Calibration in a high-dimensional setting

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

Webinar in memory of Jean-Claude Deville
May, 25 2022

Outline of my talk

- Motivation : estimation of finite population totals with large auxiliary information data-sets ;
- The calibration estimator in a high-dimensional setting ;
- Two classes of improved calibration estimators based on penalization and dimension reduction methods ;
- Simulation studies on real Irish electricity consumption data.

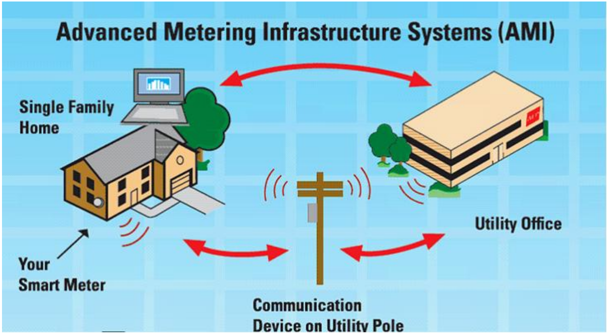
Work done in collaboration with M. Dagdoug, D. Haziza ; G. Chauvet ; H. Cardot et M.A. Shehzad

Surveys in presence of large data-sets

- Emergence of large data-sets due to digital devices which allow recording information at a very fine scale : smart meters, smartphones,. . .
- National Statistical Offices (NSO) have now access to a variety of data sources, potentially exhibiting a large number of observations on a large number of variables.
- Traditional parametric or non-parametric estimation methods may prove inefficient.

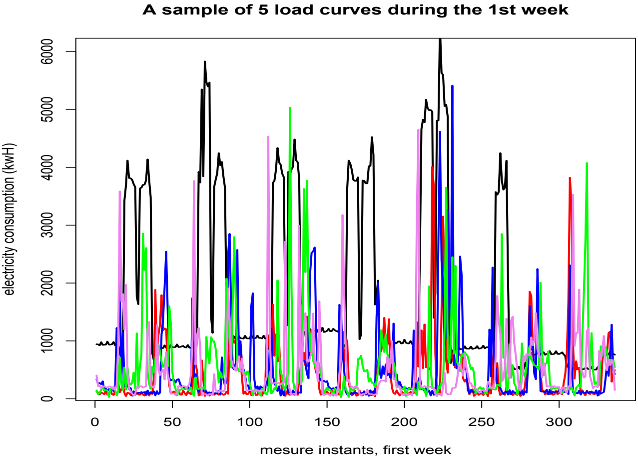
Consumption electricity recorded via smart meters

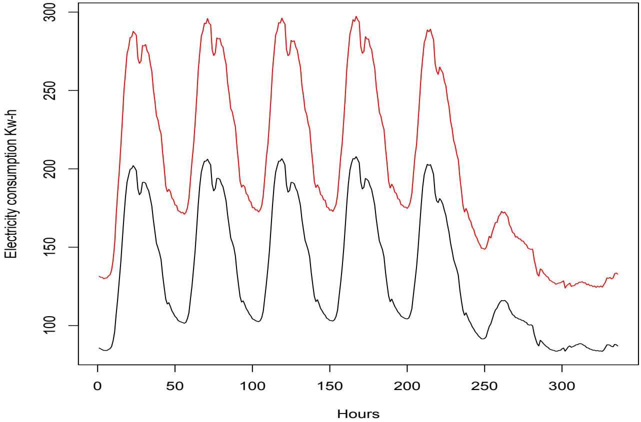
smart meter : smart device installed in households and firms capable to record and send information (electricity consumption) at a very fine scale (every minute, second)



Example 1 : a sample of 5 electricity curves

Test population : 18902 firms and the electricity consumption is recorded every 30 min. during one week.





Electricity consumption Kw-h
300
250
200
150
100
0
50
100
150
200
250
300
Hours

Population, sample

- Let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population of size N (which may be unknown) ;
- Let $s \subset U$ be a sample selected from U according to a sampling design $p(s)$;
- The inclusion probabilities are

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

- Let $\mathcal{Y}$ be the study variable and the goal is the estimation of its finite population total :

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

Horvitz-Thomson estimator and its variance

- With full response, the total t_y is estimated by the Horvitz-Thompson (HT) estimator :

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

- If $\pi_k > 0$ for all $k \in U$, then the HT estimator is design-unbiased for t_y :

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

where the expectation $\mathbb{E}_p(\cdot)$ is taken with respect to the sampling design $p(\cdot)$;

- The design-variance of $\hat{t}_{HT}$ is equal to

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

and if $\pi_{k\ell} > 0$ for all $k, \ell \in U$, it is estimated unbiasedly by

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

Auxiliary information

- Consider the auxiliary variables $X_1, \ldots, X_p$; let $\mathbf{X}$ be the auxiliary information matrix :

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

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

  - the electricity consumption recorded at each instant from the previous week ;

- In a survey framework, we may know $\mathbf{x}_k$ for all $k \in U$ (complete auxiliary information) or only on s with $\sum_{k \in U} \mathbf{x}_k$ known ;
- We may improve the Horvitz-Thompson estimator :
  - at the sampling stage by selecting individuals with π_k built by using this auxiliary information such as the stratified or the proportional to size sampling ;
  - **at the estimation stage by considering an estimator which incorporates this auxiliary information**.

The calibration approach (Deville & Sarndal, 1992)

- Build a weighted estimator of t_y :

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

with weights $w_{ks}, k \in s$ being as close as possible to the sampling weights $1/\pi_k$ and satisfying the *calibration constraints :*

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

- Several distance functions have been considered to measure the closeness between w_{ks} and $1/\pi_k$;
- Deville & Sarndal (1992) showed that the calibration estimator obtained for some distance function is asymptotically equivalent (under regularity assumptions) with the calibration estimator obtained with the chi-squared distance :

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

The calibration estimator for the chi-squared distance

- The calibration weights $w_{ks}, k \in s$ are given by

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

- The calibration estimator is equal to

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

  where $\hat{\boldsymbol{\beta}} = (\sum_{k\in s}\pi_k^{-1}\mathbf{x}_k\mathbf{x}_k^\top)^{-1}\sum_{k\in s}\pi_k^{-1}\mathbf{x}_k y_k$. It is equal to the generalized regression estimator (GREG) obtained in the model-assisted literature.

- Its variance can not be derived directly by using the classical variance formulas because of $\hat{\boldsymbol{\beta}}$; we need "approximations" techniques.

- Let $\hat{t}_{diff}$ be the generalized difference estimator defined as :

$$\hat{t}_{diff} = \hat{t}_{\mathbf{x}HT} - (\hat{t}_{\mathbf{x}HT} - t_{\mathbf{x}})^\top \tilde{\boldsymbol{\beta}}_{OLS} = \sum_{k\in U} \mathbf{x}_k^\top \tilde{\boldsymbol{\beta}}_{OLS} + \sum_{k\in s} \frac{y_k - \mathbf{x}_k^\top \tilde{\boldsymbol{\beta}}_{OLS}}{\pi_k},$$

where $\tilde{\boldsymbol{\beta}}_{OLS} = (\sum_{k\in U} \mathbf{x}_k \mathbf{x}_k^\top)^{-1} \sum_{k\in U} \mathbf{x}_k y_k$.

- Assume mild assumptions on $\mathcal{Y}$, the sampling rate and π_k, π_{kl} as well as on the auxiliary information $\left(||\mathbf{x}_k||^2 \leq C \text{ for all } k \in U\right)$;
- The calibration estimator is asymptotically equivalent to the generalized difference estimator :

$$\begin{aligned} \sqrt{n}N^{-1}\left(\hat{t}_w - t_y\right) &= \sqrt{n}N^{-1}(\hat{t}_{diff} - t_y) + o_p(1) \\ \sqrt{n}N^{-1}\left(\hat{t}_w - t_y\right) &\simeq \sqrt{n}N^{-1}(\hat{t}_{diff} - t_y) \end{aligned}$$

- The asymptotic variance of $\hat{t}_w$ is the variance of $\hat{t}_{diff}$:

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

Estimation with a large number p of auxiliary variables

We consider now that a large number p of auxiliary variables is available.

Question : do we have to consider all this high-dimensional auxiliary information ?

In a classical statistical framework, this situation has already arisen in the early 70's for the estimation of β in a linear modeling context.

Several issues have been noticed :

- for p large, problems of multi-collinearity between the X_j-variables appear ; the information contained in the $\mathbf{X}$-matrix is then redundant ;
- the OLS estimator $\tilde{\boldsymbol{\beta}}_{OLS}$ is certainly unbiased but its variance is very high in this situation ;
- $\tilde{\boldsymbol{\beta}}_{OLS}$ is in average far from $\boldsymbol{\beta}$.

In a survey sampling framework

In a survey sampling framework, Bardsley and Chambers (1984) pointed out that the model-based estimator may be inefficient if a large number of predictors is considered and Rao and Singh (1992) for the calibration estimator :

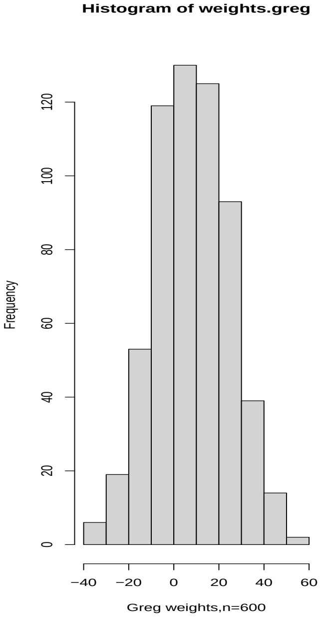
1. the weights w_{ks} used for the model-based or calibration estimators become very instable, they can be very small (even negative with the chi-square distance) or too large.
2. Difficulty to respect predetermined lower and upper bounds :

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

3. Silva and Skinner (1997) noticed on simulation studies that considering a large number of auxiliary variables increases the variance of the calibration estimators ;

Small application on Irish Electricity Data Set

- Commission for Energy Regulation (Ireland)
`http://www.cer.ie/`
- We consider a period of 14 consecutive days and a population of size $N = 6291$ individuals (households and companies) ;
- The electricity consumption is recorded every 30 min ; so, for each unit k from the population, we have $2 \times 7 \times 48 = 672$ measurement instants
- We aim at estimating the total electricity consumption of Monday of the second week :
$$t_y = \sum_{k \in U} y_k,$$
y_k is the consumption of Monday associated to smart meter k;
- Auxiliary information is the electricity consumption of each instant from the previous week, namely $p = 336$ variables :
$$X_k(t_j), j = 1 \ldots, 336, \quad k \in U.$$
- We consider a SRS of size $n = 600$ and we compute the calibration w-weights.


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

Asymptotic efficiency : $n, p \to \infty$ (Chauvet & Goga, JSPI 2022)

On suppose supplementary assumptions on $\mathbf{X}$; we suppose also that $||\mathbf{x}_k||^2 < p\tilde{C}$ for all $k \in U$.

Result

Under the assumed regularity conditions, we have :

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

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

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

  If $p^2/n \to 0$, *then*

$$\frac{\sqrt{n}}{N}\left(\hat{t}_w - t_y\right) \simeq \frac{\sqrt{n}}{N}(\hat{t}_{diff} - t_y).$$

Improving the model-assisted estimator in a high-dimensional setting

Solutions :

1. choose the most important variables by using selection variables criteria suggested for linear modeling ; however, for p very large, these methods may be time-consuming ;
2. use a generalized inverse in case of non-invertibility of $\mathbf{X}^\top\mathbf{X}$;
3. use biased-estimation methods for estimating $\boldsymbol{\beta}$:
   - penalization methods such as ridge (Bardsley and Chambers, 1984 ; Rao and Singh, 1992 ; Beaumont and Bocci, 2008 ; Guggemos and Tillé, 2010) or lasso
   - dimension reduction methods such as principal component regression (Cardot et al., 2017).

The penalized calibration

We look for weights $\mathbf{w}_s^{\text{pen}}(\lambda) = (w_{ks}^{\text{pen}}(\lambda))_{k \in s}$ such that they minimize the penalized chi-squared distance :

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

Different interpretation : we relax the calibration constraints which are no longer exactly verified :

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

- $\lambda = 0$ the constraints are exactly satisfied, we get the usual calibration estimator ;
- $\lambda \to \infty$ no constraint is satisfied, we get the Horvitz-Thompson estimator ;
- Dagdoug *al.* (2021) studied the asymptotic properties when $p \to \infty$;

The Principal Component calibration estimator (Cardot *et al.*, Stat. Sinica, 2017)

- We derive the Principal Components $\mathbf{Z}_1, \ldots, \mathbf{Z}_p$ of $\mathbf{X}$ (linear combination of $\mathbf{X}_j, j = 1, \ldots, p$, non-correlated and of maximum variance) :

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

  where $\mathbf{v}_j$ is the eigenvector associated to the largest eigenvalue λ_j of $N^{-1}\mathbf{X}^\top\mathbf{X}$;

- The new calibration variables are $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$ associated to the largest eigenvalues $\lambda_1 \geq \ldots \geq \lambda_r$ with $r << p$:

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

- The PC-weights $w_{ks}^{\mathrm{pc}}(r), k \in s$ may be obtained by calibrating on the zero totals of the first r PC, namely :

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

- The PC-calibrated estimator is given by :

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

- We estimate exactly the projection of the totals of the initial auxiliary variables on the space spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_r$:
  1. $r = 0$: we obtain the Horvitz-Thompson estimator $\hat{t}_{y\mathrm{HT}}$ which doesn't use the auxiliary information.
  2. $r = p$: we obtain the calibration estimator which uses all initial p auxiliary variables.
  3. partial calibration : we estimate exactly the totals of p_1 variables and we penalize the other $p - p_1$ variables (Bardsley and Chambers, 1984 ; Guggemos and Tillé, 2010).
  4. Cardot *et al.* (2017) studied the asymptotic properties of the PC-calibrated estimator when $r, p \to \infty$.

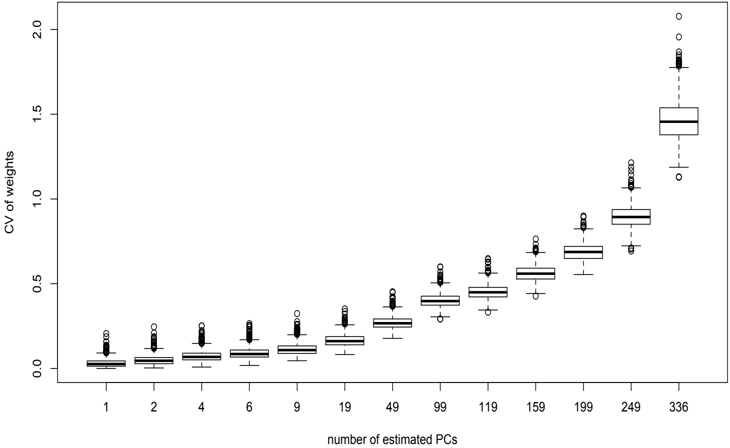
Empirical comparaison on Irish consumption data

- 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 :

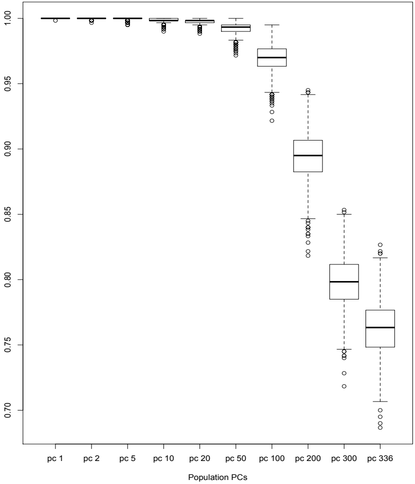
$$t_\ell = \sum_{k \in U} y_{k\ell}, \quad \ell = 1, \ldots, 7$$

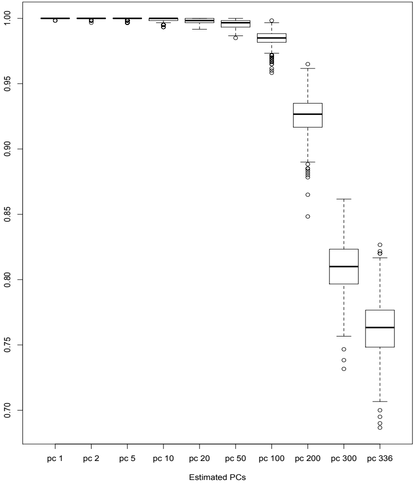
- We select a simple random sampling without replacement of size $n = 600$ and compute the PC model-assisted estimators for an increasing number r of PC variables plus the intercept.

Coefficient of variation of PC-calibration weights

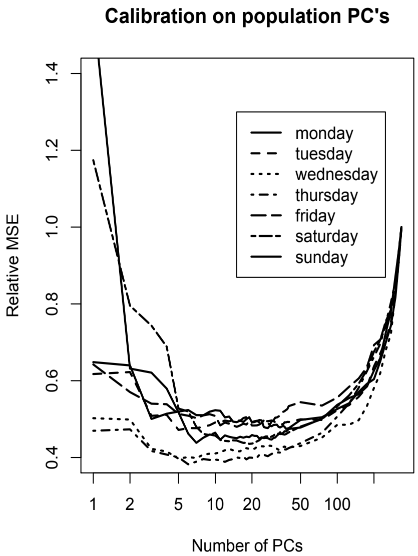


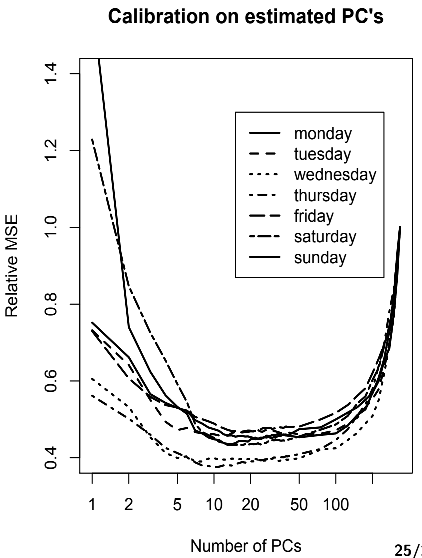
Proportion of positive weights





Relative efficiency of the PC-calibration estimator with respect to the calibration estimator





Data-driven rule for choosing the tuning parameter

- The number r of PC variables is a tuning parameter and the performance of the PC-calibration estimator depends on it ;
- Cardot *et al.* (2017) suggest selecting the largest dimension $\hat{r}$ such that all the PC weights $w_{ks}^{\mathrm{pc}}(r)$ remain positive ; it is the analogue of the strategy suggested in Bardsley and Chambers (1984) for choosing the tuning parameter λ in a ridge regression context ;
- The mean number of selected principal components with the data driven selection rule was equal to 17.3 for the population principal components and 21.3 for the sample principal components.
- The relative efficiencies with respect to the calibration estimator are given below :

	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

Conclusion

- Estimation of finite population totals with high-dimensional auxiliary data sets ;
- Traditional calibration estimator or the calibration estimator may be inefficient in this setting ; additional variability if p is very large with respect to n;
- Two classes of alternatives estimators which may be more efficient that the calibration estimator with high-dimensional auxiliary data sets. However, they need to choose tuning parameters.

Some references

- 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.