Hierarchical Bayes estimation of unit-level small area log-normal models

Aldo Gardini[1], Carlo Trivisano[1] and Enrico Fabrizi[2]
aldo.gardini2@unibo.it

[1]Dipartimento di Scienze Statistiche 'P. Fortunati', Università di Bologna
[2]DISES, Università Cattolica del S. Cuore

ITACOSM Firenze
June 5th, 2019

Outline of the presentation

1. Research motivation
2. Methodological aspects in SAE context
   - The random intercept model
   - Prior specification
3. Simulation studies
   - Model - based
   - Design - based
4. Conclusion and future work

Log-normality and Bayesian inference

Let us consider the simple **unconditional mean** estimation of an i.i.d. sample $X_1, \dots, X_n$ under log-normality assumption:

$$X_i \sim \log \mathcal{N}\left(\xi, \sigma^2\right), \; i = 1, \dots, n.$$

The functional to estimate is $\theta = \exp\{\xi + 0.5\sigma^2\}$. **Is it possible to compute its posterior mean?**

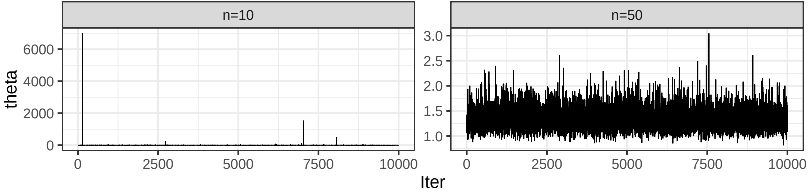
- With usual non-informative priors on σ^2, $p(\theta|\mathbf{X})$ belongs to the log-t family. Its moments coincide with evaluations of the t distribution moment generating function: **they are not finite.**
- To overcome this issue, Fabrizi and Trivisano (2012) proposed the generalized inverse Gaussian (GIG) prior for σ^2:

$$f_X(x) = \left(\frac{\gamma}{\delta}\right)^{\lambda} \frac{1}{2K_{\lambda}(\delta\gamma)} x^{\lambda-1} \exp\left\{-\frac{1}{2}\left(\frac{\delta^2}{x} + \gamma^2 x\right)\right\}, \quad x > 0.$$

What does moment infiniteness implies?

Checking the moments existence is often ignored in practice: usually MCMC methods are applied without checking for the existence conditions. **Do they send warnings about this issue?**

- *Small datasets* ($n < 15$): running reasonably long chains, anomalies can be observed.
- *Larger datasets*: the chains seem to converge to the posterior distribution without issues. **But, which is the meaning of a numerical estimate of an integral that is analytically infinite?**



Bayesian log-normal linear mixed model

Returning to the log-normal context, the same problems are registered in estimating data scale quantities from the general log-normal mixed model.
This model arises when a normal mixed model is assumed on the logarithm of the response variable $\mathbf{w} = \log \mathbf{y}$. This procedure is largely used in practice.

The model could be expressed in the following Bayesian hierarchical formulation:

$$\mathbf{w}|\mathbf{u}, \beta, \sigma^2 \sim \mathcal{N}_n \left(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}, \mathbf{I}_n\sigma^2 \right);$$
$$\mathbf{u}|\tau_1^2, ..., \tau_q^2 \sim \mathcal{N}_m \left(\mathbf{0}, \mathbf{D} \right), \ \mathbf{D} = \oplus_{s=1}^{q} \mathbf{I}_{m_s}\tau_s^2; \quad (\mu, \sigma^2) \sim p(\mu, \sigma^2);$$
$$\boldsymbol{\tau}^2 \sim p(\tau_1^2, ..., \tau_q^2).$$

The moments of the following posterior distributions require care:

- Conditioned expectations with respect to covariates and/or random effects,
- Posterior predictive distribution $p(\tilde{\mathbf{y}}|\mathbf{y})$.

Small area estimation: notation

We considered the classical small area estimation framework:

- **Population** U of N units. It is partitioned into D sub-populations $U_1, \ldots, U_D$ having dimensions $N_1, \ldots, N_D$,
- **A random sample** s with size n is drawn from U. D sub-samples $s_1, \ldots, s_D$ with size $n_1, \ldots, n_D$ are obtained.
- **The unsampled unites** $r_1, \ldots, r_D$ are $N_d - n_d$ in each area.
- The subscript s denotes quantities related to the sampled units, whereas r refers to the unsampled units characteristics.
- $\mathbf{y}_s \in \mathbb{R}^n$ denotes the observed values for the response variable, whereas $\mathbf{w}_s$ is the log-transformed vector.
- The values of p covariates are stored in $\mathbf{X}_s$.

Random intercept model definition

The classical BHF unit-level model is specified for $\mathbf{w}_s$ in the Bayesian framework:

$$\begin{aligned}
&\mathbf{w}_s|\mathbf{u},\beta,\sigma^2 \sim \mathcal{N}_n\left(\mathbf{X}_s\beta + \mathbf{Z}_s\mathbf{u},\sigma^2\mathbf{I}_n\right);\\
&\mathbf{u}|\tau^2 \sim \mathcal{N}_D\left(0,\tau^2\mathbf{I}_D\right),\ (\beta,\sigma^2) \sim p(\beta,\sigma^2);\\
&\tau^2 \sim p(\tau^2);
\end{aligned}$$

where:

- $\mathbf{Z}_s \in \mathbb{R}^{n\times D}$ is the random effects design matrix that defines the random intercept model,
- $\beta \in \mathbb{R}^p$ contains the fixed effects coefficients and $\boldsymbol{u} \in \mathbb{R}^D$ the area-specific random effects.

The target inferential quantities are the **area means**:

$$\hat{\bar{y}}_d(\boldsymbol{\theta}) = N_d^{-1}\left[\sum_{i\in s_d} y_{di} + \sum_{i\in r_d}\hat{y}_{di}\right].$$

Target quantity: area mean

In the hierarchical Bayes context, it is natural to estimate the out-of-sample elements exploiting the posterior predictive distribution:

$$\hat{\bar{Y}}_d^{HB} = \frac{1}{N_d}\left(\sum_{i\in s_d} y_{di} + \sum_{i\in r_d}\mathbb{E}\left[y_{di}|\mathbf{y}_s\right]\right).$$

Theorem (Area mean posterior moments existence)

The posterior moments of $\hat{\bar{Y}}_d^{HB}$ are defined up to order r if the prior of σ^2 includes an exponential term $\exp\{-k\sigma^2\}$ and the hyperparameters are fixed in order to satisfy the inequality:

$$k > r^2 + r^2 h_d^{max}, \quad h_d^{max} = \max_{i\in r_d} \mathbf{x}_{di}^T(\mathbf{X}_s^T\mathbf{X}_s)\mathbf{x}_{di}.$$

Target quantity: poverty measures

The hierarchical Bayes approach is appealing in case of non-linear data transformation target functionals. An example is represented by the family FGT (Foster et al. 1984) of poverty measures, that is defined for subject i of area d as:

$$F_{di,\alpha} = \left(\frac{c - y_{di}}{c} \right)^{\alpha} \mathbf{1}_{\{y_{di} < c\}}(Y_{di});$$

where c is the poverty line and $\alpha = 0, 1, 2$.

The area mean poverty measure HB estimate is:

$$\hat{F}_{d,\alpha}^{HB} = \frac{1}{N_d} \left[\sum_{i \in s_d} F_{di,\alpha} + \sum_{i \in \bar{s}_d} \mathbb{E}\left[F_{di,\alpha} | \mathbf{y}_s \right] \right].$$

Existence condition: same of previous theorem but with $r' = r\alpha$.
The prior used in Molina et al. (2014) does not preserve the posterior moments existence.

Prior specification: general ideas

- **Location parameter:** improper flat prior $p(\beta) \propto 1$.
- **Variance components:** ***how to specify a weakly informative prior preserving the moments existence conditions?***

Variance components

Distribution choice: the flexibility of the three-parameters GIG distribution has been privileged. The existence condition affects the tail parameter γ that is fixed equal to $\gamma_0^2 = (r+1)^2 + (r+1)^2 h^{max}$.

General strategy:

- Preserving the posterior moments existence;
- Saving the balance among the variance components evaluating the marginal prior on the intraclass correlation coefficient ρ. Following the idea of *uniform shrinkage prior* (Chaloner, 1987), a uniform prior on ρ is the target to reach.

Prior specification: general ideas

- **Location parameter:** improper flat prior $p(\beta) \propto 1$.
- **Variance components:** ***how to specify a weakly informative prior preserving the moments existence conditions?***

Variance components

Distribution choice: the flexibility of the three-parameters GIG distribution has been privileged. The existence condition affects the tail parameter γ that is fixed equal to $\gamma_0^2 = (r+1)^2 + (r+1)^2 h^{max}$.

General strategy:

- Preserving the posterior moments existence;
- Saving the balance among the variance components evaluating the marginal prior on the intraclass correlation coefficient ρ. Following the idea of *uniform shrinkage prior* (Chaloner, 1987), a uniform prior on ρ is the target to reach.

Prior specification: some details

Briefly, the main steps covered to formulate our proposal are:

1. Two independent GIG priors having equal hyperparameters are fixed for the variance components:

$$p(\sigma^2) \sim GIG(\lambda, \delta, \gamma_0); \quad p(\tau_s^2) \sim GIG(\lambda, \delta, \gamma_0).$$

2. The marginal prior on $\rho = \tau^2(\sigma^2 + \tau^2)^{-1}$ has the following form:

$$p(\rho) = \frac{K_{2\lambda}\left(\gamma_0^2\delta^2\left[\frac{1}{\rho} + \frac{1}{1-\rho}\right]\right)}{2\left[K_\lambda\left(\gamma_0\delta\right)\right]}\left[\rho(1-\rho)\right]^{-1}, \quad \rho \in (0,1).$$

3. To remove the tail parameter effect, the limiting case $\delta \to 0$ might be considered. If $\lambda > 0$, then the gamma distribution is obtained.
4. Fixing $\lambda = 1$ a uniform prior in the interval $[0,1]$ is assumed for ρ.

$$p(\sigma^2) \sim \mathcal{G}(1, \gamma_0/2); \quad p(\tau_s^2) \sim \mathcal{G}(1, \gamma_0/2).$$

Prior specification: some details

Briefly, the main steps covered to formulate our proposal are:

1. Two independent GIG priors having equal hyperparameters are fixed for the variance components:

$$p(\sigma^2) \sim GIG(\lambda, \delta, \gamma_0); \quad p(\tau_s^2) \sim GIG(\lambda, \delta, \gamma_0).$$

2. The marginal prior on $\rho = \tau^2(\sigma^2 + \tau^2)^{-1}$ has the following form:

$$p(\rho) = \frac{K_{2\lambda}\left(\gamma_0^2\delta^2\left[\frac{1}{\rho} + \frac{1}{1-\rho}\right]\right)}{2\left[K_\lambda\left(\gamma_0\delta\right)\right]}\left[\rho(1-\rho)\right]^{-1}, \quad \rho \in (0,1).$$

3. To remove the tail parameter effect, the limiting case $\delta \to 0$ might be considered. If $\lambda > 0$, then the gamma distribution is obtained.
4. Fixing $\lambda = 1$ a uniform prior in the interval $[0, 1]$ is assumed for ρ.

$$p(\sigma^2) \sim \mathcal{G}(1, \gamma_0/2); \quad p(\tau_s^2) \sim \mathcal{G}(1, \gamma_0/2).$$

Model based simulation setting

Evaluation of the frequentist properties of the area estimates: compared EB and HB (both under GIG priors and uniform shrinkage priors) methods. Design of the simulation study inspired by Berg et al. (2016).

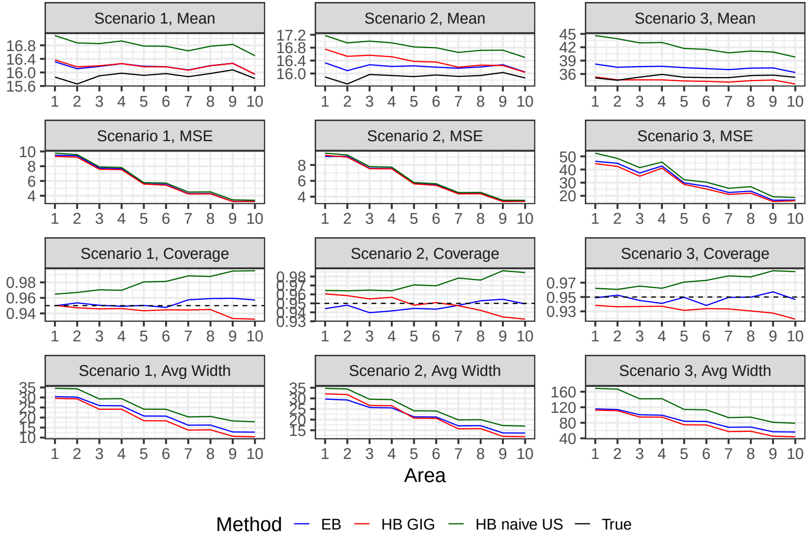
At each iteration ($B = 5000$), a population is randomly generated from the model:

$$w_{di} = \beta_0 + \beta_1 x_{di} + u_d + e_{di};$$
$$u_d \sim \mathcal{N}(0, \tau^2),\ \ e_{di} \sim \mathcal{N}(0, \sigma^2);\ \ d = 1, ..., 10;\ \ i = 1, ..., N_D,$$

and a stratified random sample is drawn. Two equivalent area for each dimension are considered, with $N_d = (41, 81, 161, 323, 645)$ and $n_d = (3, 5, 10, 20, 40)$.

Three scenarios for the variance components (σ^2, τ^2): $(0.6, 0.3)$, $(0.78, 0.12)$ and $(2, 0.5)$.

Model based simulation results



Design based simulation: AAGIS data

It is useful to investigate the design properties of the considered solutions.

The synthetic finite population of size $N = 81982$ based on the *Australian Agricultural Grazing Industries Survey* (AAGIS) has been already used to investigate log-normal SAE methods.

At each iteration ($B = 1000$), a stratified simple random sample ($n = 1686$) is drawn from the 29 areas. The variable of interest is the *annual firm cost* and the available auxiliary information is the logarithm of *firm size*. The same BHF model as before was fitted.

Simulation results:

	EB	HB_{GIG}
Average RRMSE	0.145	0.145
Average bias	0.011	0.011

Conclusion and future work

- **Bayesian analysis of mixed models on the log scale requires care:**
  - Existence of posterior moments should not be taken for granted.
  - We proposed a prior distribution that allows the moments existence for usual target quantity of the SAE framework.

- **Future work:**
  - Implement scalable MCMC algorithms in order to be able to apply Bayesian methods to large input datasets.
  - Work on approximations of the posterior predictive distribution to allow the analysis of whole population as out-of-samples.

References

Berg, E., Chandra, H., and Chambers, R. (2016). Small area estimation for lognormal data. *Analysis of Poverty Data by Small Area Estimation*, pages 279–298.

Chaloner, K. (1987). A bayesian approach to the estimation of variance components for the unbalanced one-way random model. *Technometrics*, 29(3):323–337.

Fabrizi, E. and Trivisano, C. (2012). Bayesian estimation of log-normal means with finite quadratic expected loss. *Bayesian Analysis*, 7(4):975–996.

Molina, I., Nandram, B., Rao, J., et al. (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical bayes approach. *The Annals of Applied Statistics*, 8(2):852–885.