Statistica Sinica **27** (2017), 243-260
doi:http://dx.doi.org/10.5705/ss.202014.0125

CALIBRATION AND PARTIAL CALIBRATION ON PRINCIPAL COMPONENTS WHEN THE NUMBER OF AUXILIARY VARIABLES IS LARGE

H. Cardot[1], C. Goga[1] and M.-A. Shehzad[1,2]

[1] *Université de Bourgogne Franche Comté and* [2] *Bahauddin Zakariya University*

Abstract: In survey sampling, calibration is a popular tool used to make total estimators consistent with known totals of auxiliary variables and to reduce variance. When the number of auxiliary variables is large, calibration on all the variables may lead to estimators of totals whose mean squared error (MSE) is larger than the MSE of the Horvitz-Thompson estimator even if this simple estimator does not take account of the available auxiliary information. We study a new technique based on dimension reduction through principal components that can be useful in this large dimension context. Calibration is performed on the first principal components, which can be viewed as the synthetic variables containing the most important part of the variability of the auxiliary variables. When some auxiliary variables play a more important role than others, the method can be adapted to provide an exact calibration on these variables. Some asymptotic properties are given in which the number of variables is allowed to tend to infinity with the population size. A data-driven selection criterion of the number of principal components ensuring that all the sampling weights remain positive is discussed. The methodology of the paper is illustrated, in a multipurpose context, by an application to the estimation of electricity consumption with the help of 336 auxiliary variables.

Key words and phrases: Dimension reduction, model-assisted estimation, multipurpose surveys, partial calibration, partial least squares, penalized calibration, ridge regression, survey sampling, variance approximation.

1. Introduction

Since the seminal work by Deville and Särndal (1992), calibration is one of the most popular and useful tools to improve Horvitz-Thompson estimators of totals in a design-based survey sampling framework. Roughly speaking, it consists in looking for a modification of the sampling weights so that the totals, in the population, of the auxiliary variables are perfectly estimated. Performing calibration often leads to total estimators with smaller variances and this technique is routinely used by several national statistical agencies (see Särndal (2007) for a review).

With the spread of automatic processes for data collection, as well as increasing storage capacities, it is not unusual anymore to have to analyze data coming from very large surveys with many auxiliary variables. Here, calibration on all the auxiliary variables can lead to estimators whose performances are worse than that of the simple Horvitz-Thompson estimator even if the latter does not account for any auxiliary information (see e.g., Silva and Skinner (1997)). Several difficulties arise in this context, such as instability of the calibration weights or variance inflation. There are different ways of dealing with these issues. One possibility is to choose only a subset of the auxiliary variables and to consider only the auxiliary variables that are expected to be the more pertinent, avoiding the problem of multicollinearity (see e.g., Silva and Skinner (1997); Chambers, Skinner and Wang (1999) and Clark and Chambers (2008)). Another way is to weaken exact calibration constraints to approximated ones. A class of penalized estimators has been suggested by Bardsley and Chambers (1984) in a model-based setting and extended later by Chambers (1996), Rao and Singh (1997), and Théberge (2000) in a design-based (or model-assisted) setting. Usually, some auxiliary variables play a role that is more important than others and it is required that their totals be estimated exactly. Bardsley and Chambers (1984) and Guggemos and Tillé (2010) suggested different penalized optimization problems which lead in fact to the same system of weights (see Goga, Shehzad and Vanheuverzwyn (2011)).

We present another way of dealing with this issue. Our estimator is based on dimension reduction of the auxiliary variables via principal components calibration. In multivariate statistics, principal component analysis (PCA) is a popular tool for reducing the dimension of a set of quantitative variables (see e.g., Jolliffe (2002)) by transforming the initial data set into a new set of a few uncorrelated synthetic variables, called principal components (PC), that are linear combinations of the initial variables with the largest variance. Adopting a model-assisted point of view, the PCA calibration approach can also be viewed as a GREG estimator based on Principal Components Regression (PCR). PCR can be very useful to reduce the number of covariates in a linear regression model especially when the regressors are highly correlated. As explained in Jolliffe (2002), even if PCR is a biased estimation method for estimating a regression coefficient, it is useful to overcome the problem of multicollinearity among the regressors. The method is easy to put into practice with classical softwares used for performing calibration.

A natural alternative to principal components regression is partial least squares (PLS), also a popular dimension reduction regression technique that can be useful when there is a large number of auxiliary variables that are highly correlated (see for example Swold, Sjöström and Eriksson (2001)). Other model selection techniques, such as the Lasso (Tibshirani (1996)) or the elastic net (Zou and Hastie (2005)) can be employed to deal with survey data with large number of auxiliary variables. The main drawback of these model selection techniques is that they can give survey weights that depend explicitly on the outcome variable, generally not desired in surveys, particularly in multipurpose surveys in which there can be many outcome variables under study.

The paper is structured as follows: we briefly recall in Section 2 the calibration method and the problems that can arise when the number of auxiliary variables is large. We introduce the suggested method in Section 3, and we give a model-assisted interpretation. When the values of the auxiliary variables are only known in the sample, we first estimate the PC's and then perform calibration on the first estimated principal components (Section 4). In Section 5, under mild assumptions on the sampling design and on the study and auxiliary variables, we prove that the calibration estimator on true PC's and on estimated PC's are consistent. We show in Section 6 how the method can be adapted to provide an exact calibration on variables considered by the survey statistician more important than others. Our method is illustrated in Section 7 on the estimation of the total electricity consumption for each day of a week with the help of the past consumption measured every half an hour over the previous week. A brief Section 8 gives some concluding remarks. The proofs as well as some additional results on the electricity data are available in an online supplementary file.

2. Calibration over a Large Number of Auxiliary Variables

We consider the finite population $U = \{1, \ldots, k, \ldots, N\}$ and wish to estimate the total $t_y = \sum_{k \in U} y_k$, where y_k is the value of the variable of interest $\mathcal{Y}$ for the kth unit. Let s be a random sample, with fixed size n, drawn from U according to a sampling design that assigns to unit k a known inclusion probability $\pi_k = \Pr(\mathrm{k} \in \mathrm{s})$ satisfying $\pi_k > 0$. The corresponding sampling design weight is denoted by $d_k = 1/\pi_k$. We suppose that y_k is known for all $k \in s$ (complete response).

Without auxiliary information, the total t_y is estimated unbiasedly by the Horvitz-Thompson (HT) estimator $\hat{t}_{yd} = \sum_{k \in s} d_k y_k$. Consider now p auxiliary variables, $\mathcal{X}_1, \ldots, \mathcal{X}_p$, and let $\mathbf{x}_k^T = (x_{k1}, \ldots, x_{kp})$ be the transposed vector whose elements are the values of the auxiliary variables for the kth unit. The calibration method developed by Deville and Särndal (1992) uses as effectively as possible the known population totals of $\mathcal{X}_j$, $j = 1, \ldots, p$ at the estimation stage. The calibration estimator of t_y is the weighted estimator

$$\hat{t}_{yw} = \sum_{k \in s} w_k y_k, \tag{2.1}$$

whose (calibrated) weights w_k are chosen as close as possible to the initial sampling weights d_k, according to some distance Φ_s and subject to some constraints.

More exactly,

$$(w_k)_{k\in s} = \text{argmin}_w \Phi_s(w) \tag{2.2}$$

$$\text{subject to} \quad \sum_{k\in s} w_k \mathbf{x}_k = t_{\mathbf{x}}, \tag{2.3}$$

where $w = (w_k, k \in s)$ is the vector of weights assigned to each unit in the sample, and $t_{\mathbf{x}} = \sum_{k\in U} \mathbf{x}_k$ is the vector whose elements are the known totals of $\mathcal{X}_j$ for $j = 1, \ldots, p$. Several distance functions Φ_s were studied in Deville and Särndal (1992). Under weak regularity assumptions they showed that all resulting estimators are asymptotically equivalent to the one obtained by minimizing the chi-square distance function $\Phi_s(w) = \sum_{k\in s}(w_k - d_k)^2/q_k d_k$, where the q_k's are known positive constants that can be used to take account of the variability of the observations and are unrelated to d_k. Of common use in applications are uniform weights $q_k = 1$ for all units k and we suppose, without loss of generality, that $q_k = 1$ in the following. We only consider calibration estimators derived using the chi-square distance. The calibration weights w_k, $k \in s$, are

$$w_k = d_k - d_k \mathbf{x}_k^T \left(\sum_{\ell \in s} d_\ell \mathbf{x}_\ell \mathbf{x}_\ell^T \right)^{-1} (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}), \tag{2.4}$$

where $\hat{t}_{\mathbf{x}d} = \sum_{k\in s} d_k \mathbf{x}_k$ is the HT estimator of $t_{\mathbf{x}}$, and the corresponding calibration estimator is obtained by plugging w_k in (2.1).

With a different point of view, it can be shown that the calibration estimator obtained with the chi-squared distance is equal to the generalized regression estimator (GREG) which is derived by assuming a linear regression model between the study variable $\mathcal{Y}$ and the auxiliary variables $\mathcal{X}_1, \ldots, \mathcal{X}_p$,

$$\xi: \quad y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k, \tag{2.5}$$

where $\varepsilon = (\varepsilon_k, k \in U)$ is a centered random vector with a diagonal variance matrix whose diagonal elements are $1/q_k$. Cassel, Särndal and Wretman (1976) suggested the generalized difference estimator

$$\tilde{t}_{y,\mathbf{x}}^{\text{diff}} = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \tilde{\boldsymbol{\beta}}_{\mathbf{x}}, \tag{2.6}$$

where $\tilde{\boldsymbol{\beta}}_{\mathbf{x}} = (\sum_{k\in U} \mathbf{x}_k \mathbf{x}_k^T)^{-1} \sum_{k\in U} \mathbf{x}_k y_k$ is the ordinary least squares estimator of $\boldsymbol{\beta}$. Here $\tilde{t}_{y,\mathbf{x}}^{\text{diff}}$ cannot be computed because $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}$ cannot be computed unless we have observed the whole population. We estimate $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}$ by $\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \left(\sum_{k\in s} d_k \mathbf{x}_k \mathbf{x}_k^T\right)^{-1} \sum_{k\in s} d_k \mathbf{x}_k y_k$ and obtain the GREG estimator of t_y : $\hat{t}_{yw} = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}$.

Under mild regularity assumptions, Deville and Särndal (1992) showed that the calibration estimator $\hat{t}_{yw}$ and $\tilde{t}_{y,\mathbf{x}}^{\text{diff}}$ have the same asymptotic distribution. We have $N^{-1}\left(\hat{t}_{yw}-t_y\right)=N^{-1}\left(\tilde{t}^{\text{diff}}_{y,\mathbf{x}}-t_y\right)+o_{\text{p}}(n^{-1/2})$ and, as a result, the asymptotic variance of $\hat{t}_{yw}$ is $AV(\hat{t}_{yw})=\sum_{k\in U}\sum_{\ell\in U}(\pi_{k\ell}-\pi_k\pi_\ell)d_kd_\ell(y_k-\mathbf{x}_k^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}})(y_\ell-\mathbf{x}_\ell^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}})$, where $\pi_{k\ell}=\Pr(\text{k}\in\text{s}\ \&\ \ell\in\text{s})$ is the probability that both k and ℓ are included in the sample s, and $\pi_{kk}=\pi_k$. Calibration improves the HT estimator, $AV(\hat{t}_{yw})\le V(\hat{t}_{yd})$, if the predicted values $\mathbf{x}_k^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}}$ are close enough to the y_k's, that is to say if the model (2.5) explains the variable of interest sufficiently well. Nevertheless, when a large number p of auxiliary variables is used, this no longer holds (see Silva and Skinner (1997)).

One way to circumvent the problems due to over-calibration, such as extremely large weights and variance inflation, is to weaken the exact calibration constraints to approximate ones. Then, the deviation between $\sum_{k\in s}w_k\mathbf{x}_k$ and $\sum_{k\in U}\mathbf{x}_k$ is controlled by means of a penalty. Bardsley and Chambers (1984), in a model-based setting, and Chambers (1996) and Rao and Singh (1997) in a design-based setting, suggested finding weights satisfying (2.2), subject to a quadratic constraint, as $w^{\text{pen}}(\lambda)=\arg\min_w\Phi_s(w)+\lambda^{-1}\left(\hat{t}_{\mathbf{x}w}-t_{\mathbf{x}}\right)^T\mathbf{C}\left(\hat{t}_{\mathbf{x}w}-t_{\mathbf{x}}\right)$, where $\hat{t}_{\mathbf{x}w}=\sum_{k\in s}w_k\mathbf{x}_k$, $\mathbf{C}=\text{diag}(c_j)_{j=1}^p$, and $c_j\ge 0$ is a user-specified cost associated with the jth calibration constraint. The tuning parameter $\lambda>0$ controls the trade-off between exact calibration ($\lambda\to 0$) and no calibration ($\lambda\to\infty$). With the chi-square distance, the solution is, for $k\in s$,

$$w_k^{\text{pen}}(\lambda)=d_k-d_k\mathbf{x}_k^T\left(\sum_{\ell\in s}d_\ell\mathbf{x}_\ell\mathbf{x}_\ell^T+\lambda\mathbf{C}^{-1}\right)^{-1}\left(\hat{t}_{\mathbf{x}d}-t_{\mathbf{x}}\right),$$

and the penalized calibration estimator is a GREG-type estimator, whose regression coefficient is estimated by a ridge-type estimator. For an infinite cost c_j, the jth calibration constraint is satisfied exactly (see Beaumont and Bocci (2008)). As noted in Bardsley and Chambers (1984), the risk of having negative weights (in the case of the chi-square distance) is greatly reduced by using penalized calibration. With an empirical likelihood approach, Chen, Sitter and Wu (2002) suggested replacing true totals $t_{\mathbf{x}}$ with $t_{\mathbf{x}}+\Delta(\hat{t}_{\mathbf{x}w}-t_{\mathbf{x}})$, where Δ is a diagonal matrix depending on the costs c_j and a tuning parameter controlling the deviation between $\hat{t}_{\mathbf{x}w}$ and $\hat{t}_{\mathbf{x}}$.

3. Calibration on Principal Components

We consider another class of approximately calibrated estimators that are based on dimension reduction through principal components analysis (PCA). In multivariate statistics, PCA is a popular technique for reducing the dimension of a set of quantitative variables (see e.g., Jolliffe (2002)) by extracting most of the variability of the data by projection on a low dimension space. Principal components analysis consists in transforming the initial data set into a new set of a few uncorrelated synthetic variables, called principal components (PC), that are linear combinations of the initial variables with the largest variance. The principal components are "naturally" ordered, with respect to their contribution to the total variance of the data, and the reduction of the dimension is then realized by taking only the first few PCs. PCA is particularly useful when the correlation among the variables in the dataset is strong. These new variables can be also used as auxiliary information for calibration, as noted in Goga, Shehzad and Vanheuverzwyn (2011).

Complete Auxiliary Information

We suppose without loss of generality that the auxiliary variables are centered, $N^{-1}t_{\mathbf{x}} = 0$ and, for simplicity, we do not include an intercept term in the model. In applications, an intercept term should be included. We suppose the p-dimensional vector $\mathbf{x}_k$ is known for all units $k \in U$.

Let $\mathbf{X}$ be the $N \times p$ data matrix having $\mathbf{x}_k^T, k \in U$ as rows. The variance-covariance matrix of the original variables $\mathcal{X}_1, \ldots, \mathcal{X}_p$ is $N^{-1}\mathbf{X}^T\mathbf{X}$. Let $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ be the eigenvalues of $N^{-1}\mathbf{X}^T\mathbf{X}$ associated to the corresponding orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$,

$$\frac{1}{N}\mathbf{X}^T\mathbf{X}\mathbf{v}_j = \lambda_j \mathbf{v}_j, \quad j = 1, \ldots, p. \tag{3.1}$$

For $j = 1, \ldots, p$, the jth principal component, denoted by $\mathbf{Z}_j$, is

$$\mathbf{Z}_j = \mathbf{X}\mathbf{v}_j = (z_{kj})_{k \in U}. \tag{3.2}$$

The variable $\mathbf{Z}_j$ has a (population) variance equal to $N^{-1}\sum_{k\in U} z_{kj}^2 = \lambda_j$. We consider the first r (with $r < p$) principal components, $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$. In the survey sampling framework, our goal is not to interpret $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$ but to use them as a tool to obtain calibration weights that are more stable than the calibration weights obtained with the full set of auxiliary variables. We take the principal component (PC) calibration estimator $\hat{t}_{yw}^{\mathrm{pc}}(r) = \sum_{k\in s} w_k^{\mathrm{pc}}(r) y_k$, with the vector of weights $w_k^{\mathrm{pc}}(r), k \in s$, that solve the optimization problem (2.2) subject to $\sum_{k\in s} w_k^{\mathrm{pc}}(r)\mathbf{z}_{kr} = \sum_{k\in U}\mathbf{z}_{kr}$, where $\mathbf{z}_{kr}^T = (z_{k1}, \ldots, z_{kr})$ is the vector containing the values of the first r PCs computed for the kth individual. The PC calibration weights are $w_k^{\mathrm{pc}}(r) = d_k - d_k\mathbf{z}_{kr}^T\left(\sum_{\ell\in s} d_\ell \mathbf{z}_{\ell r}\mathbf{z}_{\ell r}^T\right)^{-1}(\hat{t}_{\mathbf{z}_r d} - t_{\mathbf{z}_r}), k \in s$, where $\hat{t}_{\mathbf{z}_r d} = \sum_{k\in s} d_k \mathbf{z}_{kr}$ is the HT estimator of the total $t_{\mathbf{z}_r} = (0, \ldots, 0)$ since we have supposed that the original variables have mean zero.

The total t_y is again estimated by a GREG-type estimator that uses $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$ as auxiliary variables,

$$\hat{t}_{yw}^{\mathrm{pc}}(r) = \sum_{k\in s} w_k^{\mathrm{pc}}(r) y_k = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{z}_r d} - t_{\mathbf{z}_r}\right)^T \hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r), \tag{3.3}$$

where

$$\hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r) = \left(\sum_{k\in s} d_k \mathbf{z}_{kr}\mathbf{z}_{kr}^T\right)^{-1} \sum_{k\in s} d_k \mathbf{z}_{kr} y_k. \tag{3.4}$$

If $r = 0$, we do not take auxiliary information into account, then $\hat{t}_{yw}^{\mathrm{pc}}(0)$ is simply the HT estimator (or the Hájek estimator if the intercept term is included in the model) whereas if $r = p$, we get the calibration estimator that takes account of all the auxiliary variables.

A Model-Assisted Point of View

Consider the superpopulation model ξ presented in (2.5) and denote by $\mathbf{G} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$ the matrix whose jth column is the jth eigenvector $\mathbf{v}_j$. We can write,

$$\xi: \quad y_k = \mathbf{z}_k^T \boldsymbol{\gamma} + \varepsilon_k,$$

where $\boldsymbol{\gamma} = \mathbf{G}^T\boldsymbol{\beta}$ and $\mathbf{z}_k^T = (z_{k1}, \ldots, z_{kp})$, where z_{kj} is the value of $\mathbf{Z}_j$ for the kth unit. Principal components regression consists of considering a reduced linear regression model, denoted by ξ_r, that uses as predictors the first r principal components, $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$,

$$\xi_r: \quad y_k = \mathbf{z}_{kr}^T \boldsymbol{\gamma}(r) + \varepsilon_{kr}, \tag{3.5}$$

where $\boldsymbol{\gamma}(r)$ is a vector of r elements composed of the first r elements of $\boldsymbol{\gamma}$ and ε_{kr} is the appropriate error term of mean zero. The least squares estimation, at the population level, of $\boldsymbol{\gamma}(r)$, is

$$\tilde{\boldsymbol{\gamma}}_{\mathbf{z}}(r) = \left(\sum_{k\in U} \mathbf{z}_{kr}\mathbf{z}_{kr}^T\right)^{-1} \sum_{k\in U} \mathbf{z}_{kr} y_k, \tag{3.6}$$

which in turn can be estimated, on a sample s, by the design-based estimator $\hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r)$ given by (3.4). Thus the PC calibration estimator given in (3.3) is a GREG-type estimator assisted by the reduced model ξ_r described in (3.5). Since the principal components are centered and uncorrelated, the matrix $\left(\sum_{k\in U} \mathbf{z}_{kr}\mathbf{z}_{kr}^T\right)$ is diagonal, with diagonal elements $(\lambda_1 N, \ldots, \lambda_r N)$.

When there is a strong multicollinearity among the auxiliary variables, the ordinary least squares estimator of $\boldsymbol{\beta}$ is sensitive to small changes in $\mathbf{x}_k$ and y_k and has a large variance (see e.g., Hoerl and Kennard (1970)). To see how small eigenvalues can affect $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}$, Gunst and Mason (1977) write the least squares estimator as: $\tilde{\boldsymbol{\beta}}_{\mathbf{x}} = \left(N^{-1}\sum_{k\in U} \mathbf{x}_k\mathbf{x}_k^T\right)^{-1} \left(N^{-1}\sum_{k\in U} \mathbf{x}_k y_k\right) = \sum_{j=1}^{p} (1/\lambda_j)[\mathbf{v}_j^T(N^{-1}\sum_{k\in U} \mathbf{x}_k y_k)]\mathbf{v}_j$. Approximating the covariance matrix $N^{-1}\sum_{k\in U} \mathbf{x}_k\mathbf{x}_k^T = N^{-1}\mathbf{X}^T\mathbf{X}$ by the rank r matrix $\left(\sum_{j=1}^{r} \lambda_j \mathbf{v}_j\mathbf{v}_j^T\right)$ leads to considering the regression estimator based on the first r principal components,

$$\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r) = \sum_{j=1}^{r} \frac{1}{\lambda_j} \Big[\mathbf{v}_j^T \Big(\frac{1}{N} \sum_{k \in U} \mathbf{x}_k y_k \Big) \Big] \mathbf{v}_j. \tag{3.7}$$

Here $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)$ is obtained by subtracting from $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}$ the part of the data that belongs to the $p-r$ dimensional space with the smallest variance and performing the regression in the r-dimensional space that contains most of the variability of the data. Ridge-regression (Hoerl and Kennard (1970)), an alternative way of dealing with the multicollinearity issue, consists of adding a positive term λ to all eigenvalues $\lambda_j, j = 1, \ldots, p$, with $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\lambda) = (N^{-1}\sum_{k\in U}\mathbf{x}_k\mathbf{x}_k^T + \lambda\mathbf{I}_p)^{-1}\left(N^{-1}\sum_{k\in U}\mathbf{x}_k y_k\right) = \sum_{j=1}^{p}[1/(\lambda+\lambda_j)]\left[\mathbf{v}_j^T\left(N^{-1}\sum_{k\in U}\mathbf{x}_k y_k\right)\right]\mathbf{v}_j$, where $\mathbf{I}_p$ is the p-dimensional identity matrix. Both the ridge regression estimator $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\lambda)$ and the principal components estimator $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)$ are biased for $\boldsymbol{\beta}$ under the model ξ (Gunst and Mason (1977)).

The PC regression estimator $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)$ can be estimated under the sampling design by

$$\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r) = \mathbf{G}_r \hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r), \tag{3.8}$$

where $\hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r)$ is given in (3.4) and $\mathbf{G}_r$ is the $p \times r$ matrix whose jth column is $\mathbf{v}_j$. Using (3.8) and the fact that $\mathbf{Z}_j = \mathbf{X}\mathbf{v}_j$, we obtain that $\left(\hat{t}_{\mathbf{z}_r d} - t_{\mathbf{z}_r}\right)^T \hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r) = \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)$. Consequently $\hat{t}_{yw}^{\mathrm{pc}}(r)$ can also be written as $\hat{t}_{yw}^{\mathrm{pc}}(r) = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)$, and $\hat{t}_{yw}^{\mathrm{pc}}(r)$ may be seen as a GREG-type estimator assisted by the model ξ when $\boldsymbol{\beta}$ is estimated by $\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)$.

Calibration on the second moment of the PC variables

With complete auxiliary information, Särndal (2007) stated that “we are invited to consider $x_{kj}^2, j = 1, \ldots, p$ and other functions of x_{kj}^2 for inclusion in $\mathbf{x}_k$” especially when “the relationship to the study variable is curved”. In our case, the PC variables $\mathbf{Z}_j$ satisfy $N^{-1}\mathbf{Z}_j^T\mathbf{Z}_j = N^{-1}\sum_{k\in U} z_{kj}^2 = \lambda_j$, for all $j = 1, \ldots, p$. Thus in the presence of complete auxiliary information, the totals of squares of the PCs are known. As a consequence, if we keep the first r variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$ corresponding to the largest r eigenvalues, we can consider r additional calibration constraints on the second moment of these PCs. We look for the calibration weights $w^{\mathrm{pc}}(r)$ solving (2.2) subject to $\sum_{k\in s} w_k^{\mathrm{pc}}(r)\left(\mathbf{z}_{kr}, \mathbf{z}_{kr}^2\right)^T = \sum_{k\in U}\left(\mathbf{z}_{kr}, \mathbf{z}_{kr}^2\right)^T$, where $\mathbf{z}_{kr}^2 = (z_{k1}^2, \ldots, z_{kr}^2)$.

The estimator derived in this way is expected to perform better than the estimator calibrated only on the first moment of the principal components, though calibration on the second moment of the PCs requires r additional calibration constraints.

4. Calibration on Estimated Principal Components

We have assumed that the values of the auxiliary variables are known for all units in the population but in practice, it often happens that these variables are only known for the sampled individuals, while their population totals are known. Here we present a way to perform principal components calibration when the auxiliary variables are only observed for the units belonging to the sample.

Let $\mathbf{\Gamma} = N^{-1}\mathbf{X}^T\mathbf{X}$ be the variance-covariance matrix, estimated by

$$\hat{\mathbf{\Gamma}} = \frac{1}{\hat{N}}\sum_{k\in s} d_k(\mathbf{x}_k - \hat{\overline{\mathbf{X}}})(\mathbf{x}_k - \hat{\overline{\mathbf{X}}})^T = \frac{1}{\hat{N}}\sum_{k\in s} d_k\mathbf{x}_k\mathbf{x}_k^T - \hat{\overline{\mathbf{X}}}\hat{\overline{\mathbf{X}}}^T, \tag{4.1}$$

where $\hat{N} = \sum_{k\in s} d_k$ and $\hat{\overline{\mathbf{X}}} = \hat{N}^{-1}\sum_{k\in s} d_k\mathbf{x}_k$. Let $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p \geq 0$ be the sorted eigenvalues of $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_p$ the corresponding orthonormal eigenvectors,

$$\hat{\mathbf{\Gamma}}\hat{\mathbf{v}}_j = \hat{\lambda}_j\hat{\mathbf{v}}_j, \quad j = 1, \ldots, p. \tag{4.2}$$

The $\hat{\lambda}_j$ and $\hat{\mathbf{v}}_j$ are the design-based estimators of λ_j and $\mathbf{v}_j$, respectively, for $j = 1, \ldots, p$. It is shown in Cardot et al. (2010) that, with large samples and under classical assumptions on the first and second order inclusion probabilities π_k, π_{kl} as well as on the variables $\mathcal{X}_j$, (see the assumptions (A1)−(A6) in Section 5), the estimators $\hat{\lambda}_j$ and $\hat{\mathbf{v}}_j$ are asymptotically design unbiased and consistent for λ_j and $\mathbf{v}_j$.

The unknown population principal components $\mathbf{Z}_j$ defined in (3.2) can be approximated as $\hat{\mathbf{Z}}_j = \mathbf{X}\hat{\mathbf{v}}_j$, with $\hat{\mathbf{Z}}_j = (\hat{z}_{kj})_{k\in U}$ only known for the units in the sample. Nevertheless, the population total $t_{\hat{\mathbf{Z}}_j} = \sum_{k\in U}\hat{z}_{kj}$ is known to be zero since $t_{\hat{\mathbf{Z}}_j} = t_{\mathbf{x}}^T\hat{\mathbf{v}}_j = 0$, $j = 1, \ldots, p$. The $\hat{\mathbf{Z}}_j$ are not exactly the principal components associated with the variance-covariance matrix $\hat{\mathbf{\Gamma}}$ because the original variables are centered in the population but not necessarily in the sample.

Consider now the first r estimated principal components, $\hat{\mathbf{Z}}_1, \ldots, \hat{\mathbf{Z}}_r$, corresponding to the r largest eigenvalues, $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_r \geq 0$, and suppose that $\hat{\lambda}_r > 0$. but, for ease of notation, we will use the same r. The estimated principal component (EPC) calibration estimator of t_y is $\hat{t}_{yw}^{\text{epc}}(r) = \sum_{k\in s} w_k^{\text{epc}}(r)y_k$, where the EPC calibration weights $w_k^{\text{epc}}, k \in s$ are the solution of the optimization problem (2.2) subject to the constraints $\sum_{k\in s} w_k^{\text{epc}}(r)\hat{\mathbf{z}}_{kr} = \sum_{k\in U}\hat{\mathbf{z}}_{kr}$, where $\hat{\mathbf{z}}_{kr}^T = (\hat{z}_{k1}, \ldots, \hat{z}_{kr})$ is the vector of values of $\hat{\mathbf{Z}}_j$, $j = 1, \ldots, r$ recorded for the kth unit. With the chi-square distance function Φ_s, the EPC calibration weights $w_k^{\text{epc}}(r)$ are given by

$$w_k^{\text{epc}}(r) = d_k - d_k\hat{\mathbf{z}}_{kr}^T\left(\sum_{\ell\in s} d_\ell\hat{\mathbf{z}}_{\ell r}\hat{\mathbf{z}}_{\ell r}^T\right)^{-1}(\hat{t}_{\hat{\mathbf{z}}_r d} - t_{\hat{\mathbf{z}}_r}), \tag{4.3}$$

where $\hat{t}_{\hat{\mathbf{z}}_r d} = \sum_{k\in s} d_k \hat{\mathbf{z}}_{kr}$ is the HT estimator of the total $t_{\hat{\mathbf{z}}_r} = \sum_{k\in U} \hat{\mathbf{z}}_{kr} = 0$. The EPC calibration estimator for t_y is $\hat{t}_{yw}^{\text{epc}}(r) = \sum_{k\in U} w_k^{\text{epc}}(r) y_k = \hat{t}_{yd} - \left(\hat{t}_{\hat{\mathbf{z}}_{\mathbf{r}} d} - t_{\hat{\mathbf{z}}_{\mathbf{r}}}\right)^T \hat{\boldsymbol{\gamma}}_{\hat{\mathbf{z}}}(r)$, where $\hat{\boldsymbol{\gamma}}_{\hat{\mathbf{z}}}(r) = \left(\sum_{k\in s} d_k \hat{\mathbf{z}}_{kr} \hat{\mathbf{z}}_{kr}^T\right)^{-1} \sum_{k\in s} d_k \hat{\mathbf{z}}_{kr} y_k$. The EPC calibration estimator can also be written with respect to the population totals of the original variables, $\mathcal{X}_1, \ldots, \mathcal{X}_p$, as $\hat{t}_{yw}^{\text{epc}}(r) = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{epc}}(r)$, where $\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{epc}}(r) = \widehat{\mathbf{G}}_r \hat{\boldsymbol{\gamma}}_{\hat{\mathbf{z}}}(r)$ and $\widehat{\mathbf{G}}_r$ is the $p \times r$ matrix whose jth column is equal to $\widehat{\mathbf{v}}_j$.

5. Some Asymptotic Properties of the Principal Components Calibration Estimators

We adopt in this section the asymptotic framework of Isaki and Fuller (1982), considering a sequence of growing and nested populations U_N with size N tending to infinity and a sequence of samples s_N of size n_N drawn from U_N according to the fixed-size sampling designs $p_N(s_N)$. The sequence of subpopulations is an increasing nested one, whereas the sample sequence is not. For simplicity of notation, we drop the subscript N in the following when there is no ambiguity. The number p_N of auxiliary variables, as well as the number r_N of principal components, is allowed to tend to infinity. We need some assumptions.

(A1) $\lim_{N\to\infty} n/N = \pi \in (0,1)$.

(A2) $\pi_k > \delta > 0$ for all $k \in U_N$; $\overline{\lim}_{N\to\infty} n \max_{k\neq l} |\pi_{kl} - \pi_k \pi_l| < \infty$.

(A3) There is a constant C_y such that for all N, $(1/N)\sum_{U_N} y_k^4 < C_y$.

(A4) The largest eigenvalue λ_{1N} of $\boldsymbol{\Gamma}_N$ is bounded, $\lambda_{1N} \le C_\lambda$.

(A5) There is a contant $c > 0$ and a non decreasing sequence of integers (r_N) such that, for all $N \ge N_0$, we have $\lambda_{r_N} \ge c$.

(A6) There is a constant C_4 such that, $\forall \mathbf{v} \in \mathbb{R}^{p_N}$ satisfying $\|\mathbf{v}\| = 1$, we have $N^{-1}\sum_{k\in U_N} |\langle \mathbf{x}_k, \mathbf{v}\rangle|^4 \le C_4$.

Conditions (A1), (A2) and (A3) are classical hypotheses for asymptotics in survey sampling. Condition (A4) is closely related to a moment condition on $\mathbf{x}_k$, for $k \in U_N$. If (A4) is fulfilled, $(1/N)\sum_{k\in U_N} \|\mathbf{x}_k\|^2 = \sum_{j=1}^{p_N} \lambda_{jN} \le C_\lambda p_N$. Assumption (A5) ensures that there is no identifiability issue for the sequence $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{pc}}(r_N)$ of regression coefficients defined at the population level. It only deals with λ_{r_N} and does not prevent λ_{p_N} from being equal to zero or from being very small. The conditions (A4) and (A6) indicate that the vectors $\mathbf{x}_k$ cannot be too concentrated in one direction (see Vershynin (2012) for examples in a classical statistical inference context). The proofs of Proposition 1 and Proposition 2 are given in a Supplementary file.

We first show that the estimator based on the true principal components is consistent and we give its asymptotic variance. Note that the assumption on the eigenvalues $\lambda_r > \lambda_{r+1} \geq 0$ ensures that there is no identifiability problem of the eigenspace generated by the eigenvectors associated to the r largest eigenvalues. The condition $r_N^3/n \to 0$ prevents the number of principal components from being too large and ensures that the remainder term, whose order is $r_N^{3/2}/n$, tends to zero and is negligible compared to the main term whose order is $1/\sqrt{n}$.

Proposition 1. *If* (A1)−(A6), $\lambda_r > \lambda_{r+1} \geq 0$ *and* $r_N^3/n \to 0$ *as* N *goes to infinity, then*

$$N^{-1}(\hat{t}_{yw}^{\text{pc}}(r_N) - t_y) = N^{-1}\left(\tilde{t}_{y,\mathbf{x}}^{\text{diff}}(r_N) - t_y\right) + O_p\left(\frac{r_N^{3/2}}{n}\right),$$

and

$$\tilde{t}_{y,\mathbf{x}}^{\text{diff}}(r_N) = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{pc}}(r_N)$$

satisfies

$$N^{-1}\left(\tilde{t}_{y,\mathbf{x}}^{\text{diff}}(r_N) - t_y\right) = O_p\left(\frac{1}{\sqrt{n}}\right).$$

The condition $r_N^3/n \to 0$ could certainly be relaxed for particular sampling designs with high entropy under additional moment assumptions. Note that the asymptotic variance of $\hat{t}_{yw}^{\text{pc}}(r)$ is given by $AV(\hat{t}_{yw}^{\text{pc}}(r)) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\left(y_k - \mathbf{x}_k^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{pc}}(r)\right)\left(y_l - \mathbf{x}_l^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{pc}}(r)\right)$.

Before stating a consistency result for calibration on estimated principal components, we introduce an additional condition on the spacing between adjacent eigenvalues.

(A7) There is a constant $c_\lambda > 0$ such that $\min_{j=1,\ldots,r_N+1}(\lambda_{jN} - \lambda_{j+1,N}) \geq c_\lambda r_N$.

This assumption ensures that the r_N largest eigenvalues are nearly equidistributed in $[c, C_\lambda]$.

Proposition 2. *If* (A1)−(A7) *hold, and* $p_N^3 r_N^3/n \to 0$ *as* N *goes to infinity, then*

$$N^{-1}(\hat{t}_{yw}^{\text{epc}}(r_N) - t_y) = N^{-1}\left(\tilde{t}_{y,\mathbf{x}}^{\text{diff}}(r_N) - t_y\right) + O_p\left(\frac{p_N^{3/2} r_N^{3/2}}{n}\right).$$

A more restrictive condition on how r_N may go to infinity is imposed when the principal components are estimated: $p_N^3 r_N^3/n \to 0$ ensures that the remainder term of order $p_N^{3/2} r_N^{3/2}/n$ is negligible compared to $1/\sqrt{n}$. If p_N is bounded, one gets back to classical $\sqrt{n}$-rates of convergence whether the population principal components are known or not.

If all the second-order inclusion probabilities $\pi_{k\ell}$ are strictly positive, the asymptotic variance of $\hat{t}_{yw}^{\mathrm{pc}}(r)$ can be estimated by the Horvitz-Thompson variance estimator for the residuals $y_k - \mathbf{x}_k^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r), k \in s$,

$$\widehat{Var}(\hat{t}_{yw}^{\mathrm{pc}}(r)) = \sum_{k\in s}\sum_{\ell\in s} \frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_{k\ell}} d_k d_\ell \left(y_k - \mathbf{x}_k^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)\right)\left(y_\ell - \mathbf{x}_\ell^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)\right),$$

while the asymptotic variance of $\hat{t}_{yw}^{\mathrm{epc}}(r)$ can be estimated by the Horvitz-Thompson variance estimator for the residuals $y_k - \mathbf{x}_k^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{epc}}(r), k \in s$:

$$\widehat{Var}(\hat{t}_{yw}^{\mathrm{epc}}(r)) = \sum_{k\in s}\sum_{\ell\in s} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} d_k d_\ell (y_k - \mathbf{x}_k^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{epc}}(r))(y_\ell - \mathbf{x}_\ell^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{epc}}(r)).$$

6. Partial Calibration on Principal Components

Our calibration estimators are not designed to give the exact finite population totals of the original variables $\mathcal{X}_j$, $j = 1, \ldots, p$. In practice, it is often desired to have this property satisfied for a few important socio-demographical variables such as sex, age, or socio-professional category.

We can adapt our method to fulfill this requirement. We split the auxiliary matrix $\mathbf{X}$ into two blocks: a first block $\tilde{\mathbf{X}}_1$ containing p_1 important variables, with p_1 small compared to p, and a second block $\tilde{\mathbf{X}}_2$ containing the remaining $p_2 = p - p_1$ variables. We have $\mathbf{X} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$. The constant term will generally belong to the first block of variables.

The goal is to get calibration weights such that the totals of the p_1 auxiliary variables in $\tilde{\mathbf{X}}_1$ are estimated exactly while the totals of the p_2 remaining variables are estimated only approximately. The idea is to calibrate directly on the auxiliary variables from $\tilde{\mathbf{X}}_1$ and on the first principal components of $\tilde{\mathbf{X}}_2$, after having taken into account the fact that the variables in $\tilde{\mathbf{X}}_1$ and all their linear combinations are perfectly estimated. For that, we introduce the matrix $\mathbf{I}_N$, the N-dimensional identity matrix, and $\mathbf{P}_{\tilde{\mathbf{X}}_1} = \tilde{\mathbf{X}}_1(\tilde{\mathbf{X}}_1^T\tilde{\mathbf{X}}_1)^{-1}\tilde{\mathbf{X}}_1^T$ the orthogonal projection onto the vector space spanned by the column vectors of matrix $\tilde{\mathbf{X}}_1$. We take $\mathbf{A} = \left(\mathbf{I}_N - \mathbf{P}_{\tilde{\mathbf{X}}_1}\right)\tilde{\mathbf{X}}_2$, the projection of $\tilde{\mathbf{X}}_2$ onto the orthogonal space spanned by the column vectors of $\tilde{\mathbf{X}}_1$. Matrix $\mathbf{A}$ represents the residual part of $\tilde{\mathbf{X}}_2$ that is not "calibrated". We define the residual covariance matrix $N^{-1}\mathbf{A}^T\mathbf{A} = N^{-1}\tilde{\mathbf{X}}_2^T\left(\mathbf{I}_N - \mathbf{P}_{\tilde{\mathbf{X}}_1}\right)\tilde{\mathbf{X}}_2$, and denote by $\tilde{\lambda}_1 \geq \ldots \tilde{\lambda}_{p_2}$ its eigenvalues and by $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_{p_2}$ the corresponding orthonormal eigenvectors. Consider $\tilde{\mathbf{Z}}_j = \mathbf{A}\tilde{\mathbf{v}}_j$, for $j = 1, \ldots, p_2$, the principal components of $\mathbf{A}$. The calibration variables are $(\tilde{\mathbf{X}}_1, \tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_r)$ of zero totals and the partial principal component (PPC) calibration estimator of t_y is $\hat{t}_{yw}^{\mathrm{ppc}}(r) = \sum_{k\in s} w_k^{\mathrm{ppc}}(r) y_k$, where the PPC calibration weights $w_k^{\mathrm{ppc}}(r)$, for $k \in s$, are the solution of the optimization problem (2.2) subject to $\sum_{k\in s} w_k^{\mathrm{ppc}}(r)\left(\tilde{\mathbf{x}}_k, \tilde{\mathbf{z}}_{kr}\right)^T = \sum_{k\in U}\left(\tilde{\mathbf{x}}_k, \tilde{\mathbf{z}}_{kr}\right)^T$, where $\tilde{\mathbf{x}}_k = (\tilde{x}_{k1}, \ldots, \tilde{x}_{kp_1})$ is the vector of the values of the variables in $\tilde{\mathbf{X}}_1$ and $\mathbf{z}_{kr}^T = (\tilde{z}_{k1}, \ldots, \tilde{z}_{kr})$ is the vector whose elements are the partial principal components $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_r$ for unit k. With a different point of view, Breidt and Chauvet (2012) use, at the sampling stage, similar ideas to perform penalized balanced sampling.

If we only have a sample s at hand and we know the totals of all the calibration variables, let $\tilde{\mathbf{X}}_{s,1}$ (resp. $\tilde{\mathbf{X}}_{s,2}$) be the $n \times p_1$ (resp. $n \times p_2$) matrix containing the observed values of the auxiliary variables. We can estimate $\mathbf{A}$ by $\hat{\mathbf{A}} = \left(\mathbf{I}_n - \hat{\mathbf{P}}_{\tilde{\mathbf{X}}_{s,1}}\right)\tilde{\mathbf{X}}_{s,2}$ where $\hat{\mathbf{P}}_{\tilde{\mathbf{X}}_{s,1}} = \tilde{\mathbf{X}}_{s,1}\left(\tilde{\mathbf{X}}_{s,1}^T\mathbf{D}_s\tilde{\mathbf{X}}_{s,1}\right)^{-1}\tilde{\mathbf{X}}_{s,1}^T\mathbf{D}_s$, is the estimation of the projection onto the space generated by the columns of $\tilde{\mathbf{X}}_1$, and $\mathbf{D}_s$ is the $n \times n$ diagonal matrix with diagonal elements $d_k,\ k \in s$. Then, we can perform the principal components analysis of the projected sampled data corresponding to the variables belonging to the second group and compute the estimated principal components associated to the r largest eigenvalues as in Section 4. At last, the total estimator is calibrated on the totals of the variables in $\mathbf{X}_1$ and the first r estimated principal components.

7. Application to the Estimation of the Total Electricity Consumption

Description of the Data

We illustrate on data from the Irish Commission for Energy Regulation (CER) Smart Metering Project that was conducted in 2009−2010 (CER, 2011). The data are available on request at:
`http://www.ucd.ie/issda/data/commissionforenergyregulation/`. In this project, which focuses on energy consumption and energy regulation, about 6,000 smart meters were installed to collect every half an hour, over a period of about two years, the electricity consumption of Irish residential and business customers.

We evaluate the interest of employing dimension reduction techniques based on PCA by considering a period of 14 consecutive days and a population of $N = 6{,}291$ smart meters (households and companies). Thus, we have for each unit k in the population $(2 \times 7) \times 48 = 672$ measurement instants and we denote by $y_k(t_j), j = 1, \ldots 672$ the data corresponding to unit k, where $y_k(t_j)$ is the electricity consumption (in kW) associated to smart meter k at instant t_j. We consider a multipurpose setting and aim to estimate the mean electricity consumption of each day of the second week. For each day ℓ of the week, with $\ell \in \{1, \ldots, 7\}$, the outcome variable is $y_{k\ell} = \sum_{j=336+(\ell-1)\times 48}^{336+\ell\times 48} y_k(t_j)$ and our target is the corresponding population total, $t_\ell = \sum_{k\in U} y_{k\ell}$.

The auxiliary information is the load electricity curve of the first week. This means that we have $p = 336$ auxiliary variables, the consumption electricity levels at each of the $p = 336$ half hours of the first week. The condition number of the matrix $N^{-1}\mathbf{X}^T\mathbf{X}$, the ratio λ_1/λ_{336}, is equal to 67,055.78. The matrix is ill-conditioned and there may exist strong correlations between some of the variables used for calibration. Indeed, the first principal component explains about 63% of the variance of the 336 original variables, and about 83% of the total variability of the data is preserved by projection onto the subspace spanned by the first ten principal components.

Comparison of the estimators

To make comparisons, we drew I =1,000 samples of size $n = 600$ (the sampling fraction is about 0.095) according to a simple random sampling design without replacement and we estimated the total consumption t_ℓ over each day ℓ of the second week with the Horvitz-Thompson estimators, the calibration estimators, denoted by $\hat{t}_{\ell w}$, that take account of all the $p = 336$ auxiliary variables plus the intercept term, and the estimators calibrated on the principal components in the population (resp. in the sample) plus the constant term, denoted by $\hat{t}^{\mathrm{pc}}_{\ell w}(r)$ (resp. $\hat{t}^{\mathrm{epc}}_{\ell w}(r)$), for different values of the dimension r.

When performing principal components calibration, the dimension r plays the role of a tuning parameter. We also studied the performances of an automatic and data-drive simple rule for selecting the dimension r, selecting the largest dimension $\hat{r}$ such that all the calibrated principal component weights remain positive. This selection strategy is the analogue of the strategy suggested in Bardsley and Chambers (1984) for choosing the tuning parameter in a ridge regression context.

The distribution of the coefficient of variation (CV) of the calibration weights for the I=1,000 Monte Carlo experiments is presented in Figure 1 for different values of the dimension r. These weights do not depend on the variable of interest. It is clearly seen that those calibration weights have larger dispersion as the number of principal components used for calibration increases. Calibrating with a large number of correlated auxiliary variables may lead to instable estimations and to a lack of robustness with respect to measurement errors or misspecification in the data bases. When all the auxiliary variables were used for calibration, around 25 % of the sampling weights took negatives values, generally not desirable.

Our benchmarks were the estimators $\hat{t}_{\ell w}$ calibrated on all the $p = 336$ auxiliary variables. For each day ℓ, the performances of an estimator $\widehat{\theta}$ of the total t_ℓ were measured by considering the relative mean squared error,

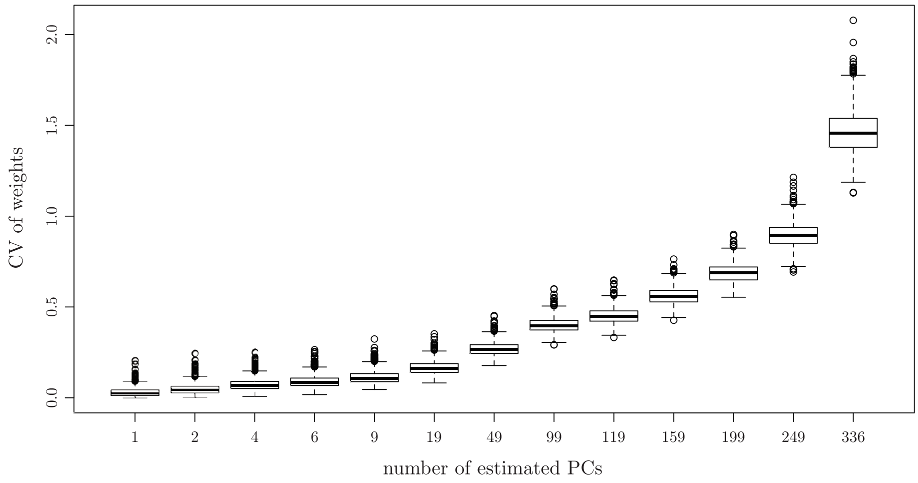

Figure 1. Distribution of the coefficient of variation (CV) of the sampling weights for different values of the dimension r. The sample size is $n = 600$.

$$R_\ell(\widehat{\theta}) = \frac{\sum_{i=1}^{I}(\widehat{\theta}^{(i)} - t_\ell)^2}{\sum_{i=1}^{I}(\widehat{t}_{\ell w}^{(i)} - t_\ell)^2}, \tag{7.1}$$

better estimators corresponding to small values of criterion $R_\ell(\widehat{\theta})$.

The values of this relative error for several values of r, as well as for the estimators obtained with the data-driven dimension selection, are given in Table 1. This relative error was also computed for the ridge-type estimators derived with the sampling weights w^{pen} given in Section 2 and a penalty $\hat{\lambda}$ chosen to be the smallest value of λ such that all the resulting weights remain positive.

First the naive Horvitz-Thompson estimator can be greatly improved, for all the days of the week, by considering an over-calibration estimator which takes account of all the (redundant) auxiliary information. Indeed, the mean square error of the HT estimator is between five and fourteen times larger than the MSE of this reference estimator. Reducing the number of effective auxiliary variables through principal components, estimated on the sample or deduced from the population, can still improve estimation compared to calibration on all the variables, and permits to divide by two the MSE. Another remarkable feature is the stability of the principal components calibration techniques with respect to the choice of the dimension r. Choosing between 5 and 100 principal components here permits to divide by two, for all the outcome variables, the MSE compared to the calibration estimator based on the whole auxiliary information.

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, explaining in each case about 85%

Table 1. Comparison of the mean relative mean squared errors of the different estimators, according to criterion (7.1).

		Days						
Estimators		monday	tuesday	wednesday	thursday	friday	saturday	sunday
Horvitz-Thompson		14.4	13.9	11.8	10.8	12.5	6.4	5.4
$\hat{t}^{\text{pc}}_{\ell w}$	$r = 1$	0.65	0.62	0.50	0.47	0.64	1.17	1.57
$\hat{t}^{\text{pc}}_{\ell w}$	$r = 2$	0.64	0.62	0.50	0.47	0.57	0.80	0.63
$\hat{t}^{\text{pc}}_{\ell w}$	$r = 5$	0.52	0.47	0.40	0.50	0.51	0.53	0.52
$\hat{t}^{\text{pc}}_{\ell w}$	$r = 50$	0.50	0.50	0.43	0.44	0.54	0.48	0.48
$\hat{t}^{\text{pc}}_{\ell w}$	$r = 100$	0.57	0.60	0.59	0.51	0.58	0.60	0.64
$\hat{t}^{\text{pc}}_{\ell w}$	$r = 200$	0.60	0.64	0.58	0.66	0.69	0.68	0.63
$\hat{t}^{\text{pc}}_{\ell w}$	$r = 300$	0.82	0.85	0.83	0.86	0.84	0.85	0.87
$\hat{t}^{\text{epc}}_{\ell w}$	$r = 1$	0.75	0.73	0.61	0.56	0.73	1.23	1.59
$\hat{t}^{\text{epc}}_{\ell w}$	$r = 2$	0.66	0.64	0.53	0.50	0.61	0.85	0.74
$\hat{t}^{\text{epc}}_{\ell w}$	$r = 5$	0.53	0.47	0.40	0.41	0.53	0.59	0.53
$\hat{t}^{\text{epc}}_{\ell w}$	$r = 50$	0.45	0.46	0.40	0.41	0.48	0.46	0.47
$\hat{t}^{\text{epc}}_{\ell w}$	$r = 100$	0.46	0.47	0.42	0.45	0.52	0.49	0.50
$\hat{t}^{\text{epc}}_{\ell w}$	$r = 200$	0.57	0.55	0.51	0.58	0.62	0.60	0.57
$\hat{t}^{\text{epc}}_{\ell w}$	$r = 300$	0.78	0.80	0.77	0.84	0.80	0.81	0.83
$\hat{t}^{\text{pc}}_{\ell w}$	$\hat{r}$, $w(\hat{r}) > 0$	0.51	0.49	0.41	0.41	0.52	0.55	0.50
$\hat{t}^{\text{epc}}_{\ell w}$	$\hat{r}$, $w(\hat{r}) > 0$	0.49	0.48	0.41	0.40	0.50	0.53	0.49
Ridge Calibration	$\hat{\lambda}$	0.44	0.46	0.40	0.41	0.48	0.48	0.43

of the variance of the original variables. As expected, the variability of the number of selected components was slightly larger when considering calibration on the estimated principal components (interquartile range of 26 versus 17 for the population principal components).

As seen in Table 1, the performances of the resulting estimators are good and comparable to the estimators based on ridge calibration with a selection rule for λ based on the same principle. The advantage of the principal components is that it permits to divide by more than 15 the final number of effective variables used for calibration and it can directly be used in classical survey sampling softwares.

8. Discussion and Concluding Remarks

Some asymptotic justifications of this dimension reduction technique are given with the number p_N of auxiliary variables and the number r_N of principal components used for calibration allowed to grow to infinity as the population size N increases. Our conditions on the asymptotic behavior of r_N appear to be rather restrictive and could probably be relaxed (see for example the results presented in Vershynin (2012) on the estimation of covariance matrices for independent observations in an infinite population). However, this would require exponential inequalities for Horvitz-Thompson estimators to very accurately control their deviations around their target.

Borrowing ideas from Marx and Smith (1990) and Wu and Sitter (2001), it would be not too difficult to extend our principal component calibration approach to deal with non-linear model calibration.

Supplementary Materials

The proofs of Propositions 1 and 2 and additional results on the electricity data are given in the online supplementary material.

Acknowledgements

The authors thank an associate editor and the two anonymous referees for comments and suggestions, and more particularly for suggesting the data-driven rule for selecting the number of components.

References

Bardsley, P. and Chambers, R. (1984). Multipurpose estimation from unbalanced samples. *Appl. Statist.* **33**, 290-299.

Beaumont, J.-F. and Bocci, C. (2008). Another look at ridge calibration. *Metron - Internat. J. Statist.* LXVI, 5-20.

Breidt, J. F. and Chauvet, G. (2012). Penalized balanced sampling. *Biometrika* **99**, 945-958.

Breidt, J. F. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Ann. Statist.* **28**, 1023-1053.

Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2010). Properties of design-based functional principal components analysis. *J. Statist. Plann. Inference* **140**, 75-91.

Cassel, C., Särndal, C.-E. and Wretman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615-620.

Chambers, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *J. Official Statist.* **12**, 3-32.

Chambers, R., Skinner, C. and Wang, S. (1999). Intelligent calibration. *Bull. Inst. Internat. Statist. Institute* **58**, 321-324.

Chen, J., Sitter, R. R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* **89**, 230-237.

Clark, R. G. and Chambers, R. L. (2008). Adaptive calibration for prediction of finite population totals. *Surv. Methodol.* **34**, 163-172.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87**, 376-382.

Goga, C., Shehzad, M. A. and Vanheuverzwyn, A. (2011). Principal component regression with survey data. Application on the French media audience. *Proceedings of the* 58*th World Statistics Congress of the International Statistical Institute*, Dublin, Ireland, 3847-3852.

Goga, C. and Shehzad, M. A. (2014). A note on partially penalized calibration. *Pakistan J. Statist.* **30**, 429-438.

Guggemos, F. and Tillé, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *J. Statist. Plann. Inference* **140**, 3199-3212.

Gunst, R. F. and Mason, R. L. (1977). Biased estimation in regression: an evaluation using mean squared error. *J. Amer. Statist. Assoc.* **72**, 616-628.

Hoerl, E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55-67.

Isaki, C. and Fuller, W. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77**, 49-61.

Jolliffe, I. T. (2002). *Principal Components Analysis.* Second Edition, Springer- Verlag. New York.

Marx, B. D. and Smith, E. P. (1990). Principal component estimation for generalized linear regression. *Biometrika* **77**, 23-31.

Rao, J. N. K. and Singh, A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, 57-65. American Statistical Association.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Surv. Methodol.* **33**, 99-119.

Silva, P. L. N. and Skinner, C. (1997). Variable selection for regression estimation in finite populations. *Surv. Methodol.* **23**, 23-32.

Swold, S., Sjöström, M. and Eriksson, L. (2001). PLS-regression : a basic tool of chemometrics. *Chemometr. Intell. Lab.* **58**, 109-130.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Théberge, A. (2000). Calibration and restricted weights. *Surv. Methodol.* **26**, 99-107.

Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *J. Theoret. Probab.* **25**, 655-686.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.* **96**, 185-193.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

IMB, UMR CNRS 5584, Université de Bourgogne Franche Comté, 9 avenue Alain Savary, Dijon, France.

E-mail: herve.cardot@u-bourgogne.fr

IMB, UMR CNRS 5584, Université de Bourgogne Franche Comté, 9 avenue Alain Savary, Dijon, France.

E-mail: camelia.goga@u-bourgogne.fr

Bahauddin Zakariya University, Bosan Road, Multan, Pakistan.

E-mail: mdjan708@gmail.com

(Received November 2014; accepted December 2015)