A Bayesian Semiparametric Fay-Herriot-type model for Small Area Estimation

Silvia Polettini

Abstract Borrowing strength in small area estimation is most often achieved through mixed effects regression models. The default normality assumption for random effects is difficult to check, as they are latent variables. Missing covariates can lead to multimodal distributions of random effects; the distribution may also be skewed. Clearly the difficulties in model checking arise for any other parametric assumption. Estimation of the random effects is crucial for predicting small area quantities, and the effect on model estimates of parametric assumptions is shown to be important [7, 2, 4]. In this paper a semiparametric Bayesian linear mixed effects model is analysed, in which the random effects are modelled through a Dirichlet process. The computational approach follows [5, 6]. The application focuses on a Fay-Herriot-type area level model; in this context, the main aim is to assess improvements in precision of small area predictions.

Key words: Dirichlet process prior, Fay-Herriot model, mixed effects regression models, small area estimation

1 Introduction and model assumptions

The simplest area level model for small area estimation can be expressed as follows:

$$\hat{\theta}_i = \theta_i + e_i \quad e_i \sim N(0, \psi_i) \quad \text{independent} \quad i = 1, \ldots, m \tag{1}$$

$$\theta_i = x_i^T \beta + v_i \quad v_i \sim N(0, \sigma_v^2) \quad \text{independent} \quad i = 1, \ldots, m \tag{2}$$

where m is the number of observed small areas, $\hat{\theta}_i$ is the direct estimator of the area characteristic (a mean or a total), with sampling error e_i (and sampling variance

Silvia Polettini
Sapienza Università di Roma, P.le A. Moro 5 – 00185 Roma, e-mail: s.polettini@gmail.com

ψ_i), θ_i is the true mean value for small area i and finally v_i is a random component accounting for heterogeneity and lack of fit. Combining the previous equations, one obtains a mixed effects linear regression model with normal random components.

For area level models, the distributional assumptions on e_i are usually justified by the properties of the direct estimators θ_i. In what follows, the sampling variances ψ_i are assumed known, as customary in most applications.

In constrast, the normality assumption for the random effects v_i has no justification other than computational convenience and is difficult to detect in practice since it involves unobservable quantities. The problem affects both frequentist and Bayesian analysis, although availability of MCMC techniques makes computational convenience less relevant in the latter framework.

The assumption of normality may fail to represent the distribution of the random effects for several reasons: missing covariates can lead to multimodal distributions; the distribution may be skewed. Accurate estimation of the random effects is crucial for predicting small area quantities; the effect on model estimates of distributional assumptions on the random effects is shown to be important [7, 4]. For instance, the presence of outliers may affect the precision of estimates and induce bias in GLMMs. Also, although point small area mean prediction is robust to deviations from normality, the precision of such predictions is decreased; also, estimation of nonlinear functionals may suffer from misrepresentation of the law of the random effects. For the reasons mentioned above, it would be important to rely on a model that has a flexible specification of the random components, so to achieve a greater flexibility and robustness against model misspecifications. For the Fay-Herriot model [3], [2] develop two robustified versions by describing the random effects by either an exponential power (EP) or a skewed EP distribution and investigate robustness of such Fay-Herriot-type models under deviations from normality. Their aim is to understand whether estimates of linear and especially nonlinear functionals such as the c.d.f. are sensitive to deviations from normality of the random effects. Although the models proposed by [2] are based on distributions that generalize the normal, yet these parametric models may fail to adequately describe the distribution of the random effects, and again the problem of checking the adequacy of these models arises. Following the work by [5], we consider a different extension of the Fay-Herriot model [3] based on Dirichlet process priors (DPP), where the distributional assumption in (2) is replaced by

$$v_i \sim G(\cdot) \quad \text{independent}, \quad i = 1, \ldots, m; \quad G \sim DP(M, N(0, \sigma_v^2)), \tag{3}$$

where $DP(M, \phi)$ stands for the Dirichlet process (DP) with precision parameter M and base measure ϕ which in the context of a generalization of the Fay-Herriot model is natural to assume to be a normal distribution. The representation above not only relaxes the normal assumption, but also provides an enlarged model for describing the random effects.

To complete the specification of the model, we introduce the following priors:

$$\sigma_v^2 \sim IG(a_1, b_1); \quad \beta \sim N(0, dI); \quad M \sim Gamma(a_2, b_2) \tag{4}$$

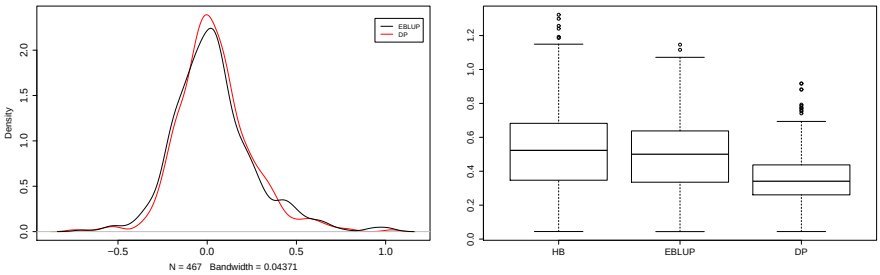

Fig. 1 Performance of the DPP model vs standard estimators: Relative estimation error (left panel) and variance (right panel); MSE is reported for the EBLUP

with fixed hyperparameters. For the DP precision parameter, we follow [6], who show that ML estimation of M, affecting the number of clusters, is the most problematic aspect of the model; a Metropolis-Hastings within Gibbs algorithm, with a Gamma candidate, produced using a Laplace approximation for the calculation of the posterior mean and variance of M is used. For a default specification of the Gamma hyperparameters for M, Dorazio [1] suggests a numeric determination of the values that result in a posterior for the total number of clusters which is closest to the uniform.

The semiparametric setting described in formulae (1–2), (3) and (4) is reported to reduce the variability of the regression parameters estimates [5], producing uniformly shorter HPD intervals than the standard normal random effects models.

2 Application and Comments

The model just described was applied to a single pseudo-sample of areas, obtained by aggregating a sample of individual records drawn from a known population of individuals using a complex sampling scheme, standard in real surveys. The target is here the estimation of the unemployment rate. A set of covariates was introduced and used without any model selection procedure. The true small area figures were known and therefore could be used to assess the estimators. For the characteristics of the sample, the random effects were not designed and therefore *a priori* there is no specific parametric family that can fully describe the random area effects.

The estimates obtained under the DPP model were compared with the EBLUP. For comparison, the standard hierarchical Bayesian (HB) model with "vague priors" is also estimated. The model formalization of the HB model coincides with the DPP except for the definition of the random effects, assumed to be normally distributed. With the "vague" prior choice (large d) HB predictions coincide with those obtained from the EBLUP (see e.g. [8]). The Bayesian estimates were obtained by running Gibbs sampling algorithms.

As expected, the EBLUP and DP prior point estimators of small area percentages perform similarly, and both agree quite well with the true figures (see Fig. 1).

To assess the model, it is important to compare the estimator above with the EBLUP with respect to measures of variability, being this one the feature where the effect of a more flexible specification of the random effects is expected. Since the model was not designed to achieve "calibration" between posterior variance and MSE (see [8], p. 238), comparing the two is not completely appropriate. We refer to the standard hierarchical Bayesian (HB) model as a benchmark. Figure 1 contrasts the posterior variances of the standard HB model across sampled areas with those of the semiparametric model; with a slight abuse of interpretation, the figure also contains the boxplot for the estimated MSE of the EBLUP. In line with [5], the posterior variance is sensibly decreased under the DPP model. Coverage properties of the model can only be assessed by a simulation study, which was not possible in our scheme since we do not have access to the population needed to draw the samples. It is only possible to investigate the fraction of HPD intervals covering the true area mean; the percentage of areas whose population mean value is covered by the .95 credibility interval was 88.8 for the DPP model and 92.1 for the HB model. In light of the limited assessment allowed by this application, the area level flexible model seems to result in accurate estimation of small area quantities.

The performance of the approach should be assessed by means of an extensive simulation. This has not been done in this paper but will be the subject of further analysis.

Acknowledgements Supported by PRIN 2008 grant "Bayesian Methods for Finite Populations".

References

1. Dorazio, R.M.: On selecting a prior for the precision parameter of Dirichlet process mixture models. Journal of Statistical Planning and Inference **139**(9), 3384–3390 (2009)
2. Fabrizi, E., Trivisano, C.: Robust linear mixed models for Small Area Estimation. Journal of Statistical Planning and Inference **140**, 433–443 (2010)
3. Fay, R., Herriot, R.: Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association **74**, 269–277 (1979)
4. Jara, A., Quintana, F.A., San Martn, E.: Linear mixed models with skew-elliptical distributions: A Bayesian approach. Computational Statistics and Data Analysis **52**, 5033–5045 (2008)
5. Kyung, M., Gill, J., Casella, G.: Estimation in dirichlet random effects models. Annals of Statistics **38**, 979–1009 (2010)
6. Kyung, M., Gill, J., Casella, G.: Sampling schemes for generalized linear dirichlet process random effects models. Statistical Methods and Applications **20**(3), 259–290 (2011)
7. Ohlssen, D.I., Sharples, L.D., Spiegelhalter, D.J.: Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. Statistic in Medicine **26**, 2088–2112 (2007)
8. Rao, J.N.K.: Small Area Estimation. Wiley series in survey methodology. John Wiley and Sons, New York (2003)