Estimation assistée par modèle dans un cadre à grande dimension pour des données d'enquête

Camelia Goga
LMB - Univ. de Bourgogne Franche-Comté
camelia.goga@univ-fcomte.fr

Séminaire en ligne sur les sondages
15 décembre 2022

Plan

- Motivation : estimation des totaux en présence d'un grand nombre de variables auxiliaires ;
- L'estimateur assisté par un modèle et l'estimateur calé du total dans ce cadre de grande dimension ;
- Deux classes d'estimateurs améliorés de type "model-assisted" basés sur des méthodes de pénalisation et de réduction de la dimension ;
- Etude par simulation sur des données réelles de consommation d'électricité de foyers et entreprises irlandaises.

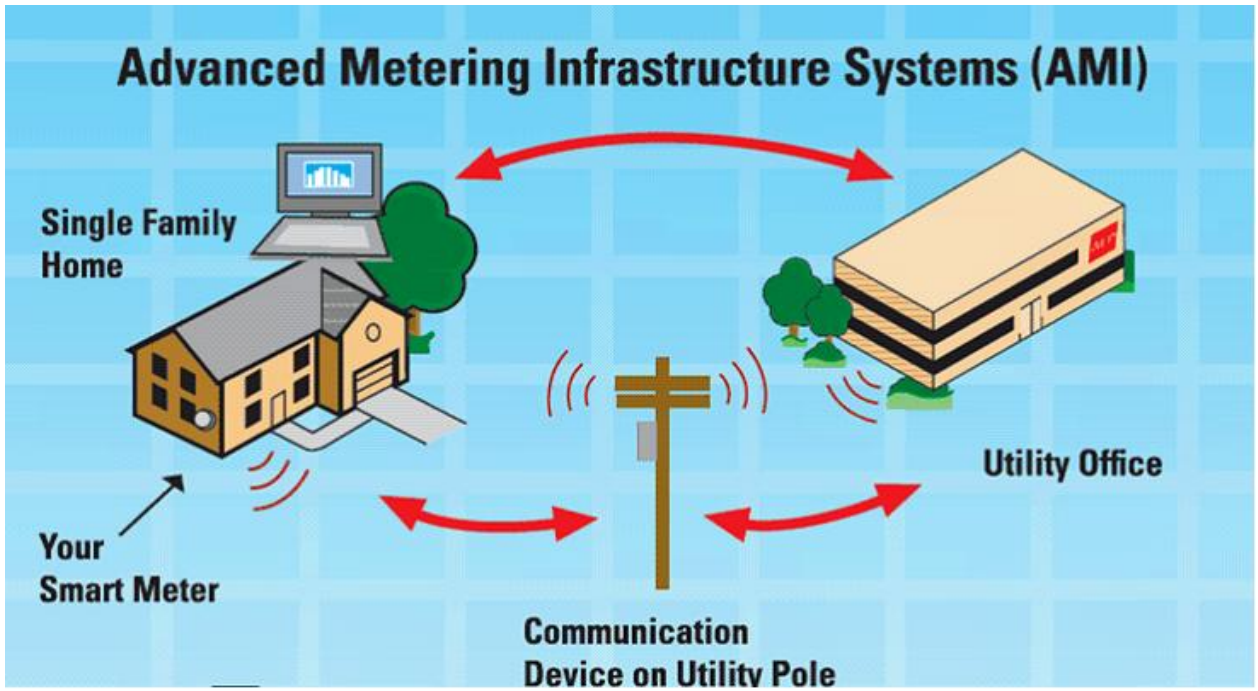
Travail en collaboration avec M. Dagdoug, D. Haziza ; G. Chauvet ; H. Cardot et M.A. Shehzad

Sondages dans de très grandes bases de données

- L'émergence de très grandes bases de données due aux capteurs digitaux (capteurs intelligents, smartphones, ...) qui permettent de collecter mais aussi de transmettre l'information à un pas très fin ;
- Les Offices Nationaux de Statistique ont accès maintenant à des nombreuses sources d'information, avec potentiellement un grand nombre de variables ;
- Les méthodes d'estimation paramétriques et non-paramétriques peuvent être inefficaces.

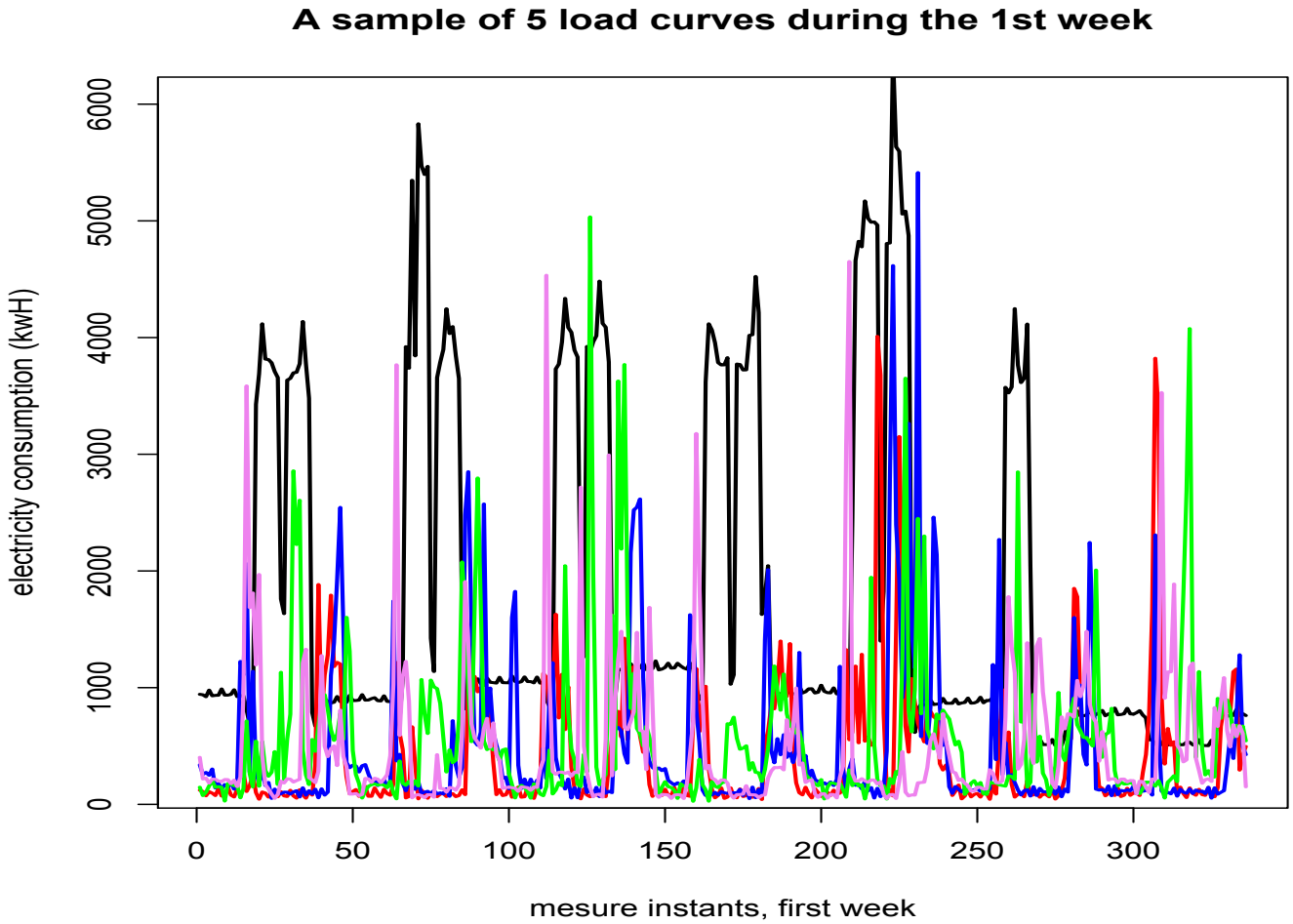
La consommation d'électricité enregistrée via des compteurs intelligents

compteur intelligent : un outil installé dans un foyer ou entreprise pour enregistrer et transmettre la consommation d'électricité à un pas potentiellement très fin (chaque minute, seconde, ...)

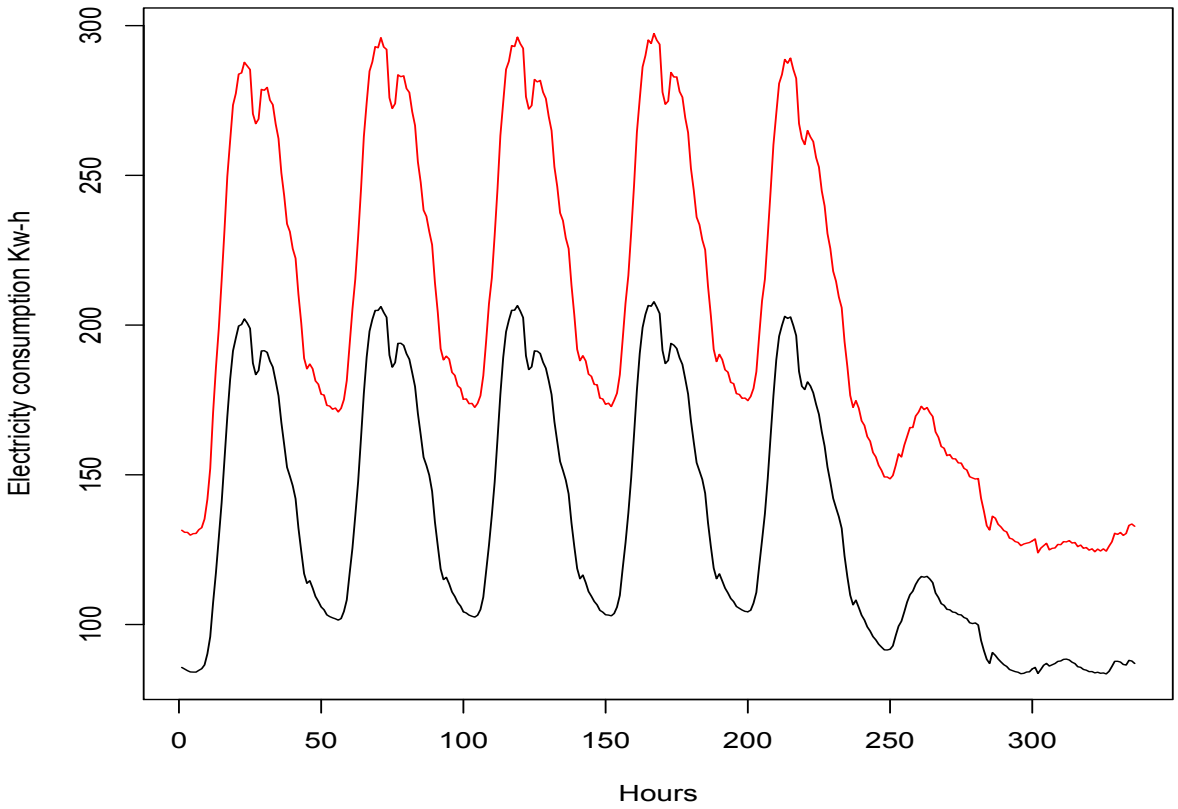


Exemple 1 : un échantillon de 5 courbes d'électricité

Population test : 18902 entreprises et la consommation est enregistrée toutes les 30 min pendant une semaine



La courbe moyenne "vs" la courbe médiane d'électricité



La courbe moyenne dans la population est en rouge et la courbe médiane en noir.

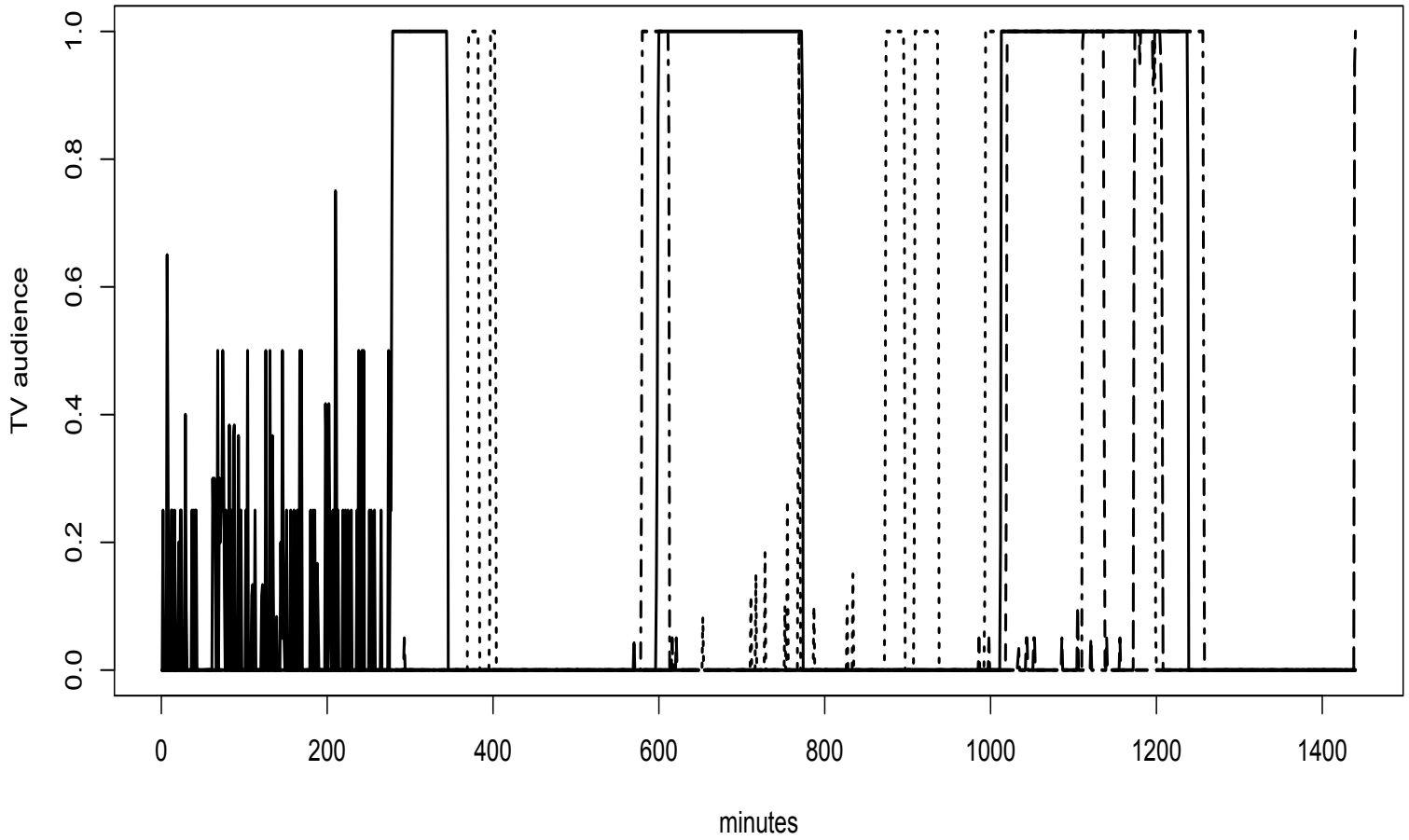
Audience enregistrée via des boitiers intelligents

Médiamétrie est l'entreprise française privée qui s'occupe avec la mesure de l'audience française à la TV, radio, internet ; les dernières années, les mesures d'audience ont été enrichies par

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

Example 2 : un échantillon de 5 courbes d'audience TV

L'audience TV est enregistrée chaque minute pendant 24h.



Population, échantillon

- Soit $U = \{1, \ldots, k, \ldots, N\}$ une population finie de taille N ;
- Soit $s \subset U$ un échantillon sélectionné dans U selon un plan de sondage $p(s)$;
- Les probabilités d'inclusion :

$$\pi_k = Pr(k \in s) = \sum_{k \in s} p(s) \quad \text{et} \quad \pi_{kl} = Pr(k, l \in s) = \sum_{k,l \in s} p(s);$$

- Soit $\mathcal{Y}$ une variable d'intérêt et l'objectif est d'estimer son total dans la population U :

$$t_y = \sum_{k \in U} y_k$$

L'estimateur d'Horvitz-Thomson du total t_y et sa variance

- En présence de données complètes, le total t_y est estimé par l'estimateur d'Horvitz-Thompson (HT) :

$$\hat{t}_{yHT} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

- Si $\pi_k > 0$ pour tous $k \in U$, alors l'estimateur HT est sans biais pour t_y :

$$\mathbb{E}_p(\hat{t}_{yHT}) = t_y,$$

où $\mathbb{E}_p(\cdot)$ est considéré par rapport au plan de sondage $p(\cdot)$;

- La variance de $\hat{t}_{HT}$ est égale à

$$\mathbb{V}_p(\hat{t}_{yHT}) = \sum_{k \in U} \sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell}$$

et si $\pi_{k\ell} > 0$ pour tous les $k, \ell \in U$, cette variance est estimée sans biais par :

$$\hat{\mathbb{V}}_p(\hat{t}_{yHT}) = \sum_{k \in s} \sum_{\ell \in s} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell}$$

Exemple : le plan aléatoire simple sans remise

Considérons un plan aléatoire simple sans remise de taille n dans U; alors, l'estimateur HT est égal à :

$$\hat{t}_{yHT} = \frac{N}{n} \sum_{k \in s} y_k$$

avec la variance

$$\mathbb{V}(\hat{t}_{yHT}) = N^2 \frac{1-f}{n} S_{yU}^2, \quad S_{yU}^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \overline{y}_U)^2$$

estimée sans biais par :

$$\hat{\mathbb{V}}(\hat{t}_{yHT}) = N^2 \frac{1-f}{n} S_{ys}^2, \quad S_{ys}^2 = \frac{1}{n-1} \sum_{k \in s} (y_k - \overline{y}_s)^2$$

Information auxiliaire

- Considérons les variables auxiliaires $X_1, \ldots, X_p$; soit $\mathbf{X}$ la matrice d'information auxiliaire :

$$\mathbf{X} = (\mathbf{X}_1 | \ldots | \mathbf{X}_p) = (\mathbf{x}_k^\top)_{k=1}^p$$

où $\mathbf{x}_k^\top = (x_{kj})_{j=1}^p$, $k \in U$;

  - la consommation d'électricité enregistrée à chaque instant de la semaine précédente ;
  - les audiences enregistrées dans le passé ;

- Dans un cadre de sondage, on peut connaitre $\mathbf{x}_k$ pour chaque $k \in U$ (information auxiliaire complète) ou uniquement pour $k \in s$ avec $\sum_{k \in U} \mathbf{x}_k$ connu ;
- L'estimateur d'Horvitz-Thompson peut être amélioré :
  - au niveau du plan d'échantillonnage en sélectionnant les individus avec des π_k qui incorporent cette information auxiliaire comme par exemple le plan stratifié ou proportionnel à la taille ;
  - **au niveau de l'estimation en considérant un estimateur qui utilise cette information auxiliaire**.

Approche basée sur un modèle : l'estimateur assisté par un modèle ou "model-assisted"

- On considère y_k comme des variables aléatoires, $k \in U$;
- On suppose le modèle de super-population :

$$\xi : \quad y_k = m(\mathbf{x}_k) + \epsilon_k,$$

m est une fonction inconnue mais lisse, $\{\epsilon_k\}_{k\in U}$ sont des variables indépendantes et de moyennes 0;

- L'estimateur de t_y assisté par un modèle :

$$\begin{aligned} \widehat{t}_{ma} &= \sum_{k\in U} \widehat{m}(\mathbf{x}_k) + \sum_{k\in s} \frac{y_k - \widehat{m}(\mathbf{x}_k)}{\pi_k} \\ &= \sum_{k\in s} \frac{y_k}{\pi_k} - \left(\sum_{k\in s} \frac{\widehat{m}(\mathbf{x}_k)}{\pi_k} - \sum_{k\in U} \widehat{m}(\mathbf{x}_k) \right) \end{aligned}$$

où $\widehat{m}$ est un estimateur de m basé sur le plan de sondage.

L'estimateur basé sur un modèle linéaire (Sarndal *et al.*, '92)

On suppose le modèle linéaire, $m(\mathbf{x}_k) = \mathbf{x}_k^\top \boldsymbol{\beta}$ avec $\boldsymbol{\beta} = (\beta_j)_{j=1}^p$:

$$\begin{aligned} y_k &= \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \qquad k = 1, \ldots, N \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \end{aligned}$$

On estime $\boldsymbol{\beta}$ en deux étapes :

1. **sous le modèle** par moindres carrés :

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{OLS} &= \text{argmin}_{\boldsymbol{\beta}} \sum_{k \in U} (y_k - \mathbf{x}_k^\top \boldsymbol{\beta})^2 \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = (\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k^\top)^{-1} \sum_{k \in U} \mathbf{x}_k y_k \end{aligned}$$

2. **sous le plan de sondage** :

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\pi &= \text{argmin}_{\boldsymbol{\beta}} \sum_{k \in s} \frac{1}{\pi_k} (y_k - \mathbf{x}_k^\top \boldsymbol{\beta})^2 \\ &= (\mathbf{X}_s^\top \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^\top \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s = (\sum_{k \in s} \pi_k^{-1} \mathbf{x}_k \mathbf{x}_k^\top)^{-1} \sum_{k \in s} \pi_k^{-1} \mathbf{x}_k y_k \end{aligned}$$

- La fonction de régression $m(\mathbf{x}_k)$ est estimée par $\widehat{m}(\mathbf{x}_k) = \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_\pi$; le total t_y est estimé par l'estimateur model-assisted ou GREG (generalized regression) :

$$\begin{aligned}\hat{t}_{GREG} &= \hat{t}_{yHT} - \left(\sum_{k\in s}\frac{\mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_\pi}{\pi_k} - \sum_{k\in U}\mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_\pi\right)\\ &= \hat{t}_{yHT} - \left(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}\right)^\top \hat{\boldsymbol{\beta}}_\pi = \sum_{k\in s} w_{ks} y_k\end{aligned}$$

avec $w_{ks} = \pi_k^{-1} - \pi_k^{-1}\mathbf{x}_k^\top\left(\sum_{k\in s}\pi_k^{-1}\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\left(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}\right)$.

- On peut écrire l'erreur d'échantillonnage somme suit :

$$\frac{1}{N}\left(\hat{t}_{GREG} - t_y\right) = \frac{1}{N}(\hat{t}_{ydiff} - t_y) - \frac{1}{N}\left(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}\right)^\top\left(\hat{\boldsymbol{\beta}}_\pi - \tilde{\boldsymbol{\beta}}_{OLS}\right)$$

où

$$\begin{aligned}\hat{t}_{ydiff} &= \hat{t}_{yHT} - \left(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}\right)^\top \tilde{\boldsymbol{\beta}}_{OLS}\\ &= \sum_{k\in U}\mathbf{x}_k^\top \tilde{\boldsymbol{\beta}}_{OLS} + \sum_{k\in s}\frac{y_k - \mathbf{x}_k^\top \tilde{\boldsymbol{\beta}}_{OLS}}{\pi_k}\end{aligned}$$

L'approche calage (Deville & Sarndal, 1992)

- Construire un estimateur pondéré de t_y :

$$\hat{t}_w = \sum_{k \in s} w_{ks} y_k$$

avec des poids $w_{ks}, k \in s$ qui soient le plus proches possible des poids de sondage $1/\pi_k$ et qui satisfont *les contraintes de calage :*

$$\sum_{k \in s} w_{ks} \mathbf{x}_k \quad = \quad \sum_{k \in U} \mathbf{x}_k$$

- Plusieurs fonctions distance ont été considérées pour mesurer la proximité entre w_{ks} et $1/\pi_k$; la distance de "chi-squared" :

$$\Psi(\mathbf{w}) = \sum_{k \in s} \frac{(w_{ks} - \pi_k^{-1})^2}{\pi_k^{-1}}$$

conduit à $w_{ks} = \pi_k^{-1} - \pi_k^{-1} \mathbf{x}_k^\top (\sum_{k \in s} \pi_k^{-1} \mathbf{x}_k \mathbf{x}_k^\top)^{-1} \left(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}\right)$ et l'estimateur par calage et l'estimateur model-assisted sont les mêmes ;

Cadre asymptotique : $n, N \to \infty$ et p fixé

On suppose le cadre asymptotic introduit par Isaki and Fuller (1982) dans lequel les tailles de la population N et de l'échantillon n tendent vers l'infini ;

L'objectif est d'obtenir la convergence et la variance asymptotique des estimateurs de type model-assisted pour $n, N \to \infty$ et p considéré fixé d'abord et ensuite, pour $p \to \infty$

Nous avons besoin de conditions de régularité supplémentaires sur :

- les probabilités d'inclusion π_k, π_{kl} :

  pour tous $k \in U$: $\pi_k \geq c_1 > 0, \lim_{n\to\infty} n \max_{k\neq l\in U} |\pi_{kl} - \pi_k\pi_l| < \infty$

- la variable y : $N^{-1}\sum_{k\in U} y_k^2 < C_0$ avec $C_0 > 0$;
- l'information auxiliaire : $||\mathbf{x}_k||^2 \leq C$ pour tous $k \in U$, avec $C > 0$;

L'efficacité asymptotique : $n, N \to \infty$ et p fixé

L'erreur de l'estimateur GREG ou model-assisted :

$$N^{-1}\left(\hat{t}_{GREG} - t_y\right) \quad = \quad N^{-1}(\hat{t}_{diff} - t_y) - N^{-1}\left(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}\right)^\top \left(\hat{\boldsymbol{\beta}}_\pi - \tilde{\boldsymbol{\beta}}_{OLS}\right)$$

Résultat

Sous les hypothèses de régularité :

- $N^{-1}(\hat{t}_{diff} - t_y) = O_p(n^{-1/2})$, $N^{-1}(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}) = O_p(n^{-1/2})$ *et*

$$\hat{\boldsymbol{\beta}}_\pi - \tilde{\boldsymbol{\beta}}_{OLS} = O_p(n^{-1/2})$$

- *L'estimateur GREG est asymptotiquement équivalent avec l'estimateur par la différence généralisée :*

$$N^{-1}\left(\hat{t}_{GREG} - t_y\right) = N^{-1}(\hat{t}_{diff} - t_y) + O_p(n^{-1})$$

La variance asymptotique de $\hat{t}_{GREG}$ est la variance de $\hat{t}_{diff}$:

$$AV_p(\hat{t}_{GREG}) = \sum_{k\in U}\sum_{\ell\in U}(\pi_{k\ell} - \pi_k\pi_\ell)\frac{y_k - \mathbf{x}_k^\top\tilde{\boldsymbol{\beta}}_{OLS}}{\pi_k}\frac{y_\ell - \mathbf{x}_\ell^\top\tilde{\boldsymbol{\beta}}_{OLS}}{\pi_\ell}.$$

Estimation en présence d'un grand nombre p de variables auxiliaires

On considère qu'un grand nombre p de variables auxiliaires est disponible.

Question : doit-on considérer toutes ces variables ?

Dans un cadre statistique classique, cette situation avait été relevée dans les années 70's dans le cadre de l'estimation du coefficient de régression β d'un modèle linéaire.

Il avait été remarqué que :

- pour p grand, des problèmes de multi-collinéarité entre les variables X_j peuvent apparaitre ; l'information contenue dans $\mathbf{X}$ est redondante ;
- l'estimateur OLS $\tilde{\boldsymbol{\beta}}_{OLS}$ est sans biais mais avec une variance très grande ;
- $\tilde{\boldsymbol{\beta}}_{OLS}$ est en moyenne très loin de β.

Dans le cadre de la théorie des sondages

Bardsley and Chambers (1984) avaient remarqué que l'estimateur basé sur un modèle peut être inefficace si un nombre très grand de prédicteurs est utilisé ; Rao and Singh (1992) avaient remarqué la même chose pour l'estimateur par calage ;

1. les poids w_{ks} utilisés dans l'estimateur MB ou par calage sont très instables ;
2. les poids w_{ks} ne satisfont plus les contraintes imposées :

$$\mathcal{L} \leq \frac{w_{ks}}{\pi_k^{-1}} \leq \mathcal{U},$$

3. Silva and Skinner (1997) ont remarqué sur des simulations que la variance de l'estimateur par calage augmentait quand le nombre de variables auxiliaires était trop grand par rapport à la taille de l'échantillon ;

Petite application sur des données d'électricité irlandaise

- Commission for Energy Regulation (Ireland)
  `http://www.cer.ie/`
- On considère une période de 14 jours consécutives et une population de taille $N = 6291$ individus (ménages et entreprises) ;
- La consommation d'électricité est enregistrée toutes les 30 min. ; on a donc, pour chaque individu la population, $2 \times 7 \times 48 = 672$ instants de mesure ;
- On veut estimer la consommation totale d'électricité de Lundi de la deuxième semaine ;

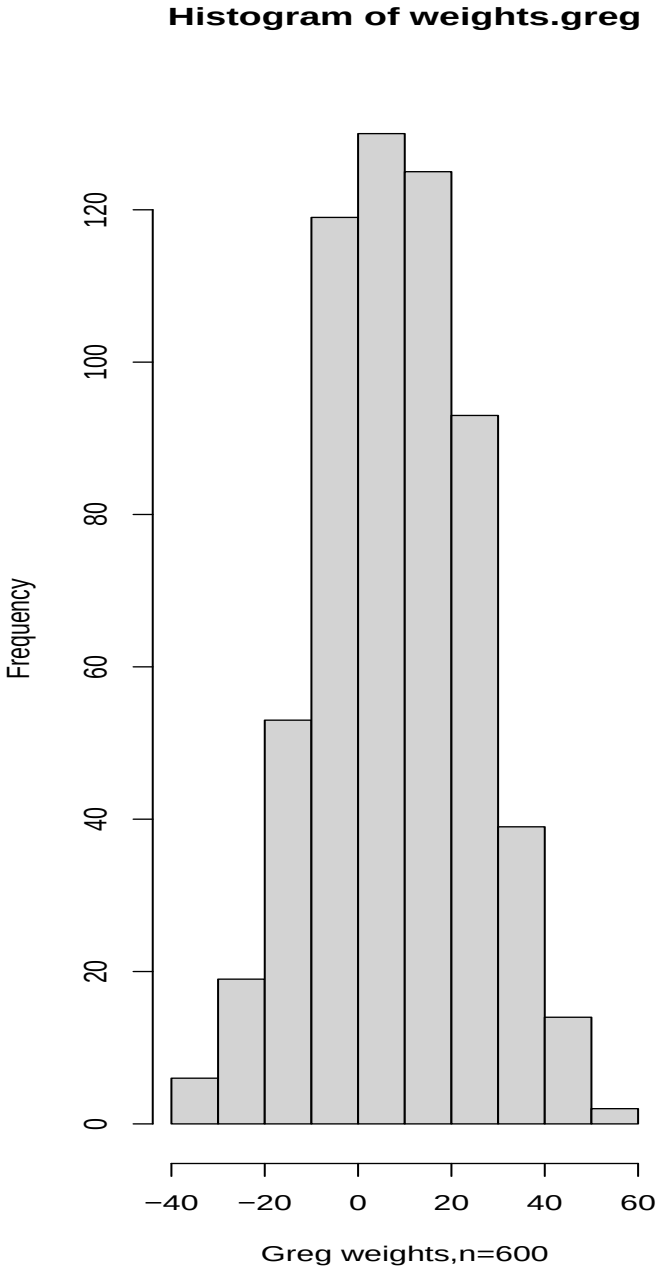
$$t_y = \sum_{k \in U} y_k,$$

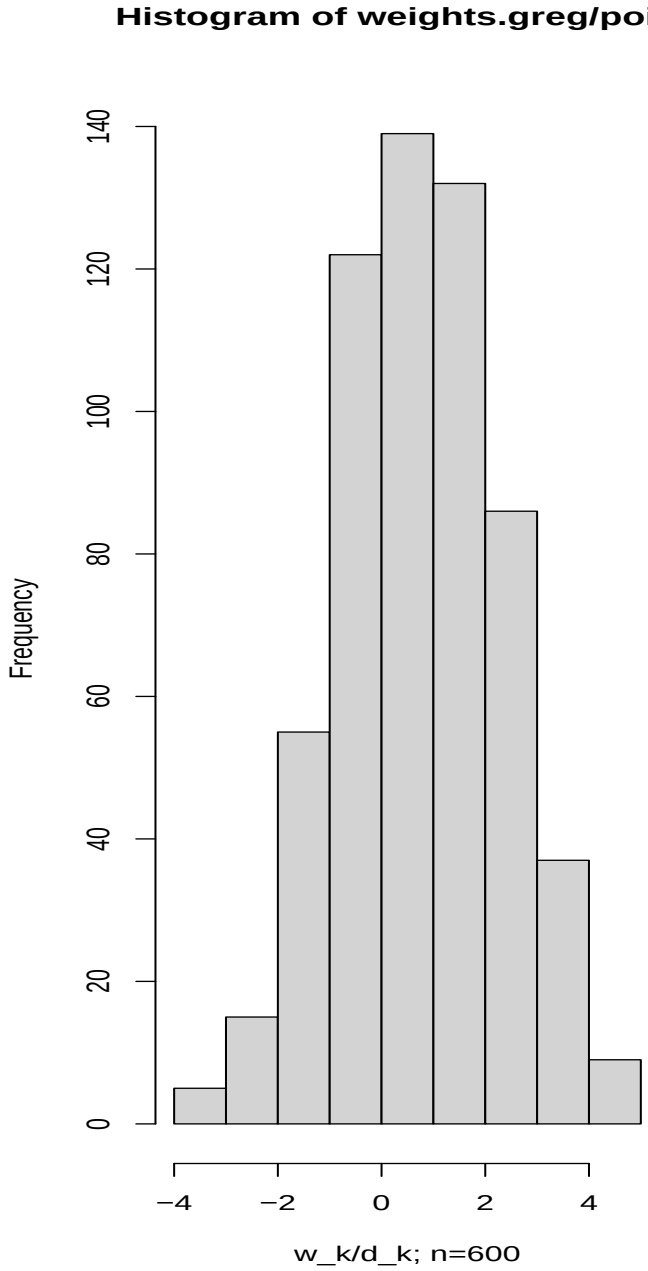
où y_k est la consommation enregistrée Lundi par le compteur k;

- L'information auxiliaire est la consommation de la semaine précédente enregistrée à chaque instant, nous avons donc $p = 336$ variables :

$$X_k(t_j), j = 1 \ldots, 336, \quad k \in U.$$

- On considère un plan SRS de taille $n = 600$ et on calcule les poids GREG


Histogram of weights.greg
Frequency
0
20
40
60
80
100
120
−40
−20
0
20
40
60
Greg weights,n=600


Histogram of weights.greg/poids
Frequency
0
20
40
60
80
100
120
140
−4
−2
0
2
4
w_k/d_k; n=600

Efficacité asymptotique : $n, p \to \infty$ (Chauvet & Goga, JSPI 2022)

On suppose des hypothèses supplémentaires sur $\mathbf{X}$; on suppose aussi que $||\mathbf{x}_k||^2 < p\tilde{C}$ pour tous $k \in U$.

Résultat

Sous les hypothèses de régularité :

- $N^{-1}(\hat{t}_{diff} - t_y) = O_p(n^{-1/2})$, $N^{-1}(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}) = O_{\mathrm{p}}(\sqrt{p/n})$ *et*

$$\hat{\boldsymbol{\beta}}_{\pi} - \tilde{\boldsymbol{\beta}}_{OLS} = O_{\mathrm{p}}\left(\sqrt{\frac{p}{n}}\right) + O_{\mathrm{p}}\left(\frac{p\sqrt{p}}{n}\right);$$

- $$\frac{1}{N}\left(\hat{t}_{GREG} - t_y\right) = \frac{1}{N}(\hat{t}_{diff} - t_y) + O_{\mathrm{p}}\left(\frac{p}{n}\right) + O_{\mathrm{p}}\left(\frac{p^2}{n\sqrt{n}}\right).$$

  Si $p^2/n \to 0$, *alors*

$$\frac{1}{N}\left(\hat{t}_{GREG} - t_y\right) = \frac{1}{N}(\hat{t}_{diff} - t_y) + o_{\mathrm{p}}\left(\frac{1}{\sqrt{n}}\right).$$

Méthodes pour améliorer l'estimateur model-assisted dans un cadre de grande dimension

Solutions :

1. choisir les variables les plus pertinentes en utilisant des critères de choix de variables, néanmoins si p est grand, ces méthodes sont très gourmande en temps de calcul ;
2. utiliser une inverse généralisée si $\mathbf{X}^\top \mathbf{X}$ est non-inversible ;
3. utiliser des méthodes d'estimation biaisée de $\boldsymbol{\beta}$:
   - méthodes de pénalization "ridge" (Bardsley and Chambers, 1984 ; Rao and Singh, 1992 ; Beaumont and Bocci, 2008 ; Guggemos and Tillé, 2010) or lasso
   - méthodes basées sur la réduction de la dimension comme "principal component regression" (Cardot et al., 2017).

L'estimateur ridge du coefficient de régression

- Hoerl and Kennard (1970) ont proposé un critère de moindres carrés avec une pénalité de type L^2 pour estimer $\boldsymbol{\beta}$:

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_\lambda &= \text{argmin}_{\boldsymbol{\beta}} \sum_{k\in U}(y_k - \mathbf{x}_k^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}\end{aligned}$$

  où $\mathbf{I}_p$ est la matrice identité de taille p. On pénalise les grandes valeurs de $\boldsymbol{\beta}$.
  - $\lambda = 0 : \tilde{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\beta}}_{OLS}$;
  - $\lambda \to \infty : \tilde{\boldsymbol{\beta}}_\lambda \to 0$
- L'estimateur de type ridge du $\boldsymbol{\beta}$ sous le plan de sondage est donné par :

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\lambda,\pi} &= \text{argmin}_{\boldsymbol{\beta}} \sum_{k\in s}\frac{1}{\pi_k}(y_k - \mathbf{x}_k^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\ &= (\mathbf{X}_s^\top \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s + \lambda \mathbf{I}_p)^{-1} \mathbf{X}_s^\top \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s\end{aligned}$$

L'estimateur de type ridge model-assisted

- Le total t_y est estimé par un estimateur model-assisted ou GREG de type ridge :

$$\begin{aligned}\hat{t}_{GREG,\lambda}^{\text{pen}} &= \sum_{k\in s}\frac{y_k}{\pi_k} - \left(\sum_{k\in s}\frac{\mathbf{x}_k}{\pi_k} - \sum_{k\in U}\mathbf{x}_k\right)^{\top}\hat{\boldsymbol{\beta}}_{\lambda,\pi} \\ &= \sum_{k\in s} w_{ks}^{\text{pen}}(\lambda) y_k\end{aligned}$$

où $w_{ks}^{\text{pen}}(\lambda) = \pi_k^{-1} - \pi_k^{-1}\mathbf{x}_k^{\top}(\mathbf{X}_s^{\top}\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s + \lambda\mathbf{I}_p)^{-1}\left(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}\right)$

  - $\lambda = 0 : \hat{t}_{GREG,0}^{\text{pen}} = \hat{t}_{GREG}$
  - $\lambda \to \infty : \hat{t}_{GREG,\lambda}^{\text{pen}} \to \hat{t}_{yHT}$

Point de vue calage : le calage pénalisé

On cherche des poids de calage $\mathbf{w}_s^{\text{pen}}(\lambda) = (w_{ks}^{\text{pen}}(\lambda))_{k\in s}$ tels qu'ils minimisent la distance de chi-deux pénalisée :

$$\mathbf{w}_s^{\text{pen}}(\lambda) = \text{argmin}_{\mathbf{w}} \sum_{k\in s} \frac{(w_{ks} - \pi_k^{-1})^2}{\pi_k^{-1}} + \frac{1}{\lambda} \left(\sum_{k\in s} w_{ks}\mathbf{x}_k - \sum_{k\in U} \mathbf{x}_k \right)^\top \left(\sum_{k\in s} w_{ks}\mathbf{x}_k - \sum_{k\in U} \mathbf{x}_k \right)$$

Différente interprétation : on relâche les contraintes de calage, on ne demande plus qu'elle soient exactement satisfaites :

$$|| \sum_{k\in s} w_{ks}\mathbf{x}_k - \sum_{k\in U} \mathbf{x}_k ||^2 \leq c^2$$

- $\lambda \to 0$ les contraintes sont satisfaites et on obtient l'estimateur par calage usuel ;
- $\lambda \to \infty$ aucune contrainte est satisfaite, on obtient l'estimateur d'Horvitz-Thompson estimator ;

Efficacité asymptotique de $\hat{t}^{\text{pen}}_{GREG,\lambda}$ (Dagdoug *et al.*, JAS, 2022)

On suppose des conditions supplémentaires sur $\mathbf{X}$; on suppose aussi que $||\mathbf{x}_k||^2 < p\tilde{C}$ pour tous les $k \in U$.

Résultat

Sous les conditions de régularité :

- $N^{-1}(\hat{t}^{\text{pen}}_{diff,\lambda} - t_y) = O_p(n^{-1/2})$, $N^{-1}(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}) = O_{\text{p}}(\sqrt{p/n})$ *and*

$$\hat{\boldsymbol{\beta}}_{\lambda,\pi} - \tilde{\boldsymbol{\beta}}_{\lambda} = O_{\text{p}}\left(\sqrt{\frac{p}{n}}\right);$$

- $$\frac{1}{N}\left(\hat{t}^{\text{pen}}_{GREG,\lambda} - t_y\right) = \frac{1}{N}(\hat{t}^{\text{pen}}_{diff,\lambda} - t_y) + O_{\text{p}}\left(\frac{p}{n}\right).$$

  Si $p^2/n \to 0$, *alors*

$$\frac{1}{N}\left(\hat{t}^{\text{pen}}_{GREG,\lambda} - t_y\right) = \frac{1}{N}(\hat{t}^{\text{pen}}_{diff,\lambda} - t_y) + o_{\text{p}}\left(\frac{1}{\sqrt{n}}\right).$$

Principal component regression (Jolliffe, 2002)

- Considérons de nouveau le modèle de superpopulation :

$$\begin{aligned} \xi: \quad y_k &= \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k, \quad k \in U \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \end{aligned}$$

- Soit $\mathbf{G} = (\mathbf{v}_1 | \ldots | \mathbf{v}_p)$ avec $\mathbf{v}_j$ les vecteurs propres de $\mathbb{V}(\mathbf{X}) = N^{-1}\mathbf{X}^T\mathbf{X}$ ($\mathbf{X}$ centrée) et

$$\mathbf{G}\mathbf{G}^\top = \mathbf{G}^\top\mathbf{G} = \mathbf{I}_p$$

- Alors, on peut re-parametriser :

$$\mathbf{X}\boldsymbol{\beta} = \underbrace{\mathbf{X}\mathbf{G}}_{\mathbf{Z}}\underbrace{\mathbf{G}^\top\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \mathbf{Z}\boldsymbol{\gamma}$$

- Le modèle peut être écrit dans la forme suivante :

$$\begin{aligned} \xi: \quad y_k &= \mathbf{z}_k^T \boldsymbol{\gamma} + \varepsilon_k, \quad k \in U \\ \mathbf{y} &= \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \end{aligned}$$

où $\mathbf{z}_k^T = \mathbf{x}_k^T\mathbf{G}$ est $k$ème ligne de $\mathbf{Z} = (\mathbf{Z}_1 | \ldots | \mathbf{Z}_p)$.

Le modèle réduit

- Les variables $\mathbf{Z}_j = \mathbf{X}\mathbf{v}_j$ sont les **principal components** de $\mathbf{X}$:
  - Ces variables sont **non-corelées** et
  - $N^{-1}\mathbf{Z}^T\mathbf{Z} = \text{diag}(\lambda_j)_{j=1}^p$ avec $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ les valeurs propres de $V(\mathbf{X}) = N^{-1}\mathbf{X}^T\mathbf{X}$.
- L'idée est de considérer un modèle réduit pour y_k avec les prédictors $\mathbf{Z}_{(r)} = (\mathbf{Z}_1|\ldots|\mathbf{Z}_r)$ correspondant aux r plus grandes valeurs propres :

$$\xi_r : \quad \mathbf{y} \quad = \quad \mathbf{Z}_{(r)}\boldsymbol{\gamma}_r + \boldsymbol{\varepsilon}_r$$

- L'estimateur OLS de $\boldsymbol{\gamma}_r$ est

$$\tilde{\boldsymbol{\gamma}}_{\mathbf{z},r} = \left(\mathbf{Z}_{(r)}^\top \mathbf{Z}_{(r)}\right)^{-1} \mathbf{Z}_{(r)}\mathbf{y}$$

et l'estimateur PC de $\boldsymbol{\beta}$ est donné par

$$\tilde{\boldsymbol{\beta}}_{\mathbf{x},r}^{\text{pc}} = (\mathbf{v}_1|\ldots,|\mathbf{v}_r)\tilde{\boldsymbol{\gamma}}_{\mathbf{z},r} = \mathbf{G}_r\tilde{\boldsymbol{\gamma}}_{\mathbf{z},r}$$

- $\tilde{\boldsymbol{\beta}}_{\mathbf{x},r}^{\text{pc}}$ est la part de $\hat{\boldsymbol{\beta}}_{OLS}$ qui appartient à l'espace de dimension r avec la plus grande the largest variance.

L'estimateur PC model-assisted (Cardot *et al.*, Stat. Sinica, 2017)

- On peut utiliser l'estimateur PC de $\boldsymbol{\beta}$ pour construire un nouvel estimateur model-assisted de t_y;
- On estime d'abord $\tilde{\boldsymbol{\beta}}^{\mathrm{pc}}_{\mathbf{x},r}$ au niveau de l'échantillon :

$$\hat{\boldsymbol{\beta}}^{\mathrm{pc}}_{\mathbf{x},r} = \mathbf{G}_r \hat{\boldsymbol{\gamma}}_{\mathbf{z},r}$$

avec

$$\hat{\boldsymbol{\gamma}}_{\mathbf{z},r} = \left(\mathbf{Z}^{\top}_{s,(r)} \boldsymbol{\Pi}_s^{-1} \mathbf{Z}_{s,(r)}\right)^{-1} \mathbf{Z}_{s,(r)} \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s$$

- Le total t_y est estimé par un estimateur GREG :

$$\begin{aligned}
\hat{t}^{\mathrm{pc}}_{GREG,r} &= \hat{t}_{yHT} - \left(\hat{t}_{\mathbf{z}_r HT} - t_{\mathbf{z}_r}\right)^T \hat{\boldsymbol{\gamma}}_{\mathbf{z},r} \\
&= \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}}\right)^T \hat{\boldsymbol{\beta}}^{\mathrm{pc}}_{\mathbf{x},r} \\
&= \sum_{k \in s} w^{\mathrm{pc}}_{ks}(r) y_k
\end{aligned}$$

- Sans information auxiliaire complète, on doit estimer d'abord $\hat{\mathbf{Z}}_j = \mathbf{X}\hat{\mathbf{v}}_j$;

Point de vue calage

- Les poids PC $w_{ks}^{\mathrm{pc}}(r), k \in s$ peuvent être obtenus par un calage sur les totaux de r premières composantes principales :

$$\sum_{k \in s} w_{ks}^{\mathrm{pc}}(r) \mathbf{z}_{kr} = \sum_{k \in U} \mathbf{z}_{kr}$$

1. $r = 0$: on obtient l'estimateur d'Horvitz-Thompson $\hat{t}_{yd}$;
2. $r = p$: on obtient l'estimateur GREG ;
3. calage partiel : on estime exactement les totaux de p_1 variables et on pénalise les autres $p - p_1$ variables (Bardsley and Chambers, 1984 ; Guggemos and Tillé, 2010).

Efficacité asymptotique de l'estimateur PC model-assisted

On suppose que $r \to \infty$;

Résultat

Sous les conditions de régularité :

- $N^{-1}(\hat{t}^{\mathrm{pc}}_{diff,r} - t_y) = O_p(n^{-1/2})$, $N^{-1}(\hat{t}_{\mathbf{z_r}HT} - t_{\mathbf{z_r}}) = O_{\mathrm{p}}(\sqrt{r/n})$ *et*

$$\hat{\boldsymbol{\gamma}}_{\mathbf{z},r} - \tilde{\boldsymbol{\gamma}}_{\mathbf{z},r} = O_{\mathrm{p}}\left(\sqrt{\frac{r}{n}}\right) + O_{\mathrm{p}}\left(\frac{r\sqrt{r}}{n}\right);$$

- $$\frac{1}{N}\left(\hat{t}^{\mathrm{pc}}_{GREG,r} - t_y\right) = \frac{1}{N}(\hat{t}^{\mathrm{pc}}_{diff,r} - t_y) + O_{\mathrm{p}}\left(\frac{r}{n}\right) + O_{\mathrm{p}}\left(\frac{r^2}{n\sqrt{n}}\right).$$

  Si $r^2/n \to 0$, *alors*

$$\frac{1}{N}\left(\hat{t}^{\mathrm{pc}}_{GREG,r} - t_y\right) = \frac{1}{N}(\hat{t}^{\mathrm{pc}}_{diff,r} - t_y) + o_{\mathrm{p}}\left(\frac{1}{\sqrt{n}}\right).$$

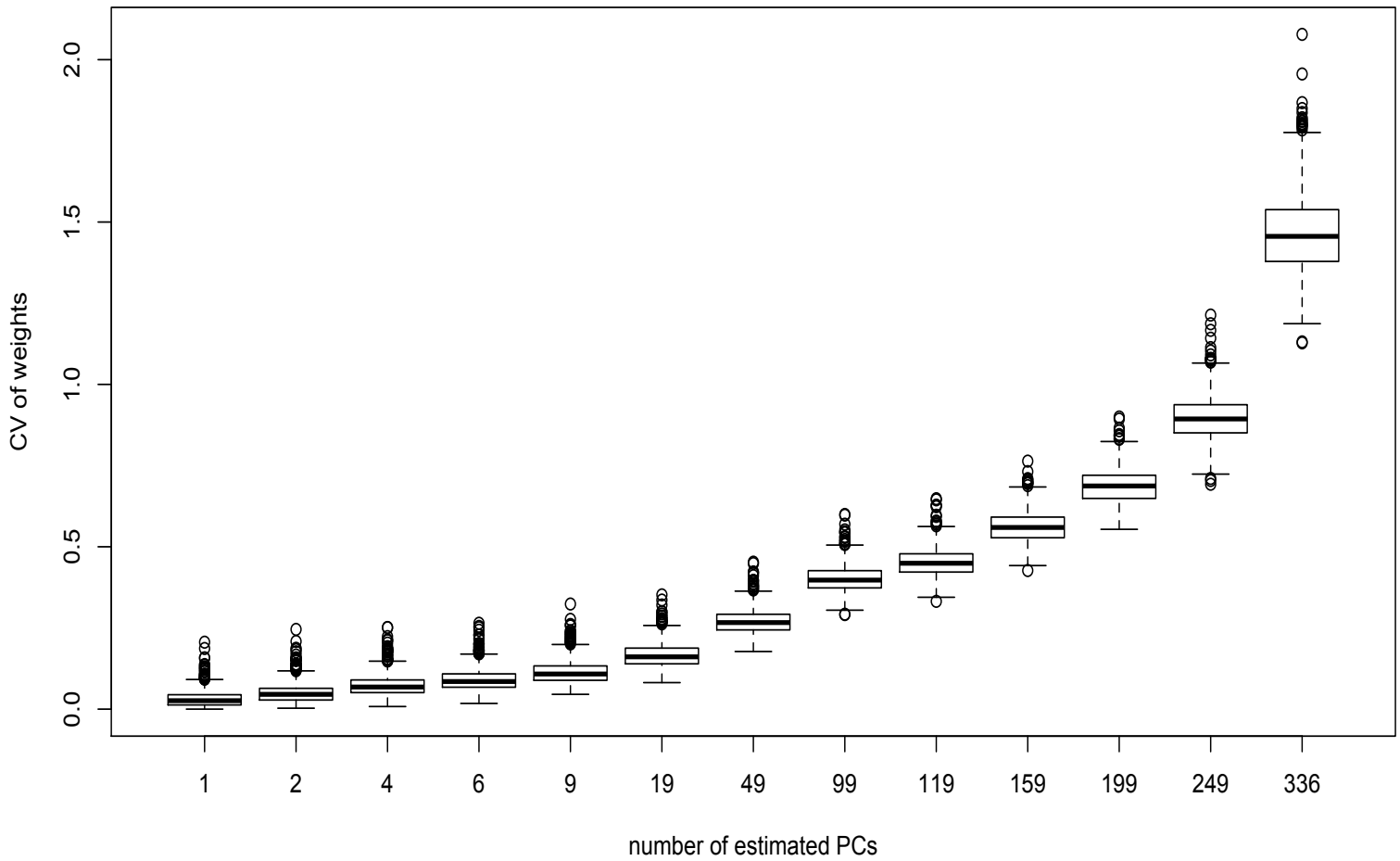
Empirical comparaison on Irish consumption data

- We consider the Irish consumption electricity data as introduced before ;
- The auxiliary variables $X_1, \ldots, X_{336}$ are highly correlated, the matrix $N^{-1}\mathbf{X}^\top\mathbf{X}$ is ill-conditioned (the conditioning number is 65055.78) ;
- The first PC variable $\mathbf{Z}_1$ explains 63% of the total variance of $\mathbf{X}$ and the first 10 PC variables explain more than 80% ;
- The goal is the estimation of the total consumption electricity of each day of the second week :

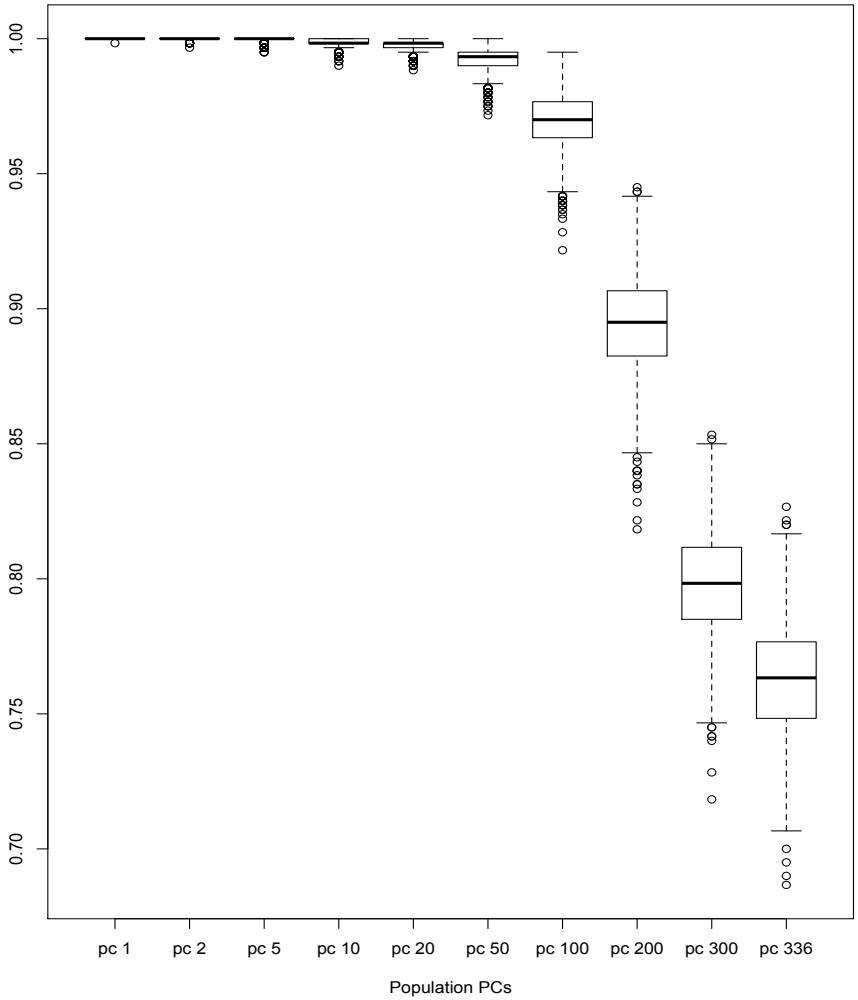
$$t_\ell = \sum_{k \in U} y_{k\ell}, \quad \ell = 1, \ldots, 7$$

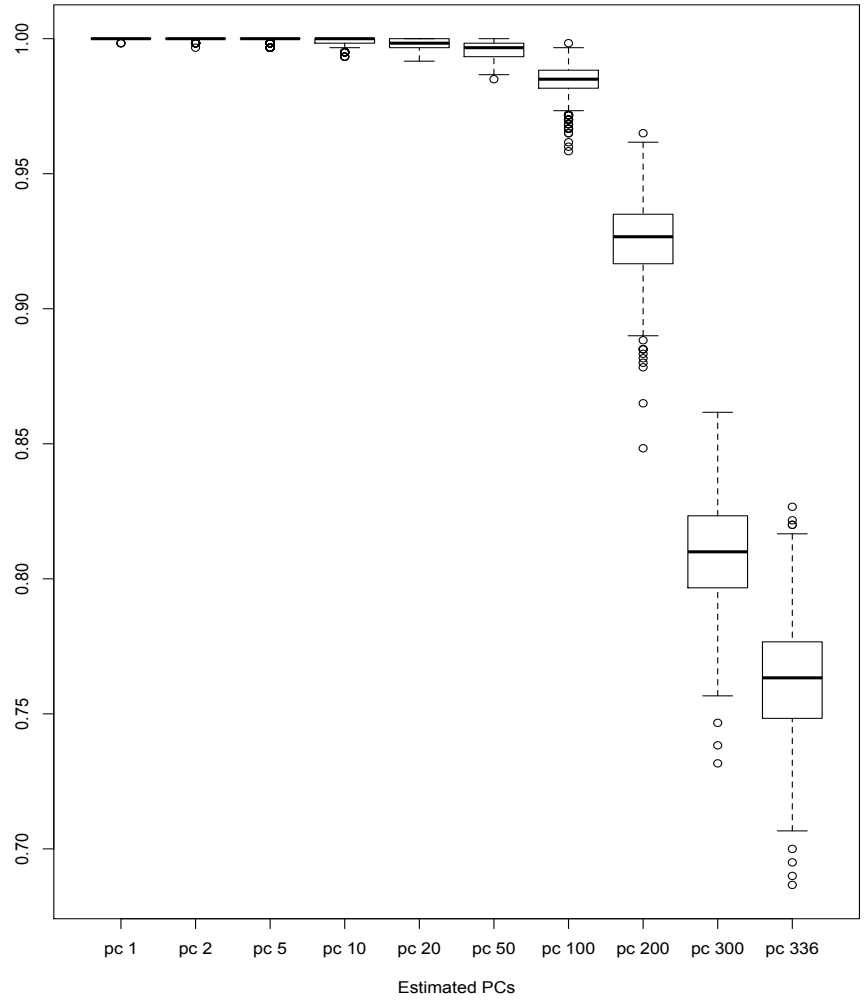
- We select a simple random sampling without replacement of size $n = 600$ and compute the PC model-assisted estimators for an increasing number r of PC variables plus the intercept.

Coefficient de variation des poids de l'estimateur PC-model assisted

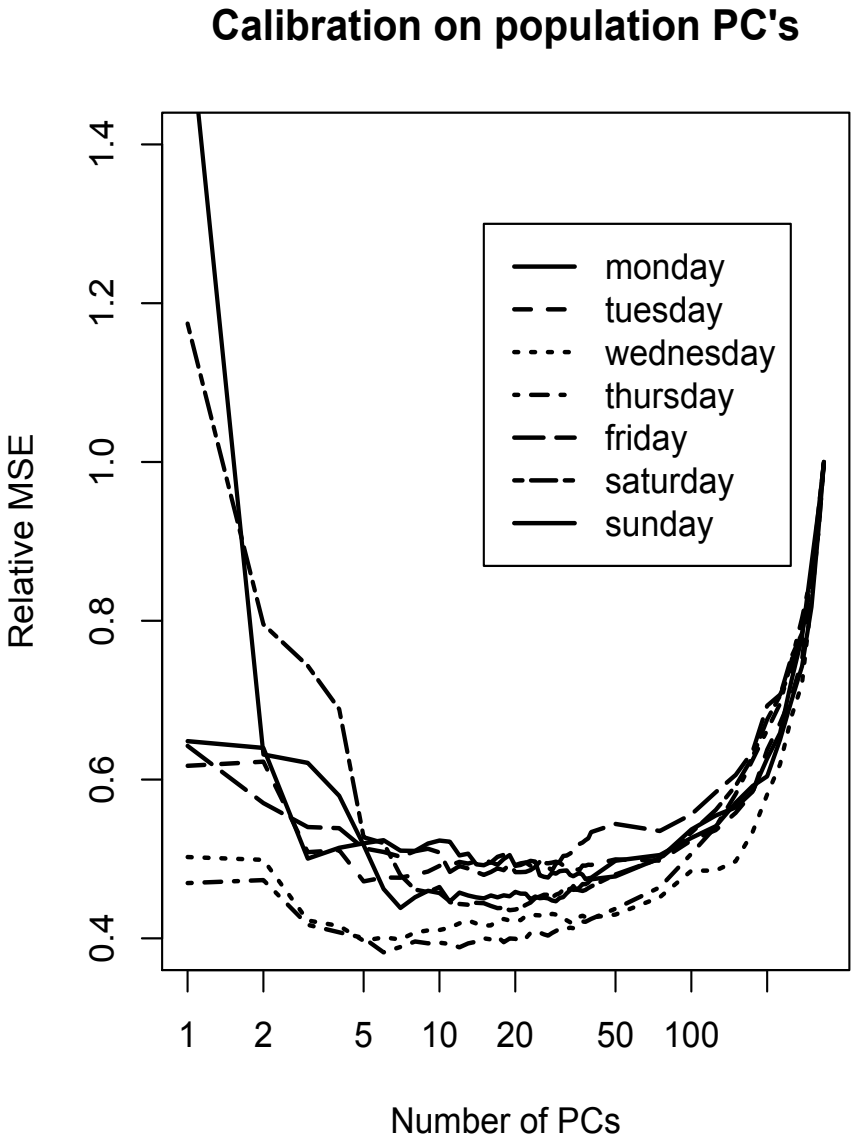


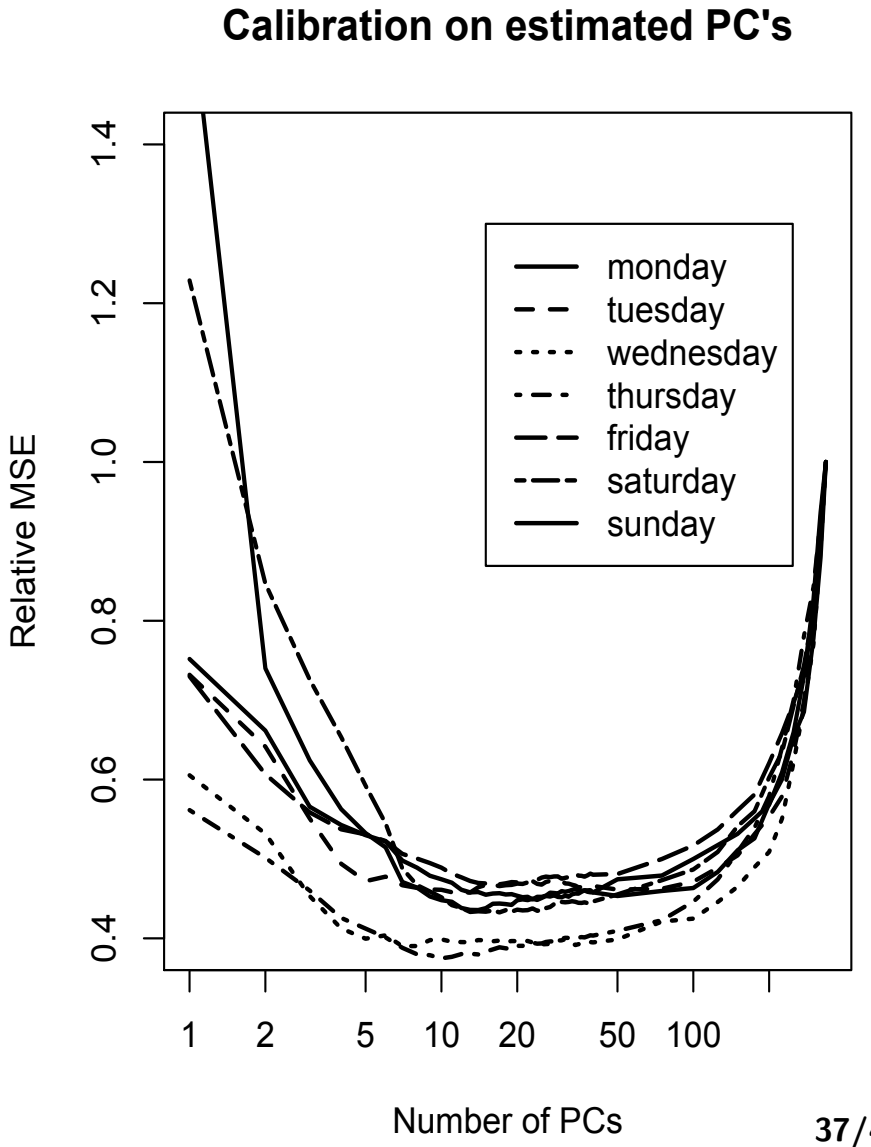
Proportion de poids positifs





Efficacité relative de l'estimateur PC-model assisted estimator par rapport à l'estimateur GREG





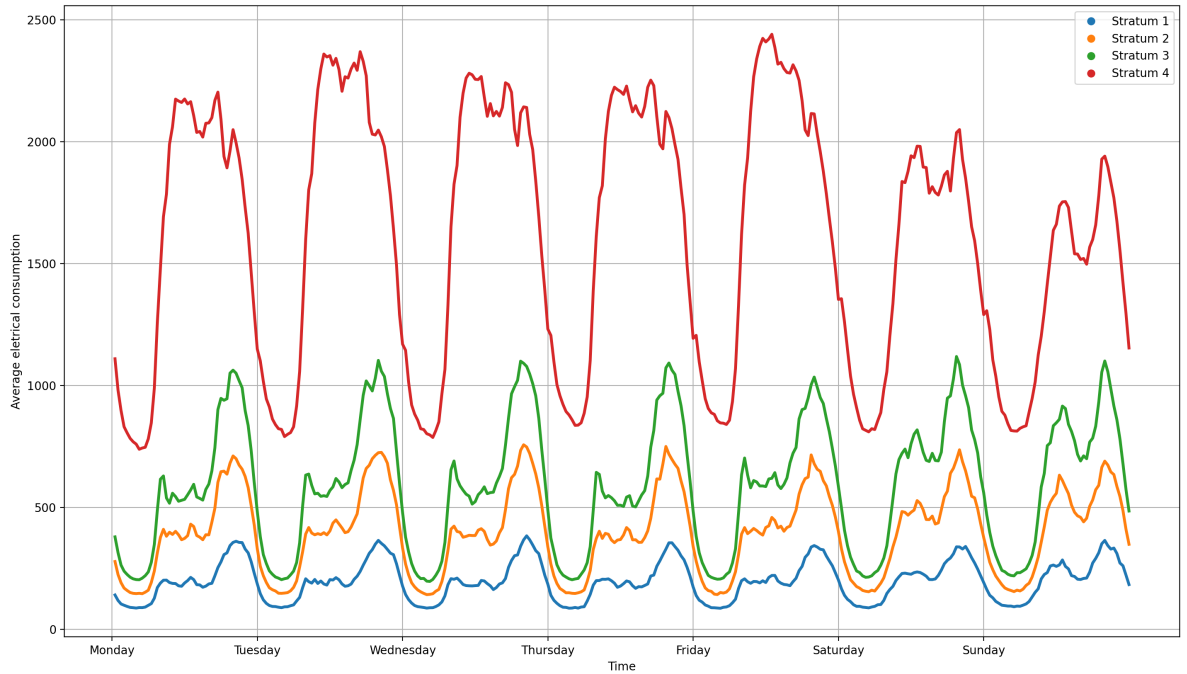
Une règle "data-driven" pour choisir le paramètre r

- Le nombre r de variables PC est un parameter et la performance de l'estimateur PC model-assisted en dépend de sa valeur ;
- Cardot *et al.* (2017) suggèrent sélectionner la plus grande dimension $\hat{r}$ telle que tous les poids PC $w_{ks}^{\mathrm{pc}}(r)$ soient positifs ; c'est l'analogue de la stratégie proposée par Bardsley and Chambers (1984) pour choisir le paramètre λ dans le cadre d'une régression ridge ;
- Le nombre moyen de PC sélectionnées avec cette méthode est de 17.3 ;
- L'efficacité relative par rapport à l'estimateur GREG :

				Days			
Estimators	mo	tu	we	thu	fri	sat	sun
HT	14.4	13.9	11.8	10.8	12.5	6.4	5.4
$\hat{t}_{\ell w}^{\mathrm{pc}}$	0.51	0.49	0.41	0.41	0.52	0.55	0.50
$\hat{t}_{\ell w}^{\mathrm{epc}}$	0.49	0.48	0.41	0.40	0.50	0.53	0.49
Ridge Calibration	0.44	0.46	0.40	0.41	0.48	0.48	0.43

Et si des modèles non-paramétriques sont utilisés ? (Dagdoug et al., JAS, 2022)

- On considère les mêmes données d'électricité irlandaise et on stratifié la population en 4 strates par rapport à la consommation de la première semaine ;
- On considère un échantillon aléatoire simple à l'intérieur de chaque strate de taille totale $n = 600$ avec l'allocation proportionnelle ;
- On désire estimer le total de la consommation d'électricité de Lundi d ela 2ème semaine et on considère plusieurs estimateurs : GREG, ridge et PC-type GREG mais aussi des estimateurs basés sur des méthodes de type machine-learning comme random forests, boosting
- On calcule l'efficacité relative par rapport à l'estimateur d'Horvitz-Thompson.



Estimator	Relative bias	Relative efficiency
GREG	0.2	9.3
PC-GREG	0.1	4.2
Ridge-GREG	0.1	4.0
Lasso-GREG	0.2	4.1
RF	-1.1	17.0
XGB	-1.7	24.9
NN5	-4.0	65.6

Conclusion

- Estimation des totaux de variables dans un contexte de grande dimension (beaucoup de variables auxiliaires) ;
- Les estimateurs traditionnels de type GREG ou par calage peuvent être inefficaces dans ce contexte ;
- Deux classes d'estimateurs peuvent être plus efficaces dans ce contexte que l'estimateur GREG ; néanmoins, ces estimateurs dépendent des paramètres qu'ils doivent être choisis en pratique.

Quelques références

- Bardsley, P. and Chambers, R. (1984), Multipurpose estimation from unbalanced samples, Applied Statistics, 33, 290-299.
- Beaumont, J.F. and Bocci, C. (2008), Another look at ridge regression, Metron-International Journal of Statistics, vol. LXVI, 5-20.
- Cardot, C., Goga, C. and Shehzad, M. A. (2017). "Calibration and Partial Calibration on Principal Components when the Number of Auxiliary Variables is large", *Statistica Sinica*, 27, 243-260.
- Chambers, R. (1996), Robust case-weighting for multi-purpose establishments surveys, Journal of official statistics, vol.12, 1996, 3-32 ;
- Chauvet, G. and Goga, C. (2021) Asymptotic efficiency of the calibration estimator in a high-dimensional data setting (to appear, *Journal of Statistical Planning and Inference*).
- Dagdoug, M., Goga, C. and Haziza, D. (2021). Model-assisted estimation in high-dimensional settings for survey data (to appear, *Journal of Applied Statistics*).
- Deville, J.-C., Särndal, C.-E., 1992. Calibration estimators in survey sampling. Journal of the American Statistical Association 87, 376–382.
- Goga, C., Shehzad, M.-A., and Vanheuverzwyn, A. (2011). Principal component regression with survey data. Application on the French media audience. Int. Statistical Inst. : Proc. 58th World Statistical Congress, 2011, Dublin (Session CPS002), 3847-3852.
- Guggemos, F. and Tillé, Y. (2010), Penalized calibration in survey sampling : Design-based estimation assisted by mixed models. Journal of Statistical Planning and Inference 140 (2010) 3199–3212.
- 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.