Calibrating on Principal Components in the Presence of Multiple Auxiliary Variables for Nonresponse Adjustment

May 29, 2016

Bernardo João Rota[1,3)] and Thomas Laitila[1,2)]

bernardo.rota@oru.se, thomas.laitila@oru.se

bernardo.rota@uem.mz, thomas.laitila@scb.se

[1)] Department of Statistics, Örebro University, 701 82 Örebro, Sweden

[2)] Department of Research and Development, Statistics Sweden, 701 89 Örebro, Sweden

[3)] Statistics, Department of Mathematics and Informatics, Eduardo Mondlane University, Maputo, Mozambique.

Abstract

Nonresponse is a major impediment to valid inference in sample surveys. In the nonresponse scenario, the driver of successful estimation is the efficient use of available auxiliary information. As electronic devices provide considerable data storage capacities, at the estimation stage it is natural for survey statisticians to face large datasets of auxiliary variables. It is unwise to use all available data as doing so may lead to poor estimators, especially if some variables are strongly correlated. Furthermore, selecting a subset of available auxiliary variables may not be the best alternative given the issues related to selection criteria. In this paper, we propose reducing the dimensions of the original set of auxiliary variables by using principal components. The use of principal components in place of the original auxiliary variables is evaluated via two calibration approaches, linear calibration using no explicit response model and propensity calibration of a known response model. For

the latter, we propose selecting components based on their canonical correlation with the model variables. The results of two simulation studies suggest that using principal components is appropriate, as it offers the great advantage of reducing the computational burden.

Key words: Weighting, Nonresponse, Calibration, Principal components

1 Introduction

When adjusting for nonresponse in sample surveys, auxiliary information plays a prominent role in successful estimation. Rizzo, Kalton and Brick (1996) note that, providing it is carefully chosen, the particular adjustment scheme used at the estimation stage is not that important. The relation with the study variable or response pattern is usually taken as a benchmark in the choice of auxiliary variables (see Särndal and Lundström, 2005, p. 110; Kreuter and Olson, 2011). Calibration estimation (Deville and Särndal, 1992), initially designed to reduce sampling error in surveys with complete response, was eventually extended to surveys affected by nonresponse, (see, e.g., Lundström and Särndal, 1999; Kott, 2006). The method relies on an efficient choice of auxiliary variables.

When many auxiliary variables are available, calibrating on all of them may lead to 'over-calibration', the term used by Guggemos and Tillé (2010). According to Särndal and Lundström (2005), a problem may arise when the candidate auxiliary vector contains variables likely to cause multicollinearity or variables with highly skewed distributions. These problems may result in a very inefficient estimator almost less efficient than, for example, the Horvitz-Thompson estimator (Cardot, Goga, and Shehzad, 2014).

Large sets of auxiliary variables have also been considered by many authors in various estimation settings, as in the following examples, Bardsley and Chambers (1984) propose a ridge-type estimator in the context of model-based estimation, an approach that relaxes the principle that the calibration weights 'exactly' reproduce the totals of known characteristics by holding only 'approximately'. Guggemos and Tillé (2010) introduce a penalized calibration estimator. Bilen, Khan, and Yadav (2010) suggest a principal component approach for reducing the multicollinearity and dimensions of the auxiliary variables in a regression context. Cardot, Goga, and Shehzad (2014) propose calibration on reduced data via principal components (PCs) in surveys with complete response.

Variable selection criteria are also suggested in the literature as an alternative way to deal with large sets of auxiliary variables and related problems. McHenry (1978) suggests an

algorithm to select the best subset of auxiliary variables in the context of multiple regression or multivariate analysis. Silva and Skinner (1997) suggest a selection criterion based on the variability of the regression estimator. Särndal and Lundström (2005, 2007) propose a selection device based on the variability of estimated inverse propensities determined under the assumption that the auxiliary variables satisfy some pre-specified condition. The variable selection is conditioned on an increase in the variability of the inverse propensities. A potential auxiliary variable must predict the key survey variables and the propensities to respond. Geuzinge, Rooijen, and Bakker (2000) propose a selection indicator based on the product of (a) the correlation between the auxiliary vector and the study variables and (b) the correlation between the auxiliary vector and the response propensity. When adjusting for nonresponse through regression estimation, Bethlehem and Schouten (2004) and Schouten (2007) propose a selection based on minimizing the maximal absolute bias of the estimator; the method relies on computing an interval for the maximal absolute bias and selecting those variables that minimize its width.

The common practice of using a subset of the full set of potential auxiliary variables and discarding others may result in the loss of important information. For example, in a regression context, it is known that the R^2 tends to decrease with the removal of regressors from the regression equation. This phenomenon can be interpreted in many ways, but in some cases is due to the loss of valuable information. Furthermore, most of the suggested selection algorithms are computationally intensive and, impractical for large sets of candidate auxiliary variables.

In this paper, we calibrate on reduced data via principal components. Thus, we account for the exponential growth in computing time due to dimensionality in the auxiliary data and most importantly, the problem of large weights due to outliers is also accounted turning the estimator more efficient. The idea was initially suggested by Cardot, Goga, and Shehzad (2014) in surveys with complete response, and we extend it to estimation in surveys affected by nonresponse. Furthermore, the ideas in Cardot, Goga, and Shehzad (2014) are centered on the Greg-type-calibration (the complete response linear calibration), while we study this and the propensity score calibration estimators in the nonresponse context. Note that the use of principal components in weighting does not stand for data interpretation, but is a tool for alleviating the problem of managing high-dimensional auxiliary data. Specifically, the PCs approach assists in the construction of new auxiliary variables from the original variables by taking into account all available candidate variables through linear combinations. Furthermore, we implement a rejection of PCs based on their canonical correlation (Hotteling, 1936) with the model variables.

Two calibration estimators are considered in the paper:

1. Linear calibration (LC) using no explicit form of response model (Särndal and Lundström, 2005)

2. Instrumental variable or propensity score calibration (PSC) with an explicit form of response model (Chang and Kott, 2008)

This suggests two sources of auxiliary information for estimation: an $\mathbf{X}_{(N\times P)}$ data matrix carrying information on the N population elements of a P-dimensional vector of auxiliary variables and an $\mathbf{H}_{(m\times L)}$ data matrix carrying information on the m respondent elements of an L-dimensional vector of instrumental variables. The LC estimator uses only the first source of auxiliary information, while the PSC combines the two sources.

The rest of the article is organized as follows: section 2 provides background information on calibration estimators for nonresponse adjustment; section 3 provides a summary theoretical framework on principal components; section 4 provides a theoretical combination of calibration estimators and principal components; section 5 provides numerical support for section 4; and the final section discusses the results.

2 Calibration Estimators

Define a finite population, U, of distinguishable units indexed by integers $1, 2, ..., k, ..., N$. A probability sample, s, of distinguishable elements indexed by integers $1, 2, ..., k, ..., n$ is drawn from U according to a probability sampling design, $p(s)$, yielding the first- and second-order inclusion probabilities, $\pi_k = P(k \in s)$ and $\pi_{kl} = P(k\&l \in s)$, respectively for all $k, l \in \{1, 2, ..., N\}$, where $\pi_{kk} = \pi_k$. Suppose that data are observed for subset $r \subset s$ with $|r| = m$. The elements of r are assumed to be generated by a random process, $q(r)$, on s. Thus, each element $k \in r$ is associated with probability $\theta_k = P(k \in r|k \in s)$. The random process $q(r)$ on a given s is usually termed a response mechanism, while θ_k is the response probability for the individual k. Here, it is assumed that events $k \in r$ and $l \in r$ for a given s are independent of one another given that $k \neq l$.

Calibration estimators were introduced by Deville and Särndal (1992) in the context of surveys with complete response; the approach was then extended to surveys affected by nonresponse. In this context, Särndal and Lundström (2005) define the calibration estimator for total $t_y = \sum_U y_k$ as,

$$\hat{t}_{ycal} = \mathbf{w}_{(r)}^t \mathbf{y}_{(r)} \tag{1}$$

where $\mathbf{w}_{(r)} = vec\{w_k\}^m$ and $\mathbf{y}_{(r)} = vec\{y_k\}^m$ are m-dimensional column vectors of calibrated weights w_k and study variable values y_k respectively. The term 'calibrated weights' means that the weights satisfy the calibration property $\mathbf{X}_{(r)}^t \mathbf{w}_{(r)} = \mathbf{T}_x$, where $\mathbf{T}_x = \sum_U \mathbf{X}_k$ and $\mathbf{X}_k$ being the transpose of the $\mathtt{k}^{th}$ line of $\mathbf{X}_{(N\times P)}$. Calibrated weights, w_k, are constructed to be as close as possible to the reciprocals of the sample inclusion probabilities, $d_k = 1/\pi_k$, according to a distance metric $\mathbf{\Omega}\left(\mathbf{w}_{(r)}; \mathbf{d}_{(r)}\right)$, while satisfying the above calibration property. Using Lagrange reasoning, calibrated weights can be derived by minimizing the following function:

$$\mathbf{\Omega}(\mathbf{w}_{(r)}; \mathbf{d}_{(r)}) + \boldsymbol{\gamma}^t \left(\mathbf{T}_x - \mathbf{X}_r^t \mathbf{w}_{(r)}\right)$$

where $\boldsymbol{\gamma}$ is a column vector of Lagrange multipliers, $\mathbf{d}_{(r)} = vec\{d_k\}^m$. The resulting calibrated weights take the form

$$w_k = d_k h(\boldsymbol{\gamma}^t \mathbf{X}_k) \tag{2}$$

where $h = \psi^{-1}$, $\psi = \partial\mathbf{\Omega}/\partial w$.

A different choice of $\mathbf{\Omega}$ leads to a different weight system, w_k. Deville and Särndal (1992) establish conditions under which any choice of distance function leads to estimators that are asymptotically equivalent to the regression estimator obtained through a Chi-square-type distance measure. Thus, the choice of distance measure may be influenced by the computational aspects or other properties of w_k, such as its non-negativity or degree of stability.

Using the Chi-square distance, i.e., $\mathbf{\Omega}(\mathbf{w}_{(r)}; \mathbf{d}_{(r)}) = \left(\mathbf{w}_{(r)} - \mathbf{d}_{(r)}\right)^t (2\mathbf{D})^{-1} \left(\mathbf{w}_{(r)} - \mathbf{d}_{(r)}\right)$, with $\mathbf{D} = diag\{d_1, d_2, ..., d_k, ..., d_m\}$, leads to the linear calibrated weights of the form

$$w_k = d_k + d_k \boldsymbol{\gamma}^t \mathbf{X}_k \tag{3}$$

where $\boldsymbol{\gamma} = \left(\mathbf{X}_{(r)}^t \mathbf{D} \mathbf{X}_{(r)}\right)^{-1} \left(\mathbf{T}_x - \mathbf{X}_{(r)}^t \mathbf{d}_{(r)}\right)$.

The linear calibration estimator for t_y is:

$$\hat{t}_{ycal} = \mathbf{w}_{(r)}^t \mathbf{y}_{(r)} = \mathbf{d}_{(r)}^t \mathbf{e}_{(r)} + \mathbf{T}_x^t \left(\mathbf{X}_{(r)}^t \mathbf{D} \mathbf{X}_{(r)}\right)^{-1} \mathbf{X}_{(r)}^t \mathbf{D} \mathbf{y}_{(r)} \tag{4}$$

where, $\mathbf{e}_{(r)} = vec\{e_k\}^m$ and $\mathbf{y}_{(r)} = vec\{y_k\}^m$ are m-dimensional column vectors of residuals $e_k = y_k - \hat{y}_k$ and study variable values y_k respectively, and $\hat{y}_k = \mathbf{X}_k^t \left(\mathbf{X}_{(r)}^t \mathbf{D} \mathbf{X}_{(r)}\right)^{-1} \mathbf{X}_{(r)}^t \mathbf{D} \mathbf{y}_{(r)}$.

In the complete response context, estimator (4) is equivalent to the GREG estimator (Särndal, Swensson and Wretman, 1992) derived under superpopulation model ξ, which

assumes a linear relationship between the survey variable, y_k, and the auxiliary vector, $\mathbf{X}_k$, given by $\xi : y_k = \boldsymbol{\beta}^t \mathbf{X}_k + \varepsilon_k$. Since, $\mathbf{X}_{(s)}^t \mathbf{d}_{(s)}$ is unbiased for $\mathbf{T}_x$, the weights (3) are in average equal to d_k which leads to zero average differences $y_k - \hat{y}_k$.

Under nonresponse, the unbiasedness property mentioned above do not generally hold. In this case auxiliary information makes a difference for the properties of the calibration estimator.

3 A brief summary of principal components

Suppose that $\mathbf{X}$ is defined as in Section 1 except that each $\mathbf{X}_j, j = 1, ..., P$ is rescaled to zero mean and unit variance, then, $\mathbf{X}^t\mathbf{X}$ is the covariance matrix of $\mathbf{X}$. Let $(\lambda_j, \mathbf{b}_j; j = 1, ..., P\ )$ be eigenvalue-eigenvector pairs of $\mathbf{X}^t\mathbf{X}$. The j^{th} principal component is given by $\mathbf{Z}_j = \mathbf{b}_j^t\mathbf{X} = \sum_{l=1}^{P} \mathbf{b}_{lj}\mathbf{X}_l$ with the properties $cov(\mathbf{Z}_j, \mathbf{Z}_i) = \begin{cases} 0, & j \neq i \\ \lambda_i, & j = i \end{cases}$, $\mathbf{b}_j$ is a P-dimensional column vector and the $\lambda_{i's}$ satisfy $\lambda_1 \geq \lambda_2 \geq, ..., \geq \lambda_P \geq 0$. The proportion of total variance accounted for by the first $R < P$ principal components is given by $\left(\sum_{i=1}^{R} \lambda_i / \sum_{i=1}^{P} \lambda_i\right) \times 100\%$.

Suppose now that $\mathbf{X} = \mathbf{X}_{(\mathbf{s})}$, that is, auxiliary data observed only at sample level. The covariance matrix of $\mathbf{X}_{(\mathbf{s})}$ is estimated without bias by $\mathbf{X}^t\mathbf{D}\mathbf{X}$, where
$\mathbf{D} = diag\{d_1, d_2, ..., d_k, ..., d_n\}$. The estimated principal components are given by $\hat{\mathbf{Z}}_j = \hat{\mathbf{b}}_j^t\mathbf{X}_{(\mathbf{s})} = \sum_{l=1}^{P} \hat{\mathbf{b}}_{lj}\mathbf{X}_{l(s)}$. The pair $(\hat{\lambda}_j, \hat{\mathbf{b}}_j; j = 1, ..., P)$ comprise the eigenvalue and eigenvector of $\mathbf{X}^t\mathbf{D}\mathbf{X}$.

4 Calibrating on principal components

The calibration estimator in the principal components setting can be derived by solving the following problem:

$$\begin{array}{l} \min \boldsymbol{\Omega}(\mathbf{w}_{(r)}^{pc}; \mathbf{d}_{(r)}) \\ sub: \ \mathbf{Z}_{(r)}^t \mathbf{w}_{(r)}^{pc} = \mathbf{T}_z \end{array}, \tag{5}$$

4.1 The linear calibration estimator based on principal components

If we follow the same reasoning that led to weights (3), we will then arrive at principal components calibrated weights given by

$$\mathbf{w}_k^{pc} = d_k - d_k \boldsymbol{\gamma}_{(pc)}^t \mathbf{Z}_k \tag{6}$$

where, $\boldsymbol{\gamma}_{(pc)} = \left(\mathbf{Z}_{(r)}^t \mathbf{D} \mathbf{Z}_{(r)}\right)^{-1} \left(\mathbf{T}_z^t - \mathbf{Z}_{(r)}^t \mathbf{d}_{(r)}\right)$ and $\mathbf{Z}_k = \{Z_{k1}, Z_{k2}, ..., Z_{kR} | R < P\}$ is the vector whose elements are the retained components. The nonresponse principal-components-based calibration estimator for t_y is given by

$$\hat{t}_{ycal(pc)} = \mathbf{d}_{(r)}^t \mathbf{e}_{(r)}^{pc} + \mathbf{T}_z^t \left(\mathbf{Z}_{(r)}^t \mathbf{D} \mathbf{Z}_{(r)}\right)^{-1} \mathbf{Z}_{(r)}^t \mathbf{D} \mathbf{y}_{(r)} \tag{7}$$

Where $\mathbf{e}_{(r)}^{pc} = vec\left\{ y_k - \mathbf{Z}_k^t \left(\mathbf{Z}_{(r)}^t \mathbf{D} \mathbf{Z}_{(r)}\right)^{-1} \mathbf{Z}_{(r)}^t \mathbf{D} \mathbf{y}_{(r)} \right\}^r$.

4.2 The propensity score calibration based on principal components

Consider a framework of unit response resulting according to a known parametric model, $\phi^{-1}(\cdot; \mathbf{H}_k)$. Observe that this model is known only up to an unknown L-dimensional vector of parameters, $\boldsymbol{\delta} = \boldsymbol{\delta}^*$, where $\boldsymbol{\delta} \in \boldsymbol{\Upsilon}$, $\dim(\mathbf{H}_k) = L \leqslant R$ and R is the number of selected PCs. Then, the model parameters can be estimated from the calibration constraint below (see Kott, 2012)

$$\mathbf{Z}_{(r)}^t \boldsymbol{\Phi}(\boldsymbol{\delta}) \mathbf{d}_{(r)} - \mathbf{T}_z = \mathbf{0} \tag{8}$$

where $\dim(\mathbf{Z}_{(r)}) = m \times R$ and $\boldsymbol{\Phi}(\boldsymbol{\delta}) = diag\left\{\phi(\boldsymbol{\delta}; \mathbf{H}_1), \phi(\boldsymbol{\delta}; \mathbf{H}_2), ..., \phi(\boldsymbol{\delta}; \mathbf{H}_k), ..., \phi(\boldsymbol{\delta}; \mathbf{H}_m)\right\}$. This is a principle suggested by Chang and Kott (2008). The solution to (8) is the minimizer of the objective function:

$$\left(\mathbf{Z}_{(r)}^t \boldsymbol{\Phi}(\boldsymbol{\delta}) \mathbf{d}_{(r)} - \mathbf{T}_z\right)^t \mathbf{W}_n \left(\mathbf{Z}_{(r)}^t \boldsymbol{\Phi}(\boldsymbol{\delta}) \mathbf{d}_{(r)} - \mathbf{T}_z\right). \tag{9}$$

When $L = R$, the form of weighting matrix $\mathbf{W}_n$ of dimension $R \times R$ is irrelevant, as system (8) is just identified, otherwise $\mathbf{W}_n$ is a suitably chosen nonnegative definite matrix. Note that $\mathbf{Z}_k$ is an R-dimensional column vector of retained principal components of P originals. Under this setting, to make the system of equations (8) feasible, the minimal

requirement is that the number of PCs in $\mathbf{Z}_k$ be at least L retained components.

Having estimated the response model parameter, $\boldsymbol{\delta}^*$, the calibration estimator for t_y (the propensity score calibration) is

$$\hat{t}_{PSC(pc)} = \sum_r d_k \phi(\hat{\boldsymbol{\delta}}^t_{(pc)} \boldsymbol{Z}_k) y_k, \tag{10}$$

where $\hat{\boldsymbol{\delta}}_{(pc)}$ is the estimated value of $\boldsymbol{\delta}$. To obtain $\hat{\boldsymbol{\delta}}_{(pc)}$, Beaumont (2006), propose an iterative procedure based on the Taylor approximation of (8), similar to the procedure suggested by Binder (1983). We apply a slightly different perspective in the estimation of $\boldsymbol{\delta}$ in (8).

Assume the following conditions to hold:

1. Function $\phi(\boldsymbol{\delta})$ is continuous and twice differentiable with respect to $\boldsymbol{\delta}$.
2. $E_{pq}\left(\mathbf{Z}^t_{(r)}\boldsymbol{\Phi}(\boldsymbol{\delta})\mathbf{d}_{(r)} - \mathbf{T}_z\right) = \mathbf{0}$ if and only if $\boldsymbol{\delta} = \boldsymbol{\delta}^*$ for all $\boldsymbol{\delta} \in \boldsymbol{\Upsilon}$
3. Set $\boldsymbol{\Upsilon}$ is a compact set .
4. $E_{pq}\left[\left(\mathbf{Z}^t_{(r)}\boldsymbol{\Phi}(\boldsymbol{\delta})\mathbf{d}_{(r)} - \mathbf{T}_z\right)\left(\mathbf{Z}^t_{(r)}\boldsymbol{\Phi}(\boldsymbol{\delta})\mathbf{d}_{(r)} - \mathbf{T}_z\right)^t\right]$ is finite
5. $\mathbf{Z}^t_{(r)}\boldsymbol{\Psi}(\boldsymbol{\delta})\mathbf{H} = \frac{\partial}{\partial\boldsymbol{\delta}}\left(\mathbf{Z}^t_{(r)}\boldsymbol{\Phi}(\boldsymbol{\delta})\mathbf{d}_{(r)} - \mathbf{T}_z\right) = \sum_r d_k\phi_1(\mathbf{H}_k;\boldsymbol{\delta})\mathbf{Z}_k\mathbf{H}^t_k$ exists and is continuous in $\boldsymbol{\Upsilon}$, where $\phi_1(\mathbf{H}_k;\boldsymbol{\delta}) = \partial\phi(\mathbf{H}_k;\boldsymbol{\delta})/\partial\boldsymbol{\delta}$ and the $m \times m$ diagonal matrix $\boldsymbol{\Psi}(\boldsymbol{\delta})$ has its $\mathtt{k}^{th}$ diagonal element given by $d_k\phi_1(\mathbf{H}_k;\boldsymbol{\delta})$
6. $\mathbf{Z}^t_{(r)}\boldsymbol{\Psi}(\boldsymbol{\delta})\mathbf{H}$ is a full-column rank matrix.

Define the quadratic distance as follows:

$$\left(\mathbf{Z}^t_{(r)}\boldsymbol{\Phi}(\boldsymbol{\delta})\mathbf{d}_{(r)} - \mathbf{T}_z\right)^t \frac{\mathbf{W}_n}{2}\left(\mathbf{Z}^t_{(r)}\boldsymbol{\Phi}(\boldsymbol{\delta})\mathbf{d}_{(r)} - \mathbf{T}_z\right) \tag{11}$$

The solution to (8) is defined as the minimizer of objective function (11). In the generalized method of moments setting, minimizing (11) is equivalent to solving the set of estimating equations defined by

$$\left(\mathbf{Z}^t_{(r)}\boldsymbol{\Psi}(\boldsymbol{\delta})\mathbf{H}\right)^t\mathbf{W}_n\left(\mathbf{Z}^t_{(r)}\boldsymbol{\Phi}(\boldsymbol{\delta})\mathbf{d}_{(r)} - \mathbf{T}_z\right) = \mathbf{0} \tag{12}$$

We use the following approximation:

$$\left(\mathbf{Z}^t_{(r)}\boldsymbol{\Phi}(\boldsymbol{\delta}^*)\mathbf{d}_{(r)} - \mathbf{T}_z\right) \approx \left(\mathbf{Z}^t_{(r)}\boldsymbol{\Phi}(\hat{\boldsymbol{\delta}}_{(pc)})\mathbf{d}_{(r)} - \mathbf{T}_z\right) + \left(\mathbf{Z}^t_{(r)}\boldsymbol{\Psi}(\hat{\boldsymbol{\delta}}_{(pc)})\mathbf{H}\right)(\boldsymbol{\delta}^* - \hat{\boldsymbol{\delta}}_{(pc)}) \tag{13}$$

Introducing equation (13) into (12) yields the following updating equation:

$$\hat{\boldsymbol{\delta}}^{1}_{(pc)} \approx \hat{\boldsymbol{\delta}}^{0}_{(pc)} + \left[\left(\mathbf{Z}^{t}_{(r)}\boldsymbol{\Psi}^{0}\mathbf{H}\right)^{t}\mathbf{W}_{n}\left(\mathbf{Z}^{t}_{(r)}\boldsymbol{\Psi}^{0}\mathbf{H}\right)\right]^{-1}\left(\mathbf{Z}^{t}_{(r)}\boldsymbol{\Psi}^{0}\mathbf{H}\right)^{t}\mathbf{W}_{n}\left(\mathbf{Z}^{\boldsymbol{t}}_{(\boldsymbol{r})}\boldsymbol{\Phi}(\boldsymbol{\delta}^{0})\mathbf{d}_{(r)} - \mathbf{T}_{z}\right) \tag{14}$$

where $\boldsymbol{\Psi}^{0} = \boldsymbol{\Psi}(\hat{\boldsymbol{\delta}}^{0}_{(pc)})$. In (10), $\hat{\boldsymbol{\delta}}_{(pc)}$ is the value of $\hat{\boldsymbol{\delta}}^{1}_{(pc)}$ obtained upon convergence of (14).

In the appendix section we provide the derivation of the asymptotic variances of the estimated coefficients of the propensity functions when population- or sample-level auxiliary information is used. A comparison of these variances shows that sample-level auxiliary information provides more accurate estimated coefficients than population-level does.

4.3 Suggested retention criterion (a canonical correlation-based criterion)

Many authors have discussed PCs retention criteria, for example, Jolliffe (1972, 1973, 1982), Cadima and Jolliffe (1995), Jolliffe, Trendafilov, and Uddin (2003), and McCabe (1984), though there is no unified recommendation on this matter (Johnson and Wichern, 2007). Common practice is based on one or combinations of the following three criteria: the eigenvalue-one, scree plot, and proportion of total variance explained criteria. Mansfield, Webster, and Gunst (1977) noted that it is common in PCs analysis for significant data variation to be accounted for by the first few components. According to these criteria, the components with small variability are excluded. Note, however that we are not concerned with interpreting PCs, instead using them as a tool for constructing new auxiliary variables that take into account all original candidate auxiliary variables.

In a canonical correlation setting, the goal is to determine sets of linearly independent vectors for two groups of variables that result in the maximum correlation between the projections of these variables onto the space spanned by these linearly independent vectors. According to Borga (2001), the correlation between two sets of multidimensional variables, if it exists, may be blurred if an inappropriate coordinate system is used to represent the variables. However, in canonical correlation, each of the two sets is linearly transformed, so that the corresponding pairs of coordinates of these transformed variables have the maximum correlation.

Observe that $\mathbf{H} = \{H_1, H_2, ..., H_L\}^t$ is an L-dimensional vector of model variables, $\mathbf{Z}_{(r)} = \left\{Z_{1(r)}, Z_{2(r)}, ..., Z_{R(r)}\right\}^t = \tilde{\boldsymbol{Z}}$ is an R-dimensional vector of retained principal components, and let $\mathbf{P_H}$ be the projection of $\mathbf{H}$ onto the space spanned by linear combinations of its

elements and suppose that $\mathbf{P}_{\tilde{\mathbf{Z}}}$ is the analogous projection of elements in $\tilde{\mathbf{Z}}$. We want to approximate the correlation ($\tilde{\rho}_{\mathbf{H},\tilde{\mathbf{Z}}}$) of sets $\mathbf{H}$ and $\tilde{\mathbf{Z}}$ by the canonical correlation defined by $\max_{\mathbf{P_H},\mathbf{P}_{\tilde{\mathbf{Z}}}} \mathbf{\Gamma}\left(\mathbf{P_H}\mathbf{H}^t, \mathbf{P}_{\tilde{\mathbf{Z}}}\tilde{\mathbf{Z}}^t\right)$.

$$\tilde{\rho}_{\mathbf{H},\tilde{\mathbf{Z}}} \equiv \max_{\mathbf{P_H},\mathbf{P}_{\tilde{\mathbf{Z}}}} \mathbf{\Gamma}\left(\mathbf{P_H}\mathbf{H}^t, \mathbf{P}_{\tilde{\mathbf{Z}}}\tilde{\mathbf{Z}}^t\right) = \max_{\mathbf{P_H},\mathbf{P}_{\tilde{\mathbf{z}}}} \frac{\left[\mathbf{P_H}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}^t\right]}{\left[\mathbf{P_H}\left(\mathbf{H}^t\mathbf{H}\right)\mathbf{P}_{\mathbf{H}}^t\right]^{1/2}\left[\mathbf{P}_{\tilde{\mathbf{Z}}}\left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}^t\right]^{1/2}} \tag{15}$$

We can equivalently reformulate (15) as

$$\begin{cases} \max\left[\mathbf{P_H}\left(\mathbf{H^t}\tilde{\mathbf{Z}}\right)\mathbf{P}_{\mathbf{H}}^{\mathbf{t}}\right] \\ sub \begin{cases} \left[\mathbf{P_H}\left(\mathbf{H^t}\mathbf{H}\right)\mathbf{P}_{\mathbf{H}}^{\mathbf{t}}\right]^{1/2} = \mathbf{I} \\ \left[\mathbf{P}_{\tilde{\mathbf{Z}}}\left(\tilde{\mathbf{Z}}^{\mathbf{t}}\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}^{\mathbf{t}}\right]^{1/2} = \mathbf{I} \end{cases} \end{cases} \tag{16}$$

Using Lagrange multiplier principle, (16) is solved by maximizing the objective function

$$\mathbf{L}(\boldsymbol{\mu},\mathbf{P}) = \left[\mathbf{P_H}\left(\mathbf{H^t}\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}^{\mathbf{t}}\right] - \left\{\boldsymbol{\mu}_{\mathbf{H}}^{\mathbf{t}}\left[\mathbf{P_H}\left(\mathbf{H^t}\mathbf{H}\right)\mathbf{P}_{\mathbf{H}}^{\mathbf{t}} - \mathbf{I}\right] - \boldsymbol{\mu}_{\tilde{\mathbf{Z}}}^{\mathbf{t}}\left[\mathbf{P}_{\tilde{\mathbf{Z}}}\left(\tilde{\mathbf{Z}}^{\mathbf{t}}\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}^{\mathbf{t}} - \mathbf{I}\right]\right\}/2$$

yielding the system of equations

$$\begin{cases} \frac{\partial L}{\partial \mathbf{P_H}} = \left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}^t - \boldsymbol{\mu}_{\mathbf{H}}^{\mathbf{t}}\left(\mathbf{H}^t\mathbf{H}\right)\mathbf{P}_{\mathbf{H}}^t = \mathbf{0} \\ \frac{\partial L}{\partial \mathbf{P}_{\tilde{\mathbf{Z}}}} = \left(\mathbf{H^t}\widetilde{\mathbf{Z}}\right)^t\mathbf{P}_{\mathbf{H}}^{\mathbf{t}} - \boldsymbol{\mu}_{\tilde{\mathbf{Z}}}^t\left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}^t = \mathbf{0} \end{cases} \tag{17}$$

Premultiplying the first equation in (17) by $\mathbf{P_H}$ and subtracting $\mathbf{P}_{\tilde{\mathbf{Z}}}$ times the second equation from the first, gives the Lagrange coefficients a solution of $\boldsymbol{\mu}_{\tilde{\mathbf{Z}}} = \boldsymbol{\mu}_{\mathbf{H}} = \boldsymbol{\mu}$.
Assuming that $\mathbf{H}^t\mathbf{H}$ is invertible, the first equation gives

$$\boldsymbol{\mu}\mathbf{P}_{\mathbf{H}}^t = \left(\mathbf{H^t}\mathbf{H}\right)^{-1}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}^t, \tag{18}$$

and after appropriate replacements in the second, we get

$\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)^t\left(\mathbf{H}^t\mathbf{H}\right)^{-1}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}^t - \boldsymbol{\mu}^2\left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}^t$, which is equivalent to writing this last equation as

$$Qx = \lambda Rx \tag{19}$$

where $Q = \left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)^t\left(\mathbf{H}^t\mathbf{H}\right)^{-1}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)$, $x = \mathbf{P}_{\tilde{\mathbf{Z}}}^t$, $\lambda = \boldsymbol{\mu}^2$ and $R = \left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)$.

Equation (19) is in the form of a generalized eigenvalue equation (Parra and Sajda, 2003). Let, $R = MM^t$ be a Cholesky decomposition of R; then (19) becomes

$$\left(M^{-1}QM^{-1t}\right) M^t x = \lambda M^t x \Leftrightarrow \tilde{Q}\tilde{x} = \lambda\tilde{x},$$

which is the standard eigenvalue equation. Solving this, we obtain a solution for $\mathbf{P}_{\tilde{\mathbf{Z}}}$, which naturally leads to a solution for $\mathbf{P}_{\mathbf{H}}$ in (18). These solutions represent the optimal projections of the variables in $\left\{\tilde{\mathbf{Z}}\right\}$ and $\{\mathbf{H}\}$ onto spaces spanned by their respective linear combinations. The coordinate systems resulting from $\mathbf{P}_{\tilde{\mathbf{Z}}}$, and $\mathbf{P}_{\mathbf{H}}$ are mutually maximally correlated. See, for example, Borga (2001) and Hardoon, Szedmak and Shawe-Taylor (2004), for more insight on canonical correlation analysis.

Our PCs selection criterion is based on the value of the canonical correlation between the PCs and the instrumental variables. The PCs are selected in order of their appearance and the canonical correlations are used to measure the representativeness of the selected components. The canonical correlations are calculated in a forward stepwise manner: the first canonical correlation is the correlation between the instrumental vector and a vector comprinsing the first PC; the second canonical correlation is the maximal correlation between the instrument vector and the vector comprising the first two PCs, and so on. The values of these canonical correlations are obtained in an increasing order. The stopping rule is based on the amount by which this correlation increases from a previous step to the actual step. If the addition of a further component to the vector of PCs does not significantly change the correlation among these two groups, then that component and the remaining components are discarded from the final auxiliary vector.

Remark 1. Unlike $\mathbf{Z^tZ}$, which is a diagonal matrix with eigenvalues of $\mathbf{X^tX}$ being its diagonal elements, matrix $\tilde{\mathbf{Z}}^{\mathbf{t}}\tilde{\mathbf{Z}}$ is no longer a diagonal since $\tilde{\mathbf{Z}}$ is made of elements of $\mathbf{Z}$ falling into response set r.

Remark 2. We maximize the relation $(\mathbf{H}, \tilde{\mathbf{Z}})$ rather than $(\mathbf{H}, \mathbf{Z})$ as the latter is impossible because information on $\mathbf{H}$ is assumed to be known at response level. The variables' distributions are generally distorted by nonresponse and the resulting correlation is expected to deviate from the true correlation. This is not of concern here, as the main goal is to guarantee at the response level selected auxiliary variables closely linked to the instruments.

5 Simulation Study

This section provides empirical illustrations of the points discussed in the previous sections. It is known, that the principal components data reduction approach is effective when the relations among the variables involved are strong. In this article, we present two simulation studies: in the first study, the structure of correlation among the variables is very strong, the first principal component alone explaining more than 90% of the total data variation, as can be observed in Figure 1; in the second study, the structure of correlation among the variables is weak, and several components are needed to meaningfully explain the total variation of the data, as illustrated by, the scree plot shown in Figure 2.

The data source for the first study is 'Unemployment and median household income for the U.S., States, and counties, 2006–2014' from the Unemployment – Bureau of Labor Statistics – LAUS data. The data are freely and publicly accessible for use at http://www.bls.gov/lau/. According to the source, 'the concepts and definitions underlying LAUS data come from the Current Population Survey (CPS), the household survey that is the official measure of the labor force for the nation. State monthly model estimates are controlled in real time to sum to national monthly labor force estimates from the CPS. These models combine current and historical data from the CPS, the Current Employment Statistics (CES) program, and State unemployment insurance (UI) systems'.

The data source for the second study is 'Small Area Income and Poverty Estimates (SAIPE)', which is a 1989, 1993, and 1995–2013 dataset, also freely and publicly accessible at http://www.census.gov/did/www/saipe/. According to the source, 'Small Area Income and Poverty Estimates (SAIPE) are produced for school districts, counties, and states. The main objective of this program is to provide updated estimates of income and poverty statistics for the administration of federal programs and the allocation of federal funds to local jurisdictions'.

5.1 Simulation setup

5.1.1 Study 1

From the data of the first study we selected 27 quantitative variables. We applied data transformation to induce the correlation among them to a desired pattern. The transformed variables are named v_1 to v_{27}. For example, from uncorrelated variables x_1 and x_2 we can generate new corresponding correlated variables $v_1 = x_1$ and $v_2 = sqrt(x_1 * x_2)$, respectively. The variable v_{27} was chosen to be the study variable and the remaining were assumed to be auxiliaries. These data correspond to our population of 3260 obser-

vations from which simple random samples without replacement were drawn. We also used a dataset from a real estate survey of 4228 sampled individuals, of whom 2445 were respondents on one of the variables. A response variable was generated by assigning values of zero and one to nonrespondents and respondents, respectively and we fitted a three-covariate logistic model with chosen variables vector $\mathbf{w} = \{1, w_1, w_2\}^t$. The correlation between w_1 and w_2 is about 0.1, while each one of w_1 and w_2 has correlation level of approximately 0.6 with the study variable. The proportion of nonzero units in the resulting binary variable is 57%. We adapted the features of this model to our study. The chosen variables, say, v_1 and v_5 have correlation level of 0.09, while having with the study variable correlation levels of 0.5 and 0.48 respectively. The model led us to an average response rate of 57%.

Recall that we base this article on two calibration approaches, the linear calibration (LC) estimator of Särndal and Lundström (2005) and the propensity score calibration (PSC) of Chang and Kott (2008). For the former estimator, we use the standard specification of auxiliary vectors, that is, $\mathbf{H}_k = \mathbf{X}_k = \{1, v_1, ..., v_{26}\}_k^t$, while the auxiliary vectors for the latter were defined as $\mathbf{X}_k = \{1, v_2, ..., v_4, v_6, ..., v_{26}\}_k^t$ and $\mathbf{H}_k = \{1, v_1, v_5\}_k^t$. The attempt to adapting the response model mentioned above to our study, led to the choice of v_1 and v_5 as instrumental or model variables.

The principal components auxiliary variables for both the LC and PSC estimators were generated from their corresponding values of $\mathbf{X}_k$. The retention criterion for the LC estimator was the proportion of total variance explained by the set of selected components. This led to the selection of three out of 26 possible components in population LC, while for the PSC estimator, the retention criterion is that suggested in subsection 3.2. Each simulation result was based on 1000 replications. All estimators under study used the same samples and same response sets. The properties of the estimators of interest are the relative bias ($Rel.bias = \frac{bias(\hat{\theta})}{\theta} * 100\%$), the standard error ($S.E. = sqrt(var(\hat{\theta}))$), and the root mean squared error ($RMSE = sqrt(bias(\hat{\theta})^2 + var(\hat{\theta}))$. The scree plot given in Figure 1 below illustrates the population correlation structure of the variables in a principal components setting.

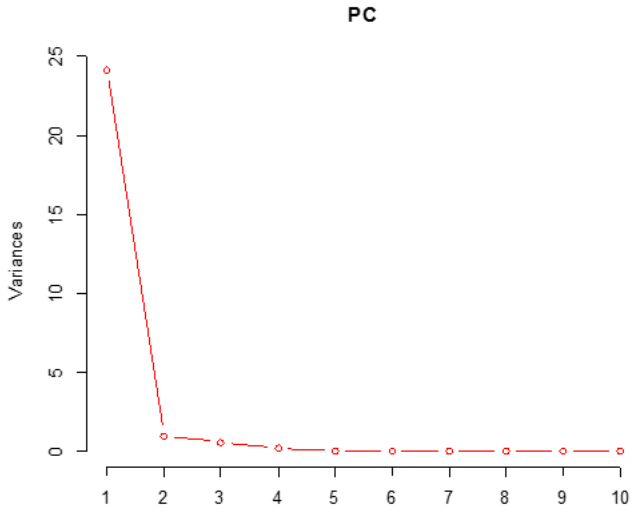

Figure 1: Scree plot of the auxiliary data in the first study.

5.1.2 Study 2

This is a replication of study 1 except that it uses the 'Small Area Income and Poverty Estimates' data-set (size 3173), from which we select some variables from the 2006 and others from the 2013 data, for a total of 19 variables, x_1 to x_{19}. The original data were square root transformed and, as in study 1, variables x_1 and x_5 are the instrumental variables while variable x_{19} was chosen as the study variable. The correlation between x_1 and x_5 is approximately 0.1, $cor(x_1, x_{19}) = 0.52$ and $cor(x_5, x_{19}) = 0.45$. The proportion of total variance explained by the selected PCs is again the retention criterion used for the LC estimator based on PCs. This criterion led to a fixed number of eight or an average number of eight retained components, depending on whether population or sample auxiliary information is used for estimation. The retention criterion for the PSC estimator based on PCs is again that described in subsection 3.2. The following is the scree plot of the principal components of the population auxiliary data used in the second simulation study.

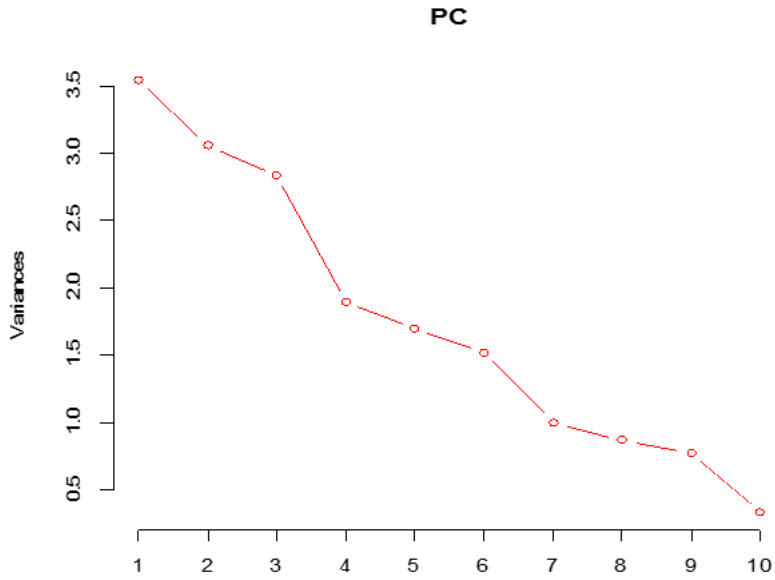

Figure 2: Scree plot of the auxiliary data in the second study

5.2 Simulation results

5.2.1 Results of study 1

The results are presented in two versions, a tabular version in Tables 1–4 and a graphic version in Figures 3–6 (the figures are in the appendix). These representations show the behaviour of each considered estimator when the sample size increases. For each table or graph, the performance of the estimator is evaluated from two perspectives: when the estimator is based on the complete original auxiliary variables (X) and when it is based on the principal components (PCs) of the auxiliary variables.

Tables 1 and 2 show the LC estimator results when auxiliary information is observed at the population and sample levels, respectively. Both tables show that, apart from the tenth line in Table 2, the relative bias, standard error, and the root mean squared error values of the principal-components-based linear calibration are smaller than the their counterparts computed based on the original auxiliary variables.

Table 1: LC on original population auxiliary variables vs. LC on population PCs – Study 1

Sample size	Properties	Estimators	
		L. Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	5.474	1.296
	S.E.	3519	935
	RMSE	8661	2094
400	Rel.bias(%)	4.771	1.224
	S.E.	3231	872
	RMSE	7616	1973
500	Rel.bias(%)	4.462	1.222
	S.E.	3083	804
	RMSE	7150	1941
600	Rel.Bias(%)	3.974	1.149
	S.E.	3135	846
	RMSE	6544	1864

Table 2: LC on original sample auxiliary variables vs. LC on sample PCs – Study 1.

Sample size	Properties	Estimator	
		L.Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	3.930	0.192
	S.E.	21,936	11,202
	RMSE	22,660	11,206
400	Rel.bias(%)	3.341	0.037
	S.E.	17,089	9621
	RMSE	17,758	9621
500	Rel.bias(%)	3.551	0.328
	S.E.	14,332	8250
	RMSE	15,224	8263
600	Rel.bias(%)	2.951	0.369
	S.E.	12,422	7608
	RMSE	13,134	7626

Tables 3 and 4 show results obtained under conditions similar to those used to obtain the results in Tables 1 and 2, except that PSC replaces LC. The results obtained by using PCs of the auxiliary variables are comparable to the obtained using original auxiliary variables, this is true in both levels of auxiliary information.

Table 3: PSC on original population auxiliary variables vs. PSC on population PCs – Study 1.

Sample size	Properties	Estimator			
		PS on X	Time (in hr)	PS on PCs	Time (in hr)
300	Rel.bias(%)	0.280	7	0.153	0.35
	S.E.	16,182		15,912	
	RMSE	16,188		15,914	
400	Rel.bias(%)	0.105	13	0.209	0.57
	S.E.	13,815		13,660	
	RMSE	13,816		13,663	
500	Rel.bias(%)	0.338	22	0.434	0.83
	S.E.	11,953		11,837	
	RMSE	11,963		11,854	
600	Rel.bias(%)	0.169	36	0.264	1.30
	S.E.	10,899		10,757	
	RMSE	10,902		10,764	

Table 4: PSC on original sample auxiliary variables vs. PSC on sample PCs – Study 1.

Sample size	Properties	Estimator			
		PS on X	Time (in hr)	PS on PCs	Time (in hr)
300	Rel.bias(%)	0.255	0.25	0.125	0.18
	S.E.	16,162		16,010	
	RMSE	16,166		16,011	
400	Rel.bias(%)	0.120	0.32	0.189	0.22
	S.E.	13,820		13,711	
	RMSE	13,821		13,713	
500	Rel.bias(%)	0.353	0.45	0.421	0.23
	S.E.	11,952		11,834	
	RMSE	11,963		11,850	
600	Rel.bias(%)	0.191	0.50	0.263	0.25
	S.E.	10,880		10,795	
	RMSE	10,884		10,801	

5.2.2 Results of study 2

Tables 5–10 below present the results of this study. The process of evaluating the estimators is similar to that used in study 1. The results of the LC, presented in Tables 5 and 6, display consistency when comparing X-and PCs- based estimators and when comparing population- and sample-based estimators.

Table 5: LC on original population auxiliary variables vs LC on population PCs – Study2.

Sample size	Properties	Estimator	
		L. Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	0.735	0.899
	S.E.	2262	2136
	MSE	2282	2168
400	Rel.bias(%)	0.810	1.077
	S.E.	1871	1798
	MSE	1901	1852
500	Rel.bias(%)	0.829	1.029
	S.E.	1558	1529
	MSE	1596	1588
600	Rel.bias(%)	0.672	0.836
	S.E.	1402	1382
	MSE	1429	1425

Table 6: LC on original sample auxiliary variables vs. LC on sample PCs – Study 2.

Sample size	Properties	Estimator	
		L. Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	0.725	0.949
	S.E.	2489	2494
	MSE	2507	2525
400	Rel.bias(%)	0.882	1.205
	S.E.	2068	2068
	MSE	2100	2129
500	Rel.bias(%)	0.841	1.104
	S.E.	1792	1814
	MSE	1825	1871
600	Rel.bias(%)	0.711	0.937
	S.E.	1639	1658
	MSE	1666	1703

The results of the PSC estimators for the second study are displayed in Tables 7 and 8. As the LC estimator, the PSC results are also consistent in terms of the type (X or PCs) and level (population or sample) of the auxiliary information used.

Table 7: PSC on original population auxiliary variables vs. PSC on population PCs – Study 2.

Sample size	Properties	Estimator	
		PSC on X	PSC on PCs
300	Rel.bias(%)	1.345	1.566
	S.E.	3748	3791
	MSE	3789	3846
400	Rel.bias(%)	1.627	1.925
	S.E.	3293	3385
	MSE	3362	3478
500	Rel.bias(%)	1.487	1.848
	S.E.	2921	2994
	MSE	2985	3091
600	Rel.bias(%)	1.757	2.072
	S.E.	2708	2846
	MSE	2804	2973

Table 8: PSC on original sample auxiliary variables vs. PSC on sample PCs – Study 2.

Sample size	Properties	Estimator	
		PSC on X	PSC on PCs
300	Rel.bias(%)	1.347	1.480
	S.E.	3634	3643
	MSE	3676	3695
400	Rel.bias(%)	1.482	1.701
	S.E.	3166	3219
	MSE	3226	3296
500	Rel.bias(%)	1.559	1.648
	S.E.	2775	2815
	MSE	2849	2897
600	Rel.bias(%)	1.778	1.908
	S.E.	2567	2651
	MSE	2672	2767

The results shown in Tables 9 and 10 comprise estimated model parameters in the PSC estimation using the data of the second study.

Table 9: Estimated model coefficients (population auxiliary information – Study 2.)

Coefficient estimates						
True coefficients	$(\delta_0, \delta_1, \delta_2)$ (1.311, -0.199, -0.083)					
Sample size	PSC on X			PSC on PCs		
	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$
300	1.129	-0.182	-0.044	1.097	-0.188	-0.026
	(0.197)	(0.008)	(0.011)	(0.247)	(0.012)	(0.014)
400	1.125	-0.174	-0.052	1.079	-0.176	-0.035
	0.149	0.006	0.008	(0.213)	(0.010)	(0.010)
500	1.147	-0.178	-0.056	1.096	-0.179	-0.038
	(0.133)	(0.005)	(0.006)	(0.182)	(0.008)	(0.009)
600	1.140	-0.178	-0.054	1.092	-0.179	-0.037
	(0.117)	(0.005)	(0.005)	(0.190)	(0.007)	(0.008)

Table 10: Estimated model coefficients (sample auxiliary information – Study 2.)

Coefficient estimates						
True coefficients	$(\delta_0, \delta_1, \delta_2)$ (1.311, -0.199, -0.083)					
Sample size	PSC on X			PSC on PCs		
	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$
300	1.139	-0.177	-0.054	1.119	-0.179	-0.046
	(0.092)	(0.003)	(0.005)	(0.122)	(0.005)	(0.005)
400	1.149	-0.175	-0.061	1.118	-0.174	-0.052
	(0.068)	(0.003)	(0.003)	(0.108)	(0.004)	(0.005)
500	1.148	-0.175	-0.061	1.132	-0.175	-0.054
	(0.063)	(0.0002	(0.003)	(0.003)	(0.094)	(0.004)
600	1.145	-0.175	-0.060	1.123	-0.174	-0.054
	(0.06)	(0.002)	(0.002)	(0.099)	(0.003)	(0.004)

6 Discussion

The results of two simulation studies are presented in the last section, and for each study we assess two calibration approaches, the LC estimator using no explicit form of response function and the PSC estimator with explicit functional form. Both estimators are evaluated using the original large set of auxiliary variables (X) and using the principal components (PCs) of the original auxiliary variables. The results of the first study are given in two versions, tabular and graphic, while the results of the second study are given in tabular form only. The graphic form enables the convenient visual inspection of the estimator behaviour, while the tabular form gives a quantitative illustration. Study 1 demonstrates that the LC estimator based on principal components auxiliary variables is always superior in terms of relative bias, standard error, and root mean squared error

(RMSE), to its counterpart using the original auxiliary information. This is true regardless of the level of the auxiliary information, that is, the population or sample levels, as demonstrated in Tables 1 and 2, respectively. There is a large discrepancy of the standard errors and RMSEs between population- and sample-based LC estimators. The RMSE of the population-based LC estimator based on the original auxiliary information ranges from 6544 to 8661 while the range for its counterpart based on sample-level auxiliary information is 13,134 to 22,660. The RMSE of the LC based on PCs auxiliary information ranges from 1864 to 2094 and from 7626 to 11,206 for population- and sample-based auxiliary information, respectively. Thus, the results differ greatly when comparing estimators of population- and sample-based auxiliary information. Considerable differences are also observed in the standard errors and RMSEs when comparing the estimators in terms of the type of auxiliary information used, that is, original X- auxiliaries and PCs-auxiliaries. This is not a surprising behaviour of the LC calibration estimator as this is a regression-type estimator.

In the response propensity calibration approach, auxiliary information is used in estimating response propensities; the estimation of population characteristics then proceeds by adjusting the design weights through multiplication by the corresponding reciprocals of the estimated propensities, which is usually called 'double weighting'. Here, it is observed that these results are more consistent and probably more realistic than the LC results. As Tables 3 and 4 illustrate, the principal-components-based estimator provides results similar to those obtained using the original auxiliary information. This is true regardless of the level of information on which the estimator is based. Furthermore, the results display consistency when comparing the properties of the corresponding estimators when population- and sample-based auxiliary data are used. The corresponding interval ranges of the RMSEs when using population-level auxiliary information are close to those when sample-level auxiliary information is used. As the sample size increases, the RMSEs tend to converge to the same level, irrespective of the type (X or PCs) or level (population or sample) of the auxiliary information. One of the major advantages of using PCs in place of the original auxiliary variables is the computational effort measured in terms of computational time; as reported in Tables 3–4, due to dimensionality reduction, the principal-components-based estimates are computed much more quickly than are the estimates based on the original auxiliary information.

Tables 5–10 report the results of the second simulation study. In contrast to the previous study, here, the LC calibration (Tables 5–6) results are consistent regardless of the type of auxiliary information used for estimation as well as when comparing the properties of the estimator across levels of information. The RMSEs of the estimators lie in virtually the same interval, regardless of the level of auxiliary information (population

or sample levels) or type (original X or PCs) . A similar observation can be made with respect to the PSC estimator in Tables 7–8. We can still compare the performances of the LC and PSC estimators as we are using the same set of auxiliary variables, however, in general this may not be a fair comparison since the estimators are conceptually different in terms of sources of auxiliary information they use.

The levels of bias are approximately the same: they are less than 0.1% in study 2 while in study 1 some differences are observed, especially in the LC estimator where the bias level attains 5.5%, as Tables 1 and 2 demonstrate. An interesting property of the auxiliary information in the PSC scheme, is the ability to appropriately estimate the response model. Tables 9–10 provide the population- and sample-based model-estimated coeffcients, and the results suggest equally good model coefficients estimates when PCs are used compared with estimates resulting from the use of the original X variables. As the results of model estimates are good, we can further improve the target estimates by performing a two-step estimation in which the products of design weigths and the reciprocal of the estimated response probabilities are used as initial weights in the linear calibration estimator. For the lack of space, we do not provide here the results of the two-step estimation.

We observe that the data structure in study 1 is an extreme case and less realistic than that in the second study. Both studies illustrate how the usage of principal components in place of original auxiliary data when adjusting for nonresponse does not lead to distorted results and has the great advantage of reducing the computational effort.

The reported PSC results based on principal components are very similar to those obtained using a fixed number of components via the eigenvalue-one rule. However, the eigenvalue-one results are worse than those of our approach based on canonical correlation for very small samples. When the sample size increases, the number of selected components converges to the number of components based on the eigenvalue-one rule. The Figure 7 in the appendix illustrates the behaviour of our components selection method using the data of the study 1.

References

Bardsley, P. and Chambers, R. L. (1984). Multipurpose estimation from unbalanced samples.*Journal of the Royal Statistical Society.Series C(Applied Statistics),*3:33,290–299.

Beaumont, J. F. (2006). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, **31**, 227–231.

Bethlehem, J. and B. Schouten (2004). Nonresponse adjustment in household surveys, *Discussion paper 04007. Statistics Netherlands, Voorburg/Heerlen, The Netherlands.*

Bilen, C., Khan, A. and Yadav, O. P. (2010). Principal components regression control for multivariate autocorrelated cascade process. *Int.J.Quality Engeneering and Technology*, **1:3**

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

Borga, M. (2001). *Canonical Correlation a Tutorial.*

Retrieved from https://www.cs.cmu.edu/˜+tom/10701_sp11/slides/CCA_tutorial.pdf

Cadima, J. and Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle components. *Journal of Applied Statistics*, **22:2**, 203-214, DOI: 10.1080/757584614

Cardot, H., Goga, C. and Shehzad, M.-A. (2014). Calibration and Partial Calibration on Principal Components when the Number of Auxiliary Variables is Large. *Xiv:1406.7686v2 [stat.ME]*

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model, *Biometrika*, **95:3**, 555–571.

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

Deville, J.C., Särndal, C. E. and Sautory, O. (1993). Generalized raking procedures in survey sampling, *Journal of the American Statistical Association*, **88:423**, 1013–1020.

Geuzinge, L., Rooijen, J. van and B.F.M. Bakker (2000). The use of administrative registers to reduce non-response bias in household surveys, *Netherlands Official Statistics,* **2000:2**, 32–39.

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.

Hardoon, D., Szedmak, S. and Shawe-Taylor, J. (2004). Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, **16**, 2639–2664.

Hotelling, A. (1939). Relation Between Two Sets of Variables. *Biometrika.* **28:4**, 321–377.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson, Prentice Hall.

Jolliffe, I. T. (1972). Discarding Variables in a Principal Component Analysis. I: Artificial Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **21:1**, 160–173.

Jolliffe, I. T. (1973). Discarding Variables in a Principal Component Analysis. II: Real Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **22:1**, 21–31.

Jolliffe, I. T. (1982). A Note on the Use of Principal Components in Regression. *Applied Statistics*, **31:3**, 300–303.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, **12:3**, 531–547.

Kott, P. S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology*, **32:2**, 133–142.

Kott, P. S. (2012). Exploring Some Uses for Instrumental-Variable Calibration Weighting. *Section on Survey Research Methods– JSM 2012*

Kreuter, F. and Olson, K. (2011). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods & Research,* **40:2**, 311–332.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*. 15:2,305–327.

Mansfield, Webster and Gunst (1977) An Analytic Variable Selection Technique for Principal Component Regression. *Applied Statistics*, **6**, 34–40.

McCabe, G. P. (1984). Principal Variables. *Technometrics,* **26:2**, 137–144.

McHenry, C. E. (1978). Computation of a Best Subset in Multivariate Analysis, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **27:3**, 291–296.

Parra, L. and Sajda, P. (2003). Blind Source Separation via Generalized Eigenvalue Decomposition. *Journal of Machine Learning Research,* **4**, 1261–1269.

Rizzo, L., Kalton, G., and Brick, M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, **22**, 43–53.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.

Särndal, C.-E. and Lundström, S. (2007). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, **24:2**, 167–191.

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

Schouten, B. (2007). A selection strategy for weighting variables under a Not-Missing-at-Random assumption, *Journal of Official Statistics,* **23**, 51–68.

Silva, N. and Skinner, C. J. (1997). Variable Selection for Regression estimation in Finite Populations. *Survey Methodology*, 23(1), 23–32.

Skinner, C. J. (1999). Calibration weighting and non-sampling errors. *Research in Official Statistics*, 2. 33–43.

Appendix

I. Asymptotic variance of the estimated coefficients of the propensity functions

Let

$$E_{pq}\left[\left(\mathbf{Z}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta}^{*})\mathbf{d}_{(r)}-\mathbf{T}_{z}\right)\left(\mathbf{Z}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta}^{*})\mathbf{d}_{(r)}-\mathbf{T}_{z}\right)^{t}\right]=\boldsymbol{\Pi}_{1}+\boldsymbol{\Pi}_{2} \tag{20}$$

where $\boldsymbol{\Pi}_1=\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\pi_l)d_kd_l\mathbf{Z}_k\mathbf{Z}_l^t$ and $\boldsymbol{\Pi}_2=\sum_U d_k(h(\mathbf{H}_k^t\boldsymbol{\delta}^*)-1)\mathbf{Z}_k\mathbf{Z}_k^t$ (see Chang and Kott, 2008).

Then

$$Avar\sqrt{n}\left(\hat{\boldsymbol{\delta}}_{(pc)}-\boldsymbol{\delta}^{*}\right)=\left[\mathbf{F}^{t}\mathbf{W}\mathbf{F}\right]^{-1}\mathbf{F}^{t}\mathbf{W}\boldsymbol{\Theta}\mathbf{W}\mathbf{F}\left[\mathbf{F}^{t}\mathbf{W}\mathbf{F}\right]^{-1}$$

with $\mathbf{F}=\mathbf{Z}^t\boldsymbol{\Psi}\mathbf{H}$ and $\boldsymbol{\Theta}=Avar\left[n^{-1/2}\left(\mathbf{Z}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta}^{*})\mathbf{d}_{(r)}-\mathbf{T}_{z}\right)\right]$

We choose

$$\mathbf{W}^{-1}=\boldsymbol{\Theta}$$

and obtain,

$$Avar\sqrt{n}\left(\hat{\boldsymbol{\delta}}_{(pc)}-\boldsymbol{\delta}^{*}\right)=\left[\left(\mathbf{Z}^{t}\boldsymbol{\Psi}\mathbf{H}\right)^{t}\boldsymbol{\Theta}^{-1}\left(\mathbf{Z}^{t}\boldsymbol{\Psi}\mathbf{H}\right)\right]^{-1} \tag{21}$$

Where, $\mathbf{W}=p\lim_{n\to\infty}\mathbf{W}_n$, is a positive definite matrix,
$(\mathbf{Z}^t\boldsymbol{\Psi}\mathbf{H})=p\lim_{n\to\infty}\frac{1}{n}E_{pq}\left(\mathbf{Z}_{(r)}^{t}\boldsymbol{\Psi}^{0}\mathbf{H}\right)$ and

$$\boldsymbol{\Theta}=p\lim_{n\to\infty}\frac{1}{n}E_{pq}\left[\left(\mathbf{Z}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta}^{*})\mathbf{d}_{(r)}-\mathbf{T}_{z}\right)\left(\mathbf{Z}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta}^{*})\mathbf{d}_{(r)}-\mathbf{T}_{z}\right)^{t}\right]$$

Alternatively, the calibration (8) is on estimated principal components, that is,

$$\hat{\mathbf{Z}}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta})\mathbf{d}_{(r)}-\hat{\mathbf{T}}_{z}=\mathbf{0} \tag{22}$$

where $\hat{\mathbf{T}}=\sum_s d_k\hat{\mathbf{Z}}_k$.
Observe that,

$$var_{pq}(\hat{\mathbf{Z}}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta}^{*})\mathbf{d}_{(r)}-\hat{\mathbf{T}}_{z})=V_1+V_2$$

where, $V_1=var_pE_q\left(\hat{\mathbf{Z}}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta}^{*})\mathbf{d}_{(r)}-\hat{\mathbf{T}}_{z}|s\right)$ and $V_2=E_pvar_q\left(\hat{\mathbf{Z}}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta}^{*})\mathbf{d}_{(r)}-\hat{\mathbf{T}}_{z}|s\right)$.

The first variance component is zero, implying that

$$E_{pq}\left[\left(\hat{\mathbf{Z}}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta}^{*})\mathbf{d}_{(r)}-\hat{\mathbf{T}}_{z}\right)\left(\hat{\mathbf{Z}}_{(r)}^{t}\boldsymbol{\Phi}(\boldsymbol{\delta}^{*})\mathbf{d}_{(r)}-\hat{\mathbf{T}}_{z}\right)^{t}\right]=\sum_{U}d_{k}(h(\mathbf{H}_{k}^{t}\boldsymbol{\delta}^{*})-1)\mathbf{Z}_{k}\mathbf{Z}_{k}^{t},$$

therefore, the sample version analogous to $(\mathbf{W})$ in (21) is
$\tilde{\mathbf{W}}=p\lim_{n\to\infty}\frac{1}{n}\sum_{U}d_{k}(h(\mathbf{H}_{k}^{t}\boldsymbol{\delta}^{*})-1)\mathbf{Z}_{k}\mathbf{Z}_{k}^{t}$.

II. Figures

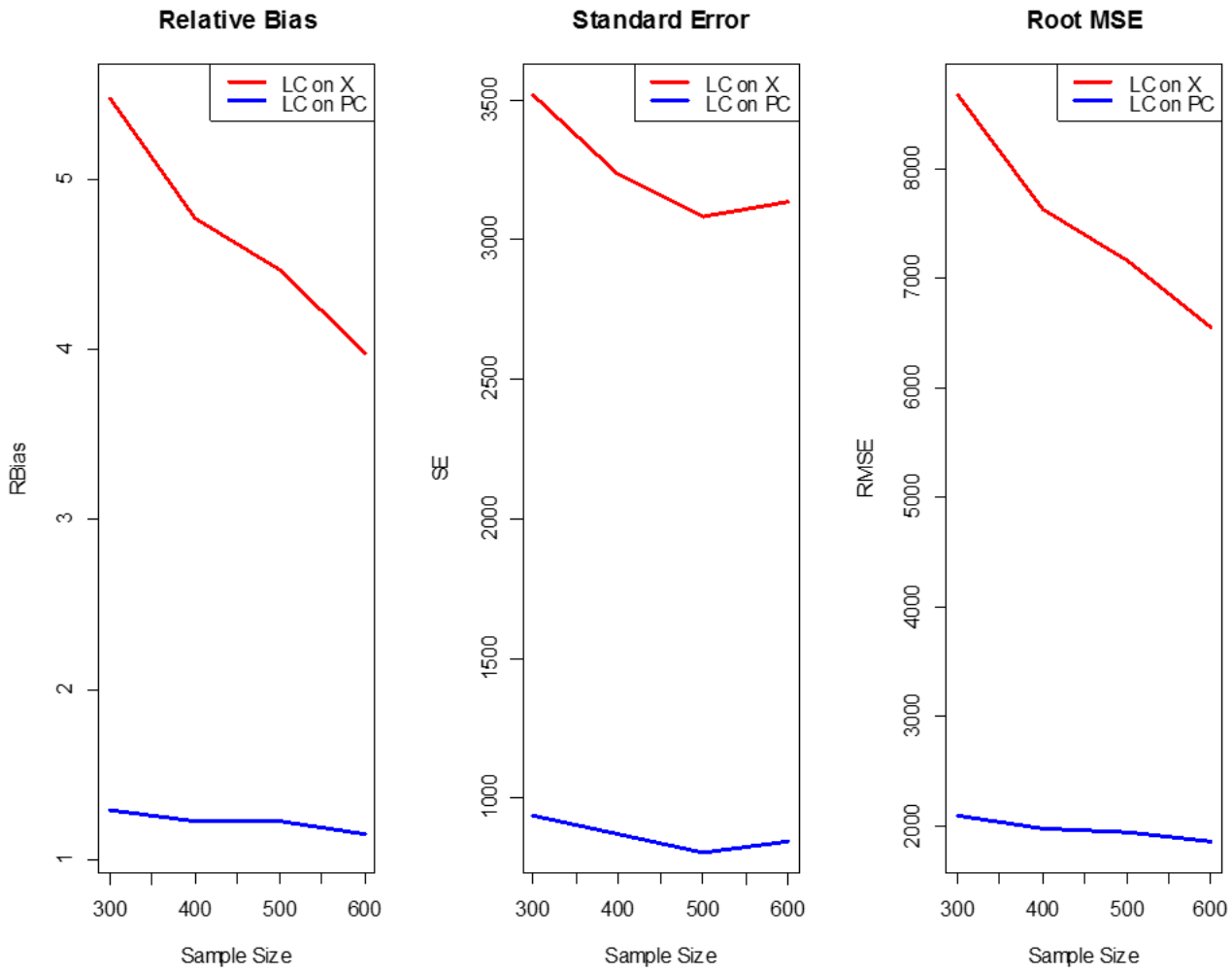

Figure 3: LC on original population auxiliary variables vs. LC on population PCs – Study 1.

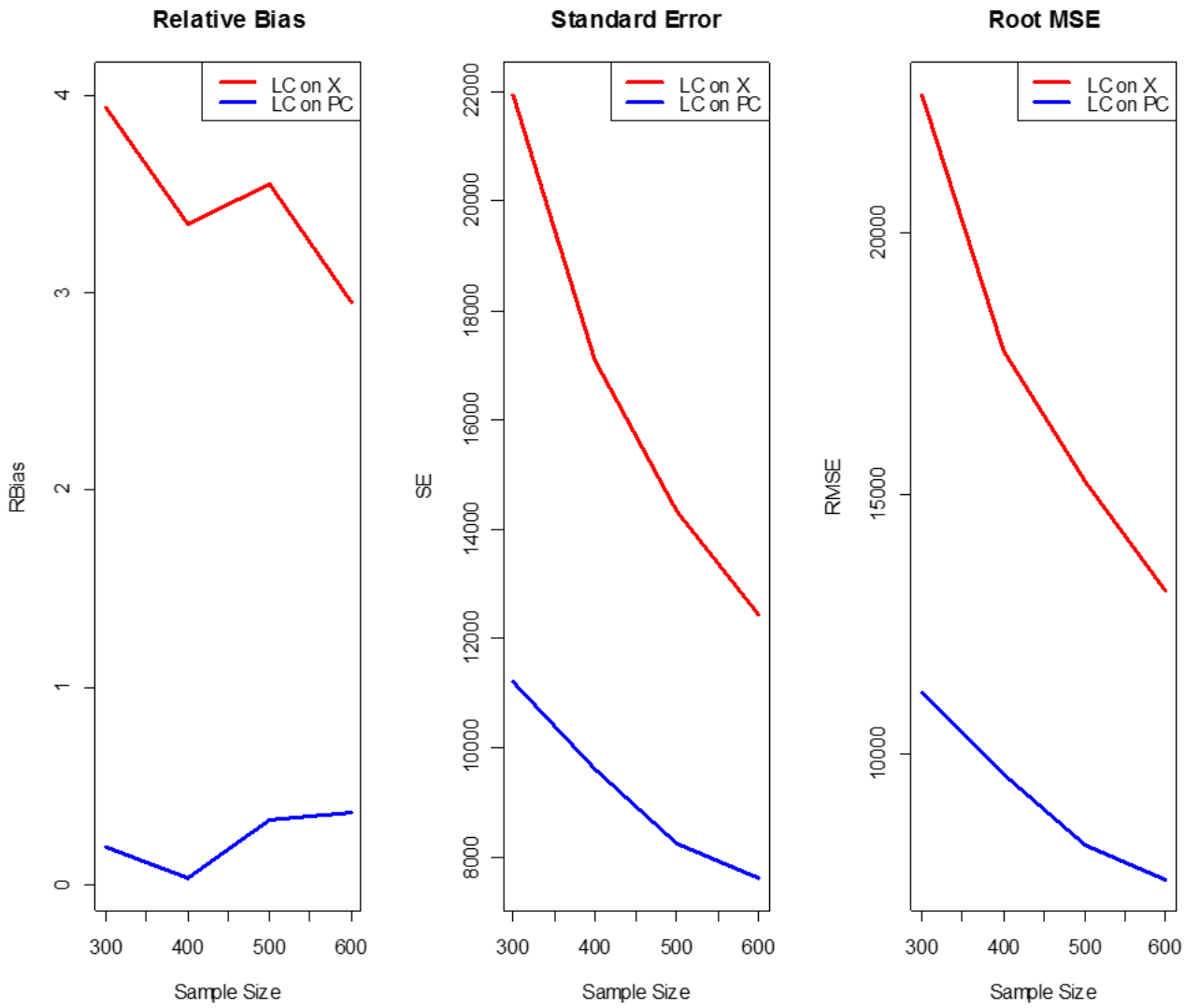

Figure 4: LC on original sample auxiliary variables vs. LC on sample PCs – Study 1.

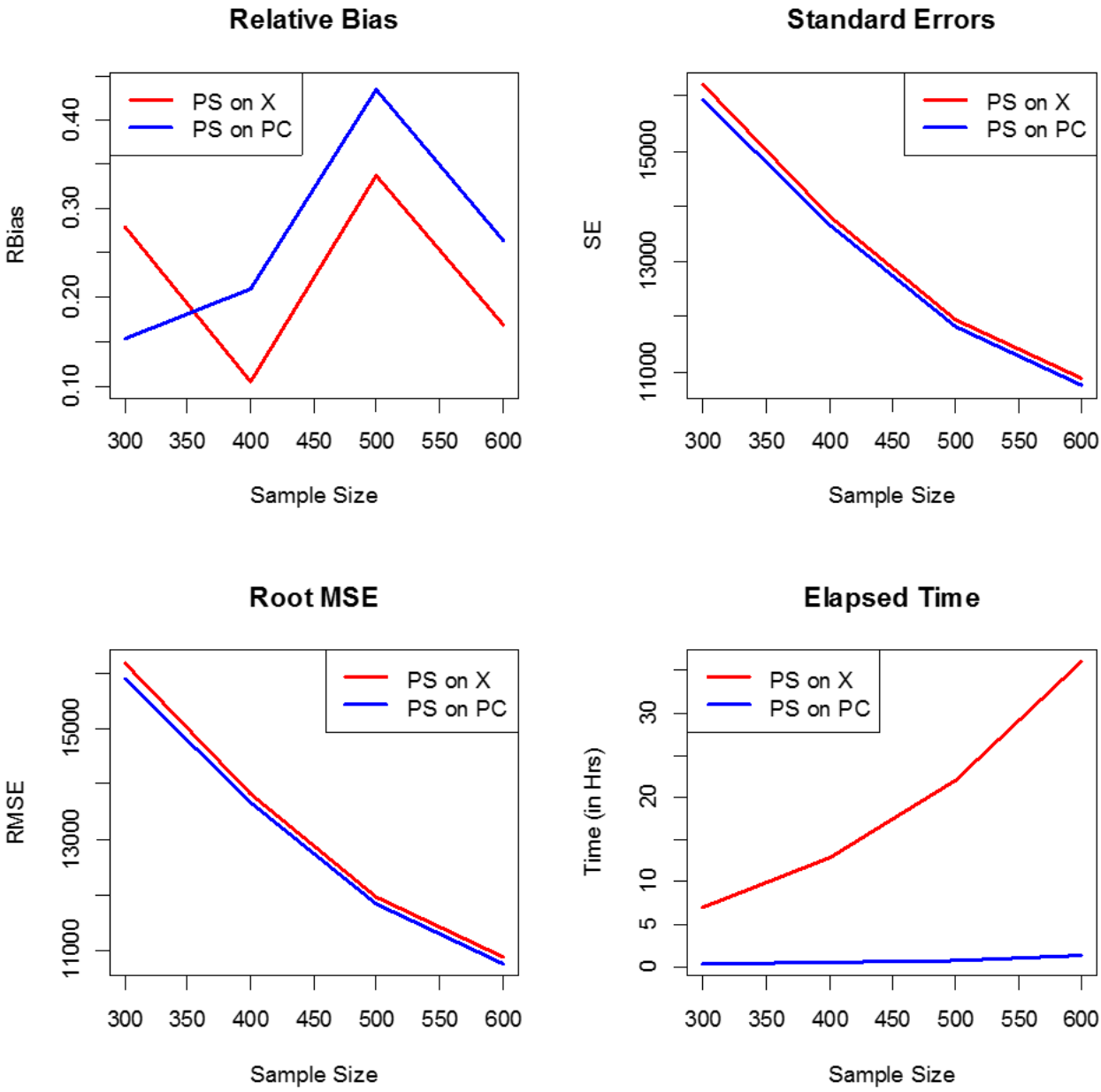

Figure 5: PSC on original population auxiliary variables vs. PSC on population PCs – Study 1.

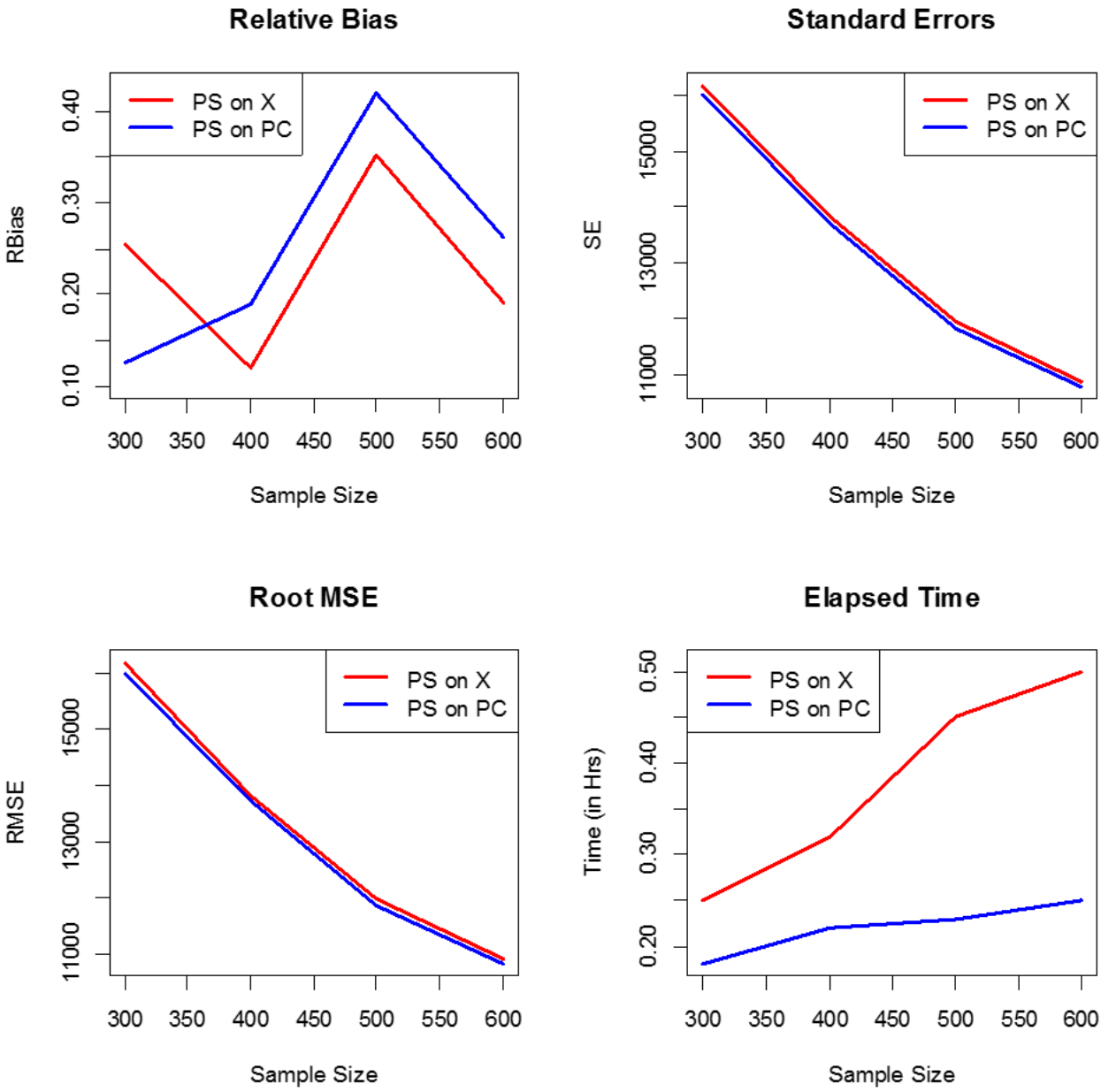

Figure 6: PSC on original sample auxiliary variables vs. PSC on sample PCs – Study 1.

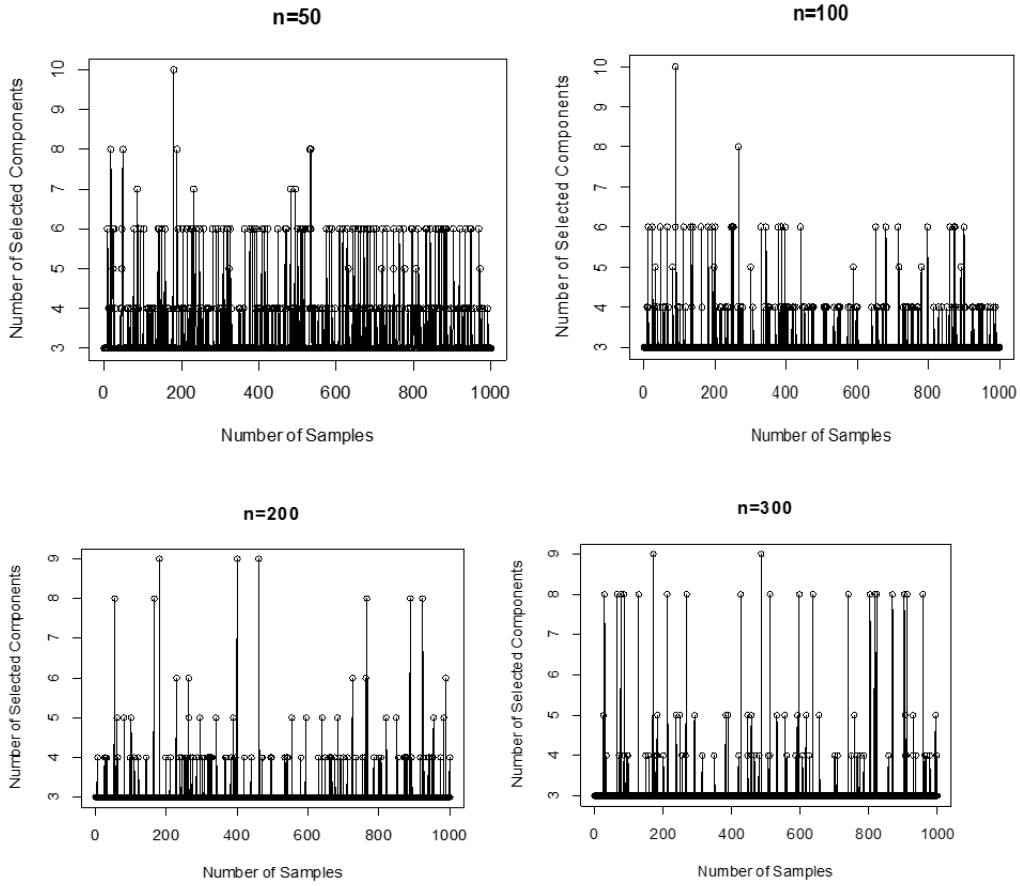

Figure 7: Behaviour of the number of selected components when sample sizes increase – Study 1.