Biometrika (2008), **95**, 2, *pp*. 514–520
© 2008 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asn008
Advance Access publication 23 April 2008

Miscellanea
A new class of average moment matching priors

BY N. GANESH

Department of Mathematics, University of Maryland, College Park, Maryland 20742, U.S.A.
ganesh@math.umd.edu

AND P. LAHIRI

Joint Program in Survey Methodology, University of Maryland, College Park, Maryland 20742, U.S.A.
plahiri@survey.umd.edu

SUMMARY

We derive a new class of priors for the variance component in the Fay–Herriot model, a mixed regression model widely used in small area estimation. This class includes the well-known uniform or superharmonic prior. Through simulation we illustrate the use of our class of priors.

Some key words: Hierarchical Bayes; Matched priors; Mixed linear model.

1. INTRODUCTION

The Fay–Herriot model (Fay & Herriot, 1979), widely used in small area estimation, consists of two levels. In Level 1, the sampling model,

$$y_i|\theta_i \sim N(\theta_i, D_i), \quad i = 1, \ldots, m,$$

independently for each i. In Level 2, the linking model,

$$\theta_i \sim N(x_i'\beta, \psi), \quad i = 1, \ldots, m,$$

also independently for each i.

Level 1 accounts for the sampling variability of the regular survey estimates y_i of true small area means θ_i. Level 2 links θ_i to a vector of p known auxiliary variables x_i, often obtained from administrative and census records. The sampling variances D_i are assumed to be known. For a comprehensive review of the theory and applications of the above model, see Rao (2003, Ch. 7).

We are primarily interested in estimating the parameters θ_i. The regression coefficients β and the variance component ψ are usually referred to as hyperparameters. In Bayesian analysis, priors, usually noninformative, are specified for the hyperparameters. Morris & Christiansen (1996) used a flat, in Lebesgue measure, prior distribution for β, and assumed the prior variance ψ to be independent of the regression coefficients and uniformly distributed over $\mathbb{R}^+$. This prior specification for the hyperparameters is often recommended; see Berger (1985) and Morris (1983b). The uniform prior on ψ, often referred to as Stein's superharmonic prior, is noninformative and is known to provide admissible minimax procedures in the context of point estimation (Morris & Christiansen, 1996). The uniform prior was also used by Morris (1983a) in obtaining a suitable measure of uncertainty of his empirical Bayes estimator.

There is a parallel empirical best linear unbiased predictor approach in which the Fay–Herriot model is viewed as a mixed regression model. The best linear unbiased predictor of θ_i is first derived and then

the unknown hyperparameters are replaced by their estimates. Different estimators of the hyperparameters arise from the maximum likelihood, residual maximum likelihood and analysis of variance methods; see Rao (2003) for further details. Under certain regularity conditions, Datta et al. (2005) showed that, for a specific area, the hierarchical Bayes estimate and the corresponding empirical best linear unbiased predictor are close in the sense that the difference between them is of order $O_P(m^{-1})$. However, the uniform prior does not ensure similarity of their respective measures of uncertainties. In other words, the difference between the expectation of the posterior variance, derived under the uniform prior, and the mean-squared error of the empirical best linear unbiased predictor is not of order $o(m^{-1})$. Datta et al. (2005) found a prior for ψ that ensures this closeness. Their prior is useful when the target of inference is a specific area mean, but it is not suitable when we are interested in simultaneous inference of more than one small area, e.g. when we are interested in multiple comparisons.

In § 2, we develop a general class of priors for ψ, which includes both the popular uniform prior and the prior proposed by Datta et al. (2005) as special cases. A prior belonging to the proposed class matches the expectation of a weighted average of the posterior variances with the corresponding weighted average of the mean-squared errors of the empirical best linear unbiased predictors, the average being taken over all small areas. Different priors can be produced by choosing the weights in the weighted average. For example, if we assign a weight of 1 to a specific area and 0 to the remaining areas, we obtain the prior of Datta et al. (2005). If the weight for a given area is proportional to the inverse of the squared sampling variance D_i, we obtain the uniform prior. Thus we have an interesting frequentist validation of the well-known uniform prior. If we are not sure about how to weigh the different areas, we can assign equal weight of $1/m$ to each area. We call the prior thus generated the average moment matching prior.

Our motivation for proposing this new class of priors is similar to those of reference priors (Bernardo, 2005) and the objective priors (Datta & Mukherjee, 2004), namely to ensure some level of frequentist validity. Objective priors may be attractive to practitioners. For example, Bayesian credible sets given by matching priors also have approximate frequentist validity. Our motivation for the new class of priors is very similar. We seek such a property in order to achieve some level of objectivity and frequentist validity in the Bayesian procedure. For the problem of estimating mean-squared error, under certain regularity conditions, the frequentist technique has the same level of precision in a higher-order asymptotic sense. However, our goal goes beyond the estimation of mean-squared error. Once the posterior is obtained, we can use it to solve different complex problems such as multiple comparison, ranking and selection, and simultaneous credible intervals, for which the frequentist methods are not well developed for the Fay–Herriot model. Bayesian methods are quite flexible in this respect. Since Bayesian methods depend on the prior choice, by doing such a matching, we are at least making sure that, for the particular case of estimation of mean-squared error, under certain regularity conditions, our method would have the same level of frequentist properties as the corresponding frequentist estimator in a higher-order asymptotic sense.

2. A NEW CLASS OF PRIORS

We seek a prior $\pi(\psi)$ such that the expectation of the weighted average of the posterior variances of θ_i matches the corresponding weighted average of the mean-squared error of the empirical best linear unbiased predictor of θ_i to a certain order; that is, given $\{w_i \geqslant 0,\ i = 1, \ldots, m$, with $\sum_{i=1}^m w_i = 1\}$, we seek a prior $\pi(\psi)$ such that

$$\sum_{i=1}^{m} w_i E[V(\theta_i \mid y) - \mathrm{MSE}\{\hat{\theta}_i(\hat{\psi})\}] = o(1/m), \tag{1}$$

where $V(\theta_i|y)$ is the posterior variance of θ_i, under the prior $\pi(\psi)$, and $E(\cdot)$ and $\mathrm{MSE}(\cdot)$ are taken with respect to the joint distribution of $\theta = (\theta_1, \ldots, \theta_m)'$ and $y = (y_1, \ldots, y_m)'$ under the Fay–Herriot model. Also, $\hat{\theta}_i(\hat{\psi})$ is the empirical best linear unbiased predictor of θ_i, i.e. $\hat{\theta}_i(\hat{\psi}) = x_i'\hat{\beta}(\hat{\psi}) + \hat{\psi}(\hat{\psi} + D_i)^{-1}\{y_i - x_i'\hat{\beta}(\hat{\psi})\}$, $\hat{\beta}(\hat{\psi}) = \{X'\Sigma^{-1}(\hat{\psi})X\}^{-1}X'\Sigma^{-1}(\hat{\psi})y$ is the best linear unbiased estimator

of β, $\Sigma(\hat{\psi}) = \text{diag}(D_1 + \hat{\psi}, \ldots, D_m + \hat{\psi})$, $X = (x_1, \ldots, x_m)'$ and $\hat{\psi}$ is the residual maximum likelihood estimator (Jiang, 1996) of ψ.

In order to satisfy (1), $\pi(\psi)$ must satisfy the differential equation

$$\frac{d\pi(\psi)}{d\psi}\frac{1}{\pi(\psi)} - 2\frac{\sum_{i=1}^m w_i D_i^2/(D_i+\psi)^3}{\sum_{i=1}^m w_i\{D_i/(D_i+\psi)\}^2} + 2\frac{\sum_{i=1}^m 1/(D_i+\psi)^3}{\sum_{i=1}^m 1/(D_i+\psi)^2} = 0; \tag{2}$$

see the Appendix.

It can be checked that the solution to (2) is given by

$$\pi(\psi) \propto \frac{\sum_{i=1}^m 1/(D_i+\psi)^2}{\sum_{i=1}^m w_i\{D_i/(D_i+\psi)\}^2}. \tag{3}$$

When the prior is given by (3), it follows from Datta et al. (2005) that, for $m - 2 > \text{rank}(X)$, the posterior distribution of θ is proper. It is interesting to note that the uniform prior is a special case of (3), corresponding to $w_i = D_i^{-2}/(\sum_{j=1}^m D_j^{-2})$. By taking $w_i = 1/m$, for $i = 1, \ldots, m$, we obtain what we refer to as the average moment matching prior,

$$\pi(\psi) \propto \frac{\sum_{i=1}^m 1/(D_i+\psi)^2}{\sum_{i=1}^m \{D_i/(D_i+\psi)\}^2}. \tag{4}$$

Note that, even though we have used the mean-squared error of the empirical best linear unbiased predictor of the θ_i's to derive (4), the average moment matching prior does not depend on these quantities.

Also, by taking $w_j = 1$ for $j = i$, and $w_j = 0$ for $j \neq i$, we obtain the following prior:

$$\pi(\psi) \propto \frac{(D_i+\psi)^2}{D_i^2}\sum_{j=1}^m \frac{1}{(D_j+\psi)^2}. \tag{5}$$

Since for the Fay–Herriot model the D_i's are assumed to be known, (5) is equivalent to the following prior, obtained by Datta et al. (2005):

$$\pi(\psi) \propto (D_i+\psi)^2\sum_{j=1}^m \frac{1}{(D_j+\psi)^2}. \tag{6}$$

The main motivation of Datta et al. (2005) was to choose a prior distribution for ψ such that the posterior variance of θ_i is second-order unbiased or nearly unbiased for the mean-squared error of the empirical best linear unbiased predictor of θ_i, i.e.

$$E\{V(\theta_i \mid y)\} = \text{MSE}\{\hat{\theta}_i(\hat{\psi})\} + o(1/m). \tag{7}$$

Note that, unless $D_i = D$ $(i = 1, \ldots, m)$, the prior for ψ is area-specific and hence it is not possible to select a prior that satisfies (7) simultaneously for $i = 1, \ldots, m$. When $D_i = D, i = 1, \ldots, m$, the uniform prior, the prior of Datta et al. (2005) and the average moment matching priors are all identical.

3. SIMULATION

In this section, we use a Monte Carlo simulation study to compare the mean-squared error, which is the same as the integrated Bayes risk, of the empirical best linear unbiased predictor of θ_i and the associated confidence interval with the corresponding hierarchical Bayes estimators and credible intervals of θ_i resulting from three different prior distributions for ψ, namely the average moment matching prior, the uniform prior and the prior suggested by Datta et al. (2005). In addition, we have also examined similar properties of the frequentist and Bayesian estimators and confidence or credible intervals for ψ.

We consider $m = 15$ small areas, which are classified into five groups, G_1, G_2, G_3, G_4 and G_5, with three small areas in each group. The sampling variances for the three small areas in groups $G_1 - G_5$ are respectively 4·0, 0·6, 0·5, 0·4 and 0·1; that is, all three small areas in group G_1 have sampling variance 4·0, all three small areas in group G_2 have sampling variance 0·6, and so on. Note that the chosen sampling variances correspond to the Type III pattern of Datta et al. (2005).

For the entire simulation $\beta = (1, 1)'$ is fixed. The single covariate x_i is generated from a Un(0, 1) distribution, and then fixed for the entire simulation study. We consider two different values, 1 and 2, for ψ. For a specific choice of ψ and prior distribution, we perform 500 simulation runs. For each run, the posterior distributions of θ, β and ψ are approximated by 10 000 runs of a Monte Carlo method using a standard accept–reject algorithm (Robert & Casella, 1999, Ch. 2). For each of the prior distributions, our objective is to compare the coverage of a 95% equal-tailed credible interval for the θ_i's and ψ, the lengths of the aforementioned credible intervals, and the mean-squared error of the θ_i's and ψ. We estimate these quantities by averaging over the 500 simulation runs. Moreover, to save space, when summarizing the results for the θ_i's, we do so by groups $G_1 - G_5$; that is, we average the coverage, length and mean-squared error over all small areas in the same group, in addition to averaging over the 500 simulation runs. In computing the average coverage, the average length of the credible interval and the average mean-squared error, the joint distribution of y and θ in the Fay–Herriot model is used. We also consider a frequentist method and compute similar summary statistics. For the frequentist method, the associated confidence interval for θ_i is given by

$$\hat{\theta}_i(\hat{\psi}) \pm 1{\cdot}96[\mathrm{mse}\{\hat{\theta}_i(\hat{\psi})\}]^{1/2},$$

where $\hat{\psi}$ is the residual maximum likelihood estimator of ψ, $\hat{\theta}_i(\hat{\psi})$ is the empirical best linear unbiased predictor of θ_i and $\mathrm{mse}\{\hat{\theta}_i(\hat{\psi})\}$ is the estimator of mean-squared error of the empirical best linear unbiased predictor of θ_i, derived by Datta & Lahiri (2000). Moreover, the frequentist confidence interval for ψ is given by

$$\hat{\psi} \pm 1{\cdot}96\{I(\hat{\psi})\}^{-1/2},$$

where $I(\psi)$ is the expected Fisher information for ψ and $I(\hat{\psi}) = \{I(\psi)\}_{\psi=\hat{\psi}}$. Note that the Datta–Rao–Smith prior given by (6) is area-specific. Hence, we need to choose a specific value of D_i to implement (6). We have chosen (6) with $D_i = 0{\cdot}4$, i.e.

$$\pi(\psi) \propto (0{\cdot}4 + \psi)^2 \sum_{j=1}^{m} \frac{1}{(D_j + \psi)^2}.$$

For $\psi = 1, 2$, Tables 1 and 2 summarize the simulation results for the frequentist method and the three choices of prior distribution. The columns G_1–G_5 summarize the average coverage, average length and average mean-squared error for groups G_1–G_5. As mentioned previously, for each group, we average over all small areas in the same group and over all 500 simulation runs. The columns labelled ψ in Tables 1 and 2 give similar summary statistics for the prior variance ψ, though, unlike for the θ_i's, the average is only taken over all simulation runs.

As can be seen from Table 1, when $\psi = 1$, in terms of estimation of ψ, significant gains can be achieved by using the average moment matching prior as opposed to the uniform prior. For example, the average length of the credible interval is 27% shorter and the average mean-squared error is reduced by 51%. However, when $\psi = 2$, see Table 2, the gains are smaller; for example, for ψ, the average length is reduced by 21% and the average mean-squared error is reduced by 42%. When estimating θ_i, for $\psi = 1$ and group G_1, by using the average moment matching prior as opposed to the uniform prior, we obtain a 10% reduction in the average length of the credible interval for approximately the same coverage and a 5% reduction in average mean-squared error. For groups G_2 and G_3, the gains are smaller, amounting to approximately a 4% reduction in average length. The results are similar when the average moment matching prior is compared to the Datta–Rao–Smith prior. When the frequentist method is compared with

Table 1. *Summary of simulation results with the frequentist method, average moment matching prior, uniform prior and Datta–Rao–Smith prior; for $m = 15$ and $\psi = 1$*

	G_1	G_2	G_3	G_4	G_5	ψ
			Frequentist method			
Coverage	0·905	0·925	0·936	0·934	0·962	0·832
Length	3·696	2·589	2·409	2·179	1·253	2·173
MSE	0·976	0·466	0·405	0·320	0·092	0·417
			Average moment matching prior			
Coverage	0·941	0·945	0·940	0·943	0·952	0·946
Length	3·987	2·506	2·356	2·135	1·193	3·630
MSE	0·988	0·460	0·402	0·317	0·092	0·865
			Uniform prior			
Coverage	0·955	0·949	0·953	0·951	0·954	0·928
Length	4·441	2·615	2·445	2·211	1·203	4·974
MSE	1·038	0·460	0·397	0·314	0·091	1·779
			Datta–Rao–Smith prior			
Coverage	0·955	0·944	0·947	0·947	0·954	0·920
Length	4·393	2·596	2·431	2·199	1·202	5·029
MSE	1·043	0·464	0·399	0·316	0·092	1·838

MSE, mean-squared error.

Table 2. *Summary of simulation results with the frequentist method, average moment matching prior, uniform prior and Datta–Rao–Smith prior; for $m = 15$ and $\psi = 2$*

	G_1	G_2	G_3	G_4	G_5	ψ
			Frequentist method			
Coverage	0·927	0·955	0·944	0·956	0·953	0·868
Length	4·694	2·796	2·578	2·319	1·234	3·615
MSE	1·543	0·478	0·451	0·343	0·098	1·037
			Average moment matching prior			
Coverage	0·949	0·952	0·941	0·951	0·952	0·944
Length	4·882	2·719	2·531	2·281	1·215	6·120
MSE	1·570	0·480	0·452	0·343	0·098	2·205
			Uniform prior			
Coverage	0·955	0·951	0·943	0·955	0·953	0·940
Length	5·222	2·772	2·571	2·317	1·220	7·789
MSE	1·616	0·472	0·449	0·342	0·097	3·785
			Datta–Rao–Smith prior			
Coverage	0·952	0·951	0·941	0·953	0·952	0·936
Length	5·219	2·773	2·571	2·316	1·220	7·919
MSE	1·622	0·476	0·451	0·342	0·097	3·952

MSE, mean-squared error.

the average moment matching prior, for groups G_2–G_5, the average moment matching prior performs slightly better in terms of average length. For group G_1 and $\psi = 1$, even though the frequentist method has shorter average length, it does not achieve the nominal coverage. When ψ is estimated, in terms of mean-squared error, the frequentist method performs better than all the hierarchical Bayes methods, but we would like to stress that the frequentist method can produce an estimate of zero for ψ, and this creates a problem in inference for a given dataset.

When θ_i is estimated, for $\psi = 2$, the gains in using the average moment matching prior as opposed to the uniform prior are smaller. For example, for group G_1, the average length is reduced by 7% and the average mean-squared error is reduced by 3%. The results are similar when the average moment matching prior is compared to the Datta–Rao–Smith prior.

In order to check the influence of the chosen sampling variances, we have repeated the study with the sampling variances for groups G_1–G_5 given by 0·7, 0·6, 0·5, 0·4 and 0·3; this is the Type I pattern of Datta et al. (2005). Note that the sampling variances are more or less equal across areas. For this sampling variance pattern and $\psi = 1, 2$, all three priors give very similar results, tables not provided; that is, when there is little or no variability in the sampling variances, the three priors give similar results.

In summary, compared to the uniform prior and Datta–Rao–Smith prior, the average moment matching prior gives better frequentist properties when there is large variability in the sampling variances and when for some small areas D_i/ψ is large. Moreover, in this case, when θ_i is estimated, the average moment matching prior has better frequentist properties for the small areas for which D_i/ψ is large; for the small areas for which D_i/ψ is small, all three priors have similar frequentist properties.

Acknowledgement

We thank Professor D. M. Titterington and referees for constructive comments and suggestions which have improved the original version of this paper. We would also like to thank Gauri Datta for pointing out a typo in the paper.

Appendix

Derivation of (2)

When $\hat{\psi}$ is the residual maximum likelihood estimator of ψ, using the approximations for $E\{V(\theta_i \mid y)\}$ and $\mathrm{MSE}\{\hat{\theta}_i(\hat{\psi})\}$ given in Datta et al. (2005), we have

$$E\{V(\theta_i \mid y)\} = g_{1i}(\psi) + g^*_{1\pi i}(\psi) + g_{2i}(\psi) + o(1/m),$$
$$\mathrm{MSE}\{\hat{\theta}_i(\hat{\psi})\} = g_{1i}(\psi) + g_{2i}(\psi) + g_{3i}(\psi) + o(1/m),$$

where

$$g_{1i}(\psi) = \frac{D_i\psi}{D_i+\psi},$$

$$g_{2i}(\psi) = \frac{D_i^2}{(D_i+\psi)^2} x_i' \left(\sum_{j=1}^m \frac{x_j x_j'}{D_j+\psi} \right)^{-1} x_i,$$

$$g^*_{1\pi i}(\psi) = 2\frac{D_i^2}{(D_i+\psi)^2} \frac{1}{\sum_{j=1}^m (D_j+\psi)^{-2}} \left\{ \frac{d\pi(\psi)}{d\psi}\frac{1}{\pi(\psi)} - \frac{1}{D_i+\psi} + 2\frac{\sum_{j=1}^m (D_j+\psi)^{-3}}{\sum_{j=1}^m (D_j+\psi)^{-2}} \right\},$$

$$g_{3i}(\psi) = 2\frac{D_i^2}{(D_i+\psi)^3} \frac{1}{\sum_{j=1}^m (D_j+\psi)^{-2}}.$$

Using condition (1), we obtain that

$$\begin{aligned}
&\frac{d\pi(\psi)}{d\psi}\frac{1}{\pi(\psi)} \sum_{i=1}^m w_i \frac{D_i^2}{(D_i+\psi)^2} \frac{1}{\sum_{i=1}^m (D_i+\psi)^{-2}} - \sum_{i=1}^m w_i \frac{D_i^2}{(D_i+\psi)^3} \frac{1}{\sum_{i=1}^m (D_i+\psi)^{-2}} \\
&\quad + 2\sum_{i=1}^m w_i \frac{D_i^2}{(D_i+\psi)^2} \frac{\sum_{i=1}^m (D_i+\psi)^{-3}}{\left\{\sum_{i=1}^m (D_i+\psi)^{-2}\right\}^2} \\
&= \sum_{i=1}^m w_i \frac{D_i^2}{(D_i+\psi)^3} \frac{1}{\sum_{i=1}^m (D_i+\psi)^{-2}}
\end{aligned}$$

and by rearranging terms we obtain (2).

REFERENCES

BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis,* 2nd ed. New York: Springer.
BERNARDO, J. M. (2005). Reference analysis. In *Handbook of Statistics*, **25**, Ed. D. K. Dey and C. R. Rao, pp. 17–90. Amsterdam: Elsevier.
DATTA, G. S. & LAHIRI, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sinica* **10**, 613–27.
DATTA, G. S. & MUKHERJEE, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. New York: Springer.
DATTA, G. S., RAO, J. N. K. & SMITH, D. D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika* **92**, 183–96.
FAY, R. E. & HERRIOT, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *J. Am. Statist. Assoc.* **74**, 269–77.
JIANG, J. (1996). REML estimation: asymptotic behavior and related topics. *Ann. Statist.* **24**, 255–86.
MORRIS, C. N. & CHRISTIANSEN, C. L. (1996). Hierarchical models for ranking and for identifying extremes with applications. In *Bayesian Statistics*, **5**, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 277–96. Oxford: Oxford University Press.
MORRIS, C. N. (1983a). Parametric empirical Bayes inference: theory and applications. *J. Am. Statist. Assoc.* **78**, 47–59.
MORRIS, C. N. (1983b). Parametric empirical Bayes confidence intervals. In *Proc. Conf. Sci. Infer. Data Anal. Robustness*, Ed. G. E. P. Box, T. Leonard and C. F. J. Wu, pp. 25–50. New York: Academic Press.
RAO, J. N. K. (2003). *Small Area Estimation*. New York: Wiley.
ROBERT, C. P. & CASELLA, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer.

[*Received August* 2006. *Revised November* 2007]