Régression en composantes principales en sondages

Camelia Goga

(en collaboration avec H. Cardot, M.-A.Shehzad et A. Vanheuverzwyn)
IMB, Université de Bourgogne-Dijon et Médiamétrie
camelia.goga@u-bourgogne.fr

7ème Colloque sur les Sondages
5-7 Novembre ENSAI

Plan de l'exposé

1. Motivation : exemple des enquêtes sur l'audience (Médiamétrie)
2. Calage sur composantes principales : information auxiliaire complète
3. Calage sur composantes principales estimées
4. Application sur les données Médiamétrie

L'enquête sur l'audience à Médiamétrie

Information auxiliaire : information externe (recensement ou une autre enquête) en lien avec les objectifs de l'enquête en cours ;
Objectif : améliorer la précision des estimations en l'utilisant au niveau de l'échantillonnage ou de l'estimation ;
Enquêtes à Médiamétrie : mesurent l'audience de la télévision en France

- **passage au numérique et l'apparition de la TNT** : les données sont de plus en plus nombreuses ;
- développement des offres numériques avec voie de retour permet de savoir à chaque instant, le nombre de boîtiers allumés sur chaque chaîne.

Comment utiliser autant d'information auxiliaire ?

Cadre et notations

- Soit la population $U = \{1, \ldots, k, \ldots, N\}$
- On sélectionne un échantillon $s \subset U$ de taille n selon $p(s)$; les probabilités d'inclusion π_k et π_{kl}
- Soit $\mathcal{Y}$ la variable d'intérêt et on veut estimer

$$t_y = \sum_U y_k$$

- $\mathcal{X}_1, \ldots, \mathcal{X}_p$ variables auxiliaires : âge, CSP, région Insee ...

$$\mathbf{x}_k' = (X_{1k}, \ldots, X_{pk}) \quad \text{pour} \quad k = 1, \ldots, N$$
$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$$

- L'estimateur d'Horvitz-Thompson (HT) sans information auxiliaire :

$$\hat{t}_{yd} = \sum_s \frac{y_k}{\pi_k} = \sum_s d_k y_k$$

Estimation en présence du sur-calage

Difficultés possibles On suppose que le nombre de variables auxiliaires (ou de calage) est très grand.

1. des poids w_k négatifs ou trop grands.
2. utiliser trop de variables de calage peut augmenter la variance (Silva and Skinner, 1997)

Solutions :

1. choisir les variables les plus pertinentes ;
2. utiliser une inverse généralisée (solution proposée dans CALMAR2) ;
3. **relâcher les équations de calage**.

Construction des composantes principales (CP)

- Soient $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ les valeurs propres de $N^{-1}\,\mathbf{X}^\top\mathbf{X}$ et $\mathbf{v}_1, \ldots, \mathbf{v}_p$ les vecteurs propres associés :

$$\left(N^{-1}\,\mathbf{X}^\top\mathbf{X}\right)\mathbf{v}_j = \lambda_j \mathbf{v}_j, \quad j = 1, \ldots, p.$$

- Les composantes principales de $\mathbf{X}$, notées $\mathbf{Z}_1, \ldots, \mathbf{Z}_p$, sont les combinaisons linéaires de variables de départ, $\mathbf{X}_1, \ldots, \mathbf{X}_p$ et qui contiennent le maximum d'information.

$$\mathbf{Z}_j = \mathbf{X}\mathbf{v}_j, \quad j = 1, \ldots, p.$$

- On ne garde que les r premières CP correspondant aux r plus grandes valeurs propres, avec $r << p$.

Utilisation de l'ACP en sondages

- Soient les nouvelles variables

$$\mathbf{Z}_1, \ldots, \mathbf{Z}_r$$

où $\mathbf{Z}_j = (z_{kj})_{k \in U}$, et les $z_{kj}, k \in U$ s'appellent les *scores* de $\mathbf{Z}_j$.

- La matrice d'information auxiliaire est maintenant donnée par

$$\boldsymbol{\mathcal{Z}} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_r) = (\mathbf{z}_k^\top)_{k \in U}$$

avec $\mathbf{z}_k^\top = (z_{k1}, \ldots, z_{kr})$.

- On réduit le problème de p variables de départ $\mathbf{X}_1, \ldots, \mathbf{X}_p$ à r nouvelles variables, $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$ conservant le maximum de variabilité.

Calage sur composantes principales

- On utilise comme variables de calage les r composantes principales,

$$\mathbf{Z}_1, \ldots, \mathbf{Z}_r$$

- **Objectif** (avec la distance de chi-deux) : trouver des poids de calage $\mathbf{w} = (w_k)_{k \in s}$ qui vérifient

$$\begin{aligned} \mathbf{w} &= \text{argmin}_{\mathbf{w}} \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k} \\ \hat{t}_{w,\mathbf{Z}_j} &= t_{\mathbf{Z}_j} \text{ pour } j = 1, \ldots, r. \end{aligned}$$

- Les équations de calage peuvent être écrites sous la forme

$$\sum_{k \in s} w_k \mathbf{z}_k = \sum_{k \in U} \mathbf{z}_k$$

L'estimateur dépend d'un paramètre r à choisir :

- $r = 0$: on obtient l'estimateur d'Horvitz-Thompson $\hat{t}_{yd}$ qui n'utilise pas d'information auxiliaire.
- $r = p$: on obtient l'estimateur calé sur les p variables de départ.
- Il faut trouver le bon compromis.

L'estimateur par calage du total, $\hat{t}_{w,PC} = \sum_{k \in s} w_k y_k$, est un estimateur par la régression généralisée :

$$\hat{t}_{w,PC} = \hat{t}_{yd} - \left(\hat{t}_{\boldsymbol{z}d} - t_{\boldsymbol{z}}\right)^{\top} \hat{\boldsymbol{\beta}}_{\boldsymbol{\mathcal{Z}}}$$

où

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\mathcal{Z}}} = \left(\sum_{k \in s} d_k \mathbf{z}_k \mathbf{z}_k^T\right)^{-1} \sum_{k \in s} d_k \mathbf{z}_k y_k.$$

Lien avec l'approche *model-assisted*

- Soit le modèle linéaire

$$\xi : \quad y_k = \mathbf{x}_k \boldsymbol{\beta} + \varepsilon_k, \quad k \in U$$

- On peut montrer que

$$\hat{t}_{w,PC} = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)' \hat{\boldsymbol{\beta}}_{PC},$$

$$\hat{\boldsymbol{\beta}}_{PC} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)' \hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}.$$

- L'estimateur $\hat{t}_{w,PC}$ est un estimateur de type *model-assisted* linéaire quand le coefficient de régression $\boldsymbol{\beta}$ est estimé par $\hat{\boldsymbol{\beta}}_{PC}$.
- L'estimateur $\hat{t}_{w,PC}$ est biaisé sous le plan $p(\cdot)$ ainsi que sous le modèle ξ, mais il est sans biais sous le plan et le modèle,

$$\text{Bias}_{p,\xi}(\hat{t}_{w,PC} - t_y) = 0.$$

Calage sur le moment d'ordre deux

- Pour des variables $\mathbf{X}_j$ standardisées, les CPs ont la propriété suivante :

$$\frac{1}{N}\mathbf{Z}_j'\mathbf{Z}_j = \frac{1}{N}\sum_{k\in U} z_{kj}^2 = \lambda_j, \quad j = 1, \ldots, p$$

- On peut considérer un calage sur le moment d'ordre deux des CPs (Ren, 2000)

$$\begin{aligned} \mathbf{w}^c &= \text{argmin}_{\mathbf{w}} \sum_s \frac{(w_k - d_k)^2}{d_k q_k} \quad \text{et} \\ \hat{t}_{w,Z_j} &= t_{Z_j}, \quad j = 1, \ldots, r \\ \hat{t}_{w,Z_j^2} &= t_{Z_j^2}, \quad j = 1, \ldots, r \end{aligned}$$

avec $\mathbf{Z}_j^2 = (z_{jk}^2)_{k\in U}$.

Estimation des composantes principales

- On a besoin de connaître $\mathbf{x}_k$ pour tous les $k \in U$ pour pouvoir déterminer les valeurs et vecteurs propres de la matrice de variance-covariance $\frac{1}{N}\mathbf{X}'\mathbf{X} = \frac{1}{N}\sum_U \mathbf{x}_k\mathbf{x}_k'$.
- Il est possible d'étendre cette approche pour le cas où on ne connaît $\mathbf{x}_k$ que sur l'échantillon mais avec des totaux $\sum_{k\in U} \mathbf{x}_k$ connus ;
- On centre les variables dans la population (on connaît leur moyennes) et on les réduit dans l'échantillon ;

- On estime la matrice de variance-covariance par

$$\hat{\mathbf{\Gamma}} = \frac{1}{\hat{N}} \sum_s d_k (\mathbf{x}_k - \hat{\overline{\mathbf{X}}})(\mathbf{x}_k - \hat{\overline{\mathbf{X}}})'$$

- Soient $\hat{\lambda}_j, \hat{\mathbf{v}}_j$ les valeurs, resp. les vecteurs propres de $\hat{\mathbf{\Gamma}}$,

$$\hat{\mathbf{\Gamma}} \hat{\mathbf{v}}_j = \hat{\lambda}_j \hat{\mathbf{v}}_j, \quad j = 1, \ldots, p$$

- On estime les composantes $\mathbf{Z}_j$ par

$$\hat{\mathbf{Z}}_j = \mathbf{X} \hat{\mathbf{v}}_j.$$

- On prend comme variables de calage $\widehat{\mathbf{Z}}_j$, $j = 1, \ldots, r$ de totaux connus.

- L'estimateur par calage sur les composantes principales estimées $\widehat{\mathbf{Z}}_j$, est donné par

$$\begin{aligned}\hat{t}^*_{w,PC} &= \hat{t}_{yd} - \left(\hat{t}_{\hat{\boldsymbol{z}}d} - t_{\hat{\boldsymbol{z}}}\right)' \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{Z}}} \\ &= \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)' \hat{\boldsymbol{\beta}}^*_{PC}\end{aligned}$$

où

$$\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{Z}}} = \left(\sum_{k\in s} d_k \hat{\mathbf{z}}_k \hat{\mathbf{z}}_k^T\right)^{-1} \sum_{k\in s} d_k \hat{\mathbf{z}}_k y_k.$$

$$\hat{\boldsymbol{\beta}}^*_{PC} = (\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_r)\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{Z}}}.$$

Hypothèses

On suppose que :
(A1) $\pi_k > \lambda > 0$ pour tous $k \in U$;
(A2) $\overline{\lim}_{N\to\infty} n \max_{k\neq l} |\pi_{kl} - \pi_k \pi_l| < \infty$.
(A3) $\lim_{N\to\infty} \frac{1}{N} \sum_U y_k^2 < \infty$.
(A4) $\lim_{N\to\infty} \frac{n}{N} = \pi \in (0,1)$.
(A5) $||\mathbf{x}_k|| < \infty$ pour tous $k \in U$.

Alors, $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{z}}} - \tilde{\boldsymbol{\beta}}_{\boldsymbol{z}} = o_p(1)$ avec $\tilde{\boldsymbol{\beta}}_{\boldsymbol{z}} = \left(\sum_{k\in U} \mathbf{z}_k \mathbf{z}_k^T\right)^{-1} \sum_{k\in U} \mathbf{z}_k y_k$ et

$$N^{-1}(\hat{t}^*_{w,PC} - t_y) = N^{-1}(\hat{t}_{diff} - t_y) + o_p(n^{-1/2})$$

où $\hat{t}_{diff} = \hat{t}_{yd} - \left(\hat{t}_{\boldsymbol{z}d} - t_{\boldsymbol{z}}\right)' \tilde{\boldsymbol{\beta}}_{\boldsymbol{z}}$.

Données Médiamétrie (1)

- Population d'étude U composée de $N = 5\ 329$ individus qui regardent une chaîne de TV sur deux semaines consécutives de septembre 2010 ;
- Le paramètre d'intérêt : la durée totale d'écoute sur lundi de la deuxième semaine ;
- La variable $\mathcal{Y}$ contient beaucoup de zéros (plus de 30%).
- On considère un échantillon s de taille $n = 500$ tiré par sondage aléatoire simple sans remise dans U.

Données Médiamétrie : Information auxiliaire

- 19 variables qualitatives de type socio-démographique (région, sexe, CSP, internet . . .) avec 52 modalités en tout ;
- age : en continu ;
- On considère de plus comme variable auxiliaire la durée totale d'écoute pour la première semaine de septembre 2010 :
  - avec cette variable, on a $R^2 = 0.50$,
  - sans cette variable, on a $R^2 = 0.11$.
- Matrice d'information auxiliaire $\mathbf{X}$ de dimension $5\,329 \times 54$;
- La matrice $\mathbf{X}$ contient l'intercept ; on calcule les CP de $\mathbf{X}$ sans l'intercept (et après avoir centré et réduit) et la matrice des CP est

$$[\mathbf{1} \quad \mathbf{Z}_1 \quad \dots \quad \mathbf{Z}_{p-1}].$$

Estimation

- On sélectionne un échantillon s de taille $n = 500$ selon un plan aléatoire simple sans remise dans U.
- On considère $n.sim = 1000$ simulations.
- L'estimateur GREG ne marche pas toujours car la matrice $\mathbf{X}_\mathbf{s}'\mathbf{\Pi}_s\mathbf{X}_\mathbf{s}$ a beaucoup de fois λ_{min} égale à zéro). On utilise l'inverse généralisée.

Comparaison entre $\hat{t}_{ridge}$, $\hat{t}_{PC}$ et $\hat{t}_{HT}$

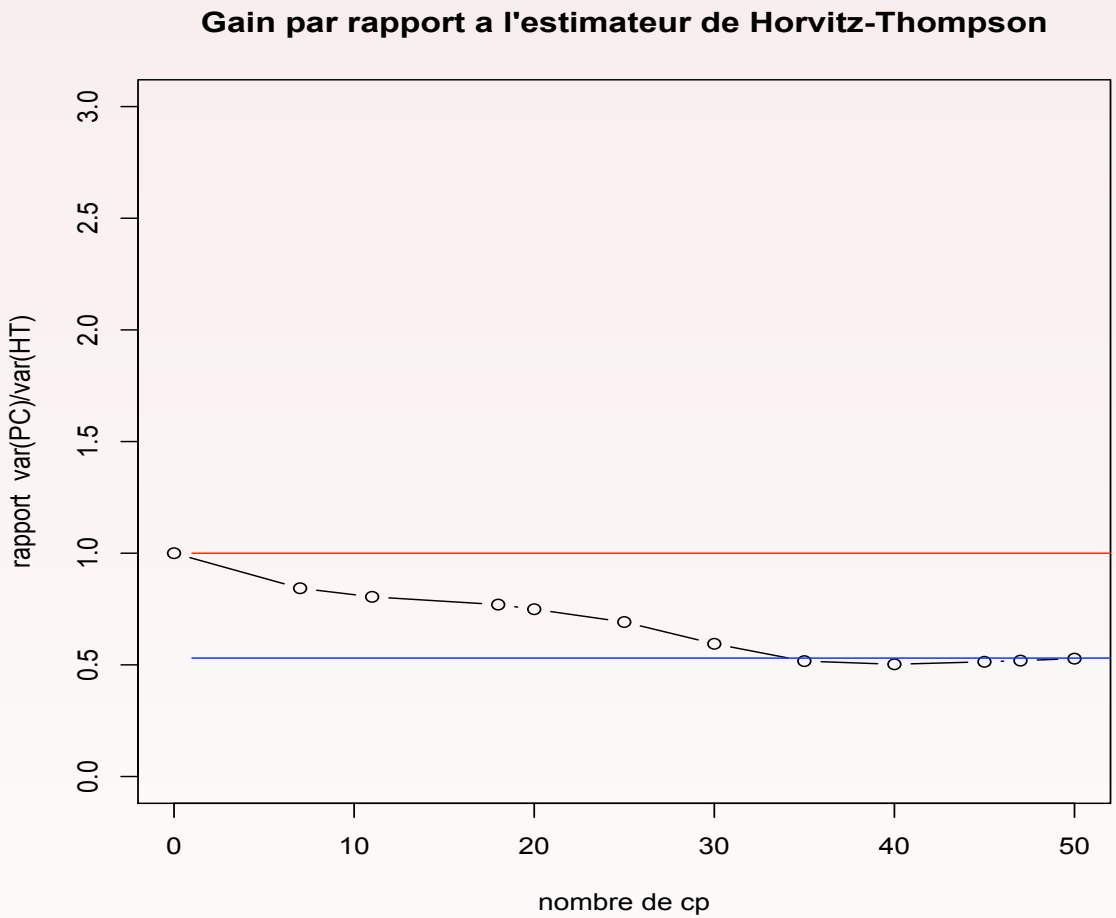
- biais relatif :

$$RB = \frac{\sum_{b=1}^{B} \hat{\theta}^{(j)}/B - t_y}{t_y}$$

  plus petit que 0.2%.

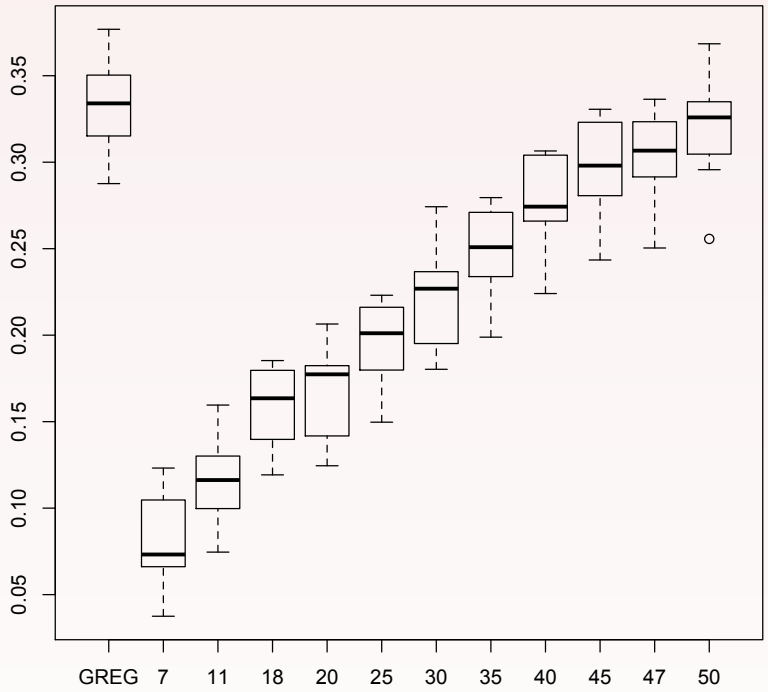
- le rapport de variances (sur les B échantillons), par rapport à l'estimateur de Horvitz-Thompson $\hat{t}_{yd}^{(b)}$,

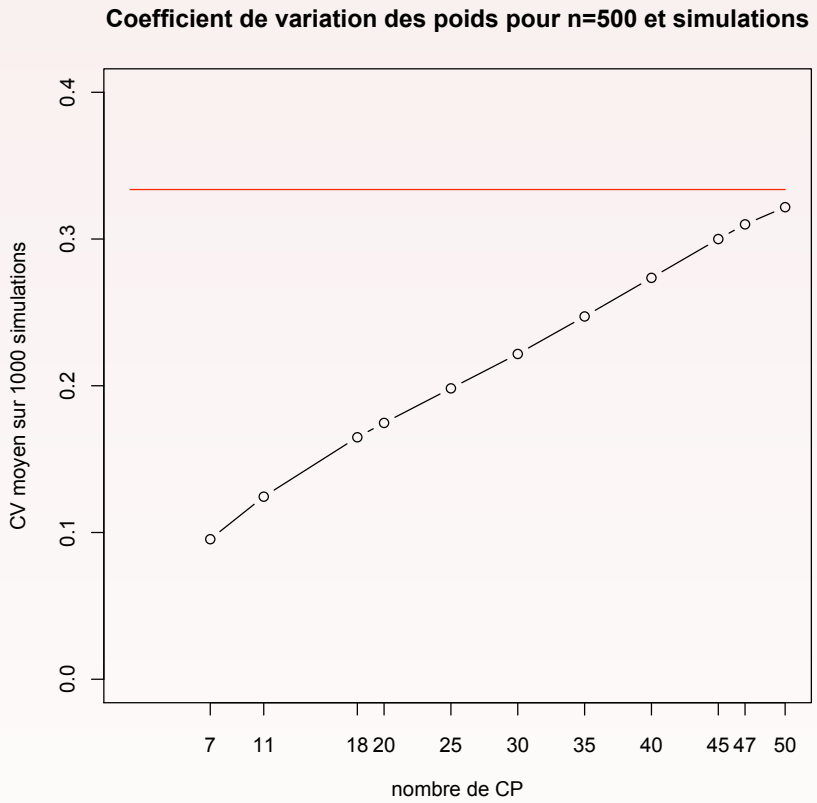
$$\frac{\text{var}_\theta}{\text{var}_{HT}} = \frac{\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - t_y)^2}{\frac{1}{B}\sum_{b=1}^{B}(\hat{t}_{yd}^{(b)} - t_y)^2}.$$



- en bleu : le rapport entre la variance de l'estimateur obtenu par calage exact (calculé avec l'inverse généralisée) et l'estimateur

Coefficient de variation des poids en fonction de r et de n





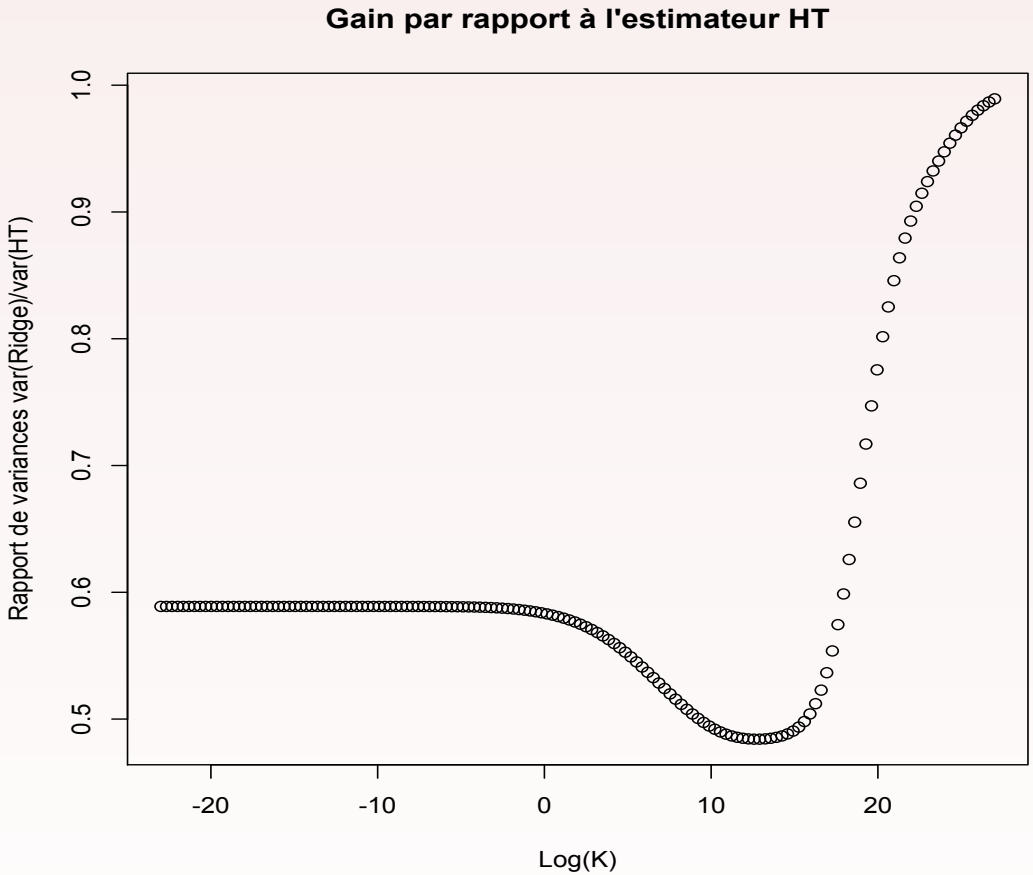
Le coefficient de variation des poids $\text{cv}(\mathbf{w}) = \dfrac{\sqrt{\text{var}(\mathbf{w})}}{\overline{\mathbf{w}}}$ est calculé pour 1000 simulations.

Etude avec calage pénalisé

- On calcule les poids ridge pour 150 valeurs de λ :

$$\lambda = 4 \cdot 2^{-h}, \quad h \in [-25, 25]$$

- On calcule l'estimateur par calage pénalisé ou ridge $\hat{t}_{w,\lambda}$ pour chaque λ et pour $B = 500$ simulations. On compare la variance de $\hat{t}_{w,\lambda}$ lors de 500 simulations avec la variance de $\hat{t}_{yd}$
- On remarque (dans le graphique $\log_2(\lambda) \in [-23, 27]$)
  1. pour λ grand, $\hat{t}_{w,\lambda}$ est aussi efficace que l'estimateur d'Horvitz-Thompson
  2. pour λ très proche de zéro, on a un gain très important.


Gain par rapport à l'estimateur HT
Rapport de variances var(Ridge)/var(HT)
0.5
0.6
0.7
0.8
0.9
1.0
-20
-10
0
10
20
Log(K)

Nous avons répété 10 fois la simulation.

- pour K petit, le gain est important ;
- pour K grand, l'estimator $\hat{t}_{ridge}$ s'approche de $\hat{t}_{HT}$.

Conclusion et perspectives

- On propose une nouvelle méthode pour réduire le grand nombre de variables auxiliaires en sondages ;
- Cette méthode est facilement mise en oeuvre ;
- Elle a l'avantage par rapport à la régression ridge de pouvoir faire un calage sur le moment d'ordre deux ;
- Elle permet d'obtenir de gains importants en termes de variance ;

Courte bibliographie

- Chambers, R. (1996), Robust case-weighting for multi-purpose establishments surveys, Journal of official statistics, vol.12, 1996, 3-32 ;
- Deville, J.-C., 1999. Variance estimation for complex statistics and estimators : linearization and residual techniques. Survey Methodology 25, 193–204.
- Deville, J.-C., Särndal, C.-E., 1992. Calibration estimators in survey sampling. Journal of the American Statistical Association 87, 376–382.
- Rao, J.N.K. and Singh, A.C. (2009), Range restricted weight calibration for survey data using ridge regression, Pakistan Journal of Statistics, 25, 371-384.
- Ren, R. (2000), Utilisation d'information auxiliaire par calage sur fonction de répartition, thèse de l'Université Paris Dauphine.
- Silva, P.L.N. and Skinner, C. (1997). Variable selection for regression estimation in finite population, Survey Methodology, 23, 23-32.
- Tillé, Y. and Guggemos, F., (2010) Penalized calibration in survey sampling : Design-based estimation assisted by mixed models. Journal of Statistical Planning and Inference 140 (2010) 3199–3212.