Sociedad de Estadística e Investigación Operativa
Test (2006) *Vol.* **15**, *No.* 1, pp. 1–96

Mixed Model Prediction and Small Area Estimation

Jiming Jiang*
Department of Statistics
University of California, USA
P. Lahiri
Joint Program in Survey Methodology
University of Maryland, USA

Abstract

Over the last three decades, mixed models have been frequently used in a wide range of small area applications. Such models offer great flexibilities in combining information from various sources, and thus are well suited for solving most small area estimation problems. The present article reviews major research developments in the classical inferential approach for linear and generalized linear mixed models that are relevant to different issues concerning small area estimation and related problems.

Key Words: Benchmarking, borrowing strength, design-consistency, mean squared errors, empirical Bayes, EBLUP, generalized linear mixed models, higher order asymptotics, resampling methods, sample surveys, variance components.

AMS subject classification: 62C12, 62C25, 62G09, 62D05, 62F11, 62F15.

1 Introduction

The term *small domain or area* typically refers to a population for which reliable statistics of interest cannot be produced due to certain limitations of the available data. Examples of domains include a geographical region (e.g. a state, county, municipality, etc.), a demographic group (e.g. a specific age × sex × race group), a demographic group within a geographic region, etc. Some of the groundwork in small area estimation related to the population counts and disease mapping research has been done by the epidemiologists and the demographers. The history of small area statistics

*Correspondence to: Jiming Jiang. Department of Statistics, University of California, Davis, CA 95616, USA. E-mail: jiang@wald.ucdavis.edu

goes back to the eleventh century England and seventeenth century Canada (Brackstone, 1987). The use of maps to understand the prevalence of a disease for small areas has been used for a long time, see Marshall (1991).

There are various reasons for the scarcity of direct reliable data on the variables of interest for small areas. In a sample survey, a small area estimation problem could arise simply due to the sampling design that aims to provide reliable data for large areas and pays little or no attention to the small areas of interest. For example, in a statewide telephone survey of sample size $4,300$ in the state of Nebraska, U.S.A., only 14 observations are available to estimate the prevalence of alcohol abuse in Boone county, a small county in Nebraska. The problem is even more severe for direct survey estimation of the prevalence for white female in the age-group 25-44 in this county since only one observation is available from the survey. See Meza et al. (2003) for details. Oversampling is often employed in surveys in order to increase sample sizes for some domains but that leaves other domains with very few samples or even no sample since the total sample size is usually fixed by the survey budget. For example, in the Third National Health and Nutrition Examination Survey (NHANES III), conducted by the U.S. National Center for Health Statistics, certain minority groups are oversampled. This design strategy resulted in a very few samples for states that do not have a large population for these minority groups.

Various design strategies that incorporate factors influencing small area data quality have been suggested in the literature (see Rao, 2003, pp. 21-24 and the references therein). While these changes in a sampling design generally improve on the small area estimation, the problem still remains because of the budget and other practical constraints. In some situations, administrative data can be used to produce small area statistics. Such a statistic does not suffer from the sampling errors, but is often subject to measurement errors, resulting in poor quality of small area statistics. For example, law enforcement records are likely to underreport small area crime statistics since many crimes are not reported to the law enforcement authorities. Even the count for certain subgroups (e.g., illegal immigrants, certain minority groups, etc.) of the population compiled from the census could be of poor quality because of nonresponse and issues related to hard-to-find populations. In a disease mapping problem, small area problem arises simply because of the small population size.

Small area statistics are needed in regional planning, apportionment of congressional seats, and in the allocation of funds in many government pro-

grams and thus the importance of producing reliable small area statistics cannot be over-emphasized. For example, in both developed and developing countries, governmental policies increasingly demand income and poverty estimates for small areas. In fact, in the U.S. more than $130 billion of federal funds per year are allocated based on these estimates. In addition, states utilize these small area estimates to divide federal funds and their own funds to areas within the state. These funds cover a wide range of community necessities and services including education, public health, and numerous others. Therefore, there is a growing need to refine the manner in which these estimates are taken to provide an increased level of precision. In response to the growing need for these estimates, the Census Bureau formed a committee on Small Area Income and Poverty Estimates (SAIPE). This committee was created in the early 1990's with the goal of providing more timely and precise estimates than those from the data in a decennial census in accordance with the need for these estimates in order to allocate governmental funds. For example, the "Improving America's Schools Act" requires SAIPE estimates of poor school-age children (under 18) for counties as well as school districts in order to allocate more than $7 billion annually for the educationally disadvantaged students. SAIPE provides estimates of income and poverty at state, county, and district levels. For example, it provides estimates of all poor persons, poor children under the age of 5, poor children under the age of 18, poor children aged 5-17, and median income of households for states. See Citro et al. (1997) for details. In Philippines, poverty statistics are used to implement, monitor and evaluate different projects. Their national government also uses these statistics to determine specific provinces and municipalities that need more government funds. See Tabunda and Abanilla (1997) and Juan-Albacea (2004) for details. Mapping a disease incidence over different small areas are useful in allocation of government resources to various geographical areas and also to identify factors (such as existence of a nuclear reactor near a site) causing a disease.

In absence of adequate direct information, any improved estimation procedure calls for *implicit* or *explicit* models that combine information from related sources. In order to reduce the sampling errors in small area statistics derived from sample survey data, one can combine information from the sample survey, various administrative/census records and even previous surveys. If nonsampling errors are of concern, e.g., when administrative or census records are used to produce small area statistics, a relatively smaller

survey with better control of the nonsampling errors can be considered in conjuction with the larger survey or administrative records. In 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.

Once relevant sources of information are identified for a particular small area problem of interest, a decision needs to be made as to which method to apply to combine information from different sources. Various indirect methods that combine information using implicit models have been discussed in Rao (2003). Formal evaluation for such methods is problematic since the model involved is not spelled out. An explicit model is useful in small area estimation since this gives the users an idea of the data generation process and how different information sources are combined. These methods permit formal model building process, including model selection and model diagnostics, and provide a measure of uncertainty of the point estimator or predictor under a reasonable working model.

Mixed effects models, or mixed models, are particularly suitable for small area estimation because of its flexibility in effectively combining different sources of information and explaining different sources of errors. Mixed models typically incorporate area-specific random effects that explain the additional between area variations in the data not explained by the fixed effects part of the model. 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 variable(s).

In Section 2, we explain the rationale behind using mixed models in small area estimation and related problems by considering two well-known and widely used, yet simple, mixed models. In Section 3, we describe a general linear mixed model that includes the two models described in Section 2 as special cases. We describe the generalized linear mixed models (GLMM) in Section 4. This covers most, if not all, small area applications considered in the literature, including the general linear mixed models and spatial models.

There are primarily two different approaches to inference using mixed models: (i) the classical prediction approach like the empirical best prediction (EBP) approach; and (ii) hierarchical Bayesian approach. Although the empirical Bayes (EB) approach has also been used, the EB estimators

are identical to the EBP's, and the same inferential methods are applicable to both the EB and EBP's. Thus, in this paper, we shall use the terms EB and EBP's interchangeably. The hierarchical Bayes approach is straightforward in the sense that the relevant posterior distribution, once obtained, can be used for all inferential purposes. However, it requires a specification of a prior distribution on the hyperparameters or model parameters such as the regression coefficients and the variance components. The method requires checking the propriety of the posterior distribution of the parameter(s) of interest and the convergence of the computer intensive Markov Chain Monte Carlo (MCMC) method often used to approximate the posterior distribution. An early application of such an approach in small area estimation with finite population can be found in Ghosh and Lahiri (1992). This was extended to a general linear mixed model by Datta and Ghosh (1991). For a hierarchical Bayes approach under a generalized linear mixed model, see Malec et al. (1997) and Ghosh et al. (1998). We shall not review hierarchical Bayes approach in this review paper and refer the interested readers to Ghosh and Meeden (1997) and Rao (2003).

A large number of conferences and workshops on small area estimation were held. A list of conferences and their proceedings is given in Ghosh and Rao (1994) and Rao (2003). Several review papers including Rao (1986, 1999, 2001), Chaudhuri (1994), Ghosh and Rao (1994), Marker (1999), Pfeffermann (2002), Lahiri and Meza (2002) were published. Two books on small area estimation written by Mukhopadhyay (1998) and Rao (2003) are also available. All these activities helped the dissemination of small area research and the understanding of small area methodologies and applications. The purpose of this article is to provide a review of the classical prediction theory and related topics in mixed models with emphasize on small area estimation. Because of the great volume of research in mixed models, we cannot claim the completeness of the literature review on mixed models and refer the interested readers to the upcoming monograph by Jiang (2006b).

2 Area-specific versus unit specific mixed models

In the small area context, mixed models are generally classified into two broad classes based on the data availability on the response variable of interest: (i) area level models where information on the response variable is available only at the small area level; and (ii) unit level models where

information on the response variable is available at the unit or respondent level. In the following two subsections, we explain these two important types of mixed models with the help of examples. In the first case, area level auxiliary information is used. In the second case, auxiliary information can be used at the area and/or unit levels. Each of these two types of linear mixed models can be extended to the corresponding special case of the generalized linear mixed model, depending on the process that generates the response variable.

2.1 Area level mixed models

In order to estimate the per-capita income of small places with population size less than 1000, Fay and Herriot (1979) used the following two-level Bayesian model:

The Fay-Herriot model:

- Level 1 (sampling model): $y_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, D_i),\ i = 1, \cdots, m$;
- Level 2 (linking model): $\theta_i \overset{\text{ind}}{\sim} N(x_i'\beta, A),\ i = 1, \cdots, m$.

In the above model, Level 1 is used to account for the sampling variability of the regular survey estimates y_i of true small area means. Level 2 links the true small area means θ_i to a vector of p known auxiliary variables x_i, often obtained from various administrative and census records. The parameters β and A of the linking model are generally unknown and are estimated from the available data. In order to estimate the sampling variability D_i, Fay and Herriot (1979) employed generalized variance function (GVF; see Wolter, 1985) method that uses some external information in addition to the survey data. See Hinrichs (2003), and Gershunskaya and Lahiri (2005) for recent developments on variance estimation.

We note that the Fay-Herriot model can be viewed as an area level mixed regression model.

$$y_i \quad = \quad \theta_i + e_i = x_i'\beta + v_i + e_i,\ i = 1, \cdots, m,$$

where v_i's and e_i's are independent with $v_i \overset{i.i.d.}{\sim} N(0, A)$ and $e_i \overset{ind}{\sim} N(0, D_i)$. Note that the area-specific random effect v_i is used to relate the true per-capita income (θ_i) to the auxiliary variables (x_i) obtained from the census,

housing and Internal Revenue Service records. In other words, Fay and Herriot (1979) used random effects in order to capture the additional area-specific effects not explained by the area-specific auxiliary variables. This is achieved at the expense of an additional unknown variance component A to be estimated from the data. In contrast, the corresponding regression model without random effects fails to capture this additional area-specific variability. Using the U.S. census data, Fay and Herriot (1979) demonstrated that their EB estimator (also an empirical best linear unbiased predictor (EBLUP)) performed better than the direct survey estimator and a synthetic estimator used earlier by the U.S. Census Bureau.

Prior to Fay and Herriot (1979), mixed model applications in small area problem can be found in the baseball data example of Efron and Morris (1975) and the false alarm probability estimation example of Carter and Rolph (1974). In the baseball data example, $x_i'\beta = \mu$, i.e., θ_i were treated as exchangeable, and $D_i = 1$. We shall return to this example in Section 7. In both the examples, the well-known arc-sine transformation on the sample proportion was taken which justified the assumption of known sampling variance D_i of the transformed proportion. Like Fay and Herriot (1979), these two applications implement mixed models using the empirical Bayes method.

2.2 Unit level mixed models

To estimate areas planted with corn and soybeans for twelve counties (small areas) of North-Central Iowa, Battese et al. (1988) used a nested error regression model to combine information from a firm survey and satellite data that provide information on the number of pixels planted with corn and soybeans for each county. They used the following mixed model:

Nested error regression model:

$$y_{ij} = \beta_{0i} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + e_{ij},$$

where y_{ij} is the number of hectares of corn (soybeans) in the jth segment of the ith county as reported in the firm survey and x_{1ij} (x_{2ij}) is the number of *pixels* classified as corn (soybeans) in the jth segment of the ith county. A *pixel* (a term for *picture elements*) is the unit for which satellite information is recorded. In addition, the data also include $\bar{X}_{1i}$ ($\bar{X}_{2i}$), the mean number

of pixels per segment classified as corn (soybeans) for county i. In this example, the intercepts β_{0i} are area specific and are assumed to be *i.i.d.* with mean β_0 and variance σ_v^2. The pure errors e_{ij} are assumed to be *i.i.d.* with mean 0 and variance σ_e^2.

In this application, sample sizes for the counties are very small, ranging from 1 to 6, making the sample means highly variable. Also, the variance estimation by standard sample survey method is not possible for certain counties with just one observation. The random county specific intercept term is useful in capturing any additional area-specific variability not explained by the county specific covariates. In contrast, the usual fixed effects regression model that assumes equal intercept for all the counties fails to account for this additional county specific variability. The above nested error regression model can be extended by assuming exchangeability for the slopes, in addition to the intercepts. Such a model is called a random coefficient model, a model that has also been used in the analysis of longitudinal data (e.g., Diggle et al., 1996).

The two mixed models considered in this section are special cases of a general linear mixed model described in the next section.

3 Linear mixed models

The best way to understand a linear mixed model, also referred to as a mixed linear model in some earlier papers, is to first recall a linear regression model. The latter can be expressed as $y = X\beta + \epsilon$, where y is a vector of observations, X is a matrix of known covariates, β is a vector of unknown regression coefficients and ϵ is a vector of (unobservable random) errors. In this model, the regression coefficients are considered fixed. However, there are cases (e.g. the two cases mentioned in Section 2) in which it makes sense to assume that some of these coefficients are random. These cases typically occur when the observations are correlated. For example, in medical studies, repeated measures are often collected from the same individuals over time. It may be reasonable to assume that the observations for the same individual are correlated, especially if they are collected within a short time interval.

A general linear mixed model may be expressed as

$$y = X\beta + Z\alpha + \epsilon, \tag{3.1}$$

where y is a vector of observations, X is a matrix of known covariates, β is a vector of unknown regression coefficients, which are often called the fixed effects, Z is a known matrix, α is a vector of random effects and ϵ is a vector of errors. Note that both α and ϵ are unobservable. Compared with the linear regression model, it is clear that the difference is $Z\alpha$, which may take many different forms, and thus creates a rich class of models. The basic assumptions for (3.1) are that the random effects and errors have mean zero and finite variances. Typically, the covariance matrices $G = \mathrm{Var}(\alpha)$ and $R = \mathrm{Var}(\epsilon)$ involve some unknown dispersion parameters, or variance components. It is also assumed that α and ϵ are uncorrelated.

If, in addition, the normality assumption is made, the model is called a Gaussian linear mixed model, or simply Gaussian mixed model. A Gaussian mixed model may be defined with or without random effects. In the latter case, it is assumed that the (marginal) distribution of y is multivariate normal with mean $X\beta$ and covariance matrix V, which is specified up to a vector ϑ of variance components, i.e., $y \sim N(X\beta, V)$, where $V = V(\vartheta)$. Such a model is also called a *marginal model*.

3.1 Estimation of variance components

An important problem in the analysis of a linear mixed models is the estimation of the variance components. In many cases, such as in quantitative genetics, the variance components are of main interest (e.g., Shaw, 1987). In some other cases, the variance components themselves are not of main interest, but they need to be estimated in order to assess the variability of estimators or predictors of other quantities, such as the fixed and random effects, which are of main interest. Such examples include the analysis of longitudinal data (e.g., Diggle et al., 1996) and small area estimation (see Rao, 2003). Some of the earlier methods in mixed model analysis did not require the normality assumption. These include the analysis of variance (ANOVA) method, or Henderson's methods (Henderson, 1953), and minimum norm quadratic unbiased estimation (MINQUE) method, proposed by C. R. Rao (e.g., Rao, 1972). However, the ANOVA method is known to produce inefficient estimators of the variance components for unbalanced data. In addition, this method could produce estimates falling outside the parameter space, such as negative estimates of variances. The MINQUE method, on the other hand, depends on some initial values of the variance components. If it is run iteratively, each time using the current values to

update the initial values, the procedure, called I-MINQUE, is known to produce estimators that are identical to the REML estimators, if they belong to the parameter space (e.g., Searle et al., 1992, Section 11.3). However, the latter property is not guaranteed.

If normality is assumed, the efficient estimators of the variance components are the maximum likelihood estimators (MLE). However, this method was not considered seriously in linear mixed models until Hartley and Rao (1967). The main reason was that, unlike the ANOVA estimator, the MLE under a linear mixed model was not easy to handle computationally in the earlier days. There was also an issue regarding the asymptotic behavior of the MLE, because, unlike the traditional *i.i.d.* case, the observations are correlated under a linear mixed model. Both the computational and asymptotic issues were addressed by the Hartley-Rao paper. Goldstein (1986) proposed an iterative generalized least squares method for estimating the fixed effects and variance components. The method is popular and easy to implement, and it produces the MLE upon convergence. Asymptotic properties of the MLE were further studied by Miller (1977) for a wider class of models.

The MLE of the variance components are, in general, biased. Such a bias does not vanish as the sample size increases, if the number of the fixed effects is proportional to the sample size. In fact, in the latter case the MLE is inconsistent, as the Neyman-Scott example showed (Neyman and Scott, 1948). Furthermore, in some cases such as animal genetics the parameters of main interest are the variance components, the fixed effects being treated as nuisance parameters. It would be nice to have a method that can focus on the variance components without having to simultaneously estimate the nuisance parameters. Thompson (1962) proposed a method, which was later extended by Patterson and Thompson (1971), known as restricted or residual maximum likelihood, or REML in either case. The idea is to consider a transformed data that is orthogonal to the design matrix of the fixed effects, that is, X. In other words, each component of the transformed data vector is an *error contrast*. To formulate the REML procedure, let the dimensions of y and β be n and p, respectively. Without loss of generality, assume that $\text{rank}(X) = p$. Let A be a $n\times(n-p)$ matrix of full rank such that $A'X = 0$. The REML estimators of the variance components are simply the MLE based on $z = A'y$. It is easy to show that the REML estimators do not depend on the choice of A. Furthermore, several authors have argued that there is no loss of information by using REML in estimating the variance

components (e.g., Patterson and Thompson (1971), Harville (1977), Jiang (1996)). Different derivations of REML were given by Harville (1974), Barndorff-Nielsen (1983), Verbyla (1990), Heyde (1994), and Jiang (1996). In addition, several review articles on REML are available, see Harville (1977), Khuri and Sahai (1985), Robinson (1987) and Speed (1997).

The ML and REML methods are developed under the normality assumption. However, the normality assumption is likely to be violated in real life. For example, Lange and Ryan (1989) gave several examples showing that non-normality of the random effects is, indeed, encountered in practice. Due to such concerns, some researchers have taken a quasi-likelihood approach. The idea is to use the Gaussian ML or REML estimators in nonnormal situations. See Richardson and Welsh (1994), Jiang (1996, 1997b), Heyde (1994, 1997), among others. Throughout this review, these estimators will also be called ML and REML estimators, even if normality does not hold. Jiang (1996, 1997b) established consistency and asymptotic normality of the REML estimators in nonnormal situations, and gave necessary and sufficient conditions for similar asymptotic properties for the MLE. These conditions are in terms of the rate at which p goes to infinity with n. In particular, Jiang (1996) derived the asymptotic covariance matrix (ACM) of the REML estimator of ϑ, the vector of variance components. See Jiang (1998a) for an ML analogue.

The ACM is important to various inferences about the model, including interval estimation and hypothesis testing. Unfortunately, the ACM under nonnormality involves parameters other than the variance components, for example, the third and fourth moments of the random effects. Note that standard procedures such as ML and REML do not produce estimators of these additional parameters. Jiang (2006c) proposed a new method for estimating the ACM. The author noted that a typical element of ACM can be decomposed as $S_1 + S_2$, where S_1 is in the form of $\mathrm{E}(\cdots)$ and $\cdots$ is a sum of terms involving the variance components and products of centralized data (i.e., data subtract its mean); and S_2 depends only on the variance components. If the expectation in S_1 is taken, the higher moments (up to 4th) will occur, which is something the author wanted to avoid. Instead, he proposed to estimate S_1 by an observed S_1, that is, by removing the expectation sign and relacing the parameters involved in $\cdots$ by, say, their REML estimators. On the other hand, S_2 can be estimated by replacing the variance components by their REML estimators, say. The method is thus known as *partially observed information*. See Jiang (2006c) for details.

3.2 Prediction of random effects

The prediction of random effects, or mixed effects in a more general context, has a long history, dating back to C. R. Henderson's early work on animal breeding (e.g., Henderson, 1948). The best known method for the prediction of mixed effects is the best linear unbiased prediction, or BLUP. See Robinson (1991) for an overview of the BLUP with examples and applications.

A mixed effect may be expressed as $\eta = b'\beta + a'\alpha$, where a, b are known vectors. If the fixed effects and variance components are both known, the best predictor (BP) for η, under the normality assumption, is given by $\tilde{\eta}$, where $\tilde{\eta}$ is η with α replace by $\mathrm{E}(\alpha|y)$, i.e., $\tilde{\alpha} = GZ'V^{-1}(y - X\beta)$, $V = \mathrm{Var}(y) = V(\vartheta)$ and ϑ is the vector of variance components. In absence of the normality assumption, $\tilde{\eta}$ is the best linear predictor of η in the sense that it minimizes the mean squared prediction error (MSPE) of a predictor that is linear in y (e.g., Searle et al., 1992, §7.3). Of course, β is unknown in practice. It is then customary to replace β by $\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$, which is the MLE of β under normality, provided that ϑ is known. The result is BLUP, in other words, the BLUP of η is given by $\tilde{\eta}$ with β replaced by $\tilde{\beta}$.

The original derivation of the BLUP was given by Henderson (1950) where he proposed to find the "maximum likelihood estimates" of the random effects. Of course, these are not the MLE in the usual sense, because the random effects are different from fixed parameters. Later, Henderson (1973) showed that the BLUP is, indeed, the best linear unbiased predictor, that is, (i) it is linear in y; (ii) its expected value is equal to that of η; and (iii) it minimizes the MSPE among all predictors satisfying (i) and (ii). Different derivations of the BLUP were also given by Harville (1990) and Jiang (1997a), among others. In a related work, Pfeffermann (1984) developed a simple method of obtaining the BLUP of fixed and stochastic (random) coefficients. The method can be implemented with any regression routine. It also produces predictors of the random effects in nonsampled areas, in the same run, even when the random effects are correlated.

The expression of the BLUP involves ϑ, the vector of variance components, which is typically unknown in practice. It is customary to replace ϑ by a consistent estimator, $\hat{\vartheta}$. The resulting predictor is often called empirical BLUP, or EBLUP, denoted by $\hat{\eta}$. Kackar and Harville (1981) showed that, if $\hat{\vartheta}$ is an even and translation invariant estimator and the data is

symmetrically distributed, EBLUP remains unbiased. Some of the well-known estimators of ϑ, including ANOVA, ML and REML estimators, are even and translation invariant. In their arguments, however, Kackar and Harville assumed the existence of $\mathrm{E}(\hat{\eta})$, which is not obvious. The existence of the expected value was later proved by Jiang (2000a).

Harville (1991) considered a one-way random effects model and showed that, in this case, the EBLUP is identical to an empirical Bayes (EB) estimator. He further noted that while much of the work on EB was carried out by professional statisticians in relatively simple cases, such as the one-way random effects model, EBLUP was used by practitioners under relatively complex models. One of the areas in which EBLUP has been extensively used and studied is small area estimation.

To illustrate the EBLUP procedure, consider the Fay-Herriot model introduced in Subsection 2.1. The best predictor (BP) of θ_i is given by:

$$\hat{\theta}_i(y_i;\phi) \;=\; (1-B_i)y_i + B_i x_i'\beta,$$

where $B_i = D_i/(D_i+A)$, and $\phi = (\beta, A)'$. Replacing β by the weighted least square estimator given by

$$\hat{\beta}(A) \;=\; \left(\sum_{i=1}^{m}\frac{1}{D_i+A}x_i x_i'\right)^{-1}\left(\sum_{i=1}^{m}\frac{1}{D_i+A}x_i y_i\right),$$

we obtain the following BLUP of θ_i:

$$\hat{\theta}_i(y_i;A) \;=\; (1-B_i)y_i + B_i x_i'\hat{\beta}(A).$$

Replacing A by $\hat{A}$, an estimator of A, we get $\hat{\theta}_i(y_i;\hat{A})$, an EBLUP of θ_i.

3.3 Assessing uncertainty of EBLUP

While EBLUP is fairly easy to obtain, estimation of its mean squared prediction error (MSPE) is a challenging problem. On the other hand, estimation of the MSPE of EBLUP is of significant practical interest. Kackar and Harville (1984) provided an approximation to the MSPE of EBLUP under the Gaussian linear mixed model (3.1), taking into account the variability of $\hat{\vartheta}$, and proposed an estimator of $\mathrm{MSPE}(\hat{\eta})$ based on this approximation. But the approximation is somewhat heuristic, and the accuracy

of the approximation and the associated MSPE estimator was not studied. In a pioneering paper, Prasad and Rao (1990) studied the accuracy of a second-order approximation to MSPE($\hat{\eta}$) for a longitudinal mixed linear model that covers both the Fay-Herriot model and the nested error regression model introduced in Section 2. Datta and Lahiri (2000) considered the same longitudinal model but extended the Prasad-Rao results in order to cover a wide range of variance component estimators, including the ML and REML estimators. Das et al. (2004) considered a more general linear mixed models but their asymptotic set–up differs from that of Datta and Lahiri (2000). For example, for Gaussian mixed ANOVA models with REML estimation of variance components, Das et al. (2004) obtained: MSPE($\hat{\eta}$) $= g_1(\vartheta) + g_2(\vartheta) + g_3(\vartheta) + o(d_*^{-2})$, where

$$\begin{aligned} g_1(\vartheta) &= a'(G - GZ'V^{-1}ZG)a, \\ g_2(\vartheta) &= (b - X'V^{-1}ZGa)'(X'V^{-1}X)^{-1}(b - X'V^{-1}ZGa), \\ g_3(\vartheta) &= \mathrm{tr}\left[\{(\partial/\partial\vartheta')V^{-1}ZGa\}'V\{(\partial/\partial\vartheta')V^{-1}ZGa\}H^{-1}\right], \end{aligned}$$

$H = \mathrm{E}(\partial^2 l_{\mathrm{R}}/\partial\vartheta\partial\vartheta')$, $d_* = \min_{1\le i\le s} d_i$ with $d_i = \|Z_i'PZ_i\|_2$, l_{R} is the restricted log-likelihood, Z_i is the design matrix for the ith random effect factor, and

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}.$$

Based on the approximation, the authors obtained an estimator of MSPE($\hat{\eta}$) which is unbiased up to the second order. More specifically, an estimator $\widehat{\mathrm{MSPE}}(\hat{\eta})$ was obtained such that $\mathrm{E}\{\widehat{\mathrm{MSPE}}(\hat{\eta})\} = \mathrm{MSPE}(\hat{\eta}) + o(d_*^{-2})$.

Alternatively, Jiang, Lahiri, and Wan (2002) proposed a jackknife method which led to second-order approximation and estimation of the MSPE of EBLUP in the case of longitudinal linear mixed models. Their method also applies to estimation of MSPE of EBP under longitudinal GLMMs (see Subsection 4.4). As noted by Bell (2001), the jackknife estimator of Jiang, Lahiri, and Wan (2002) could take negative values. Chen and Lahiri (2002), however, noticed that this is not a severe problem and can be easily rectified. They also considered a weighted version of the Jiang-Lahiri-Wan jackknife which improved the efficiency in certain situations. Parametric bootstrap methods have also been considered. See Butar (1997), Butar and Lahiri (2003), Pfeffermann and Tiller (2002), Lahiri (2003a) and Glickman and Pfeffermann (2004). Singh et al. (1998) compared the Bayesian and non-Bayesian methods. For a conditional approach, see Fuller (1990) and Booth and Hobert (1998).

3.4 Example

It is useful to revisit the Fay-Herriot model. Lahiri (2003b) compared different measures of uncertainty under this model. By the Kackar-Harville identity (Kackar and Harville, 1984), we have

$$\text{MSPE}[\hat{\theta}_i(y_i;\hat{A})] \;=\; g_{1i}(A) + g_{2i}(A) + G_{3i}(A), \tag{3.2}$$

where

$$\begin{aligned} g_{1i}(A) &= \frac{AD_i}{A+D_i}, \\ g_{2i}(A) &= \frac{D_i^2}{(A+D_i)^2} x_i' \left(\sum_{j=1}^{m} \frac{1}{A+D_j} x_j x_j' \right)^{-1} x_i, \\ G_{3i}(A) &= \mathrm{E}\{\hat{\theta}_i(y_i;\hat{A}) - \hat{\theta}_i(y_i;A)\}^2. \end{aligned}$$

In the above, $g_{1i}(A) + g_{2i}(A)$ is the MSPE of the BLUP and $G_{3i}(A)$ is the additional uncertainty due to the estimation of the variance component A.

A naive MSPE estimator is obtained by estimating the MSPE of the BLUP and is given by:

$$\text{mspe}_i^{\text{N}} \;=\; g_{1i}(\hat{A}) + g_{2i}(\hat{A}).$$

Intuitively, this naive MSPE estimator is likely to underestimate the true MSPE since it fails to incorporate the additional uncertainty due to the estimation of A. In fact, Prasad and Rao (1990) showed that the order of this underestimation is $O(m^{-1})$ under certain regularity conditions. Interestingly, the naive MSPE estimator even underestimates the true MSPE of the BLUP, the order of underestimation being $O(m^{-1})$.

Prasad and Rao (1990) proposed the following MSPE estimator when A is estimated by the usual method of moments:

$$\text{mspe}_i^{\text{PR}} \;=\; g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2g_{3i}(\hat{A}),$$

where $g_{3i} = \{2D_i^2/m^2(\hat{A}+D_i)^3\} \sum_{j=1}^{m} (\hat{A}+D_j)^2$. Under the same regularity conditions, the bias of $\text{mspe}_i^{\text{PR}}$ is of the order $o(m^{-1})$.

For the Fay-Herriot model, the jackknife MSPE estimator proposed by Jiang et al. (2002) reduces to:

$$\begin{aligned}\text{mspe}_i^{\text{JLW}} &= g_{1i}(\hat{A}) - \frac{m-1}{m}\sum_{u=1}^{m}\left\{g_{1i}(\hat{A}_{-u}) - g_{1i}(\hat{A})\right\} \\ &\quad + \frac{m-1}{m}\sum_{u=1}^{m}\left\{\tilde{\theta}_i(y_i;\hat{A}_{-u}) - \hat{\theta}_i(y_i;\hat{A})\right\}^2,\end{aligned}$$

where

$$\begin{aligned}\tilde{\theta}_i(y_i;\hat{A}_{-u}) &= \frac{D_i}{\hat{A}_{-u}+D_i}x_i'\hat{\beta}_{-u} + \frac{\hat{A}_{-u}}{\hat{A}_{-u}+D_i}y_i, \\ \hat{\beta}_{-u} &= \left(\sum_{j\neq u}\frac{D_j}{\hat{A}_{-u}+D_j}x_jx_j'\right)^{-1}\sum_{j\neq u}\frac{D_j}{\hat{A}_{-u}+D_j}x_jy_j.\end{aligned}$$

For the Fay-Herriot case, the weighted jackknife MSPE estimator suggested by Chen and Lahiri (2003a) is given by:

$$\begin{aligned}\text{mspe}_i^{\text{CL}} &= g_{1i}(\hat{A}) + g_{2i}(\hat{A}) \\ &\quad - \sum_{u=1}^{m} w_u\left[g_{1i}(\hat{A}_{-u}) + g_{2i}(\hat{A}_{-u}) - \{g_{1i}(\hat{A}) + g_{2i}(\hat{A})\}\right] \\ &\quad + \sum_{u=1}^{m} w_u\left\{\hat{\theta}_i(y_i;\hat{A}_{-u}) - \hat{\theta}_i(y_i;\hat{A})\right\}^2.\end{aligned}$$

Chen and Lahiri (2003a) suggested two choices of weights: $w_u = (m-1)/m$ and $w_u = x_u'\left(\sum_{j=1}^{m} x_jx_j'\right)x_u$. Note that $\text{mspe}_i^{\text{CL}}$ is different from $\text{mspe}_i^{\text{JLW}}$ in two respects. First, Chen and Lahiri (2003a) used more exact calculations by exploiting the Kackar-Harville identity which is valid under the normality assumption. Secondly, the method also adjusts $g_{2i}(\hat{A})$ for bias. Although in the standard second-order asymptotic sense this adjustment is not needed, we may not ignore this bias correction when the relative contribution from $g_{2i}(A)$ is significant (see Chen and Lahiri, 2003b).

Butar and Lahiri (2003) proposed the following parametric bootstrap MSPE estimator:

$$\begin{aligned}\text{mspe}_i^{\text{BL}} =& g_{1i}(\hat{A}) + g_{2i}(\hat{A}) - E_\star\left[g_{1i}(\hat{A}^\star) + g_{2i}(\hat{A}^\star) - \{g_{1i}(\hat{A}) + g_{2i}(\hat{A})\}\right] \\ &+ \text{E}_\star\left\{\hat{\theta}_i(y_i;\hat{A}^\star) - \hat{\theta}_i(y_i;\hat{A})\right\}^2.\end{aligned}$$

In the above, $\mathrm{E}_\star$ is the bootstrap expectation, i.e., the expectation with respect to the Fay-Herriot model with β and A replaced by $\hat{\beta}$ and $\hat{A}$ respectively. We obtain $\hat{A}^\star$ using the formula for $\hat{A}$ with the original sample replaced by the bootstrap sample. In practice, Monte Carlo methods are employed to approximate the bootstrap expectations. Note that because of the bias correction term, both the jackknife and parametric bootstrap methods could produce negative estimates. This was first observed by Bell (2001) in the context of jackknife MSPE estimator. But, Chen and Lahiri (2005) noted that this can be easily corrected. They recommended the following MSPE estimator in case $\mathrm{mspe}_i^{\mathrm{CL}}$ yields a negative value:

$$\mathrm{mse}_i^{\mathrm{ACL}} \;=\; g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + \frac{D_i^2}{(\hat{A}+D_i)^3} v_{\mathrm{WJ}} + \frac{D_i^2}{(\hat{A}+D_i)^4} r_i^2 v_{\mathrm{WJ}}.$$

where $r_i = y_i - x_i'\hat{\beta}$, the residual for the ith area, and $v_{\mathrm{WJ}} = \sum_{u=1}^m w_u(\hat{A}_{-u} - \hat{A})^2$, a weighted jackknife variance estimator of $\hat{A}$. Similar corrections can be made to $\mathrm{mspe}_i^{\mathrm{JLW}}$ and $\mathrm{mspe}_i^{\mathrm{BL}}$. All the MSPE estimators, except the naive one, has the same second-order unbiasedness property under the same regularity conditions. However, in different simulation studies, the performances of different MSPE estimators varied, depending on several factors such as the value of m, D_i/A, variations in D_i/A, presence of a outlier in D_i/A, leverage $h_i = x_i'(X^TX)^{-1}x_i$, and the method of estimation of A.

Among the MSPE estimators discussed, the derivation of all MSPE estimators other than the Jiang-Lahiri-Wan jackknife estimator requires the assumption of normality. Although the Prasad-Rao MSPE estimator was originally derived under the assumption of normality, Lahiri and Rao (1995) showed that it is quite robust under non-normality of the random effects. An unpublished work by Chen et al. (2005) suggests that one needs information on the kurtosis of the sampling errors under a purely non-normal Fay-Herriot model when A is estimated by the method of moments. Estimation of kurtosis for the model error may be also needed when A is estimated by a different method such as the estimator proposed by Fay and Herriot (1979). For a special case of the Fay-Herriot model when $D_i = D$ and $x_i'\beta = \mu$ $(i = 1, \cdots, m)$, Butar and Lahiri (2003) showed that their parametric bootstrap MSPE estimator is identical to a measure of uncertainty proposed by Morris (1983a) up to the order $O(m^{-1})$ if an unbiased estimator of $B = D/(A+D)$ is chosen in the EBLUP formula. This is also true for the Chen-Lahiri jackknife MSPE estimator. Thus, the parametric bootstrap and the Chen-Lahiri jackknife MSPE estimators are close to a

Bayesian solution since Morris (1983a) obtained his uncertainty measure by approximating the posterior variance using flat or uniform priors on μ and B. In this context, see Datta et al. (2005). For the above model with non-normality and the standard untruncated unbiased quadratic estimator of A, Lahiri (1995) appoximated $\text{mspe}_i^{\text{JLW}}$ up to the order $O(m^{-1})$ and obtained the following result:

$$\begin{aligned}\text{mspe}_i^{\text{JLW}} &\doteq g_1(\hat{A}) + g_2(\hat{A}) + \frac{D^2}{m(\hat{A}+D)}(b_2-1) \\ &\quad + \frac{D^2}{m(\hat{A}+D)^2}(b_2-1)r_i^2 - \frac{2D^2}{m(\hat{A}+D)^{3/2}}\sqrt{b_1}r_i,\end{aligned}$$

where $b_1 = m_3^2/(\hat{A}+D)^3$, $b_2 = m_4/(\hat{A}+D)^2$ and $r_i = y_i - \bar{y}$. Here b_1 and b_2 can be viewed as estimated skewness and kurtosis for the marginal distribution of y_i's. Under normality, $b_1 \approx 0$ and $b_2 \approx 3$ and so in this case $\text{mspe}_i^{\text{JLW}}$ reduces to

$$\text{mspe}_i^{\text{JLW}} \doteq g_1(\hat{A}) + g_2(\hat{A}) + \frac{2D^2}{m(\hat{A}+D)} + \frac{2D^2}{m(\hat{A}+D)^2}r_i^2.$$

It is reasonable to expect that, in this case, $\text{mspe}_i^{\text{JLW}}$ is identical to $\text{mspe}_i^{\text{CL}}$ and $\text{mspe}_i^{\text{BL}}$ up to the order $O(m^{-1})$.

Furthermore, we may define the first level (or conditional) mean squared prediction error as: $\text{CMSPE}(\hat{\theta}_i) = \text{E}\{(\hat{\theta}_i - \theta_i)^2]|\theta_i\}$, where the expectation is taken over the first level of the model, i.e., conditioning on the θ_i's. Rivest and Belmonte (1999) proposed an unbiased estimator of CMSPE. Hwang and Rao(1987; private communication) obtained a similar unbiased estimator earlier and, using a Monte Carlo simulation study, showed that the Prasad-Rao MSPE estimator is more stable than such an unbiased estimator of CMSPE. Interestingly, their simulation results showed that the Prasad-Rao MSPE estimator tracks the CMSPE very well even under a moderate deviation from the Level 2 model. However, the Prasad-Rao MSPE estimator could perform poorly compared to the CMSPE unbiased estimator for an outlying small area. This may be due to the fact that the Prasad-Rao MSPE estimator is not area-specific in terms of the main variable. The jackknife and the parametric bootstrap MSPE estimators are likely to perform better than the Prasad-Rao estimator in this situation as both of them are area-specific with respect to the main variable, i.e., y_i.

However, this conjecture remains to be validated through a Monte Carlo simulation study.

In the above discussions, we assume that D_i's are known as in Fay and Herriot (1979). However, sometimes it is reasonable (e.g., when simple random sampling is used to select sample in each area) to assume that $D_i = \sigma_e^2/n_i$, where n_i is the sample size for the ith area and σ_e^2 is the within area variability. In the above discussions, it is assumed that n_i's are all bounded and m is large. However, in some small area situations, both n_i and m could be large (perhaps, m is much larger than n_i). In such a situation, the usual second-order MSPE estimators may not be very meaningful and one may instead decide to keep all terms of different orders, such as $O(m^{-1}n_i^{-\alpha})$. This requires a different type of asymptotics and is likely to produce new MSPE estimators. When σ_e^2 is unknown and information on the within area variability for each area is given, one can apply results for the nested area regression model.

4 Generalized linear mixed models

For the most part, linear mixed models have been used in situations where the observations are continuous. However, in many cases the observations are discrete or categorical. McCullagh and Nelder (1989) proposed an extension of linear models, called generalized linear models, or GLMs. They noted that the key elements of a classical linear model, that is, a linear regression model, are that (i) the observations are independent; (ii) the mean of the observation is a linear function of some covariates; and (iii) the variance of the observation is a constant. The extension to GLM consists of modification of (ii) and (iii) above, by that (ii)' the mean of the observation is associated with a linear function of some covariates through a link function; and (iii)' the variance of the observation is a function of the mean. GLM covers a variety of models that include *normal*, *binomial*, *Poisson* and *multinomial* as special cases.

The GLM is, however, not general enough to cover the case when the observations are dependent, a case often encountered in practice. For example, if $y_{i1}, \ldots, y_{i10}$ indicate whether or not the ith individual (person) visited a doctor during each of the past 10 years, that is, $y_{ij} = 1$ if the ith individual visited a doctor in the jth year, and $y_{ij} = 0$ otherwise, then the responses from the same individual are likely to be correlated. On the

other hand, the responses are binary instead of continuous. As mentioned earlier, linear mixed models do not apply to such cases. Therefore, we need an extension of the linear mixed model that applies to correlated responses that are discrete or categorical.

4.1 From linear mixed model to GLMM

To motivate the extension, let us first consider an alternative expression of the Gaussian mixed model. Suppose that, given a vector of random effects, α, the observations $y_1, \ldots, y_n$ are (conditionally) independent such that $y_i|\alpha \sim N(x_i'\beta + z_i'\alpha, \tau^2)$, where x_i and z_i are known vectors, β is a vector of regression coefficients, and τ^2 is an unknown variance. Furthermore, suppose that α is multivariate normal with mean 0 and covariance matrix G, which depends on a vector ϑ of unknown variance components. Let X and Z be the matrices whose ith rows are x_i' and z_i', respectively. It is easy to see that this leads to the linear mixed model (3.1) with normality and $R = \tau^2 I$.

The key elements in the above that define a Gaussian mixed model are: (i) conditional independence given the random effects with a conditional distribution; (ii) the distribution of the random effects. These are also the essential parts of a generalized linear mixed model, or GLMM. Suppose that given a vector of random effects, α, the responses, $y_1, \ldots, y_n$ are conditionally independent such that the conditional distribution of y_i given α is a member of the exponential family with pdf

$$f_i(y_i|\alpha) \quad = \quad \exp\left\{\frac{y_i\xi_i - b(\xi_i)}{a_i(\phi)} + c_i(y_i, \phi)\right\}, \tag{4.1}$$

where $b(\cdot)$, $a_i(\cdot)$, $c_i(\cdot,\cdot)$ are known functions, and ϕ is a dispersion parameter which may or may not be known. Here ξ_i is associated with $\mu_i = \mathrm{E}(y_i|\alpha)$, which, in turn, is associated with a linear predictor

$$\eta_i \quad = \quad x_i'\beta + z_i'\alpha, \tag{4.2}$$

through a known link function $g(\cdot)$ such that

$$g(\mu_i) \quad = \quad \eta_i. \tag{4.3}$$

Here x_i and z_i are known vectors, and β is a vector of unknown parameters (the fixed effects). Furthermore, it is assumed that $\alpha \sim N(0, G)$, where

the covariance matrix G may depend on a vector ϑ of unknown variance components. Note that, according to the properties of exponential family, one has $b'(\xi_i) = \mu_i$. In particular, under the so-called *canonical* link, one has $\xi_i = \eta_i$, that is, $g = h^{-1}$, where $h(\cdot) = b'(\cdot)$.

One of the earlier examples of GLMM was given by McCullagh and Nelder (1989, Section 14.5) involving some salamander mating experiments. The GLMM models have since received considerable attention because of their usefulness in various fields, including biology, medical research and surveys. See, for example, Breslow and Clayton (1993), Lee and Nelder (1996), Christiansen and Morris (1997), Malec et al. (1997), Ghosh et al. (1998), Lahiri and Maiti (2002) and Meza (2003).

Despite the usefulness of these models, inference about GLMM has encountered difficulties. This is because, unlike linear mixed models, the likelihood function under a GLMM typically does not have a closed-form expression (with, of course, the exception of the normal case). In fact, such a likelihood may involve high-dimensional integrals which cannot be evaluated analytically. To understand the computational difficulties, consider the following simple example. Suppose that, given the random effects $u_1, \ldots, u_{m_1}$ and $v_1, \ldots, v_{m_2}$, binary responses y_{ij}, $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$ are conditionally independent such that, with $p_{ij} = \mathrm{P}(y_{ij} = 1|u, v)$, $\mathrm{logit}(p_{ij}) = \mu + u_i + v_j$, where μ is an unknown parameter, $u = (u_i)_{1 \le i \le m_1}$ and $v = (v_j)_{1 \le j \le m_2}$. Furthermore, the random effects $u_1, \ldots, u_{m_1}$ and $v_1, \ldots, v_{m_2}$ are independent such that $u_i \sim N(0, \sigma_1^2)$, $v_j \sim N(0, \sigma_2^2)$, where σ_1^2 and σ_2^2 are unknown. The unknown parameters involved in this model are $\psi = (\mu, \sigma_1^2, \sigma_2^2)'$. It can be shown that the likelihood function under this model for estimating ψ can be expressed as

$$
\begin{aligned}
& c - \frac{m_1}{2}\log(\sigma_1^2) - \frac{m_2}{2}\log(\sigma_2^2) + \mu y_{\cdot\cdot} \\
& + \log \int \cdots \int \left[\prod_{i=1}^{m_1} \prod_{j=1}^{m_2} \{1 + \exp(\mu + u_i + v_j)\}^{-1} \right] \\
& \times \exp\left(\sum_{i=1}^{m_1} u_i y_{i\cdot} + \sum_{j=1}^{m_2} v_j y_{\cdot j} - \frac{1}{2\sigma_1^2}\sum_{i=1}^{m_1} u_i^2 - \frac{1}{2\sigma_2^2}\sum_{j=1}^{m_2} v_j^2 \right) \\
& \times du_1 \cdots du_{m_1} dv_1 \cdots dv_{m_2},
\end{aligned}
$$

where c is a constant, $y_{\cdot\cdot} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} y_{ij}$, $y_{i\cdot} = \sum_{j=1}^{m_2} y_{ij}$ and $y_{\cdot j} = \sum_{i=1}^{m_1} y_{ij}$. The multi-dimensional integral involved here has no closed-form

expression, and it cannot be further simplified. Moreover, such an integral is difficult to evaluate even numerically. For example, if $m_1 = m_2 = 40$, the dimension of the integral will be 80. To make it even worse, the integrand involves a product of 1600 terms with each term less than one. Such a product is numerically zero, making it difficult to be evaluated with Monte Carlo method. To overcome the computational difficulties, two main approaches to inference about GLMM have been proposed. The first approach focuses on developing computational method for the maximum likelihood. The second one tries to avoid the computational difficulties of the likelihood-based inference by considering approximate or other alternative methods.

4.2 Likelihood-based inference

For relatively simple models, the likelihood function may be evaluated by numerical integration techniques. See, for example, Hinde (1982), Crouch and Spiegelman (1990). However, numerical integration is generally intractable in GLMM if the dimension of integrals involved is greater than two. Alternatively, the integrals may be evaluated by Monte Carlo methods. It should be pointed out that, for problems involving irreducibly high-dimensional integrals, naive Monte Carlo usually does not work. For example, the high-dimensional integral in the example above cannot be evaluated by a naive Monte Carlo method. (This is because a product of 1,600 terms with each term less than one is numerically zero, thus, an *i.i.d.* sum of such terms will not yield anything but zero without a huge simulation size!)

McCulloch (1994) proposed a Monte Carlo EM algorithm, in which the E-step was implemented by a Gibbs Markov chain. Later, McCulloch (1997) improved his earlier algorithm by replacing the Gibbs sampler with a Metropolis-Hastings algorithm to fit more general models. Subsequently, Booth and Hobert (1999) proposed two new Monte Carlo EM algorithms. Unlike McCulloch (1994, 1997), Booth and Hobert used *i.i.d.* sampling to construct Monte Carlo approximation at the E-step. More specifically, the authors proposed two methods to generate the Monte Carlo samples. The first method used rejection sampling from the conditional distribution of the random effects given the data; the second used a multivariate-t importance sampling. They pointed out that the two methods are complementary in that rejection sampling is more efficient in small sample situation while importance sampling works better in large sample case. Furthermore, the

authors proposed a rule that automatically increases the Monte Carlo sample size as the algorithm proceeds, whenever necessary. They showed that the new algorithms have the following advantages over the Markov chain methods. First, the assessment of the Monte Carlo errors is straightforward when *i.i.d.* samples are used. Such an assessment is critical for the automated method. A related theoretical advantage is that conditions for the central limit theorem in the *i.i.d.* case is much easier to verify than under a Markov chain. In terms of the computational speed, they used an example to show that the rejection and importance sampling methods are about 2.5 times and 30 times faster, respectively, than the Metropolis-Hastings sampling method of McCulloch (1997).

In a related development, Zeger and Karim (1991) used Gibbs sampling to approximate the Bayesian posterior means of the parameters under a GLMM. Also see Malec et al. (1997), Ghosh et al. (1998). As noted by Zeger and Karim, their posterior distribution of the parameters reduces to the likelihood function if a flat prior is used. However, such a prior may lead to an improper posterior (e.g. Natarajan and McCulloch, 1995).

Recently, Song et al. (2005) proposed a new method, which they called maximization by parts. Their objective was to overcome some of the computational difficulties in the maximum likelihood estimation. Write the log-likelihood function as $l(\psi) = l_w(\psi) + l_e(\psi)$, where ψ is the vector of parameters. The initial estimator, $\hat{\psi}_1$, is a solution to $\dot{l}_w(\psi) = 0$. Then, use the equation $\dot{l}_w(\psi) = -\dot{l}_e(\hat{\psi}_1)$ to update the estimator, where $\dot{l}$ represents the vector of first derivatives (and, similarly, $\ddot{l}$, the matrix of second derivatives). If the iterative algorithm converges, the limit is a solution to the likelihood equation. The method is potentially useful in computing the MLE in some cases of GLMM, where the dimension of integrals involved is not very high. In fact, as pointed out by Jiang (2006a), the ideal situation for the method to work is when $\ddot{l}$ is much more difficult to deal with, numerically or analytically, than $\dot{l}$. However, in some cases of GLMM, e.g., the salamander mating problem (McCullagh and Nelder, 1989, §14.5), even $\dot{l}$ is difficult enough to evaluate.

4.3 Non-likelihood-based inference

4.3.1 Approximate inference

When the exact likelihood function is difficult to compute, approximation becomes one of the natural alternatives. A well-known method is Laplace approximation (LA) to integrals. Several authors have used this method to approximate the likelihood function, and then treat the approximate likelihood as the true likelihood for inference about GLMM. The method may also be viewed as estimation of both fixed and random effects via maximization of the joint density functions of the observations and random effects, penalized quasi-likelihood (PQL), or maximum hierarchical likelihood. See, Schall (1991), Breslow and Clayton (1993), Wolfinger and O'Connell (1993), McGilchrist (1994), Lee and Nelder (1996), and Lin and Breslow (1996), among others. Unfortunately, the LA-based methods are known to have some unsatisfactory properties. In particular, the resulting estimators are inconsistent under standard asymptotic assumptions (e.g., Jiang, 1998b). For example, consider a problem from small area estimation, in which a sample of size n_i is available from the ith small area, $i = 1, \ldots, m$. Jiang (1999) showed that, in this situation, the method of joint estimation of fixed and random effects, which is asymptotically equivalent to the LA-based methods, will result in consistent estimators of the fixed effects only when $m/N \to 0$, where $N = \sum_{i=1}^{m} n_i$. Thus, if $n_i = k$, $1 \le i \le m$, the sample size for each small area has to be large, a situation unrealistic in many small area estimation problems. Furthermore, Lin and Breslow (1996) showed that their PQL method works well when the variances of the random effects are close to zero; otherwise, the bias can be substantial. Also see Kuk (1995).

4.3.2 Estimating equations

Generalized estimating equations (GEE) have been used in the analysis of longitudinal data (Liang and Zeger, 1986; Prentice, 1988). Such a method applies to a special class of GLMM, in which the observations are independently clustered. In other words, the covariance matrix of the observations is block-diagonal. Jiang (1998b) proposed estimating equations that apply to GLMMs not necessarily having a block-diagonal covariance structure, such as the one in the salamander mating problem. He showed that the

estimators, which are solution to the estimating equations, are consistent, but inefficient. Later, Jiang and Zhang (2001) proposed a two-step procedure to obtain more efficient estimators. They considered a broader class of models, which does not require full specification of the conditional distribution of the responses given the random effects. Therefore, the method applies to a broader class of models than the GLMMs. To illustrate their method, let S be a vector of base statistics. The first-step estimator of ψ, $\tilde{\psi}$, is a solution to the estimating equation $B\{S - u(\psi)\} = 0$, where ψ is the vector of parameters under the GLMM which include fixed effects and variance components, B is a known matrix, and $u(\psi) = \mathrm{E}_\psi(S)$. As it turns out, if the base statistics are chosen as those in Jiang (1998b), then, to obtain the first-step estimator of ψ, one only needs to assume the form of the conditional mean of the responses given the random effects. For example, if the responses y_i are sums of binary observations, one only needs to assume that $\mathrm{logit}\{\mathrm{E}(y_i|\alpha)\} = x_i'\beta + z_i'\alpha$. It was shown that, under mild conditions, the first-step estimator is consistent. In fact, the estimator of Jiang (1998b) is a special case of the first-step estimator. On the other hand, the efficiency of the first-step estimator may be improved by choosing a "good" B. It can be shown that the optimal B, in the sense of minimum asymptotic variance, is $B^* = U'V^{-1}$, where $U = \partial u/\partial\psi'$ and $V = \mathrm{Var}_\psi(S)$. It is clear that B^* depends on ψ, i.e., $B^* = B^*(\psi)$, hence is not available. However, one may replace ψ by $\tilde{\psi}$, the first-step estimator. This results in the second-step estimator, which is a solution to $\tilde{B}^*\{S - u(\psi)\} = 0$, where $\tilde{B}^* = B^*(\tilde{\psi})$. To obtain the second-step estimator, one needs to specify the forms of the first two conditional moments of the responses given the random effects. Jiang and Zhang (2001) showed that, under regularity conditions, the second-step estimator is asymptotically as efficient as the solution to the optimal estimating equation with B^*. Furthermore, the authors reported results from two simulated examples; in each case the second-step estimator had about 40% reduction of the MSE compared to the first-step estimator.

4.4 Prediction of random effects

The method of joint estimation of the fixed and random effects discussed in Subsection 4.3.1 provides estimators, or predictors, of the random effects. Jiang et al. (2001a) took another look at the method as maximization of *a posterior* of the random effects (under a non-informative prior). The authors also applied the method to a small area estimation problem.

The joint estimates of the fixed and random effects are typically obtained by solving a system of nonlinear equations

$$\frac{\partial l_{\rm J}}{\partial \beta} = 0, \quad \frac{\partial l_{\rm J}}{\partial \alpha} = 0, \tag{4.4}$$

where $l_{\rm J}$ is the logarithm of the joint density. However, in many cases the number of random effects involved is quite large. For example, in the salamander-mating problem (McCullagh and Nelder, 1989, Section 14.5), the number of random effects associated with the female and male salamanders is 80. In the NHIS problem discussed in Malec et al. (1997), the number of random effects corresponding to the small areas is over 600. This means that one has to solve a large system of nonlinear equations. It is well-known that standard methods of solving nonlinear systems such as the Newton-Raphson method is inefficient in such problems. Jiang (2000b) proposed a nonlinear Gauss-Seidel algorithm for effectively solving (4.4). The author proved global convergence of the algorithm. Alternatively, Breslow and Clayton (1993) proposed an iterative procedure for solving (4.4) by modifying the Fisher scoring algorithm. An attractive feature of the latter procedure is that it exploits a close correspondence with the well-known mixed model equations (Henderson et al., 1959).

One of the main areas in which prediction of random effects, or mixed effects, plays an important role is small area estimation. In this context, a method known as empirical best prediction (EBP) has been developed. Consider a GLMM suitable for small area estimation such that conditional on a vector of random effects, $\alpha_i = (\alpha_{ij})_{1\le j\le r}$, responses $y_{i1}, \ldots, y_{in_i}$ are independent with density

$$f(y_{ij}|\alpha_i) \;=\; \exp\left[\left(\frac{a_{ij}}{\phi}\right)\{y_{ij}\xi_{ij} - b(\xi_{ij})\} + c\left(y_{ij}, \frac{\phi}{a_{ij}}\right)\right],$$

where $b(\cdot)$ and $c(\cdot,\cdot)$ are functions associated with the exponential family (McCullagh and Nelder, 1989, §2), ϕ is a dispersion parameter, a_{ij} is a weight such that $a_{ij} = 1$ for ungrouped data; $a_{ij} = l_{ij}$ for grouped data when the average is considered as response and l_{ij} is the group size; and $a_{ij} = l_{ij}^{-1}$ when the sum of individual responses is considered. Furthermore, ξ_{ij} is associated with a linear function

$$\eta_{ij} \;=\; x_{ij}'\beta + z_{ij}'\alpha_i$$

through a link function $g(\cdot)$, i.e., $g(\xi_{ij}) = \eta_{ij}$, or $\xi_{ij} = h(\eta_{ij})$, where $h = g^{-1}$. Here x_{ij} and z_{ij} are known vectors, and β a vector of unknown regression

coefficients. Finally, suppose that $\alpha_1, \ldots, \alpha_m$ are independent with density $f_\vartheta(\cdot)$, where ϑ is a vector of variance components. Let $\psi = (\beta', \vartheta')'$, and $\varphi = (\psi', \phi)$. Note that in some cases such as *binomial* and *Poisson* the dispersion parameter ϕ is known, so ψ represents the vector of all unknown parameters.

Consider the problem of predicting a mixed effect of the following form:

$$\zeta \;=\; \zeta(\beta, \alpha_S),$$

where S is a subset of $\{1, \ldots, m\}$, and $\alpha_S = (\alpha_i)_{i \in S}$. Let $y_S = (y_i)_{i \in S}$, where $y_i = (y_{ij})_{1 \le j \le n_i}$. Under the above model, the best predictor (BP) of ζ, in the sense of minimum MSPE, is given by

$$\begin{aligned}
\tilde{\zeta} &= \mathrm{E}(\zeta | y) \\
&= \mathrm{E}(\zeta(\beta, \alpha_S) | y_S) \\
&= \frac{\int \zeta(\beta, \alpha_S) f(y_S | \alpha_S) f_\vartheta(\alpha_S) d\alpha_S}{\int f(y_S | \alpha_S) f_\vartheta(\alpha_S) d\alpha_S} \\
&= \frac{\int \zeta(\beta, \alpha_S) \exp\{\phi^{-1} \sum_{i \in S} s_i(\beta, \alpha_i)\} \prod_{i \in S} f_\vartheta(\alpha_i) \prod_{i \in S} d\alpha_i}{\prod_{i \in S} \int \exp\{\phi^{-1} s_i(\beta, v)\} f_\vartheta(v) dv},
\end{aligned}$$

where $s_i(\beta, v) = \sum_{j=1}^{n_i} a_{ij} [y_{ij} h(x'_{ij}\beta + z'_{ij} v) - b\{h(x_{ij}\beta + z_{ij} v)\}]$. The dimension of integrals involved in the denominator on the right side is $r = \dim(\alpha_i)$, while that of the numerator is at most sr, where $s = |S|$, the cardinality of S. When r and s are relatively small, such integrals may be evaluated by Monte Carlo methods, provided that ψ (φ) is known. For example, suppose that $\alpha_i \sim N\{0, V(\vartheta)\}$, where $V(\vartheta)$ is a covariance matrix depending on ϑ, and that $S = \{i\}$. Then, we have

$$\begin{aligned}
\tilde{\zeta} &= \frac{\int \xi(\beta, v) \exp\{\phi^{-1} s_i(\beta, v)\} f_\vartheta(v) dv}{\int \exp\{\phi^{-1} s_i(\beta, v)\} f_\vartheta(v) dv} \\
&\approx \frac{\sum_{l=1}^{L} \zeta(\beta, v_l) \exp\{\phi^{-1} s_i(\beta, v_l)\}}{\sum_{l=1}^{L} \exp\{\phi^{-1} s_i(\beta, v_l)\}},
\end{aligned}$$

where $f_\vartheta(v)$ is the density of $N\{0, V(\vartheta)\}$, and $v_1, \ldots, v_L$ are generated independently from $N\{0, V(\vartheta)\}$.

Note that the BP depends on both y_S and φ, that is, $\tilde{\zeta} = u(y_S, \varphi)$. Since φ is usually unknown, it is customary to replace φ by a consistent

estimator, say, $\hat{\varphi}$. The result is called empirical best predictor, or EBP, given by

$$\hat{\zeta} \quad = \quad u(y_S, \hat{\varphi}). \tag{4.5}$$

In practice, it is desirable not only to compute the EBP but also to assess its variation. A measure of the variation is the MSPE, defined by $\mathrm{MSPE}(\hat{\zeta}) = \mathrm{E}(\hat{\zeta}-\zeta)^2$. Unfortunately, the latter may be difficult to evaluate. In some cases, an expression of the MSPE of $\tilde{\zeta}$, not that of $\hat{\zeta}$, may be obtained, say, $\mathrm{MSPE}(\tilde{\zeta}) = b(\varphi)$. Then, a naive estimator of the MSPE of $\hat{\zeta}$ is obtained as $b(\hat{\varphi})$. However, this underestimates the true MSPE. To see this, note the following decomposition of the MSPE:

$$\begin{aligned} \mathrm{MSPE}(\hat{\zeta}) &= \mathrm{MSPE}(\tilde{\zeta}) + \mathrm{E}(\hat{\zeta} - \tilde{\zeta})^2 \\ &= b(\varphi) + \mathrm{E}(\hat{\zeta} - \tilde{\zeta})^2. \end{aligned} \tag{4.6}$$

It is clear that the naive estimator simply ignores the second term on the right side of (4.6), and therefore underestimates the true MSPE.

Jiang (2003) developed a method based on Taylor series expansion that gives an estimate whose bias is corrected to the second-order. The method may be regarded as an extension of the Prasad-Rao method for estimating the MSPE of EBLUP in linear mixed models. Consider, for simplicity, the case that ϕ is known (e.g., *binomial*, *Poisson*), so that $b(\varphi) = b(\psi)$ in (4.6). Then, the estimator may be expressed as

$$\widehat{\mathrm{MSPE}}(\hat{\zeta}) \quad = \quad b(\hat{\psi}) + m^{-1}\{e(\hat{\psi}) - B(\hat{\psi})\}, \tag{4.7}$$

where the functions $e(\cdot)$ and $B(\cdot)$ are some functions (see below for a special case). Here by second-order correctness we mean that the estimator has the property that

$$\mathrm{E}\{\widehat{\mathrm{MSPE}}(\hat{\zeta}) - \mathrm{MSPE}(\hat{\zeta})\} \quad = \quad o(m^{-1}). \tag{4.8}$$

Note that, if $\widehat{\mathrm{MSPE}}(\hat{\zeta})$ in (4.8) is replaced by the naive estimator $b(\hat{\vartheta})$, which is the first term on the right side of (4.7), the right side of (4.8) will have to be replaced by $O(m^{-1})$. In other words, the naive estimator is correct to the first order, not the second one.

Alternatively, the jackknife method of Jiang et al. (2002, see Subsection 3.3) can be used here to obtain an estimate of the MSPE that is correct to the second order.

4.5 Example

We illustrate the EBP method described in the previous subsection for the case of binary responses in small area estimation. Jiang and Lahiri (2001) considered the following mixed logistic model for small area estimation. Suppose that, conditional on an area-specific random effect α_i, binary responses y_{ij}, $1 \le j \le n_i$ are independent with

$$\text{logit}\{\text{P}(y_{ij} = 1|\alpha_i)\} \quad = \quad x'_{ij}\beta + \alpha_i,$$

where x_{ij} is a vector of known covariates and β a vector of unknown regression coefficients. Furthermore, $\alpha_1, \ldots, \alpha_m$ are independent and distributed as $N(0, \sigma^2)$, where σ^2 is an unknown variance.

Suppose that the problem of interest is to predict α_i, the small area-specific random effect. By (4.5), the EBP is $\hat{\alpha}_i = u_i(y_{i\cdot}, \hat{\psi})$, where $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$, $\psi = (\beta', \sigma)'$,

$$u_i(y_{i\cdot}, \psi) \quad = \quad \sigma \frac{\text{E}[\xi \exp\{s_i(y_{i\cdot}, \sigma\xi, \beta)\}]}{\text{E}[\exp\{s_i(y_{i\cdot}, \sigma\xi, \beta)\}]}$$

with $s_i(k, v, \beta) = kv - \sum_{j=1}^{n_i} \log\{1 + \exp(x'_{ij}\beta + v)\}$ and $\xi \sim N(0, 1)$.

To see the behavior of u_i, note that $u_i(y_{i\cdot}, \psi)/\sigma \to 0$ as $\sigma \to 0$. Now consider a special case in which $x_{ij} = x_i$, that is, the covariates are at the small area (e.g., county) level. Then, it can be shown (Jiang and Lahiri, 2001) that, as $\sigma \to \infty$,

$$u_i(y_{i\cdot}, \psi) \quad \longrightarrow \quad \sum_{k=1}^{y_{i\cdot}-1} \left(\frac{1}{k}\right) - \sum_{k=1}^{n_i - y_{i\cdot}-1} \left(\frac{1}{k}\right) - x'_i\beta.$$

To see what the expression means, note that when n is large, $\sum_{k=1}^{n-1}(1/k) \sim \log(n) + C$, where C is Euler's constant. Therefore, as $\sigma \to \infty$, we have

$$u_i(y_{i\cdot}, \psi) \quad \approx \quad \text{logit}(\bar{y}_{i\cdot}) - x'_i\beta.$$

Finally, it can be shown that, as $m \to \infty$ and $n_i \to \infty$, we have

$$\hat{\alpha}_i - \alpha_i \quad = \quad O_{\text{P}}(m^{-1/2}) + O_{\text{P}}(n_i^{-1/2}).$$

Now consider the estimation of the MSPE of $\hat{\alpha}_i$. It can be shown that, in this case, the terms $b(\psi)$ and $e(\psi)$ in (4.7) have the following expressions:

$$\begin{aligned} b(\psi) &= \sigma^2 - \sum_{k=0}^{n_i} u_i^2(k,\psi) p_i(k,\psi), \\ e(\psi) &= \sum_{k=0}^{n_i} \left(\frac{\partial u_i}{\partial \psi'}\right) V(\psi) \left(\frac{\partial u_i}{\partial \psi}\right) p_i(k,\psi), \end{aligned}$$

where

$$\begin{aligned} p_i(k,\psi) =& \mathrm{P}(y_{i\cdot} = k) \\ =& \sum_{z \in S(n_i,k)} \exp\left(\sum_{j=1}^{n_i} z_j x_{ij}'\beta\right) \mathrm{E}[\exp\{s_i(z_\cdot, \sigma\xi, \beta)\}] \end{aligned}$$

with $S(l,k) = \{z = (z_1, \ldots, z_l) \in \{0,1\}^l : z_\cdot = z_1 + \cdots + z_l = k\}$.

Next, we consider the prediction of the mixed effect $p_i = \mathrm{P}(y_{ij} = 1 | \alpha_i)$. For simplicity, suppose that the covariates are at the small area level, i.e., $x_{ij} = x_i$. Then, we have

$$p_i = \frac{\exp(x_i'\beta + \alpha_i)}{1 + \exp(x_i'\beta + \alpha_i)}.$$

The EBP of p_i is given by

$$\begin{aligned} \hat{p}_i &= u_i(y_{i\cdot}, \hat{\psi}) \\ &= \exp(x_i'\hat{\beta}) \\ &\quad \times \frac{\mathrm{E}\exp[(y_{i\cdot}+1)\hat{\sigma}\xi - (n_i+1)\log\{1+\exp(x_i'\hat{\beta}+\hat{\sigma}\xi)\}]}{\mathrm{E}\exp[y_{i\cdot}\hat{\sigma}\xi - n_i\log\{1+\exp(x_i'\hat{\beta}+\hat{\sigma}\xi)\}]}, \end{aligned} \tag{4.9}$$

where the expectations are taken with respect to $\xi \sim N(0,1)$. Note that the EBP is not p_i with β and α_i replaced respectively by $\hat{\beta}$ and $\hat{\alpha}_i$.

On the other hand, a naive predictor of p_i is $\bar{y}_{i\cdot} = y_{i\cdot}/n_i$. Although the EBP given by (4.9) is not difficult to compute (e.g., by Monte Carlo method), it does not have a closed form. So, a natural question is: how much is the gain in using EBP over the naive predictor? To answer this question, we consider the relative savings loss (RSL) introduced by Efron

and Morris (1973). In the current case, the RSL is given by

$$\begin{aligned} \text{RSL} &= \frac{\text{MSPE}(\hat{p}_i) - \text{MSPE}(\tilde{p}_i)}{\text{MSPE}(\bar{y}_{i\cdot}) - \text{MSPE}(\tilde{p}_i)} \\ &= \frac{\text{E}(\hat{p}_i - \tilde{p}_i)^2}{\text{E}(\bar{y}_{i\cdot} - \tilde{p}_i)^2}, \end{aligned} \tag{4.10}$$

where $\tilde{p}_i$ is the BP of p_i. It can be shown that the numerator on the right side of (4.10) is $O(m^{-1})$, while

$$\begin{aligned} \text{the denominator} &= \sum_{k=0}^{n_i} \left\{ \frac{k}{n_i} - u_i(k,\psi) \right\}^2 p_i(k,\psi) \\ &\geq \{u_i(0,\psi)\}^2 p_i(0,\psi). \end{aligned} \tag{4.11}$$

If n_i is bounded, the right side of (4.11) has a positive lower bound. Therefore, RSL $\to 0$ as $m \to \infty$. In fact, the convergence rate is $O(m^{-1})$. So, the complication of EBP is worthwhile.

5 Different adjustments to EBP

Given adequate time and resources, it is reasonable to expect that a researcher can identify a good working model which is the key step in improving small area statistics no matter what model-based method is used. However, this is quite a formidable task in a complex survey environment requiring small area statistics for a large number of variables on an ongoing basis. A reasonable approach might be to find a relatively simple robust mixed model - one that can be scientifically justified and is likely to work for many variables at different points of time. Even after taking such a caution, there is no guarantee that it will work reasonably well in all situations and all time. Thus, it is reasonable for a survey practitioner to expect a few more precautions before producing officials statistics that are so important for the understanding of the state of economy, health, education, and the overall well being of a country. We describe below a few adjustments to the EBP that have been considered in the literature.

5.1 Benchmarking

A typical data file from a complex survey contains survey weight for each respondent. A survey weight associated with a respondent represents a

certain number of units in the population that is sampled. There are many factors that go into the construction of survey weights, the first factor being the basic weight that is needed to compensate for the unequal probability of selection in order to avoid the selection bias. Basic weights are usually calibrated to make the sample representative of the finite population in the sense that the distributions of the survey weights with respect to certain variables, say various socio-demographic variables, match the corresponding known distributions for the finite population. Finally, the nonresponse factor is introduced to compensate for the nonresponse bias. It simply inflates the survey weight for a respondent who is similar to one or more nonrespondents in terms of a few variables (e.g., socio-demographic variables) for which information is available both for the respondents and the nonrespondents. Since survey weights contain a great deal of information on the selection and nonresponse biases for the sample and information about some population characteristics, they are used in estimation and other inferences from complex survey data. For example, if a survey oversamples a particular minority group, it is likely that the sample will have a disproportionate number of that minority group compared to other groups of the finite population. In such a case, a weighted mean is much more sensible than the ordinary sample mean.

The weighted survey estimator for a large area is generally quite robust and does not need a model for its justification. They have been found to be reasonable in various empirical evaluation checks using administrative and census data. A small area estimator satisfies the benchmarking property if it can generate the usual desirable large area weighted estimator by an appropriate aggregation. To illustrate the property, let y_{ij} (w_{ij}) denote the value of a response variable (survey weight) for the jth unit of the small area i belonging to a large area ($i = 1, \cdots, m;\ j = 1, \cdots, n_i$). Then $\hat{T} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} y_{ij}$ is the usual survey weighted estimator of the finite population total $T = \sum_{i=1}^{m} T_i$, with $T_i = \sum_{j=1}^{N_i} y_{ij}$, where $N_i = \sum_{j=1}^{n_i} w_{ij}$ is the population size for the ith area. Note that $\hat{T}_i = \sum_{j=1}^{n_i} w_{ij} y_{ij}$, the usual survey weighted estimator for area i ($i = 1, \cdots, m$), satisfies the benchmarking property since $\sum_{i=1}^{m} \hat{T}_i = \hat{T}$; but as we discussed earlier they are very inefficient in producing reliable small area statistics. The BP or EBP method usually does not enjoy the benchmarking property. To see this, let us consider a special case of the nested error regression model discussed in Subsection 2.2 with $x_{ij} = \mu$ and known model parameters. In this case, using Ghosh and Meeden (1986) and Ghosh and Lahiri (1987),

the BP for T_i can be expressed as

$$\hat{T}_i^{BP} = \sum_{j=1}^{n_i} w_{ij}^{BP} y_{ij} + f_i B_i N_i \mu,$$

where w_{ij}^{BP}, the BP weight, is given by

$$w_{ij}^{BP} = \frac{1 - f_i B_i}{1 - f_i},$$

where $f_i = 1 - n_i/N_i$, the finite population correction factor for the ith small area, and $B_i = \sigma_e^2/(\sigma_e^2 + n_i\sigma_v^2)$. Evidently, in general $d = \hat{T}^{BP} - \hat{T} \neq 0$, where $\hat{T}^{\mathrm{BP}} = \sum_{i=1}^{m} \hat{T}_i^{\mathrm{BP}}$. Thus, unlike $\hat{T}_i$'s, the BP's for the small areas do not usually satisfy the benchmarking requirement. In fact, sometimes the difference d could be considerably large. If this happens, it is hard to justify the use of EBP's, especially to the policy makers who are usually comfortable with the weighted means because of its simplicity of interpretation. Note that $\hat{T}_i^{BP}$, when multiplied by the ratio $\hat{T}/\hat{T}^{\mathrm{BP}}$, yields small area estimators that satisfy the benchmarking requirement. Such simple ratio adjustment is known as raking in the sample survey literature.

Sometimes it is desirable to achieve the benchmarking property in two or more dimensions in which case generalized raking or proportional iterated fitting (IPF) (see Ireland and Kullback, 1968) has been employed. Fay and Herriot (1979) provided one practical example for this kind of adjustments to the EBP's. They used a two-way raking method to adjust the small area EBP's of the per-capita income for small places so that they add up to the county level per-capita income and also the per-capita income for a large region by three different sizes. Evidently, raking destroys the good model optimality properties of the BP and EBP. But, Fay and Herriot (1979) demonstrated that even after such raking adjustments to their EBP the latter performed well compared to the survey weighted estimator or a synthetic (regression) estimator that the Census Bureau used.

5.2 Design-consistency

According to the design-consistency property, the difference between an estimator and the corresponding true finite population parameter converges in probability to zero for a large sample. The convergence is with respect

to the sample design under the standard set of regularity conditions (see Isaki and Fuller, 1982), and is valid irrespective of the model that may have been used to generate the estimator or predictor. The weighted mean has the desirable design-consistency property.

In a typical small area estimation setting, we are usually interested in the simultaneous estimation of a number of areas or domains, some of them could be large planned areas. For example, in a nationwide survey, the sample size for a large subnational area (e.g. a large state like California in a US national survey) is likely to be larger than that for a smaller subnational area (e.g. a small state like Arkansas). The model-assisted approach is one where one uses a working mixed model to derive a small area predictor in such a way that it achieves the desirable design-consistency property. This is done with a hope of getting some protection against a possible model failures at the expense of possible loss of efficiency. Needless to say, the design-consistency of a model-assisted procedure makes it closer to the survey weighted estimator for large areas, an estimator popular among some survey data users.

Note that generally $\hat{T}_i^{BP} - \hat{T}_i$ does not converge to 0 in the usual sense of asymptotics for infinite population sampling and hence the BP is not design consistent. Särndal (1984) proposed a model-assisted method for small area estimation but used simple regression type fixed effects models. Kott (1989) was the first to recognize the utility of a random effects model in a model-assisted approach for domain estimation. Using the same one-way random effects model, Prasad and Rao (1999) provided a pseudo-EBLUP domain estimator and demonstrated the superiority of their estimator over the Kott's estimator. They considered the following reduced model on $\bar{y}_{iw} = \hat{T}_i/N_i$, the usual survey weighted estimator of the small area mean T_i/N_i:

$$\bar{y}_{iw} \quad = \quad \mu + v_i + e_{iw},$$

where $e_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} e_{ij}$, $\tilde{w}_{ij} = w_{ij}/N_i$ and $e_{iw} \overset{ind}{\sim} (0, \delta_{iw}\sigma_e^2)$ with $\delta_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij}^2$. Instead of deriving the BLUP of the finite population mean for area i using the full model on y_{ij}, they derived the following BP of $\theta_i = \mu + v_i$ under the above reduced model as

$$\hat{\theta}_i^{\mathrm{PBP}} \quad = \quad (1 - B_{iw})\bar{y}_{iw} + B_{iw}\mu,$$

where $B_{iw} = \sigma_e^2\delta_{iw}/(\sigma_e^2\delta_{iw}+\sigma_v^2)$. They also provided a MSPE estimator and showed that their MSPE estimator is more stable than that of Kott's. See

You and Rao (2002) for an extension of Prasad and Rao (1999), where they achieved the desirable benchmarking properties of their pseudo-EBLUP. Li (2005) extended this idea of simultaneously achieving design-consistency and benchmarking properties.

We note that the MSPE estimators proposed by Prasad and Rao (1999) and You and Rao (2002) are not second-order correct, a desirable property in small area estimation (Prasad and Rao, 1990). Furthermore, the approach considered by Prasad and Rao (1999) is not readily extendable to GLMMs because of the difficulty in deriving the reduced model for the survey weighted estimator from the mixed model assumed on the unit level observations. In addition, the one-way random effects model considered in Kott (1989) and Prasad and Rao (1999) is clearly not appropriate for a complex sampling design such as a stratified multi-stage sampling design. Recently, Jiang and Lahiri (2006) put forward a unified method for producing model-assisted small area predictors using a general mixed model that includes longitudinal linear mixed models and GLMMs as special cases. The method essentially amounts to assuming a mixed model on the survey weighted estimator itself, rather than a mixed model on the observations on the unit level, and then obtaining the EBP for the expectation of the survey weighted estimator conditional on the random effects. This is in sharp contrast with the pseudo-EBLUP method of Prasad and Rao (1999), who derived a model for the design consistent estimator using a one-way random effects superpopulation model for all the units (both observed and unobserved) of the finite population. Jiang and Lahiri (2006) also provided a second-order unbiased MSPE estimator.

5.3 Limited translation method

Although the overall performance of the EBP is expected to be better than the survey weighted estimator when the model assumed is reasonable, it may perform miserably for a particular small area due to a variety of reasons (e.g., poor auxiliary data for the small area or an outlier). Following the suggestions of Efron and Morris (1972), Fay and Herriot (1979) considered the following type of adjustment to the EBP:

$$\hat{\theta}_i \;=\; \begin{cases} \hat{\theta}_i^{EBP}, & \text{if } y_i - c_i\sqrt{D_i} \le \hat{\theta}_i^{EBP} \le y_i + c_i\sqrt{D_i}, \\ y_i - c_i\sqrt{D_i}, & \text{if } \hat{\theta}_i^{EBP} < y_i - c_i\sqrt{D_i}, \\ y_i + c_i\sqrt{D_i}, & \text{if } \hat{\theta}_i^{EBP} > y_i + c_i\sqrt{D_i}, \end{cases}$$

where c_i is a suitably chosen constant. For example, $c_i = 1$ in the application considered by Fay and Herriot (1979). The main idea is to limit the deviation of the EBP from the survey weighted estimator. This reduces the loss in efficiency if the model fails for a particular area without losing the benefit of modelling for the areas where the model works reasonably well. In the baseball example, such an adjustment protected a good player (Roberto Clemente) and a bad hitter. See Efron (1975).

6 Other issues

6.1 Prediction interval

Interval prediction is a useful data analysis tool since this integrates the concepts of both point prediction and hypothesis testing in an effective manner. Jeske and Harville (1988) considered prediction intervals for a mixed effect, assuming the parametric distribution of α and $y - \mathrm{E}(y)$. Jiang and Zhang (2002) used a distribution-free method for constructing prediction intervals for a future observation under a non-Gaussian linear mixed model.

In the context of small area estimation, the mean of area i, say θ_i, is treated as a mixed effect and an interval, say $I_i(\alpha)$, is called a $100(1-\alpha)\%$ prediction interval for θ_i if $\mathrm{P}(I_i(\alpha) \ni \theta_i) = 1 - \alpha$, where the probability is with respect to the marginal distribution induced by the mixed model. In the empirical Bayes context (see Morris, 1983a,b), this is also referred to as an empirical Bayes confidence interval for θ_i and in this review paper we do not distinguish between the two different nomenclatures. Mostly, these prediction intervals are of the form $\hat{\theta}_i \pm qs(\hat{\theta}_i)$, where $s(\hat{\theta}_i)$ is a measure of uncertainty of an arbitrary point predictor $\hat{\theta}_i$ and q is chosen suitably in an effort to attain good coverage probability, i.e., at least greater than the nominal value of $1 - \alpha$.

Research on prediction intervals in the small area context is limited except in some special cases involving normal distributions. For our review, we consider the Fay-Herriot model which received the most attention so far as the interval prediction is concerned. For this example, the direct prediction interval is given by

$$I_i^D(\alpha) : y_i \pm z_{\alpha/2}\sqrt{D_i},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percent point of $N(0,1)$. Obviously, for this prediction interval, the coverage probability is $1-\alpha$. However, it is not efficient since its average length is too large to make any reasonable conclusion. This is due to the high variability of the point predictor y_i. Cox (1975) initiated the idea of developing the empirical Bayes confidence intervals. In the current context, his suggestion generates the following prediction interval:

$$I_i^C(\alpha) : \hat{\theta}_i(y_i; \hat{A}) \pm z_{\alpha/2}\sqrt{\hat{g}_{1i}}.$$

When m is large, under certain regularity conditions $\mathrm{P}(I_i^C(\alpha)) = 1-\alpha + O(m^{-1})$. Thus, this prediction interval attains the desired coverage probability asymptotically but the margin of error is of the order $O(m^{-1})$, not accurate enough in most small area applications. Intuitively, this could be due to the fact that the construction of the prediction interval does not take into account the additional errors incurred by the estimation of model parameters. In fact, in a simulation study, Chatterjee and Lahiri (2002) reported that the coverage probability could be as low as 31% when the nominal value is 95%.

For a special case of the Fay-Herriot model with common mean and equal sampling variances $D_i = D$, Morris (1983b) used a different measure of uncertainty for the empirical Bayes estimator that incorporates the additional uncertainty due to the estimation of the model parameters. In a different work, Morris (1983a) considered a variation of his (1983b) empirical Bayes confidence interval where he used a hierarchical Bayes type point estimator in place of the previously used empirical Bayes estimator and conjectured with some evidence that the coverage probability for his interval is at least $1-\alpha$. He also noted that the coverage probability tends to $1-\alpha$ as m goes to ∞ or D goes to zero. Basu et al. (2003) showed that the empirical Bayes confidence interval proposed by Morris (1983b) does not improve the coverage probability in the sense that the margin of error of the coverage probability remains $O(m^{-1})$. However, using a Taylor series expansion of the coverage probability, Basu et al. (2003) obtained an expression for the order $O(m^{-1})$ term which was then used to calibrate the length of Morris' empirical Bayes confidence interval in order to reduce the margin of error to $o(m^{-1})$. They also studied a prediction interval proposed by Carlin and Louis (1996, p. 98) and showed that the Carlin-Louis interval has a coverage bias of the order $o(m^{-1})$. As a by-product of this analytical result, they obtained an explicit expression for an alternative prediction interval attaining a coverage probability up to the same

order of approximation. Basu et al. (2003) showed that their modification to Morris' proposal and analytic implementation of Carlin and Louis' idea enjoy the same coverage probability and expected length, even under the higher order asymptotics retaining terms of the order $O(m^{-1})$. Using a multi-level model, Nandram (1999) obtained an empirical Bayes confidence interval for a small area mean and showed that asymptotically it converges to the nominal coverage probability. However, he did not study the higher order asymptotic properties of his interval.

Datta et al. (2002) used an approach similar to Basu et al. (2003) in order to calibrate the Cox-type prediction interval for a more general Fay-Herriot model with covariates but with equal sampling variances. The coverage error for their interval is of the order $O(m^{-1.5})$. Chatterjee and Lahiri (2002) used a parametric bootstrap method to obtain a bootstrap prediction interval for the most general case of the Fay-Herriot model. They showed that the margin of error for the coverage probability is of the order $O(m^{-1.5})$. For a special case of the Fay-Herriot model with no covariate and equal sampling variances, Laird and Louis (1987) also proposed a prediction interval based on parametric bootstrap. However, their interval is different from that of Chatterjee and Lahiri (2002) in the sense that the former used Morris' approximation while the latter used bootstrap interval in the usual way. Higher order asymptotic properties of the Laird-Louis parametric bootstrap is not available. Carlin and Gelfand (1990, 1991) proposed a bootstrap approximation of the empirical Bayes confidence intervals conditional on some statistic which is not ancillary. Hill (1990) suggested a general framework for constructing an empirical Bayes confidence interval conditional on some suitable ancillary statistic. In a simple normal-normal setting this matches with an exact hierarchical Bayes confidence interval. Datta et al. (2002) followed up Hill's idea in proposing an empirical Bayes confidence interval which is correct up to $O(m^{-1})$, a property not noted by Hill (1990).

6.2 Model building

Model building is probably the most complex part in any scientific investigation that uses a statistical model, and the complexity increases with the increase of the number and nature of different sources of uncertainties that make up the data. A good interaction between the subject matter expert and the statistician helps the model building process immensely,

but eventually the observed data is to be used in an intelligent manner to determine a reasonable model useful for a particular statistical inference. Like any other good model building process, small area model building is essentially an iterative procedure that involves selection of a model from a class of plausible models, model fitting, and model diagnostics. A class of models is first identified. One or more appropriate statistical methods are applied to select an initial model from the class of plausible models. The selected model is then fitted and any possible deficiency of the assumed model are checked through appropriate model diagnostics. Model diagnostics could reveal certain deficiencies of the initial model which helps determining the next model to be fitted and checked. The iterative process is then continued until a model is found satisfactory to both the analyst and the subject matter expert.

Mixed models are used extensively in small area inference. However, the mixed model formulation is too general. The first task of an analyst is to understand the type of the mixed model that is relevant for a specific small area application. Two main sources of variations are generally considered in selecting a mixed model: (i) within area and (ii) between area variations. Multi-level models, a special case of the general mixed model, are frequently used to explain both the within and between area variations in most small area applications. Since the number of small areas and factors is typically large, the model building process is complicated. Unlike a fixed effects model, the selection and diagnostics for the random effects in a multi-level model are much more difficult since the random effects are unobserved.

A suitable graphical device always help in multi-level model selection. For example, Battese et al. (1988) used a single graph to study the relationship between their dependent and independent variables for each of the small areas. They observed a linear relationship where the intercepts appeared to vary across area. However, the number of sample for a given area is too small to fit different intercepts for different areas. As a compromise, Battese et al. (1988) considered intercepts to be exchangeable in their nested error regression model. For the same data set, one could raise several questions leading to other multi-level models. For example, do the slopes of the regression coefficients vary across areas as well? Should one assume area-specific sampling variances as in Arora et al. (1997) or Kleffe and Rao (1992)? In another example from the Small Area Income and Poverty Program (SAIPE) of the U.S. Census Bureau, the Fay-Herriot model has been used. Here a two level model comes very naturally where the first level

is used to explain the sampling variabilities of the survey estimates that need to be smoothed. The second level is used to explain the variations of the true poverty rates in terms of a number of administrative variables. How does one select these administrative variables in presence of sampling variabilities of the survey estimates that are modelled? This is certainly an important issue. See Lahiri and Li (2005) for further discussions of this type of problems.

There is a huge literature on model selection. Readers are referred to the Institute of Mathematical Statistics Monograph edited by Lahiri (2001) for four long review papers on the subject. While there is an extensive literature on parameter estimation in mixed models, mixed model selection has received relatively less attention. This is a topic of current research interest. Datta and Lahiri (2001) discussed a model selection method for choosing between a fixed effect model and a random effect model, a topic of interest in small area applications. Jiang and Rao (2003) proposed two methods for linear mixed model selection in general. The first method selects variables associated with the fixed effects, assuming that the part of the model involving the random effects is not subject to selection; the second method carries out model selection to the random part as well. However, the Jiang-Rao methods have yet to be used in a small area context. In the complex survey context, Fabrizi and Lahiri (2004) proposed a robust approximation to the Bayesian Information Criterion (BIC) that can be applied to certain small area problems with complex survey data. See Meza and Lahiri (2005) for an extension of the well-known C_p statistic to the small area setting. Further research is needed in this important area.

There has been some significant development in linear mixed model diagnostics. See Beckman et al. (1987), Dempster and Ryan (1985), Battese et al. (1988), Lange and Ryan (1989), Calvin and Sedransk (1991), Pinheiro and Bates (2000), Christensen et al. (1992), among others. In informal checking of mixed models, the construction of appropriate residuals is the key step. However, these tools should be used with care and one should pay attention to the special multi-level structure of the small area data. For example, an overall residual plot, i.e. one that does not pay any attention to the small area identifier, obtained from a nested error regression model may not reveal any abnormalities; but when plotted separately for different areas it may indicate a lack of fit. Alternately, an overall residual plot may reveal an outlier but this may be due to an important small area effect that has been left out in the mixed model. Futhermore, certain

formal tests of goodness-of-fit were considered in the literature. Jiang et al. (2001b) developed an asymptotic theory of Pearson's χ^2-test with estimated cell frequencies for formally checking the distributional assumptions. The procedure requires splitting the data into two parts, one used for estimation and the other for testing. Jiang (2001) developed a χ^2 goodness-of-fit test that applies to a general mixed ANOVA model, which does not require data splitting. We refer to Rao (2003) who described some of the methods discussed in this paper.

We end this section where we began. We need a systematic iterative method for building good multi-level small area model. A good reference is Christiansen and Morris (1996), who discussed such a model building process for a Poisson multi-level model and applied it to modeling patients mortality rates in heart transplant patients. The S-PLUS program PRIMM that was used in the paper is available through Statlib. See Christiansen and Morris (1994) for some details about the program.

7 Baseball example revisited

In this section, we revisit the baseball data example given in Efron and Morris (1975). This data set has been analyzed by several authors in the past, including Efron and Morris (1975), Efron (1975), Morris (1983b), Gelman et al. (1995), Datta and Lahiri (2000), Rao (2003), Lahiri and Li (2005), among others. Efron and Morris (1975) used this data set to demonstrate the performance of their empirical Bayes and limited translation empirical Bayes estimators derived using an exchangeable prior in the presence of an outlying observation. They first obtained the batting average of Roberto Clemente, an extremely good hitter, from New York Times dated April 26, 1970 when he had already batted $n = 45$ times. The batting average of a player is just the proportion of hits among the number at-bats. They selected 17 other major league baseball players who had also batted 45 times from the April 26 and May 2, 1970 issues of the New York Times. They considered the problem of predicting the batting averages of all the 18 players for the remainder of the 1970 season based on their batting averages for first 45 at bats. This is a good example for checking the effect of an outlier on the efficiency of an EB estimation with an exchangeable prior. Gelman et al. (1995) provided additional data for this estimation problem and included important auxiliary data like the batting average of

each player through the end of 1969 season. We review the problem of predicting the batting averages of all the 18 players for the entire 1970 season, instead of predicting the the batting averages for the remainder of the 1970 season as Efron and Morris (1975) originally considered.

For the player i ($i = 1, \cdots, m$), let p_i and π_i be the batting average for first 45 'at-bats' and the true season batting average of the 1970 season. Note that p_i is the direct maximum likelihood (also unbiased) estimator of π_i under the assumption that conditional on π_i, the number of hits for first n at bats, np_i, follows a binomial distribution with number of trials n and success probability π_i, $i = 1, \cdots, m$.

Efron and Morris (1975) considered the following standard arc-sine transformation:

$$y_i \quad = \quad \sqrt{n}\arcsin(2p_i - 1)$$

and then assumed the following model

$$y_i|\theta_i \quad \overset{\text{ind}}{\sim} \quad N(\theta_i, 1), \text{i=}1, \cdots, m,$$

where $\theta_i = \sqrt{n}\arcsin(2\pi_i - 1)$. There could a criticism about the validity of the above approximation. However, both Efron and Morris (1975) and Gelman et al. (1995) noted that this is not a serious concern given the moderate sample size of 45. The data analysis by Lahiri and Li (2005) supports this conjecture. Efron and Morris (1975) assumed exchangeability of the θ_i's and used the two-level Fay-Herriot model, given in Subsection 2.1, without any covariate and equal sampling variances (i.e. 1).

Gelman et al. (1995) noted the possibility of an extra-binomial variation in the number of hits. The outcomes from successive bats could be correlated and the probability of hits may change across bats due to injury to the player and other external reasons not given in the data set. However, there is no way to check these assumptions because of the unavailability of such data. Assuming Level 1 is reasonable, Lahiri and Li (2005) checked the validity of the above model through graphical tools. To this end, they used the following standardized residual:

$$e_i \quad = \quad \frac{y_i - \bar{y}}{s},$$

where $s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$ is the usual sample variance. Note that marginally $y_i \overset{i.i.d.}{\sim} N(\mu, 1+A)$, where $\mu = \mathrm{E}(\theta_i)$. Under this marginal model,

$\mathrm{E}(e_i) = 0$, and $\mathrm{Var}(e_i) \approx 1$ for large m. Note that the e_i's are not strictly independent. Thus, if the model is reasonable, a plot of the standardized residuals versus the players is expected to fluctuate randomly around 0. Otherwise, we might suspect the adequacy of the two-level model. However, random fluctuation of the residuals may not reveal certain systematic patterns of the data. For example, Lahiri and Li (2005) noted that the residuals, when plotted against players arranged in increasing order of the previous batting averages, did reveal a linear regression pattern, something not apparent when the same residuals were plotted against players arranged in an arbitrary order. This is probably questioning the exchangeability assumption in the Efron-Morris model, a fact we knew earlier because of the intentional inclusion of an extremely good hitter.

Let p_{i0} be the batting average of player i through the end of 1969 season and $x_i = \sqrt{n}\arcsin(2p_{i0} - 1)$, $i = 1, \cdots, m$. We plot y and θ vs x in Figure 1(a) and 1(b) respectively. This probably explains the systematic pattern of the residuals mentioned in the previous paragraph. We also note the striking similarity of the two graphs 1(a) and 1(b). While Roberto Clemente seems like an outlier with respect to y, θ, or x, player L. Alvarado appears to be an outlier in the sense that his current batting average is much better than his previous batting average. He influences the regression fit quite a bit. For example, the BIC for the two-level model reduced from 55 to 44 when Alvarado was dropped from the model. Further investigation reveals that this player is a rookie and batted only 51 times through the end of 1969 season compared to other players in the dataset, making his previous batting average information not very useful. The BIC for the Fay-Herriot model with and without the auxiliary data are almost the same (54.9 and 55.3 respectively), a fact not expected at the beginning of the data analysis. For each player, Figure 2 displays true season batting average (True) along with the sample proportion (Direct) and different predictors obtained from the simple regression model (Reg), empirical Bayes estimator using the Efron-Morris model (EM) and the Fay-Herriot model that uses previous batting average as a covariate (EMReg). In spite of more or less similar BIC values and a presence of an outlier in the regression, Figure 2 shows that EMReg did a good job in predicting the batting averages of Clemente and Alvarado, two different types of outliers. Further details on this data analysis is given in Lahiri and Li (2005).

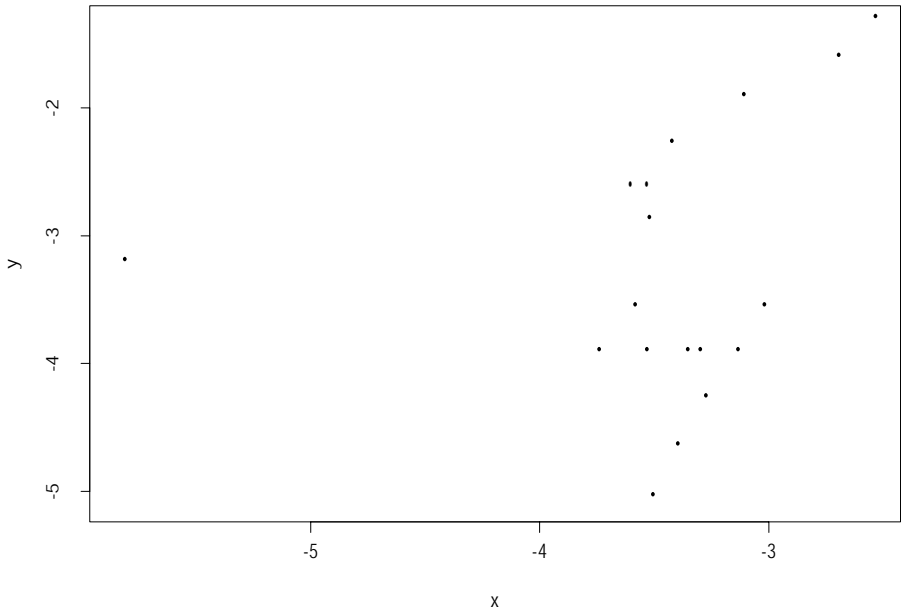

Figure 1(a):y versus x

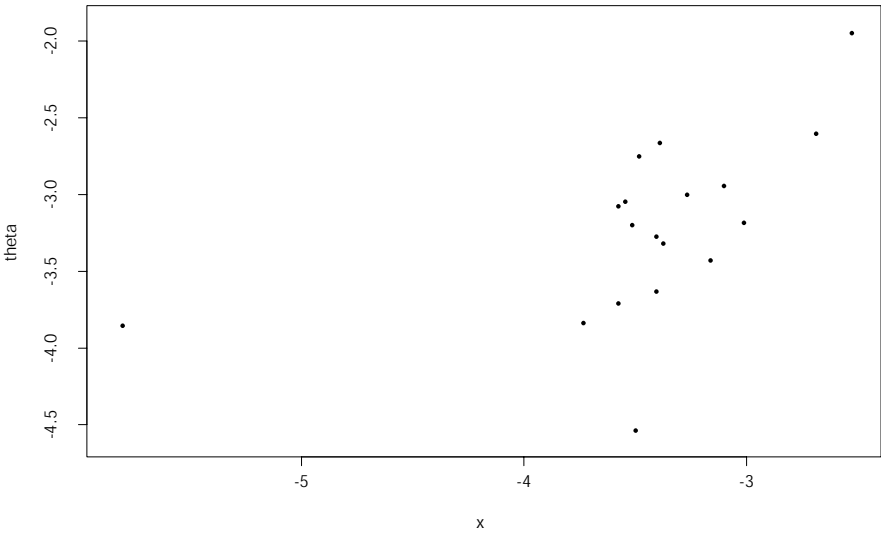

Figure 1(b):θ versus x

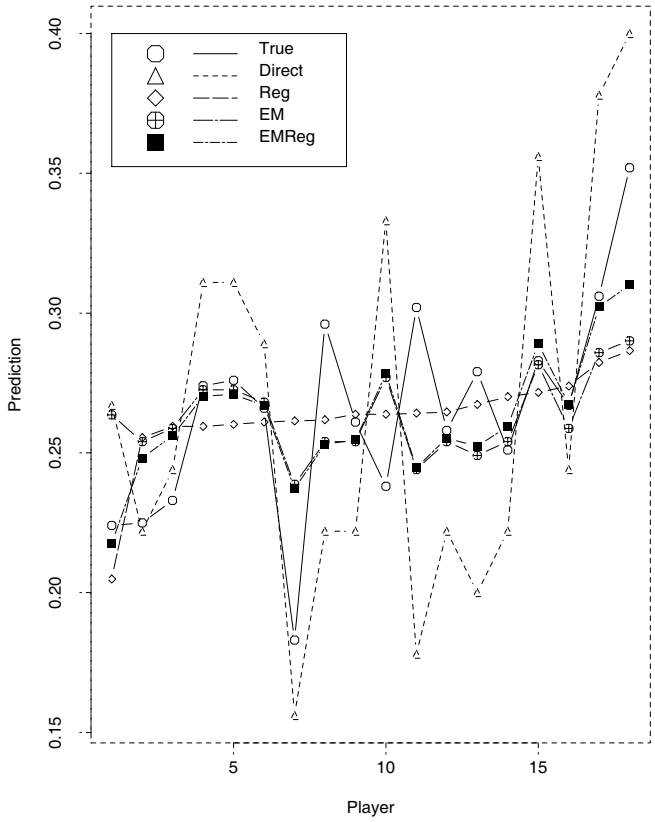

Figure 2: *Baseball example*

References

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

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70:343–365.

Basu, R., Ghosh, J. K., and Mukerjee, R. (2003). Empirical Bayes prediction intervals in a normal regression model: Higher order asymptotics. *Statistics & Probability Letters*, 63:197–203.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 80:28–36.

Beckman, R. J., Nachtsheim, C. J., and Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, 29:413–426.

Bell, W. (2001). Discussion with "Jackknife in the Fay-Herriott model with an example". In *Proceeding of the Seminar on Funding Opportunity in Survey Research*, pp. 98–104.

Booth, J. G. and Hobert, J. P. (1998). Standard errors of predictors in generalized linear mixed models. *Journal of the American Statistical Association*, 93:262–272.

Booth, J. G. and Hobert, J. P. (1999). Maximum generalized linear mixed model likelihood with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society. Series B*, 61:265–285.

Brackstone, G. J. (1987). Small area data: policy issues and technical challenges. In R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh, eds., *Small Area Statistics*, pp. 3–20. John Wiley & Sons, New York.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Butar, F. B. (1997). *Empirical Bayes Methods in Survey Sampling*. PhD Thesis, Dept. of Mathematics and Statistics, University of Nebraska-Lincoln.

Butar, F. B. and Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small area estimators. Model selection, model diagnostics, empirical Bayes and hierarchical Bayes, Special Issue II (Lincoln, NE, 1999). *Journal of Statistical Planning and Inference*, 112(1–2):63–76.

Calvin, J. A. and Sedransk, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *Journal of the American Statistical Association*, 86:36–48.

Carlin, B. P. and Gelfand, A. (1990). Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association*, 85:105–114.

Carlin, B. P. and Gelfand, A. (1991). A sample reuse method for accurate parametric empirical Bayes confidence intervals. *Journal of the Royal Statistical Society. Series B*, 53:189–200.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London.

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

Chatterjee, S. and Lahiri, P. (2002). Parametric bootstrap confidence interval in small area estimation problems. Unpublished manuscript.

Chaudhuri, A. (1994). Small domain statistics: a review. *Statistica Neerlandica*, 48:215–236.

Chen, S. and Lahiri, P. (2002). A weighted jackknife MSPE estimator in small area estimation. In *Proceedings of the Survey Research Methods Section*, pp. 473–477. American Statistical Association.

Chen, S. and Lahiri, P. (2003a). A comparison of different MSPE estimators of EBLUP for the Fay-Herriot model. In *Proceedings of the Survey Research Methods Section*, pp. 905–911. American Statistical Association.

Chen, S. and Lahiri, P. (2003b). Jackknifing the mean square prediction error of EBLUP for the Fay-Herriot model. Unpublished manuscript.

Chen, S. and Lahiri, P. (2005). On mean squared prediction error estimation in small area estimation problems. In *Proceedings of the Survey Research Methods Section*. American Statistical Association. To appear.

Chen, S., Lahiri, P., and Rao, J. N. K. (2005). On mean square prediction error estimation for a non-normal Fay-Herriot model. Unpublished manuscript.

Christensen, R., Pearson, L. M., and Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, 34:38–45.

Christiansen, C. L. and Morris, C. N. (1994). README document, June 7, 1994, a seven-page latex tutorial for PRIMM. Available through StatLib, `stat@lib.stat.cmu.edu`.

Christiansen, C. L. and Morris, C. N. (1996). Hierarchical Poisson regression modeling. In D. Berry and D. Stangl, eds., *Bayesian Biostatistics*, pp. 467–501. Marcel Dekker, New York.

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

Citro, C. F., Cohen, M. L., Kalton, G., and West, K. K. (1997). *Small-area estimates of school-age children in poverty: Interim report I-evaluation of 1993 county estimates for title I allocation*. National Academy Press, Washington.

Cox, D. R. (1975). Prediction intervals and empirical Bayes confidence intervals. In J. Gani, ed., *Perspectives in Probability and Statistics (Papers in honor of M. S. Bartlett)*, pp. 47–55. National Academic Press, Washington.

Crouch, E. A. C. and Spiegelman, D. (1990). The evaluation of integrals of the form $\int f(t)\exp(-t^2)dt$: Application to logistic normal models. *Journal of the American Statistical Association*, 85:464–469.

Das, K., Jiang, J., and Rao, J. N. K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32:818–840.

Datta, G. S. and Ghosh, M. (1991). Bayesian prediction in linear models: applications to small area estimation. *The Annals of Statistics*, 19:1746–1770.

Datta, G. S., Ghosh, M., Smith, D. D., and Lahiri, P. (2002). On an asymptotic theory of conditional and unconditional coverage probabilities of empirical Bayes confidence intervals. *Scandinavian Journal of Statistics*, 29:139–152.

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

Datta, G. S. and Lahiri, P. (2001). Discussions on a paper by Efron and Gous. In *Model Selection*, Vol. 38 of *IMS Lecture Notes/Monograph*.

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

Dempster, A. P. and Ryan, L. M. (1985). Weighted normal plots. *Journal of the American Statistical Association*, 80:845–850.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1996). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Efron, B. (1975). Biased versus unbiased estimation. *Advances in Mathematics*, 16:259–277.

Efron, B. and Morris, C. N. (1972). Limiting the risk of Bayes and empirical Bayes estimators. Part II: the empirical Bayes case. *Journal of the American Statistical Association*, 67:130–139.

Efron, B. and Morris, C. N. (1973). Stein's estimation rule and its competitors- an empirical Bayes approach. *Journal of the American Statistical Association*, 68:117–130.

Efron, B. and Morris, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70:311–319.

Fabrizi, E. and Lahiri, P. (2004). A new approximation to the Bayes information criterion in finite population sampling. Unpublished manuscript.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74:269–277.

Fuller, W. A. (1990). Prediction of true values for the measurement error model. *Contemporary Mathematics*, 112:41–57.

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

Gershunskaya, J. B. and Lahiri, P. (2005). Variance estimation for domains in the U.S. Current Employment Statistics Program. In *Proceedings of the Survey Research Methods Section*. American Statistical Association. To appear.

Ghosh, M. and Lahiri, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 82:1153–1162.

Ghosh, M. and Lahiri, P. (1992). A hierarchical Bayes approach to small area estimation with auxiliary information (with discussion). In *Bayesian Analysis in Statistics and Econometrics*, pp. 107–125. Springer–Verlag, Berlin.

Ghosh, M. and Meeden, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association*, 81:1058–1062.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman and Hall, London.

Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93:273–282.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9:55–93.

Glickman, H. and Pfeffermann, D. (2004). Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. In *Proceedings of the Survey Research Methods Section*. American Statistical Association. To appear.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, 73:43–56.

Hartley, H. O. and 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 components estimation and related problems. *Journal of the American Statistical Association*, 72:320–340.

Harville, D. A. (1990). BLUP (best linear unbiased prediction) and beyond. In D. Gianola and K. Hammond, eds., *Advances in Statistical Methods for Genetic Improvement of Livestock*, pp. 239–276. Springer–Verlag, New York.

Harville, D. A. (1991). Comment on Robinson: estimation of random effects. *Statistical Science*, 6:35–39.

Henderson, C. R. (1948). *Estimation of general, specific and maternal combining abilities in crosses among inbred lines of swine*. PhD Thesis, Iowa State University, Ames, Iowa.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *The Annals of Mathematical Statistics*, 21:309–310.

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

Henderson, C. R. (1973). Sire evaluation and genetic trends. In *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush*, pp. 10–41. Amer. Soc. Animal Sci. - Amer. Dairy Sci. Assoc. - Poultry Sci. Assoc., Champaign, IL.

Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15:192–218.

Heyde, C. C. (1994). A quasi-likelihood approach to the REML estimating equations. *Statistics & Probability Letters*, 21:381–384.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application*. Springer–Verlag, New York.

Hill, J. R. (1990). A general framework for model-based statistics. *Biometrika*, 77:115–26.

Hinde, J. (1982). Compound Poisson regression models. In R. Gilchrist, ed., *GLIM 82: Proceedings of the International Conference on Generalized Linear Models*, pp. 109–121. Springer–Verlag, Berlin.

Hinrichs, P. E. (2003). *Consumer expenditure estimation incorporating generalized variance functions in hierarchical Bayes models*. PhD Thesis, University of Nebraska-Lincoln.

Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55:179–188.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77:89–96.

Jeske, D. R. and Harville, D. A. (1988). Prediction-interval procedures and (fixed-effects) confidence-interval procedures for mixed linear models. *Communications in Statistics. Theory and Methods*, 17:1053–1087.

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

Jiang, J. (1997a). A derivation of BLUP — best linear unbiased predictor. *Statistics & Probability Letters*, 32:321–324.

Jiang, J. (1997b). Wald consistency and the method of sieves in REML estimation. *The Annals of Statistics*, 25:1781–1803.

Jiang, J. (1998a). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Statistica Sinica*, 8:861–885.

Jiang, J. (1998b). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, 93:720–729.

Jiang, J. (1999). Conditional inference about generalized linear mixed models. *The Annals of Statistics*, 27:1974–2007.

Jiang, J. (2000a). A matrix inequality and its statistical applications. *Linear Algebra and its Applications*, 307:131–144.

Jiang, J. (2000b). A nonlinear Gauss-Seidel algorithm for inference about GLMM. *Computational Statistics*, 15:229–241.

Jiang, J. (2001). Goodness-of-fit tests for mixed model diagnostics. *The Annals of Statistics*, 29:1137–1164.

Jiang, J. (2003). Empirical best prediction for small area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111:117–127.

Jiang, J. (2006a). Comment on Song, Fan and Kalbfleisch: Maximization by parts in likelihood inference. *Journal of the American Statistical Association*. In press.

Jiang, J. (2006b). Linear and generalized linear mixed models and their applications. Springer–Verlag, working monograph.

Jiang, J. (2006c). Partially observed information and inference about non-Gaussian mixed linear models. *The Annals of Statistics*. In press.

Jiang, J., Jia, H., and Chen, H. (2001a). Maximum posterior estimation of random effects in generalized linear mixed models. *Statistica Sinica*, 11:97–120.

Jiang, J. and 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. and Lahiri, P. (2006). Estimation of finite population domain means - a model assisted empirical best prediction approach. *Journal of the American Statistical Association*. In press.

Jiang, J., Lahiri, P., and Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics*, 30:1782–1810.

Jiang, J., Lahiri, P., and Wu, C. H. (2001b). A generalization of Pearson's χ^2 goodness-of-fit test with estimated cell frequencies. *Sankhyā. Series A*, 63:260–276.

Jiang, J. and Rao, J. S. (2003). Consistent procedures for mixed linear model selection. *Sankhyā. Series A*, 65:23–42.

Jiang, J. and Zhang, W. (2001). Robust estimation in generalized linear mixed models. *Biometrika*, 88:753–765.

Jiang, J. and Zhang, W. (2002). Distribution-free prediction intervals in mixed linear models. *Statistica Sinica*, 12:537–553.

Juan-Albacea, Z. V. (2004). Small area estimation of sub-national poverty incidence. In *Proceedings of the ADB-GTZ-CEPA Regional Conference on Poverty Monitoring*. Manila, Philippines.

Kackar, R. N. and Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics. Theory and Methods*, 10:1249–1261.

Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79:853–862.

Khuri, A. I. and Sahai, H. (1985). Variance components analysis: a selective literature survey. *International Statistical Review*, 53:279–300.

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

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

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society. Series B*, 57:395–407.

Lahiri, P. (1995). A jaccknife measure of uncertainty of linear empirical Bayes estimators. Unpublished manuscript.

Lahiri, P. (2001). *Model Selection*, Vol. 38 of *IMS Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Beachwood, Ohio.

Lahiri, P. (2003a). On the impact of bootstrap in survey sampling and small area estimation. *Statistical Science*, 18:199–210.

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

Lahiri, P. and Li, H. (2005). A fresh look at baseball data analysis. Unpublished report.

Lahiri, P. and Maiti, T. (2002). Empirical Bayes estimation of relative risks in disease mapping. *Calcutta Statistical Association Bulletin*, 53:211–212.

Lahiri, P. and Meza, J. L. (2002). Small-area estimation. In A. H. El-Shaarawi and W. W. Piegorsch, eds., *Encyclopedia of Environmetric*, Vol. 4, pp. 2010–2014. John Wiley & Sons, Chichester.

Lahiri, P. and Rao, J. N. K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90:758–766.

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82:739–757.

Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, 17:624–642.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society. Series B*, 58:619–678.

Li, H. (2005). Comparison of different small area estimators for complex survey data. Unpublished manuscript.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91:1007–1016.

Malec, D., Sedransk, J., Moriarity, C. L., and 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.

Marker, D. A. (1999). Organization of small area estimators using generalized regression framework. *Journal of Official Statistics*, 15:1–24.

Marshall, R. J. (1991). A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society. Series A*, 154:421–441.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2nd ed.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89:330–335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170.

McGilchrist, C. A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society. Series B*, 56:61–69.

Meza, J. L. (2003). Empirical Bayes estimation smoothing of relative risks in disease mapping. *Journal of Statistical Planning and Inference*, 112:43–62.

Meza, J. L., Chen, S., and Lahiri, P. (2003). Estimation of lifetime alcohol abuse for Nebraska counties. Unpublished manuscript.

Meza, J. L. and Lahiri, P. (2005). A note on the Cp statistic under the nested error regression model. *Survey Methodology*, 31:105–109.

Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of analysis of variance. *The Annals of Statistics*, 5:746–762.

Morris, C. N. (1983a). Parametric empirical Bayes confidence intervals. In *Scientific Inference, Data Analysis and Robustness*, Vol. 4, pp. 25–.50. Academic Press, New York.

Morris, C. N. (1983b). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78:47–59.

Mukhopadhyay, P. (1998). *Small Area Estimation in Survey Sampling*. Narosa Publishing House, New Delhi.

Nandram, B. (1999). An empirical Bayes prediction interval for the finite population mean of a small area. *Statistica Sinica*, 9:325–343.

Nandram, B. and Sedransk, J. (1993). Empirical Bayes estimation for the finite population mean on the current occasion. *Journal of the American Statistical Association*, 88:994–1000.

Natarajan, R. and McCulloch, C. E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika*, 82:639–643.

National Research Council (2000). *Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology*. National Academic Press, Washington.

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

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

Pfeffermann, D. (1984). On extensions of the Gauss-Markov theorem to the case of stochastic regression coefficient. *Journal of the Royal Statistical Society. Series B*, 46:139–148.

Pfeffermann, D. (2002). Small area estimation - new developments and directions. *International Statistical Review*, 70:125–143.

Pfeffermann, D. and Tiller, R. (2002). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. Technical Report, U.S. Bureau of Labor Statistics.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. Springer–Verlag, New York, 2nd ed.

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

Prasad, N. G. N. and Rao, J. N. K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*, 25:67–72.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44:1033–1048.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67:112–115.

Rao, J. N. K. (1986). Synthetic estimators, SPREE and the best model based predictors. In *Proceedings of the Conference on Survey Methods in Agriculture*, pp. 1–16. U.S. Dept. of Agricultur, Washington, D.C.

Rao, J. N. K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25:175–186.

Rao, J. N. K. (2001). EB and EBLUP in small area estimation. In S. E. Ahmed and N. Reid, eds., *Empirical Bayes and Likelihoodinference*, Vol. 148 of *Lecture Notes in Statistics*, pp. 33–43. Springer–Verlag, New York.

Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley & Sons, New York.

Richardson, A. M. and Welsh, A. H. (1994). Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. *Australian Journal of Statistics*, 36:31–43.

Rivest, L. P. and Belmonte, E. (1999). The conditional mean square errors of small area estimators in survey sampling. Technical Report, Laval University.

Robinson, D. L. (1987). Estimation and use of variance components. *Journal of the Royal Statistical Society. Series D. The Statistician*, 36:3–14.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science*, 6:15–51.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78:719–727.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. John Wiley & Sons, New York.

Shaw, R. G. (1987). Maximum-likelihood approaches applied to quantitative genetics of natural populations. *Evolution*, 41:812–826.

Singh, A. C., Stukel, D. M., and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society. Series B*, 60:377–396.

Song, P. X.-K., Fan, Y., and Kalbfleisch, J. D. (2005). Maximization by parts in likelihood inference (with discussion). *Journal of the American Statistical Association*, 100(472):1145–1158.

Speed, T. P. (1997). Restricted maximum likelihood (REML). In S. Kotz, C. B. Read, and D. L. Banks, eds., *Encyclopedia of Statistical Sciences*, Vol. 1, pp. 472–481. John Wiley & Sons, New York.

Särndal, C. E. (1984). Design-consistent versus model-dependent estimators for small domainss. *Journal of the American Statistical Association*, 79:624–6315.

Tabunda, A. M. L. and Abanilla, J. R. Y. (1997). A look at EBLUP and other small area estimates of poverty incidence at the city/municipality level. In *Proceedings of 10th SRTC Annual Conference*, pp. 73–97. Statistical Research and Training Center, Quezon City, Philippines.

Thompson, J., W. A. (1962). The problem of negative estimates of variance components. *Annals of Mathematical Statistics*, 33:273–289.

Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics*, 32:227–230.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–vi243.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer–Verlag, New York.

You, Y. and Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30(3):431–439.

Zeger, S. L. and Karim, R. M. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, 86:79–86.

DISCUSSION

Peter Hall
Mathematical Sciences Institute
Australian National University, Australia

This authoritative paper provides a review, at once broad and deep, of prediction methods for small-area inference. As the authors note, although this fascinating field of statistics is centuries old, it is as vital as ever. Indeed, the results of small-area analyses continue to motivate substantial budgetary commitments by many nations, and so are inevitably the subject of discussion by the media as well as in national statistical offices.

Professors Jiang and Lahiri are the architects of recent, very important methodology for small-area prediction, and in particular have made substantial contributions to sample reuse techniques there. Bootstrap methods can be taken a long way in this problem. Depending to some extent on the nature and quantity of the data, the bootstrap can provide adaptive and robust approaches to quantifying the variability of predictors, a topic to which Professors Jiang and Lahiri devote significant attention in their review.

Bootstrap techniques for small-area prediction can be either parametric or nonparametric. In this brief discussion we shall focus on the latter type, suggesting approaches to assessing predictor performance that are alternative to those treated by Hall and Maiti (2006a).

When replicated data are available, a Fay-Herriot model for the observations (X_{ij}, Y_{ij}) might assume the form,

$$Y_{ij} = f_i(\beta) + U_{ij} + V_i\,, \quad 1 \le i \le n\,, \quad 1 \le j \le n_i\,. \tag{1}$$

It is assumed that each $n_i \ge 2$; that f_i denotes a known, smooth function of the explanatory variables $X_{i1}, \ldots, X_{in_i}$; β is a vector of unknown parameters; U_{ij} and V_i are independent and distributed as U and V, respectively, both having zero mean; and $\Theta_i = f_i(\beta) + V_i$ denotes the true value of the small-area mean.

In this setting the distributions, rather than just the variances, of U and V may be estimated nonparametrically. One approach, valid under the assumption that the distribution of U has a real-valued, positive characteristic function χ, say, is to observe that for each pair (j_1, j_2) for which $1 \le j_1 < j_2 \le n_i$, we have:

$$Y_{ij_1} - Y_{ij_2} = U_{ij_1} - U_{ij_2} \,.$$

Under the assumptions made about χ, this random variable has characteristic function χ^2, and so $\chi(t)^2$ can be estimated by

$$\widehat{\chi}(t)^2 = \max \left[\frac{2}{N} \sum_{i=1}^{n} \sum_{1 \le j_1 < j_2 \le n_i} \Re \exp \left\{ \sqrt{-1}\, t \, (Y_{ij_1} - Y_{ij_2}) \right\}, 0 \right],$$

where $N = \sum_i n_i(n_i - 1)$ and $\Re\, z$ denotes the real part of z. (A weighted version of this estimator can give better results.) We take $\widehat{\chi}$, the positive square root of $\widehat{\chi}^2$, to be our estimator of χ, and compute an estimator, $\widetilde{F}_U$ say, of the distribution function F_U of U, by regularised Fourier inversion of $\widehat{\chi}$. Finally we "monotonise" $\widetilde{F}_U$, obtaining an estimator $\widehat{F}_U$ which is nondecreasing from zero to one. For example, we might take

$$\widehat{F}_U(v) = \max \left[\min \left\{ \sup_{x \le v} \widetilde{F}_U(x), 1 \right\}, 0 \right].$$

An alternative way of estimating the distribution of U is to modify methods suggested by Li and Vuong (1998) in a quite different context. These and other approaches permit the assumption that χ is real and positive to be removed. See also El-Amroui and Goffinet (1991) and Beran and Hall (1992).

The distribution function $\widehat{F}_U$ can be used to generate random variables having a distribution closely approximating F_U. To do the same for the distribution of V we should first estimate that quantity's distribution. To this end, note that $\widehat{W}_{ij} \equiv Y_{ij} - f_i(\hat{\beta})$ is a good approximation to $W_{ij} = U_{ij} + V_i$; here, $\hat{\beta}$ denotes the least-squares estimator of β, computed from the model (1). Using a deconvolution method, starting from the data $\widehat{W}_{ij}$ and employing the estimator $\widehat{F}_U$, we may compute an estimator, $\widehat{F}_V$ say, of the distribution function of V.

By sampling from the distributions $\widehat{F}_U$ and $\widehat{F}_V$, after centring at their respective means, we may generate bootstrap variables U^*_{ij} and V^*_i, and

thence $\Theta_i^* = f_i(\hat{\beta}) + V_i^*$ and $Y_{ij}^* = \Theta_i^* + U_{ij}^*$. From these bootstrap data we can produce predictors of Θ_i in the "bootstrap world". A bootstrap estimator of mean-squared prediction error is given by the mean-squared error of these bootstrap predictors, conditional on the original data pairs (X_{ij}, Y_{ij}). Other descriptions of prediction error, such as prediction regions, can also be constructed using bootstrap methods. Versions of these methods that apply to more general models can also be developed, for example in the setting where, in (1), U_{ij} is replaced by $\eta_i U_{ij}$ and η_i denotes a known standard deviation (for example, a known function of the explanatory variables $X_{i1}, \ldots, X_{in_i}$).

However, in the setting of small-area inference there may well remain, in a broad range of very important problems, significant advantages to using parametric approaches. After all, small-area methods are motivated when data are scarce, and it is exactly in such cases that informed parametric techniques can enjoy statistical advantages over their more adaptive nonparametric cousins, which are relatively demanding of data. Parametric bootstrap methods therefore have an important role to play. Contributions to that topic, often founded on Bayesian approaches, include those of Butar and Lahiri (2003), Lahiri (2003a), Meza (2003) and Hall and Maiti (2006b).

Additional references

Beran, R. and Hall, P. (1992). Estimating coefficient distributions in random coefficient regressions. *The Annals of Statistics*, 20:1970–1984.

El-Amroui, A. and Goffinet, B. (1991). Estimation of the density of g given observations of $y = g + e$. *Biometrical Journal*, 33:347–355.

Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics*. To appear.

Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small-area prediction. *Journal of the Royal Statistical Society. Series B*. To appear.

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis*, 65:139–165.

Domingo Morales
Operations Research Center
University Miguel Hernández of Elche, Spain

Small Area Estimation (SAE) is a topic of great interest in many fields, like public statistics, agriculture or disease mapping, because of the increasing demand of data at disaggregated level. Under the prediction approach SAE procedures use models to borrow strength from auxiliary variables, neighbor regions or past observations. At the same time target parameters are defined in finite population, so most branches of statistics converge in this field and makes SAE so attractive. In the last 15-20 years a lot of research has been done in SAE and still this is a rich field in research problems. Many review papers and two books have been published but, because of the dynamic state of the art in SAE, present review covers a great deal of recently published new material.

Professors Jiang and Lahiri have succeed in providing an excellent overview of the prediction theory in linear and generalized linear models. They also have put special emphasis in SAE and related topics. They introduce area-specific and unit-specific mixed models. They consider linear and generalized linear mixed models paying attention in both cases to model parameters estimation, prediction of random effects and assessing uncertainty of EBLUP and EBP. The paper analyzes also several adjustments to EBP and treats the problem of obtaining prediction intervals and building models. In what follows I will point out some issues that may complement the review.

On behalf of their flexibility, nonparametric models have been scarcely used in SAE. In the near future we will probably see a lot of researchers investigating applications of let say kernel regression, splines, orthogonal series or nearest neighbor methods to SAE. However, semiparametric regression models, and in particular extensions of the parametric generalized mixed models, will have in my opinion an increasing role in SAE. In fact this kind of models compromises two aims, flexibility and simplicity of statistical procedures. So theory of predicting random effects and of assessing uncertainty should be adapted or newly developed for the possibly new families of models.

Prediction theory produce much simpler SAE procedures for linear mixed models than for generalized linear mixed models. In the first case problems of predicting random effects and of assessing uncertainty are solved via EBLUP and formulas for $g_1 - g_4$. In the second case everything becomes more complicated to practitioners: election of the fitting algorithm, construction of best predictors of a possibly nonlinear mixed effect which do not have a closed form (cf. Section 4.4) and assessing its uncertainty. At this point one would like to see results of simulations experiments illustrating the gain of using EBP with respect to other naive or more straightforward methods. For example, in Example of Section 4.5 one should be tempted to estimate p_i with

$$\widehat{p_i} = \frac{1}{N_i}\left(y_{i.} + \sum_{j \in s_{ir}} \frac{\exp\{x'_{ij}\widehat{\beta} + \widehat{\alpha}_i\}}{1 + \exp\{x'_{ij}\widehat{\beta} + \widehat{\alpha}_i\}}\right),$$

or, in the case $x_{ij} = x_i$, with

$$\widehat{p_i} = \frac{1}{N_i}\left(y_{i.} + (N_i - n_i)\frac{\exp\{x'_{i}\widehat{\beta} + \widehat{\alpha}_i\}}{1 + \exp\{x'_{i}\widehat{\beta} + \widehat{\alpha}_i\}}\right),$$

or even with the plug-in estimator $\widehat{p}_i = p_i(\widehat{\beta}, \widehat{\alpha}_i)$. The question to be answered here is how much is lost by using these simpler estimators. Or equivalently when it is really worthwhile to use the EBP of Jiang (2003). The problem of assessing the uncertainty was also treated by him and formulas for MSPE and order of convergence are given in (4.7) and (4.8) of review. Here it is interesting to check numerically the behavior of (4.7) against resampling methods (for example, parametric bootstrap or jackknife) via simulation experiments in a similar way as for Prasad-Rao method has been done. In González-Manteiga et al. (2005) related topics are treated.

Calibrated weighting methods for estimation of survey population characteristics are widely used in Statistical Offices. Main reason is that weighted survey estimators for large areas are generally quire reliable. Benchmarking property, guaranteeing that estimations of, let say, totals of large areas can be obtained by summing up estimated totals of its contained small areas, is thus desirable to have consistency of published data. However model-based estimators do not have in general this property. At this point one can cite the working paper of Chambers (2005), where he explores the possibility of using mixed models to construct calibrated sample weights

and investigates the performance of the corresponding weighted survey estimator together with the EBLUP in both population levels: small and large areas.

I agree with the authors that model building and diagnosis are one the most complex parts in any scientific investigation that uses a statistical model. So I welcome Section 6.2 and at the same time I thank Professors Jiang and Lahiri for witting the paper and for their effort in reviewing the given references.

Finally I would like to thank the editors of TEST for giving me the opportunity to read and discuss on this paper.

Additional references

Chambers, R. L. (2005). Calibrated weighting for small area estimation. Methodology Working Paper Series M05/04, Southampton Statistical Sciences Research Institute, University of Southampton, United Kingdom.

González-Manteiga, W., Lombardía, M. J., Morales, D., and Santamaría, L. (2005). Estimation of the mean squared error of predictors of small area linear parameters under logistic mixed model. Submitted.

Danny Pfeffermann
The Hebrew University of Jerusalem, Israel
and University of Southampton, U.K.

This article provides a very through and comprehensive account of the literature on mixed model inference, with particular attention to small area estimation (SAE). The theory and application of mixed models to which Professors Jiang and Lahiri have made major contributions is far from being simple, particularly when dealing with generalized linear mixed models, but the authors found the way to describe it in a very friendly manner, without compromising on important theoretical and technical details.

In the discussion that follows I first make some comments on topics overviewed in the paper and then I consider briefly two other classes of

models that haven't been overviewed. I concentrate mostly on SAE problems.

1 Use of hierarchical Bayes approach

Professors Jiang and Lahiri mention the use of the full Bayesian approach in the introduction but it is important to list the advantages of this approach. Bayesian inference generates observations from the joint posterior distribution of the small area parameters and thus broadens the possibilities of statistical inference far and beyond what can be done, or at least has been done so far under the frequentist approach. In particular, the use of this approach enables simple computation of the posterior variance of the small area predictors and accurate confidence intervals, without the need to rely on large sample approximations, and it permits also ranking of the true area means, thus addressing the 'triple goal' estimation challenge discussed in Rao (2003) book.

Clearly, the application of this approach is not without a price and Professors Jiang and Lahiri mention some of the problems underlying its use, including the specification of prior distributions for the unknown model hyper-parameters. In most applications the prior distributions are noninformative, and a major problem that I have encountered is that with a small number of sampled areas, different noninformative prior distributions for the variance components can yield different variance components estimates and different predictors and MSE estimates. As far as I can tell, this is still by and large an open problem in Bayesian inference.

2 Use of design-based methods for SAE

Design based methods are mentioned in this paper in various connections and the following points are worth bearing in mind.

1. There exists a founded design-based theory for small area estimation, but only for areas with sampled units (see below). In particular, direct design based estimators that only use the outcome values in the area under consideration are unbiased, or approximately unbiased over repeated sampling (with the population values held fixed).

The obvious disadvantage of these estimators is their large variances when the area sample sizes are small, unless in cases where auxiliary variables with good prediction power are available.

2. Indirect estimators like synthetic or combined estimators borrow strength across areas, but they could be severely biased over repeated sampling, thus inflating the MSE. Estimating the MSE is very problematic in such cases.

3. Another major drawback of design-based inference is that there is no design-based theory of prediction for nonsampled areas. In many practical applications, estimators and MSE estimators are required for many areas, of which only some are represented in the sample. For example, in the SAIPE and NHANES III surveys described in the paper, the small areas are counties and there are samples from only about 35% of the counties in the first survey and only 3% of the counties in the second survey. This problem exists irrespective of the sample sizes within the sampled areas. Classical design-based theory is not suited for prediction problems.

The use of models permits computing optimal predictors for nonsampled areas under the model, but notice that for models for which the random effects are independent between the areas, the optimal predictor for a nonsampled area is a again a *synthetic* estimator. Thus, synthetic estimators play an important role even with the use of mixed models that contain area random effects.

3 Estimation of the MSE of the EBLUP

Professors Jiang and Lahiri discuss alternative methods of estimating the variance components and hence the MSE of the resulting EBLUP. The MSE estimators considered have bias of correct order, but it is important to emphasize that different methods of estimating the variance components result in different precision of the EBLUP and different precision of the EBLUP MSE estimators. Datta et al. (2005) illustrate the better performance of the Fay Herriot predictor when estimating the variance component by the Fay Herriot method as compared to estimating it by the method of Fitting of Constants. Pfeffermann and Glickman (2004) extend the empirical

illustrations in Datta et al. (2005) by considering other distributions for the model error terms and other MSE estimators, like the Jackknife and bootstrap estimators, and obtain similar results, with another important finding that an EBLUP MSE estimator with a lower bias than another EBLUP MSE estimator does not necessarily have a lower MSE. This finding is explained by the fact that different variance component estimators and different MSE estimators require different bias corrections, thus inflating the MSE in different ways.

The asymptotic variance of the variance components estimators can generally be compared, but comparing the MSE of MSE estimators analytically may not be feasible, particularly for generalized linear mixed models. However, the MSE of the MSE estimators can and should be explored empirically, thus providing users with better guidelines of which variance components and MSE estimators to use in practical applications.

4 Benchmarking of model dependent small area estimators

Benchmarking the small area estimators is important for two reasons. First, it guarantees consistency in publication in the sense that the model-based small area predictors add up to the published design based estimator in a large area that comprises the small areas and for which the design-based estimator is deemed accurate. Second, it provides some protection against model breakdown; see Pfeffermann and Tiller (2005b) for a striking real example. A common benchmarking procedure mentioned in the paper, known as raking (or prorating) is defined as

$$\hat{\theta}_{i,\,\text{rake}} = \hat{\theta}_{i,\,\text{model}} \frac{\sum_i w_i \hat{\theta}_{i,\,\text{design}}}{\sum_i w_i \hat{\theta}_{i,\,\text{model}}},$$

where $\sum_i w_i \hat{\theta}_{i,\,\text{design}}$ is the design based estimator in the large area. The use of this procedure is sensible, but estimation of the MSE of the raked estimator seems complicated and can probably be best implemented by use of resampling methods.

For the unit level mixed model defined in Section 2.2, You and Rao (2002) show that by modelling the weighted estimator $\bar{y}_{iw} = \frac{\sum_j w_{ij} y_{ij}}{\sum_j w_{ij}}$, the resulting small area predictors satisfy the benchmarking property without further adjustments, thus solving also the problem of MSE estimation of

the benchmarked estimators. However, the use of weighted estimators as the input data instead of the individual observations y_{ij} may result in loss of efficiency, particularly under generalized linear mixed models. Also, as Professors Jiang and Lahiri point out, it may not be possible in this case to extract the model holding for the weighted estimator from the model holding for the individual observations.

Pfeffermann and Barnard (1991) and Pfeffermann and Tiller (2005b) force the benchmarking constraints by including them among the model equations, thus allowing the computation of the benchmarked estimators and their MSEs (assuming known variance components) as part of the model fitting process. The first article shows that there are no small area predictors with smallest variance in the class of all the linear unbiased predictors satisfying the benchmarking constraints. Note that the raked estimator considered earlier does not belong to this class. Pfeffermann and Tiller (2005a) propose parametric and nonparametric bootstrap procedures for estimating the MSE of the benchmarked estimators that account for model hyper-parameters estimation.

5 Design consistency of model dependent small area estimators

Advocating design consistency in SAE is in my view irrelevant. The problem of SAE is the small sample sizes and it makes therefore little sense to consider the case of large samples. The real issue requiring weighting or other related procedures is selection bias under informative sampling or non-response, where the observed data are sampled with differential probabilities that are related to the outcome values. Using the direct weighted estimator as the input data in a model based framework instead of the individual observations is one way to protect against such bias, but at the expense of possibly inflating the variance. See the last comment for an alternative approach that permits modelling the individual observations.

6 Use of time series models for SAE

Professors Jiang and Lahiri didn't consider times series models, but notice that all the linear cross-sectional models overviewed in this article can be

presented in state-space form, and all the times series, cross-sectional models proposed in the literature for SAE are linear mixed models. See Sallas and Harville (1981) for other illuminating examples. What distinguishes time series models from the mixed models considered by Professors Jiang and Lahiri is that the relationship between the random effects operating in a given area in different time points is also modeled (say, as an autoregressive process), which enables borrowing information from past surveys.

Many of the surveys used for SAE are monthly or quarterly surveys, which calls for the use of past data to further strengthen the estimator in any given time point. My own experience shows that the use of past data often yields much better predictors than the use of cross-sectional data, but one can attempt modelling jointly the time series and cross-sectional relationships between the small area parameter of interests and between the sampling errors. See Pfeffermann (2002) and Rao (2003) for review of time-series, cross-sectional models that have been used for SAE. The use of the Kalman filter or smoothing algorithms produces the small area estimators and their 'naive' MSE estimators for any given time point. Pfeffermann and Tiller (2005a) developed parametric and nonparametric bootstrap procedures for MSE estimation of correct order that account for the estimation of the model hyper-parameters.

7 SAE under informative sampling

Basically, all the models considered in the literature for SAE assume, at least implicitly, that the model holding for the population values holds also for the sample data. This is true even for studies where the input data are the weighted estimators $\bar{y}_{iw}$. As already mentioned, it is often the case that samples are taken from only part of the areas and it is also common that units sampled within the selected areas are sampled (or respond) with unequal probabilities. When the sampling of areas (within the areas) is with unequal selection probabilities that are related to the true area means (the outcome values), the sampling process is 'informative' and the model holding for the sample data is then different from the population model. Ignoring the sampling process in such situations biases the predictors and the MSE estimators.

The use of weighted estimators like $\bar{y}_{iw}$ is one possible way of accounting for informative sampling within the selected areas, but the variances of

the weighted estimators have to be computed in this case under the randomization (repeated sampling) distribution. As mentioned before, the use of weighted estimators may result in loss of efficiency and for generalized linear mixed models the model holding for the weighted estimators cannot usually be extracted from the model holding for the individual observations.

Pfeffermann and Sverchkov (2005) propose a general approach to deal with the problem of informative sampling. The approach consists of modelling the distribution of the random effects for the sampled areas and the distribution of the observed outcome values, and then predicting the area means for the nonsampled areas and the outcome values for nonsampled units within the selected areas, using the relationships between the *sample distribution* and the *sample-complement distribution* that they developed in previous articles. The use of this approach is shown to perform well in a simulation study both in term of point prediction and confidence intervals.

Additional references

Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9:73–84.

Pfeffermann, D. and Glickman, H. (2004). Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. In *Proceedings of the Survey Research Methods Section*. American Statistical Association.

Pfeffermann, D. and Sverchkov, M. (2005). Small area estimation under informative probability sampling of areas and within the selected areas. Working Paper Series M05/18, Southampton Statistical Sciences Research Institute, University of Southampton, U.K.

Pfeffermann, D. and Tiller, R. (2005a). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26:893–916.

Pfeffermann, D. and Tiller, R. (2005b). Small area estimation with state-space models subject to benchmark constraints. Working Paper Series M05/14, Southampton Statistical Sciences Research Institute, University of Southampton, U.K.

Sallas, W. H. and Harville, D. A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76:860–869.

Carl N. Morris
Statistics Department
Harvard University, USA

Analyses of small area data abound with predictions that underlie major decisions. While multi-level models provide clear advantages for describing such data, their analyses are complex and must be based on the most reliable fitting procedures. This review of over 160 papers that relate to small area mixed model prediction theory, to which Professors Jiang and Lahiri themselves have contributed mightily, provides a much-needed overview for small area researchers, and it will inspire the involvement of others. Explicitly and implicitly, the authors add their own valuable interpretations on the many methods and approaches available to small area practitioners.

No review of this length could address all the multi-level literature relevant to small-area prediction. The approach to fitting hierarchical models via Bayesian simulation techniques receives little attention here. Multi-level models that incorporate spatial auto-correlation and neighborhood information are skipped. In the latter case, aren't distance considerations very helpful for small-area data? Or, is it simpler and just as effective to fit separate (but with m smaller) models in different political areas? Handling topological features is complex, but it is facilitated in two-level models, as those features only need Level II specifications. In the former case, de-emphasizing Bayesian simulations is natural in a review aimed at methods simple enough to permit evaluations. Simple models and methods developed for repeated use have several advantages: they enhance the understanding and confidence of practitioners; they reduce computer time needed to analyze large data sets and to facilitate iterative model-building; they aid fast computations needed for frequency evaluations; and they enable theoretical and mathematical evaluations.

As Professors Jiang and Lahiri observe, some widely-used approaches, including the MLE, can work poorly for a limited number m of groups. For three reasons, the MLE yields biased estimates of A ("A" is the between

groups variance defined by the notation of Section 3). Unfortunately, all three reasons produce biases in the same non-conservative direction, i.e. they all underestimate A. The problem is exacerbated, as Professors Jiang and Lahiri point out, because MLE calculations also underassess the variance components of mixed-model predictions, this problem being worst when the number of small areas (m) is minimal, and/or when A is small. First, bias occurs because the likelihood function for A almost always is right-skewed, putting MLE estimates of A to the left of the center of the likelihood - substantially so if A is small. Second, the convex dependence on A of a shrinkage factor $B = D/(D + A)$ further increases this non-conservative bias, pushing shrinkage estimates yet more optimistically toward 1.00. Third, the MLE does not account for uncertainty in estimates of fitted first-order (mean) Level II parameters, which further biases the MLE of a shrinkage factor. As Professors Jiang and Lahiri explain, REML corrections are recommended to combat the third bias and the variance component "$g_2(A)$" in Section 3.4.

The first two biases can be addressed via an easy adjustment, as follows. Think of the REML-adjusted likelihood $L(A)$ as the (marginal) posterior density of A (it is, for the prior recommended below). Then maximizing $L(A)$ occurs at the posterior mode. In the setting of Section 3, $L(A)$ is right-skewed, noticeably so if m is not large, so the mean of A exceeds its mode. Maximizing $A * L(A)$ approximates the mean of A better than the does the mode. This multiplication by A corrects for underestimation of A and also for convexity of B_i. It derives from an extension of MLE methods, termed "ADM" (adjustment for density maximization), applied in this case to estimate shrinkage factors. Unlike estimates from kindred MLE and REML maximizations, ADM estimates of A are always positive and shrinkage factors automatically are constrained to [0, 1]. ADM also addresses the second flaw above because it is set up to approximate the mean of a shrinkage constant, not the mean of A, so it addresses directly the nonlinear dependence of B_i on A, i.e. with $B_i = D_i/(A+D_i)$. In effect, A is over-estimated by just enough to allow for this curvature.

The ADM approach, developed but not named in (Morris, 1988), allows for fitting any Pearson family of distributions via maximization and 2^{nd} derivatives. Because shrinkage factors are constrained to [0, 1], Beta distributions are preferred to Normal distributions. In fact, when m is not large (and sometimes even when m is large), the Normal approximations of shrinkage factor distributions fall noticeably outside [0, 1]. Parallel to

MLE/REML fitting of shrinkage factors, the ADM also uses two derivatives of the likelihood, but now to determine the two parameters of an approximating Beta distribution.

The formal Bayes rules that stem from Stein's superharmonic prior which specifies A to have a flat prior distribution on $[0, \infty)$, have good frequency properties when used in two-level Normal models. With this prior, $L(A)$ actually becomes the posterior density of A and underpins the motivation to maximize $A * L(A)$. Morris and Tang, in (Tang, 2002, Harvard dissertation) show that the resulting ADM shrinkage factor estimates of B_i work much better than those based on MLE and REML methods. That is, for Normal models, point and interval estimates stemming from an ADM application with the superharmonic prior have much better accuracies and coverages in frequency evaluations than MLE and REML estimates. The variance components "g_3" that Professors Jiang and Lahiri review in Section 3.4 are handled directly by this approach, leading to simple prediction intervals that meet or exceed their nominal coverage requirements, even for small m. Graphs in (Tang, 2002) in the James-Stein setting show that the resulting nominal 95% confidence intervals have coverages in repeated sampling that meet or exceed 95% for any $m \geq 4$ and for all values of $A > 0$.

It is very helpful that the review also covers many papers and methods for non-Normal responses, including GLMMs for exponential families. For those models too, the REML/ADM improvement on maximum likelihood applies when conjugate prior distributions are specified at Level II. Shrinkage constants exist if conjugate priors are specified at Level II and then the posterior means are linear in the shrinkages. For NEF-QVF distributions that use conjugate priors, the posterior variances are (at most) quadratic functions of the shrinkages B_i (Morris, 1983). Conjugate priors also mitigate Professors Jiang and Lahiri's observation in Section 4 that difficult integrals are required for such models.

Christiansen and Morris (1997) develop and apply the ADM approach to Poisson models. The general strategy described in Section 2 there includes: how the ADM is used to approximate shrinkage factors; what multiple of the REML likelihood is used before maximization; and how to extend the superharmonic prior (as a limit of uniform shrinkage priors). The nominal 95% confidence intervals determined by this approach to the two-level Poisson model are shown in Christiansen and Morris (1997) to have almost exact 95% frequency coverages for $m = 15$, as vividly illustrated in Fig-

ure 1 there. By contrast, in the same application, commonly-used empirical Bayes interval estimates based on MLE and REML calculations that claim 95% coverages are badly over-optimistic – actual coverages being only 65% and 75%.

Approximations and evaluations in multi-level models that are based on asymptotics for large m won't distinguish among good methods and bad methods when m is small. In many applications, m cannot be large. Even when many small areas are under consideration, the number of small areas may have to be limited in order to justify the Level II exchangeability assumptions governing distributions of the m Level I parameters. (Of course exchangeability is only required of Level II error terms, perhaps after a linear regression.) The models described assume all m Level II (potentially distinct) variances to be the same, i.e. equal to A. Similarly, taking advantage of neighborhood information may limit m because heterogeneity (arising from farther locations) inhibits shrinkages, risking forfeiture of the advantages of borrowing information across small areas. Professors Jiang and Lahiri recognize this exchangeability concern by reviewing model checking, and by describing limited translation methods that minimize the negative effects of shrinking legitimate outliers. Can large numbers m of small areas be combined safely in most applications? In actual small area applications, how often is m large enough to justify asymptotics in m?

Professors Jiang and Lahiri's insightful review reveals the dilemmas concerning the number m of individual areas being grouped. This invokes the fascinating interplay between the adequacies of models, of model-fitting, of asymptotics, and of computation that arises in all of statistical theory.

Additional references

Morris, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics*, 11(2):515–529.

Morris, C. N. (1988). Approximating posterior distributions and posterior moments. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds., *Bayesian Statistics*, Vol. 3, pp. 327–344. Oxford University Press, Oxford.

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

J. N. K. Rao
School of Mathematics and Statistics
Carleton University, Canada

This paper provides an excellent overview and appraisal of mixed model prediction in the context of small area estimation. The latter topic has received considerable attention in recent years due to growing demand for reliable small area statistics. Professors Jiang and Lahiri are among the leading researchers in small area estimation.

Section 3.1 discusses the estimation of variance components in a linear mixed model and gives a concise account of Jiang's (2006c) new method for estimating the asymptotic covariance matrix (ACM) of REML estimators under non-normality, avoiding the evaluation of third and fourth moments of the random effects. The new method should be practically useful in diverse areas of applications of linear mixed models since the estimated ACM is often used for inference on variance components. Section 3.2 deals with the important topic of EBLUP estimation of small area parameters that can be expressed as linear combinations of fixed and random effects in the model. The focus is on mean squared prediction error (MSPE) estimation and this is nicely done in Section 3.4 using the basic Fay-Herriot area level model. However, general results of Das et al. (2004) on MSPE estimation under the general linear mixed model are also mentioned for the benefit of the reader.

The re-sampling MSPE estimators (the jackknife, the weighted jackknife and the parametric bootstrap) are useful additions to the original Prasad and Rao (1990) nearly unbiased MSPE estimator. The authors claim that the jackknife MSPE estimators do not require the assumption of normality unlike the other estimators. This is not necessarily true because orthogonal decomposition of MSPE is needed for getting the jackknife estimator. The orthogonal decomposition may not hold under non-normality because of a covariance term of the order $O(m^{-1})$. The linearization method of Chen et al. (2005) handles the non-normal case using

information on the kurtosis of the sampling errors and model errors, as noted by the authors.

The authors refer to my unpublished work (Hwang and Rao, 1987) on the estimation of the first level MSPE, denoted CMSPE. Note that CMPE refers to the expectation of squared error with respect to sampling errors treating the θ_i's as fixed, unlike MSPE. Also, the CMSPE is different from the conditional MSPE of Fuller (1990). The latter is obtained by averaging over both levels but conditioning on the area-specific survey estimator y_i. Hwang and Rao have shown that the Prasad-Rao MSPE estimator is more stable than an unbiased estimator of CMSPE but it could perform poorly in tracking CMSPE for an outlying small area. Professors Jiang and Lahiri note that the poor performance of the Prasad-Rao MSPE estimator under an outlier could be due to the fact that the Prasad-Rao MSPE estimator is not area specific in terms of the direct survey estimator y_i. Unfortunately, area-specific MSPE estimators, such as the jackknife and its variations, do not help in this respect because the leading term, $g_{1i}(\hat{A})$, in the MSPE estimators is not area-specific unlike in the CMSPE estimator; only the lower order term estimating $g_{3i}(A)$ is area-specific. It is interesting to note in this connection that an area-specific estimator of $g_{3i}(A)$ is simply obtained by writing $g_{3i}(A)$ as

$$g_{3i}(A) = D_i^2(D_i + A)^{-4}E(y_i - x_i'\hat{\beta})^2V(\hat{A})$$

and then estimating the middle expectation term by $(y_i - x_i'\hat{\beta})^2$, as noted by Rao (2001). The resulting estimator $g_{3i}^*(\hat{A}, y_i)$ is area-specific but less stable than the Prasad-Rao $g_{3i}(\hat{A})$, but its instability will not affect the corresponding MSPE estimator since it is a lower order term. It is also interesting to note that this area-specific MSPE estimator is closely related to the area-specific estimator of Fuller (1990) obtained by conditioning on y_i only (Rao, 2003, p. 132).

Section 4 gives a nice account of small area estimation under generalized linear mixed models (GLMM) that are relevant for unit level models, particularly for binary responses. Professors Jiang and Lahiri discuss the difficulties in computing ML estimators of model parameters due to high dimensional integration. However, Monte Carlo EM methods of McCulloch (1997) and Booth and Hobert (1999) should work well for small area estimation because the GLMM models that are used are fairly simple. Simpler non-likelihood methods studied in Section 4.3 are also useful and may be needed in more complex small area problems. Example 4.5 demonstrates

how to obtain the EBP of small area proportion p_i under a nested error logistic regression model for binary response, and shows that the EBP can lead to significant gains in efficiency over the naïve direct estimator $\bar{y}_i$. The authors also make the important point that the EBP is not equal to p_i with β and the random effect α_i replaced by their estimators.

Regarding MSPE estimation, Jiang (2003) extended the Prasad-Rao linearization method to GLMM models and Jiang et al. (2002) used the jackknife method, but the leading term of the MSPE estimators is not area-specific. Rao (2003, p. 199) proposed a simple modification to the jackknife that leads to a computationally simpler MSPE estimator with area-specific leading term. Booth and Hobert (1998) criticized the unconditional approach under the GLMM models by arguing that the posterior variance of the small area parameter (function of fixed and random effects) given the area-specific data and model parameters is area-specific, unlike in the linear mixed model case. They proposed to estimate the conditional MSPE given the area-specific data, similar to Fuller (1990) for the linear case. The area-specific MSPE estimator of Rao (2003) is simpler, and simulation results by Lohr and Rao (2003) indicated that it has small conditional bias unlike the unconditional jackknife, but its unconditional coefficient of variation is slightly larger.

Section 5 deals with issues related to adjustments to EBP. In particular, the use of survey weights in area level models to ensure design-consistency and benchmarking is discussed. Professors Jiang and Lahiri refer to You and Rao (2002) where a pseudo-EBLUP for a nested error linear regression model was proposed. Those small area estimators satisfy the benchmarking property in the sense of adding up to a reliable direct estimator at a large area level. Professors Jiang and Lahiri note that the estimator of MSPE given in the You-Rao paper is not second correct to second order unless the survey weights within areas are equal. This is indeed true and recently Torabi and Rao (2005) worked out a second order correct MSPE estimator of the pseudo-EBLUP by estimating the covariance term in MSPE, following the general approach outlined in Jiang and Lahiri (2006). However, empirical results with unequal weights within areas suggested that the effect of the covariance term on MSPE is very small when comparing Prasad-Rao type estimators. On the other hand, an area specific version of the correct estimator was significantly smaller than the corresponding area specific version of the You-Rao estimator.

Interval estimation of small area parameters is an important problem. The parametric boostrap method of Chatterjee and Lahiri (2002) for interval estimation under the Fay-Herriot area level model, mentioned in Section 6.1, looks promising. The interval is based on the EBLUP estimator and the associated naïve estimator of MSPE, and the margin of error for the coverage probability associated with the interval is of lower order than $O(m^{-1})$. Extensions to other small area models would be useful in practice. In Section 6.2, Professors Jiang and Lahiri give a nice account of model building in small area estimation and note that it is essentially an iterative process involving selection of a model, fitting of the selected model and model diagnostics. I agree with the authors that a good interaction between the subject matter specialist and the statistician is essential for the model building process.

Finally, Section 7 revisits the famous baseball example of Efron and Morris (1975). It provides interesting insights, including the validity of the exchangeability assumption in the presence of outlier (Roberto Clemente) and the use of previous end of season batting average for the previous year as an auxiliary variable in the Fay-Herriot model. In this connection, I noted (Rao, 2003, p. 67) that the relative overall accuracy of the James-Stein estimator (or the EBLUP under exchangeability) is in fact larger than that of the limited translation estimator (Section 5.3) which limits the effect of Clemente on the overall accuracy.

I congratulate the authors for preparing an insightful and informative overview paper on the important topic of small area estimation.

Additional references

Lohr, S. and Rao, J. N. K. (2003). Resampling methods for MSE estimation with nonlinear small area models. In *Proceedings of Statistics Canada Symposium 2003*.

Torabi, M. and Rao, J. N. K. (2005). Mean squared error estimators of small area means using survey weights. In *Proceedings of the Statistical Society of Canada Annual Meeting, Survey Methods Section*.

John L. Eltinge
U.S. Bureau of Labor Statistics, USA

Introduction: Modeling approaches, utility functions and inferential risk in small domain estimation

Professors Jiang and Lahiri have produced a very solid and accessible review of mixed model prediction and related applications to small area estimation. This brief discussion will highlight some areas in which special population structures and the practical needs of some data users might lead to extensions of this literature. Section 1 considers the use of sparse-effect models. Section 2 discusses complex utility functions. Section 3 suggests some alternative approaches to evaluation of inferential risk.

1 Use of sparse-effect models for small domains defined by the cross-classification variables

First, one brief note on modeling. In some cases, the small areas (or "small domains") of interest are defined by the intersection of several cross-classification variables. See, e.g., Zhang and Chambers (2004) and references cited therein. In some such cases, it may be worthwhile to combine the mixed-model approaches emphasized by Professors Jiang and Lahiri with simple sparse-effect modeling and diagnostic approaches developed primarily in the literature on experimental design, e.g., Daniel (1959), Brenneman and Nair (2001), McGrath and Lin (2001) and references cited therein. In particular, adaptation of these approaches to the small-domain setting may provide a useful framework through which to study the extent to which estimation and inference methods are robust against the undetected presence of a sparse main effect or low-order interaction associated with the abovementioned cross-classification variables. Similarly, the abovementioned experimental-design literature includes diagnostics for detection of sparse effects in mean functions and dispersion functions, and one could consider extension of some of these diagnostics to small-domain settings.

2 Complex utility functions and empty cells

Second, when government statistical agencies produce small area estimates, those estimates will often be used by a wide range of stakeholders. The inferential goals of these stakeholders, and the impact of estimation error on their utility functions, can be markedly heterogeneous. Consequently, in development and evaluation of a small area estimation procedure, it is worthwhile to consider several distinct approaches.

In keeping with the predominant thrust of the literature, Professors Jiang and Lahiri focus primarily on squared-error loss and related quantities like the mean squared error, integrated over an appropriate space. This is a very reasonable approach to evaluation of small area estimation, but some data users have utility functions that deviate substantially from standard squared-error loss. For example, some data users have highly asymmetric utility functions, i.e., for an error of a given magnitude ε, a stakeholder may incur a much larger loss if the error is positive than if it is negative, or vice versa. Within the social sciences, the literature on prospect theory (e.g., Kahneman and Tversky, 1979, 1984; Tversky and Kahneman, 1992, and references cited therein) suggests that this phenomenon is a fairly common characteristic of human behavior. Consequently, it is worthwhile to consider measures of estimator performance that account explicitly for asymmetries in stakeholders' loss functions. In addition, it would be of interest to study the extent to which frequentist or Bayesian approaches could be tuned to accommodate these asymmetries.

Another form of asymmetric loss function arises with the decision of whether or not to publish estimates for a given set of small domains. Since such estimates are often published in tabular form, with each domain corresponding to a given cell in a cross-classified table, the decision not to publish a given estimate (due to problematic data quality) is sometimes known colloquially as the "empty cell problem". The literature on small domain estimation generally has focused attention on loss functions associated with the errors of published estimates. It would be of interest to extend this approach to incorporate losses associated with the decision to leave some cells empty. This would be one step toward embedding the traditional evaluation of the quality of small domain estimators within a broader framework of survey data quality, as discussed, e.g., in Brackstone (1999).

In the United States and some other countries, complex utility functions can also arise when small area estimates are used in formula-based allocation of federal funds to states or localities. Some of this allocation work is used to allocate a fixed pool of funds, leading to a multivariate utility function with one row associated with each potential funding recipient, and with corresponding asymmetries across the competing parties. Also, some cases involve periodic repetition of the estimation process accompanied by re-allocation of funds, but with "hold harmless" provisions that limit the extent to which changes in estimates over time will affect the allocation of funds. Such "hold harmless" cases involve utility functions that may be quite different for, respectively, the initial and updated estimation tasks. See, e.g., Zaslavsky and Schirm (2002) and Aldridge (2002).

3 Characterization and quantification of inferential risk in exploratory analyses

Again in keeping with the predominant literature, Professors Jiang and Lahiri have focused primarily on cases in which stakeholders wish to produce point estimates or related confidence sets for a relatively well-defined set of population means or closely related quantities. Shen and Louis (1998, 2000) consider an alternative "triple goal" approach in which stakeholders may wish to estimate the abovementioned means, rank those means, and construct an associated histogram. Under this approach, one may encounter patterns of inferential risk that differ somewhat from those encountered in the customary small area estimation literature.

In addition, anecdotal evidence and direct observation suggests that some stakeholders use published sets of small area estimates in a highly exploratory and conditional form. For such cases, application and interpretation of standard testwise and experimentwise error rates may be problematic. It would be worthwhile to study the extent to which false discovery rates and related concepts (e.g., Benjamini and Hochberg, 1995; Black, 2004; Storey, 2002, and references cited therein) may provide better tools through which we could characterize and quantify inferential risk for such stakeholders.

Additional references

Aldridge, D. K. (2002). Special supplemental nutrition program for Women, Infants, and Children (WIC) funding formula evolution. *Journal of Official Statistics*, 18(3):429–439.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57:289–300.

Black, M. A. (2004). A note on the adaptive control of false discovery rates. *Journal of the Royal Statistical Society. Series B*, 66:297–304.

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, 25:139–149.

Brenneman, W. A. and Nair, V. N. (2001). Methods for identifying dispersion effects in unreplicated factorial experiments: A critical analysis and proposed strategies. *Technometrics*, 43:388–405.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1:311–341.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–291.

Kahneman, D. and Tversky, A. (1984). Choices, values and frames. *American Psychologist*, 39:341–350.

McGrath, R. N. and Lin, D. K. J. (2001). Testing multiple dispersion effects in unreplicated fractional factorial designs. *Technometrics*, 43:406–414.

Shen, W. and Louis, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society. Series B*, 60:455–471.

Shen, W. and Louis, T. A. (2000). Triple-goal estimates for disease mapping. *Statistics in Medicine*, 17:2295–2308.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, 64:479–498.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323.

Zaslavsky, A. M. and Schirm, A. L. (2002). Interactions between survey estimates and federal funding formulas. *Journal of Official Statistics*, 18:371–391.

Zhang, L. C. and Chambers, R. L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society. Series B*, 66:479–496.

Rejoinder by J. Jiang and P. Lahiri

We thank the discussants, Professors Eltinge, Hall, Morales, Morris, Pfeffermann and Rao for their generous and stimulating discussions. Their discussions help clarifying various underlying issues and concepts and supplement our paper by new approaches and topics that we did not cover. There are some common issues raised by different discussants. However, since they come from different perspectives, we find it convenient to organize our rejoinder around each discussant separately. Our nonresponse to any particular issue simply implies that we have nothing to add to the issue.

1 Reply to Professor Hall

The nonparametric bootstrap method suggested by Professor Hall is indeed interesting and needs to be compared with other existing methods. In surveys, robustness is an important issue and the practitioners are always interested in efficient nonparametric methods. However, due to scarce data at the small area level, some level of smoothing [either implicit or explicit] is needed. Thus we agree with Professor Hall that parametric bootstrap is likely to play a greater role than the traditional nonparametric bootstrap in the small area estimation and other related problems.

For the Fay-Herriot model, Lahiri and Rao (1995) showed that the well-known Prasad-Rao MSPE estimator, a Taylor series or delta method

derived under normality, has a surprising robustness property under the non-normality of random effects when ANOVA method is used to estimate the variance component. However, in the late 80's Professors Lahiri and Rao, in an unpublished work, observed that such robustness property of the Prasad-Rao MSPE estimator does not hold for the nested error model or even the Fay-Herriot model when the normality of the sampling errors does not hold. In addition, Chen et al. (2005) observed that the robustness of the Prasad-Rao MSPE estimator depends on the variance component estimation method. In any case, for the non-normal mixed linear models, robust MSPE estimators can be proposed using the delta method, but this involves estimation of kurtosis, in addition to the estimation of the variance components. Jiang's (2003) method for estimation of kurtosis for mixed models should be useful for such problems. For robust MSPE estimators that use the delta method, see Chen et al. (2005) and Ganesh et al. (2005).

The nonparametric bootstrap method proposed by Hall and Maiti (2006a) clearly offers an alternative to the robust delta method mentioned above and the jackknife method of Jiang et al. (2002). Unlike the jackknife method, the robust delta and the nonparametric bootstrap methods rely on the estimation of the kurtosis terms. While they are all second-order unbiased under their respective non-normal mixed models, the degree of their robustness and efficiency relative to the simple normality-based Prasad-Rao method or the linearlized jackknife method of Chen and Lahiri (2005) needs to be assessed through Monte Carlo simulations.

While Professor Hall's nonparametric bootstrap is promising in constructing confidence intervals for the low dimensional hyperparameters like the regression coefficients and the variance components, its efficiency in solving the prediction interval problem for the high dimensional parameters like the θ_i's needs further study. The nonparametric model permits the generation of bootstrap histogram based on a synthetic or the regression model but does not permit the conditional distribution of θ_i given the data. As a result, the nonparametric bootstrap prediction interval for θ_i is likely to give less weight to the area specific data on the response variable than the corresponding parametric bootstrap of Chatterjee and Lahiri (2002) or Chatterjee et al. (2006). Construction of the area specific bootstrap histogram appears to be important in achieving good area specific coverage.

2 Reply to Professor Morales

First of all, we share Professor Morales' view that many branches of statistics have converged to the field of small area estimation.

In our view, a completely nonparametric method is not very effective in combining different sources of information. However, it is possible to evaluate small area estimators such as an EBP under a nonparametric setting. In this context jackknife (Jiang et al., 2002) and nonparametric bootstrap (Pfeffermann and Glickman, 2004; Hall and Maiti, 2006a) approaches have been considered. See Chen et al. (2005) and Ganesh et al. (2005) for certain robust delta methods. We agree with you that an appropriate semiparametric approach is promising in combining information in small area estimation problems. Such an approach has been considered by Opsomer et al. (2004). However, more research is needed in this interesting area.

We agree with you that generalized linear mixed models (GLMMs) provide the flexibility and simplicity in modelling small area data. However, inferences based on GLMMs are far more challenging computationally than those based on linear mixed models. Such computational difficulties have limited, to some extent, the use of GLMMs in SAE. However, we agree with Professor Morales that GLMMs are expected to play an increasing role in SAE as more effective computational procedures and software become available. The jackknife (see Jiang et al., 2002) and the parametric bootstrap methods (see Glickman and Pfeffermann, 2004; Lahiri, 2003a) are well posed to handle GLMMs for SAE.

We agree with Professor Morales that numerical comparisons are needed in order to understand the benefit of using more sophisticated methods, such as EBP. There are papers which considered such comparisons. See Rao (2003) and the references therein. Finally, the additional reference provided by Professor Morales on calibrated weighting methods and benchmarking should be helpful to the readers.

3 Reply to Professor Pfeffermann

Professor Pfeffermann's discussion on the hierarchical Bayesian method should be helpful to the readers. He correctly listed the major advantages and disadvantages of the hierarchical Bayes methods in small area estimation problems. The choice of priors for the hyperparameters (including the variance components) is indeed an important problem. Different objective

priors, including reference priors (see Bernardo, 2005) and matching priors (see Datta and Mukherjee, 2004; Datta and Sweeting, 2005), have been proposed in the Bayesian literature. The objective priors may be attractive to the practitioners. For example, Bayesian credible sets given by matching priors have approximate frequentist validity as well. Datta et al. (2005) and Ganesh and Lahiri (2005) who considered certain objective priors in solving small estimation problems. In addition, we refer to the work of Chen (2001) who evaluated the unconditional mean squared error properties of eight different priors for the variance component A. We reiterate that the choice of priors is critical in any Bayesian analysis and more research is needed in this important area.

Professor Pfeffermann's discussion on the design-based and indirect methods such as the synthetic or combined methods are indeed helpful. For the unsampled small areas, one has the option of synthetic method, as Professor Pfeffermann noted.

It is true that a mixed model with independent area specific random effects essentially produces a synthetic estimator. Using a spatial model for the area specific random effects, one can produce a model-based estimator that is less synthetic since the information on the study variable can be used from the neighboring areas. If the study is repeated over time, and if data from a previous survey are available for the areas for which sample is not available for the current time, one can use a spatio-temporal model to produce an estimator for unsampled area that is dependent on both the neighboring areas and the past data for the area under consideration. Thus, modelling approach is much more flexible than the corresponding design-based synthetic approach.

Professor Pfeffermann has correctly pointed out that the efficiency of an EBLUP and its MSE estimator depends on the type of the variance component estimators used. In this context, we reiterate that although the ANOVA, ML, REML and the Fay-Herriot type estimators are all consistent for large m, for a given sample they are all subject to producing zero estimates for the variance components. When this happens, the benefit of an EBLUP is lost and the EBLUP reduces to the synthetic estimator. In the context of the Small Area Income and Poverty (SAIPE) project of the U.S. Census Bureau, Bell (2001) used this argument in favor of the Bayesian methods. Certain possible frequentist's solutions to this problem were discussed in Lahiri (2003b). The solutions include consideration of a

measure of central tendency other than the mode (e.g., mean or median) of a normalized residual likelihood.

Professor Pfeffermann has raised an important issue related to the evaluation of different MSE estimators available in the literature. We agree that the MSE's of MSE estimators may be compared through simulations, but theoretical comparison could be a formidable task. We may add that one could also compare the MSE estimators by comparing the performances of the corresponding prediction intervals of the small area parameter(s) in terms of the average length and the coverage probability.

Professor Pfeffermann provided a lively discussion on benchmarking and design consistency. In our opinion, these two concepts are similar in the sense that they both tend to give more importance to the survey data than the model-based method that does not possesses these properties. However, unlike benchmarking, the design consistency does not ensure exact match of the model-based estimator to the corresponding design-based estimator at the higher level of aggregation. When the sample size for a domain is small, an efficient design-consistent model-based method is likely to be very different from the corresponding design-based method; but it will be closer to the design-based method for a domain with large sample size. Even though design consistency would not have practical effect to areas with small sample sizes, knowing that a method is design-consistent may enhance confidence in using such a method (because one knows that there is protection, at least in large sample, in case of possible model failure).

The design-consistency and benchmarking should not undermine the need for a good model. These two properties need to be ensured only after a reasonable model is selected. As Professor Pfeffermann says, the main issue is how to incorporate survey weights in the small area models. This topic has generated quite a few interesting papers in the general context of finite population sampling. Professor Pfeffermann's discussion, especially on informative sampling which we did not cover, is a valuable addition to our paper. In addition to the papers cited in our paper and Professor Pfeffermann's discussion, we also refer to two nice papers by Ghosh and Sinha (1990) and Little (2004).

We agree with Professor Pfeffermann that time series model could play an important role in small area estimation. In fact, such models can be covered by the general theory of Datta and Lahiri (2000), Butar and Lahiri (2003), Jiang et al. (2002) and Das et al. (2004) as long as the observations

across areas are independent. For applications of time series and cross-sectional models that use Taylor series method, see Rao and Yu (1994) and Datta et al. (2002). Datta et al. (1999) assessed the efficacy of cross-sectional, time series and joint cross-sectional and times series approach in the context of estimation of the state level unemployment rates using a hierarchical Bayesian approach. The related parametric and nonparametric implementation of the Kalman filter approach of Pfeffermann and Tiller (2005a) clearly offers a useful and promising alternative to the existing theory.

4 Reply to Professor Morris

As Professor Morris observes, we reviewed small area estimation methods that are simpler and less computer intensive than the Bayesian methods implemented via Monte Carlo Markov Chain (MCMC) methods. We did that primarily to bring the literature on such popular simple methods up to date. Readers interested in the Bayesian methods in small area estimation are referred to Nandram (2000) and Rao (2003).

We feel that a hierarchical Bayesian method needs to be suitably adapted before it can be applied to a situation that needs large volume of small area statistics for an on-going data generation process. For this, two points are important: (i) a careful specification of the noninformative prior distribution for the hyperparameters that ensures proper posterior and good frequentist properties of the resulting Bayesian method, and (ii) a good approximation to the posterior distribution that provides simple and transparent solution. The approximate Bayesian approach of Morris (1988) based on ADM that retains most of the good features of the corresponding fully Bayesian method and yet simple to implement and interpret should be considered seriously by the small area data users. In addition to its simplicity, such a method has been shown to have good frequentist's properties. As Professor Morris observes, unlike the ML or REML, the ADM method avoids undesirable zero variance component estimates without going for a fully Bayesian MCMC-based method.

In analyzing large scale sample survey data, attention is often restricted to producing point estimates and the associated standard errors. However, there is a growing demand to produce confidence intervals. We agree with Professor Morris that the naive empirical Bayes approach could have terrible coverage properties. Various solutions to overcome this problem include

the ADM method and other methods mentioned in Section 6.1 of the paper. It will be interesting to assess performances of the ADM method with these alternative methods, especially the parametric bootstrap, both in terms of the frequentist's coverage and the average length, at least by Monte Carlo simulations.

Professor Morris' discussion on approximations and evaluations based on asymptotics in m is intriguing. First of all, we share his concerns. The quality of approximations and evaluations of any method (classical or Bayesian) rests on the implicit or explicit assumptions involved and we feel that one good way of evaluating them is through the Monte Carlo simulations by varying the factor(s) affecting the approximations and evaluations. A related question is: how do we compare a classical method with a Bayesian method? In our view, the models used in the Bayesian and the corresponding frequentist analyzes differ only in the specification of the noninformative parts of the multi-level model. Thus, it seems reasonable to consider an evaluation measure based on the joint marginal distribution of all random variables having proper informative distributions. For the classical methods (e.g., methods based on asymptotics in m), the quality of approximations and evaluations may depend on m, but for a Bayesian method the quality may depend on other factors such as D_i/A, variations in D_i/A, presence of outlier in D_i/A, etc.

How large m is needed to justify an asymptotically justified method? This is an important question and needs to be evaluated before an application of such methods. The answer seems to depend on the relative magnitude of the B_i. If B_i is small, the classical methods may perform well even for small m; but when B_i is close to 1, one needs a very large m. We feel that the main problem in the classical method is not so much the approximation of the MSPE, but the estimation of the variance component A since, as Professor Morris noted, it can produce a zero estimate in which case the EBLUP or EB overshrunks and yields the synthetic or the regression estimate. For the EBP methods to work, it is critical that a good estimator of A is found. Lahiri (2003b) has some suggestions in this respect. For example, one can consider mean or median of a normalized residual likelihood for A which, like the Bayesian methods, would guarantee a positive estimate without the introduction of an additional level of prior distribution for A.

The answer to Professor Morris' other question about the magnitude of m obviously depends on the situation. In absence of any auxiliary information, the number of small areas available for building an exchangeable model is likely to be small. When one or more strongly related area specific auxiliary variables are available, the demand for exchangeable models is likely to diminish since such auxiliary variable(s) could be used to relax assumptions on both the mean and the variance of Level II model. In turn, the number of small areas m available for data analyzes increases.

In the context of the Small Area Income and Poverty Estimates (SAIPE) project of the U.S. Census Bureau, the number of counties (small areas) used in fitting a Fay-Herriot type model for school age children in poverty for different years are: 1,028 in 1989; 1,184 in 1993 (see National Research Council, Report 1, p. 65), and 985 in 1995 (see National Research Council, Report 3, p. 14). In the context of the U.S. Bureau of Labor Statistics Current Employment Statistics (CES) program, Lahiri and Li (2006) fitted a two-level model with m=2040 small areas (formed by the combination of different industry types, states and months).

When strongly related auxiliary data are not available, one can possibly form groups of small areas such that the small areas within each group are exchangeable with respect to the study variable. We refer to Malec and Sedransk (1992) for an interesting Baysian treatment to construct such groups of small areas. We agree with Professor Morris that the spatial modelling in this situation is appealing since it increases the effective number of small areas that can be used in combining information. Although an exchangeable model or a spatial model is justifiable statistically, there could be a non-statistical issue as to whether or not one can combine information from different small areas, especially when the purpose is to compare them. Here, area specific auxiliary data seems to be badly needed.

5 Reply to Professor Rao

Professor Rao's clarification of different mean squared prediction error definitions should be helpful to the readers. He has nicely pointed out how different types of area specific MSPE estimators can be generated - this should complement our paper.

We feel that further explanation of the Jiang-Lahiri-Wan jackknife is needed. Note that in order to obtain the jackknife formula, Jiang et al.

(2002) used an identity (not the Kacker-Harville identity) that is valid for certain non-normal situations. Their jackknife formula can be applied as long as one can compute the best predictor (BP). Thus, under a complete specification of the multi-level densities (not necessarily normal), this formula is valid. There is another interesting non-normal situation where the JLW jackknife is valid. Note that EBLUP can be motivated from the BP which can be obtained using the posterior linearity assumption. So a natural question is: can a MSPE estimator be produced under the assumption of posterior linearity? The jackknife estimator proposed by JLW is valid under this assumption. In fact for a simple case, as given in Section 3.4, the JLW jackknife formula involves the kurtosis and skewness terms. The Chen-Lahiri jackknife and weighted jackknifes (different from the JLW jackknife), however, have been shown to be valid under normality. Lahiri and Rao (1995), Chen et al. (2005) and Ganesh et al. (2005) suggested yet another interesting way to generate robust MSPE estimator of EBLUP, based on the existence of certain moments. The work of Torabi and Rao (2005) is an useful addition to the small area literature.

As Professor Rao points out, interval estimation of small area parameters is an important problem and an extension of the Chatterjee-Lahiri method for the Fay-Herriot model to other small area models would be helpful. Chatterjee et al. (2006) showed that the Chatterjee-Lahiri parametric bootstrap methods can be extended to a very general normal mixed linear model considered in Das et al. (2004). Normality can be relaxed but that requires a calibration type bootstrap, a method that is more cumbersome than the method one can use for the normal case.

6 Reply to Professor Eltinge

Professor Eltinge has raised a number of challenging issues in small area estimation faced by the data users. We agree that mixed model approaches and the related model diagnostics need to explored for cross-classified data. The new problem here is the correlation across small areas, a topic not fully explored in the classical prediction approach.

We agree that the small area methods have focused primarily on squared error loss functions and that small area methods need to be evaluated and compared with the corresponding design-based methods for its robustness with respect to different asymmetric loss functions. Mathematics become complicated for the asymmetric loss functions, but one can at least explore

the simulation-based methods. In case of wide range of stakeholders, one possibility is to save the conditional distribution of the small area parameters given the data so that different small area statistics can be computed for different loss functions. The Bayesian methods seem to be straightforward in this respect, at least conceptually, but appropriate algorithms need to be developed. Parametric bootstrap method is likely to find a role here. But research is needed in this interesting area.

The "empty cell problem" is indeed a difficult issue. In this case, one of the main challenges for the statistician is to develop a class of implied loss functions. This should be a great research area where statisticians can collaborate with the data users. Professor Eltinge's discussion on the complexity of utility functions is quite informative.

We have not covered "triple–goal" approach, and would refer the interested readers to Rao (2003), in addition to Shen and Louis (1998, 2000). As mentioned by Shen and Louis, this problem can be viewed as an inferential problem in presence of three specific loss functions. This approach is useful when one set of small area estimates is needed that can be used for three different purposes more efficiently than the other existing approaches.

We agree with Professor Eltinge that the use of small area published data in a highly conditional form is problematic. Such "data snooping" is very tempting and there is a definite need for developing appropriate statistical methods to facilitate the data users. As far as we know, this area is wide open in small area estimation. Recently, Ganesh and Lahiri (2005) extended the multiple comparison method in small area estimation. However, statistical methods needs to be developed for other complex "data snooping" purposes.

Within the discipline of statistics, the small area research is advancing on different topics such as generalized linear mixed models, spatial models, hierarchical Bayes, empirical best linear theory, linear empirical Bayes, variance component estimation, resampling methods, higher order asymptotics, and statistical computing. Furthermore, the small area research has contributed significantly to the literature on multi-level modelling widely used by researchers in social, behavioral and health sciences. Because of

the interest of researchers in different fields (e.g., statistics, epidemiology, social sciences, health sciences, etc.), small area estimation will remain one of the most intriguing research areas for many years to come.

Acknowledgements

We thank Professors Domingo Morales, Carl Morris, Danny Pfeffermann, and J. N. K. Rao for bringing to our attentions a number of typographical and other errors and making a number of constructive suggestions on the original version of our paper, Dr. William Bell for providing details of the U.S. Census Bureau's SAIPE program, Professor Rahul Mukerjee for discussion on objective priors, Professors Snigdhansu Chatterjee and J. N. K. Rao for some useful discussions on small area interval prediction, and Professor John Eltinge for explaining some practical aspects of small area estimation. We thank Ms Huilin Li for producing the graphs in Section 7. Finally, we thank the editors for giving us an opportunity to write this paper.

Additional references

Bernardo, J. M. (2005). Reference analysis. In D. K. Dey and C. R. Rao, eds., *Bayesian Thinking, Modeling and Computation*, Vol. 25 of *Handbook of Statistics*, pp. 17–90. North Holland, Amsterdam.

Chatterjee, S., Lahiri, P., and Li, H. (2006). Prediction interval estimation for small areas. Unpublished manuscript.

Chen, S. (2001). *Empirical Best Prediction and Hierarchical Bayes Methods in Small-Area Estimation*. PhD Thesis, Univ. of Nebraska-Lincoln.

Datta, G. S., Lahiri, P., and Maiti, T. (2002). Empirical Bayes estimation of median income of four person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102:83–97.

Datta, G. S., Lahiri, P., Maiti, T., and Lu, K. L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94:1074–1082.

Datta, G. S. and Mukherjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*, Vol. 178 of *Lecture Notes in Statistics*. Springer–Verlag, New York.

Datta, G. S. and Sweeting, T. J. (2005). Probability matching priors. In D. K. Dey and C. R. Rao, eds., *Bayesian Thinking, Modeling and Computation*, Vol. 25 of *Handbook of Statistics*, pp. 91–114. North Holland, Amsterdam.

Ganesh, N. and Lahiri, P. (2005). Multiple comparisons for small areas. Unpublished manuscript.

Ganesh, N., Lahiri, P., and Rao, J. N. K. (2005). Mean squared error estimation under nonnormal nested error models. Unpublished manuscript.

Ghosh, M. and Sinha, B. K. (1990). On the consistency between model-based and design-based estimators in survey sampling. *Communications in Statistics. Theory and Methods*, 19(2):689–702.

Jiang, J. (2003). Empirical method of moments and its applications. *Journal of Statistical Planning and Inference*, 115:69–84.

Lahiri, P. and Li, T. (2006). Estimation of the change in total employment using the U.S. current employment statistics survey. Unpublished manuscript.

Little, R. J. A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99:546–556.

Malec, D. and Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika*, 79:593–601.

Nandram, B. (2000). Bayesian generalized linear models for inference about small areas. In D. K. Dey, S. K. Ghosh, and B. K. Mallick, eds., *Generalized linear models: A Bayesian Perspective*, Chap. 6, pp. 91–114. Marcel Dekker, New York.

National Research Council, Report 1 (1997). *Small Area Estimates of Children in Poverty: Interim Report 1, Evaluation of 1993 Estimates for Title I Allocations*. National Academic Press, Washington.

National Research Council, Report 3 (1999). *Small Area Estimates of Children in Poverty: Interim Report 3, Evaluation of 1995 County and School District Estimates for Title I Allocations*. National Academic Press, Washington.

Opsomer, J. D., Breidt, F. J., Claeskens, G., Kauermann, G., and Ranalli, G. (2004). Nonparametric small area estimation using penalized spline regression. In *Proceedings of the Survey Research Methods Section*. American Statistical Association. To appear.

Rao, J. N. K. and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22:511–528.