Estimation avec triple objectif appliquée à l'enquête sur la population active des États Unis (U.S. Current Population Survey)

Daniel Bonnéry [1] & Yang Cheng [2] & Partha Lahiri [3]

[1] *Joint Program in Survey Methodology University of Maryland 1218M LeFrak Hall, College Park, MD 20742 dbonnery@umd.edu*

[2] *yang.cheng@census.gov U.S. Census Bureau 4600 Silver Hill Road, Washington DC, DC 20233*

[3] *Joint Program in Survey Methodology University of Maryland 1218M LeFrak Hall, College Park, MD 20742 plahiri@umd.edu*

Résumé Nous présentons une méthode d'estimation sur petits domaines avec objectif triple appliquée à l'enquête sur la population active des États Unis (CPS), en adaptant la méthode de Shen and Louis (1998) d'estimations simultanées sur petits domaines à partir de sondages complexes. Le but principql de cette méthode est de produire une série d'estimateurs (un pour chaque domaine), à partir desquels pourront être produits trois estimateurs, des moyennes par petits domaines, de la fonction de répartition et des rangs qui seront simultanément bons. Des simulations et l'utilisation de méthode de Monte Carlo nous permettent de comparer les estimateurs "tripe objectifs" des estimateurs obtenus à partir des estimateurs directs, des espérances à posteriori et des estimateurs empiriques/hiérarchiques contraints, développés par Louis (1984), Lahiri (1992) et Ghosh (1992). Nous implémentons notre méthodologie en utilisant des méthodes de Monte Carlo par chaîne de Markov (MCMC) et étudions l'impact de cette méthode sur l'estimation du taux de chomage à partir des données du CPS, et de données administratives.

Mots-clés. Petits domaines, Enquête emploi, sondages complexes.

1 Estimation avec triple objectif

Soit $\hat{\pi}_i$ l'estimateur direct du taux de chomage π_i pour le ie petit domaine ($i = 1, \cdots, m$). Nous souhaitons produire des estimateurs triple objectif de $\pi = (\pi_1, \cdots, \pi_m)$. Soit $\hat{\theta}_i = \arcsin(\sqrt{\hat{\pi}_i})$. Etudions le modèle de travail suivant:

Model: Pour $i = 1, \cdots, m$,

$$\begin{aligned} (i) \quad & \hat{\theta}_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, \psi_i); \\ (ii) \quad & \theta_i|\beta, A \overset{\text{ind}}{\sim} N(x_i^T\beta, A), i = 1, \cdots, m; \\ (iii) \quad & f(\beta, A) \propto 1. \end{aligned}$$

Ci-dessus, $\psi_i = \frac{1}{4n_i}$, avec n_i la taille effective de l'échantillon dans le petit domaine i. Nous utilisons $n_i = \frac{\tilde{n}_i}{\text{deff}}$, où $\tilde{n}_i$ est la taille d'échantillon pour le domaine i et deff est l'estimateur de l'effet de plan pour le grands domaine qui contient le petit domaine i. La procédure pour obtenir des estimateurs d'objectif triple, selon Shen and Louis (1998), est la suivante:

D'abord, nous obtenons un estimateur de la distribution empirique de π. La distribution empirique de π est définie par:

$$F_m(t) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{I}\{\pi_i \leq t\}, \tag{1}$$

où $t \in \mathbb{R}$ et $\mathcal{I}$ est la fonction indicatrice. Pour le risque quadratique intégré:

$$\text{ISEL}(F_m, \tilde{F}_m) = \int \left[F_m(t) - \tilde{F}_m(t)\right]^2 dt, \tag{2}$$

l'estimateur de Bayes de la distribution empirique est donné par

$$\hat{F}_m(t) = E\left[F_m(t)|\hat{\theta}\right] = \frac{1}{m}\sum_{i=1}^{m} P(\pi_i \leq t|\hat{\theta}). \tag{3}$$

Ensuite, nous estimons le rang du paramètre π. Le rang est défini par

$$R_i = \text{rang}(\pi_i) = \sum_{j=1}^{m} \mathcal{I}\{\pi_i \geq \pi_j\}. \tag{4}$$

Pour le risque quadratique, défini par

$$\text{RSEL}(R, \tilde{R}) = \frac{1}{m}\sum_{i=1}^{m}(R_i - \tilde{R}_i)^2, \tag{5}$$

l'estimateur de Bayes de R_i est donné par

$$\bar{R}_i = E(R_i|\hat{\theta}) = \sum_{j=1}^{m} P(\pi_i \geq \pi_j|\hat{\theta}). \tag{6}$$

L'estimateur $\bar{R}_i$ peut prendre des valeurs non entières, c'est pourquoi nous utilisons le second estimateur:

$$\hat{R}_i = \text{rang}(\bar{R}_i|R), i = \ldots, m. \tag{7}$$

Enfin; nous générons un ensemble d'estimateurs de niveaux :

$$\hat{\pi}_i^{TG} = \hat{F}_m^{-1}\left(\frac{2\hat{R}_i - 1}{2m}\right), i = 1, \ldots, m. \tag{8}$$

Nous proposons de résoudre le problème avec un échantillonage de Gibbs. Pour l'implémentation, nous utilisons le modèle hiérarchique suivant :

(a) $\theta_i|\beta, A, \hat{\theta} \overset{\text{ind}}{\sim} N\left[(1-B_i)\hat{\theta}_i + B_i x_i'\beta, \frac{\psi_i A}{A+\psi_i}\right],\ i = 1, \cdots, m$

(b) $\beta|\theta, A, \hat{\theta} \sim N\left[(X^TX)^{-1}X^T\theta, A(X^TX)^{-1}\right]$

(c) $A|\beta, \theta, \hat{\theta} \sim IG\left[\frac{1}{2}\sum(\theta_i - x_i^T\beta)^2, \frac{m-2}{2}\right],$

où $B_i = \frac{\psi_i}{A+\psi_i},\ (i = 1, \cdots, m)$ et IG représente une distribution inverse-gamma.

Echantillonage de Gibbs :

(i) Générer $\theta_i^{(1)},\ i = 1, \cdots, m$, depuis (a), en utilisant $\beta^{(0)}$ & $A^{2(0)}$ comme valeurs de départ. Obtenir $\pi_i^{(1)} = \sin^2\left(\theta_i^{(1)}\right),\ i = 1, \cdots, m.$

(ii) Générer $\beta^{(1)}$ depuis (b) en utilisant $\theta^{(1)}$ & $A^{2(0)}$.

(iii) Générer $A^{2(1)}$ depuis (c), en utilisant $\theta^{(1)}$ & $\beta^{(1)}$.

Les étapes (i)-(iii) correspondent à un cycle. Nous réalisons plusieurs cycles. Les échantillons obtenus après avoir écarté les t premiers "burn-in" cycles, i.e.

$$\left\{\beta^{(t+r)}, A^{2(t+r)}, \pi^{(t+r)},\ r = 1, \cdots, R\right\},$$

sont considérés un échantillon de taille R de la distribution a posteriori de β, A, π.

La densité à posteriori de π est approximée pqr

$$\left\{\pi^{(t+r)},\ r = 1, \cdots, R\right\}.$$

En particulier, nous faisons l'approximation:

$$\hat{F}_m(t) \approx \frac{1}{m}\sum_{i=1}^{m}\left\{\frac{1}{R}\sum_{r=1}^{R}\mathcal{I}\left[\pi_i^{(r)} \leq t\right]\right\}, \tag{9}$$

$$\bar{R}_i = \approx \sum_{j=1}^{m}\left\{\frac{1}{R}\sum_{r=1}^{R}\mathcal{I}\left[\pi_i^{(r)} \leq \pi_j^{(r)}\right]\right\}. \tag{10}$$

2 Estimation du taux de chomage pour les Etats Unis

Nous utilisons les données mensuelles de l'enquete CPS pour estimer le taux de chomage pour chaque état entre janvier 2009 à décembre 2012. L'enquete CPS est conduite par le Bureau du Census des Etats Unis et ses échantillons mensuels sont constitués de 72,000 foyers et sont collectés pour 729 aires géographiques, qui correspondent à plus de 1,000 contés aui couvrent chaque état plus le District de Colombie. Pour plus d'information, voir `(http://www.bls.gov/cps/)`.

Nous proposons d'appliquer la méthode triple objectif dans le cas d'un modèle de séries temporelles et transversal.

Nous observons:

- $\hat{\pi}_{t,s}$, les estimateurs direct du taux d'emploi $\pi_{t,s}$ pour le mois t et l'état s, pour chaque mois entre jqnvier 1990 et décembre 2013 ($t = 1, \ldots, T$, $s = 1, \ldots S$).
- un estimateur basé sur plan $\hat{\Sigma}$ de $\mathrm{Var}\left[\hat{\pi} \mid \pi\right]$.
- $X_{t,s}$ le nombre moyen de demandes faites pour l'assurance des chômeurs.

Nous considérons le modèle de travail suivant :
niveau 1:

$$\hat{\pi} \mid \pi \sim \mathscr{N}\left(\pi, \Sigma\right),$$

où Σ est remplacé par son estimateur $\hat{\Sigma}$.
niveau 2:

$$\pi_{t,s} = \mathrm{logit}^{-1}\left(\mu + \alpha_s + X_{t,s}\beta + \eta_{t,s}\right),\ t = 1, \ldots, T,\ s = 1, \ldots, S.$$

niveau 3:

$$\begin{cases} \eta_{t,s} = \eta_{t-1,s} + \zeta_{t,s} \\ \zeta_{t,s} \sim \mathscr{N}(0, \sigma^2) \end{cases}$$

Nous utilisons des a priori plats sur μ, β, α, et σ^2.

D'abord en utilisant des chaînes de Markov Monte-Carlo, nous calculons

$$\bar{R}_{s,T} = \mathrm{E}\left[R_{s,T} | \hat{\pi}\right],$$

l'estimateur de $R_{s,T}$ donné par

$$R_{s,T} = \mathrm{rang}(\pi_{s,T}) = \sum_{s'=1}^{S} \mathbb{1}\{\pi_{s,T} \geq \pi_{s',T}\},$$

et $\hat{R}_{s,T} = \mathrm{rang}\left(\bar{R}_{s,T}\right)$.

Ensuite, nous calculons

$$\hat{F}_{s,T}(\alpha) = \mathrm{E}\left[F_{s,T}(\alpha)|\hat{\pi}\right],$$

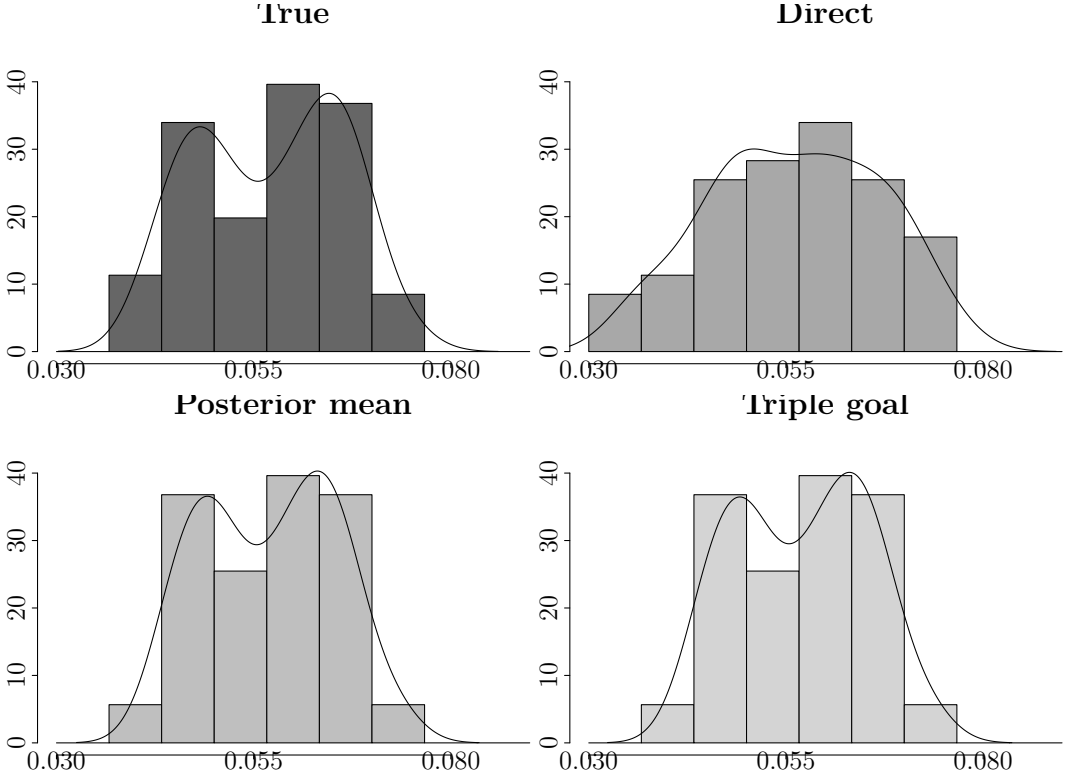
l'estimateur bayésien de la distribution empirique $F_{s,T}$ donné par

$$F_{s,T}(\alpha) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\{\pi_{s,T} \leq \alpha\}.$$

Enfin, nous estimons $\pi_{s,T}$ par

$$\hat{\pi}^{\mathrm{TG}}_{s,T} = F^{-1}_{s,T}\left(\frac{2\hat{R}_{s,T}-1}{2S}\right).$$

Pour les simulations, nous estimons les paramêtres μ, α, β via EM. Nous choisissons σ^2 arbitrairement petit. Nous utilisons ces estimateurs pour générer π; $\hat{\pi}$ selon le modèle précédent, et étudions les estimateurs triple objectif Nous obtenons le résultat suivant :



- Pour $\hat{\pi} = \hat{\pi}^{\mathrm{Dir}},\ \hat{\pi}^{PM},\ \hat{\pi}^{\mathrm{TG}}$, nous calculons les indicateurs de performance suivants:

Root Average Squared Deviation (RASD):

$$\sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\pi}_{s,T}-\pi_{s,T})^2}$$

Root Integrated Squared Error Loss (RISEL):

$$\sqrt{\int\left[F_m(t)-\tilde{F}_m(t)\right]^2 dt}$$

Variance Ratio (VR):

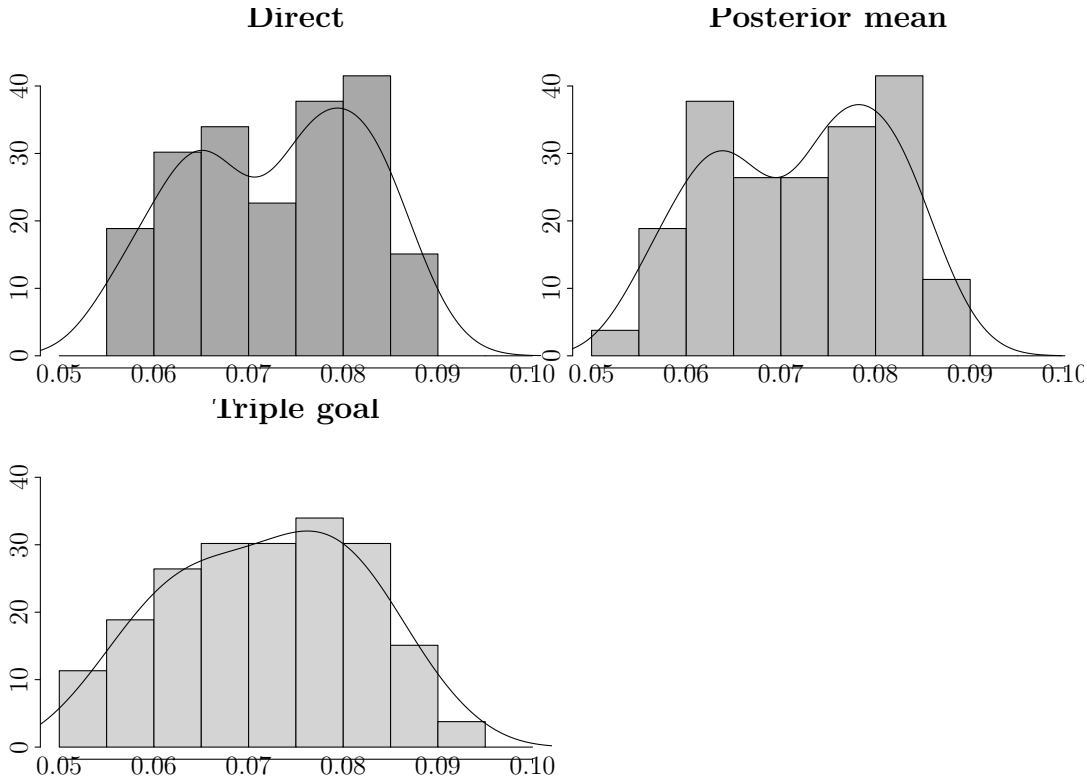
$$\frac{\sum_{s=1}^{S}(\hat{\pi}_{s,T}-\bar{\hat{\pi}}_{.,T})^2}{\sum_{s=1}^{S}(\pi_{s,T}-\bar{\pi}_{.,T})^2}$$

Root Rank Average Squared Deviation (RRASD):

$$\sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{R}_{s,T}-R_{s,T})^2}.$$

	RASD	RRASD	VR	RISEL
direct	0.00535	7.95506	1.33895	0.01032
post. mean	0.00122	1.87335	0.83362	0.00681
tri-goal	0.00122	1.87335	0.83570	0.00655

Nous appliquons ensuite la méthode d'objectif triple aux vraies données, et obtenons le graphique suivant :



Le modèle de travail étudié s'applique mal aux données. Par la suite, nous appliquerons un modèle plus adapté (voir Pfeffermann & Tiller (2006)). Nous étudions aussi d'autres méthodes de calcul de distribution a posteriori, pour palier à la lenteur de l'algorithme de Gibbs, particulièrement problématique ici, compte tenu de la dimension des paramêtres à étudier.

Bibliographie

References

Shen, W. and Louis, T. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of Royal Statistical Society, Series B*, 60:455–471.

Datta, G. S, Lahiri, P., Maiti, T., & Lu, K. L. 1999. Hierarchical Bayes estimation of unemployment rates for the states of the US. *Journal of the American ...*, **94**(448), 1074–1082.

Pfeffermann, D., & Tiller, R. 2006a. Small-Area Estimation With State-Space Models Subject to Benchmark Constraints. *Journal of the American Statistical Association*, **101**(476), 1387–1397.

Ha, N. S. 2013. *Hierarchical Bayesian Estimation of Small Area Means Using Complex Survey Data.*