Calibration and Partial Calibration on Principal Components when the Number of Auxiliary Variables is Large

Cardot, H.(1), Goga, C.(1) and Shehzad, M.-A.(1),(2)

(1) IMB, Université de Bourgogne, DIJON, FRANCE
(2) Bahauddin Zakariya University, MULTAN, PAKISTAN

Abstract: When the number of auxiliary variables is large, calibration on all the variables may lead to estimators of totals that are not as effective as the Horvitz-Thompson estimator even if this simple estimator does not take at all account of the available auxiliary information. We propose in this paper 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 the others, the method can be adapted to provide an exact calibration on these important variables. Finally, the interest of considering a small number of calibration variables built thanks to principal components analysis is illustrated on the estimation of total electricity consumption over one week with the help of 336 auxiliary variables consisting of the past consumption measured every half an hour over the previous week.

Key words and phrases: calibration on estimated auxiliary variables; dimension reduction; model-assisted estimation; partial calibration; partial least squares; penalized calibration; variance approximation.

1. Introduction

With the spread of automatic process for data collection as well as increasing storage capacities, it is not unusual anymore to have to analyze data coming form very large surveys with many auxiliary variables. In this context, calibration on all the auxiliary variables may lead to estimators whose performances are worse than the simple Horvitz-Thompson estimator even if this latter estimator does not take into account at all of the auxiliary information (see e.g. Silva and Skinner, 1997).

M.-A. Shehzad was supported by the HEC (Higher Education Commission) of Pakistan
Calibration over a very large number of auxiliary variables has been called over-calibration by Guggemos and Tillé (2010). Several difficulties arise in this context such as instability of the calibration weights, variance inflation and there are different ways of dealing with this issue. 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; Skinner and Silva, 1997; Chambers, Skinner and Wang, 1999 and Clark and Chambers, 2008). Another way is to relax the calibration constraints, meaning that the too restrictive requirement of being exactly calibrated is dropped off and replaced by the requirement of being only approximately calibrated. 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, 2009) and Théberge (2000) in a design-based (or model-assisted) setting.

We present in this paper another way of dealing with this issue. Our estimator is based on dimension reduction via principal components calibration. In multivariate statistics, the principal component analysis (PCA) is one of most popular technique 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), which are linear combinations of the initial variables with the largest variance. PCA may be used in a regression context, called principal component regression (PCR) especially when the regressors are (almost) linearly related. 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. However, as far as we know, its application in a survey sampling setting is new. The method we suggest consists in reducing the number of auxiliary variables by considering a small number of PC’s and by performing calibration on these new synthetic variables. The method is easy to put into practice with a software computer used for performing calibration, such as CALMAR used at the French National Statistical Institut (Insee) since, with centered data, these new calibration variables are also centered.

Note that a natural concurrent to principal components regression is partial
least squares (PLS) which is also a popular dimension reduction technique that can be useful when there is a large number of auxiliary variables that are highly correlated (see for example Swold et al. 2001 for a pedagogical presentation of PLS). As noted in Frank and Friedman (1993) it often has similar prediction errors to principal components regression. Furthermore, a major drawback of calibration on PLS scores would be that the calibration weights would depend on the variable of interest. This is generally not a desired property since then different sampling weights should be considered when there are different variables of interest as it is the case in multipurpose surveys. Consequently PLS calibration has not been considered in this work.

The paper is structured as follows: we briefly recall in Section 2 the calibration method as it was suggested by Deville and Särndal (1992) and the problems which may arise when the number of auxiliary variables is large. We introduce in Section 3 the suggested method and we give a model-assisted interpretation. Interestingly, with the chi-squared distance, the calibration estimator on PC’s may be written as a GREG-type estimator with respect to the initial auxiliary variables but with a modified regression coefficient estimator. Usually, some auxiliary variables often play a role that is more important than the others and we show how the method can be adapted to provide an exact calibration on these important variables. Bardsley and Chambers (1984) and Guggemos and Tillé (2010) were confronted with a similar situation in the context of ridge-type estimators. When the auxiliary information is not complete, i.e. the values of the auxiliary variables are only known in the sample, we first estimate the PC’s and then we 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 as well as on estimated PC’s are asymptotical unbiased and consistent. We show also that these estimators are asymptotically equivalent to the generalized difference estimator computed with a PCR regression coefficient estimator. Variance estimators are also presented. Our method is illustrated in Section 7 on the estimation of total electricity consumption over one week with the help of the past consumption measured every half an hour over the previous week. Some practical guidelines for choosing the number of PC’s are given. Finally, a brief
Section 8 gives some concluding remarks as well as some directions that would deserve further investigation. The proofs are gathered in an Appendix.

2. Estimation of finite population totals with calibration over a large number of auxiliary variables

We consider the finite population $U = \{1, \ldots, k, \ldots, N\}$ and we wish to estimate the total

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

where $y_k$ is the value of the variable of interest $Y$ for the $k$th unit. Let $s$ be a random sample drawn from $U$ according to a sampling design that assigns to unit $k$ a known inclusion probability $\pi_k = \Pr(k \in s) > 0$. 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, $X_1, \ldots, X_p$, and let $x_k^T = (x_{k1}, \ldots, x_{kp})$ be the transposed vector whose elements are the values of the auxiliary variables for the $k$th unit. The calibration method has been developed by Deville and Särndal (1992) to use as effectively as possible the known population totals of $X_j, j = 1, \ldots, p$ at the estimation stage. The calibration estimator of $t_y$ is a weighted estimator

$$\hat{t}_{yw} = \sum_{k \in s} w_k y_k,$$  \hspace{1cm} (2.1)

where the calibration weights $w_k$ are chosen so that they are as close as possible to the initial sampling weights $d_k$, according to some distance $\Phi$. Estimator (2.1) also satisfies the calibration constraints, that is to say

$$\sum_{k \in s} w_k x_k = t_x,$$  \hspace{1cm} (2.2)
where \( t_x = \sum_{k \in U} x_k \) is the vector whose elements are the known totals of \( X_j \) for \( j = 1, \ldots, p \). Several distance functions \( \Phi_s \) have been studied in Deville and Särndal (1992). Under weak regularity assumptions these authors have shown that all resulting estimators are asymptotically equivalent to the one obtained by minimizing the chi-square distance function

\[
\Phi_s(w) = \sum_{k \in s} \frac{(w_k - d_k)^2}{q_k d_k},
\]

(2.3)

where \( w = (w_k, k \in s) \) is the vector of weights assigned to each unit in the sample and the \( q_k \)'s are known positive constants used to control the variability of the observations and are unrelated to \( d_k \). A common use in the applications is to consider uniform weights \( q_k = 1 \) for all units \( k \) and we will suppose, without loss of generality, that \( q_k = 1 \) in the following. The resulting calibration weights \( w_k, k \in s \), are

\[
w_k = d_k - d_k x_k^T \left( \sum_{k \in s} d_k x_k x_k^T \right)^{-1} (\hat{t}_{xd} - t_x)
\]

(2.4)

and the corresponding calibration estimator \( \hat{t}_w \) is

\[
\hat{t}_w = \hat{t}_{yd} - (\hat{t}_{xd} - t_x)^T \hat{\beta}_x,
\]

(2.5)

where \( \hat{t}_{xd} = \sum_{k \in s} d_k x_k \) is the HT estimator of \( t_x \) and \( \hat{\beta}_x = (\sum_{k \in s} d_k x_k x_k^T)^{-1} \sum_{k \in s} d_k x_k y_k \).

It can also be shown that the previous calibration estimator is in fact equal to the generalized regression estimator (GREG) which is derived by assuming a linear regression model between the study variable \( Y \) and the auxiliary variables \( X_1, \ldots, X_p \),

\[
x : y_k = x_k^T \beta + \varepsilon_k,
\]

(2.6)

where \( \varepsilon = (\varepsilon_k, k \in U) \) is a centered random vector with a diagonal variance matrix, whose diagonal elements are equal to \( 1/q_k \).

Under mild regularity assumptions, Deville and Särndal (1992) have proven that the calibration estimator \( \hat{t}_w \) and the difference estimator

\[
\hat{t}_{y,x}^{\text{diff}} = \hat{t}_{yd} - (\hat{t}_{xd} - t_x)^T \hat{\beta}_x,
\]

(2.7)
where \( \tilde{\beta}_x = (\sum_{k \in U} x_k x_k^T)^{-1} \sum_{k \in U} x_k y_k \), have the same asymptotic distribution. More precisely, we have

\[
\frac{1}{N} (\hat{t}_{yw} - t_y) = \frac{1}{N} (\hat{t}_{y,x} - t_y) + o_p(n^{-1/2}).
\]

(2.8)

As a result, the asymptotic variance of \( \hat{t}_w \) is

\[
AV(\hat{t}_{yw}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) d_k d_l (y_k - x_k^T \tilde{\beta}_x)(y_l - x_l^T \tilde{\beta}_x).
\]

(2.9)

The calibration estimator will improve the HT estimator, namely \( AV(\hat{t}_{yw}) \leq V(\hat{t}_{yd}) \), if the predicted values \( x_k^T \tilde{\beta}_x \) are close enough to the \( y_k \)'s, that is to say if the model \( \xi \) stated in (2.6) explains sufficiently well the variable of interest. Nevertheless, when a very large number \( p \) of auxiliary variables is used, this result is no longer true as it was remarked by Silva and Skinner (1997) in a simulation study. More precisely, it has been shown under classical hypotheses (see Chauvet and Goga, 2013) that

\[
\frac{1}{N} (\hat{t}_{xd} - t_x)^T (\hat{\beta}_x - \tilde{\beta}_x) = O_p \left( \frac{p^2}{n} \right),
\]

which means that the error between the calibration estimator and the generalized difference estimator depends on the number \( p \) of auxiliary variables, and thus, if \( p \) is large, approximation (2.8) should be written,

\[
\frac{1}{N} (\hat{t}_{yw} - t_y) = \frac{1}{N} (\hat{t}_{y,x} - t_y) + O_p \left( \frac{p^2}{n} \right).
\]

Consequently, performing calibration with as many auxiliary variables as possible may not be the best strategy and this may lead to calibrated estimators whose performances are poor. The asymptotic approximation to the variance of the calibration estimator given in (2.9) is valid only when \( p = O(n^a) \) with \( a < 1/2 \), so that the number of auxiliary variables should not be too large compared to the sample size. Otherwise, the variability of \( (\hat{t}_{xd} - t_x)^T (\hat{\beta}_x - \tilde{\beta}_x) \) is not negligible anymore and it must also be taken into account in the variance approximation.

Usually, with calibration, the user requires that the weight ratios \( w_k/d_k \) are restricted to lie between pre-specified lower and upper bounds, called also range restrictions. The main reason for doing that is to avoid negative or extremely large weights. Nevertheless, the calibration weights derived for many auxiliary
variables may be very unstable and very large, implying that the range restrictions are more difficult to be satisfied. To cope with this issue, several modifications of the distance have been suggested in the literature (Deville and Särndal, 1992; Jayasuriya and Valliant, 1996 and Singh and Mohl, 1996), but as Beaumont and Bocci (2008) remarked, these methods “are all iterative and may not yield a solution even if the range restrictions are mild”.

Note also that in the extreme case in which matrix $\sum_{k \in s} d_k x_k x_k^T$ is not a full rank matrix, calibration weights cannot be computed directly with (2.4) because there is an identifiability issue that is due to multicollinearity and a generalized inverse of previous matrix should be used. Théberge (1999) considered the minimum norm least squares method and the Moore-Penrose inverse matrix to derive weights in presence of multicollinearity among regressors.

One way to circumvent the problems due to over-calibration is to relax the calibration constraints, meaning that the too restrictive requirement of being exactly calibrated as in (2.2) is dropped off and replaced by the requirement of being only approximately calibrated. Then, the deviation between $\sum_{k \in s} w_k x_k$ and $\sum_{k \in U} x_k$ is controlled by means of a penalty. A class of penalized estimators was suggested by Bardsley and Chambers (1984) in a model-based setting and extended later by Chambers (1996) and by Rao and Singh (1997, 2009) in a design-based (or model-assisted) setting. These approaches lead to a class of model-based or GREG-type estimators that use regression coefficients estimated by ridge-type estimators.

In a design-based setting, the vector penalized calibration weights $w_{\text{pen}}$ is obtained as the solution of the following penalized minimization problem:

$$w_{\text{pen}}(\lambda) = \arg \min_w \Phi_s(w) + \lambda^{-1} (\hat{t}_w - t_x)^T C (\hat{t}_w - t_x),$$  \hspace{1cm} (2.10)

where $\hat{t}_w = \sum_{k \in s} w_k x_k$, $C = \text{diag}(c_j)_{j=1}^{p+1}$, and $c_j \geq 0$ is a user-specified cost associated with the $j$th calibration constraint. The scale factor $\lambda > 0$ controls the trade off between the calibration constraints and the distance of the calibration weights from the original sampling weights. With the chi-square distance, the solution of (2.10) is, for $k \in s$,

$$w_{\text{pen}}^k(\lambda) = d_k - d_k x_k^T \left( \sum_{k \in s} d_k x_k x_k^T + \lambda C^{-1} \right)^{-1} (\hat{t}_d - t_x),$$  \hspace{1cm} (2.11)
and the penalized calibration estimator is a GREG type estimator, as given in (2.5), whose regression coefficients are estimated by a ridge-type estimator (Hoerl and Kennard, 1970),

\[
\hat{\beta}_x(\lambda) = \left( \sum_{k \in s} d_k x_k x_k^T + \lambda C^{-1} \right)^{-1} \sum_{k \in s} d_k x_k y_k.
\]

Beaumont and Bocci (2008) considered the optimization problem (2.10) with a general distance. Considering a quadratic penalty in (2.10) leads to adding the diagonal matrix \( \lambda C^{-1} \) to \( \sum_{k \in s} d_k x_k x_k^T \). One can remark that with an infinite cost \( c_j \) in (2.10) or \( c_j = 0 \) in (2.11), we obtain that the \( j \)th calibration constraint is satisfied exactly. Note that different algorithms have been studied in the literature to select good values, with data driven procedures, of the ridge parameter \( \lambda \) (see Barsley and Chambers, 1984; Beaumont & Bocci, 2008 and Guggemos & Tillé, 2010). As noted in Barsley and Chambers (1984), the risk of having negative weights (in the case of the chi-square distance) is greatly reduced by using penalized calibration. In a calibration approach, the penalized estimator is design-consistent for any fixed value of \( \lambda \) and its asymptotic variance is equal to the variance of the generalized difference estimator (2.7) with a ridge-type regression estimator \( \hat{\beta}_x(\lambda) = \left( \sum_{k \in U} x_k x_k^T + \lambda C^{-1} \right)^{-1} \sum_{k \in U} x_k y_k \).

3. Calibration on Principal Components

We consider in this work another class of approximately calibrated estimators which are based on dimension reduction through principal components analysis (PCA). In multivariate statistics PCA is one of most 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), which 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 of 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 et al. (2011) and Shehzad (2012).

**Complete Auxiliary Information**

We suppose without loss of generality that the auxiliary variables are centered, namely $\frac{1}{N} \mathbf{t}_X = 0$ and to avoid heavy notations we do not include an intercept term in the model. Note that in applications this intercept term should be included. We also suppose that the auxiliary information is complete, that is to say the $p$-dimensional vector $\mathbf{x}_k$ is known for all the 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 initial variables $X_1, \ldots, X_p$ is given by $N^{-1}\mathbf{X}^T \mathbf{X}$. 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 $j$th principal component, denoted by $Z_j$, is defined as follows

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

We will only consider the first $r$ (with $r < p$) principal components, $Z_1, \ldots, Z_r$, which correspond to the $r$ largest eigenvalues. In a survey sampling framework, the goal is not to give interpretations of these new variables $Z_1, \ldots, Z_r$ as it is the custom in PCA. These variables serve as a tool to obtain calibration weights which are more stable than the calibration weights that would have been obtained with the whole set of auxiliary variables.

More exactly, we want to find the principal component (PC) calibration estimator

$$\hat{\rho}_{yw}^{pc}(r) = \sum_{k \in s} w_k^{pc}(r) y_k,$$

where the vector of PC calibration weights, which depends on the number $r$ of principal components used for calibration, is the solution of the following optimization problem

$$w^{pc}(r) = \arg \min_{w} \Phi_s(w),$$

subject to

$$\sum_{k \in s} w_k^{pc}(r) z_{kr} = \sum_{k \in U} z_{kr},$$
where $\mathbf{z}_{kr}^T = (z_{k1}, \ldots, z_{kr})$ is the vector containing the values of the $r$ first PCs computed for the $k$-th individual. Considering the chi-square distance function $\Phi_s$, defined in (2.3), the PC calibration weights $w^\text{pc}_k(r)$’s are given by

$$w^\text{pc}_k(r) = d_k - d_k \mathbf{z}_{kr}^T \left( \sum_{k \in s} d_k \mathbf{z}_{kr} \mathbf{z}_{kr}^T \right)^{-1} \left( \hat{t}_{z,d} - t_{z,r} \right), \quad (3.3)$$

where $\hat{t}_{z,d} = \sum_{k \in s} d_k \mathbf{z}_{kr}$ is the HT estimator of the total $t_{z,r} = (0, \ldots, 0)$ since we have supposed that the original variables have mean zero so that the principal components are also centered variables.

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

$$\hat{t}^\text{pc}_{yw}(r) = \sum_{k \in s} w^\text{pc}_k(r) y_k = \hat{t}_{yd} - \left( \hat{t}_{z,r} - t_{z,r} \right)^T \hat{\gamma}_z(r), \quad (3.4)$$

where

$$\hat{\gamma}_z(r) = \left( \sum_{k \in s} d_k \mathbf{z}_{kr} \mathbf{z}_{kr}^T \right)^{-1} \sum_{k \in s} d_k \mathbf{z}_{kr} y_k. \quad (3.5)$$

The PC calibration estimator $\hat{t}^\text{pc}_{yw}(r)$ depends on the number $r$ of the PC variables and we can note that if $r = 0$, that is to say if we do not take auxiliary information into account, then $\hat{t}^\text{pc}_{yw}(0)$ 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 which takes account of all the auxiliary variables.

### A Model-Assisted Point of View

Consider again the superpopulation model $\xi$ presented in (2.6) and denote by $\mathbf{G} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$ the matrix whose $j$th column is the $j$th eigenvector $\mathbf{v}_j$. Model $\xi$ may be written in the equivalent form

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

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

$$\xi_r: \quad y_k = \mathbf{z}_{kr}^T \gamma(r) + \varepsilon_{kr}, \quad (3.6)$$
 Calibration on Principal Components

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

$$\hat{\gamma}_z(r) = \left( \sum_{k \in U} z_{kr} z_{kr}^T \right)^{-1} \sum_{k \in U} z_{kr} y_k \quad (3.7)$$

which in turn can be estimated, on a sample $s$, by the design-based estimator $\hat{\gamma}_z(r)$ given by (3.5). We can see now that the PC calibration estimator given in (3.4) is in fact equal to a GREG-type estimator assisted by the reduced model $\xi_r$ described in (3.6). Note also that since the principal components are centered and uncorrelated, matrix $\left( \sum_{k \in U} z_{kr} z_{kr}^T \right)$ is diagonal, with diagonal elements $(\lambda_1 N, \ldots, \lambda_r N)$.

When there is strong multicollinearity among the auxiliary variables, it is well known that the ordinary least squares estimator of $\beta$

$$\hat{\beta}_x = \left( \frac{1}{N} X^T X \right)^{-1} \frac{1}{N} X^T y$$

is very sensitive to small changes in $x_k$ and $y_k$ and it has a very large variance (see e.g. Hoerl and Kennard, 1970). To see better how small eigenvalues may affect $\hat{\beta}_x$, Gunst and Mason (1977) write previous estimator as follows:

$$\hat{\beta}_x = \left( \sum_{j=1}^{p} \frac{1}{\lambda_j} v_j v_j^T \right) \left( \frac{1}{N} \sum_{k \in U} x_k y_k \right) = \sum_{j=1}^{p} \frac{1}{\lambda_j} \left[ v_j^T \left( \frac{1}{N} \sum_{k \in U} x_k y_k \right) \right] v_j.$$

Approximating the covariance matrix $\frac{1}{N} X^T X$ by the rank $r$ matrix $\left( \sum_{j=1}^{r} \lambda_j v_j v_j^T \right)$ leads to consider the following approximation to the regression estimator that is based on the $r$ first principal components,

$$\hat{\beta}_x^{pc} (r) = G_r \hat{\gamma}_z(r),$$

$$= \sum_{j=1}^{r} \frac{1}{\lambda_j} \left[ v_j^T \left( \frac{1}{N} \sum_{k \in U} x_k y_k \right) \right] v_j \quad (3.8)$$
where \( G_r = (v_1, \ldots, v_r) \). This means that \( \tilde{\beta}_{\text{pc}}^r(r) \) is obtained by subtracting from \( \tilde{\beta}_{\text{pc}}^r \) the part of the data that belongs to the \( p-r \) dimensional space with the smallest variance and by performing the regression in the \( r \) dimensional space that contain most of the variability of the data. Note that ridge-regression (Hoerl and Kennard, 1970) which is an alternative way of dealing with the multicollinearity issue, consists in adding a positive term \( \lambda \) to all eigenvalues \( \lambda_j, j = 1, \ldots, p \). Both the ridge regression estimator and the principal components estimator \( \tilde{\beta}_{\text{pc}}^r(r) \) are biased for \( \beta \) under model \( \xi \) (Gunst and Mason, 1977).

The PC regression estimator \( \tilde{\beta}_{\text{pc}}^r(r) \) can estimated under the sampling design by

\[
\tilde{\beta}_{\text{pc}}^r(r) = G_r \hat{\gamma}_z(r),
\]

(3.9)

where \( \hat{\gamma}_z(r) \) is given in (3.5). Using relation (3.9) and the fact that \( Z_j = Xv_j \), we obtain that

\[
(\hat{t}_{z_d} - t_z)(\hat{\gamma}_z)(r) = (\hat{t}_{x_d} - t_x)(\tilde{\beta}_{\text{pc}}^r(r)).
\]

Consequently \( \hat{\ell}_{yw}(r) \) can also be written in the following form

\[
\hat{\ell}_{yw}(r) = \hat{t}_{yd} - (\hat{t}_{x_d} - t_x) \hat{\beta}_{\text{pc}}^r(r)
\]

and \( \hat{\ell}_{yw}(r) \) may be seen as a GREG-type estimator assisted by the model \( \xi \) when \( \beta \) is estimated by \( \tilde{\beta}_{\text{pc}}^r(r) \).

**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 \( x_k^2 \) especially when “the relationship to the study variable is curved”. Calibration on higher-order moments of the auxiliary variables has also been studied by Ren (2000). In our case, the PC variables \( Z_j \) have the following property:

\[
\frac{1}{N} Z_j^T Z_j = \frac{1}{N} \sum_{k \in U} z_{kj}^2 = \lambda_j, \quad \text{for all } j = 1, \ldots, p.
\]

This means that in presence of complete auxiliary information, the totals of squares of the PCs are known. As a consequence, if we keep the first \( r \) variables \( Z_1, \ldots, Z_r \) corresponding to the largest \( r \) eigenvalues, we can consider \( r \)
additional calibration constraints on the second moment of these PCs.

Let $Z_{j}^{2} = (z_{kj}^{2})_{k \in U}$ for all $j = 1, \ldots, p$. We want to find the calibration weights $w^{pc}(r)$ that satisfy the following optimization problem

$$w^{pc}(r) = \arg \min_{w} \Phi_{s}(w),$$

subject to

$$\sum_{k \in s} w^{pc}_{k}(r) \begin{pmatrix} z_{kr} \\ z_{kr}^{2} \end{pmatrix} = \sum_{k \in U} \begin{pmatrix} z_{k,r} \\ z_{k,r}^{2} \end{pmatrix}$$

where $z_{kr}^{2} = (z_{kr}^{21}, \ldots, z_{kr}^{2p})$.

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

4. Calibration on Estimated Principal Components

The approach presented in the above sections supposes that the auxiliary variables $X_{j}$, for $j = 1, \ldots, p$ are known for all units $k$ in the population. In practice, it often happens that the variables $X_{j}$ are only known for the sampled individuals, but their population totals are known. Then, it is not possible anymore to compute the eigenvalues and the eigenvectors of the population variance-covariance matrix. We present in this section a way to perform principal components calibration when the auxiliary variables are only observed for the units belonging to the sample.

Let $\hat{\mathbf{\Gamma}}$ be the design-based estimator of the variance-covariance matrix $\mathbf{\Gamma} = \frac{1}{N} \mathbf{X}^{T} \mathbf{X}$ with

$$\hat{\mathbf{\Gamma}} = \frac{1}{\hat{N}} \sum_{k \in s} \frac{1}{\pi_{k}} (\mathbf{x}_{k} - \hat{\mathbf{X}})(\mathbf{x}_{k} - \hat{\mathbf{X}})^{T} = \frac{1}{\hat{N}} \sum_{k \in s} \frac{1}{\pi_{k}} \mathbf{x}_{k} \mathbf{x}_{k}^{T} - \hat{\mathbf{X}} \hat{\mathbf{X}}^{T}$$

(4.1)

where $\hat{N} = \sum_{s} \frac{1}{\pi_{k}}$ and $\hat{\mathbf{X}} = \frac{1}{\hat{N}} \sum_{s} \frac{1}{\pi_{k}} \mathbf{x}_{k}$. 
Let \( \hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p \geq 0 \) be the sorted eigenvalues of \( \hat{\Gamma} \) and \( \hat{v}_1, \ldots, \hat{v}_p \) the corresponding orthonormal eigenvectors,

\[
\hat{\Gamma} \hat{v}_j = \hat{\lambda}_j \hat{v}_j, \quad j = 1, \ldots, p.
\] (4.2)

We have that \( \hat{\lambda}_j \) and \( \hat{v}_j \) are the design-based estimators of \( \lambda_j \) and respectively, \( v_j \) 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 \( \pi_k, \pi_{kl} \) as well as on the variables \( X_j \), (see the assumptions (A1)-(A4) in Section 5), that the estimators \( \hat{\lambda}_j \) and \( \hat{v}_j \) are asymptotically design unbiased and consistent for \( \lambda_j \) and respectively, for \( v_j \). More precisely,

\[
E_p \left( \max_j |\hat{\lambda}_j - \lambda_j| \right)^2 = O(n^{-1}),
\]

and for each fixed \( j \),

\[
E_p \left( \|\hat{v}_j - v_j\|^2 \right) = O(n^{-1}), \quad j = 1, \ldots, p
\]

where \( \| \cdot \| \) is the usual Euclidean norm and \( \hat{v}_j \), which is defined up to sign change, is chosen to satisfy \( \hat{v}_j^T v_j \geq 0 \).

The unknown population principal components \( Z_j \) defined in (3.2) can be approximated as follows

\[
\hat{Z}_j = X \hat{v}_j,
\]

reminding that \( \hat{Z}_j = (\hat{z}_{kj})_{k \in U} \) is only known for the units in the sample. Nevertheless, its population total \( t_{Z_j} = \sum_{k \in U} \hat{z}_{kj} \) is known and is equal to zero since

\[
t_{\hat{Z}_j} = t_X^T \hat{v}_j = 0, \quad j = 1, \ldots, p.
\]

Note also that \( \hat{Z}_j \) are not the principal components associated with the variance-covariance matrix \( \hat{\Gamma} \). Consider now the \( r \) first estimated principal components \( \hat{Z}_1, \ldots, \hat{Z}_r \), corresponding to the \( r \) largest eigenvalues \( \hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_r \geq 0 \) and suppose that \( \hat{\lambda}_r > 0 \). 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 vector of EPC calibration weights \( w_{\text{epc}} = (w_{\text{epc}}^k, k \in s) \) is the solution of the following optimization problem

\[
w_{\text{epc}}(r) = \arg \min_w \Phi_s(w),
\]

subject to the constraints

\[
\sum_{k \in s} w_{\text{epc}}^k(r) \hat{z}_{kr} = \sum_{k \in U} \hat{z}_{kr},
\]

where \( \hat{z}_{kr}^T = (\hat{z}_{k1}, \ldots, \hat{z}_{kr}) \) is the vector of values of \( \hat{Z}_j, j = 1, \ldots, r \) recorded for the \( k \)th unit. With the chi-square distance function \( \Phi_s \) defined in (2.3), the EPC calibration weights \( w_{\text{epc}}^k(r) \) are given by

\[
w_{\text{epc}}^k(r) = d_k - d_k \hat{z}_{kr}^T \left( \sum_{k \in s} d_k \hat{z}_{kr} \hat{z}_{kr}^T \right)^{-1} \left( \hat{t}_{\hat{z}_d} - t_{\hat{z}_r} \right), \tag{4.3}
\]

where \( \hat{t}_{\hat{z}_d} = \sum_{k \in s} d_k \hat{z}_{kr} \) is the HT estimator of the total \( t_{\hat{z}_r} = \sum_{k \in U} \hat{z}_{kr} = 0 \).

The EPC calibration estimator for \( t_y \) is given by

\[
\hat{t}_{y_{\text{epc}}}^w(r) = \sum_{k \in U} w_{\text{epc}}^k(r) y_k = \hat{t}_{yd} - (\hat{t}_{\hat{z}_d} - t_{\hat{z}_r})^T \gamma_{\hat{z}}(r) \tag{4.4}
\]

where

\[
\gamma_{\hat{z}}(r) = \left( \sum_{k \in s} d_k \hat{z}_{kr} \hat{z}_{kr}^T \right)^{-1} \sum_{k \in s} d_k \hat{z}_{kr} y_k.
\]

The above relation may also be written with respect to the population totals of the initial variables, \( X_1, \ldots, X_p \), as follows

\[
\hat{t}_{y_{\text{epc}}}^w(r) = \hat{t}_{yd} - (\hat{t}_{\hat{z}_d} - t_{\hat{z}})^T \hat{\beta}_{\text{epc}}^X(r) \tag{4.5}
\]

where

\[
\hat{\beta}_{\text{epc}}^X(r) = (\hat{\upsilon}_1, \ldots, \hat{\upsilon}_r) \gamma_{\hat{z}}(r).
\]

5. Some Asymptotic Properties of the Principal Components Calibration Estimators
We consider in this section the asymptotic framework of Isaki and Fuller (1982) which assumes that the population and the sample sizes go to infinity. The number $p$ of auxiliary variables is supposed to be fixed. The proofs are postponed in an Appendix.

We suppose that the following classical assumptions hold.

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

(A2) $\pi_k > \lambda > 0$ for all $k \in U$; $\lim_{N \to \infty} n \max_{k \neq l} |\pi_{kl} - \pi_k \pi_l| < \infty$.

(A3) $\lim_{N \to \infty} \frac{1}{N} \sum_{U} y_k^2 < \infty$.

(A4) $||x_k|| < \infty$ for all $k \in U$, where $|| \cdot ||$ is the Euclidean norm.

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, $\lambda_1 \geq \cdots \geq \lambda_r$.

**Proposition 1.** Assume that (A1)-(A4) hold and that $\lambda_r > \lambda_{r+1} \geq 0$. Then, $\hat{\gamma}_z(r) - \tilde{\gamma}_z(r) = O_p(n^{-1/2})$ and

$$N^{-1}(\hat{\gamma}_{yw}(r) - t_y) = N^{-1}(\hat{\gamma}_{yw}(r) - t_y) + o_p(n^{-1/2}),$$

where $\hat{\gamma} = \hat{\gamma}_{yw}(r) - (\hat{\gamma}_{xw}(r) - t_x)\beta_{pc}^x(r)$.

The asymptotic variance of $\hat{\gamma}_{yw}(r)$ is given by

$$AV(\hat{\gamma}_{yw}(r)) = \sum_{k \in U} \sum_{k \in U} (\pi_{kl} - \pi_k \pi_l) d_k d_l \left( y_k - x_k^T \hat{\beta}_{pc}^x(r) \right) \left( y_l - x_l^T \hat{\beta}_{pc}^x(r) \right). \quad (5.1)$$

Note that if furthermore all the second-order inclusion probabilities are strictly positive, namely if $\pi_{kl} = \Pr(k \& l \in s) > 0$, then the variance of $\hat{\gamma}_{yw}(r)$ may be estimated by:

$$\hat{Var}(\hat{\gamma}_{yw}(r)) = \sum_{k \in s} \sum_{k \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} d_k d_l \left( y_k - x_k^T \hat{\beta}_{pc}^x(r) \right) \left( y_l - x_l^T \hat{\beta}_{pc}^x(r) \right).$$
The asymptotic variance may also be written with respect to the principal components $Z_j$,

$$AV(\hat{t}_{yw}(r)) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) d_k d_l (y_k - z_{kr}^T \hat{\gamma}_z(r)) (y_l - z_{lr}^T \hat{\gamma}_z(r)). \quad (5.2)$$

Thus, the variance can be estimated as follows,

$$\hat{\text{Var}}(\hat{t}_{yw}(r)) = \sum_{k \in s} \sum_{l \in s} \pi_{kl} - \pi_k \pi_l \pi_{kl} d_k d_l (y_k - z_{kr}^T \hat{\gamma}_z(r)) (y_l - z_{lr}^T \hat{\gamma}_z(r)).$$

We present now an intermediate result which states that we consistently estimate the regression coefficient based on the estimated principal components.

**Proposition 2.** Assume that (A1)-(A4) hold and that $\lambda_r > \lambda_r + 1 \geq 0$. We have $\hat{\gamma}_z(r) - \hat{\gamma}_z(r) = O_p(n^{-1/2})$. As a consequence, $\hat{\beta}_x^{\text{epc}}(r) - \hat{\beta}_x^{\text{epc}}(r) = O_p(n^{-1/2})$.

The last result simply shows that estimating the principal components does not change the asymptotic behavior of the principal component calibration estimator.

**Proposition 3.** Assume that (A1)-(A4) hold and that $\lambda_r > \lambda_r + 1 \geq 0$. We have

$$\frac{1}{N} \left( \hat{t}_{yw}(r) - t_y \right) = \frac{1}{N} \left( \hat{t}_{\text{diff}}(r) - t_y \right) + o_p(n^{-1/2}).$$

The asymptotic variance of $\hat{t}_{yw}^{\text{epc}}(r)$ is the variance of $\hat{t}_{\text{diff}}(r)$ given in (5.2).

Again if all the second order inclusion probabilities satisfy $\pi_{kl} = \Pr(k \& l \in s) > 0$, the variance $AV(\hat{t}_{yw}^{\text{epc}}(r))$ may be estimated by:

$$\hat{\text{Var}}(\hat{t}_{yw}^{\text{epc}}(r)) = \sum_{k \in s} \sum_{l \in s} \pi_{kl} - \pi_k \pi_l \pi_{kl} d_k d_l (y_k - z_{kr}^T \hat{\gamma}_z(r)) (y_l - z_{lr}^T \hat{\gamma}_z(r)) = \sum_{k \in s} \sum_{l \in s} \pi_{kl} - \pi_k \pi_l \pi_{kl} d_k d_l (y_k - x_k^T \hat{\beta}_x^{\text{epc}}(r)) (y_l - x_l^T \hat{\beta}_x^{\text{epc}}(r)).$$

Remark that even if both estimators $\hat{t}_{yw}(r)$ and $\hat{t}_{yw}^{\text{epc}}(r)$ have the same asymptotic behavior, the variance estimators are different.

6. Partial Calibration on Principal Components

The calibration estimators derived before are not designed to give the exact finite population totals of the initial variables $X_j$, $j = 1, \ldots, p$. In practice, it
is often desired to have this property satisfied for a few but important socio-demographical variables such as sex, age or socio-professional category.

We can adapt the method presented in previous section in order to fulfill this requirement. We split the auxiliary matrix $\mathbf{X}$ into two blocks: a first block $\tilde{\mathbf{X}}_1$ containing the $p_1$ 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 $\mathbf{X} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$. Note that 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 $p - p_1$ remaining variables are estimated only approximately. The idea is to calibrate directly on the auxiliary variables from $\tilde{\mathbf{X}}_1$ and on 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. We introduce for that the matrix $\mathbf{I}_N$ which is 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 span by the column vectors of matrix $\tilde{\mathbf{X}}_1$. We also define matrix $\mathbf{A}$,

$$\mathbf{A} = \left( \mathbf{I}_N - \mathbf{P}_{\tilde{\mathbf{X}}_1} \right) \tilde{\mathbf{X}}_2,$$

which is the projection of $\tilde{\mathbf{X}}_2$ onto the orthogonal space span 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" when considering an estimator of the total of $\mathcal{Y}$ calibrated on the variables in $\tilde{\mathbf{X}}_1$. We define the residual covariance matrix

$$\frac{1}{N} \mathbf{A}^T \mathbf{A} = \frac{1}{N} \tilde{\mathbf{X}}_2^T \left( \mathbf{I}_N - \mathbf{P}_{\tilde{\mathbf{X}}_1} \right) \tilde{\mathbf{X}}_2$$

and denote by $\tilde{\lambda}_1 \geq \ldots \geq \tilde{\lambda}_{p_2}$ its eigenvalues and by $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_{p_2}$ the corresponding orthonormal eigenvectors. Consider now, for $j = 1, \ldots, p - p_1$,

$$\tilde{\mathbf{Z}}_j = \mathbf{A} \tilde{\mathbf{v}}_j$$

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

$$\hat{t}_{yw}^{ppc}(r) = \sum_{k \in s} w_k^{ppc}(r) y_k,$$
where the PPC calibration weights $w_{ppc}^k(r)$, for $k \in s$, are the solution of the following optimization problem

$$w^{ppc}(r) = \arg\min_w \Phi_s(w),$$

subject to

$$\sum_{k \in s} w_{ppc}^k(r) \begin{pmatrix} \tilde{x}_k \\ \tilde{z}_{kr} \end{pmatrix} = \sum_{k \in U} \begin{pmatrix} \tilde{x}_k \\ \tilde{z}_{kr} \end{pmatrix},$$

where $\tilde{x}_k = (\tilde{x}_{k1}, \ldots, \tilde{x}_{kp1})$ is the vector of the values of the variables in $\tilde{X}_1$ and $\tilde{z}_{kr}^T = (\tilde{z}_{k1}, \ldots, \tilde{z}_{kr})$ is the vector whose elements are the partial principal components $\tilde{Z}_1, \ldots, \tilde{Z}_r$ for unit $k$. Note that with a different point of view, Breidt and Chauvet (2012) use, at the sampling stage, similar ideas in order to perform penalized balanced sampling.

7. Application to the estimation of the total electricity consumption

Description of the data

We illustrate the interest of using principal components calibration on data from the Irish Commission for Energy Regulation (CER) Smart Metering Project that has been conducted in 2009-2010 (CER, 2011).

In this project, which focuses on energy consumption and energy regulation, about 6000 smart meters have been installed in order 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 reduction dimension techniques based on PCA by considering a period of 14 consecutive days and a population of $N = 6291$ 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 $t_j$. Our

The data are available on request at the address: http://www.ucd.ie/issda/data/commissionforenergyregulation/
variable of interest is the total electricity consumption over the second week,

\[ t_y = \sum_{k \in \mathcal{U}} \sum_{j=336}^{672} y_k(t_j). \]

The auxiliary information is the load electricity curve of the first week. This means that we have \( p = 336 \) auxiliary variables, which are the consumption electricity levels at each of the \( p = 336 \) half hours of the first week. A sample of 5 auxiliary information curves is drawn in Figure 7.1.

The condition number of \( N^{-1}X^TX \) is 67055.78, which means that the matrix is really ill-conditioned and there are strong relationships between the calibration variables. The first principal component explains about 63% of the variance 336 initial variables and about 83% of the total variability of the data is preserved by projection onto the subspace span by the first ten principal components. Reducing the dimension should improve the performances of calibration.

**Comparison of the estimators**

To make comparisons we draw \( I = 1000 \) samples of size \( n = 600 \) (the sampling fraction is about 0.1) according to a simple random sampling design without replacement and we estimate the total consumption \( t_y \) over the second week with the following estimators:

- the Horvitz-Thompson estimator, denoted by \( \hat{t}_{yd} \),
- the calibration estimator \( \hat{t}_{yw} \) that takes account of all the \( p = 336 \) auxiliary variables plus the intercept term.
- the estimated principal components calibration estimator \( \hat{t}_{epc}(r) \) that takes account of \( r \) estimated PC plus the intercept term. The dimension \( r \) plays the role of a tuning parameter.

We have represented in Figure 7.2 the coefficient of variation (CV) of the calibration weights for the \( I=1000 \) Monte Carlo experiments for different values of the dimension \( r \). It is clearly seen that the weights have larger dispersion and are more heterogeneous as the number \( r \) of principal components used for calibration becomes large.

The accuracy of the estimators is then evaluated by comparing their mean square errors to the mean square error of the calibration estimator \( \hat{t}_{yw} \) that takes
Figure 7.1: A sample of 5 electricity load curves observed every half an hour during the first week.
Figure 7.2: Distribution of the coefficient of variation (CV) of the sampling weights for different values of the dimension $r$. The sample size is $n = 600$. 
account of all the auxiliary variables. The relative mean square error is defined as follows,

\[
R(\hat{\theta}) = \frac{\sum_{i=1}^{I}(\hat{\theta}(i) - t_y)^2}{\sum_{i=1}^{I}(\hat{t}_{yw}(i) - t_y)^2},
\]

(7.1)

where \(\hat{\theta}\) is the HT estimator \(\hat{t}_{yd}\) or the EPC calibration estimator \(\hat{t}_{yw}^{epc}(r)\). Better estimation procedures correspond to smaller values of criterion \(R(\hat{\theta})\). The HT estimator conducts very bad since \(R(\hat{t}_{yd}) = 23.3\).

We have computed this ratio for several values of \(r\), (see Figure 7.3), starting from \(r = 1\) to \(r = 336\) which leads to the calibration estimator \(\hat{t}_{yw}\). We remark that this ratio is roughly decreasing with a minimum value of about 0.4 for \(r = 6\) principal components and then, it is increasing up to 1 as the number of PC increases. An interesting feature is the stability of the principal components calibration technique with respect to the choice of the dimension \(r\). Indeed, in this application, choosing between 2 and 99 principal components permits to divide, at least by two, the mean squared error of estimation compared to the calibration estimator based on the whole auxiliary information.

8. Discussion and concluding remarks

A simple dimension reduction technique based on principal components calibration has been studied in this article. It provides an effective technique for approximate calibration when the number of auxiliary variables is large that can improve significantly the estimation compared to calibration on the whole set of initial auxiliary variables. Furthermore this simple technique can also be modified so that calibration can be exact for a set of a few important auxiliary variables. We have also noted in previous Section that a bad choice of the number of principal components which are used as calibration variables may not have dramatic consequences. Nevertheless, finding automatic data-driven procedures that could help in choosing a reasonable values for the dimension \(r\) is of real interest. Criterions based on cross-validation or penalized cross-validation for principal components regression (see for example Frank and Friedman, 1993) could certainly be adapted to our finite population context. This clearly deserves further investigation but
Figure 7.3: Evolution of the relative MSE, defined in (7.1), for different values of the dimension $r$ and a sample size $n = 600$. The horizontal axis is at a log scale.
is beyond the scope of the present work.

From a more theoretical point of view, it would be interesting to examine what happens when the number $p$ of auxiliary variables is also allowed to tend to infinity when the sample size grows. Different situations about the asymptotic behavior of the smallest eigenvalue of matrix $\frac{1}{Np}X^TX$ may be distinguished. According to the fact that it tends to zero or not, different conclusions on how the number of principal components should be chosen may be drawn. This difficult problem is related to functional data analysis and inverse problems techniques and is beyond the scope of the paper.

Appendix : Proofs

Proof of Proposition 1

We may write

$$N^{-1}(\hat{\beta}_{pc}^{x}(r) - \hat{\beta}_{pc}^{d}(r)) = N^{-1}(\hat{\beta}_{pc}^{x}(r) - \hat{\beta}_{pc}^{d}(r))$$

and following the same lines as in Breidt and Opsomer (2000), we obtain with assumptions (A1)-(A4) that $N^{-1}(\hat{\beta}_{pc}^{x}(r) - \hat{\beta}_{pc}^{d}(r)) = O_p(n^{-1/2})$. See also the proof of result 2 below. □

Proof of Proposition 2

We consider, for simplicity, that $q_k = 1$ for all $k \in U$. We show first that

$$\frac{1}{N} \left( \sum_{k \in S} d_k \tilde{z}_{kr}^T \tilde{z}_{kr}^T - \sum_{k \in U} z_{kr}^T z_{kr}^T \right) = O_p(n^{-1/2}).$$

Let $G_r = (v_1, \ldots, v_r)$ be the $N \times r$ matrix whose columns are the $r$ orthonormal eigenvectors associated to the $r$ largest eigenvalues. We have that $(Z_1, \ldots, Z_r) = XG_r$, and

$$\frac{1}{N} \sum_{k \in U} z_{kr}^T z_{kr}^T = G_r^T \left( \frac{1}{N} X^TX \right) G_r = \text{diag}(\lambda_j^r)_{j=1}^r := \Lambda_r.$$
We also denote by $\hat{G}_r = (\hat{v}_1, \ldots, \hat{v}_r)$ the matrix whose columns are the orthonormal eigenvectors of $\hat{\Gamma}$ associated to the $r$ largest eigenvalues, $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_r \geq 0$. Note that these eigenvectors are unique up to sign change and, for $j = 1, \ldots, r$, we choose $\hat{v}_j$ such that $\langle \hat{v}_j, v_j \rangle \geq 0$.

With (4.1) we have that
\[
\frac{1}{N} \sum_{k \in s} d_k x_k x_k^T = \frac{\hat{N}}{N} \left( \hat{\Gamma} + \hat{X} \hat{X}^T \right)
\]
and
\[
\frac{1}{N} \sum_{k \in s} d_k z_{kr} z_{kr}^T = G_r^T \left( \frac{1}{N} \sum_{k \in s} d_k x_k x_k^T \right) G_r
\]
\[
= \left( \frac{\hat{N}}{N} - 1 \right) \hat{\Lambda}_r + \hat{\Lambda}_r + G_r^T \left( \frac{1}{N} \sum_{k \in s} d_k x_k - \bar{x} \right) \bar{X}^T G_r,
\]
where $\hat{\Lambda}_r$ is the diagonal matrix whose diagonal elements are $\hat{\lambda}_j$ for $j = 1, \ldots, r$. Since $\left( \frac{\hat{N}}{N} - 1 \right) = O_p(n^{-1/2})$ and
\[
\left\| G_r^T \left( \frac{1}{N} \sum_{k \in s} d_k x_k - \bar{x} \right) \bar{X}^T G_r \right\|_2^2 \leq \left\| G_r^T \left( \frac{1}{N} \sum_{k \in s} d_k x_k - \bar{x} \right) \right\|_2^2 \left\| \bar{X}^T G_r \right\|_2^2
\]
\[
= \sum_{j=1}^r < \hat{v}_j, \frac{1}{N} \sum_{k \in s} d_k x_k - \bar{x} >^2 \sum_{j=1}^r < \hat{v}_j, \bar{X}^T >^2
\]
\[
\leq \left\| \frac{1}{N} \sum_{k \in s} d_k x_k - \bar{x} \right\|^2 \left\| \bar{X}^T \right\|^2
\]
\[
= O_p(n^{-1}),
\]
we have that
\[
\frac{1}{N} \sum_{k \in s} d_k z_{kr} z_{kr}^T = \hat{\Lambda}_r + O_p(n^{-1/2}).
\]

Moreover, under assumptions (A1)-(A3), it can be shown that (see Cardot et al., 2010)
\[
\max_{j=1, \ldots, r} |\hat{\lambda}_j - \lambda_j| = O_p(n^{-1/2})
\]
and thus, $\|\hat{\Lambda}_r - \Lambda_r\| = O_p(n^{-1/2})$, for any matrix norm $\|\cdot\|$. This means that
\[
\frac{1}{N} \left( \sum_{k \in s} d_k \hat{z}_{kr} \hat{z}_{kr}^T - \sum_{k \in U} z_{kr} z_{kr}^T \right) = O_p(n^{-1/2}).
\]

Since $\lambda_r > \lambda_{r+1} \geq 0$ and $|\hat{\lambda}_r - \lambda_r| = O_p(n^{-1/2})$, we have that $1/\hat{\lambda}_r = 1/\lambda_r + O_p(n^{-1/2})$ and
\[
\left\| \left( \frac{1}{N} \sum_{k \in s} d_k \hat{z}_{kr} \hat{z}_{kr}^T \right)^{-1} - \left( \frac{1}{N} \sum_{k \in U} z_{kr} z_{kr}^T \right)^{-1} \right\| \leq \left\| \left( \frac{1}{N} \sum_{k \in s} d_k \hat{z}_{kr} \hat{z}_{kr}^T \right)^{-1} \right\| \times \left\| \left( \frac{1}{N} \sum_{k \in U} z_{kr} z_{kr}^T \right)^{-1} \right\| 
\]
\[
\leq \frac{1}{\lambda_r} \left\| \sum_{k \in s} d_k \hat{z}_{kr} \hat{z}_{kr}^T \right\| - \left\| \sum_{k \in U} z_{kr} z_{kr}^T \right\| \frac{1}{\lambda_r} = O_p(n^{-1/2}).
\]

Hence,
\[
\hat{\gamma}_z(r) - \tilde{\gamma}_z(r) = \left( \frac{1}{N} \sum_{k \in s} d_k \hat{z}_{kr} \hat{z}_{kr}^T \right)^{-1} \sum_{k \in s} d_k \hat{z}_{kr} y_k - \left( \frac{1}{N} \sum_{k \in U} z_{kr} z_{kr}^T \right)^{-1} \sum_{k \in U} z_{kr} y_k
\]
\[
= O_p(n^{-1/2}).
\]

Considering now $\hat{\beta}_x^{epc}(r) = \hat{G}_x^T \hat{\gamma}_z(r)$ and $\hat{\beta}_x^{dc}(r) = G_x^T \hat{\gamma}_z(r)$, we show with similar arguments that $\hat{\beta}_x^{epc}(r) - \hat{\beta}_x^{dc}(r) = O_p(n^{-1/2})$.

\[\square\]

**Proof of Proposition 3**

We may write
\[
\frac{1}{N} (i_{yw}^{epc} (r) - t_y) = \frac{1}{N} (i_{yw,x}^{diff} (r) - t_y) + \frac{1}{N} (i_{xd} - t_x)^T \left( \hat{\beta}_x^{epc}(r) - \hat{\beta}_x^{dc}(r) \right)
\]
\[
= \frac{1}{N} (i_{yw,x}^{diff} (r) - t_y) + o_p(n^{-1/2}).
\]

and the results follows from the above proposition. \[\square\]
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 calibration. *Metron-International Journal of Statistics*, LXVI: 5-20.

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

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

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

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

Chambers, R., Skinner, C. and Wang, S. (1999). Intelligent calibration. *Bulletin of the International Statistical Institute*, 58(2), 321-324.

Chauvet, G. and Goga, C. (2013). Bootstrap variance estimation and variable selection, preprint.

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

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87, 376-382.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109-148.

Goga, C., Shehzad, M.A., and Vanheuverzwyn, A. (2011). Principal component regression with survey data. application on the french media audience. *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*, Dublin, Ireland, 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**, 3199-3212.

Gunst, R.F. and Mason, R.L. (1977). Biased estimation in regression: an evaluation using mean squared error. *Journal of the American Statistical Association*, **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.

Jayasuriya, B. and Valliant, R. (1996). An application of restricted regression estimation in a household survey. *Survey Methodology*, **22**, 127-137.

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

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, American Statistical Association*, 57-65.

Rao, J.N.K. and Singh, A.C. (2009). Range Restricted Weight Calibration for Survey Data Using Ridge Regression. *Pakistan Journal of Statistics*, 25(4), 371-384.

Särndal, C.E., Swensson, B., and Wretman J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York Inc.

Ren, R. (2000). *Utilisation de l’information auxiliaire par calage sur fonction de répartition*. PhD, Université Paris Dauphine, France.

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

Skinner, C. and Silva, P.L.N. (1997). Variable selection for regression estimation in the presence of nonresponse. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 76-81.
Shehzad, M.-A. (2012). *Penalization and auxiliary information reduction methods in surveys*. Doctoral dissertation, Université de Bourgogne, France. Available at [http://tel.archives-ouvertes.fr/tel-00812880/](http://tel.archives-ouvertes.fr/tel-00812880/).

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

Singh, A. C. and Mohl, C. (1996). Understanding calibration in survey sampling. *Survey Methodology, 22*, 107-115.

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

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association, 94*, 635-644.

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

Cardot, H.
IMB, UMR CNRS 5584, Université de Bourgogne, 9 avenue Alain Savary, Dijon, France
herve.cardot@u-bourgogne.fr

Goga, C.
IMB, UMR CNRS 5584, Université de Bourgogne, 9 avenue Alain Savary, Dijon, France
camelia.goga@u-bourgogne.fr

Shehzad, M.-A.
Bahauddin Zakariya University, Bosan Road Multan, Pakistan.
mdjan708@gmail.com