Empirical Likelihood for Unit Level Small Area Estimation

Sanjay Chaudhuri,
National University of Singapore

Joint with Malay Ghosh, University of Florida, Yan Liyuan and Yin Teng, NUS.

Introduction

- In small area estimation one needs to increase the effective sample size from a particular area by utilising information from from neighbouring areas, or areas similar in characteristics.

- Empirical and hierarchical Bayesian methods provide a natural means for utilising such information.

- Virtually the entire Bayesian and also some of the frequentist small area literature is based on parametric likelihoods. In most cases normal likelihood is used.

- Both conjugate and non-conjugate priors are used and the prior parameters are usually estimated as in an empirical Bayesian analysis or modelled as in a hierarchical Bayesian analysis.

- The frequentist literature often does not assume normality, but they need to assume linearity of the predictors in order to estimate their standard errors.

Our Objective

- We explore an alternative semi-parametric Bayesian approach for small area estimation based on empirical likelihoods.
- Our prior is parametric, but the likelihood is obtained from the empirical distribution estimated from the data under certain constraints.
- Bayesian empirical likelihood for small area estimation can handle continuous and discrete data in a unified manner.
- Both area and unit level models can be handled as well.
- In this talk we look at the unit level models.

Unit Level Analysis

- We consider an unit level direct estimates y_{ij}, with $j = 1, \ldots, n_i$, $i = 1, \ldots, m$ for m small areas.

- In standard parametric Bayesian analysis based on regular one-parameter exponential family models for the ith area one assumes:

$$y_{ij}|\eta_{ij} \overset{ind}{\sim} \exp[\phi_i^{-1}\{\eta_{ij}y_{ij} - \psi(\eta_{ij})\} + c(y_{ij}, \phi_{ij})] \quad \text{for } j = 1, 2, \ldots, n_i, \tag{1}$$

$$\theta_{ij} = x_{ij}^t\beta + u_i, \qquad E\left(u_i \mid \beta, A\right) = 0, \qquad \text{var}\left(u_i \mid \beta, A\right) = A^2, \tag{2}$$

  where $\theta_{ij} = h(\eta_{ij})$, with a strictly increasing link function h. Variable u denotes the random effects. The scale parameters $\phi_i (> 0)$ are assumed to be known.

- Recall that the first and second Bartlett identities applied to (1) implies:

$$\begin{aligned} E(y_{ij}|\eta_{ij}) &= \psi'(\eta_{ij}) = \psi' \cdot h^{-1}(\theta_{ij}) = k\left(\theta_{ij}\right), \\ V(y_{ij}|\eta_{ij}) &= \phi_{ij}\psi''(\eta_{ij}) = \phi_{ij}\left(\psi'' \cdot h^{-1}\right)\left(\theta_{ij}\right) = V\left(\theta_{ij}\right). \end{aligned} \tag{3}$$

Semi-parametric Bayesian Methodology

- We don't specify any parametric likelihood but estimate it through a constrained empirical distribution function.

- Suppose $\theta_i = (\theta_{i1}, \ldots, \theta_{in_i})^T$ is the parameter vector for the ith area and $w_i = (w_{i1}, \ldots, w_{in_i})^T$ is the vector of possible jumps at the points $y_{i1}, y_{i2}, \ldots, y_{in_i}$, determining the empirical distribution function. For all $w_i \in \Delta_{n_i - 1}$, the n_i dimensional simplex. Let

$$\mathcal{W}_{\theta_i} = \left\{ w_i \ : \ \sum_{j=1}^{n_i} w_{ij}\{y_{ij} - k(\theta_{ij})\} = 0, \sum_{j=1}^{n_i} w_{ij} \left[\frac{\{y_{ij} - k(\theta_{ij})\}^2}{V(\theta_{ij})} - 1 \right] = 0 \right\}. \tag{4}$$

  For a given θ_i, the contribution of the ith area to likelihood is defined as:

$$l(\theta_i) = \prod_{j=1}^{n_i} \widehat{w}_{ij}(\theta_i), \text{ where } \widehat{w}_i(\theta_i) = \arg \max_{w_i \in \mathcal{W}_{\theta_i}} \sum_{i=1}^{n_i} f(w_{ij}(\theta_i)), \tag{5}$$

  for some specified function f.

- For a fixed θ_i, $\mathcal{W}_{\theta_i}$ is convex, so any concave f will have a unique maxima.

Empirical Likelihood

- For a given θ_i, $l(\theta_i)$ equals the empirical likelihood ($EL\,(\theta_i)$) when $f(w_{ij}(\theta_i)) = \log(w_{ij}(\theta_i))$.

- Thus $EL(\theta_i) = \max_{w \in \mathcal{W}_{\theta_i}} \prod_{i=1}^{n_i} w_{ij}\,(\theta_i)$.

- Owen (2001) viewed empirical likelihood as the constrained maximum of a non-parametric "likelihood". He showed that under the truth and iid setting the usual Wilks statistics obtained from empirical likelihood has an asymptotic chi-squared distribution.

- Following a general result of Monahan and Boos(1992), Lazar(2003) justifies the use of empirical likelihood in Bayesian inference. See also Fang and Mukherjee (2006) and Mukherjee (2008).

- $EL(\theta)$ can also be expressed (see Qin and Lawless (1994)) as profile empirical likelihood of θ. Thus the posterior resulting from it can be interpreted as a posterior-profile likelihood.

- Grendar and Judge (2009) chose a Bayesian interpretation of $EL(\theta)$ to prove a version of the Sanov's theorem.

Exponentially Tilted Empirical Likelihood

- The exponentially tilted empirical likelihood ($ET(\theta)$) is obtained when $f(w_{ij}(\theta_i)) = -w_{ij}(\theta_i)\log(w_{ij}(\theta_i))$.
- The exponentially tilted empirical likelihood maximises an entropy.
- Schennach (2005) introduced $ET(\theta_i)$ and showed that it is obtained naturally from certain priors in a non-parametric Bayesian context.
- One can also view $ET(\theta_i)$ as an profile likelihood of θ_i. However in general it would be different from $EL(\theta_i)$.
- Both EL and ET can be viewed as empirical discrepancy measures [Schennach (2007)] and generalised empirical likelihoods [Smith (1997)].
- Any member of the Cressie-Read family can be considered. All won't produce an interpretable likelihood. Further, Grendar and Judge (2009) shows that in the Bayesian setting non-likelihood members may not be consistent.

Separate Unit Level Formulation

- A natural way to construct the likelihood for the whole data is to multiply the likelihoods for each area.

- That is, for a given θ the likelihood in this case is given by:

$$l(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \widehat{w}_{ij}(\theta), \tag{6}$$

  where for each $i = 1, 2, \ldots, m$,

$$\widehat{w}_{ij}(\theta) = \arg \max_{w_i \in \mathcal{W}_{\theta i}} \sum_{j=1}^{n_i} f(w_{ij}(\theta)). \tag{7}$$

- The separate formulation estimates the cumulative distribution for each area individually.

- This can be implemented if the number of observed units in each area is relatively large.

Prior Specification and Bayesian Formulation

- There are several ways to introduce prior information in EL based formulation.
- One can specify priors on the weights (eg. Grendar and Judge [2009]), which would impose a prior on θ.
- For applications a more natural way is to specify priors for θ_i.
- The weights are non-negative and bounded by 1, so any proper prior would produce a proper posterior.
- It is not clear how improper priors would behave.
- Once the prior on θ, $\pi(\theta)$ is specified, the posterior can be calculated up to a constant. In particular we write:

$$\Pi(\theta) \propto \left\{ \prod_{i=1}^{m} \prod_{j=1}^{n_i} \widehat{w}_{ij}(\theta) \right\} \pi(\theta).$$

- The normalising constant cannot be easily calculated.

Computational Issues

- The posterior cannot be expressed in an analytic form. We use Markov chain Monte Carlo to simulate observations from the posterior.

- θ is generated from the joint prior.

- The weights are estimated by maximising the appropriate objective functions under corresponding constraints. This can be done using methods due to Owen or Chen, Sitter and Wu. Computationally it is not very demanding.

- Observations from the posterior distribution of θ is generated by a Markov chain Monte Carlo procedure.

- MCMC procedures require certain amount of thought. This is particularly true if the sample sizes are false and the constraints determining the weights are feasible in a small region.

- This region depends on the proposed value of θ and cannot be determined except in a few easy situations.

Fowlkes, Freeny and Landwehr data

- Fowlkes, Freeny and Landwehr [1988] studied a data set from the job satisfaction survey.

- The response is binary and the subjects are categorised in 3 age, 2 gender, 2 race and 7 broad regions.

- The parameter of interest is the probability of job satisfaction p_{ik}, where i denotes the region and k is one of age$\times$gender$\times$race cell.

- The natural parameters are modelled as (see Fowlkes, Freeny, Landwehr [1988], hosh and Natarajan [1999]):

$$\theta_{ik} = \log\left(p_{ik}/(1-p_{ik})\right) = \mu + \alpha_a + \gamma_g + \eta_r + (\gamma\eta)_{sr} + v_i,$$
$$V_{ik} = p_{ik}\left(1-p_{ik}\right).$$

- The response and the auxiliary variables are binary but the usual EL method is still valid.

Fowlkes, Freeny and Landwehr Data (Results)

	Sample		HB model		EL Method	
Category	est	se	est	se	est	se
White, < 35, M	0.600	0.023	0.606	0.019	0.607	0.011
White, < 35, F	0.538	0.028	0.562	0.024	0.665	0.023
White, $35 - 44$, M	0.666	0.026	0.660	0.021	0.625	0.024
White, $35 - 44$, F	0.667	0.043	0.645	0.029	0.682	0.020
White, > 44, M	0.664	0.023	0.645	0.019	0.682	0.020
White, > 44, F	0.733	0.036	0.703	0.026	0.728	0.020
Other, < 35, M	0.643	0.064	0.658	0.033	0.678	0.034
Other, < 35, F	0.610	0.076	0.549	0.038	0.560	0.031
Other, $35 - 44$, M	0.563	0.124	0.685	0.037	0.695	0.036
Other, $35 - 44$, F	0.688	0.116	0.582	0.041	0.580	0.032
Other, > 44, M	0.842	0.084	0.748	0.033	0.739	0.031
Other, > 44, F	0.444	0.166	0.620	0.041	0.632	0.032

- The HB fits are from Ghosh and Natarajan [1999].
- We put in following prior knowledge: $v_i \mid A \sim N\left(0, A^2\right)$, $A \sim |t_5|$ and $\beta \mid A \sim N\left(0, A^2 N (X^T X)^{-1}/g\right)$, where $N = \sum_{i=1}^{7} n_i$ ans $g = 0.0005$.

Battese, Harter and Fuller Data set (1988)

- The separate unit level model requires at least three data points in each area. For numerical comfort we may need many more.

- Battese, Harter and Fuller (1988) analyses a unit level survey and Satelite data for Corn and Soybean in 12 counties in Iowa.

- Their proposed random effects model is:

  $y_{ij} = \theta_{ij} + \epsilon_{ij},\ \theta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + v_i \quad i = 1, 2, \ldots, T,\ j = 1, 2, \ldots, n_i$

  where y_{ij} is the reported area under Corn (Soybean) in the jth unit of the ith county, x_{1ij} and x_{2ij} is respectively the No. of pixels under Corn and Soybean identified from the satelite data in that unit. v_i i.i.d. $N\left(0, A^2\right)$ is the area level random effect for the ith area and ϵ_{ij} i.i.d. $N\left(0, \sigma^2\right)$ is the error for the ij cell. The random variables v_i and ϵ_{ij} are assumed to be independent for all i and j.

- Out of twelve, eight counties have less than three observations. There are only one observation in three counties.

- Obviously, the separate unit level model cannot be applied in this case.

Joint Unit Level Formulation

- We propose a joint formulation of all the counties.
- Suppose $\theta = \{\theta_{ij}\}$ and $w = \{w_{11}, w_{12}, \ldots, w_{Tn_T}\}$ be the possible jumps at points y_{ij}, $i = 1, \ldots, T$ and $j = 1, \ldots, n_i$, $n = \sum_{i=1}^{T} n_i$. As before $w \in \Delta_{n-1}$. Let

$$\mathcal{W}_{\theta,\sigma} = \left\{ w \ : \ \sum_{i=1}^{T}\sum_{j=1}^{n_i} w_{ij}\left(y_{ij} - \theta_{ij}\right) = 0, \sum_{i=1}^{T}\sum_{j=1}^{n_i} w_{ij}\left\{\frac{\left(y_{ij} - \theta_{ij}\right)^2}{\sigma^2} - 1\right\} = 0 \right\}.$$

- For a given θ and σ^2, the likelihood is defined as:

$$l\left(\theta\right) = \prod_{i=1}^{T}\prod_{j=1}^{n_i} \widehat{w}_{ij}\left(\theta, \sigma\right),$$

where $\widehat{w}\left(\theta, \sigma\right) = arg\max_{w \in \mathcal{W}_{\theta,\sigma}} \sum_{i=1}^{T}\sum_{j=1}^{n_i} f\left(w_{ij}\left(\theta, \sigma\right)\right)$ for some specified function f.

- The vector w represents the jumps in the empirical joint distribution over all counties and all observed units.

Results for the BHF data

County	Units	Corn				Soybean			
		BHF		EL		BHF		EL	
		pr	se	pr	se	pr	se	pr	se
Cerro Gordo	1	122.2	9.6	122.89	3.68	77.8	12.0	90.61	4.53
Hamilton	1	126.3	9.5	123.67	3.73	94.8	11.8	93.70	4.50
Worth	1	106.2	9.3	119.21	3.66	86.9	11.5	97.52	4.22
Humboldt	2	108.0	8.1	117.86	3.71	79.7	9.7	103.90	4.46
Franklin	3	145.0	6.5	130.42	4.16	65.2	7.6	89.47	5.38
Pochahontas	3	112.6	6.6	104.46	5.16	113.8	7.7	116.10	6.76
Winnebago	3	112.4	6.6	122.24	3.64	98.5	7.7	88.85	4.50
Wright	3	122.1	6.7	121.00	3.67	122.8	7.8	104.42	4.31
Webster	4	115.8	5.8	106.11	5.00	109.6	6.7	115.92	6.49
Hancock	5	124.3	5.3	127.67	3.73	101.0	6.2	94.50	4.72
Kossuth	5	106.3	5.2	121.84	3.54	119.9	6.1	97.36	4.20
Hardin	5	143.6	5.7	134.17	4.69	74.9	6.6	84.77	6.25

- The priors were specified as follows:

- $A^2 \sim IG\,(5/2, 10/2)$, $\sigma^2 \sim IG\,(5/2, 10/2)$, $\alpha \sim \chi^2_5$, $v_i \mid G \sim G$, $G \mid A, \alpha \sim \mathcal{DP}\left(\alpha, N(0, A^2)\right)$ and $\beta \mid \sigma^2 \sim N\left(\hat{\beta}_{OLS}, 0.00004\sigma^2 I\right)$.

Relationship Between Two Formulations

- The main difference is that in the joint formulation we assume that $\sum_{i=1}^{T}\sum_{j=1}^{n_i} w_{ij} = 1$, whereas in the previous formulation we insisted that $\sum_{j=1}^{n_i} w_{ij} = 1$ holds for all $i = 1, 2, \ldots, m$.

- At the likelihood level, the separate formulation treats each county independently, whereas joint formulation allows some dependence.

- However, if all the constraints are of the form $\sum_{j=i}^{n_i} w_{ij} h_i(\theta_{ij}, V_{ij}) = 0$, for EL, the joint formulation and the separate formulations are equivalent.

- To see this define $p_i = \sum_{j=i}^{n_i} w_{ij}$ and $w_{ij} = p_i v_{ij}$. It is clear that $\sum_{j=i}^{n_i} v_{ij} = 1$ and $\sum_{i=1}^{T} p_i = 1$. For EL, the likelihood factors and since there is no parameter constraint on p_i, $\widehat{p}_i = 1/n_i$.

- We can put area level constraints on p_i as well. This may come handy in benchmarking and multiple level models.

Discussion

- We show that empirical likelihood based Bayesian methodology provides a viable alternative to fully parametric procedures in small area estimation.
- Both area and unit level models can be considered.
- A variety of models and priors can be used. A proper prior will always give a proper posterior.
- A one parameter exponential family is not needed either.
- The same method can be applied to general mixed effects models.
- It is not clear how an improper prior would behave.
- In many cases, efficient Markov Chain Monte Carlo methods are required.