Bayesian Analysis (2012) **7**, Number 2, pp. 363–384

Objective Bayesian Analysis of a Measurement Error Small Area Model

Serena Arima*, Gauri S. Datta† and Brunero Liseo‡

Abstract. We consider small area estimation under a nested error linear regression model with measurement errors in the covariates. We propose an objective Bayesian analysis of the model to estimate the finite population means of the small areas. In particular, we derive Jeffreys' prior for model parameters. We also show that Jeffreys' prior, which is improper, leads, under very general conditions, to a proper posterior distribution. We have also performed a simulation study where we have compared the Bayes estimates of the finite population means under the Jeffreys' prior with other Bayesian estimates obtained via the use of the standard flat prior and with non-Bayesian estimates, i.e., the corresponding empirical Bayes estimates and the direct estimates.

Keywords: Bayesian inference, Jeffreys' prior, small area model.

1 Introduction

In recent years, small area estimation has emerged as an important area of statistics as private and public agencies try to extract the maximum information from sample survey data. Sample surveys are generally designed to provide estimates of totals and means of variables of interest for large subpopulations or domains. However, governments are more and more interested in obtaining statistical summaries for smaller domains such as states, provinces, or different racial and/or ethnic subgroups. These domains are called small areas. In recent years, demand for reliable estimates for small area means has greatly increased due to their growing use in formulating policies and programmes, allocating government funds, regional planning and other uses. Policy makers are often interested in targeting areas with particular needs in order to conduct specific actions: for example, areas with the highest unemployment rates can be selected to carry out training programmes to improve the possibilities of finding a job or becoming self-employed. Other examples include poverty counts of school-age children at county levels, income for small places and so on (see Rao (2003), chapter 5).

The simplest approach to the small area estimation problem is to consider *direct estimators*, that is estimating the variable of interest using the domain-specific sample data. However, it is well known that the domain sample sizes are rarely large enough to support reliable and accurate direct estimators (Rao (2003); Ghosh and Rao (1994)). Small area estimation tackles the problem of providing reliable estimates of one or

*Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome, Italy, serena.arima@uniroma1.it

†Department of Statistics, University of Georgia, Athens, GA, gauri@uga.edu

‡Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome, Italy, brunero.liseo@uniroma1.it

 DOI:10.1214/12-BA712

several variables of interest in areas where the information available on those variables is, on its own, not sufficient to provide accurate direct estimate. Estimates for all areas are produced using the sample and some additional auxiliary information which should be available for all small areas.

Indirect estimators are often employed in order to increase the effective domain sample size by borrowing strength from the related areas using linking models, census, administrative data and other auxiliary variables associated with the small areas. Depending on the type of data available, small area models are classified into two types: area-level and unit-level. A comprehensive account of model-based small area estimation under area-level and unit-level models is given in Rao (2003). In this paper we focus on a unit-level nested error linear regression model with an area-level covariate subject to measurement error. As documented in Ghosh et al. (2006), this model is more realistic than other simpler models since it is often not possible to obtain exact measurements of important covariates. Ghosh et al. (2006) propose a Bayesian hierarchical model for estimating finite population small area means. They did not discuss the issue of the choice of the priors, rather they used a conjugate prior, which is particularly suitable for computation. However, the use of a conjugate prior leaves the problem of the selection of the hyperparameter values unsolved. Also, since small area models are often used by national institutes of statistics a certain degree of objectivity in the procedure seems required.

To this end, we compute the Jeffreys' prior, and we show that the corresponding posterior distribution is proper, at least when more than six small areas are involved in the analysis. Bayesian estimators obtained via the use of the Jeffreys' prior usually have excellent frequentist properties.

The paper is organized as follows. In Section 2 we introduce the small area model with measurement error in the covariates and give a brief overview of the classical estimation methods. In Section 3 we introduce the Bayesian estimation of small area means with flat prior distributions as in Ghosh et al. (2006); in Section 4 the Jeffreys' prior is derived and it is shown that the resulting posterior distribution is proper. In Section 5 a small simulation study comparing different estimation methods is reported. In Section 6 the estimation method is applied to real data. We conclude with a brief discussion in Section 7.

2 Unit-level model with measurement error in an area-level covariate

Consider a finite population, whose units are divided into m small areas. The population size of the i-th area is N_i, $i = 1, \ldots, m$. Let Y_{ij} be the value of the variable of interest associated with the j-th unit ($j = 1, \ldots, N_i$) in the i-th area ($i = 1, \ldots, m$). A random sample of size $n_i \geq 1$ is drawn from the i-th area population and the sample data are denoted by y_{ij} ($i = 1, \ldots, m$; $j = 1, \ldots, n_i$). For each area, the value of a covariate x_i is observed with error on each of the n_i units, so that the following nested linear regression

model with an area-level covariate X is defined for $j = 1, \ldots, n_i, i = 1, \ldots, m$ as

$$\begin{aligned} y_{ij} &= b_0 + b_1 x_i + v_i + e_{ij}, \\ X_{ij} &= x_i + u_{ij}. \end{aligned} \tag{1}$$

In the above model, n_i represents the sample size of the i-th domain, X_{ij} is the observed covariate value for the j-th unit in area i and x_i is the unknown true "area-specific" covariate value associated with y_{ij}. For simplicity, we are assuming x_i is a scalar, although the extension to the multiple regression case is mathematically simple. Further, v_i is the area-level random effect, e_{ij} is the random error and u_{ij} is the measurement error affecting the covariate x_i. Finally, random variables e_{ij}, v_i, u_{ij} and x_i are assumed to be mutually independent with

$$e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2);\ u_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_u^2);\ v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2);\ x_i \overset{\text{iid}}{\sim} N(\mu_x, \sigma_x^2).$$

Our model is a structural measurement error model. The vector of model parameters is denoted by $\boldsymbol{\phi} = (b_0, b_1, \sigma_e^2, \sigma_u^2, \sigma_v^2, \sigma_x^2, \mu_x)$. The typical goal of a finite population small area analysis is estimation of the true area-level means (TM), that is,

$$\gamma_i = \frac{1}{N_i} \sum_{i=1}^{N_i} Y_{ij}, \qquad i = 1, \ldots, m. \tag{2}$$

In Ghosh et al. (2006) it is shown that, given the model parameters, the best (or Bayes) predictor of γ_i, under a squared error loss, using the observed data $\mathbf{y_i} = (y_{i1}, \ldots, y_{in_i})$ is given by

$$\hat{\gamma_i}^B = (1 - f_i B_i)\bar{y_i} + f_i B_i (b_0 + b_1 \mu_x), \tag{3}$$

where $f_i = \frac{N_i - n_i}{N_i}$ is the finite population correction factor, $\bar{y_i} = n_i^{-1} \sum_{i=1}^{n_i} y_{ij}$ and $B_i = \sigma_e^2/(\sigma_e^2 + n_i(\sigma_v^2 + b_1^2 \sigma_x^2))$. Since the estimator in (3) involves unknown parameters, Ghosh et al. (2006) proposed the following empirical Bayes estimator

$$\hat{\gamma_i}^{EB} = (1 - f_i \hat{B}_i)\bar{y_i} + f_i \hat{B}_i (\hat{b}_0 + \hat{b}_1 \bar{x}) = (1 - f_i \hat{B}_i)\bar{y_i} + f_i \hat{B}_i \bar{y} \tag{4}$$

where $\bar{x} = \sum_{i=1}^m n_i \bar{x}_i / n_T$, $\bar{x}_i = n_i^{-1} \sum_{i=1}^{n_i} x_{ij}$, $\bar{y} = \sum_{i=1}^m n_i \bar{y}_i / n_T$, $\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$ and $\hat{b}_1 = (1 - MSW_x/MSB_x)^{-1} \tilde{b}_1$ where $\tilde{b}_1 = \frac{\sum_{i=1}^m n_i \bar{y}_i (\bar{X}_i - \bar{X})}{\sum_{i=1}^m n_i (\bar{X}_i - \bar{X})^2}$.
Moreover $\hat{B}_i = MSW_y/(MSW_y + n_i \hat{\eta}_m)$, with

$$MSW_y = (n_T - m)^{-1} \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2;\ MSB_y = (m-1)^{-1} \sum_{i=1}^m n_i (\bar{y}_i - \bar{y})^2;$$
$$\hat{\eta}_m = max[0, (MSB_y - MSW_y)(m-1)/g_m]$$

and $g_m = n_T - n_T^{-1} \sum_{i=1}^m n_i^2$. Ghosh et al. (2006) proved that the estimator in (4) is asymptotically optimal in the sense that

$$\lim_{m \to +\infty} m^{-1} \sum_{i=1}^m E\left[\left(\hat{\gamma_i}^B - \hat{\gamma_i}^{EB}\right)^2\right] = 0.$$

However, as noticed in Torabi et al. (2009), this estimator can be improved in several aspects. First, the Bayes predictor in equation (3) does not consider all the sample information provided by the covariate: more precisely, the expression $\hat{\gamma}_i^{EB}$ is obtained without conditioning upon the observed values of the X_{ij}'s, thus ignoring a relevant piece of information. Second, similarly to any empirical Bayes estimator, $\hat{\gamma}_i^{B}$ does not fully exploit the hierarchical structure of the model; as such, it can only produce point estimates of the unknown parameters and approximate standard errors. In order to account for the first issue, Torabi et al. (2009) proposed a fully efficient Bayes predictor by taking into account all the available data and then derived its empirical Bayes version by replacing the unknown parameters with their moment estimators. They also proved that the new estimator is asymptotically optimal and provided an estimate of the mean squared prediction error. However, neither estimator accounts for the hierarchical nature of the model; to this end one needs to introduce a hierarchical Bayes procedure as in Ghosh et al. (2006).

3 Hierarchical Bayesian model

Ghosh et al. (2006) introduced a hierarchical model to predict the population strata means γ_i $(i = 1, \ldots, m)$, when a covariate X is subject to measurement error. The model can be written in the usual multi-stage way:

S1 $y_{ij} = \theta_i + e_{ij} \quad j = 1, \ldots, n_i; \quad i = 1, \ldots, m$, with $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$;

S2 $\theta_i = b_0 + b_1 x_i + v_i \quad i = 1, \ldots, m$, with $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ and $X_{ij} = x_i + u_{ij}$, with $u_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$;

S3 $x_i \overset{\text{iid}}{\sim} N(\mu_x, \sigma_x^2), \quad i = 1, \ldots, m$;

S4 $b_0, b_1, \sigma_e^2, \sigma_u^2, \sigma_v^2, \sigma_x^2, \mu_x$ are, loosely speaking, mutually independent with flat priors over location parameters and inverse gamma distributions over the scale parameters.

Ghosh et al. (2006) proved that, under general conditions, the resulting posterior distribution of the parameter vector is proper. Given the conjugacy of the prior densities, it is straightforward to get a posterior sample via Gibbs sampling. Consequently, a hierarchical Bayesian estimate of the finite population means can be easily obtained.

Ghosh et al. (2006) did not discuss in detail the issues related to the choice of the hyperparameters in the inverse gamma distributions. They suggested to select vague priors corresponding to small values of the hyperparameters. However, this is not a real solution, especially in official statistics applications, where some form of objectivity is required both from a scientific and a legal perspective. Without going into the details of the philosophical debate on the meaning of "objectivity" in statistics, we believe that the possibility of having a Bayesian solution largely acceptable by official statisticians would be very important for practical applications. Noninformative priors on hyperparameters

are routinely used in hierarchical linear mixed models. The usual situation is that, while the resulting posterior means and the empirical Bayes predictors are in agreement with the empirical Bayes approach, the corresponding frequentist and Bayesian measures of uncertainty are not necessarily close. In order to obtain nonsubjective priors for the hyperparameters in the hierarchical Bayes model, Datta and Rao (2010) derived both the quantile matching and the Mean Squared Error (MSE) matching criteria of a resulting hierarchical Bayes procedure to determine suitable nonsubjective priors. The matching prior in Datta and Rao (2010), when the goal is to infer the small area mean θ_i, depends on the sampling variance D_i for the i-th area. Or for unit level models, the matching prior depends on the area sample size n_i. This implies that the same prior will not continue to have good frequentist properties for *all* small areas.

In this paper we derive the Jeffreys' prior for the model in (1). It is known that Jeffreys' prior is a reasonable objective prior when the entire parameter vector is *the* parameter of interest. In applications related to our model the usual parameter of interest is the area level population total (or mean), which depends on the whole parameter vector: in other words we are interested in a function of the parameter vector, a situation not very different from a predictive scenario. In the next section we will derive the Jeffreys' prior and we also prove that, under very general conditions, the resulting posterior distribution will be proper. The expression of the prior is quite complicated. However, a relatively simple Markov chain Monte Carlo algorithm, based on the ARMS (Adaptive Rejection Metropolis Sampling) strategy, can be implemented to produce a posterior sample.

4 Jeffreys' prior for hierarchical Bayesian model with measurement error

The original model is defined by equation (1). We reparametrize the model as follows. Let $t_i = x_i - \mu_x \sim N(0, \sigma_x^2)$ and set $\nu = b_0 + b_1\mu_x$. Then

$$b_0 + b_1 x_i + v_i = \nu + b_1 t_i + v_i, \tag{5}$$

and consequently,

$$Z_{ij} = \begin{pmatrix} Y_{ij} \\ X_{ij} \end{pmatrix} = \begin{pmatrix} \nu + s_i + e_{ij} \\ \mu_x + t_i + u_{ij} \end{pmatrix}$$

where $s_i = v_i + b_1 t_i$. It can be easily checked that $(s_i, t_i)' \sim N(0_2, \Sigma_R)$ with

$$\Sigma_R = \begin{pmatrix} b_1^2\sigma_x^2 + \sigma_v^2 & b_1\sigma_x^2 \\ b_1\sigma_x^2 & \sigma_x^2 \end{pmatrix}. \tag{6}$$

The likelihood function depends upon the observed data via the joint sufficient statistics

$$\bar{Z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij} \sim N\left(\begin{pmatrix} \nu \\ \mu_x \end{pmatrix}; \Sigma_R + \frac{1}{n_i}B\right), \quad i = 1, \ldots, m,$$

and $S_i = \sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_i)(Z_{ij} - \bar{Z}_i)' \sim W_2(n_i - 1, B)$, a Wishart distribution. Here $B = \mathbf{diag}\sigma_e^2, \sigma_u^2$. Then, setting $\lambda = (\nu, \mu_x)'$ and $T_i = \Sigma_R + \frac{1}{n_i}B$, the likelihood function will be proportional to

$$L(\lambda, \Sigma_R, B) \propto |B|^{-\frac{n_T - m}{2}} \prod_{i=1}^{m} |T_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{tr}B^{-1}\sum_{i=1}^{m} S_i\right) \exp\left(-\frac{1}{2}\sum_{i=1}^{m}(\bar{z}_i - \lambda)'T_i^{-1}(\bar{z}_i - \lambda)\right). \tag{7}$$

The Jeffreys' prior can be obtained as the square root of the determinant of the 7 by 7 Fisher Information matrix I. An easy calculation shows that I is block diagonal, that is the location parameter λ is orthogonal to both Σ_R and B; also the λ-block does not depend on λ itself. Following the suggestion of Jeffreys (1961), as reported in Kass and Wasserman (1996), the location parameter λ should be treated separately assuming a uniform prior on both the coordinates. The same result would be obtained by using the reference prior approach with (Σ_R, B) as the parameter of interest. Indeed we are in the situation where λ is a location parameter. Then the Jeffreys' prior will be

$$\pi^J(\lambda, \theta) \propto det(I(\theta))^{-1/2}$$

where $I(\theta)$ is computed taking λ as fixed and $\theta = (\sigma_e^2, \sigma_u^2, \sigma_v^2, \sigma_x^2, b_1)$. In our case $I(\theta)$ does not actually depend on λ and, incidentally, the result obtained by using the above approach or by first integrating λ and then computing the Fisher information matrix on the integrated likelihood coincide. After integrating out λ with respect to a uniform prior, the integrated log-likelihood function $\ell(\Sigma_R, B)$ is

$$\begin{aligned}\ell(\Sigma_R, B) &= -\frac{n_T - m}{2}\log|B| - \frac{1}{2}\sum_{i=1}^{m}\log|T_i| - \frac{1}{2}\log|\sum_i T_i^{-1}| - \frac{1}{2}\mathbf{tr}B^{-1}\sum_{i=1}^{m} S_i \\ &- \frac{1}{2}\sum_{i=1}^{m}\bar{z}_i'T_i^{-1}\bar{z}_i + \frac{1}{2}\left(\sum_i T_i^{-1}\bar{z}_i\right)'\left(\sum_i T_i^{-1}\right)^{-1}\left(\sum_i T_i^{-1}\bar{z}_i\right).\end{aligned}$$

The Jeffreys' prior can be then obtained as the square root of the determinant of the 5 by 5 Fisher information matrix I, whose elements are

$$
\begin{aligned}
I[\Sigma_R^{(11)}, \Sigma_R^{(11)}] &= \tfrac{1}{2}\sum_{i=1}^m \left(\Psi_i^{(11)}\right)^2 \\
I[\Sigma_R^{(11)}, \Sigma_R^{(12)}] &= \sum_{i=1}^m \Psi_i^{(11)}\Psi_i^{(12)} \\
I[\Sigma_R^{(11)}, \Sigma_R^{(22)}] &= \tfrac{1}{2}\sum_{i=1}^m \left(\Psi_i^{(12)}\right)^2 \\
I[\Sigma_R^{(11)}, B^{(11)}] &= \tfrac{1}{2}\sum_{i=1}^m \tfrac{1}{n_i}\left(\Psi_i^{(11)}\right)^2 \\
I[\Sigma_R^{(11)}, B^{(22)}] &= \tfrac{1}{2}\sum_{i=1}^m \tfrac{1}{n_i}\left(\Psi_i^{(12)}\right)^2 \\
I[\Sigma_R^{(12)}, \Sigma_R^{(22)}] &= \sum_{i=1}^m \Psi_i^{(12)}\Psi_i^{(22)} \\
I[\Sigma_R^{(12)}, B^{(11)}] &= \sum_{i=1}^m \tfrac{1}{n_i}\Psi_i^{(12)}\Psi_i^{(11)} \\
I[\Sigma_R^{(12)}, B^{(22)}] &= \sum_{i=1}^m \tfrac{1}{n_i}\Psi_i^{(12)}\Psi_i^{(22)} \\
I[\Sigma_R^{(22)}, \Sigma_R^{(22)}] &= \tfrac{1}{2}\sum_{i=1}^m \left(\Psi_i^{(22)}\right)^2 \\
I[\Sigma_R^{(22)}, B^{(11)}] &= \tfrac{1}{2}\sum_{i=1}^m \tfrac{1}{n_i}\left(\Psi_i^{(12)}\right)^2 \\
I[\Sigma_R^{(22)}, B^{(22)}] &= \tfrac{1}{2}\sum_{i=1}^m \tfrac{1}{n_i}\left(\Psi_i^{(22)}\right)^2 \\
I[\Sigma_R^{(12)}, \Sigma_R^{(12)}] &= \sum_{i=1}^m \left(\Psi_i^{(12)}\right)^2 + \Psi_i^{(11)}\Psi_i^{(22)} \\
I[B^{(11)}, B^{(22)}] &= \tfrac{1}{2}\sum_{i=1}^m \tfrac{1}{n_i^2}\left(\Psi_i^{(12)}\right)^2 \\
I[B^{(22)}, B^{(22)}] &= \tfrac{1}{2}\tfrac{n_T - m}{(B^{(22)})^2} + \tfrac{1}{2}\sum_{i=1}^m \tfrac{1}{n_i^2}\left(\Psi_i^{(22)}\right)^2 \\
I[B^{(11)}, B^{(11)}] &= \tfrac{1}{2}\tfrac{n_T - m}{(B^{(11)})^2} + \tfrac{1}{2}\sum_{i=1}^m \tfrac{1}{n_i^2}\left(\Psi_i^{(11)}\right)^2
\end{aligned}
$$

where we denote by Ψ_i the inverse of T_i and by $\Psi_i^{(hj)}$ the element of row h and column j of the matrix Ψ_i.

Our main result is that the Jeffreys' prior always produces a proper posterior and it can be routinely used as an objective prior for small area models with structural measurement error.

Theorem: *When $m > 6$, the Jeffreys' prior produces a proper posterior distribution.*

Proof: See Appendix.

5 Simulation study

We conducted a simulation study to compare the performance of the proposed Jeffreys' prior (JP) with flat priors (FP) proposed in Ghosh et al. (2006) and with the empirical Bayes estimator (EB). To this end, following the simulation scheme in Ghosh et al. (2006) and the suggestions of the referees, we simulate data under four scenarios based on different choices of prior parameters and different sample sizes. In Scenario 1 we created a finite super-population of size 1400 spread across $m = 12$ strata of sizes N_i respectively equal to 50, 250, 50, 100, 200, 150, 50, 150, 100, 150, 100 and 50. The

responses y_{ij} are generated under the super-population model as considered in this study with $b_0 = 100$, $\mu_x = 19.4$, $b_1 = 2$, $\sigma_e^2 = 100$, $\sigma_u^2 = 16$, $\sigma_x^2 = 2737$ and $\sigma_v^2 = 25$. A 2% simple random sample was used to generate samples from each stratum. Accordingly, the sample sizes n_i for the 12 strata are given, respectively, by 1, 5, 1, 2, 4, 3, 1, 3, 2, 3, 2 and 1.

In Scenario 2, sensitivity about the sample sizes is investigated. A super-population of size 140000 is considered and a 5% simple random sample was used to generate samples from each stratum. The same prior parameter values of Scenario 1 are used.

In Scenarios 3 and 4, sensitivity of variability parameters is studied. In Scenario 3, we increase the variability of the data fixing $\sigma_e^2 = 1000$. On the other hand, in Scenario 4, small values have been fixed for variability parameters: in particular, $\sigma_e^2 = 1$, $\sigma_u^2 = 5$, $\sigma_x^2 = 1$ and $\sigma_v^2 = 1$. For the other parameters, the same values of Scenario 1 have been used. For each scenario, we generated $R = 300$ independent sets of normal variates $u_i^{(r)}, i = 1, ..., m$, $v_i^{(r)}, i = 1, ..., m$, $e_{ij}^{(r)}, i = 1, ..., m, j = 1, ..., N_i$ with mean 0 and variances σ_u^2, σ_v^2 and σ_e^2. We also generated true $x_i^{(r)}, i = 1, ..., m$ from a normal distribution with mean μ and variance σ_x^2. Using u_i, e_{ij}, v_i, x_i the values $Y_{ij}^{(r)}$ and $X_{ij}^{(r)}$ have been computed according to the model in equation (1). The r-th simulated population mean of the i-th area is given by $\gamma_i^{(r)} = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}^{(r)}$. From each simulated population, we generated simple random samples $y_{ij}^{(r)}, j = 1, ..., n_i; i = 1, ..., m$ and $X_{ij}^{(r)}, j = 1, ..., n_i; i = 1, ..., m$. Using simulated values, we computed the sample mean $\hat{\gamma}_i^{SM(r)} = \frac{1}{n_i} \sum_{i=1}^{n_i} y_{ij}^{(r)}$. The empirical Bayes estimates $\hat{\gamma}_i^{EB(r)}$ of the true mean $\gamma_i^{(r)}$ have been computed as in equation (4).

Similarly, for each simulated population, we estimated the true means according to the hierarchical model proposed in Section 3. We computed $\hat{\gamma}_i^{FP(r)}$ when flat priors have been used for the unknown parameters as in Ghosh et al. (2006). With respect to the hyperparameters of the prior distributions we considered uniform $(-\infty; +\infty)$ priors for b_0, b_1 and μ_x. We also fixed the parameters of the inverse gamma distributions of the variance components $a_e, b_e, a_u, b_u, a_v, b_v, a_x, b_x$ all equal to 0.002. Different values of the inverse gamma parameters give similar results. We also computed $\hat{\gamma}_i^{JP(r)}$ when the proposed Jeffreys' prior is applied. Markov Chain Monte Carlo (MCMC) algorithms have been used in order to sample from the joint posterior distribution: in particular, location parameters $\lambda = (b_0, \mu)$ have been updated using a standard Gibbs sampling scheme, since the full conditional distribution is

$$\lambda|\theta \sim N_2 \left((\sum_{i=1}^{m} T_i^{-1})^{-1} (\sum_{i=1}^{m} T_i^{-1} \bar{z}_i); (\sum_{i=1}^{m} T_i^{-1})^{-1} \right)$$

where $\theta = (\sigma_e^2, \sigma_u^2, \sigma_v^2, \sigma_x^2, b_1)$. On the other hand, for the parameters θ, for which the Jeffreys' prior has been derived, full conditional distributions cannot be obtained in a closed form: we produce samples from the joint posterior distribution using the Adaptive Rejection Metropolis Sampling (ARMS) algorithm originally proposed in Gilks et al. (1995). We obtain $2K = 50000$ samples from the posterior distribution and

half of them are discarded as burn-in. The estimator of the true mean of area i of the r-th simulated data set is derived as $\hat{\gamma}_i^{JP(r)} = \frac{1}{K}\sum_{k=1}^{K}\hat{\gamma}_{ik}^{JP(r)}$ where $\hat{\gamma}_{ik}^{JP(r)} = (1 - f_i B_{ik}^{(r)})\bar{y}_i^{(r)} + f_i B_{ik}^{(r)}(b_{0k}^{(r)} + b_{1k}^{(r)}\mu_k^{(r)})$ $(i = 1, ...m; k = 1, ..., K; r = 1, ..., R)$. The estimator $\hat{\gamma}_i^{FP(r)}$ has been computed similarly. It must be noted that for the Bayesian hierarchical model with flat priors (FP), a Gibbs sampler algorithm can be adopted in order to generate samples from the joint posterior distribution (Ghosh et al. (2006)). However, for fair comparison with JP, we used the ARMS algorithm with equal starting values and with the same boundary constraints.
For each estimation method we then computed the empirical mean squared prediction error (EMSPE) as follows

$$EMSPE(\hat{\gamma}_i) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\gamma}_i^{(r)} - \gamma_i^{(r)})^2 \tag{8}$$

Tables 1, 2, 3 and 4 report the sample sizes and the empirical mean squared prediction error computed with the different methods for each county. The counties are denoted by $i = 1, ..., 12$ in the tables.

i	n_i	SM	EB	FP	JP
1	1	74.28	32.75	30.72	28.92
2	5	23.26	19.64	17.87	16.17
3	1	73.93	48.84	42.06	42.43
4	2	47.98	28.81	22.93	18.53
5	4	20.50	15.97	15.25	16.69
6	3	39.59	31.72	19.62	19.80
7	1	94.73	25.06	20.58	21.54
8	3	30.26	17.29	14.02	13.82
9	2	48.89	25.79	21.87	21.72
10	3	37.99	24.46	19.63	18.19
11	2	50.98	30.46	21.88	21.99
12	1	89.11	19.08	17.49	17.45

Table 1: Simulation study: scenario 1

i	n_i	SM	EB	FP	JP
1	25	3.74	3.94	3.86	3.66
2	125	0.96	0.91	0.90	0.90
3	25	4.79	3.52	3.44	3.41
4	50	2.01	2.20	2.11	2.06
5	100	1.16	1.05	1.07	1.04
6	75	1.59	1.61	1.59	1.56
7	25	3.75	4.91	3.61	3.85
8	75	1.13	1.19	1.14	1.14
9	50	1.54	1.73	1.57	1.61
10	75	1.34	1.55	1.33	1.34
11	50	1.77	1.90	1.81	1.83
12	25	3.52	4.33	3.57	3.45

Table 2: Simulation study: scenario 2

Table 1 shows that according to the EMSPE, the estimators based on the hierarchical Bayesian model are doing better than both the sample mean and the EB estimator. More precisely, FP and JP estimators are doing better than both the sample means and the EB estimator in 11 of the 12 counties. However, comparing the two model based estimators, JP is doing better than the FP in seven of the 12 counties. With respect to Scenario 2, Table 2 clearly shows that all methods perform better in terms of EMPSE with respect to the previous scenario because of the larger sample size. Also the same conclusion can be drawn in terms of relative performance of the different methods. FP and JP perform similarly and overcome both EB and SM. In Scenario 3 the larger data variability makes EB and SM perform much worse than the two model-based estimators that continue to behave similarly (see Table 3). On the other hand, an

i	n_i	SM	EB	FP	JP
1	1	951.08	109.36	75.95	73.91
2	5	165.49	62.96	54.79	48.16
3	1	923.58	121.32	76.51	77.05
4	2	553.12	94.76	60.98	60.90
5	4	169.18	67.31	55.14	54.35
6	3	314.89	95.39	56.38	57.05
7	1	890.34	120.96	72.70	73.97
8	3	352.18	91.60	61.63	60.69
9	2	627.41	94.11	64.90	64.92
10	3	352.04	101.98	82.10	82.14
11	2	475.79	89.60	53.27	53.77
12	1	1017.80	143.64	83.36	82.84

Table 3: Simulation study: scenario 3

i	n_i	SM	EB	FP	JP
1	1	2.06	0.94	0.96	0.90
2	5	0.44	0.38	0.36	0.33
3	1	1.80	0.73	0.74	0.70
4	2	0.95	0.68	0.69	0.60
5	4	0.48	0.42	0.42	0.37
6	3	0.74	0.55	0.51	0.48
7	1	1.99	0.90	0.88	0.88
8	3	0.71	0.53	0.51	0.46
9	2	1.06	0.66	0.62	0.61
10	3	0.66	0.55	0.54	0.48
11	2	0.98	0.63	0.64	0.61
12	1	1.96	0.86	0.92	0.83

Table 4: Simulation study: scenario 4

interesting conclusion can be drawn looking at Table 4: when the variability is low, as in the case of Scenario 4, the estimates obtained with the hierarchical model with flat prior show larger error than those obtained when Jeffreys' prior is involved. Moreover, in some situations, FP performs even worse than EB. The variability of the posterior estimates when vague priors are used is not surprising: interesting discussions about the consequences of using flat priors in mixed models can be found in Gelman (2006) and Hobert and Casella (1996).
The main message of this paper is that one can simply adopt a Bayesian approach based on Jeffreys' prior in order to provide an easy-to-use inferential procedure which is optimal, from a frequentist perspective and that, therefore, it should be easily adopted by official agencies.

6 Data analysis

We use the well known crop data by Battese et al. (1988) for a comparative analysis. Sample surveys have been designed to estimate crop areas for large regions: predicting crop areas for small areas such as counties is a difficult task due to the lack of availability of data from farm surveys in these areas. In their paper, Battese et al. (1988) considered data for 12 counties in Iowa, obtained from the 1978 June Enumerative Survey of the US Department of Agriculture as well as from the satellite LANDSAT during the 1978 growing season. The purpose was to predict the area under soy bean and corn in these counties.

We consider prediction of soy bean data using soy bean pixels only as covariates. In order to apply the proposed model, we incorporate possible measurement errors in the values of the covariates. As Ghosh et al. (2006) we generate copies of x-observations and then apply the model described in Section 3. The procedure is repeated 100 times and finally the estimates and their standard errors are computed using the averaging

principle. Table 5 provides the predicted hectares and the estimated standard errors of the predicted hectares for each county when the hierarchical Bayesian model is applied using flat priors ($\hat{\gamma}_i^{(FP)}$) and using the proposed Jeffreys' prior ($\hat{\gamma}_i^{(JP)}$). Standard errors of the direct estimators are reported in the last column of the table. It appears from the table that in the presence of measurement errors, the predicted values based on the Jeffreys' prior result in an error reduction with respect to those obtained using flat priors. Both methods clearly outperform the direct estimates.

County	k	$\hat{\gamma}_i^{(FP)}$	$\hat{\gamma}_i^{(JP)}$	$SE(\hat{\gamma}_i^{(FP)})$	$SE(\hat{\gamma}_i^{(JP)})$	$SE(\bar{y}_i)$
Cerro Gordo	1	31.00	36.70	19.90	18.20	29.10
Hamilton	1	101.90	100.70	19.90	18.20	29.10
Worth	1	100.10	99.10	19.90	18.20	29.10
Humboldt	2	44.20	47.00	14.10	12.90	20.60
Franklin	3	56.90	58.40	11.50	10.50	16.80
Pocahontas	3	115.50	114.40	11.50	10.50	16.80
Winnebago	3	88.90	88.90	11.50	10.50	16.80
Wright	3	97.00	96.70	11.50	10.50	16.80
Webster	4	111.00	110.30	10.00	9.10	14.60
Hancock	5	115.60	114.80	8.90	8.10	13.00
Kossuth	5	115.90	115.10	8.90	8.10	13.00
Hardin	5	101.10	100.70	8.90	8.10	13.00

Table 5: Predicted hectares of soy bean with corresponding standard errors, using soy bean pixels as the only covariate. $\hat{\gamma}_i^{(FP)}$ and $\hat{\gamma}_i^{(JP)}$ correspond respectively to the predicted hectares obtained with hierarchical Bayesian model with flat priors and with the proposed Jeffreys' prior. The last column shows the standard error of the direct estimates.

6.1 Computational details

The publicly available R software (RDevelopmentCoreTeam (2011)) has been used for implementation. R code is available on the web page of the first author[1]. As remarked by one of the referees, the ARMS algorithm is known to be sensible to the choice of the boundary values. Plausible boundary values can be chosen taking into account frequentist parameter estimates. We defined boundary constraints by considering the point frequentist estimation ± ten times their standard error. This procedure allows us to have plausible and sufficiently large boundary constraints.
Chain convergence has been ascertained by visual inspection using standard convergence diagnostic tools, such as trace plots and autocorrelation plots. Figure 1 and Figure 2 show the trace plot and the autocorrelation function of parameter σ_v^2 and b_1 estimated using Battese data. Also for the other parameters convergence diagnostic tools do not provide any convergence warnings.

[1] https://sites.google.com/site/webpageserenaarima/publications/ba-rcode

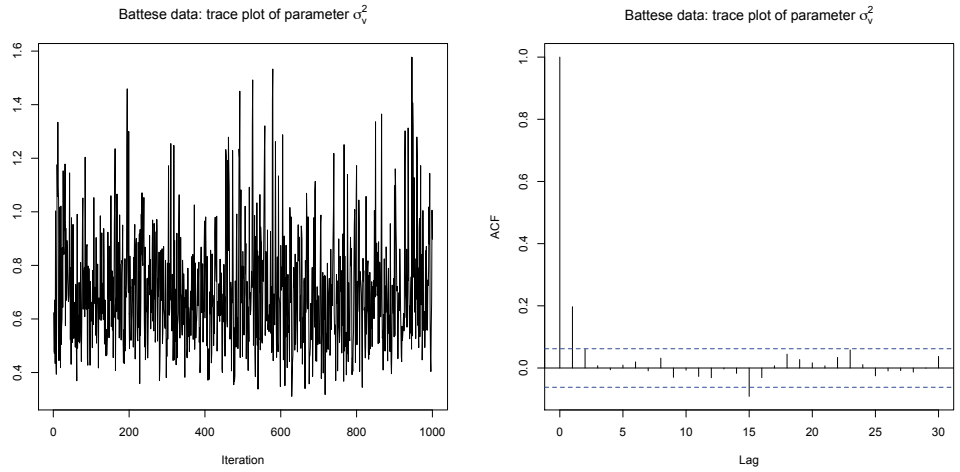

Figure 1: Battese data: trace plot and autocorrelation function of parameter σ_v^2.

With respect to the computational cost, we used ARMS for both vague priors and Jeffreys' prior. When using ARMS as the sampling method, flat priors and the proposed Jeffreys' prior have clearly the same computational cost. However, when Gibbs sampling can be applied (and this is the case for FP), the computational cost is clearly reduced: for 1000 simulations, the computational cost of ARMS is more or less three times the computational cost of the Gibbs sampling.

7 Summary and conclusion

In this paper we have proposed an objective Bayesian analysis of small area models with measurement error in the covariates. We have derived the Jeffreys' prior for the unknown parameters and we have shown that this prior leads, under very general conditions, to a well defined proper posterior distribution. Jeffreys' prior is not a conjugate prior, so ad-hoc Markov Chain Monte Carlo algorithms have been proposed in order to produce a sample from the joint posterior distribution. The estimated small area means have been compared with those obtained with a Bayesian hierarchical model with vague priors, the empirical Bayes estimates and the direct estimates. As revealed in our simulation study, according to the EMSPE, estimators based on the hierarchical Bayesian model usually perform better than sample means and the EB estimators. Estimators based on vague priors and Jeffreys' prior are very similar and they also have similar EMSPEs, although the estimator based on Jeffreys' prior is in some cases more precise than the one based on vague priors. We stress the fact that although producing similar results, the ideas behind the two approaches are very different: any subjective prior could be criticized on the ground that final inference could (even strongly) depend on the values of the hyperparameters and the use of vague priors (with ad-hoc values for the

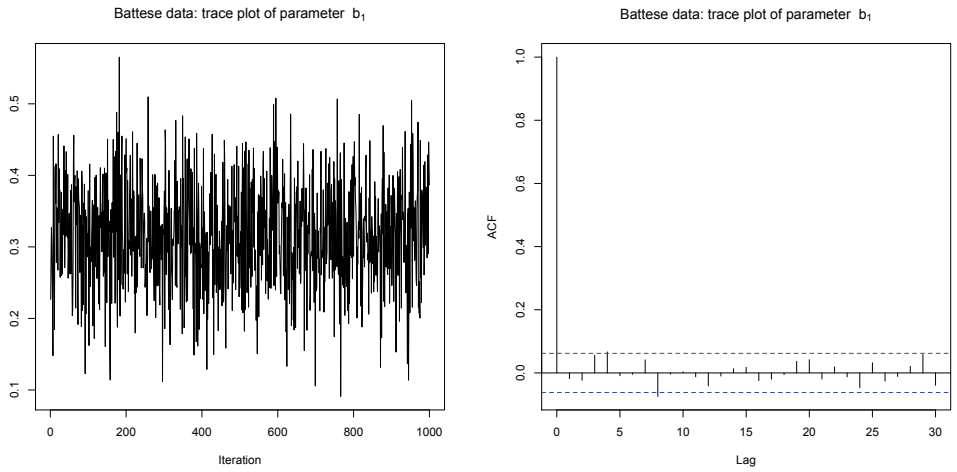

Figure 2: Battese data: trace plot and autocorrelation function of parameter b_1.

hyperparameters) cannot be considered a real solution to this problem. The subjectivity related to the prior choices makes the official statisticians more suspicious with respect to the Bayesian approach. We believe that the proposed objective Bayesian approach would be largely acceptable by official statisticians and would lead to a more effective dissemination of the Bayesian methodologies in this field.

8 Proof of Theorem

Note that the posterior distribution of (Σ_R, B) is

$$\pi(\Sigma_R, B|Y) \propto \pi^J(\Sigma_R, B)L(\Sigma_R, B)$$

where $L(\Sigma_R, B)$ is the integrated likelihood, $\pi^J(\Sigma_R, B)$ is the Jeffreys' prior defined in Section 4, and

$$B = \begin{pmatrix} \sigma_e^2 & 0 \\ 0 & \sigma_u^2 \end{pmatrix}, \Sigma_R = \begin{pmatrix} b_1^2\sigma_x^2 + \sigma_v^2 & b_1\sigma_x \\ b_1\sigma_x & \sigma_x^2 \end{pmatrix} = \begin{pmatrix} \sigma_{R_{11}} & \sigma_{R_{12}} \\ \sigma_{R_{12}} & \sigma_{R_{22}} \end{pmatrix}.$$

We will prove the theorem by showing that there exists a positive integrable function $f(\Sigma_R, B)$ such that $\pi^J(\Sigma_R, B)L(\Sigma_R, B) \leq f(\Sigma_R, B)$ uniformly. The lengthy proof has been split into 3 lemmas as follows:

- *Lemma 1*: we provide a non-negative function $h(\Sigma_R, B)$ such that $\pi^J(\cdot) \leq h(\cdot)$;

- *Lemma 2*: we provide a non-negative function $g(\Sigma_R, B)$ such that the integrated (with respect to λ) likelihood function $L(\cdot)$ is bounded above by $g(\cdot)$;
- *Lemma 3*: we show that $f(\Sigma_R, B) = h(\Sigma_R, B)g(\Sigma_R, B)$ is integrable over the parameter space.

Lemma 1 *There exists a function $h \geq 0$ such that $\pi^J(\Sigma_R, B) \leq h(\Sigma_R, B)$.*

Proof: Let $T_i = \Sigma_R + \frac{1}{n_i}B$. Let Ψ_i be the inverse of T_i and $\Psi_i^{(hj)}$ the (h, j) element of Ψ_i. Positive definiteness implies that $(\Psi_i^{(12)})^2 \leq \Psi_i^{(11)}\Psi_i^{(22)}$. An upper bound for the Jeffreys' prior can be obtained by deriving an upper bound for each diagonal element of the Fisher information matrix.
Note that

$$\Psi_i^{(11)} = \frac{\sigma_{R_{22}} + n_i^{-1}\sigma_u^2}{\left(\sigma_{R_{11}} + n_i^{-1}\sigma_e^2\right)\left(\sigma_{R_{22}} + n_i^{-1}\sigma_u^2\right) - (\sigma_{R_{12}})^2}$$

$$\Psi_i^{(22)} = \frac{\sigma_{R_{11}} + n_i^{-1}\sigma_e^2}{\left(\sigma_{R_{11}} + n_i^{-1}\sigma_e^2\right)\left(\sigma_{R_{22}} + n_i^{-1}\sigma_u^2\right) - (\sigma_{R_{12}})^2}.$$

Both $\Psi_i^{(11)}$ and $\Psi_i^{(22)}$ are increasing functions of n_i, and

$$I[\Sigma_R^{(11)}, \Sigma_R^{(11)}] = \frac{1}{2}\sum_{i=1}^{m}\left(\Psi_i^{(11)}\right)^2 \leq \frac{m}{2}\left[\frac{\sigma_{R_{22}} + n^{*-1}\sigma_u^2}{\left(\sigma_{R_{11}} + n^{*-1}\sigma_e^2\right)\left(\sigma_{R_{22}} + n^{*-1}\sigma_u^2\right) - (\sigma_{R_{12}})^2}\right]^2 \tag{9}$$

$$I[\Sigma_R^{(22)}, \Sigma_R^{(22)}] = \frac{1}{2}\sum_{i=1}^{m}\left(\Psi_i^{(22)}\right)^2 \leq \frac{m}{2}\left[\frac{\sigma_{R_{11}} + n^{*-1}\sigma_e^2}{\left(\sigma_{R_{11}} + n^{*-1}\sigma_e^2\right)\left(\sigma_{R_{22}} + n^{*-1}\sigma_u^2\right) - (\sigma_{R_{12}})^2}\right]^2 \tag{10}$$

where $n^* = \max_{1 \leq i \leq m} n_i$. Similarly,

$$\begin{aligned} I[\Sigma_R^{(12)}, \Sigma_R^{(12)}] &= \sum_{i=1}^{m}\left[(\Psi_i^{(12)})^2 + \Psi_i^{(11)}\Psi_i^{(22)}\right] \leq 2\sum_{i=1}^{m}\Psi_i^{(11)}\Psi_i^{(22)} \\ &\leq 2m\frac{(\sigma_{R_{22}} + n^{*-1}\sigma_u^2)(\sigma_{R_{11}} + n^{*-1}\sigma_e^2)}{[(\sigma_{R_{22}} + n^{*-1}\sigma_u^2)(\sigma_{R_{11}} + n^{*-1}\sigma_e^2) - (\sigma_{R_{12}})^2]^2}. \end{aligned} \tag{11}$$

The two diagonal elements in the Fisher information matrix corresponding to σ_e^2 and σ_u^2 are given by

$$\begin{aligned} I[B^{(11)}, B^{(11)}] &= \frac{1}{2}\sum_{i=1}^{m}\left[\frac{1}{n_i^2}(\Psi_i^{(11)})^2 + \frac{n_i - 1}{\sigma_e^4}\right] \\ I[B^{(22)}, B^{(22)}] &= \frac{1}{2}\sum_{i=1}^{m}\left[\frac{1}{n_i^2}(\Psi_i^{(22)})^2 + \frac{n_i - 1}{\sigma_u^4}\right]. \end{aligned}$$

Taking the derivative with respect to n_i, one sees that $\Psi_i^{(11)}/n_i$ is decreasing in n_i and it is maximized at $n_i = 1$. The same happens to $\Psi_i^{(22)}/n_i$ and an upper bound for $I[B^{(11)}, B^{(11)}]$ and $I[B^{(22)}, B^{(22)}]$ is given by

$$I[B^{(11)}, B^{(11)}] \leq \frac{m}{2}\left[\left(\frac{\sigma_{R_{22}} + \sigma_u^2}{(\sigma_{R_{11}} + \sigma_e^2)(\sigma_{R_{22}} + \sigma_u^2) - \sigma_{R_{12}}^2}\right)^2 + \frac{n^* - 1}{\sigma_e^4}\right] \tag{12}$$

$$I[B^{(22)}, B^{(22)}] \leq \frac{m}{2}\left[\left(\frac{\sigma_{R_{11}} + \sigma_e^2}{(\sigma_{R_{11}} + \sigma_e^2)(\sigma_{R_{22}} + \sigma_u^2) - \sigma_{R_{12}}^2}\right)^2 + \frac{n^* - 1}{\sigma_u^4}\right]. \tag{13}$$

For simplicity of notation, we denote $I[\Sigma_R^{(11)}, \Sigma_R^{(11)}] = I_{\sigma_{R_{11}}}$, etc. From (9), (10) and (11) we get

$$\left(I_{\sigma_{R_{11}}} I_{\sigma_{R_{22}}} I_{\sigma_{R_{12}}}\right)^{\frac{1}{2}} \leq const\left[\frac{(\sigma_{R_{22}} + \sigma_u^2)^{\frac{3}{2}}(\sigma_{R_{11}} + \sigma_e^2)^{\frac{3}{2}}}{\left[(\sigma_{R_{11}} + \sigma_e^2)(\sigma_{R_{22}} + \sigma_u^2) - \sigma_{R_{12}}^2\right]^3}\right] \tag{14}$$

Let

$$\omega_{11} = \sigma_{R_{11}}\sigma_e^{-2}, \omega_{22} = \sigma_{R_{22}}\sigma_u^{-2}, \omega_{12} = \sigma_{R_{12}}\left(\sigma_e\sigma_u\right)^{-1}; \tag{15}$$

then

$$\begin{aligned}(I[B^{(11)}, B^{(11)}]I[B^{(22)}, B^{(22)}])^{\frac{1}{2}} \leq \frac{const}{\sigma_e^2\sigma_u^2}\Bigg[k_1 + \frac{(1+\omega_{11})(1+\omega_{22})}{[(1+\omega_{11})(1+\omega_{22}) - \omega_{12}^2]^2} \\ + \frac{k_2(1+\omega_{22})}{(1+\omega_{11})(1+\omega_{22}) - \omega_{12}^2} + \frac{k_3(1+\omega_{11})}{(1+\omega_{11})(1+\omega_{22}) - \omega_{12}^2}\Bigg].\end{aligned} \tag{16}$$

Using (14) and (17), an upper bound for the Jeffreys' prior is then

$$\begin{aligned}h(\sigma_e, \sigma_u, \omega) \propto \frac{1}{\sigma_e^5\sigma_u^5}\Bigg[\frac{k_1(1+\omega_{11})^{\frac{3}{2}}(1+\omega_{22})^{\frac{3}{2}}}{[(1+\omega_{11})(1+\omega_{22}) - \omega_{12}^2]^3} + \frac{(1+\omega_{11})^{\frac{5}{2}}(1+\omega_{22})^{\frac{5}{2}}}{[(1+\omega_{11})(1+\omega_{22}) - \omega_{12}^2]^5} \\ + \frac{k_2(1+\omega_{11})^{\frac{3}{2}}(1+\omega_{22})^{\frac{5}{2}}}{[(1+\omega_{11})(1+\omega_{22}) - \omega_{12}^2]^4} + \frac{k_3(1+\omega_{11})^{\frac{5}{2}}(1+\omega_{22})^{\frac{3}{2}}}{[(1+\omega_{11})(1+\omega_{22}) - \omega_{12}^2]^4}\Bigg]\end{aligned} \tag{17}$$

which proves Lemma 1.

Lemma 2: *There exists a function g such that $L(\Sigma_R, B|Y) \leq g(\Sigma_R, B|Y)$*

Proof. We rewrite the model as follows:

$$\begin{aligned}Y_{1ij} &= X_{1ij}^T\beta_1 + v_{1i} + e_{1ij} \\ Y_{2ij} &= X_{2ij}^T\beta_2 + v_{2i} + e_{2ij}\end{aligned}$$

where

$$\begin{array}{rcl} Y_{1i} & = & \mathrm{col}_{1\le j\le n_i} Y_{1ij} \quad Y_1 = \mathrm{col}_{1\le i\le m} Y_{1i} \quad v_1 = \mathrm{col}_{1\le i\le m} v_{1i} \quad e_{1i} = \mathrm{col}_{1\le j\le n_i} e_{1ij} \\ X_{1i} & = & \mathrm{col}_{1\le j\le n_i} X_{1ij}^T \quad T_1 = \mathrm{col}_{1\le i\le m} X_{1i} \quad e_1 = \mathrm{col}_{1\le i\le m} \quad Z_1 = \oplus 1_{n_i} \end{array}$$

Similarly, Y_2, v_2, X_2, Z_2 and e_2 are defined. Let $e_{1ij} \overset{\text{iid}}{\sim} N(0, \sigma_1^2)$; $e_{2ij} \overset{\text{iid}}{\sim} N(0, \sigma_2^2)$; $(v_{1i}, v_{2i})' \sim N_2(0, \Sigma_V)$ $(i = 1, \ldots, m;\ j = 1, \ldots, n_i)$, β_1 and β_2 are vectors of length p_1 and p_2 respectively and

$$Z = \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix},$$

$$X = \begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix},$$

where X is a $(2n \times (p_1 + p_2))$ matrix of rank $p_1 + p_2$. Then the original model can be written as a linear mixed model

$$Y = X\beta + Zv + e \tag{18}$$

with $e \sim N(0, R)$, $v \sim N(0, G)$ for positive definite (p.d.) matrices R and G, where $R = \mathrm{diag}(\sigma_1^2 I_n, \sigma_2^2 I_n)$, $\sum_{i=1}^m n_i = n$ and I_n is the n-dimensional identity matrix. Notice that $\mathrm{Var} Y = ZGZ^T + R = \Sigma$. The likelihood function can be written as

$$L(\Sigma, \beta) = |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(Y - X\beta)^T \Sigma^{-1} (Y - X\beta)\right). \tag{19}$$

Let $K = \left(\Sigma^{-1} - \Sigma^{-1} X \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1}\right)$. A standard calculation shows that, after adopting a uniform prior for β, the integrated likelihood function for Σ is

$$\tilde{L}(\Sigma) \propto |\Sigma|^{-\frac{1}{2}} |X^T \Sigma^{-1} X|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} Y^T K Y\right). \tag{20}$$

We provide an upper bound for each of the three factors of the likelihood function, namely $|\Sigma|^{-\frac{1}{2}}$, $|X^T \Sigma^{-1} X|^{-\frac{1}{2}}$ and $\exp\left(-\frac{1}{2} Y^T K Y\right)$.
Suppose $\lambda_{(m)} (> 0)$ is the largest eigenvalue of G so that $G \le \lambda_{(m)} I_{2m}$. Then for $c > 1$

$$\Sigma = ZGZ^T + R \le R + \lambda_{(m)} ZZ^T \le R + c\lambda_{(m)} ZZ^T$$

$$\begin{array}{rcl} \implies \Sigma + c\lambda_{(m)} XX^T & \le & R + c\lambda_{(m)} (XX^T + ZZ^T) \\ \implies Y^T [\Sigma + c\lambda_{(m)} XX^T]^{-1} Y & \ge & Y^T [R + c\lambda_{(m)} WW^T]^{-1} Y, \end{array} \tag{21}$$

where $W = (X, Z)$. After some algebra it is possible to show that $WW^T = FHF^T$, where $F = \mathbf{diag} F_1, F_2$ is a full-column rank matrix and $H = \mathbf{diag} H_1, H_2$ is a p.d. matrix.

For suitable matrices A, B, U, V, using the identity

$$(A+UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1}+VA^{-1}U)^{-1}VA^{-1}$$

and taking $c \to +\infty$ one gets

$$(\Sigma + c\lambda_{(m)}XX^T)^{-1} \to \Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} = K.$$

By applying the same formula to $(R + c\lambda_{(m)}FHF^T)^{-1}$ and taking $c \to +\infty$ one gets from (21) that

$$Y^TKY \geq Y^T[R^{-1} - R^{-1}F(F^TR^{-1}F)^{-1}F^TR^{-1}]Y. \tag{22}$$

Since $R = \mathbf{diag}\sigma_1^2 I_n, \sigma_2^2 I_n$, it follows that

$$F^TR^{-1}F = \mathbf{diag}\sigma_1^{-2}F_1^TF_1, \sigma_2^{-2}F_2^TF_2, \quad F^TR^{-1}Y = \begin{pmatrix}\sigma_1^{-2}F_1^TY_1\\ \sigma_2^{-2}F_2^TY_2\end{pmatrix}$$

and $(F^TR^{-1}F)^{-1} = \mathbf{diag}\sigma_1^2(F_1^TF_1)^{-1}, \sigma_2^2(F_2^TF_2)^{-1}$. Hence,

$$Y^T[R^{-1} - R^{-1}F(F^TR^{-1}F)^{-1}F^TR^{-1}]Y = \frac{s_1^2}{\sigma_1^2} + \frac{s_2^2}{\sigma_2^2},$$

where $s_1^2 = Y_1^T[I - F_1(F_1^TF_1)^{-1}F_1^T]Y_1$ and $s_2^2 = Y_2^T[I - F_2(F_2^TF_2)^{-1}F_2^T]Y_2$. Thus

$$\exp\left(-\frac{1}{2}Y^TKY\right) \leq \exp\left(-\frac{s_1^2}{2\sigma_1^2} - \frac{s_2^2}{2\sigma_1^2}\right). \tag{23}$$

To evaluate $|\Sigma|$ and $|X^T\Sigma^{-1}X|$, we arrange the elements of Y as shown below:
$Y_{ij} = \binom{Y_{1ij}}{Y_{2ij}}$, $X_{ij} = \begin{pmatrix}x_{1ij}^T & 0^T\\ 0^T & x_{2ij}^T\end{pmatrix}$, $Y_i = \mathrm{col}_{i\leq j\leq n_i}Y_{ij}$, $X_i = \mathrm{col}_{i\leq j\leq n_i}X_{ij}$.
Let $s_i = \binom{v_{1i}}{v_{2i}}$, $e_{ij} = \binom{e_{1ij}}{e_{2ij}}$. Write $Var(s_i) = \Sigma_V$ and $Var(e_{ij}) = \begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix} = \Sigma_e$.
Let $\Sigma_i = Var(Y_i) = I_{n_i} \otimes \Sigma_e + J_{n_i} \otimes \Sigma_V$ where $\otimes$ is the Kronecker product operator and J_k is a k squared matrix of 1s. As proved in Rao (1973) (Problem 1.3, p.68)

$$|\Sigma_i| = |\Sigma_e|^{n_i-1}|\Sigma_e + n_i\Sigma_V| = |\Sigma_e|^{n_i}|I_2 + n_i\Sigma_e^{-1/2}\Sigma_V\Sigma_e^{-1/2}|.$$

Hence

$$|\Sigma| \quad = \quad \prod_{i=1}^m |\Sigma_i| = |\Sigma_e|^n \prod_{i=1}^m |I_2 + n_i\Sigma_e^{-1/2}\Sigma_V\Sigma_e^{-1/2}|.$$

In order to define an upper bound for $|X^T\Sigma^{-1}X| = \sum_{i=1}^m X_i^T\Sigma_i^{-1}X_i$, we note that:

$$\Sigma_i^{-1} = \left(I_{n_i} - \frac{1}{n_i}J_{n_i}\right) \otimes \Sigma_e^{-1} + \frac{J_{n_i}}{n_i} \otimes (\Sigma_e + n_i\Sigma_v)^{-1}.$$

Let O_i be an $n_i \times n_i$ orthogonal matrix such that

$$O_i \left(I_{n_i} - \frac{1}{n_i} J_{n_i} \right) O_i^T = \mathbf{diag} 1, 1, \ldots 1, 0; \quad O_i \left(\frac{1}{n_i} J_{n_i} \right) O_i^T = \mathbf{diag} 0, 0, \ldots, 0, 1 \quad (24)$$

and set $Q_i = O_i \otimes I_2$. By (24), and noting that $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$, it follows that

$$\begin{aligned} Q_i \Sigma_i^{-1} Q_i^T &= Q_i \left[\left(I_{n_i} - \frac{1}{n_i} J_{ni} \right) \otimes \Sigma_e^{-1} + \frac{J_{n_i}}{n_i} \otimes (\Sigma_e + n_i \Sigma_v)^{-1} \right] Q_i^T \\ &= \mathbf{diag} \Sigma_e^{-1}, \Sigma_e^{-1}, \ldots, \Sigma_e^{-1}, (\Sigma_e + n_i \Sigma_v)^{-1}. \end{aligned}$$

Then

$$X_i^T \Sigma_i^{-1} X_i = c_i^T \mathbf{diag} \Sigma_e^{-1}, \Sigma_e^{-1}, \ldots, \Sigma_e^{-1}, (\Sigma_e + n_i \Sigma_V)^{-1} c_i$$

where $c_i = (c_{i1}^T, \ldots, c_{in_i}^T)^T = Q_i X_i$. Hence,

$$X_i^T \Sigma_i^{-1} X_i = \sum_{k=1}^{n_i - 1} c_{ik}^T \Sigma_e^{-1} c_{ik} + c_{in_i}^T (\Sigma_e + n_i \Sigma_V)^{-1} c_{in_i}.$$

Note that $c_{ik}^T \Sigma_e^{-1} c_{ik} = \mathbf{diag} \sigma_1^{-2} \alpha_{ik1} \alpha_{ik1}^T, \sigma_2^{-2} \alpha_{ik2} \alpha_{ik2}^T$ where $\alpha_{iku} = \sum_{j=1}^{n_i} o_{ikj} X_{uij}$ $(u = 1, 2)$. It follows that

$$X_i^T \Sigma_i^{-1} X_i \geq \sum_{k=1}^{n_i - 1} c_{ik}^T \Sigma_e^{-1} c_{ik} = \mathbf{diag} \sigma_1^{-2} \sum_{k=1}^{n_i - 1} \alpha_{ik1} \alpha_{ik1}^T, \sigma_2^{-2} \sum_{k=1}^{n_i - 1} \alpha_{ik2} \alpha_{ik2}^T.$$

Since X is a $2n \times (p_1 + p_2)$ matrix of rank $p_1 + p_2$, we can conclude that if $n - m \geq p_1, n - m \geq p_2$

$$|X^T \Sigma^{-1} X|^{-\frac{1}{2}} \leq \text{const} \;\; \sigma_1^{p_1} \sigma_2^{p_2}. \quad (25)$$

Also

$$|\Sigma|^{-1/2} = \sigma_1^{-n} \sigma_2^{-n} \prod_{i=1}^{m} |I_2 + n_i \Sigma_e^{-1/2} \Sigma_V \Sigma_e^{-1/2}|^{-1/2}. \quad (26)$$

Hence

$$\begin{aligned} |\Sigma|^{-\frac{1}{2}} |X^T \Sigma^{-1} X|^{-\frac{1}{2}} &\leq \text{const} \;\; \sigma_1^{-(n-p_1)} \sigma_2^{-(n-p_2)} \prod_{i=1}^{m} |I_2 + n_i \Sigma_e^{-\frac{1}{2}} \Sigma_V \Sigma_e^{-\frac{1}{2}}|^{-\frac{1}{2}} \\ &\leq \text{const} \;\; \sigma_1^{-(n-p_1)} \sigma_2^{-(n-p_2)} \prod_{i=1}^{m} |I_2 + \Sigma_e^{-\frac{1}{2}} \Sigma_V \Sigma_e^{-\frac{1}{2}}|^{-\frac{1}{2}} \\ &= \text{const} \;\; \sigma_1^{-(n-p_1)} \sigma_2^{-(n-p_2)} |I_2 + \Sigma_e^{-\frac{1}{2}} \Sigma_V \Sigma_e^{-\frac{1}{2}}|^{-\frac{m}{2}}. \quad (27) \end{aligned}$$

From (27) and (23), we can conclude that

$$\begin{aligned} & |\Sigma|^{-\frac{1}{2}} |X^T \Sigma^{-1} X|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} Y^T K Y \right) \\ \leq \;& \text{const} \;\; \sigma_1^{-(n-p_1)} \sigma_2^{-(n-p_2)} |I_2 + \Sigma_e^{-\frac{1}{2}} \Sigma_V \Sigma_e^{-\frac{1}{2}}|^{-\frac{m}{2}} \exp\left(-\frac{s_1^2}{2\sigma_1^2} - \frac{s_2^2}{2\sigma_1^2} \right) \\ = \;& g(\Sigma_R, B|Y). \quad (28) \end{aligned}$$

Lemma 3 *The function $f = g \cdot h$ is a non-negative integrable function.*
Proof: From (17) and (28), an upper bound function f for the posterior density is

$$\begin{aligned} & |\Sigma|^{-\frac{1}{2}} |X^T\Sigma^{-1}X|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}Y^TKY\right)\pi^J(\Sigma_R, B) \\ \leq\ & \sigma_1^{-(n-p_1+5)}\sigma_2^{-(n-p_2+5)}|I_2+\Omega|^{-\frac{m}{2}} \exp\left(-\frac{s_1^2}{2\sigma_1^2}-\frac{s_2^2}{2\sigma_1^2}\right) \\ \times\ & \left[\frac{k_1(1+\omega_{11})^{\frac{3}{2}}(1+\omega_{22})^{\frac{3}{2}}}{|I_2+\Omega|^3}+\frac{(1+\omega_{11})^{\frac{5}{2}}(1+\omega_{22})^{\frac{5}{2}}}{|I_2+\Omega|^5}\right. \\ +\ & \left.\frac{k_2(1+\omega_{11})^{\frac{3}{2}}(1+\omega_{22})^{\frac{5}{2}}}{|I_2+\Omega|^4}+\frac{k_3(1+\omega_{11})^{\frac{5}{2}}(1+\omega_{22})^{\frac{3}{2}}}{|I_2+\Omega|^4}\right]. \end{aligned} \tag{29}$$

where $\Omega = \left(\begin{smallmatrix}\omega_{11} & \omega_{12}\\ \omega_{12} & \omega_{22}\end{smallmatrix}\right)$ is a p.d. matrix. We need to check the integrability of the function in (29). It suffices to prove that the quantity

$$\frac{\sigma_1^{-(n-p_1+5)}\sigma_2^{-(n-p_2+5)}}{|I_2+\Omega|^{\frac{m}{2}+3}} \exp\left(-\frac{s_1^2}{2\sigma_1^2}-\frac{s_2^2}{2\sigma_1^2}\right)(1+w_{11})^{\frac{a}{2}}(1+w_{22})^{\frac{a}{2}} \tag{30}$$

is integrable, with $a = 5$. Taking the integral with respect to σ_1^2 and σ_2^2 one can see that integrability is achieved if

$$n - p_j > 0,\ j = 1, 2,$$

that is when $n > max(p_1, p_2)$. Now we consider the integrability of

$$h(\Omega) = |I_2+\Omega|^{-\frac{m}{2}-3}(1+\omega_{11})^{\frac{a}{2}}(1+\omega_{22})^{\frac{a}{2}}.$$

For $a > 0$,

$$(1+\omega_{11})^{\frac{a}{2}}(1+\omega_{22})^{\frac{a}{2}} \leq (1+\omega_{11}+\omega_{22})^a.$$

Thus

$$h(\Omega) \leq |I_2+\Omega|^{-\frac{m}{2}-3}(1+\omega_{11}+\omega_{22})^a = h^*(\Omega).$$

Let $l_1 > l_2 > 0$ be the eigenvalues of Ω. Then $|I_2+\Omega| = (1+l_1)(1+l_2)$ and $1+\omega_{11}+\omega_{22} = 1+l_1+l_2$. It follows that $h^*(\Omega) = (1+l_1)(1+l_2)^{-\frac{m}{2}-3}(1+l_1+l_2)^a$. Then

$$\begin{aligned} \int h^*(\Omega)d\Omega &= \iint\limits_{0\leq l_2\leq l_1\leq\infty} \frac{(1+l_1+l_2)^a}{[(1+l_1)(1+l_2)]^{\frac{m}{2}+3}}dl_1dl_2 \\ &= \int_0^\infty\int_0^\infty \frac{[(1+z)(1+t)+z]^a}{(1+z)^{m+5}(1+t)^{\frac{m}{2}+3}}dzdt \\ &= \sum_{j=0}^a \binom{a}{j}\int_0^\infty \frac{z^{a-j}}{(1+z)^{m-j+5}}dz\int_0^\infty \frac{1}{(1+t)^{m/2-j+3}}dt, \end{aligned} \tag{31}$$

where we have used the change of variable $l_2 = z, \quad l_1 = z+(1+z)t$. Hence the integral (31) is finite if and only if

$$m > (2a - 4) \text{ and } m - a + 4 > 0.$$

It follows that the posterior is proper provided that $m > 6$ and $n - \max(p_1, p_2) > 0$. This concludes the proof of Theorem 1.

References

Battese, G., Harter, R., and Fuller, W. (1988). "An error components model for prediction of county crop areas using survey and satellite data." *Journal of the American Statistical Association*, 83: 23–36. 372

Datta, G. and Rao, J. (2010). "The choice of nonsubjective priors on hyperparameters for hierarchical Bayes models." In *Frontiers of Statistical Decition Making and Bayesian Analysis: In Honor of James O. Berger*, 237–247. 367

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1: 515–533. 372

Ghosh, M. and Rao, J. (1994). "Small area estimation: an appraisal." *Statistical Sciences*, 9: 55–93. 363

Ghosh, M., Sinha, K., and Kim, D. (2006). "Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models." *Scandinavian Journal of Statistics*, 33: 591–608. 364, 365, 366, 369, 370, 371, 372

Gilks, W., Best, N., and Tan, K. (1995). "Adaptive rejection Metropolis sampling." *Applied Statistics*, 44: 455–472. 370

Hobert, J. and Casella, G. (1996). "The effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models." *Journal of the Americal Statistical Association*, 91(436): 1461–1473. 372

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press, 3rd edition. 368

Kass, R. and Wasserman, L. (1996). "The selection of prior distribution via formal rules." *Journal of the American Statistical Association*, 91: 1343–1370. 368

Rao, C. (1973). *Linear Statistical Inference and Its Applications*. USA: Wiley. 379

Rao, J. (2003). *Small Area Estimation*. Hoboken, New Jersey: Wiley. 363, 364

RDevelopmentCoreTeam (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. 373

Torabi, M., Datta, G., and Rao, J. (2009). "Empirical Bayes estimation of small area means under nested error linear regression model with measurement errors in the covariates." *Scandinavian Journal of Statistics*, 36: 355–368. 366

Acknowledgments

G.S. Datta thanks the Italian PRIN 2008 grant - CEFF37;002, *Bayesian Methods for Finite Populations* for travel support. He was also partially supported through the grants DMS 0914603 from the National Science Foundation and H98230-11-1-0208 from the National Security Agency. The work of Brunero Liseo and Serena Arima has been fully supported by MIUR grant 2008 CEFF37;002, *Bayesian Methods for Finite Populations*, Italy.