approx \bar{X}_i{}'\beta + v_i, \tag{3.2}$$

where $\bar{X}_i$ is the vector of known population means for the auxiliary variables. This follows from the assumption that the sum of random errors over the population units is negligible (see Battese et al., 1988). They first find the BLUP of the right side

of (3.2), assuming the variance components to be known and then plugged in the classical estimates of σ_e^2 and σ_v^2. They consider Henderson's method III (Henderson, 1953) to estimate the variance components. Note that in their method, $\hat{\sigma_v^2}$ can be negative, in which case they truncate it zero. Datta & Ghosh (1991), using the Battese et al. (1988) data set, applied diffuse priors for the variance components with suitably chosen parameters in the inverse gamma distribution. In other words, they used the inverse gamma $IG(a, b)$ priors for both σ_e^2 and $\lambda = \sigma_v^2/\sigma_e^2$, where a is the shape parameter and b is the scale parameter. But they chose $a = 0$ and $b = 0.005$ to reflect lack of prior information. The winBUGS developers (Spiegelhalter et al., 1997, 2002) also recommended use of inverse gamma, with small parameter values $[IG(0.001, 0.001)]$, prior for variance components. In the context of simple two-level normal model (as in (3.1) but without covariates), Gelman (2006) noted that for datasets in which low values of σ_v^2 are possible, inferences under inverse gamma prior are very sensitive to the choice of small values for its parameters. Our simulation results (Section 3.4) also supports Gelman's claim.

We can also write the small area mean $\bar{Y}_i$ as

$$\begin{aligned} \bar{Y}_i &= \frac{1}{N_i}\left\{\sum_{j\in s_i} y_{ij} + \sum_{j\notin s_i} y_{ij}\right\} \\ &= \frac{1}{N_i}\left\{n_i\bar{y}_{is} + (N_i - n_i)\,\bar{y}_{ins}\right\} \\ &= f_i\bar{y}_{is} + (1 - f_i)\,\bar{y}_{ins}, \end{aligned} \tag{3.3}$$

where $f_i = n_i/N_i$ is the sampling rate, $1 - f_i$ is the finite population correction, $\bar{y}_{is}$ and $\bar{y}_{ins}$ are the means of the sampled and nonsampled units, respectively, for the ith area. From (3.3), we can say that the prediction of $\bar{Y}_i$ is equivalent to the

prediction of $\bar{y}_{ins}$ given the sample. The right hand side of (3.2) can be considered as a good approximation of the right hand side of (3.3) if $f_i \approx 0$.

In this chapter, we apply a hierarchical Bayesian approach to estimate the small area means (3.3) for finite populations. However, as an intermediate step, we need to estimate the parameters involved in the prediction model. We propose a new prior distribution for the variance component ratio λ (a standard uniform prior will be considered for the regression coefficients and σ_e^2). This prior is chosen so that the resulting posterior mode of λ is always positive and at the same time the new prior leads to estimators of the shrinkage factors and small area means having good frequentist properties. Since the posterior mode of λ, using the new prior, always lies in the interior of the parameter space, we can easily apply the Laplace approximation to obtain various posterior moments.

3.3 Hierarchical Bayes Estimation of Finite Population Means

A hierarchical Bayesian version of the nested error regression model (3.1) is given by

- Level 1: $y_{ij}|\beta, v_i, \sigma_e^2 \overset{ind}{\sim} N(x'_{ij}\beta + v_i, \sigma_e^2);\ j = 1, \ldots, N_i,\ i = 1, \ldots, m$
- Level 2: $v_i|\sigma_v^2 \overset{iid}{\sim} N(0, \sigma_v^2)$

Our aim is to find the Bayes estimator of (3.3) and its measure of uncertainty, which are given by $E(\bar{Y}_i|\mathbf{y})$ and $V(\bar{Y}_i|\mathbf{y})$ respectively, where $\mathbf{y}$ is the vector of sampled values of y. Whether we are in Bayesian or frequentist paradigm depends on whether we assume a prior on the hyperparameters $(\beta, \sigma_v^2, \sigma_e^2)$ at the third level. The two

levels of the Bayesian model given above can be combined into a single linear mixed model (3.1). That is why the empirical Bayes approach is considered as a frequentist approach by many researchers (Jiang & Lahiri 2006b, p 4-5; Datta & Ghosh 1991, p 1748), it does not assume any prior distribution for the hyperparameters. Instead, the hyperparameters are estimated using some classical method. To obtain the Bayesian summary statistics for $\bar{Y}_i$, we proceed as follows (assuming that the sample values follow the same model above, as discussed in Section 3.2):

First we write the likelihood using level 1 and level 2 of the hierarchical model as:

$$L(\mathbf{y}|\beta, \mathbf{v}, \sigma_e^2, \sigma_v^2) \propto (\sigma_e^2)^{-\frac{n}{2}} (\sigma_v^2)^{-\frac{m}{2}} \exp\left[-\frac{1}{2}\left\{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - x'_{ij}\beta - v_i)^2}{\sigma_e^2} + \frac{\sum_{i=1}^{m} v_i^2}{\sigma_v^2}\right\}\right] \tag{3.4}$$

Now, following Ghosh & Lahiri (1989), and Datta & Ghosh (1991), we specify prior on σ_e^2 and λ at level 3 as:

- Level 3: $\pi(\beta, \sigma_e^2, \lambda) \propto \pi(\sigma_e^2)\pi(\lambda)$,

where $\lambda = \sigma_v^2/\sigma_e^2$. This completes our hierarchical Bayesian model. Combining the likelihood (3.4) and the prior at level 3, we write the joint distribution of $\mathbf{y},\ \beta,\ \mathbf{v} = (v_1, \ldots, v_m),\ \sigma_e^2,\ \lambda$ as

$$\begin{aligned} f(\mathbf{y}, \beta, \mathbf{v}, \sigma_e^2, \lambda) \quad &\propto \quad (\sigma_e^2)^{-\frac{n}{2}} \lambda^{-\frac{m}{2}} (\sigma_e^2)^{-\frac{m}{2}} \\ &\exp\left[-\frac{1}{2}\left\{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - x'_{ij}\beta - v_i)^2}{\sigma_e^2} + \frac{\sum_{i=1}^{m} v_i^2}{\lambda\sigma_e^2}\right\}\right] \pi(\lambda) \end{aligned} \tag{3.5}$$

In (3.5), we consider a flat prior (uniform prior) on σ_e^2 i.e., we consider $\pi(\sigma_e^2) \propto 1,\ \sigma_e^2 \in [0, \infty]$. Typically, enough data will be available to estimate σ_e^2 precisely.

Moreover, the classical ANOVA estimate of σ_e^2 is always positive. As a result, we do not explore the choice of priors on σ_e^2. Any reasonable noninformative prior usually works well for σ_e^2 (Gelman, 2006).

After collecting terms involving v_i from the exponent of (3.5), we can say that

$$v_i|\mathbf{y}, \beta, \sigma_e^2, \lambda \overset{ind}{\sim} N\left[(1-B_i)(\bar{y}_{is} - \bar{x}'_{is}\beta),\ \frac{\sigma_e^2}{n_i}(1-B_i)\right], \tag{3.6}$$

where $B_i = 1/(1+n_i\lambda)$. Integrating out v_i from (3.5), using matrix notation, we write the joint distribution (without the normalizing constant) of $\mathbf{y},\ \beta,\ \sigma_e^2,\ \lambda$ as

$$\begin{aligned}
& f(\mathbf{y}, \beta, \sigma_e^2, \lambda) \\
\propto\ & \prod_{i=1}^{m} \left\{(1+n_i\lambda)^{-1/2}\right\} \left(\sigma_e^2\right)^{-n/2} \pi(\lambda) \\
& \exp\left[-\frac{1}{2\sigma_e^2}\sum_{i=1}^{m}\left\{y_i'y_i + \beta' X_i'X_i\beta - 2y_i'X_i\beta - \frac{\lambda}{1+\lambda n_i}\left(y_i'J_iy_i + \beta'X_i'J_iX_i\beta - 2\beta'X_i'J_iy_i\right)\right\}\right] \\
=\ & \prod_{i=1}^{m} \left\{(1+n_i\lambda)^{-1/2}\right\} \left(\sigma_e^2\right)^{-n/2} \pi(\lambda) \\
& \exp\left[-\frac{1}{2\sigma_e^2}\left\{(\beta - \hat{\beta}_\lambda)'\Sigma_\lambda^{-1}(\beta - \hat{\beta}_\lambda) + T_\lambda\right\}\right],
\end{aligned} \tag{3.7}$$

where $T_\lambda = \sum_{i=1}^m y_i'\Sigma_i y_i - \left(\sum_{i=1}^m X_i'\Sigma_i y_i\right)' \left(\sum_{i=1}^m X_i'\Sigma_i X_i\right)^{-1} \left(\sum_{i=1}^m X_i'\Sigma_i y_i\right)$. From the expression (3.7), it follows that

$$\beta|\mathbf{y}, \sigma_e^2, \lambda \sim N_p\left[\hat{\beta}_\lambda,\ \sigma_e^2\Sigma_\lambda\right], \tag{3.8}$$

where $\hat{\beta}_\lambda = \left(\sum_{i=1}^m X_i'\Sigma_i X_i\right)^{-1}\left(\sum_{i=1}^m X_i'\Sigma_i y_i\right)$, $\Sigma_\lambda = \left(\sum_{i=1}^m X_i'\Sigma_i X_i\right)^{-1}$. The area specific terms are defined as follows: $\Sigma_i = I_i - \frac{\lambda}{1+\lambda n_i}J_i = (I_i + \lambda J_i)^{-1}$, I_i is the identity matrix of order n_i, J_i is a $n_i \times n_i$ matrix with all the elements equal to 1, X_i is the $n_i \times p$ matrix of covariates, y_i is the vector (of length n_i) of observations for the ith area. The subscript λ in both $\hat{\beta}_\lambda$ and Σ_λ indicates the dependence of the

terms on the ratio of variance component $\lambda = \sigma_v^2/\sigma_e^2$ alone. But one should note that $V(\beta|\mathbf{y}, \sigma_e^2, \lambda) = \sigma_e^2\Sigma_\lambda$ depends on both λ and σ_e^2. One should also note that Σ_i and Σ_λ are different notations.

Integrating out β from (3.7), we find

$$f(\mathbf{y}, \sigma_e^2, \lambda) \;\propto\; \prod_{i=1}^{m} \left\{(1+n_i\lambda)^{-1/2}\right\} \left(\sigma_e^2\right)^{-(n/2-p/2)} |\Sigma_\lambda|^{1/2} \exp\left[-\frac{1}{2\sigma_e^2}T_\lambda\right] \pi(\lambda) \tag{3.9}$$

From the equation (3.9), it follows that

$$\sigma_e^2|\mathbf{y}, \lambda \;\sim\; IG\left(\frac{n-p-2}{2},\; \frac{1}{2}T_\lambda\right), \tag{3.10}$$

where $IG(a, b)$ is the inverse gamma distribution with shape parameter a and scale parameter b. A random variable Z has an inverse Gamma distribution $IG(a, b)$, if its pdf is given by $f(z) \propto z^{-(a+1)} \exp[-b/z];\; z > 0$. Using the property of inverse gamma distribution, we express the conditional posterior mean and variance of σ_e^2 given λ as

$$E(\sigma_e^2|\mathbf{y}, \lambda) = \frac{T_\lambda}{n-p-4} \tag{3.11}$$

$$V(\sigma_e^2|\mathbf{y}, \lambda) = \frac{2T_\lambda^2}{(n-p-4)^2(n-p-6)} \tag{3.12}$$

Integrating out σ_e^2 from (3.9), we can write the posterior distribution of a single variance component λ as

$$f(\lambda|\mathbf{y}) \propto \prod_{i=1}^{m} \left\{(1+n_i\lambda)^{-1/2}\right\} |\Sigma_\lambda|^{1/2} T_\lambda^{-(n-p-2)/2} \pi(\lambda) \tag{3.13}$$

To present the results in a unified way, we restate (3.13) using matrix notation as

$$f(\lambda|\mathbf{y}) \propto |\Sigma|^{1/2} \; |X'\Sigma X|^{-1/2} (Y'Q_\lambda Y)^{-(n-p-2)/2} \; \pi(\lambda), \tag{3.14}$$

where $\Sigma = \bigoplus_{i=1}^{m} \Sigma_i$, $X = col_{1\le i\le m} col_{1\le j\le n_i} x'_{ij}$, $Y = col_{1\le i\le m} col_{1\le j\le n_i} y_{ij}$, $T_\lambda = Y'Q_\lambda Y$, $Q_\lambda = \Sigma - \Sigma X(X'\Sigma X)^{-1}X'\Sigma$. This formulation is presented for the ease of writing codes in simulation and data analysis.

3.3.1 Posterior moments of finite population mean when λ known

We derive $E(\bar{Y}_i|\mathbf{y})$ and $V(\bar{Y}_i|\mathbf{y})$ in two steps. First, we find analytical expressions for these quantities when λ is known. Then using the Laplace approximation, we obtain the final results. Here we consider the definition of $\bar{Y}_i$ as given in (3.3). Recall that $f_i = n_i/N_i$ is the sampling rate, and $1 - f_i = fpc_i (< 1)$ is the finite population correction for the ith small area. Using the iterative expectation and variance technique and the results given in (3.6), (3.8), and (3.11), it is not very difficult to get

$$\begin{aligned} & E(\bar{Y}_i|\mathbf{y},\lambda) \\ = \; & (1 - fpc_i B_i)\bar{y}_{is} + fpc_i(\bar{X}'_{ins} - (1-B_i)\bar{x}'_{is})\hat{\beta}_\lambda \\ = \; & g_i(\lambda), \text{ say} \qquad (3.15) \\ & V(\bar{Y}_i|\mathbf{y},\lambda) \\ = \; & \frac{T_\lambda}{n-p-4}\left[\frac{fpc_i}{N_i} + \frac{(fpc_i)^2(1-B_i)}{n_i} + (A_i - D_i)'\Sigma_\lambda(A_i - D_i)\right] \\ = \; & h_i(\lambda), \text{ say}, \qquad (3.16) \end{aligned}$$

where $A_i = fpc_i\bar{X}_{ins}$, $D_i = fpc_i B_i \bar{x}_{is}$, $A_i - D_i = fpc_i\{\bar{X}_{ins} - (1-B_i)\bar{x}_{is}\}$. $\hat{\beta}_\lambda$ and Σ_λ is defined in (3.8). Note that Σ_λ is different from Σ in (3.14). Now let's consider some special cases.

Common mean model, Unbalanced case

Suppose there are no covariates and hence we have a common mean μ, say, at the second level and the first level mean is v_i according to the Bayesian model described at the beginning of Section 3.3. If we consider the definition of $\bar{Y}_i$ as given in equation (3.3), then modifying (3.15) we find $E(\bar{Y}_i|\mathbf{y},\lambda) = (1-fpc_iB_i)\bar{y}_{is}+B_ifpc_i\hat{\mu}$, where $\hat{\mu} = \sum_{i=1}^{m} B_in_i\bar{y}_{is}/\sum_{i=1}^{m} n_iB_i$. Now consider the approximate definition of the finite population mean as given in Battese et al. i.e., $\bar{Y}_i \approx \mu + v_i$, then $E(\bar{Y}_i|\mathbf{y},\lambda) = (1-B_i)\bar{y}_{is} + B_i\hat{\mu}$. A simple comparison between the above two expectations reveals that, when λ is known, the Bayes estimate of finite population mean $\bar{Y}_i$ assigns more weight to the direct estimate $\bar{y}_{is}$ than the Bayes estimate of $\bar{Y}_i \approx \mu + v_i$ does, since $fpc_i < 1$. This fact was also observed by Ghosh (2008) in the context of a simple exchangeable model.

Common mean model, Balanced case, and $f_i \approx 0$

Here, besides the common mean assumption, we assume that $n_i = k\ \forall i, n = mk$ and hence $B_i = B = 1/(1+k\lambda)$. Then the posterior distribution of λ is given by:

$$f(\lambda|\mathbf{y}) \propto \frac{(1+k\lambda)^{-(m-1)/2}}{(SSW + B.SSB)^{(n-3)/2}}\,\pi(\lambda),$$

where $SSW = \sum_i\sum_j(y_{ij}-\bar{y}_{is})^2$ and $SSB = k\sum_i(\bar{y}_{is}-\hat{\mu})^2$; $\hat{\mu} = \frac{1}{m}\sum_i\bar{y}_{is}$ are the usual definition of within sum of square and between sum of square used in the balanced one-way ANOVA. After some modification and simplification of (3.15) and (3.16), we find $E(\bar{Y}_i|\mathbf{y},\lambda) = (1-B)\bar{y}_{is}+B\hat{\mu}$ and $V(\bar{Y}_i|\mathbf{y},\lambda) = (\lambda B+\frac{B}{n})\left\{\frac{SSW+B.SSB}{n-5}\right\}$.

3.3.2 Choice of Prior on λ

To obtain the posterior mean and variance of $\bar{Y}_i$ from (3.15) and (3.16), we need to perform one-dimensional integral with respect to the posterior distribution of λ. For that we need to assume a prior distribution on λ. The uniform prior on λ is noninformative and yields a posterior distribution of λ for which the mode is identical to the residual maximum likelihood (REML) estimator of λ, following the arguments discussed in Chapter 2. The posterior mode of λ with uniform prior or equivalently, the REML estimator of λ can be zero for a particular application. In many practical applications, the maximum likelihood (ML) or restricted maximum likelihood (REML) estimates of hyperparameters occur at the boundary point. For example, in a two-level poisson-gamma model, the ML estimate of the variance component can occur at infinity (Christiansen & Morris, 1997). For a basic area level normal-normal model the ML or REML estimate of the variance component can be zero (Bell 1999, Li & Lahiri 2008, Chapter 2 of this dissertation). In order to prevent such boundary solutions in the context of discrete data analysis, the Latent GOLD Choice software uses Dirichlet priors for the latent and response probabilities (Vermunt & Magidson, 2005). When the hyperparameter estimates occur at the boundary points, it presents unreasonable implications for the small area estimators and their measure of uncertainty, e.g., an estimator of $\bar{Y}_i$ would put all the weights to the synthetic estimator and none to the direct area specific estimator [see (3.15)]. This is not desirable for an area with lots of sample.

Following Li & Lahiri (2008), we match the posterior distribution of λ given

in (3.13) to the adjusted profile likelihood function of λ to obtain an appropriate prior distribution of λ. This prior leads to a posterior distribution of λ for which the mode is always positive and results in estimators of the shrinkage factor and small area mean that have good frequentist properties. This follows from the theory on ADM (see Li & Lahiri 2008, and also Chapter 2, for further details). By profile likelihood, we mean the likelihood of λ that does not account for the loss of degrees of freedom due to the estimation of regression coefficient β. This is given by

$$L_p(\lambda) \propto \prod_{i=1}^{m} \left\{(1+n_i\lambda)^{-1/2}\right\}\ T_\lambda^{-(n-2)/2} \tag{3.17}$$

Note that, to obtain the profile likelihood (3.17), we consider β as the only nuisance parameter. We plug in the ML estimate of β in (3.7) [without the term $\pi(\lambda)$] to obtain the joint profile likelihood of λ and σ_e^2. Then integrating out σ_e^2, we derive the marginal profile likelihood of λ. For more on profile likelihood, see Cox & Reid (1992, 1993). This leads to the prior

$$\pi(\lambda) \propto \lambda \left|\sum_{i=1}^{m}(X_i'\Sigma_i X_i)\right|^{1/2} T_\lambda^{-p/2};\ \lambda > 0. \tag{3.18}$$

To simplify the prior for some special cases, we proceed as follows:

$$\begin{aligned}
\Sigma_i &= I_i - \frac{\lambda}{1+\lambda n_i} J_i \\
&= (I_i - \bar{J}_i) + \bar{J}_i - \frac{\lambda}{1+\lambda n_i} 1_i 1_i' \\
&= C_i + 1_i \left(\frac{1}{n_i} - \frac{\lambda}{1+\lambda n_i}\right) 1_i' \\
&= C_i + \frac{1}{1+\lambda n_i}\bar{J}_i,
\end{aligned}$$

where $\bar{J}_i = \frac{1}{n_i} J_i,\ C_i = I_i - \bar{J}_i$. See Searle et al., 1992, Appendix M, for properties of I and J matrices. Thus, we can write $X_i'\Sigma_i X_i = X_i' C_i X_i + \frac{1}{1+\lambda n_i} X_i' \bar{J}_i X_i$. For the

common mean model, balanced case (i.e. $p = 1,\ n_i = k\ \forall\ i$), $X_i'C_iX_i = 0$, $\frac{1}{1+\lambda n_i}X_i'\bar{J}_iX_i = \frac{k}{1+\lambda k}$, and $T_\lambda = SSW + \frac{SSB}{1+k\lambda}$, where SSW and SSB is defined in page 91. Hence, leaving the proportionality constant aside we can say $\pi(\lambda) \propto \frac{\lambda}{\sqrt{SSB+SSW(1+\lambda k)}}$. For the **common mean model, unbalanced case**: $X_i'C_iX_i = 0$, $\frac{1}{1+\lambda n_i}X_i'\bar{J}_iX_i = \frac{n_i}{1+\lambda n_i}$, and $T_\lambda = (SSW+SSB)^{-1/2}$, where SSW and SSB is defined in page 96. Hence, in this case the prior is given by $\pi(\lambda) \propto \lambda\ \sqrt{\sum_{i=1}^{m}\frac{n_i}{1+\lambda n_i}}(SSW + SSB)^{-1/2}$.

3.3.3 Propriety of Posterior

The prior $\pi(\sigma_e^2, \lambda) \propto \pi(\lambda)$ with $\pi(\lambda)$ given in equation (3.18), leads to a proper joint posterior distribution $f(\sigma_e^2, \lambda|\mathbf{y})$, if

(i) $n > m + Rank(H)$, where $H = R_{Z_1}X$, $R_{Z_1} = I_n - Z_1(Z_1'Z_1)^{-1}Z_1'$, and $Z_1 = \bigoplus_{i=1}^{m}(col_{1\leq j\leq n_i}1)$, a block diagonal matrix of 1_i. Also recall that m is the number of small areas, $n = \sum_{i=1}^{m} n_i$, total sample size.

(ii) $m > 4$

(iii) $n > p + 2$, where p is the number of covariates in the model.

For an outline of the proof see the Appendix of this chapter.

3.3.4 Laplace Approximation

The posterior moments of $\bar{Y}_i$, using the prior in (3.18), are not in a closed-form, but can be obtained either by numerical integration or by the Monte Carlo Markov

Chain (MCMC) method. The slow computation speed of the MCMC method (besides examining the convergence) does not permit its evaluation by repeated use in simulation. Thus, for convenient implementation and evaluation of our hierarchical Bayes method, we approximate the posterior moments of $\bar{Y}_i$ using Laplace's method. Laplace approximation method has been applied by many authors in the context of Bayesian analysis (Tierney & Kadane 1986; Tierney et al. 1989; Kass & Steffey 1989; Butar & Lahiri 2002; Datta et al. 2005). To obtain the asymptotic variance-covariance matrix of the parameters of the model, the Latent GOLD Choice software (Vermunt & Magidson, 2005) uses the negative inverse of the Hessian matrix, where the Hessian matrix is obtained by computing the second order derivative of the log-posterior density instead of the log-likelihood of the parameters without any prior assumption on the model parameters. This approach can be viewed as a first order Laplace approximation to the posterior variance of the model parameters.

The first order Laplace approximation of the posterior moments of $\bar{Y}_i$ is given by

$$E(\bar{Y}_i|\mathbf{y}) = g_i(\hat{\lambda}) + O(1/m) \tag{3.19}$$

$$V(\bar{Y}_i|\mathbf{y}) = h_i(\hat{\lambda}) + \left\{g_i'(\hat{\lambda})\right\}^2 \frac{1}{i_0} + O(1/m^2) \tag{3.20}$$

where $\hat{\lambda}$ is the posterior mode obtained by maximizing the posterior distribution of λ with $\pi(\lambda)$ given by (3.18). Our prior always leads to a positive estimate of the mode. $g_i(\lambda)$ and $h_i(\lambda)$ are defined in (3.15) and (3.16) respectively. $g_i'(\lambda)$ is the first order derivative of $g_i(\lambda)$ with respect to λ. The $\hat{\lambda}$ inside the parenthesis indicates the the function is evaluated at $\hat{\lambda}$. The quantity i_0 is the negative second derivative

of log-posterior of λ evaluated at $\hat{\lambda}$. i_0 should be bounded away from zero. It should be noted that that both the approximations have relative error of order $(1/m)$. For details on Laplace approximation, see Tierney et al. (1989); Kass & Steffey (1989). Below we derive the analytical expressions of the posterior moments of $\bar{Y}_i$ using equation (3.19) and (3.20) under some special cases.

Common mean model, Balanced case, and $f_i \approx 0$

$$E(\bar{Y}_i|\mathbf{y}) = (1-\hat{B})\bar{y}_{is} + \hat{B}\hat{\mu} + O(1/m)$$

$$V(\bar{Y}_i|\mathbf{y}) = \left(\hat{\lambda}\hat{B} + \frac{\hat{B}}{n}\right)\left\{\frac{SSW + \hat{B}.SSB}{n-5}\right\} + (\bar{y}_{is}-\hat{\mu})^2 V(B|\mathbf{y}) + O(1/m^2),$$

where $V(B|\mathbf{y}) = \left\{\frac{\hat{B}(1-\hat{B})}{\hat{\lambda}}\right\}^2 1/i_0$, $\hat{B}$ is B evaluated at $\hat{\lambda}$.

Common mean model, Unbalanced case, and $f_i \approx 0$

$$E(\bar{Y}_i|\mathbf{y}) = (1-\hat{B}_i)\bar{y}_{is} + \hat{B}_i\hat{\hat{\mu}} + O(1/m),$$

where $\hat{\hat{\mu}}$ is $\hat{\mu}$ evaluated at $\hat{B}_i$, $\hat{B}_i$ is B_i evaluated at $\hat{\lambda}$, $\hat{\mu} = \sum_{i=1}^m B_i n_i \bar{y}_{is} / \sum_{i=1}^m n_i B_i$. Unlike the balanced case, here $\hat{\mu}$ is a function of λ.

$$V(\bar{Y}_i|\mathbf{y}) = \left(\frac{1-\hat{B}_i}{n_i} + \frac{\hat{B}_i^2}{\sum n_i \hat{B}_i}\right)\left\{\frac{SSW + \widehat{SSB}}{n-5}\right\} + \left(\bar{y}_{is} - \hat{\hat{\mu}}\right)^2 V(B_i|\mathbf{y}) + O(1/m^2),$$

where $SSW = \sum_i \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{is})^2$ and $SSB = \sum_i n_i B_i (\bar{y}_{is} - \hat{\mu})^2$. Unlike the balanced case, SSB is a function of λ in this situation. $\widehat{SSB}$ is SSB evaluated at $\hat{\lambda}$. $V(B_i|\mathbf{y}) = \left\{\frac{\hat{B}_i(1-\hat{B}_i)}{\hat{\lambda}}\right\}^2 1/i_0$.

To obtain the second order Laplace approximation to the posterior mean and variance of $\bar{Y}_i$, we apply the fully exponential form suggested by Tierney & Kadane (1986). This has the advantage of requiring only the first two derivatives of the

corresponding posterior distribution to achieve a second order approximation, unlike the standard form which needs fourth and fifth derivatives of the log-likelihood (Kass et al., 1988, p 265). The second order approximation to the posterior mean is given by

$$E(\bar{Y}_i|\mathbf{y}) = \frac{i_0}{i_0^*}\exp\left[m\left\{L^*(\hat{\lambda}^*) - L(\hat{\lambda})\right\}\right], \tag{3.21}$$

where $L = \frac{1}{m}\log f(\lambda|\mathbf{y}), L^* = \frac{1}{m}\{\log f(\lambda|\mathbf{y}) + \log g_i(\lambda)\}$, and $\hat{\lambda}^*$ is the maximizer of L^*. i_0^* is the negative second derivative of L^* evaluated at $\hat{\lambda}^*$. The second order approximation of posterior variance is given by

$$V(\bar{Y}_i|\mathbf{y}) = T_1 + T_2, \tag{3.22}$$

where T_1 = right hand side (RHS) of (3.21) with L^* as $\frac{1}{m}\{\log f(\lambda|\mathbf{y}) + \log h_i(\lambda)\}$. T_2 = right hand side of (3.21) with L^* as $\frac{1}{m}\{\log f(\lambda|\mathbf{y}) + \log g_i^2(\lambda)\} - \{$ RHS of (3.21)$\}^2$. To apply the fully exponential form, we need $g_i(\lambda)$ and $h_i(\lambda)$ to be positive functions of λ. Both of the above second order approximations are of relative order $O(1/m^2)$.

3.4 Simulation Study

This section compares the performance of the approximate hierarchical Bayesian (HB) approach using our new prior on λ with several other existing choices. The HB approach is approximate as it is implemented through Laplace approximation. Moreover, we consider one classical approach in estimating small area means that uses the REML method to estimate the variance component λ. Here we study the frequentist properties of the resulting Bayes procedure by a Monte Carlo simulation study under the assumption of a balanced set-up, common mean (μ) model

with $f_i \approx 0$. The frequentist properties of the estimators are derived from the distribution of the estimates obtained from repeated simulations of observations following the statistical model assuming the hyperparameters to be known. For our simulation experiment, we set $m = 10, \mu = 1, k = 6$, i.e., $n = 60$. We explore several choices of λ, by varying the values of σ_e^2 and σ_v^2. We generate $N = 5,000$ replicates $\{(y_{ij}, \theta_i = \mu + v_i), j = 1, \ldots, k;\ i = 1, \ldots, m\}$ using the simplified nested error regression model to study various frequentist properties of our approximate hierarchical Bayesian methods.

Table 3.1: Percentage of trials that posterior mode of $\lambda = 0$

$\lambda = \sigma_v^2/\sigma_e^2$	REML	NEW	WIN	DG	JEFF
$1/5 = 0.2$	11.42	0	0	0	19.46
$1/2 = 0.5$	1.92	0	0	0	3.82
$1/1 = 1$	0.38	0	0	0	0.70
$5/5 = 1$	0.14	0	0	0	0.40
$2/1 = 2$	0.04	0	0	0	0.06

Among the methods considered to estimate λ in the simulation, the REML estimate turns out to be zero in some cases, which is highly undesirable. This phenomenon results in a full shrinkage for the estimator of small area mean. Also the posterior mode of λ, obtained using the Jeffrey's prior (JEFF) on λ and σ_e^2, is zero in some simulation runs. We consider the Jeffrey's prior as suggested by Datta & Smith (2003): $\pi(\lambda, \sigma_e^2) \propto \{\sigma_e^2(1 + k\lambda)\}^{-1}$. But the percentage of time a zero mode is obtained decreases with the increase of true value of λ. The posterior mode of λ resulting from the other three priors, viz. NEW: $\pi(\lambda) \propto \frac{\lambda}{\sqrt{SSB + SSW(1+\lambda k)}}$ (with uniform prior on σ_e^2), WinBUGS default priors (WIN): $IG(0.001, 0.001)$ on

both σ_e^2 and λ, Datta & Ghosh (1991) prior (DG): $IG(0, 0.005)$ on both σ_e^2 and λ, never yields zero estimates of the variance component. The REML estimate can also be viewed as the posterior mode of λ when a uniform prior is used for both σ_e^2 and λ. Note that, this particular choice of priors lead to a proper posterior when $Rank\,(Z_1' C_n Z_1) > 2$ and $n > 5$; $Z_1 = \bigoplus_{i=1}^{m} (col_{1 \le j \le k} 1)$, a block diagonal matrix of 1_k and $C_n = I_n - \bar{J}_n$. This typically holds true in practical applications (but does not hold for $m = 3$, Rao 2003). The above results follow from the necessary and sufficient conditions given by Hobert & Casella (1996) for the propriety of the joint posterior for a general class of priors of which the uniform is a particular case.

To compare the performance of different priors for the variance components, in estimating the shrinkage factor and the small area mean, we present the results in the form of average in the tables below. Averaging across 10 small areas makes sense since the same model holds for all 10 areas in the simulations. The new prior performs very well in estimating the shrinkage factor B (see Table 3.2), both in terms of bias and MSE (compared to all the other estimators), irrespective of the value of λ. This supports Morris and Li-Lahiri since the Laplace first order approximation to the posterior mean of B is essentially the plugged-in estimator of B, in an empirical Bayes set-up. The performance of the new prior is even better than the REML method although REML does a better job in estimating λ (Table not shown here). This is not counter intuitive as it follows from Jenson's inequality, B being a convex function of λ. Even an unbiased estimator of λ would yield positive bias, if we plug in the estimate of λ to estimate B. In estimating B, Jeffrey's prior (JEFF) does much better than that of the inverse gamma priors (WIN and DG) on λ and σ_e^2.

Table 3.2: Bias (and MSE) in percent of different estimators of the shrinkage factor B for different choices of λ

$\lambda = \sigma_v^2/\sigma_e^2$	REML	NEW	WIN	DG	JEFF
$1/5 = 0.2$	10.16(7.24)	-7.79(2.47)	38.02(21.54)	35.25(19.07)	19.18(9.99)
$1/2 = 0.5$	7.96(4.28)	-0.71(1.30)	27.60(18.23)	25.95(16.28)	14.68(6.77)
$1/1 = 1$	4.78(1.76)	0.69(0.69)	13.50(7.08)	12.97(6.41)	8.93(2.91)
$5/5 = 1$	4.35(1.53)	0.40(0.62)	12.79(6.56)	12.28(5.91)	8.46(2.66)
$2/1 = 2$	2.64(0.55)	0.76(0.27)	6.07(1.87)	5.94(1.70)	4.94(0.96)

Table 3.3: MSE of different estimators of the small area mean for different choices of λ

$\lambda = \sigma_v^2/\sigma_e^2$	REML	NEW	WIN	DG	JEFF
$1/5 = 0.2$	57.12	54.80	78.04	74.36	59.66
$1/2 = 0.5$	28.44	27.27	38.56	37.03	29.87
$1/1 = 1$	15.45	15.11	17.75	17.46	15.97
$5/5 = 1$	75.09	73.60	85.87	84.52	77.62
$2/1 = 2$	15.97	15.82	16.72	16.64	16.27

As expected the priors WIN ($IG(0.001, 0.001)$) and DG ($IG(0, 0.001)$) perform in the same way in estimating λ, B, and small area mean θ_i. For small values of λ, the performance of these two priors is very poor in estimating B and θ_i. The MSE of θ_i is higher for these priors relative to the new prior, see Table 3.3. Gelman (2006) also argued that for datasets in which low values of λ are possible, inferences under inverse-gamma prior are very sensitive to the choice of small values for its parameters.

The new prior performs reasonably well in estimating θ_i, irrespective of the choice of λ. The MSE of the estimator NEW of θ_i is close to REML and JEFF but always less, and much less than that of WIN and DG estimator (see Table 3.3). The

Table 3.4: Coverage properties of different estimators of the small area mean for different choices of λ

	Coverage probability of 95% CI				Average length of CI			
$\lambda = \sigma_v^2/\sigma_e^2$	REML	NEW	WIN	DG	REML	NEW	WIN	DG
$1/5 = 0.2$	0.86	0.95	0.62	0.67	2.50	3.03	1.80	1.90
$1/1 = 1$	0.94	0.95	0.91	0.91	1.50	1.54	1.49	1.50
$5/5 = 1$	0.94	0.95	0.91	0.91	3.36	3.44	3.34	3.36
$2/1 = 2$	0.95	0.95	0.95	0.95	1.57	1.58	1.60	1.61

differences tend to be smaller as λ increases. It is mentionable that the simulation results are invariant of σ_e^2, it depends only on λ. Any change in the values of σ_e^2 reflects the scale of the data. For example, when $\lambda = 1$ and $\sigma_e^2 = 5$, the MSE of θ_i for all the five methods is very close to five times the MSE of θ_i when $\lambda = 1$ and $\sigma_e^2 = 1$, the differences can probably be explained by the Monte Carlo simulation error.

We also examined the coverage properties (Table 3.4) of the 95% confidence interval for θ_i, comparing the four methods: REML, NEW, WIN, DG. For REML, we consider the usual $EB \pm 1.96\sqrt{MSE}$ type interval. For the hierarchical Bayesian approaches, we first obtain the posterior variance of θ_i using the Laplace first order approximation and then assume normality to obtain the intervals. For the Jeffrey's prior, it was not possible to obtain positive posterior variance in all simulation runs. In many cases, the posterior mode of λ under the Jeffrey's prior turned out to be negative, in which case we truncate it to zero. For those cases, the negative Hessian at zero is not positive, as zero is not the posterior mode. A positive value of the negative Hessian is required to guarantee a positive posterior variance of θ_i, see the

equation (3.20). That is why we decided to drop the Jeffrey's prior while obtaining the confidence intervals for small area means. This in turn establishes the superiority of our prior which always produces positive posterior mode, and hence makes the implementation of Laplace approximation straightforward in a Bayesian context.

The 95% confidence interval for θ_i resulting from the new prior, in all cases, has a true coverage rate more than or equal to 0.95. This result shows that the Laplace approximation accurately approximates the posterior variance of θ_i taking into account all sources of uncertainties. On the contrary, the Laplace approximation does not work well for the other two priors (WIN and DG). The coverage rates for these two priors are very poor for small values of λ, but seems to improve as λ increases. When $\lambda = 0.2$, the true coverage probability corresponding to the REML method is less (0.86) than the nominal value (0.95). This is most likely due to the higher percentage of zero estimate of the posterior mode of λ, leading to the underestimation of the MSE.

3.5 Data Analysis

In this section, we carry out an empirical application. Battese et al. (1988) used a nested error linear regression model to predict area under corn and soybeans for 12 counties in north central Iowa using data from the 1978 June Enumerative Survey as well as LANDSAT satellite data. Each county was divided into area segments, and data were collected on a sample of segments by interviewing farm operators. The number of sample segments, n_i, in a county ranged from 1 to 5, with

a total sample size of 36, while the population count of segments N_i in the counties ranged from 394 to 687. This dataset, thus, provides an opportunity to apply small area estimation techniques in an unbalanced set-up. Because of the negligible sampling fractions, Battese et al. estimated the small area means $\bar{Y}_i \approx \bar{X}_i'\beta + v_i$, we also follow that assumption in this data analysis. Unit level auxiliary data, in the form of the number of pixels classified as corn and soybeans, are available for all the segments in the population from the LANDSAT satellite readings. So, this dataset allows us to use three unit level covariates, as opposed to applying a simple common mean model. The dataset can be found in Table 1 of Battese et al. (1988). One sample segment in Hardin county was deleted from the prediction procedure because the corn area for that segment looked erroneous to Battese et al. in a preliminary analysis. They used model checking criteria to validate the assumed nested error regression model (3.1) with $p = 3$.

We use their model in a hierarchical Bayesian approach to estimate area under corn for each of 12 counties, whereas they used a frequentist approach (EBLUP) to analyze the data. They obtained classical ANOVA estimates of the variance components ($\hat{\sigma}_v^2 = 140, \hat{\sigma}_e^2 = 150$) for the corn data. According to our notation, this yields a classical estimate for $\lambda = 0.93$. When the value of λ is less than 1, our simulation study suggests that there can be substantial improvements in the estimation of small area means under the hierarchical Bayesian model using our new prior. In this section, we compare the exact posterior moments obtained after performing an one-dimensional numerical integration using our new prior to the approximate HB approach using Laplace approximation. Recall that, to obtain

the posterior moments of $\bar{Y}_i$ from $g_i(\lambda)$ and $h_i(\lambda)$, we need to perform an one-dimensional integration. Datta & Ghosh (1991) and Rao (2003) also considered the HB approaches to analyze this dataset, but they used different priors for the variance components (the inverse-gamma prior, with a choice of small parameter values).

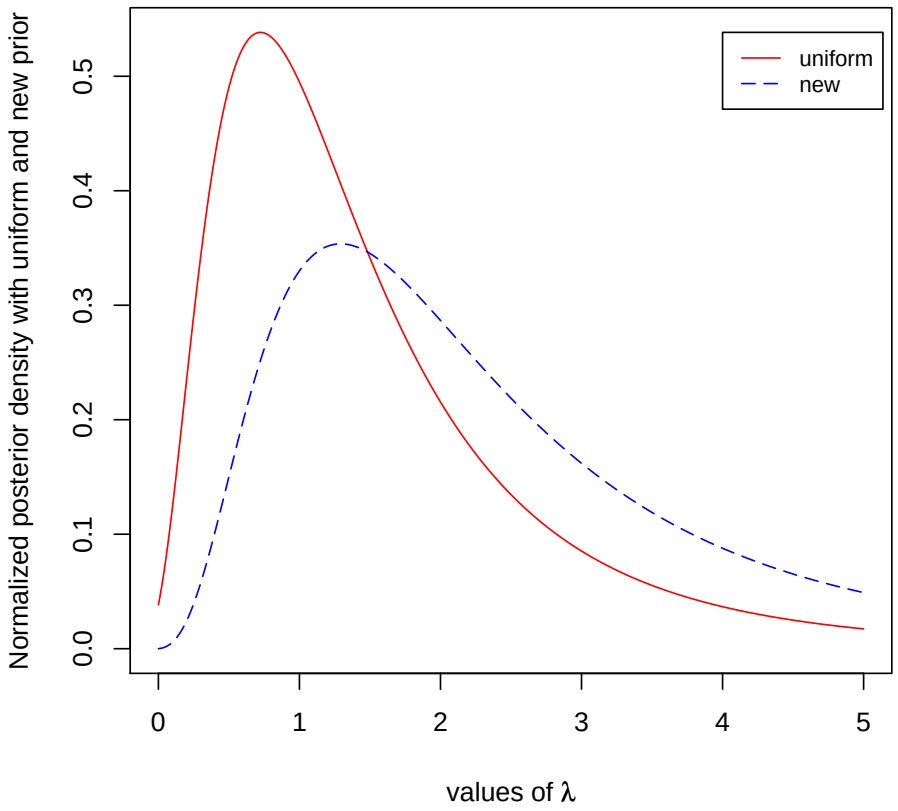

Figure 3.1: Plot of the posterior density of λ using Battese et al. (1988) data: a comparison between uniform prior and proposed prior for λ under nested error regression model

In Figure 3.1, we draw two different posterior densities of λ using two different priors on λ: uniform prior and $\pi(\lambda)$ as defined in (3.18). From the plot we can see that use of the new prior pushes the posterior mode (1.29) to the right, compared

to that of uniform prior (0.725).

Table 3.5: Different point estimates of mean hectares of corn

County	n_i	Naïve	BHF	EXACT	L:o1	L:o2
1	1	122.2	122.2	121.2	121.7	121.2
2	1	126.2	126.5	127.4	126.8	127.4
3	1	106.8	105.5	102.8	104.9	102.8
4	2	108.5	107.6	105.9	107.1	105.9
5	3	144.1	145.3	145.8	145.1	145.9
6	3	112.1	112.9	113.6	112.9	113.6
7	3	112.8	112.1	111.4	112.0	111.4
8	3	122.0	122.1	121.8	121.9	121.8
9	4	115.3	116.1	116.5	116.0	116.5
10	5	124.4	124.2	124.5	124.5	124.5
11	5	106.9	106.1	105.5	106.1	105.6
12	5	143.0	143.5	144.4	143.8	144.4

In Table 3.5, we present several point estimates of the mean hectares under corn for the 12 counties. The Naïve and BHF estimators are taken from Battese et al. (1988) paper. Both these estimators are variants of EBLUP, having the form $\hat{\bar{Y}}_i \equiv \bar{X}_i'\hat{\beta} + (\bar{y}_{is} - \bar{x}_{is}'\hat{\beta})\hat{g}_i$, where $\hat{\beta}$ is an estimate of β as defined in (3.8) with plugged-in estimates of variance component, obtained by some classical method. Setting $\hat{g}_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_i}$ produces the Naïve estimator. An alternative estimator of g_i (defined in their appendix) leads to the BHF estimator. These two estimators are derived from a frequentist point of view. The last three estimators are derived under the Bayesian paradigm using our new prior. The EXACT estimator is the posterior mean of $\bar{Y}_i$ obtained using numerical integration technique. L:o1 and L:o2 are respectively the Laplace first and second order approximation to the posterior mean of $\bar{Y}_i$ as discussed in Section 3.3.4. The second column of the table presents

the sample size in each county. It is quite evident that the point estimates of the small area means are more or less similar.

Table 3.6 displays the measure of uncertainty associated with the estimates in Table 3.5. The Naïve MSE estimator does not take into account the uncertainty in estimating the variance components, but the frequentist MSE of the BHF estimator incorporates all sources of uncertainty in estimating the small area mean (see Battese et al., 1988, appendix), as do all three Bayesian measures. The MSE estimates corresponding to the Naïve estimator are consistently smaller than all other estimators across the counties. In general, the measures of uncertainties associated with the estimators of Table 3.5 decrease with increasing sample sizes. Comparing the last two columns of Table 3.6, with the EXACT column, we can emphatically say that the Laplace approximation to the posterior variance of $\bar{Y}_i$ is performing very well. For a formal evaluation of Laplace method, see Figure 3.2.

Table 3.6: Measure of uncertainty of the estimators of mean hectares of corn

County	n_i	Naïve	BHF	EXACT	L:o1	L:o2
1	1	9.1	10.3	10.5	10.1	10.5
2	1	9.0	10.2	10.2	9.9	10.3
3	1	8.8	10.0	10.5	10.5	10.5
4	2	7.6	8.5	8.5	8.6	8.5
5	3	6.2	6.8	6.7	6.8	6.6
6	3	6.2	6.8	6.6	6.8	6.6
7	3	6.2	6.8	6.6	6.8	6.6
8	3	6.3	6.9	6.6	6.7	6.6
9	4	5.5	6.0	5.8	5.9	5.8
10	5	5.1	5.5	5.2	5.4	5.3
11	5	5.0	5.4	5.3	5.5	5.3
12	5	5.4	5.8	5.7	5.8	5.7

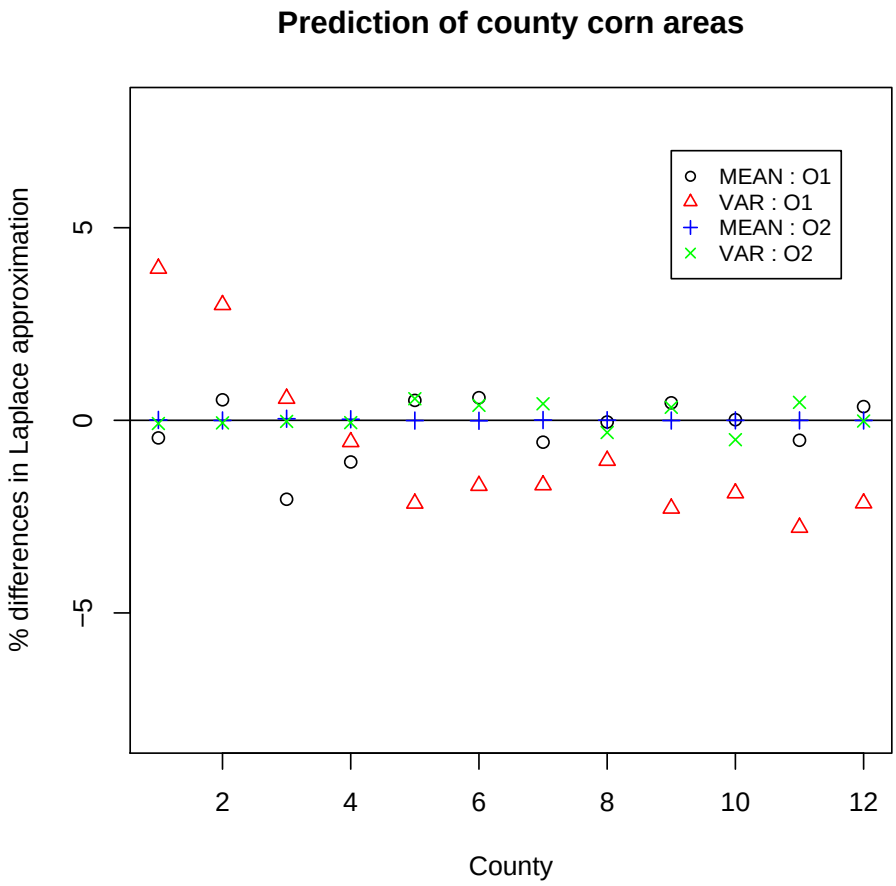

Figure 3.2: Evaluation of Laplace approximation using new prior for λ: percent difference from EXACT (numerical integration) as a summary measure

To measure the precision of Laplace method, we compute the percentage difference as the summary statistics. Mathematically, this can be defined as {(exact-approximate)/exact}×100. We treat the numerical integration results as the exact posterior moments. Figure 3.2 shows that both the first and second order approximation values of the mean are quite close to the exact value, although the second order values are more accurate as the percentage differences lie on the zero line for almost all counties. As far as the measure of uncertainty is concerned, the first order Laplace approximation overestimates the variance (the percentage difference being

negative), excepts for the first three counties. The second order Laplace approximation performs very well in estimating the true posterior variance, the percentage differences are close to the zero line for all counties.

3.6 Concluding Remarks

In this chapter we applied unit level linear mixed model to estimate the finite population means for small areas. As an inferential procedure we used Bayesian approach that needs specification of prior for the hyperparameters. Following some objective criteria, we obtain a prior distribution for the ratio (λ) of variance components, along with a standard flat prior on the regression coefficient. To approximate the posterior moments of small area means, we apply Laplace method. Our choice of prior avoids the extreme skewness, usually present in the posterior distribution of variance components. This property leads to more accurate Laplace approximation. Our simulation study shows that the resulting approximate Bayes estimators (with new prior) of small area means have good frequentist properties such as MSE and coverage rate.

3.7 Appendix

The propriety condition given in Section 3.3.3 follows from Smith (2001), and Datta & Smith (2003). An outline of the proof is given below. For details see the above mentioned papers. The joint posterior distribution of σ_e^2 and λ, obtained by

using the prior (3.18) in (3.9), and matrix notation, is given by

$$f(\sigma_e^2, \lambda|\mathbf{y}) \propto \lambda\ |\Sigma|^{1/2}\ \left(\sigma_e^2\right)^{-(n/2-p/2)}\ \exp\left[-\frac{T_\lambda}{2\sigma_e^2}\right] T_\lambda^{-p/2} \tag{A.1}$$

Step 1: An upper bound for $|\Sigma|^{1/2}$

$$\begin{aligned} & 1+\lambda n_i\ \geq 1+\lambda n^*\ \geq 1+\lambda,\ n^* = \min\{n_i\} \\ \Rightarrow\ & \frac{1}{1+\lambda n_i} \leq \frac{1}{1+\lambda} \\ \Rightarrow\ & \prod_{i=1}^{m} \frac{1}{1+\lambda n_i} \leq \prod_{i=1}^{m} \frac{1}{1+\lambda} \\ \Rightarrow\ & |\Sigma|^{1/2} \leq (1+\lambda)^{-m/2} \end{aligned} \tag{A.2}$$

Step 2: Find $\lim_{\lambda\to\infty} T_\lambda$

It is possible to show that T_λ is a nonincreasing function of λ (For a proof see Smith 2001). Thus, $T_\lambda \geq \lim_{\lambda\to\infty} T_\lambda$. Now,

$$\lim_{\lambda\to\infty} T_\lambda = s = Y'[R_{Z_1} - H(H'H)^{-1}H']Y, \tag{A.3}$$

where $R_{Z_1} = I_n - Z_1(Z_1'Z_1)^{-1}Z_1'$, $H = R_{Z_1}X$. Using (A.2) and (A.3) in (A.1), we can write

$$f(\sigma_e^2, \lambda|\mathbf{y}) \leq \lambda(1+\lambda)^{-m/2}\ \left(\sigma_e^2\right)^{-(n/2-p/2)}\ \exp\left[-\frac{s}{2\sigma_e^2}\right]\ s^{-p/2} \tag{A.4}$$

We need to show that the right hand side of (A.4) is integrable. If $n > m+Rank(H)$, then $s > 0$. For s away from 0, $s^{-p/2}$ is bounded for any p. Now, given $s > 0$, we need to find conditions which satisfy

$$\int_0^\infty \frac{\lambda}{(1+\lambda)^{m/2}} d\lambda\ \int_0^\infty \frac{\exp\left[-\frac{s}{2\sigma_e^2}\right]}{(\sigma_e^2)^{\frac{n-p-2}{2}+1}} d\sigma_e^2 < \infty \tag{A.5}$$

The first integral is integrable if $m/2 > 2 \Rightarrow m > 4$. The second one is integrable if $\frac{n-p-2}{2} > 0 \Rightarrow n > p+2$.

Chapter 4

Hierarchical Bayes Estimation of Binary Data for Small Areas

4.1 Introduction

Surveys are usually designed to produce reliable estimates of various characteristics of interest for large geographic areas. For example, the National Health Interview Survey (NHIS) is designed to produce precise estimates of finite population parameters related to health issues (e.g., the probability of at least one visit to a doctor within the past 12 months) for the entire United States but not for small geographical areas. However, for effective planning of health, social and other services and for apportioning government funds, there is a growing demand to produce similar estimates for small geographic areas and subpopulations. Usually, reliable direct estimates cannot be obtained for small areas using national survey data due to the small sample sizes in small areas. In the absence of adequate direct information for small areas, it is customary to borrow strength from related sources to form indirect estimators that increase the effective sample size and hence reduce the sampling errors of the estimators. Such indirect estimators are usually based on implicit or explicit models which combine information from the sample survey, various administrative/census records and previous surveys. For various small area estimation methods along with its application in different fields, we refer readers to Ghosh & Rao (1994); Rao (2003), and the long review paper by Jiang & Lahiri

(2006b).

Most survey data are binary or categorical in nature, hence the problem of estimating rates and proportions has received considerable attention in the recent past. Dempster & Tomberlin (1980) proposed an empirical Bayes method for estimating census undercount for local areas using mixed logistic regression models. Stasny (1991) considered an empirical Bayes analysis for the proportion of individuals having a particular characteristic and the probabilities of response (response rates) within subgroups of the population with application to the National Crime Survey (NCS). Stroud (1991) developed a general hierarchical Bayes methodology for univariate natural exponential families with quadratic variance functions (NEFQVF), which includes binomial distribution for binary data. Stroud (1994) provided a comprehensive treatment of binary survey data, encompassing simple random, stratified, cluster and two-stage sampling, as well as two-stage sampling within strata. He & Sun (1998) developed hierarchical Bayesian estimators for hunting success rates per trip for the counties of Missouri. They extended their methodology to include spatial correlations among neighboring subareas (He & Sun, 2000). Malec et al. (1997) provided estimates of population proportion for small geographical areas using the National Health Interview Survey (NHIS) data. To include all sources of variation in the model, they carried out a hierarchical Bayesian analysis. Farrell et al. (1997) developed an empirical Bayes methods based on a second order Taylor series expansion to obtain model based predictions that requires only local-area summary statistics for both continuous and categorical auxiliary variables. Jiang & Lahiri (2001) provided a frequentist's alternative to the hierarchical Bayes methods for

small area estimation with binary data. Specifically, they obtained the best predictor (BP) and empirical best predictor (EBP) of the small area specific random effect using a mixed logistic model and studied different asymptotic properties of the proposed BP and EBP. For a comprehensive review of empirical Bayes and hierarchical Bayes methods in analyzing binary data for small areas, readers are referred to Rao (2003).

In this chapter, we develop a new hierarchical Bayesian methodology to analyze binary data for small areas. We implement the Bayesian procedure using the Laplace approximation. In order for the Laplace approximation work well, we need the posterior mode of the hyperparameters to fall inside the range of the parameter space. We choose a prior for one of the hyperparameters very carefully so that it avoids the extreme skewness, if any, present in the posterior distribution of the hyperparameters. The prior is chosen following the guidelines given by Morris (2006); Li & Lahiri (2008).

4.2 Model and Methodology

Let y_{ij} be the value of the jth unit belonging to the ith area. y_{ij} can take on the value 1 or 0 depending on whether the individual possesses the relevant characteristic or not. The particular characteristic may be visiting a physician at least once during the past 12 months, hunting a turkey successfully in one trip to hunting, whether an individual selected in the sample for a particular survey responds or not, whether an individual belongs to a particular age $\times$ race $\times$ sex group

or not, whether an individuals' household income is below the poverty threshold or not, etc. Also let θ_i be the true proportion of having a particular characteristic in the population (small area). We use the following hierarchical model to analyze the data $\{y_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, m\}$.

Notation: Throughout this paper we use square brackets to express any distribution in terms of its mean and variance, whereas round parentheses denotes standard parameter representation.

- Level 1: $y_{ij}|\theta_i \overset{ind}{\sim} \text{Bernoulli}[\theta_i, \theta_i(1-\theta_i)]$
- Level 2: $\theta_i|\gamma, \beta \overset{ind}{\sim} \text{Beta}[\mu_i, \gamma\mu_i(1-\mu_i)]$,

where $\text{logit}(\mu_i) = x_i'\beta$, x_i is the vector of area-level covariates and $\beta,\ -\infty < \beta < \infty$ is a $p \times 1$ vector of unknown regression coefficients. In the context of a sample survey, suppose we want to obtain the response probability for a particular age $\times$ race $\times$ sex group; then the particular age, race and sex can be considered as area-level covariates in the analysis. $0 < \gamma < 1$ is the term needed to capture the skewness and kurtosis of the Beta distribution. Our aim is to find a precise estimate of the true proportion for the small areas with a reliable measure of uncertainty which takes into account all sources of error using a hierarchical Bayesian approach. In other words, we aim to find the posterior mean and variance of $\theta_i,\ i = 1, \ldots, m$. In a subjective Bayesian analysis, the hyperparameters β and γ are considered to be known. In the absence of subjective input of the user, it is a common practice to use some noninformative prior. Researchers often prefer noninformative priors for the hyperparameters to let the data to dominate the posterior distribution and this choice also makes the

Bayesian analysis comparable to the frequentist analysis (Albert, 1988).

The above hierarchical model closely follows the model proposed by Stasny (1991); Stroud (1994); He & Sun (1998), with respect to the first two levels (we come to the third level of prior for the hyperparameters later). He & Sun (1998) used this model to estimate the hunting success rates at the county level using data from Missouri Turkey Hunting Survey (MTHS) 1994 spring season. Their hierarchical Bayes model has the advantage of borrowing strength from all counties to estimate the success rates for individual counties. But their model, along with the model considered by Stasny and Stroud, does not include covariates which might have had an influence on hunting success rate. For their application, the success rate may depend on some set of covariates, such as regional productivity, hunting regulations, forest cover, distance from a large metropolitan area, and population size (He & Sun, 1998). Our model permits including covariates in the analysis.

4.2.1 Hyperparameters Known

We reduce the data $\mathbf{y_i} = (y_{i1}, \ldots, y_{in_i})$ for the ith area to the sample total $y_i = \sum_{j=1}^{n_i} y_{ij};\ i = 1, \ldots, m$, noting that y_i is minimal sufficient statistic for the level 1 model. In terms of the usual shape and scale parameter representation of the Beta density, we can restate the level 2 model, together with level 1 model as follows:

- Level 1: $y_i|\theta_i \overset{ind}{\sim} \text{Binomial}(n_i, \theta_i)$
- Level 2: $\theta_i|\gamma, \beta \overset{ind}{\sim} \text{Beta}(a_i, b_i);\ a_i > 0, b_i > 0,$

where $a_i = \mu_i \frac{1-\gamma}{\gamma}$ and $b_i = (1-\mu_i)\frac{1-\gamma}{\gamma}$. Based on the present hierarchical model, the joint distribution of data $\mathbf{y} = (\mathbf{y_1}, \ldots, \mathbf{y_m})$, $\theta = (\theta_1, \ldots, \theta_m)$, β, γ is given by

$$\begin{aligned} f(\mathbf{y}, \theta, \beta, \gamma) &= \prod_{i=1}^{m} \{f(y_i|\theta_i).f(\theta_i|\gamma, \beta)\} \\ &\propto \prod_{i=1}^{m} \left\{\theta_i^{y_i}(1-\theta_i)^{n_i - y_i}\right\} \left\{\theta_i^{a_i - 1}(1-\theta_i)^{b_i - 1}\right\} \\ &= \prod_{i=1}^{m} \theta_i^{y_i + a_i - 1}(1-\theta_i)^{n_i - y_i + b_i - 1} \end{aligned} \tag{4.1}$$

From the joint density (4.1), we can say that the true proportion θ_i, given β, γ, and the data follows Beta distribution independently, $i = 1, \ldots, m$ or equivalently,

$$\theta_i|\beta, \gamma, \mathbf{y} \overset{ind}{\sim} \text{Beta}(y_i + a_i, n_i - y_i + b_i) \tag{4.2}$$

When the hyperparameters are known, the conditional posterior mean and variance of θ_i are given by

$$E(\theta_i|\mathbf{y}, \beta, \tau) = g_i(\beta, \tau) = B_i\mu_i + (1 - B_i)\bar{y}_i \tag{4.3}$$

and

$$V(\theta_i|\mathbf{y}, \beta, \tau) = h_i(\beta, \tau) = \frac{\tau}{1 + n_i\tau + \tau}\, g_i(\beta, \tau)\left\{1 - g_i(\beta, \tau)\right\}, \tag{4.4}$$

where $\bar{y}_i = \frac{1}{n_i}\sum_j y_{ij}$ is the sample proportion (the direct estimate of θ_i), and $B_i = \frac{1}{1+n_i\tau}$ is the shrinkage factor. The parameter τ, $0 < \tau < \infty$ is a transformation of the parameter γ, defined as $\tau = \frac{\gamma}{1-\gamma}$. The shrinkage parameter B_i depends only on the shape parameter τ, unlike the situation in Christiansen & Morris (1997) in the context of a poisson-gamma model, where B_i depended on both the location parameter and the squared coefficient of variation.

The conditional posterior mean (in other words, the best predictor) of $\theta_i,\ i = 1, \ldots, m$, which is linear in the data, can be viewed as a weighted average of the prior mean μ_i and the sample mean $\bar{y}_i$, the weight being the shrinkage factor. Note that, B_i is a decreasing function of the sample size n_i for the ith area. Hence, for fixed τ, the best predictor of θ_i gives more weight to the direct estimate for areas with large n_i compare to the areas having small sample size. This makes (4.3) an intuitively appealing estimator of θ_i. For small values of τ, i.e. when $\tau \to 0$, we have full shrinkage towards the second level mean or, equivalently the best predictor puts no weight to the direct estimate which is highly undesirable for areas having relatively large sample size.

Researchers (Malec et al., 1997; Jiang & Lahiri, 2001; Farrell et al., 1997; He & Sun, 2000) often use a logit model in level 2 with a random effect term in the linear predictor along with the fixed effect term $x_i'\beta$. But our beta prior permits a simple expression of the conditional posterior mean of θ_i, unlike considering logit model at the second level, and in the form of a shrinkage estimator, heavily preferred in the small area literature. This holds because the beta distribution is conjugate to the binomial likelihood. Also the Bayes estimator obtained in this way has the minimax property under squared error loss (for details, see Morris 1983a). In the context of count data analysis, conjugate hierarchical Poisson-Gamma models have been considered by Natarajan et al. (1998), Christiansen & Morris (1997), among many others. Unlike Natarajan et al. (1998), the modeling assumptions of Christiansen & Morris broadened the range of applications by permitting a regression specification at level 2. Our paper extends the idea of Christiansen & Morris (1997) in the context

of analyzing binary data.

4.2.2 Hyperparameters are not Known

In practice, the hyperparameters are not known. In an empirical Bayesian approach, the hyperparameters β and γ or equivalently β and τ are estimated from the marginal distribution of data and then plugged in to (4.3) and (4.4) to obtain an estimator of θ_i along with a measure of uncertainty. This näive approach ignores the uncertainty in estimating the hyperparameters. Integrating (4.1), with respect to $\theta_i,\ i = 1, \ldots, m$, we can say that $y_i|\beta, \tau \overset{ind}{\sim}$ Beta-binomial with probability mass function

$$f(y_i|\beta, \tau) \propto \frac{\Gamma(y_i + a_i)\Gamma(n_i - y_i + b_i)}{\Gamma(a_i + b_i + n_i)} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)},\ y_i = 0, \ldots, n_i. \tag{4.5}$$

The use of Beta-binomial distribution in estimating proportions can be found in literature (see Kleinman, 1973). In practical experience, it is possible to find greater variation in binomial proportion data than would be expected under the binomial distribution. In such situations the Beta-binomial distribution is used. The mean and variance of beta-binomial distribution is given by

$$E(y_i|\beta, \tau) = n_i\mu_i$$

$$V(y_i|\beta, \tau) = n_i\mu_i(1 - \mu_i) + \frac{n_i(n_i - 1)\mu_i(1 - \mu_i)}{1 + \frac{1}{\tau}}.$$

When $\tau \to 0$, the beta-binomial distribution (4.5) is approximately the binomial distribution. From (4.5), we can write the log likelihood of β and τ as

$$l(\beta, \tau) \propto \sum_{i=1}^{m} \left[\sum_{h=0}^{y_i - 1} \log(\mu_i + h\tau) + \sum_{h=0}^{n_i - y_i - 1} \log(1 - \mu_i + h\tau) - \sum_{h=0}^{n_i - 1} \log(1 + h\tau) \right] \tag{4.6}$$

To obtain (4.6) from (4.5), we use the property of the gamma function: $\Gamma(x) = (x-1)\Gamma(x-1)$ and cancel terms common in both the numerator and denominator of (4.5); then express a_i and b_i in terms of μ_i and τ (Recall that μ_i is a function of β). If $y_i = 0$, the first term of (4.6) is taken to be zero. If $y_i = n_i$, the 2nd term is taken to be zero. Several authors (Kleinman 1973, Tamura & Young 1986, Tamura & Young 1987, among others) have tried to estimate the parameters of a Beta-binomial distribution using maximum-likelihood (ML) method, the method-of-moments, or some other modified methods. The ML estimates of a Beta-binomial distribution cannot be obtained as a closed form solution.

Tamura & Young (1986), in the context of tumor incidences of animals, showed that the maximum likelihood estimator of η ($\equiv 1/\tau$, in our notation) was biased to the right for low tumor probabilities for the small number and size of historical control groups usually available in chronic rodent bioassays. Tamura & Young (1987) reached the same conclusion with the help of alveolar-bronchiolar adenomas in mice data analysis. Following their argument, we can say that the maximum likelihood estimate of τ can be biased towards zero for selected applications. When that happens, the empirical Bayes strategy tends to put more weight to the regression synthetic part as compared to the direct part, irrespective of the sample

size for small areas. This phenomenon is not at all desirable. In many practical applications, maximum likelihood (ML) or restricted maximum likelihood (REML) estimates of hyperparameters occur at the boundary point. For example, in two-level Poisson-gamma model the ML estimate of the variance component can be infinity (Christiansen & Morris, 1997), for a basic area level Normal-normal model the ML or REML estimate of variance component can be zero (Bell 1999, Li & Lahiri 2008, Chapter 2 of this dissertation). For this reason we don't recommend empirical Bayes method as a general solution for analyzing binary data.

4.2.3 Choice of Prior on the Hyperparameters

In a hierarchical Bayesian approach, we need to consider some prior distribution on the hyperparameters instead of estimating it from the marginal distribution of data, as in an empirical Bayesian analysis. In this approach, to obtain $E(\theta_i|\mathbf{y})$ and $V(\theta_i|\mathbf{y})$ from (4.3) and (4.4) respectively, we need to use an iterative expectation and variance technique as follows:

$$\begin{aligned} E(\theta_i|\mathbf{y}) &= E\{E(\theta_i|\mathbf{y},\beta,\tau)|\mathbf{y}\} \\ &= E\{g_i(\beta,\tau)|\mathbf{y}\} \qquad (4.7) \\ V(\theta_i|\mathbf{y}) &= E\{V(\theta_i|\mathbf{y},\beta,\tau)|\mathbf{y}\} + V\{E(\theta_i|\mathbf{y},\beta,\tau)|\mathbf{y}\} \\ &= E\{h_i(\beta,\tau)|\mathbf{y}\} + V\{g_i(\beta,\tau)|\mathbf{y}\} \qquad (4.8) \end{aligned}$$

The expectation and variance in (4.7) and (4.8) are with respect to the posterior distribution of the hyperparameters β and τ. Note that here, unlike linear mixed model, the regression coefficient β cannot be integrated out from the joint distri-

bution (4.1). This makes it difficult to apply numerical integration technique, in case we need to perform high dimensional integration depending on the number of covariates involved in a particular data analysis. More complication arises due to the third level of prior distribution.

In the absence of subjective prior information, Stroud (1991, 1994) used some non-informative priors such as the uniform prior, which is proportional to a constant, and Jeffreys' prior, which is proportional to the determinant of the Fisher information matrix of the parameters. Improper noninformative priors should be used with caution because they may not result in a proper posterior. In the context of a hierarchical poisson-gamma model, Natarajan et al. (1998) used Jeffrey's prior at the third level, with a suitable modification to avoid improper posterior distribution in their analysis. He & Sun (1998) used a proper gamma distribution for the hyperparameters instead of Jeffreys or uniform prior.

Natarajan et al. and He & Sun, both implemented their models through Monte Carlo Markov Chain (MCMC) techniques, such as Gibbs sampling (Gelfand & Smith, 1990) and rejection sampling (Gilks & Wild, 1992). The slow computational speed of MCMC methods usually does not permit to evaluate the frequentist properties of the resulting Bayes rule (Christiansen & Morris, 1997). Moreover, one should not overlook the convergence issues of the MCMC methods. Although the use of the MCMC method is justified by the ergodic theorem, in practice results from a MCMC run can depend heavily on several factors (see Chapter 1, Section 1.4.4). All these factors are carefully examined in a Bayesian analysis. If the Bayesian methodology is to be carried out routinely by someone with minimal knowledge of

the sophisticated MCMC method, then the convergence of the MCMC technique may not be checked properly, which may lead to unreasonable conclusions. Also, the MCMC methods do not provide any analytical expressions of the posterior mean and variance. For these reasons many researchers prefer to apply some approximate methods, such as the adjusted density method (Christiansen & Morris 1997; Morris 1988; Tang 2002), or the Laplace approximation (Tierney & Kadane 1986; Tierney et al. 1989; Kass & Steffey 1989; Butar & Lahiri 2002; Datta et al. 2005; Natarajan et al. 1998), to carry out a Bayesian analysis.

We consider Laplace approximation to implement our hierarchical model to analyze binary data. The Laplace method does not work well if the posterior distribution of the hyperparameters is extremely skewed (Natarajan et al., 1998). We choose the prior for the hyperparameters carefully so that we can avoid the extreme skewness of the posterior distribution. In other words, this will ensure that the posterior mode of the hyperparameters falls inside the range of the parameter space. To ensure that the mode of a hyperparameter occurs at a finite value for a Poisson-gamma model, Christiansen & Morris (1997) used a proper prior distribution for the corresponding parameter. To avoid zero posterior mode for the variance components in area and unit-level Normal-normal models, in Chapter 2 and 3 of this dissertation, we proposed new prior distributions for the variance components and obtained good frequentist properties to the resulting rules. In order to prevent boundary solutions for the ML estimates, in the context of discrete data analysis, the Latent GOLD Choice software use Dirichlet priors for the latent and response probabilities (Vermunt & Magidson, 2005). For the same reason, we propose a noninformative

prior $\pi(\tau),\ \tau > 0$ on τ in this chapter, following some criteria. The assumption of improper uniform prior on β i.e. $\pi(\beta) \propto 1;\ \beta \in R^p$ is usually accepted as it provides good repeated sampling properties (Christiansen & Morris, 1997).

If we consider uniform priors for both β and τ, then the joint posterior distribution of β and τ is equivalent to the likelihood function of β and τ as defined in (4.6), after taking a logarithm. But as we discussed in the previous paragraph, this may result in a boundary posterior mode of τ. As a remedy, Morris (2006) suggested using adjusted REML likelihood. Li & Lahiri (2008), while studying the frequentist properties of the variance component in the context of normal-normal hierarchical model, have observed that the adjusted REML likelihood leads to an estimator with an overestimation problem. Use of adjusted ML overcomes this problem. For further details see Li & Lahiri (2008), Chapter 2 of this thesis. Following their criteria we recommend to use the following prior on τ

$$\pi(\tau) \propto \tau,\ \tau > 0. \tag{4.9}$$

We don't provide any condition for the propriety of the posterior resulting from this prior. But we hope this will lead to a proper joint posterior distribution of β and τ. Our conjecture is based on the fact that this type of prior on variance component leads to a proper posterior in the context of hierarchical linear mixed model which follows from the necessary and sufficient conditions given by Hobert & Casella (1996), for the propriety of the joint posterior for a general class of priors of which the prior (4.9) is a particular case. Also for a proof of proper posterior using this type of prior for a variance component in a basic area-level model see Appendix

A of Chapter 2.

4.3 Laplace Approximation

For convenient implementation and evaluation of our hierarchical Bayes method, we approximate the posterior moments of θ_i using Laplace's method. Many authors used Laplace approximation method in the context of Bayesian analysis (Tierney & Kadane 1986; Tierney et al. 1989; Kass & Steffey 1989; Butar & Lahiri 2002; Datta et al. 2005; Natarajan et al. 1998). To obtain the asymptotic variance-covariance matrix of the parameters of the model, the Latent GOLD Choice software (Vermunt & Magidson, 2005) uses the negative inverse of the Hessian matrix, where the Hessian matrix is obtained by computing the second order derivative of the log-posterior density (instead of the log-likelihood of the parameters without any prior assumption on the model parameters). This approach can be viewed as a first order Laplace approximation to the posterior variance of the model parameters.

The logarithm of the joint posterior density $f(\beta, \tau|\mathbf{y})$ of β and τ is given by

$$L(\beta,\tau) \propto \sum_{i=1}^{m}\left[\sum_{h=0}^{y_i-1}\log(\mu_i+h\tau)+\sum_{h=0}^{n_i-y_i-1}\log(1-\mu_i+h\tau)-\sum_{h=0}^{n_i-1}\log(1+h\tau)\right]+\log\tau \tag{4.10}$$

The first order Laplace approximation of the posterior moments of θ_i is given by

$$E(\theta_i|\mathbf{y}) = g_i(\hat{\beta},\hat{\tau}) + O(1/m) \tag{4.11}$$

$$V(\theta_i|\mathbf{y}) = h_i(\hat{\beta},\hat{\tau}) + \left\{Dg_i(\hat{\beta},\hat{\tau})\right\}' \hat{\Sigma} \left\{Dg_i(\hat{\beta},\hat{\tau})\right\} + O(1/m^2) \tag{4.12}$$

where $\hat{\beta}$ and $\hat{\tau}$ maximize $f(\beta, \tau|\mathbf{y})$ or equivalently, are the solution of the equation (4.10) whose first derivatives are set equal to zero. $g_i(\beta, \tau)$ and $h_i(\beta, \tau)$ are

defined in (4.3) and (4.4) respectively. $Dg_i(\hat{\beta},\hat{\tau})$ is the vector of first order partial derivatives of $g_i(\beta,\tau)$ with respect to β and τ, evaluated at $\hat{\beta},\hat{\tau}$ ($\hat{\beta},\hat{\tau}$ inside the parenthesis indicates that the corresponding function is evaluated at $\hat{\beta},\hat{\tau}$). $\hat{\Sigma}$ is the inverse of the negative Hessian of $L(\beta,\tau)$, evaluated at $\hat{\beta},\hat{\tau}$. It should be noted that both the approximations have relative error of order $O(1/m)$. For details on first order Laplace approximation, see Tierney et al. (1989); Kass & Steffey (1989). Note that the second term in the variance expression (4.12) captures the additional uncertainty in estimating the hyperparameters. The following derivatives may be useful to apply some iterative procedure to obtain the posterior mode $\hat{\beta}$ and $\hat{\tau}$ and the corresponding Hessian matrix. To better understand the following expressions recall that $\text{logit}(\mu_i) = x_i'\beta$.

$$\frac{\partial L(\beta,\tau)}{\partial \beta} = \sum_{i=1}^{m} \mu_i(1-\mu_i)\left[\sum_{h=0}^{y_i-1}\frac{1}{\mu_i+h\tau} - \sum_{h=0}^{n_i-y_i-1}\frac{1}{1-\mu_i+h\tau}\right]x_i$$

$$\frac{\partial^2 L(\beta,\tau)}{\partial \beta \beta'} = \sum_{i=1}^{m} \mu_i(1-\mu_i)\left[\sum_{h=0}^{y_i-1}\frac{h\tau(1-2\mu_i)-\mu_i^2}{(\mu_i+h\tau)^2} - \sum_{h=0}^{n_i-y_i-1}\frac{h\tau(1-2\mu_i)+(1-\mu_i)^2}{(1-\mu_i+h\tau)^2}\right]x_i x_i'$$

$$\frac{\partial^2 L(\beta,\tau)}{\partial \beta \partial \tau} = \sum_{i=1}^{m} \mu_i(1-\mu_i)\left[-\sum_{h=0}^{y_i-1}\frac{h}{(\mu_i+h\tau)^2} + \sum_{h=0}^{n_i-y_i-1}\frac{h}{(1-\mu_i+h\tau)^2}\right]x_i$$

$$\frac{\partial L(\beta,\tau)}{\partial \tau} = \sum_{i=1}^{m}\left[\sum_{h=0}^{y_i-1}\frac{h}{\mu_i+h\tau} + \sum_{h=0}^{n_i-y_i-1}\frac{h}{1-\mu_i+h\tau} - \sum_{h=0}^{n_i-1}\frac{h}{1+h\tau}\right] + \frac{1}{\tau}$$

$$\frac{\partial^2 L(\beta,\tau)}{\partial \tau^2} = \sum_{i=1}^{m}\left[-\sum_{h=0}^{y_i-1}\frac{h^2}{(\mu_i+h\tau)^2} - \sum_{h=0}^{n_i-y_i-1}\frac{h^2}{(1-\mu_i+h\tau)^2} + \sum_{h=0}^{n_i-1}\frac{h^2}{(1+h\tau)^2}\right] - \frac{1}{\tau^2}$$

The Newton Raphson algorithm for solving simultaneous equations is given by:

$$\begin{pmatrix} \beta^{(k+1)} \\ \tau^{(k+1)} \end{pmatrix} = \begin{pmatrix} \beta^{(k)} \\ \tau^{(k)} \end{pmatrix} - \begin{bmatrix} \frac{\partial^2 L(\beta,\tau)}{\partial\beta\beta'} & \frac{\partial^2 L(\beta,\tau)}{\partial\beta\partial\tau} \\ \frac{\partial^2 L(\beta,\tau)}{\partial\beta\partial\tau} & \frac{\partial^2 L(\beta,\tau)}{\partial\tau^2} \end{bmatrix}^{-1}_{\beta^{(k)},\tau^{(k)}} \begin{pmatrix} \frac{\partial L(\beta,\tau)}{\partial\beta} \\ \frac{\partial L(\beta,\tau)}{\partial\tau} \end{pmatrix}_{\beta^{(k)},\tau^{(k)}}$$

The convergence of Newton-Raphson method depends heavily on the initial values of β and τ. For some initial values of τ, the algorithm does not converge. We need to consider some objective criteria (such as using method-of-moment estimates as initial values).

To obtain the second order Laplace approximation to the posterior mean and variance of θ_i, we apply the fully exponential form suggested by Tierney & Kadane (1986). This has the advantage of requiring only the first two derivatives of the corresponding posterior distribution to achieve a second order approximation, unlike the standard form which needs fourth and fifth derivatives of the log-likelihood (Kass et al., 1988, p. 265). The second order approximation of posterior mean is given by

$$E(\theta_i|\mathbf{y}) = \left\{ \frac{|\hat{\Sigma}^*|}{|\hat{\Sigma}|} \right\}^{1/2} \exp\left[L^*(\hat{\beta}^*, \hat{\tau}^*) - L(\hat{\beta}, \hat{\tau}) \right], \tag{4.13}$$

where $L^*(\beta,\tau) = L(\beta,\tau) + \log\, g_i(\beta,\tau)$, and $\hat{\beta}^*$ and $\hat{\tau}^*$ maximize L^*. The $\hat{\beta}^*, \hat{\tau}^*$ inside the parenthesis indicates that the corresponding function is evaluated at $\hat{\beta}^*, \hat{\tau}^*$. $\hat{\Sigma}^*$ is the inverse of the negative Hessian of $L^*(\beta,\tau)$ evaluated at $\hat{\beta}^*, \hat{\tau}^*$.

The second order approximation of posterior variance is given by

$$V(\theta_i|\mathbf{y}) = T_1 + T_2, \tag{4.14}$$

where T_1 = right hand side (RHS) of (4.13) with $L^*(\beta,\tau) = L(\beta,\tau) + \log\, h_i(\beta,\tau)$. T_2 = right hand side of (4.13) with $L^*(\beta,\tau)$ as $L(\beta,\tau) + \log\, g_i^2(\beta,\tau) - \{$ RHS of (4.13)$\}^2$.

To apply the fully exponential form, we need $g_i(\beta, \tau)$ and $h_i(\beta, \tau)$ to be positive functions of β, τ. Both of the above second order approximations are of relative order $O(1/m^2)$.

4.4 Data Analysis

In this section we carry out two empirical applications, one of which includes covariates in the analysis.

4.4.1 Missouri Turkey Hunting Survey Data

We apply our hierarchical model on the Missouri Turkey Hunting Survey (MTHS) 1994 spring season data. This data have been analyzed earlier by He & Sun (1998). The dataset can be found in Table 1 of their paper. In this dataset, n_i is the total number of hunting trips to the ith area, y_i is the number of successful individuals among the sample of n_i. We define θ_i as the probability of success for each individual in the ith area. Our aim is to obtain $E(\theta_i|\mathbf{y})$ and $V(\theta_i|\mathbf{y})$. He & Sun (1998) used an exchangeable model at the second level of their hierarchical model, although they agreed that the success rate may depend on some set of covariates. This dataset does not include any covariates. If covariates were available, to include it in the analysis one could easily apply our model to estimate the hunting success rate for different counties in Missouri. At the third level, they used subjective proper prior in the form of a gamma distribution for the hyperparameters. We prefer to use the noninformative prior $\pi(\tau) \propto \tau$ for τ in our analysis.

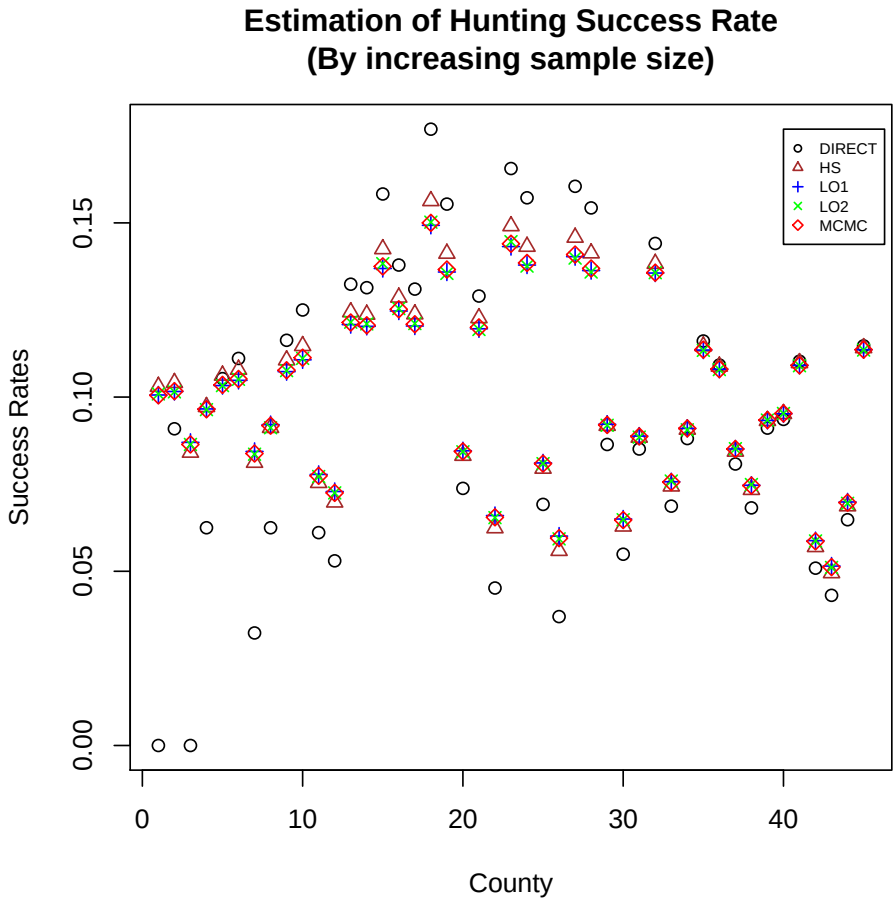

Figure 4.1: Different Point Estimates of Hunting Success Rates

We have 114 counties (small areas) in this dataset. Presenting the results for all these counties in a tabular form may not be very informative. That's why we present the results with the help of figures. In the figures, we use only 45 counties. We select 10 counties with small sizes (ranging from 2 to 48), 19 counties having moderate sample size (ranging from 131-162), and 16 large counties with sample size in the range of 328-802. In other words, in the figures we present a representative sample as far as the sample size of the counties is concerned. In Figure 4.1, we plot five point estimates of the hunting success rate. Five estimators considered in this plot are: sample mean (DIRECT), the hierarchical Bayes estimate discussed in the He & Sun

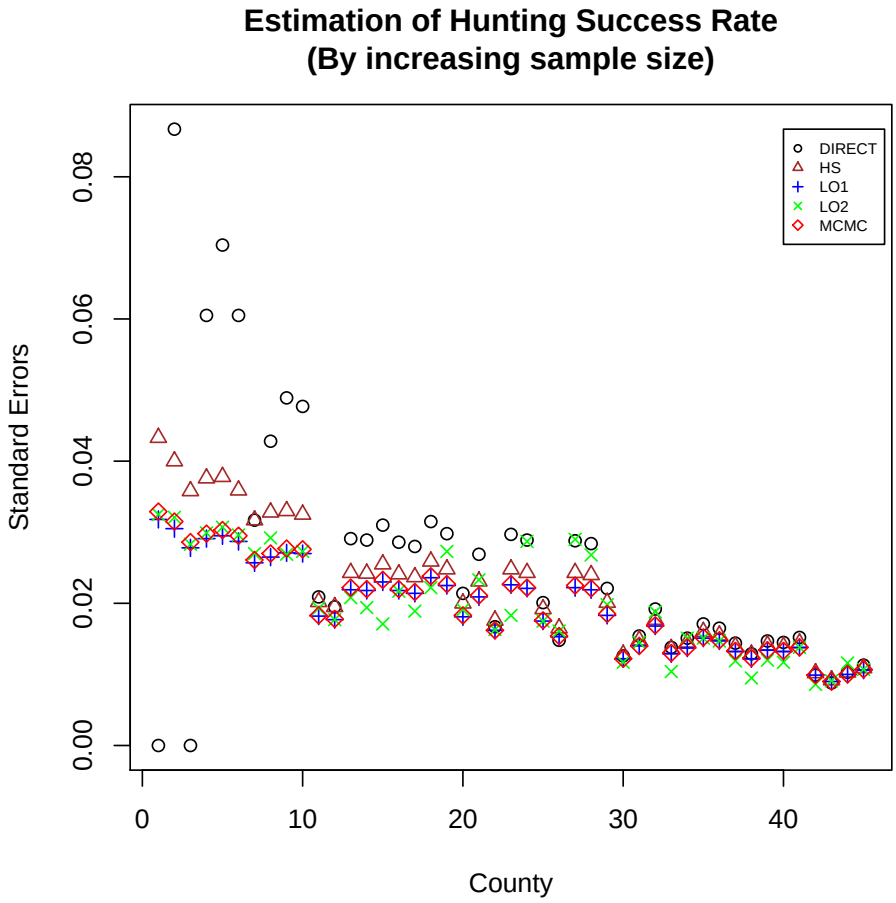

Figure 4.2: Standard Errors Associated with Different Point Estimates: Turkey Hunting Data

(HS), Laplace first order approximation to the posterior mean of θ_i using the new prior for τ (LO1), Laplace second order approximation (LO2). The estimate MCMC is obtained from the `BRugs` (Thomas et al., 2006) output using the new prior. For small counties (left side of the plot), we can evidently see that the direct estimates are quite different from all the model based estimates, the difference decreases with the increase in the sample size. All the model based estimates are more or less identical. For some counties (having moderate sample size), HS estimate appears to be closer to the direct estimate than the estimates obtained using our model. Comparing

the estimates obtained using Laplace approximation to the MCMC output we can say that Laplace approximation is performing well in estimating the hunting success rate. Figure 4.2 exhibits the standard errors (SEs) assosiated with the estimators considered in Figure 4.1. The SEs decrease with the increase in the sample size for all the estimators. For small counties, the extremely high SEs (compared to others) underscore the fact that for small areas model based methods should be preferred to the direct estimator. For counties having small and moderately large sample size, the SEs associated with the HS estimators are larger compared to that of our estimator. Here LO1 and LO2 represent the Laplace first and second order approximation to the posterior variance of θ_i, using our new prior.

In Figure 4.3, we formally evaluate Laplace approximation using the estimates and its measure of uncertainties for all 114 counties. To measure the precision of Laplace method, we compute the percentage difference as the summary statistics. Mathematically, this can be defined as $\{$(exact-approximate)/exact$\}\times 100$. We treat the output from the `BRugs` package as the exact posterior moments. Figure 4.3 shows that both the first and second order approximation values of the mean are quite close to the exact value. The percentage difference values lie on the zero line for almost all counties. The first order approximation to the posterior variance works very well, the absolute difference being less than 5% for all counties. There is a slight tendency toward under-estimation, the percentage difference being positive for all the counties. The second order approximation to the posterior variance does not perform very well. This result seems to be very surprising.

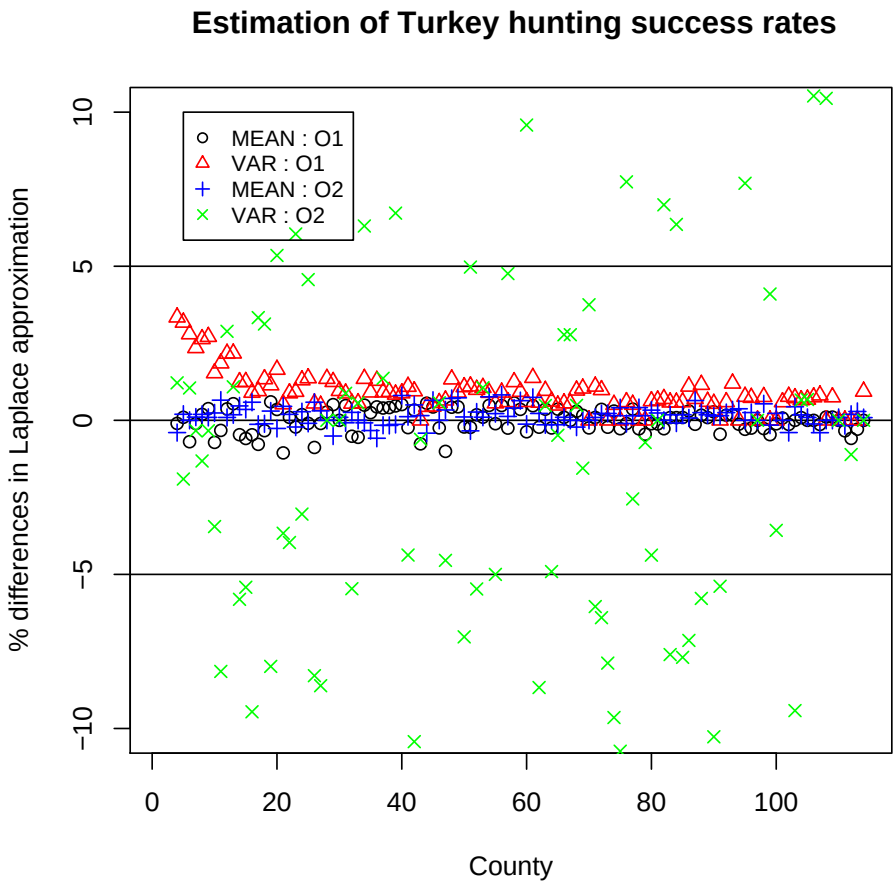

Figure 4.3: Evaluation of Laplace Approximation using Turkey Hunting Data: Percent Difference from MCMC as a Summary Measure

4.4.2 Baseball Data

In this section, we revisit the baseball data example given in Efron & Morris (1975). This data set has been analyzed by several researchers in the past, including Efron & Morris (1975); Morris (1983b); Gelman et al. (1995); Datta & Lahiri (2000); Rao (2003); Jiang & Lahiri (2006b), among others. Efron & Morris (1975) used this dataset to demonstrate the performance of their empirical Bayes method with an exchangeable prior in the presence of an outlying observation. They showed that the James and Stein's (1961) estimator can be derived from an empirical Bayes

context. Efron & Morris (1975) considered the problem of estimating the batting averages of 18 baseball players (small areas, in our terminology). Although this is a proportion estimation problem, they used hierarchical normal-normal model to analyze the data, using suitable transformation (arcsin). At the end, the estimates are transformed back to obtain the estimates of proportion. They did not include any auxiliary variable in their model, instead they used an exchangeable prior at the second level.

Gelman et al. (1995) provided additional data for this estimation problem and included important covariates like the batting average of each player in the previous season (1969), and the number of times at bat in the 1969 season. The dataset is given in the Appendix. We review the problem of estimating the batting averages of all the 18 players for the entire 1970 season. We apply our hierarchical model on this dataset and study the utility of using covariates in the analysis. We have a sample of size 45 for each player. Let n_i be the number of times at bat for the ith player. Then, $n_i = 45\ \forall i$. Let y_i be the number of hits among the number at-bats for the player i. Also let θ_i be the true batting average for the 1970 season for player i, which is known. The dataset given in Appendix differs from that of Efron & Morris, 1975 (their Table 1) with respect to the true values (besides having no covariate), as they considered the problem of predicting the batting averages for the remainder of the 1970 season.

In Figure 4.4, we compare four different estimators of the 1970 season batting averages along with the true value. These four estimators are: sample proportion (DIRECT), Efron and Morris 1975 empirical Bayes estimator (EM), Laplace

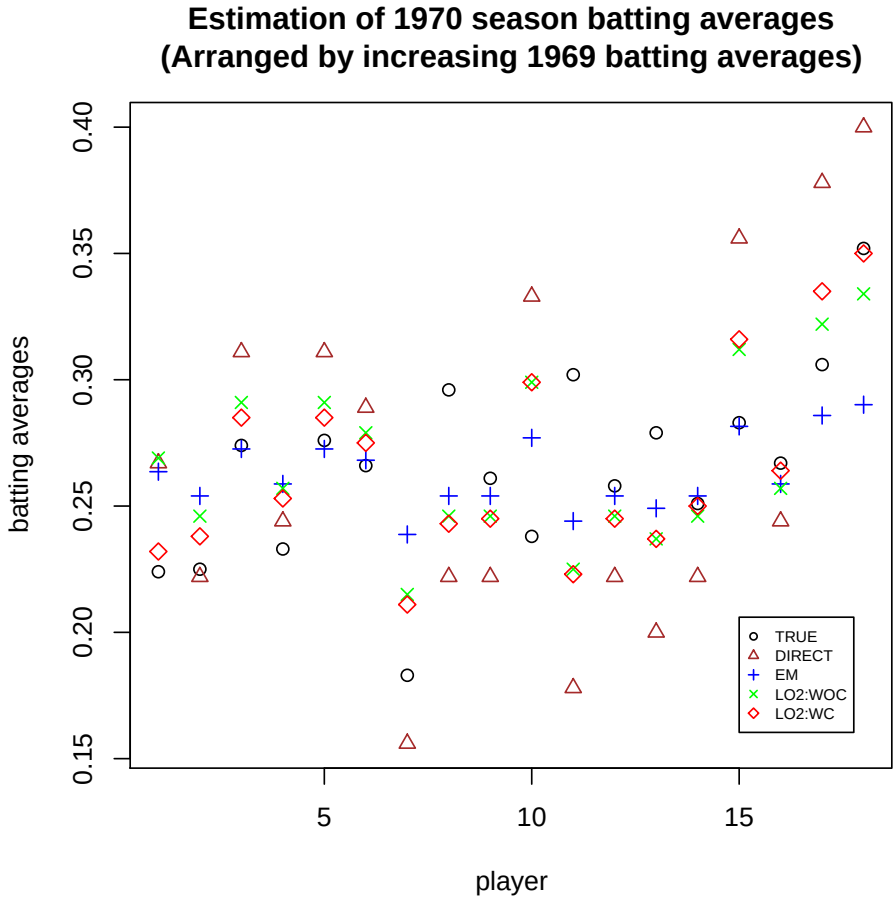

Figure 4.4: Comparison of Different Point Estimates: Baseball Data

second order approximation to the posterior mean of true batting average using the previous year batting average as a covariate (LO2:WC), Laplace second order approximation to the posterior mean of true batting average without covariate (LO2:WOC). LO2:WC and LO2:WOC have been obtained using our new hierarchical model. The players are arranged in increasing order of previous batting averages in the plot. Clemente, an extremely good hitter, is undoubtedly an outlier. Jiang & Lahiri (2006b) noted that the player Alvarado is also an outlier in the sense that his current batting average is much better than his previous batting average. For further discussion on this see Jiang & Lahiri (2006b, p. 42-44). From the plot we

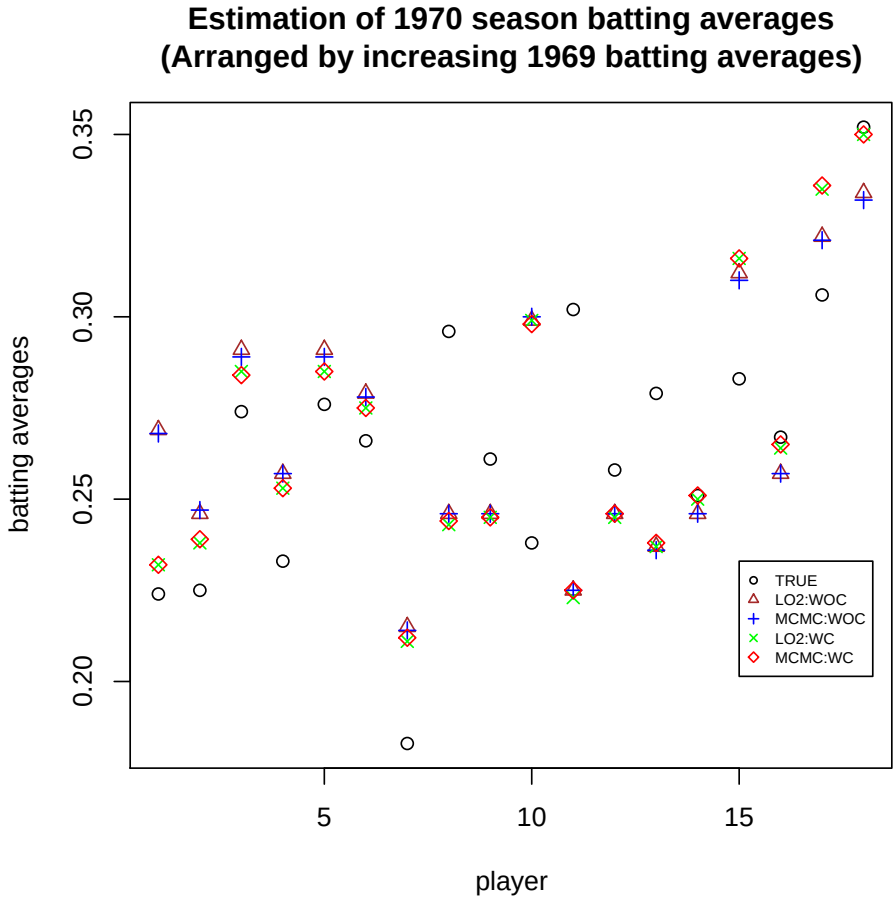

Figure 4.5: Comparison of Point Estimates: Laplace Approximation vs MCMC

can say that LO2:WC, the estimator which uses covariates in the analysis, does a great job in predicting the batting averages of Clemente and Alvarado, two different types of outliers. For the first (Alvarado) and the last (Clemente) player in the plot, the LO2:WC and TRUE values are very close. This fact was also noted by Jiang & Lahiri (2006b), who used covariates at the second level of the normal hierarchical model. The performance of the sample proportion (DIRECT), which is unbiased and the maximum likelihood estimator under the binomial assumption, is very poor. For the values of the estimates of all the players along with their standard errors see Table 4.1 and Table 4.2.

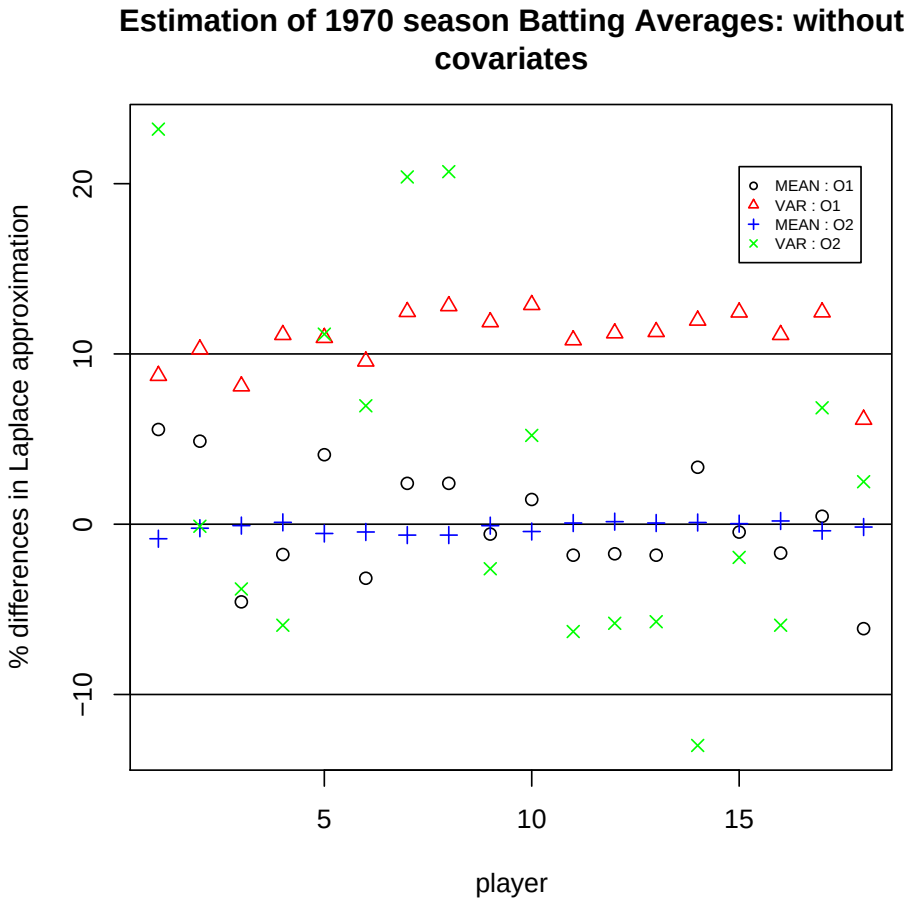

Figure 4.6: Evaluation of Laplace Approximation without using Covariate: Percent Difference from MCMC as a Summary Measure

In Figure 4.5, we compare the second order Laplace approximation to the MCMC output. Here also we carried out the analysis twice, once using the covariate (WC) and another time without using it (WOC). The Laplace approximation works well as the LO2:WC values coincide with the MCMC:WC and LO2:WOC values coincide with the MCMC:WOC for most of the players. As noted earlier, inclusion of the previous year batting average in the analysis has a prominent effect in estimating the batting averages of two outliers. Also it (WC) does a better job for some other players as well.

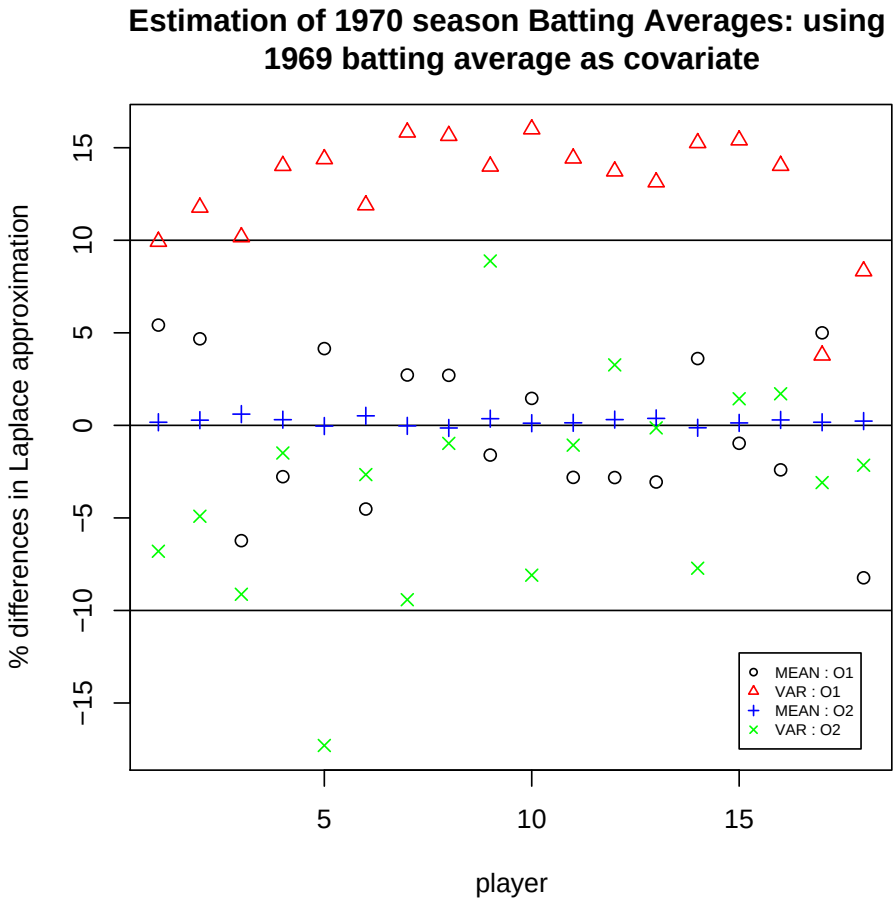

Figure 4.7: Evaluation of Laplace Approximation using Covariate: Percent Difference from MCMC as a Summary Measure

In the next two figures (Figure 4.6 and Figure 4.7), we formally evaluate Laplace approximation using the estimates and its measure of uncertainties for all the 18 players. To measure the precision of Laplace method, we compute the percentage difference as the summary statistics. Mathematically, this can be defined as {(exact-approximate)/exact}×100. We treat the output from the BRugs package as the exact posterior moments. In both the figures (one with covariate another without covariate), we can see that both the first and second order approximation values of the mean are quite close to the exact value, although the second order

values are more accurate as the percentage difference values lie on the zero line for almost all players. Use of covariate leads to better second order approximation of the posterior variance if we make a comparison between Figure 4.7 and Figure 4.6.

Table 4.1: Results from the Baseball Data Analysis: without using Covariate

Player	Direct	True	LO1	LO2	MCMC	se.LO1	se.LO2	se.MCMC
Clemente	0.400	0.352	0.313	0.334	0.332	0.050	0.042	0.055
F.Robins	0.378	0.306	0.305	0.322	0.321	0.048	0.053	0.053
Munson	0.178	0.302	0.235	0.225	0.225	0.042	0.048	0.046
Scott	0.222	0.296	0.251	0.246	0.246	0.041	0.049	0.046
F.Howard	0.356	0.283	0.298	0.312	0.310	0.046	0.046	0.051
Campaner	0.200	0.279	0.243	0.237	0.236	0.041	0.043	0.046
Spencer	0.311	0.276	0.282	0.291	0.289	0.042	0.039	0.049
Berry	0.311	0.274	0.282	0.291	0.289	0.042	0.039	0.049
Swoboda	0.244	0.267	0.258	0.257	0.257	0.041	0.047	0.046
Kessinge	0.289	0.266	0.274	0.279	0.278	0.041	0.045	0.048
E.Rodrig	0.222	0.261	0.251	0.246	0.246	0.041	0.049	0.046
Williams	0.222	0.258	0.251	0.246	0.246	0.041	0.049	0.046
Unser	0.222	0.251	0.251	0.246	0.246	0.041	0.049	0.046
Johnston	0.333	0.238	0.290	0.299	0.300	0.044	0.056	0.050
Santo	0.244	0.233	0.258	0.257	0.257	0.041	0.047	0.046
Petrocel	0.222	0.225	0.251	0.246	0.247	0.041	0.049	0.046
Alvarado	0.267	0.224	0.267	0.269	0.268	0.041	0.043	0.047
Alvis	0.156	0.183	0.228	0.215	0.214	0.044	0.045	0.047

Table 4.2: Results from the Baseball Data Analysis: using 1969 Batting Average as Covariate

Player	Direct	True	LO1	LO2	MCMC	se.LO1	se.LO2	se.MCMC
Clemente	0.400	0.352	0.332	0.350	0.350	0.051	0.061	0.057
F.Robins	0.378	0.306	0.320	0.335	0.336	0.048	0.057	0.055
Munson	0.178	0.302	0.239	0.223	0.225	0.041	0.050	0.046
Scott	0.222	0.296	0.251	0.243	0.244	0.039	0.046	0.046
F.Howard	0.356	0.283	0.302	0.316	0.316	0.044	0.060	0.052
Campaner	0.200	0.279	0.249	0.237	0.238	0.041	0.047	0.046
Spencer	0.311	0.276	0.277	0.285	0.285	0.041	0.053	0.049
Berry	0.311	0.274	0.277	0.285	0.284	0.041	0.049	0.049
Swoboda	0.244	0.267	0.269	0.264	0.265	0.041	0.043	0.048
Kessinge	0.289	0.266	0.271	0.275	0.275	0.040	0.051	0.047
E.Rodrig	0.222	0.261	0.252	0.245	0.245	0.039	0.046	0.046
Williams	0.222	0.258	0.253	0.245	0.246	0.039	0.044	0.046
Unser	0.222	0.251	0.258	0.250	0.251	0.040	0.047	0.047
Johnston	0.333	0.238	0.288	0.299	0.298	0.042	0.054	0.050
Santo	0.244	0.233	0.255	0.253	0.253	0.039	0.045	0.046
Petrocel	0.222	0.225	0.245	0.238	0.239	0.039	0.045	0.046
Alvarado	0.267	0.224	0.221	0.232	0.232	0.056	0.060	0.059
Alvis	0.156	0.183	0.229	0.211	0.212	0.042	0.047	0.046

4.5 Concluding Remarks

In this chapter, we propose hierarchical models to estimate small area proportions. The linear normal mixed models, considered in Chapter 2 and 3 are usually not applicable to analyze binary data. Our proposed hierarchical binomial-beta model leads to a simple expression for the best predictor of true small area proportion, as opposed to considering mixed logistic model, usually applied in small area estimation to borrow strength from other areas, in the presence of covariates in the analysis. We recommended to use an improper noninformative prior for the shape parameter of the beta density at the third level of our hierarchical model that would keep the posterior mode within the parameter space. In future, we will study the

frequentist properties of the resulting Bayes estimators of the small area proportions using the prior we recommend in this chapter.

4.6 Appendix

4.6.1 Baseball Data

Player	Direct	x1	x2	true
Clemente	0.400	0.314	8142	0.352
F.Robins	0.378	0.303	7542	0.306
Munson	0.178	0.256	86	0.302
Scott	0.222	0.250	2065	0.296
F.Howard	0.356	0.275	4826	0.283
Campaner	0.200	0.264	3210	0.279
Spencer	0.311	0.246	2244	0.276
Berry	0.311	0.244	454	0.274
Swoboda	0.244	0.281	5658	0.267
Kessinge	0.289	0.248	2753	0.266
E.Rodrig	0.222	0.255	2281	0.261
Williams	0.222	0.257	1216	0.258
Unser	0.222	0.271	888	0.251
Johnston	0.333	0.255	1139	0.238
Santo	0.244	0.244	1967	0.233
Petrocel	0.222	0.234	291	0.225
Alvarado	0.267	0.118	51	0.224
Alvis	0.156	0.249	3514	0.183

4.6.2 BRugs model specification without covariates

```
model
{
for(i in 1:m){
y[i] ~ dbin(theta[i], n[i]) ##n[i] should be greater than 1 for all i.
theta[i] ~ dbeta(a, b)
}
a <- mu/tau
b <- (1-mu)/tau
mu ~ dunif(0,1)
## to specify new prior for tau
dummy <- 0
dummy ~ dgeneric(ll)
ll <- log(tau)
tau ~ dflat()T(0,)
}
```

4.6.3 BRugs model specification including covariates

```
model
{
for(i in 1:m){
y[i] ~ dbin(theta[i], n[i]) ##n[i] should be greater than 1 for all i.
lp[i] <- inprod(X[i,], beta[])
mu[i] <- exp(lp[i])/(1+exp(lp[i]))
a[i] <- mu[i]/tau
b[i] <- (1-mu[i])/tau
theta[i] ~ dbeta(a[i], b[i])
}
beta[1] ~ dflat()
beta[2] ~ dflat()
dummy <- 0
dummy ~ dgeneric(ll)
ll <- log(tau)
tau ~ dflat()T(0,)
}
```

Chapter 5

Concluding Remarks and Future Research

In this dissertation, we have developed new hierarchical Bayesian methods that are useful for analyzing both discrete and continuous data for small area estimation. Throughout the dissertation, our main goal has been to choose prior distributions for the hyperparameters that offer good frequentist properties and at the same time provide accurate approximation to the complex posterior distributions by the Laplace method.

The extremely skewed posterior distribution of the variance component is an unavoidable consequence of the asymmetry in the parameter space, with variance parameters restricted to be positive. Our prior choice avoids the extreme skewness of the posterior distribution. As a result, the Laplace approximation usually works well, a result that contradicts some earlier research. We studied the frequentist properties of the Bayes estimator. Our simulation results show that the frequentist properties (e.g. MSE, coverage) of the Bayes estimator of the true small area quantity (θ_i, say) corresponding to our proposed prior is better than the popular choice of uniform prior and some other methods.

Although we have developed our hierarchical Bayes methods to address small area estimation problems, there is a great potential for using such models in other important applications. For example, in surveys there is a great deal of interest in

estimating intra-interviewer correlation and the associated interviewer effects. The current statistical literature on interviewer effects focus mostly on point estimation where the ANOVA method is the usual tool. The ANOVA method can yield an intra-interviewer correlation estimate outside the parameter range. We intend to explore the methods proposed in the dissertation in addressing both point estimation and interval estimation for making inference about the intra-interviewer correlation and the interviewer effects. Another application of potential interest is the disease mapping problem where it is necessary to smooth prevalence of certain diseases across geography. In the future, we would like to extend our methodology to include hierarchical Poisson models and spatial models so we can address the important disease mapping problem.

Many large scale national surveys employ complex sample designs involving several layers of stratification and clustering. In order to capture the variability of such complex survey data, one needs more complex generalized linear mixed models than the ones considered in the dissertation. In the future, we would like to find a prior distribution for variance components using a general linear mixed model, that would retain all the useful properties we observed in case of relatively simple models. We also want to exploit some other objective criteria to choose a prior for the variance components. One of them is to match the expectation of the posterior variance of θ_i to the frequentist MSE, following Datta et al. (2005) and Ganesh & Lahiri (2008). Unlike them, we would like to consider frequentist MSE as the mean squared error of empirical Bayes estimator of θ_i when ADM method (Morris, 2006; Li & Lahiri, 2008) is used to estimate the variance components.

Bibliography

Albert, J. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *Journal of the American Statistical Association* 83 1037–1044.

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

Arora, V., Lahiri, P. & Mukherjee, K. (1997). Empirical Bayes estimation of finite population means from complex surveys. *Journal of the American Statistical Association* 92 1555–1562.

Battese, G., Harter, R. & Fuller, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* 83 28–36.

Bell, W. (1999). Accounting for Uncertainty About Variances in Small Area Estimation. *Bulletin of the International Statistical Institute* 52.

Bell, W. & Otto, M. (1992). Bayesian Assessment of Uncertainty in Seasonal Adjustment with Sampling Error Present. Tech. rep.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

Brooks, S. & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* 7 434–455.

Butar, F. B. & Lahiri, P. (2002). Empirical Bayes estimation of several population means and variances under random sampling variances model. *J. Statist. Plann. Inference* 102 59–69.

Carter, G. & Rolph, J. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association* 69 880–885.

Chatterjee, S., Lahiri, P. & Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Annals of Statistics* 36 1221.

Christiansen, C. L. & Morris, C. N. (1997). Hierarchical Poisson regression modeling. *Journal of the American Statistical Association* 92 618–632.

Citro, C. & Kalton, G., eds. (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*. Panel on Estimates of Poverty for Small Geographic Area, Committee on National Statistics, Washington, DC: National Academy Press.

Cox, D. R. & Reid, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika* 79 408–411.

Cox, D. R. & Reid, N. (1993). A note on the calculation of adjusted profile likelihood. *J. Roy. Statist. Soc. Ser. B* 55 467–471.

Datta, G. S. & Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics* 19 1748–1770.

Datta, G. S. & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sinica* 10 613–627.

Datta, G. S., Lahiri, P., Maiti, T. & Lu, K. L. (1999). Hierarchical Bayes Estimation of Unemployment Rates for the States of the US Journal of the American Statistical Association. *J. Amer. Statist. Assoc.* 94 1074–82.

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

Datta, G. S. & Smith, D. (2003). On propriety of posterior distributions of variance components in small area estimation. *Journal of Statistical Planning and Inference* 112 175–183.

Dempster, A. & Tomberlin, T. (1980). The analysis of census undercount from a postenumeration survey. In *Proceedings of the Conference on Census Undercount.* 88–94.

Dempster, A. P., Rubin, D. B. & Tsutakawa, R. K. (1981). Estimation in covariance components models. *J. Amer. Statist. Assoc.* 76 341–353.

Efron, B. & Morris, C. (1975). Data analysis using Steins estimator and its generalizations. *Journal of the American Statistical Association* 70 311–319.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *J. Roy. Statist. Soc. Ser. B* 31 195–233.

Erkanli, A. (1994). Laplace Approximations for Posterior Expectations When the Mode Occurs at the Boundary of the Parameter Space. *Journal of the American Statistical Association* 89 250–250.

Farrell, P. J., MacGibbon, B. & Tomberlin, T. J. (1997). Empirical bayes small-area estimation using logistic regression models and summary statistics. *Journal of Business & Economic Statistics* 15 101–108.

Fay, R. & Train, G. (1995). Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties. In *American Statistical Association, Proceedings of the Section on Government Statistics.*

Fay, R. E., III & Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* 74 269–277.

Ganesh, N. & Lahiri, P. (2008). A new class of average moment matching priors. *Biometrika* 95 514.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* 85 398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1 515–534.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

Gelman, A. & Rubin, D. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 7 457–457.

Gershunskaya, J. B. & Lahiri, P. (2005). Variance estimation for domains in the u.s. current employment statistics program. In *American Statistical Association, Proceedings of the Survey Research Methods Section*. 3044–3051.

Ghosh, M. (2008). Bayesian developments in surveys. Tech. rep.

Ghosh, M. & Lahiri, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *J. Amer. Statist. Assoc.* 82 1153–1162.

Ghosh, M. & Lahiri, P. (1989). A hierarchical Bayes approach to small area estimation with auxiliary information. In *Proceedings of the Joint Indo-U.S. Workshop on Bayesian Inference in Statistics and Econometrics*.

Ghosh, M. & Meeden, G. (1986). Empirical Bayes estimation in finite population sampling. *J. Amer. Statist. Assoc.* 81 1058–1062.

Ghosh, M. & Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statist. Sci.* 9 55–93.

Gilbert, P. (2008). *numDeriv: Accurate Numerical Derivatives*. R package version 2006.4-1, URL `http://www.bank-banque-canada.ca/pgilbert`.

Gilks, W. R. & Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *J. Royal Statist. Soc. Ser. C (Applied Statistics)* 337–348.

Hall, P. & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 68 221–238.

Hartley, H. O. & Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* 54 93–108.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* 61 383–385.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* 72 320–340.

He, Z. & Sun, D. (1998). Hierarchical Bayes estimation of hunting success rates. *Environmental and Ecological Statistics* 5 223–236.

He, Z. & Sun, D. (2000). Hierarchical bayes estimation of hunting success rates with spatial correlations. *Biometrics* 56 360–367.

Heady, P. & Clarke, P., eds. (2003). *Model-based small area estimation series No. 2*. Small Area Estimation Project Report. Office for National Statistics, U.K.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics* 9 226–252.

Hinrichs, P. (2003). *Consumer Expenditure Estimation Incorporating Generalized Variance Functions in Hierarchical Bayes Models*. Ph.D. thesis, University of Nebraska-Lincoln.

Hobert, J. & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* 91 1461–1473.

James, W. & Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*. Berkeley, Calif.: Univ. California Press, 361–379.

Jiang, J. (1996). REML estimation: Asymptotic behavior and related topics. *Annals of Statistics* 24 255–286.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer.

Jiang, J. & Lahiri, P. (2001). Empirical Best Prediction for Small Area Inference with Binary Data. *Annals of the Institute of Statistical Mathematics* 53 217–243.

Jiang, J. & Lahiri, P. (2006a). Estimation of Finite Population Domain Means: A Model-Assisted Empirical Best Prediction Approach. *Journal of the American Statistical Association* 101 301–311.

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

Kass, R., Tierney, L. & Kadane, J. (1988). Asymptotics in bayesian computation. In *Bayesian Statistics 3*. Oxford University Press, 261–278.

Kass, R. E. & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.* 84 717–726.

Kish, L. (1965). *Survey sampling*. John Wiley New York.

Kleffe, J. & Rao, J. N. K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *J. Multivariate Anal.* 43 1–15.

Kleinman, J. (1973). Proportions with extraneous variance: single and independent samples. *Journal of the American Statistical Association* 68 46–54.

Kott, P. (1989). Robust small domain estimation using random effects modeling. *Survey Methodology* 15 3–12.

Lahiri, P. (2003). A review of empirical best linear unbiased prediction for the Fay-Herriot small-area model. *The Philippine Statistician* 52 1–15.

Lahiri, P. & Mukherjee, K. (2007). On the design-consistency property of hierarchical Bayes estimators in finite population sampling. *Ann. Statist.* 35 724–737.

Li, H. (2007). *Small Area Estimation: An Empirical Best Linear Unbiased Prediction Approach*. Ph.D. thesis, Dept. of Mathematics, University of Maryland, College Park.

Li, H. & Lahiri, P. (2008). Adjusted density maximization method: An application to the small area estimation problem. Tech. rep.

Malec, D. & Sedransk, J. (1985). Bayesian inference for finite population parameters in multistage cluster sampling. *Journal of the American Statistical Association* 80 897–902.

Malec, D., Sedransk, J., Moriarity, C. L. & LeClere, F. B. (1997). Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association* 92 815–826.

Marshall, R. (1991). Mapping disease and mortality rates using empirical bayes estimators. *JR Stat Soc Ser C Appl Stat* 40 283–94.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall/CRC.

McCulloch, C. E. (2003). *Generalized linear mixed models*. NSF-CBMS Regional Conference Series in Probability and Statistics, 7. Beachwood, OH: Institute of Mathematical Statistics.

Morris, C. N. (1983a). Natural exponential families with quadratic variance functions: statistical theory. *Ann. Statist.* 11 515–529.

Morris, C. N. (1983b). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* 78 47–65.

Morris, C. N. (1988). Approximating posterior distributions and posterior moments. In *Bayesian statistics, 3 (Valencia, 1987)*, Oxford Sci. Publ. New York: Oxford Univ. Press, 327–344.

Morris, C. N. (2006). Discussion of Mixed model prediction and small area estimation. *Test* 15 1–96.

Morris, C. N. & Christiansen, C. L. (1996). Hierarchical models for ranking and for identifying extremes, with applications. In *Bayesian statistics, 5 (Alicante, 1994)*, Oxford Sci. Publ. New York: Oxford Univ. Press, 277–296.

Moura, F. & Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology* 25 73–80.

Nandram, B. (1999). An empirical bayes prediction interval for the finite population mean of a small area. *Statistica Sinica* 9 325–344.

Nandram, B. & Sedransk, J. (1993). Empirical Bayes estimation of the finite population mean on the current occasion. *J. Amer. Statist. Assoc.* 88 994–1000.

Natarajan, K., Ghosh, M. & Maiti, T. (1998). Hierarchical bayes quality measurement plan. *Communications in Statistics-Simulation and Computation* 27 199–214.

Neyman, J. & Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica* 16 1–32.

Otto, M. & Bell, W. (1995). Sampling error modelling of poverty and income statistics for states. In *American Statistical Association, Proceedings of the Section on Government Statistics*. 160–165.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58 545–554.

Pfeffermann, D. & Nathan, G. (1981). Regression analysis of data from a cluster sample. *Journal of the American Statistical Association* 76 681–689.

Prasad, N. & Rao, J. (1999). On Robust Small Area Estimation Using a Simple Random Effects Model. *Survey Methodology* 25 67–72.

Prasad, N. G. N. & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc.* 85 163–171.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Rao, C. R. (1973). *Linear statistical inference and its applications*. John Wiley & Sons, New York-London-Sydney, 2nd ed. Wiley Series in Probability and Mathematical Statistics.

Rao, J. N. K. (2003). *Small area estimation*. Wiley Series in Survey Methodology. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons].

Robert, C. & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

Royall, R. M. (1971). Linear regression models in finite population sampling theory. In *Foundations of statistical inference (Proc. Sympos., Univ. Waterloo, Waterloo, Ont., 1970)*. Rinehart and Winston of Canada, Toronto, Ont.: Holt, 259–279.

Scott, A. & Smith, T. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association* 64 830–840.

Searle, S. R., Casella, G. & McCulloch, C. E. (1992). *Variance components*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc. , A Wiley-Interscience Publication.

Smith, D. D. (2001). *Bayesian and minimum hellinger distance approaches to inference with applications*. Ph.D. thesis, Dept. of Statistics, University of Georgia.

Spiegelhalter, D., Best, N., Carlin, B. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal Of The Royal Statistical Society Series B* 64 583–639.

Spiegelhalter, D., Thomas, A., Best, N. & Gilks, W. (1997). *BUGS: Bayesian inference using Gibbs sampling, Version 0.6*. Biostatistics Unit. Cambridge: MRC.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2007). OpenBUGS user manual, version 3.0. 2. *MRC Biostatistics Unit, Cambridge* .

Stasny, E. A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: An example from the national crime survey. *Journal of the American Statistical Association* 86 296–303.

Stroud, T. (1991). Hierarchical baxes predictive means and variances with application to sample survey inference. *Communications in Statistics-Theory and Methods* 20 13–36.

Stroud, T. W. F. (1994). Bayesian analysis of binary survey data. *Canad. J. Statist.* 22 33–45.

Tamura, R. & Young, S. (1986). The incorporation of historical control information in tests of proportions: simulation study of Tarone's procedure. *Biometrics* 42 343–9.

Tamura, R. & Young, S. (1987). A stabilized moment estimator for the beta-binomial distribution. *Biometrics* 43 813–824.

Tang, R. (2002). *Fitting and Evaluating Certain Two-Level Hierarchical Models*. Ph.D. thesis, Dept. of Statistics, Harvard University.

Thomas, A., O'Hara, B., Ligges, U. & Sturtz, S. (2006). Making bugs open. *R News* 6 12–17. URL `http://cran.r-project.org/doc/Rnews/`.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* 81 82–86.

Tierney, L., Kass, R. E. & Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* 84 710–716.

Valliant, R., Dorfman, A. & Royall, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley.

Vermunt, J. & Magidson, J. (2005). *LATENT GOLD® 4.0*.

Wolter, K. M. (1985). *Introduction to variance estimation*. Springer Series in Statistics. New York: Springer-Verlag.

You, Y. & Rao, J. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics= Revue Canadienne de Statistique* 431.

You, Y. & Rao, J. N. K. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *J. Statist. Plann. Inference* 111 197–208.