Received: 00 Month 0000 | Revised: 00 Month 0000 | Accepted: 00 Month 0000

DOI: xxx/xxxx

ARTICLE TYPE

Reduction of optimal calibration dimension with a new optimal auxiliary vector for calibrated estimators of the distribution function.

Sergio Martínez*[1] | María del Mar Rueda[2] | María Dolores Illescas[3]

[1]Department of Mathematics, University of Almería, Almería, Spain
[2]Department of Statistics and Operations Research, University of Granada, Granada, Spain
[3]Department of Economics and Business, University of Almería, Almería, Spain

Correspondence
*Sergio Martínez. emailspuertas@ual.es

Present Address
Department of Mathematics. University of Almería. La Cañada de San Urbano, 04120, Almería, Spain

Summary

The calibration method[1] has been widely used to incorporate auxiliary information in the estimation of various parameters. Specifically,[2] adapted this method to estimate the distribution function, although their proposal is computationally simple, its efficiency depends on the selection of an auxiliary vector of points. This work deals with the problem of selecting the calibration auxiliary vector that minimize the asymptotic variance of the calibration estimator of distribution function. The optimal dimension of the optimal auxiliary vector is reduced considerably with respect to previous studies[3] so that with a smaller set of points the minimum of the asymptotic variance can be reached, which in turn allows to improve the efficiency of the estimates.

KEYWORDS:
Survey sampling, distribution function, auxiliary information, calibration

1 | INTRODUCTION

In sample surveys, auxiliary population information is sometimes used in the estimation stage to increase the precision of the estimators of a mean or total population. Previous literature has investigated the use of auxiliary information to improve the estimation of a finite population mean, however, previous studies have considered to a lesser extent the development of efficient methods to estimate the distribution function and the finite population quantiles by incorporating the auxiliary information. The estimation of finite population distribution function is an important issue because the distribution function can be more useful than means and totals[4].Through the finite population distribution function, parameters such as population quantiles can be obtained. More specifically, in economics, many indicators used in the poverty analisys are based on quantiles, since they analyze variables with skewed distributions such as income, and in such cases the median is a more suitable location measure than the mean. Moreover, poverty studies incorporate the analysis of wage inequality and income distribution thorugh percentile ratios[5,6,7].

In the last decade, the well-known calibration estimation method to estimate the population total[1] has been employed to develop new estimators which incorporates the auxiliary information available and it has become an important field of research in survey sampling[2,8,9,10,11].

Previous works[2,12,13] use different implementations of the calibration approach to obtain estimators of the distribution function and the quantiles. Under a general superpopulation model[14] propose a model-calibrated estimators that is optimal under a chosen model with respect to the anticipated variance. Although[14] considers a general sampling design, its proposal does not produce an estimator with the properties of a genuine distribution function unless the weight system is obtained by using a point t_0 for any t value, which restricts the efficiency of the estimator to a neighborhood of t_0. Additionally, the proposal[14] requires the estimation of certain superpopulation parameters that depend on the study variable, which may restrict its applicability in some cases and also require additional conditions on the sampling design to maintain the asymptotic behavior of the proposed estimator[15].

Nonparametric regression ([16] and[2]), is also used for model-calibration estimation of the distribution function.[17] propose a new estimator for the distribution function that integrates ideas from model calibration and penalized calibration. The method[2] is computationally simple and it employs the calibration method by minimizing the chi-square distance subject to calibration equations that require the use of arbitrarily fixed values. One drawback of these estimators is that their efficiency depends on selected points. Under simple random sampling, the problem of optimal selection points in order to obtain the best estimation has been treated in previous works[3,18,19,20]. In fact, the work[3] obtained the optimal dimension and the optimal auxiliary vector for the estimator of the distribution function proposed in the work[2] and although this proposal do not generate a unique weight system that is optimal for each point t, it produces an estimator that is computationally simple and is a genuine distribution function that can be used directly in the estimation of quantiles and poverty measures[21].

In many situations, the optimal auxiliary vector has a very high dimension, which makes the calibration process difficult and can also affect the efficiency of the estimator. Performing calibration with a high dimensional auxiliary dataset can be several problems: the variance of the calibration estimator can be increases and the optimisation procedure may fail.[22] showed that if too many auxiliary variables are used, the bias of the calibrated estimator increases and can become nonnegligible compared to the variance (over-calibration). Recently[23] theoretically prove that over-calibration may deteriorate the efficiency of the estimates. Various procedures have been suggested for variable selection.[22] computed the mean squared error (MSE) for all possible subsets of quantitative auxiliary variables and then chose the one producing the smallest MSE. Later,[24] used forward and stepwise selection based on the difference between the MSE of the prediction for two nested sets of variables. Alternatively, the least absolute shrinkage and selection operator (LASSO) ([25]) might be considered for selecting the best subsets. Once the best set of regressors has been selected, the calibration is performed on these variables alone. Another approach to consider is that of penalised calibration ([26]), which takes account of auxiliary information by attaching more or less importance according to its presumed explanatory power for the variable of interest. In a different way,[27] and[28] suggested applying principal component analysis for quantitative auxiliary variables in order to achieve a strong dimension reduction. These works are oriented to the estimation of linear parameters.

In this work, we intend to analyze whether it is possible to reduce the optimal dimension of the auxiliary vector proposed in the previous work[3]. The remainder of the article is organized as follow. After introducing the problem of distribution function estimation in Section 2 with the method proposed in research work[2] and the optimal auxiliary vector proposed in the previous work[3], in Section 3 we will analyze the conditions under which we can reduce the dimension of the optimal auxiliary vector. Then, Section 4 proposes a new calibration estimator based on the results of Section 3. Section 5 reports the results of an extensive simulation study run on a set of synthetic and real finite populations in which the performance of the proposed class of estimators is investigated for finite size samples. Section 6 provides some conclusions.

2 | CALIBRATION ESTIMATION OF THE DISTRIBUTION FUNCTION AND OPTIMAL AUXILIARY VECTOR

Let $U = \{1, \dots, N\}$ a finite population composed of N different units and let $s = \{1, 2, \dots, n\}$ a random sample of size n selected using a specified sampling design $p(\cdot)$ with first and second-order inclusion probabilities $\pi_k > 0$ and $\pi_{kl} > 0$ $k, l \in U$ respectively and $d_k = \pi_k^{-1}$ denotes the sampling design-basic weight for unit $k \in U$. Let y_k be the study variable and $\mathbf{x}_k^{'} = (x_{1k}, \dots, x_{Jk})$ be a vector of auxiliary variables at unit k. We assume that value $\mathbf{x}_k$ is available for all population units whereas the value y_k is available only for sample units. The distribution function $F_y(t)$ for the study variable y is defined as follow:

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k) \tag{1}$$

with

$$\Delta(t - y_k) = \begin{cases} 1 & \text{si } t \geq y_k \\ 0 & \text{si } t < y_k. \end{cases}$$

A design-based estimator of the distribution function $F_y(t)$ is the Horvitz–Thompson estimator, defined by

$$\widehat{F}_{YHT}(t) = \frac{1}{N} \sum_{k \in s} d_k \Delta(t - y_k). \tag{2}$$

The estimator $\widehat{F}_{YHT}(t)$ is unbiased, but it does not incorporate the auxiliary information provided by the auxiliary vector $\mathbf{x}$.

Several authors[2,13,29,30] have incorporated the auxiliary information to obtain new estimators of $F_y(t)$ through the calibration method[1]. The proposal[2] applies the calibration procedure from a pseudo-variable

$$g_k = (\widehat{\beta})' \mathbf{x}_k \text{ for } k = 1, 2, ...N \tag{3}$$

$$\widehat{\beta} = \left(\sum_{k \in s} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \cdot \sum_{k \in s} d_k \mathbf{x}_k y_k \tag{4}$$

With the variable g, the basic weights d_k are replaced by new calibrated weights ω_k through the minimization of the chi-square distance measure

$$\Phi_s = \sum_{k \in s} \frac{(\omega_k - d_k)^2}{d_k q_k} \tag{5}$$

subject to the calibration constrains

$$\frac{1}{N} \sum_{k \in s} \omega_k \Delta(t_j - g_k) = F_g(t_j) \quad j = 1, 2, \dots, P \tag{6}$$

where $F_g(t_j)$ denotes the finite distribution function of the pseudo-variable g_k evaluated at the points $t_j, \quad j = 1, 2, \dots, P$. We assume, with no loss in generality, $t_1 < t_2 < \dots t_P$. The values q_k are known positive constants unrelated to d_k. Following[2], we assume that the matrix T given by:

$$T = \sum_{k \in s} d_k q_k \Delta(\mathbf{t_g} - g_k) \Delta(\mathbf{t_g} - g_k)'$$

is nonsingular. With this calibration procedure, the calibration estimator obtained is:

$$\widehat{F}_{yc}(t) = \widehat{F}_{YHT}(t) + \left(F_g(\mathbf{t_g}) - \widehat{F}_{GHT}(\mathbf{t_g}) \right)' \cdot \widehat{D}(\mathbf{t_g}) \tag{7}$$

where

$$\widehat{D}(\mathbf{t_g}) = T^{-1} \cdot \sum_{k \in s} d_k q_k \Delta(\mathbf{t_g} - g_k) \Delta(t - y_k)$$

and $\widehat{F}_{GHT}(\mathbf{t_g})$ is the Horvitz-Thompson estimator of $F_g(\mathbf{t_g})$ evaluated at $\mathbf{t_g} = (t_1, \dots, t_P)'$.

The calibration estimator $\widehat{F}_{yc}(t)$ has the following asymptotic variance[2]:

$$AV(\widehat{F}_{yc}(t)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (d_k E_k)(d_l E_l) \tag{8}$$

where $E_k = \Delta(t - y_k) - \Delta(\mathbf{t_g} - g_k) \cdot D(\mathbf{t_g})$, with

$$D(\mathbf{t_g}) = \left(\sum_{k \in U} q_k \Delta(\mathbf{t_g} - g_k) \Delta(\mathbf{t_g} - g_k)' \right)^{-1} \cdot \left(\sum_{k \in U} q_k \Delta(\mathbf{t_g} - g_k) \Delta(t - y_k) \right). \tag{9}$$

As a consequence, the behavior of the estimator $\widehat{F}_{yc}(t)$ and its precision depends on the selection of the vector $\mathbf{t_g}$.

Previous works[3,19,20] treated, under simple random sampling without replacement and $q_k = c$ for all $k \in U$, the optimal selection of the vector $\mathbf{t_g}$ in order to minimize the asymptotic variance (8). In fact,[3] established the optimal dimension of $\mathbf{t_g}$ and its optimal value, for a given value t, through the definition of the sets:

$$A_t = \{ g_k : k \in U; y_k \le t \} = \{ a_1^t, a_2^t, \dots, a_{M_t}^t \} \quad \text{with} \quad a_h^t < a_{h+1}^t \quad \text{for} \quad h = 1, \dots, M_t - 1 \tag{10}$$

where M_t is the number of elements in the set A_t and

$$B_t = \{ b_1^t, b_2^t, \dots, b_{M_t}^t \} \tag{11}$$

with

$$\begin{array}{ll} b_1^t = \max_{l \in U_1} \{ g_l \} & \text{where } U_1 = \{ l \in U : g_l < a_1^t \} \\ b_h^t = \max_{l \in U_h} \{ g_l \} & \text{where } U_h = \{ l \in U : a_{h-1}^t < g_l < a_h^t \} \quad h = 2, 3, \dots, M_t \end{array}$$

and $b^t_h < b^t_{h+1}$ for $h = 1, \dots, M_t - 1$.

Thus,[3] established that the auxiliary vector $\mathbf{t}_g$ has optimal dimension $P = 2M_t$ if b^t_h exists for $h = 1, \dots, M_t$ and the optimal value of $\mathbf{t}_g$ is given by

$$\mathbf{t_{OPT}}(t) = (b^t_1, a^t_1, \dots, b^t_{M_t}, a^t_{M_t}). \tag{12}$$

If there are some values $j^t_1, j^t_2, \dots j^t_{p_t} \in \{1, \dots, M_t\}$; such as $b^t_{j_h}$ does not exits for $h = 1, 2, \dots p_t$ with $p_t \leq M_t$ and $j^t_h \neq j^t_q$ if $h \neq q$, the optimal dimension is given by $P = 2M_t - p_t$ and the optimal auxiliary vector $\mathbf{t_{OP}}$ is:

$$\mathbf{t_{OP}}(t) = (b^t_1, a^t_1 \dots, b^t_{j_1-1}, a^t_{j_1-1}, a^t_{j_1}, b^t_{j_1+1}, \dots, b^t_{j_h-1}, a^t_{j_h-1}, a^t_{j_h}, b^t_{j_h+1}, \dots b^t_{M_t}, a^t_{M_t}). \tag{13}$$

In the next section, we will analyze if the minimum of the asymptotic variance can be reached with a vector of less dimension and we will establish conditions under which the dimension of the optimal vector $\mathbf{t_{OPT}}(t)$, can be reduced under simple random sampling without replacement.

3 | DIMENSION REDUCTION OF THE OPTIMAL AUXILIARY VECTOR

In this section, we will analyze the conditions under which the dimension of the optimal vector $\mathbf{t_{OPT}}(t)$ can be reduced, that is, we will analyze the existence of a vector with a smaller dimension than $\mathbf{t_{OPT}}(t)$ that allows obtaining the minimum value of the asymptotic variance of the estimator $\widehat{F}_{yc}(t)$.

For the minimization of the asymptotic variance (8), we consider it as a function of a vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_P)$ of dimension P:

$$AV(\widehat{F}_{yc}(t)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl}(d_k \Gamma_k)(d_l \Gamma_l) \tag{14}$$

with $\Gamma_k = \Delta(t - y_k) - \Delta(\boldsymbol{\gamma} - g_k) \cdot D(\boldsymbol{\gamma})$, with $D(\boldsymbol{\gamma})$ given by (9).

Following[20], under simple random sampling without replacement and $q_k = c$ for all units in the population, the minimization of (14) is equivalent to the minimization of the function:

$$Q_t(\boldsymbol{\gamma}) = Q_t(\gamma_1, \dots, \gamma_P) = 2NF_y(t) \cdot K_t(\gamma_P) - \sum_{j=1}^{P} \frac{\left(K_t(\gamma_j) - K_t(\gamma_{j-1})\right)^2}{(F_g(\gamma_j) - F_g(\gamma_{j-1}))} - \left(K_t(\gamma_P)\right)^2 \tag{15}$$

with $K_t(\gamma_j) = \sum_{k \in U} \Delta(\gamma_j - g_k)\Delta(t - y_k)$, where we suppose that $F_g(\gamma_0) = 0$ and $K_t(\gamma_0) = 0$.

As mentioned above,[3] established the optimal dimension P and the minimum of (14) is reached at $\boldsymbol{\gamma} = \mathbf{t_{OP}}(t)$. However, there are cases where the optimal dimension has a high value so the calibration procedure has a plenty of constraints which raises the computational cost for calculating the estimator. For example, if we consider $t = y_{max}$ where

$$y_{max} = \max_{k \in U} y_k$$

the optimal auxiliary vector $\mathbf{t_{OP}}(t) = (a_1, a_2, \dots, a_M)$ can be reduced to the auxiliary vector $\boldsymbol{\gamma} = (a_M)$ (see Appendix A.1). Consequently, the optimal dimension can be reducted from M to 1.

In a similar way, we try to reduce the dimension of the auxiliary vector to reach the minimum of $Q_t(\boldsymbol{\gamma})$. For it, given a value t for which we want to estimate $F_y(t)$, we consider the sets A_M, A_t and B_t given by (A1); (10) and (11) respectively and for each $a_i \in A_M$ we define:

$$r_i = \text{Frequency of the } a_i$$

For the value t, we have:

$$A_t = \{a^t_1, a^t_2, \dots, a^t_{M_t}\} = \{a_{f^t_1}, a_{f^t_2}, \dots, a_{f^t_{M_t}}\}$$

where

$$\{f^t_1, f^t_2, \dots, f^t_{M_t}\} \subseteq \{1, 2, \dots, M\} \text{ and } f^t_1 < f^t_2 < \dots < f^t_{M_t}.$$

Similarly, we consider the following set:

$$C_t = \{g_k \, : \, k \in U; y_k > t\} = \{c^t_1, c^t_2, \dots, c^t_{S_t}\} = \{a_{l^t_1}, a_{l^t_2}, \dots, a_{l^t_{S_t}}\}$$

with

$$\{l_1^t, l_2^t, \dots, l_{S_t}^t\} \subseteq \{1, 2, \dots, M\} \text{ and } l_1^t < l_2^t < \cdots < l_{S_t}^t.$$

It is clear that $A_t \cup C_t = A_M$ and since for two different units k and j can be possible that $g_j = g_k = a_i$ and $y_k > t$ and $y_j < t$, not necessarily $A_t \cap C_t = \emptyset$. For the sets A_t and C_t we define:

$$p_i^t\text{=Frequency of the } a_i^t \text{ in } A_t$$
$$q_i^t\text{=Frequency of the } c_i^t \text{ in } C_t.$$

Next, we consider the following sets:

$$D_t = \{c_i \in C_t \, : \, q_i^t = r_i\} \tag{16}$$
$$Z_t = \{a_i^t \in A_t \, : \, q_i^t = 0\} = \{a_i^t \in A_t \, : \, a_i^t \notin C_t\} = A_t - C_t \tag{17}$$
$$F_t = \{a_i^t \in A_t \, : \, 0 < q_i^t < r_i\}. \tag{18}$$

It is easy to see that $D_t = A_M - A_t$ and consequently $A_t \cap D_t = \emptyset$. Futhermore, $B_t \subseteq D_t$; $A_t = Z_t \cup F_t$ and $Z_t \cap F_t = \emptyset$.

Firstly, if we suppose that $D_t = A_M$, we have $A_t = \emptyset$ and consequently $y_k > t$, $\forall k \in U$. In this case $F_y(t) = 0$ and we can calibrate with any auxiliary vector since

$$\widehat{F}_{yc}(t) = \frac{1}{N} \sum_{k \in s} \omega_k \Delta(t - y_k) = 0$$

regardless of the auxiliary vector, so we can calibrate with $\mathbf{t_{OP}}(t) = a_M$ with dimension 1.

Secondly, if we suppose that $D_t = \emptyset$, then $B_t = \emptyset$ and $A_t = A_M$. In this case, following[3] the optimal auxiliary vector $\mathbf{t_{OP}}(t) = (a_1, a_2, \dots, a_M)$.

Since $A_t = Z_t \cap F_t = A_M$, if we suppose that $Z_t = A_t = A_M$ then $t > y_k$ $\forall k \in U$ and this case is like the case where $t = y_{max}$ and although the optimal auxiliary vector is $\mathbf{t_{OP}}(t) = (a_1, a_2, \dots, a_M)$, we can reach the minimum value of $Q_\gamma(t)$ with the auxiliary vector $\gamma = (a_M)$.

On the other hand, if we consider that $Z_t = \emptyset$ and $F_t = A_M$ there is not reduction in the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ (see Appendix A.2).

Next, if we suppose that $Z_t \neq A_t = A_M$ and $F_t \neq A_t = A_M$ then there is a set $I_{F_t} = \{j_1, j_2, \dots, j_l\} \subseteq \{1, 2, \dots M\}$ such that $a_{j_i} \in F_t$ and therefore $q_{j_i}^t \neq 0$ for $i = 1, 2, \dots, l$.

Now, if we consider that $j_1 > 1$; $j_i - 1 > j_{(i-1)}$ for all $i = 2, \dots l$ and $j_l < M$; then for $h = 1, \dots j_1 - 1$; $q_h^t = 0$ and we have:

$$K_t(a_1) = \sum_{k \in U} \Delta(a_1 - g_k)\Delta(t - y_k) = NF_g(a_1)$$
$$\vdots$$
$$K_t(a_{(j_1-1)}) = \sum_{k \in U} \Delta(a_{j_1-1} - g_k)\Delta(t - y_k) = NF_g(a_{j_1-1}).$$

Similarly, for $j_i, \dots j_{i+1} - 1$ with $i = 1, 2, \dots l - 1$ we have:

$$K_t(a_{j_i}) = \sum_{k \in U} \Delta(a_{j_i} - g_k)\Delta(t - y_k) = NF_g(a_{j_i}) - \sum_{h=1}^{i} q_{j_h}^t$$
$$\vdots$$
$$K_t(a_{(j_{(i+1)}-1)}) = \sum_{k \in U} \Delta(a_{(j_{(i+1)}-1)} - g_k)\Delta(t - y_k) = NF_g(a_{(j_{(i+1)}-1)}) - \sum_{h=1}^{i} q_{j_h}^t$$

and finally, for $j_l, \dots, M$

$$K_t(a_{j_l}) = \sum_{k \in U} \Delta(a_{j_l} - g_k)\Delta(t - y_k) = NF_g(a_{j_l}) - \sum_{h=1}^{l} q_{j_h}^t$$
$$\vdots$$
$$K_t(a_M) = \sum_{k \in U} \Delta(a_M - g_k)\Delta(t - y_k) = NF_g(a_M) - \sum_{h=1}^{l} q_{j_h}^t = NF_y(t).$$

The minimum of $Q_t(\gamma)$ reached at the optimum auxiliary vector $\mathbf{t_{OP}}(t)$ is given by:

$$Q_t(\mathbf{t_{OP}}(t)) = (NF_y(t))^2 - \sum_{j=1}^{M} \frac{\left(K_t(a_j - K_t(a_{j-1}))\right)^2}{F_g(a_j) - F_g(a_{j-1})} =$$

$$= (NF_y(t))^2 - N^2 \cdot \sum_{\substack{j=1 \\ j \notin \{j_1,\ldots,j_l\}}}^{M} \frac{\left(F_g(a_j) - F_g(a_{j-1}))\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \sum_{j \in \{j_1,\ldots,j_l\}} \frac{\left(NF_g(a_j) - NF_g(a_{j-1})) - q_j^t\right)^2}{F_g(a_j) - F_g(a_{j-1})}$$

$$= (NF_y(t))^2 - N^2 \cdot \sum_{\substack{j=1 \\ j \notin I_{F_t}}}^{M} \left(F_g(a_j) - F_g(a_{j-1}))\right) - N^2 \cdot \sum_{j \in I_{F_t}} \left(F_g(a_j) - F_g(a_{j-1})\right) + 2N \sum_{j \in I_{F_t}} q_j^t - \sum_{j \in I_{F_t}} \frac{\left(q_j^t)\right)^2}{F_g(a_j) - F_g(a_{j-1})} =$$

$$= (NF_y(t))^2 - N^2 \cdot \sum_{j=1}^{M} \left(F_g(a_j) - F_g(a_{j-1}))\right) + 2N \sum_{j \in I_{F_t}} q_j^t - \sum_{j \in I_{F_t}} \frac{\left(q_j^t\right)^2}{F_g(a_j) - F_g(a_{j-1})} =$$

$$= (NF_y(t))^2 - N^2 + 2N \sum_{j \in I_{F_t}} q_j^t - \sum_{j \in I_{F_t}} \frac{\left(q_j^t\right)^2}{F_g(a_j) - F_g(a_{j-1})}. \tag{19}$$

The same value can be reached with the auxiliary vector $\gamma = (a_{(j_1-1)}, a_{j_1}, \ldots a_{(j_l-1)}, a_{j_l}, a_M)$. To see it, if we set a_{j0} such as $F_g(a_{j0}) = 0$ and we replace the vector γ in (15), we have:

$$Q_t(\gamma) = (N \cdot F_y(t))^2 - \sum_{h=1}^{l} \frac{\left(NF_g(a_{(j_h-1)}) - NF_g(a_{j_{(h-1)}}))\right)^2}{F_g(a_{(j_h-1)}) - F_g(a_{j_{(h-1)}})} - \sum_{h=1}^{l} \frac{\left(NF_g(a_{j_h}) - NF_g(a_{(j_h-1)}) - q_{j_h}^t)\right)^2}{F_g(a_{j_h}) - F_g(a_{(j_h-1)})}$$

$$- \frac{\left(NF_g(a_{(M)}) - NF_g(a_{j_l}))\right)^2}{F_g(a_M) - F_g(a_{j_l})} = (N \cdot F_y(t))^2 - N^2 \sum_{h=1}^{l} F_g(a_{(j_h-1)}) - F_g(a_{j_{(h-1)}}) - N^2 \sum_{h=1}^{l} F_g(a_{j_h}) - F_g(a_{(j_h-1)})$$

$$+2N \sum_{h=1}^{l} q_{j_h}^t - \sum_{h=1}^{l} \frac{\left(q_{j_h}^t\right)^2}{F_g(a_{j_h}) - F_g(a_{(j_h-1)})} - N^2(F_g(a_M) - F_g(a_{j_l}))$$

$$= 2N \sum_{h=1}^{l} q_{j_h}^t - \sum_{h=1}^{l} \frac{\left(q_{j_h}^t\right)^2}{F_g(a_{j_h}) - F_g(a_{(j_h-1)})}$$

and the auxiliary vector γ, with less dimension than $\mathbf{t_{OP}}(t)$, attain the minimum of $Q_t(\gamma)$.

Previously, we suppose that $a_{j_1} > a_1$ and $a_{j_l} < a_M$. If $a_{j_1} = a_1$ then it is easy to see that the minimum can be obtain at $\gamma = (a_1, a_{(j_2-1)}, a_{j_2}, \ldots a_{(j_l-1)}, a_{j_l}, a_M)$ that has less dimension than in the case $a_{j_1} > a_1$. In a similar way, if $a_{j_l} = a_M$ the minimum can be attained at $\gamma = (a_{j_1}, a_{(j_1-1)}, a_{j_2}, \ldots a_{(j_l-1)}, a_M)$.
Finally, we have assumed that $j_i - 1 > j_{(i-1)}$ for all $i = 2, \ldots l$. If there is a $h \in \{1, 2, \ldots l\}$ that $j_h - 1 = j_{(h-1)}$ it is easy to see that the minimum value of $Q_t(\gamma)$ is reached at $\gamma = (a_{(j_1-1)}, a_{j_1}, \ldots a_{j_{(h-1)}}, a_{j_h}, \ldots, a_{(j_l-1)}, a_{j_l}, a_M)$ with less dimension than in the case $j_i - 1 > j_{(i-1)}$ for all $i = 2, \ldots l$. Therefore, if $D_t = \emptyset$ we can reduce the optimal dimension when $F_t \neq A_t = A_M$.

Next, we consider the case where $D_t \neq \emptyset$ and $D_t \neq A_M$. Because $A_t = A_M - D_t$, we have $A_t \neq \emptyset$ and $A_t \neq A_M$. Therefore:

$$A_t = \{a_1^t, a_2^t, \ldots, a_{M_t}^t\} = \{a_{f_1^t}, a_{f_2^t}, \ldots, a_{f_{M_t}^t}\}$$

where $\{f_1^t, f_2^t, \ldots, f_{M_t}^t\} \subseteq \{1, 2, \ldots M\}$.
In this case, if we suppose that $B_t = \emptyset$ then $f_1^t = 1, \ldots, f_{M_t}^t = M_t$ and

$$A_t = \{a_1, a_2, \ldots a_{M_t}\} \quad ; \quad D_t = \{a_{M_t+1}, \ldots, a_M\}.$$

To see it, if we suppose that $f_1^t > 1$ then $a_{f_1^t} > a_{(f_1^t-1)} \geq a_1$ and consequently the set

$$U_1 = \{l \in U \,:\, g_l < a_1^t\} = \{l \in U \,:\, g_l < a_{f_1^t}\} \neq \emptyset$$

and $b_1^t = a_{(f_1^t-1)}$. Thus, $B_t \neq \emptyset$ (Contradiction). As a consequence, $a_{f_1^t} = a_1$.

If we suppose that for $i \in \{2, \dots, M_t\}$ such as $f^t_{(i-1)} = i - 1$ and we suppose that $f^t_i > i$ then $a_{f^t_{(i-1)}} = a_{(i-1)}$ and $a_{f^t_i} > a_i > a_{(i-1)} = a_{f^t_{(i-1)}}$. The set U_i is given by:

$$U_i = \{l \in U \, : \, a^t_{(i-1)} < g_l < a^t_i\} = \{l \in U \, : \, a_{f_{(i-1)}} < g_l < a_{f_i}\} \neq \emptyset$$

and $b^t_i = a_{(f^t_i - 1)}$. Thus, $B_t \neq \emptyset$ (Contradiction again). As a consequence, if $f^t_{(i-1)} = i - 1$ implies that $f^t_i = i$ for $i \in \{2, \dots, M_t\}$ and we have:

$$A_t = \{a_1, a_2, \dots a_{M_t}\}$$

If $M_t = M$ it is clear that $A_t = A_M$ and $D_t = \emptyset$ (Contradiction again). Therefore, $M_t < M$ and

$$D_t = A_M - A_t = \{a_{(M_t+1)}, a_{(M_t+2)}, \dots a_M\}$$

The optimal auxiliary vector is given by $\mathbf{t_{OP}}(t) = (a_1, \dots, a_{M_t})$ and in a similar way that in the previous cases, we can proof that if $F_t = A_t$, there is not a reduction in the optimal dimension. If $Z_t = A_t$, then we can attain the minimum of $Q_t(\gamma)$ at $\gamma = (a_{M_t})$. Finally, if $Z_t \neq A_t$ and $F_t \neq A_t$, and we suppose that

$$F_t = \{a_{j_1}, a_{j_2}, \dots a_{j_l}\}$$

we can reached the minimum value of $Q_\gamma(t)$ at $\gamma = (a_{(j_1-1)}, a_{j_1}, \dots, a_{(j_l-1)}, a_{j_l}, a_{M_t})$.

Next, under the assumption $D_t \neq \emptyset$ and $D_t \neq A_M$, we consider $B_t \neq \emptyset$. If we assume that b^t_h exists for $h = 1, \dots, M_t$, following the proposal[3], the optimal auxiliary vector $\mathbf{t}_{OPT}(t)$ is given by (12) and there is not a possible reduction in the dimension. On the other hand, if there are some values $j^t_1, j^t_2, \dots j^t_{p_t} \in \{1, \dots, M_t\}$; such as $b^t_{j_h}$ does not exits for $h = 1, 2, \dots p_t$ with $p_t \leq M_t$ and $j^t_h \neq j^t_q$ if $h \neq q$, the optimal dimension is given by $P = 2M_t - p_t$ and the optimal auxiliary vector $\mathbf{t_{OP}}$ is given by (13). Analogously, there are $p_1, p_2, \dots, p_{l_t} \in \{1, 2, \dots, M_t\}$ such as $b^t_{f_{p_h}}$ exists and following[3] there is not a reduction between the points $b^t_{f_{p_h}}$ and $a^t_{f_{p_h}}$. Alternatively, the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ can be expressed as follows

$$\mathbf{t_{OP}}(t) = (a_{f^t_1}, \dots, a_{f^t_{(p_1-1)}}, b_{f^t_{p_1}}, a_{f^t_{p_1}}, a_{f^t_{(p_1+1)}}, \dots, a_{f^t_{(p_h-1)}}, b_{f^t_{p_h}}, a_{f^t_{p_h}}, a_{f^t_{(p_h+1)}}, \dots, a_{f^t_{M_t}}) \tag{20}$$

If we suppose that $p_1 = 1$ then the value $b^t_{f_1}$ exists and consequently:

$$U_1 = \{l \in U \, : \, g_l < a^t_1\} = \{l \in U \, : \, g_l < a_{f^t_1}\} \neq \emptyset$$

then, $a_{f^t_1} > a_1$ and $f^t_1 > 1$. As a consequence $a_i \notin A_t$ with $i = 1, \dots f^t_1 - 1$ and $\{a_1, \dots a_{(f^t_1-1)}\} \subseteq D_t$ which implies that $b_{f^t_1} = a_{(f^t_1-1)}$. In this case, we have:

$$K_t(b_{f^t_1}) = \sum_{k \in U} \Delta(b_{f^t_1} - g_k)\Delta(t - y_k) = 0$$

$$K_t(a_{f^t_1}) = \sum_{k \in U} \Delta(a_{f^t_1} - g_k)\Delta(t - y_k) = NF_g(a_{f^t_1}) - \sum_{h=1}^{f^t_1 - 1} r_h - q^t_1 = NF_g(a_{f^t_1}) - H^t_1$$

and following[3] there is no possibility to reduce the number of points here.

On the other hand, if we suppose that $p_1 > 1$, for $i \in \{1, 2, \dots p_1 - 1\}$ the value $b^t_{f_i}$ does not exist and:

$$U_1 = \{l \in U \, : \, g_l < a^t_1\} = \{l \in U \, : \, g_l < a_{f^t_1}\} = \emptyset$$

$$U_i = \{l \in U \, : \, a^t_{(i-1)} < g_l < a^t_i\} = \{l \in U \, : \, a_{f^t_{(i-1)}} < g_l < a_{f^t_i}\} = \emptyset \text{ for } i = 2, \dots, p_1 - 1$$

therefore $a_{f^t_1} = a_1$ and $a_{f^t_i} = a_{(f_{(i-1)}+1)}$ for $i = 2, \dots, p_1 - 1$ and then:

$$a_{f^t_1} = a_1; a_{f^t_2} = a_2; \dots; a_{f^t_{(p_1-1)}} = a_{(p_1-1)}.$$

Thus, if $p_1 > 1$ we have $\{a_1, \dots, a_{(p_1-1)}\} \subseteq A_t$ and consequently the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ given by (20) can be expressed as follows:

$$\mathbf{t_{OP}}(t) = (a_1, \dots, a_{(p_1-1)}, b_{f^t_{p_1}}, a_{f^t_{p_1}}, \mathbf{t}_R) =$$

where $\mathbf{t}_R$ denotes

$$\mathbf{t}_R = (a_{f^t_{(p_1+1)}}, \dots, a_{f^t_{(p_h-1)}}, b_{f^t_{p_h}}, a_{f^t_{p_h}}, a_{f^t_{(p_h+1)}}, \dots, a_{f^t_{M_t}}).$$

On the contrary, for p_1 we have

$$U_{p_1} = \{l \in U \,:\, a^t_{(p_1-1)} < g_l < a^t_{p_1}\} = \{l \in U \,:\, a_{f^t_{(p_1-1)}} < g_l < a_{f^t_{p_1}}\} =$$
$$= \{l \in U \,:\, a_{(p_1-1)} < g_l < a_{f^t_{p_1}}\} \neq \emptyset$$

and then $a_{f^t_{p_1}} > a_{p_1}$ and $f^t_{p_1} > p_1$ which implies that

$$\{a_{p_1}, a_{(p_1+1)}, \dots a_{(f^t_{p_1}-1)}\} \subseteq D_t.$$

We consider the following sets:

$$A_{p_1} = \{a_1, \dots, a_{(p_1-1)}\} \tag{21}$$
$$Z_{p_1} = \{a_i \in A_{p_1} \,:\, q^t_i = 0\} \tag{22}$$
$$F_{p_1} = \{a_i \in A_{p_1} \,:\, 0 < q^t_i < r_i\} \tag{23}$$

Similarly to the previous cases, if we suppose that $Z_{p_1} = A_{p_1}$ and $F_{p_1} = \emptyset$, it is easy to see that we can delete a_i for $i = 1, 2, \dots p_1 - 2$ from the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ and we can calibrate only with the value $a_{(p_1-1)}$. To see it, it is clear that:

$$K_t(a_1) = \sum_{k \in U} \Delta(a_1 - g_k)\Delta(t - y_k) = NF_g(a_1)$$
$$\vdots$$
$$K_t(a_{(p_1-1)}) = \sum_{k \in U} \Delta(a_{j_1-1} - g_k)\Delta(t - y_k) = NF_g(a_{(p_1-1)}).$$

Because $\{a_{p_1}, a_{(p_1+1)}, \dots a_{(f_{p_1}-1)}\} \subseteq D_t$ it is easy to see that:

$$K_t(b_{f^t_{p_1}}) = \sum_{k \in U} \Delta(b_{f^t_{p_1}} - g_k)\Delta(t - y_k) = K_t(a_{(p_1-1)}) = NF_g(a_{(p_1-1)})$$
$$K_t(a_{f^t_{p_1}}) = \sum_{k \in U} \Delta(a_{f^t_{p_1}} - g_k)\Delta(t - y_k) = NF_g(a_{f^t_{p_1}}) - \sum_{h=p_1}^{f_{p_1}-1} r_h - q^t_{f_{p_1}} = NF_g(a_{f^t_{p_1}}) - H^t_{p_1}.$$

The minimum value of $Q_t(\gamma)$ at $\mathbf{t_{OP}}(t)$ can be expressed as follows:

$$Q_t(\mathbf{t_{OP}}(t)) = Q_t(a_1, \dots, a_{(p_1-1)}, b_{f^t_{p_1}}, a_{f^t_{p_1}}, \mathbf{t}_R) =$$
$$= Q_t(\mathbf{t}_R) - \sum_{j=1}^{p_1-1} \frac{\left(K_t(a_j) - K_t(a_{j-1})\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \frac{\left(K_t(b_{f^t_{p_1}}) - K_t(a_{(p_1-1)})\right)^2}{F_g(b_{f^t_{p_1}}) - F_g(a_{(p_1-1)})} - \frac{\left(K_t(a_{f^t_{p_1}}) - K_t(b_{f^t_{p_1}})\right)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})} =$$
$$= Q_t(\mathbf{t}_R) - N^2 \sum_{j=1}^{p_1-1} \left(F_g(a_j) - F_g(a_{j-1})\right) - \frac{\left(NF_g(a_{f^t_{p_1}}) - NF_g(a_{(p_1-1)}) - H^t_{p_1}\right)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})} =$$
$$= Q_t(\mathbf{t}_R) - N^2 F_g(a_{(p_1-1)}) - \frac{\left(NF_g(a_{f^t_{p_1}}) - NF_g(a_{(p_1-1)}) - H^t_{p_1}\right)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})}.$$

If we calibrate with the auxiliary vector $\gamma = (a_{(p_1-1)}, b_{f^t_{p_1}}, a_{f^t_{p_1}}, \mathbf{t}_R)$, we obtain the same value:

$$Q_t(\gamma) = Q_t(a_{(p_1-1)}, b_{f^t_{p_1}}, a_{f^t_{p_1}}, \mathbf{t}_R) =$$
$$= Q_t(\mathbf{t}_R) - \frac{\left(K_t(a_{(p_1-1)}))\right)^2}{F_g(a_{(p_1-1)})} - \frac{\left(K_t(b_{f^t_{p_1}}) - K_t(a_{(p_1-1)})\right)^2}{F_g(b_{f^t_{p_1}}) - F_g(a_{(p_1-1)})} - \frac{\left(K_t(a_{f^t_{p_1}}) - K_t(b_{f^t_{p_1}})\right)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})} =$$
$$= Q_t(\mathbf{t}_R) - N^2 F_g(a_{(p_1-1)}) - \frac{\left(NF_g(a_{f^t_{p_1}}) - NF_g(a_{(p_1-1)}) - H^t_{p_1}\right)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})}.$$

Thus, if $Z_{p_1} = A_{p_1}$ we can reduce the dimension of the auxiliary vector to attain the minimum value of $Q_t(\gamma)$.

Now, if we suppose that $F_{p_1} = A_{p_1}$ and $Z_{p_1} = \emptyset$, as in the previous cases, we cannot reduce the dimension of the subvector $(a_1, \dots, a_{(p_1-1)})$ in the auxiliary vector $(\mathbf{t_{OP}}(t))$.

Next, if we suppose that $Z_{p_1} \neq A_{p_1}$ and $F_{p_1} \neq A_{p_1}$ then there is a set $I_{F_{p_1}} = \{j_1^1, j_2^2, \dots, j_{l_1}^1\} \subseteq \{1, 2, \dots p_1 - 1\}$ such that $a_{j_h^1} \in F_{p_1}$ and therefore $q_{j_h^1}^t \neq 0$ for $h = 1, 2, \dots, l_1$.

Now, if we consider that $j_1^1 > 1$; $j_h^1 - 1 > j_{(h-1)}^1$ for all $h = 2, \dots l_1$ and $j_{l_1}^1 < p_1 - 1$; then for $v = 1, \dots j_1^1 - 1$; $q_v^t = 0$ and we have:

$$K_t(a_1) = \sum_{k \in U} \Delta(a_1 - g_k)\Delta(t - y_k) = NF_g(a_1)$$
$$\vdots$$
$$K_t(a_{(j_1^1-1)}) = \sum_{k \in U} \Delta(a_{j_1^1-1} - g_k)\Delta(t - y_k) = NF_g(a_{j_1^1-1})$$

Similarly, for $j_h^1, \dots j_{h+1}^1 - 1$ with $h = 1, 2, \dots l_1 - 1$; we have:

$$K_t(a_{j_h^1}) = \sum_{k \in U} \Delta(a_{j_h^1} - g_k)\Delta(t - y_k) = NF_g(a_{j_h^1}) - \sum_{v=1}^{h} q_{j_v^1}^t$$
$$\vdots$$
$$K_t(a_{(j_{(h+1)}^1-1)}) = \sum_{k \in U} \Delta(a_{(j_{(h+1)}^1-1)} - g_k)\Delta(t - y_k) = NF_g(a_{(j_{(h+1)}^1-1)}) - \sum_{v=1}^{h} q_{j_v^1}^t$$

and finally, for $j_{l_1}^1, \dots, p_1 - 1$

$$K_t(a_{j_{l_1}^1}) = \sum_{k \in U} \Delta(a_{j_{l_1}^1} - g_k)\Delta(t - y_k) = NF_g(a_{j_{l_1}^1}) - \sum_{v=1}^{l_1^1} q_{j_v^1}^t = NF_g(a_{j_{l_1}^1}) - L_1^t$$
$$\vdots$$
$$K_t(a_{(p_1-1)}) = \sum_{k \in U} \Delta(a_{(p_1-1)} - g_k)\Delta(t - y_k) = NF_g(a_{(p_1-1)}) - \sum_{v=1}^{l_1^1} q_{j_v^1}^t = NF_g(a_{(p_1-1)}) - L_1^t.$$

Again, because $\{a_{p_1}, a_{(p_1+1)}, \dots a_{(f_{p_1}-1)}\} \subseteq D_t$ we have:

$$K_t(b_{f_{p_1}^t}) = \sum_{k \in U} \Delta(b_{f_{p_1}^t} - g_k)\Delta(t - y_k) = K_t(a_{(p_1-1)}) = NF_g(a_{(p_1-1)}) - L_1^t$$

$$K_t(a_{f_{p_1}^t}) = \sum_{k \in U} \Delta(a_{f_{p_1}^t} - g_k)\Delta(t - y_k) = NF_g(a_{f_{p_1}^t}) - L_1^t - \sum_{h=p_1}^{f_{p_1}-1} r_h - q_{f_{p_1}}^t = NF_g(a_{f_{p_1}^t}) - L_1^t - H_{p_1}^t$$

The minimum of $Q_t(\gamma)$ at $(\mathbf{t_{OP}}(t))$ is given by:

$$Q_t(\mathbf{t_{OP}}(t)) = Q_t(a_1, \dots, a_{(p_1-1)}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R) =$$

$$= Q_t(\mathbf{t}_R) - \sum_{j=1}^{p_1-1} \frac{\left(K_t(a_j) - K_t(a_{j-1})\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \frac{\left(K_t(b_{f_{p_1}^t}) - K_t(a_{(p_1-1)})\right)^2}{F_g(b_{f_{p_1}^t}) - F_g(a_{(p_1-1)})} - \frac{\left(K_t(a_{f_{p_1}^t}) - K_t(b_{f_{p_1}^t})\right)^2}{F_g(a_{f_{p_1}^t}) - F_g(b_{f_{p_1}^t})} =$$

$$= Q_t(\mathbf{t}_R) - N^2 \sum_{j=1}^{p_1-1} \left(F_g(a_j) - F_g(a_{j-1})\right) + 2N \cdot L_1^t - \sum_{v=1}^{l_1^1} \frac{\left(q_{j_v^1}^t\right)^2}{F_g(a_{j_v^1}) - F_g(a_{(j_1^v-1)})} - \frac{\left(NF_g(a_{f_{p_1}^t}) - NF_g(a_{(p_1-1)}) - H_{p_1}^t\right)^2}{F_g(a_{f_{p_1}^t}) - F_g(b_{f_{p_1}^t})} =$$

$$= Q_t(\mathbf{t}_R) - N^2 F_g(a_{(p_1-1)}) + 2N \cdot L_1^t - \sum_{v=1}^{l_1^1} \frac{\left(q_{j_v^1}^t\right)^2}{F_g(a_{j_v^1}) - F_g(a_{(j_v^1-1)})} - \frac{\left(NF_g(a_{f_{p_1}^t}) - NF_g(a_{(p_1-1)}) - H_{p_1}^t\right)^2}{F_g(a_{f_{p_1}^t}) - F_g(b_{f_{p_1}^t})}.$$

If we calibrate with the following auxiliary vector:

$$\gamma = \left(a_{(j_1^1-1)}, a_{j_1^1}, \dots, a_{(j_{l_1}^1-1)}, a_{j_{l_1}^1}, a_{(p_1-1)}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R\right)$$

in a similar way to the previous cases, it is easy to see that:

$$Q_t\left(a_{(j_1^1-1)}, a_{j_1^1}, \dots, a_{(j_{l_1}^1-1)}, a_{j_{l_1}^1}, a_{(p_1-1)}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R\right) =$$

$$= Q_t(\mathbf{t}_R) - N^2 F_g(a_{(p_1-1)}) + 2N \cdot L_1^t - \sum_{v=1}^{l_1^1} \frac{\left(q_{j_v^1}^t\right)^2}{F_g(a_{j_v^1}) - F_g(a_{(j_v^1-1)})} - \frac{\left(NF_g(a_{f_{p_1}^t}) - NF_g(a_{(p_1-1)}) - H_{p_1}^t\right)^2}{F_g(a_{f_{p_1}^t}) - F_g(b_{f_{p_1}^t})}.$$

Then, if $Z_{p_1} \neq A_{p_1}$ and $F_{p_1} \neq A_{p_1}$ again we can reduce the dimension of the optimal vector $\mathbf{t_{OP}}(t)$.

Finally, we have assumed that $j_1^1 > 1$; $j_h^1 - 1 > j_{(h-1)}^1$ for all $h = 2, \dots l_1$ and $j_{l_1}^1 < p_1 - 1$. If there is a $h \in \{1, 2, \dots l\}$ that $j_h^1 - 1 = j_{(h-1)}^1$ it is easy to see that the minimum value of $Q_t(\gamma)$ is reached at

$$\gamma = \left(a_{(j_1^1-1)}, a_{j_1^1}, \dots a_{j_{(h-1)}^1}, a_{j_h^1}, \dots, a_{(j_{l_1}^1-1)}, a_{j_{l_1}^1}, a_{(p_1-1)}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R\right)$$

with less dimension than in the case $j_h^1 - 1 > j_{(h-1)}^1$ for all $h = 2, \dots l_1$. Similarly, if we suppose that $j_1^1 = 1$ then the minimum is obtained with

$$\gamma = \left(a_{j_1^1}, a_{(j_2^1-1)}, a_{j_2^1}, \dots a_{j_{(h-1)}^1}, a_{j_h^1}, \dots, a_{(j_{l_1}^1-1)}, a_{j_{l_1}^1}, a_{(p_1-1)}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R\right)$$

again, with less dimension than in the case $j_1^1 > 1$. In a similar way, if $j_{l_1}^1 = p_1 - 1$, the minimum of $Q_t(\gamma)$ is reached at

$$\gamma = \left(a_{(j_1^1-1)}, a_{j_1^1}, \dots a_{j_{(h-1)}^1}, a_{j_h^1}, \dots, a_{(j_{l_1}^1-1)}, a_{j_{l_1}^1}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R\right)$$

again, with less dimension than in the case $j_{l_1}^1 < p_1 - 1$.

Now, if we consider p_i with $i = 2, \dots, l_t$, we can extend the analysis for the dimension reduction of the optimal auxiliary vector in a similar way to the case p_1.(see Appendix A.3).

4 | THE NEW OPTIMAL ESTIMATOR WITH THE NEW OPTIMAL VECTOR

In the previous section we have theoretically demonstrated that the optimal vector $\mathbf{t_{OPT}}(t)$ proposed in[3] can be reduced in its dimension and we can minimize the variance given by (8) with a new optimal vector $\mathbf{t_{NEWOPT}}(t)$ of lower dimension. As with the original optimal vector $\mathbf{t_{OPT}}(t)$, the new vector $\mathbf{t_{NEWOPT}}(t)$ depends on unknown population values and therefore needs to be estimated. For it, in the same way as in[3], from the sample versions of the sets A_t, B_t, C_t, D_t, Z_t and F_t and the sample version of the function $Q_t(\gamma)$, an estimate $\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t)$ of the vector $\mathbf{t_{NEWOPT}}(t)$ can be obtained and we can define a new calibrated estimator $\widehat{F}_{CALNEWOPT}(t)$ for the distribution function $F_y(t)$, given by:

$$\widehat{F}_{CALNEWOPT}(t) = \widehat{F}_{YHT}(t) + \left(F_g(\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t)) - \widehat{F}_{GHT}(\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t))\right)' \cdot \widehat{D}\left(\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t)\right) \tag{24}$$

where

$$\widehat{D}\left(\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t)\right) = T^{-1} \cdot \sum_{k \in s} d_k q_k \Delta(\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t) - g_k) \Delta(t - y_k)$$

5 | SIMULATION STUDY

In this section, a simulation study was conducted to compare the performance of the proposed optimal estimator with other alternative estimators for the distribution function $F(t)$. The simulation study was programmed in R software [version 4.1.0] and it was necessary to develop a new code to calculate the estimators included in the simulation study. The precision of the proposed new optimal calibration estimator $\widehat{F}_{CALNEWOPT}(t)$ was compared with the following estimators, the Horvitz Thompson estimator, $\widehat{F}_{HT}$, the difference estimator[31], $\widehat{F}_D(t)$, the ratio estimator[31] $\widehat{F}_R(t)$, the Chambers-Dunstan estimator[32] $\widehat{F}_{CD}(t)$, the Rao-Kovar-Mantel estimator[31] $\widehat{F}_{RKM}(t)$, the calibration estimator[2] with $t_1 = Q_g(0.5)$, the population median, as point for calibration, $\widehat{F}_{CAL}(t)$, the calibration estimator[2] with three points $t_1 = Q_g(0.25)$, $t_2 = Q_g(0.5)$ and $t_3 = Q_g(0.75)$, the population quartiles, as points for calibration, $\widehat{F}_{CAL3}(t)$, the calibration estimator with one optimal point[18], $\widehat{F}_{CALMAX}(t)$ and finally the previous optimal calibration estimator[3] $\widehat{F}_{CALOPT}(t)$.

Both real populations and simulated populations were considered for the simulation study. Specifically, we considered a real population included in The R Datasets Package called DNase that provides data collected from an ELISA assay for recombinant

DNase protein in rat serum with population size $N = 176$. In addition, two simulated population called Simh and Simser were considered. The first one, Simh, is a population of size $N = 5000$ generated from the following superpopulation model:

$$y_k = 8 - 7.82/x_k + \epsilon_k$$

where x is a sample from a discrete uniform distribution in $\{1, 2, \ldots, 100\}$ and ϵ_k's are i.i.d. random variables from $N(0, 0.5/x_k)$.

The simulated population Simser is a population of size $N = 5882$ generated from the following superpopulation model:

$$y_k = x_k^2 + \epsilon_k$$

where x is a sample from a discrete uniform distribution in $\{-100, -99, \ldots, 100\}$ and ϵ_k's are i.i.d. random variables from $N(0, 10)$.

For each population included in the simulation study, we drawn by simple random sampling without replacement 1000 samples of several sizes. For each sample, we estimated the distribution function $F(t)$ through all the estimators considered in the study at 11 different values of t, namely the quantiles $Q_y(\alpha)$ for α=0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8 and 0.9.

To measure the performance of each estimator included in the study, we considered the average relative bias (AVRB) and the average relative efficiency (AVRE), defined as follow:

$$\text{AVRB}(t) = \frac{1}{11}\sum_{q=1}^{11} |\text{RB}(t_q)|, \quad \text{AVRE}(t) = \frac{1}{11}\sum_{q=1}^{11} \text{RE}(t_q)$$

where RB and RE are defined as

$$\text{RB}(t) = \frac{1}{B}\sum_{b=1}^{B} \frac{\hat{F}(t)_b - F_y(t)}{F_y(t)} \quad \text{and} \quad \text{RE}(t) = \frac{MSE[\hat{F}(t)]}{MSE[\hat{F}_{HT}(t)]}, \tag{25}$$

where b indexes the bth simulation run, $\hat{F}(t)$ is an estimator for the distribution function, $MSE[\hat{F}(t)] = B^{-1}\sum_{b=1}^{B}[\hat{F}(t)_b - F_y(t)]^2$ is the empirical mean square error for $\hat{F}(t)$ and $MSE[\hat{F}_{HT}(t)]$ is similarly defined for the Horvitz-Thompson estimator.

Given that the new estimator proposal $\hat{F}_{CALNEWOPT}(t)$ and the estimator $\hat{F}_{CALOPT}(t)$ are based on the minimization of (15), it is possible that their behavior in terms of efficiency is similar and therefore it is necessary to analyze their behavior in greater detail. A reduced dimensionality in the auxiliary information set may reduce numerical issues in optimization procedures and also avoid the presence of unstable calibration weights (both negative weights and huge weights). Therefore, for each of the eleven estimation points t_q, we compared the dimension of the optimum auxiliary vector used in each estimators $\hat{F}_{CALOPT}(t)$ and $\hat{F}_{CALNEWOPT}(t)$. For it, we considered the mean dimension and the variance of the dimension:

$$\text{MD}(\hat{F}(t_q)) = \frac{1}{B}\sum_{b=1}^{B} \text{DIM}(\mathbf{t_{qopt}}), \quad \text{VD}(\hat{F}(t_q)) = \frac{1}{B}\sum_{b=1}^{B}\left(\text{DIM}(\mathbf{t_{qopt}}) - \text{MD}t_q\right)^2$$

where $\hat{F}$ can be $\hat{F}_{CALOPT}(t)$ or $\hat{F}_{CALNEWOPT}(t)$, $\mathbf{t_{qopt}}$ denote the optimum auxiliary vector used with the point t_q and $\text{DIM}(\mathbf{t_{qopt}})$ denote the dimension of $(\mathbf{t_{qopt}})$.

Additionally, because a limited number of variables may reduce the execution time to resolve the calibration procedure, we compared for each estimation point the execution time in calculating the estimators using the following measure:

$$\text{RT}(t) = \frac{\text{TIME}(\hat{F}_{CALNEWOPT}(t))}{\text{TIME}(\hat{F}_{CALOPT}(t))}$$

where $\text{TIME}(\hat{F}_{CALOPT}(t))$ and $\text{TIME}(\hat{F}_{CALNEWOPT}(t))$ denote the running time for calculating $\hat{F}_{CALOPT}(t)$ and $\hat{F}_{CALNEWOPT}(t)$ respectively.

For the population DNase, Table 1 gives the values of AVRB and AVRE whereas Table 2 gives the values of MD; VD and RT. With respect to the results obtained for the bias and efficiency analysis, the estimators $\hat{F}_{CALOPT}(t)$ and $\hat{F}_{CALNEWOPT}(t)$ show the same behavior. Thus, both estimators present an adequate bias value, although for all sample sizes there are estimators that present a lower bias. In relation to efficiency, both estimators are clearly the most efficient. From results in Table 2 , as expected, the estimator $\hat{F}_{CALNEWOPT}(t)$ always presents a smaller dimension of the auxiliary vector used, but this reduction is quite modest for all sample sizes and therefore the reduction obtained in the execution time is also quite modest. This may be because the set F_t has a cardinal similar to the set A_t or the set B_t also has a cardinal similar to the set A_t. Consequently, the new proposal $\widehat{F}_{CALNEWOPT}(t)$ only achieves small reductions in the dimension of the auxiliary information used with respect to $\widehat{F}_{CALOPT}(t)$, that produces slight improvements in execution time and it is not considerable enough to improve the asymptotic behavior[23], although it does not deteriorate it either.

TABLE 1 Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared. Population: DNase.

	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 30$		$n = 32$		$n = 35$		$n = 37$	
$\widehat{F}_{HT}$	0.0064	1	0.0063	1	0.0037	1	0.0030	1
$\widehat{F}_{D}$	0.0136	0.4898	0.0219	0.4867	0.0173	0.4642	0.0153	0.4703
$\widehat{F}_{R}$	0.0083	0.4412	0.0040	0.4343	0.0061	0.4247	0.0024	0.4335
$\widehat{F}_{CD}$	0.1869	0.8930	0.1878	0.9150	0.1814	0.9441	0.1783	0.9749
$\widehat{F}_{RKM}$	0.0032	0.4128	0.0103	0.4124	0.0063	0.4023	0.0055	0.4068
$\widehat{F}_{CAL}$	0.0066	0.8575	0.0090	0.8377	0.0029	0.8881	0.0012	0.8436
$\widehat{F}_{CAL3}$	0.0046	0.3401	0.0035	0.3468	0.0053	0.3322	0.0015	0.3261
$\widehat{F}_{CALMAX}$	0.0057	0.2102	0.0079	0.2247	0.0052	0.1981	0.0050	0.1991
$\widehat{F}_{CALOPT}$	0.0047	0.1895	0.0059	0.1928	0.0030	0.1676	0.0024	0.1629
$\widehat{F}_{CALNEWOPT}$	0.0047	0.1895	0.0059	0.1928	0.0030	0.1676	0.0024	0.1629
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 40$		$n = 42$		$n = 45$		$n = 47$	
$\widehat{F}_{HT}$	0.0056	1	0.0032	1	0.0044	1	0.0052	1
$\widehat{F}_{D}$	0.0160	0.4620	0.0102	0.4670	0.0089	0.4675	0.0108	0.4481
$\widehat{F}_{R}$	0.0018	0.4241	0.0033	0.4380	0.0045	0.4389	0.0041	0.4160
$\widehat{F}_{CD}$	0.1759	1.0417	0.1703	1.0957	0.1625	1.1100	0.1598	1.1054
$\widehat{F}_{RKM}$	0.0076	0.4012	0.0023	0.4070	0.0011	0.4136	0.0017	0.4014
$\widehat{F}_{CAL}$	0.0049	0.8386	0.0032	0.8394	0.0034	0.9319	0.0034	0.8700
$\widehat{F}_{CAL3}$	0.0008	0.3397	0.0024	0.3466	0.0033	0.3442	0.0022	0.3273
$\widehat{F}_{CALMAX}$	0.0043	0.1833	0.0028	0.1938	0.0031	0.1962	0.0036	0.1972
$\widehat{F}_{CALOPT}$	0.0022	0.1530	0.0013	0.1582	0.0011	0.1557	0.0019	0.1516
$\widehat{F}_{CALNEWOPT}$	0.0022	0.1530	0.0013	0.1582	0.0011	0.1557	0.0019	0.1516

Tables 3 and 4 provide the results obtained for the Simh population. From the results of Table 3 (AVRB and AVRE) we can again observe that $\widehat{F}_{CALOPT}(t)$ and $\widehat{F}_{CALNEWOPT}(t)$ have the same behavior and they are the most efficient estimators for all sample sizes. Also, they also present a less bias for most of sample sizes. Additionally, we can highlight the bias and efficiency problems of $\widehat{F}_{CD}(t)$ because this estimator is biased when the relationship between y and x is not linear. As in the previous case, the dimension reduction analysis (Table 4) is essential to find out if $\widehat{F}_{CALNEWOPT}(t)$ is a better alternative than $\widehat{F}_{CALOPT}(t)$. In this case, from the results of Table 4 , we can verify that again there is a slight reduction in the optimal vector used in $\widehat{F}_{CALNEWOPT}(t)$ for the smalls and medium quantiles. On the contrary, there is a moderate reduction for the higher quantiles where the dimension of the optimal vector used in $\widehat{F}_{CALOPT}(t)$ has a value between 10 and 20 while the dimension for $\widehat{F}_{CALNEWOPT}(t)$ remains between 2 and 5 for all sample sizes. Due to this reduction, the new estimator $\widehat{F}_{CALNEWOPT}(t)$ provides a considerable benefit in execution time, especially in the higher quantiles but as in the previous case, this moderate reduction in the dimension of the auxiliary information used in the calibration procedure does not allow an improvement in the asymptotic efficiency. Probably, in this case we have a considerable cardinal for the set Z_t, although the set F_t is not empty.

TABLE 2 Average dimension(MD), variance dimension (VD) and comparison of execution time (RT) of the estimators $\widehat{F}_{CALOPT}$ and $\widehat{F}_{CALNEWOPT}$. Population: DNase.

	$n = 30$					$n = 32$				
	MD		VD		RT	MD		VD		RT
	OPT	*NEWOPT*	*OPT*	*NEWOPT*		*OPT*	*NEWOPT*	*OPT*	*NEWOPT*	
t_1	1	1	0	0	1	1	1	0	0	1
t_2	1.929	1.698	0.257	0.459	0.875	1.920	1.742	0.271	0.438	0.556
t_3	1.976	1	0.153	0	0.600	1.978	1	0.147	0	0.217
t_4	2.781	1.738	0.419	0.440	0.846	2.807	1.761	0.405	0.427	0.467
t_5	3.658	1.668	0.499	0.471	0.813	3.665	1.670	0.491	0.470	0.895
t_6	3.951	1	0.225	0	0.238	3.960	1	0.196	0	0.412
t_7	4.915	1.515	0.286	0.500	0.375	4.939	1.530	0.244	0.499	0.849
t_8	5.883	1.739	0.343	0.439	0.579	5.897	1.763	0.307	0.425	0.261
t_9	5.921	1	0.288	0	0.556	5.938	1	0.241	0	0.600
t_{10}	6.723	1.733	0.476	0.443	0.571	6.765	1.756	0.452	0.430	0.909
t_{11}	7.502	1.573	0.580	0.495	0.536	7.578	1.632	0.541	0.483	0.950

	$n = 35$					$n = 37$				
	MD		VD		RT	MD		VD		RT
	OPT	*NEWOPT*	*OPT*	*NEWOPT*		*OPT*	*NEWOPT*	*OPT*	*NEWOPT*	
t_1	1	1	0	0	1	1	1	0	0	0.187
t_2	1.951	1.816	0.216	0.388	0.972	1.960	1.815	0.196	0.388	0.974
t_3	1.989	1	0.104	0	0.379	1.990	1	0.100	0	0.250
t_4	2.853	1.803	0.357	0.398	0.762	2.876	1.847	0.330	0.360	0.941
t_5	3.714	1.714	0.465	0.452	0.650	3.749	1.742	0.438	0.438	0.615
t_6	3.977	1	0.150	0	0.375	3.986	1	0.118	0	0.833
t_7	4.962	1.608	0.196	0.488	0.524	4.971	1.572	0.168	0.495	0.773
t_8	5.936	1.817	0.249	0.387	0.515	5.948	1.803	0.231	0.398	0.599
t_9	5.969	1	0.173	0	0.125	5.975	1	0.162	0	0.471
t_{10}	6.823	1.810	0.382	0.392	0.769	6.865	1.864	0.353	0.343	0.905
t_{11}	7.650	1.671	0.494	0.470	0.667	7.654	1.677	0.497	0.468	0.714

	$n = 40$					$n = 42$				
	MD		VD		RT	MD		VD		RT
	OPT	*NEWOPT*	*OPT*	*NEWOPT*		*OPT*	*NEWOPT*	*OPT*	*NEWOPT*	
t_1	1	1	0	0	1	1	0	0	1	
t_2	1.983	1.862	0.129	0.345	0.760	1.982	1.889	0.133	0.314	0.982
t_3	1.998	1	0.0447	0	0.818	1.996	1	0.063	0	0.353
t_4	2.898	1.871	0.303	0.335	0.706	2.920	1.895	0.271	0.307	0.753
t_5	3.793	1.789	0.405	0.408	0.278	3.782	1.781	0.413	0.414	0.647
t_6	3.996	1	0.063	0	0.346	3.993	1	0.083	0	0.500
t_7	4.989	1.613	0.104	0.487	0.615	4.989	1.665	0.104	0.472	0.698
t_8	5.974	1.869	0.159	0.338	0.587	5.979	1.880	0.143	0.325	0.440
t_9	5.991	1	0.094	0	0	5.992	1	0.089	0.403	0.214
t_{10}	6.909	1.882	0.291	0.323	0.814	6.893	1.875	0.309	0.331	0.692
t_{11}	7.731	1.735	0.446	0.441	0.773	7.750	1.756	0.433	0.430	0.400

	$n = 45$					$n = 47$				
	MD		VD		RT	MD		VD		RT
	OPT	*NEWOPT*	*OPT*	*NEWOPT*		*OPT*	*NEWOPT*	*OPT*	*NEWOPT*	
t_1	1	1	0	0	1	1	1	0	0	1
t_2	1.989	1.905	0.104	0.293	0.991	1.993	1.920	0.083	0.271	0.882
t_3	1.998	1	0.045	0	0.467	1.999	1	0.032	0	0.600
t_4	2.929	1.914	0.257	0.281	0.793	2.937	1.929	0.243	0.257	0.733
t_5	3.834	1.829	0.372	0.377	0.882	3.863	1.863	0.347	0.344	0.805
t_6	3.994	1	0.077	0	0.083	3.997	1	0.055	0	0.556
t_7	4.990	1.698	0.100	0.459	0.786	4.995	1.705	0.071	0.456	0.529
t_8	5.981	1.908	0.137	0.289	0.450	5.982	1.920	0.133	0.271	0.857
t_9	5.993	1	0.083	0	0.136	5.995	1	0.071	0	0.368
t_{10}	6.928	1.919	0.262	0.273	0.875	6.931	1.918	0.254	0.275	0.615
t_{11}	7.790	1.794	0.415	0.405	0.782	7.782	1.784	0.416	0.412	0.643

TABLE 3 Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared. Population: Simh.

	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 75$		$n = 100$		$n = 125$		$n = 150$	
$\widehat{F}_{HT}$	0.0066	1	0.0018	1	0.0036	1	0.0034	1
$\widehat{F}_{D}$	0.0075	0.5584	0.0016	0.5648	0.0027	0.5809	0.0041	0.5616
$\widehat{F}_{R}$	0.0078	1.2980	0.0018	1.2368	0.0031	1.2648	0.0041	1.2889
$\widehat{F}_{CD}$	0.3645	9.9689	0.3708	13.4606	0.3662	16.9590	0.3651	19.7880
$\widehat{F}_{RKM}$	0.0080	0.3990	0.0055	0.3822	0.0038	0.3990	0.0044	0.3789
$\widehat{F}_{CAL}$	0.0059	0.9369	0.0012	0.8205	0.0016	0.8794	0.0037	0.8845
$\widehat{F}_{CAL3}$	0.0028	0.3639	0.0015	0.3642	0.0008	0.3749	0.0037	0.3545
$\widehat{F}_{CALMAX}$	0.0012	0.2112	0.0008	0.1961	0.0016	0.1924	0.0009	0.1862
$\widehat{F}_{CALOPT}$	0.0030	0.1800	0.0012	0.1591	0.0006	0.1555	0.0010	0.1441
$\widehat{F}_{CALNEWOPT}$	0.0030	0.1800	0.0012	0.1591	0.0006	0.1555	0.0010	0.1441
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 175$		$n = 200$		$n = 250$		$n = 300$	
$\widehat{F}_{HT}$	0.0025	1	0.0037	1	0.0026	1	0.0015	1
$\widehat{F}_{D}$	0.0025	0.5626	0.0032	0.5822	0.0013	0.5593	0.0011	0.5653
$\widehat{F}_{R}$	0.0023	1.3001	0.0044	1.2704	0.0021	1.3077	0.0009	1.2659
$\widehat{F}_{CD}$	0.3653	23.6861	0.3743	26.3930	0.3659	32.5335	0.3632	38.7995
$\widehat{F}_{RKM}$	0.0040	0.3886	0.0024	0.3926	0.0020	0.3770	0.0011	0.3891
$\widehat{F}_{CAL}$	0.0021	0.8935	0.0048	0.8732	0.0011	0.8603	0.0012	0.8778
$\widehat{F}_{CAL3}$	0.0020	0.3780	0.0022	0.3721	0.0011	0.3480	0.0007	0.3598
$\widehat{F}_{CALMAX}$	0.0007	0.1920	0.0008	0.1853	0.0008	0.1758	0.0006	0.1725
$\widehat{F}_{CALOPT}$	0.0003	0.1486	0.0007	0.1438	0.0008	0.1303	0.0005	0.1313
$\widehat{F}_{CALNEWOPT}$	0.0003	0.1486	0.0007	0.1438	0.0008	0.1303	0.0005	0.1313

For the Simser population, Tables 5 and 6 provide the results of the simulation study. In this case, Table 5 shows that $\widehat{F}_{CALOPT}(t)$ and $\widehat{F}_{CALNEWOPT}(t)$ do not present the same behavior and $\widehat{F}_{CALNEWOPT}(t)$ is the one with the least bias and the best efficiency of all the estimators included in the simulation study and it produce a considerable improvement in efficiency with respect to $\widehat{F}_{CALOPT}(t)$. For some of the estimators included in the simulation study, we can observe a worse efficiency than $\widehat{F}_{HT}(t)$, which may be caused by the absence of a linear relationship between y and x. Regarding dimension reduction analysis and efficiency in execution time, Table 6 shows that for all quantiles, the optimal vector of $\widehat{F}_{CALOPT}(t)$ increase its size when the sample size increases, especially in the larger quantiles, where we can observe very high dimensional optimal vectors. On the other hand, the dimension of the optimal vector for $\widehat{F}_{CALNEWOPT}(t)$ remains stable for all sample sizes and it always shows a value below 7 and in most cases its value is between 3 and 5. It represents a quite considerable reduction of the dimension that causes a quite remarkable improvement in execution, especially in the high quantiles and according to previous studies[23] this remarkable reduction allows $\widehat{F}_{CALNEWOPT}(t)$ to achieve an improvement in efficiency with respect to $\widehat{F}_{CALOPT}(t)$. In this case, the cardinal of the set Z_t is probably very high and it is similar to the cardinal of the set A_t, which implies that the set F_t has few elements.

TABLE 4 Average dimension(MD), variance dimension (VD) and comparison of execution time (RT) of the estimators $\widehat{F}_{CALOPT}$ and $\widehat{F}_{CALNEWOPT}$. Population: Simh.

	$n = 75$					$n = 100$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	2.355	0.543	1.395	0.489	0.557	2.471	0.527	1.489	0.500	0.780
t_2	4.168	0.529	1.44	0.584	0.400	4.313	0.503	1.597	0.655	0.299
t_3	5.18	0.560	1.55	0.639	0.710	5.341	0.522	1.739	0.676	0.541
t_4	6.289	0.616	1.73	0.697	0.364	6.47	0.551	1.91	0.696	0.335
t_5	8.552	0.657	2.277	0.739	0.519	8.769	0.533	2.471	0.675	0.343
t_6	10.509	0.702	2.416	0.795	0.504	10.78	0.633	2.681	0.811	0.327
t_7	12.514	0.863	2.831	0.998	0.367	12.833	0.769	3.162	0.956	0.388
t_8	14.475	0.822	2.923	0.956	0.242	14.807	0.645	3.239	0.877	0.351
t_9	15.664	0.928	3.1	1.0213	0.359	16.087	0.811	3.5	0.944	0.294
t_{10}	16.71	0.978	3.375	1.143	0.249	17.163	0.923	3.827	1.140	0.273
t_{11}	18.704	0.969	3.519	1.187	0.337	19.178	0.750	3.934	1.163	0.294
	$n = 125$					$n = 150$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	2.605	0.493	1.608	0.488	0.620	2.652	0.477	1.651	0.477	0.516
t_2	4.392	0.493	1.695	0.671	0.610	4.471	0.501	1.841	0.692	0.291
t_3	5.414	0.500	1.852	0.709	0.554	5.492	0.500	1.981	0.704	0.627
t_4	6.58	0.5018	2.041	0.714	0.488	6.673	0.469	2.224	0.687	0.659
t_5	8.883	0.456	2.64	0.628	0.339	8.961	0.414	2.775	0.573	0.409
t_6	10.928	0.541	2.865	0.734	0.488	11.058	0.517	3.033	0.717	0.377
t_7	12.988	0.658	3.415	0.892	0.343	13.092	0.621	3.578	0.876	0.342
t_8	14.95	0.586	3.412	0.824	0.208	15.09	0.520	3.657	0.812	0.275
t_9	16.277	0.753	3.698	0.936	0.308	16.394	0.708	3.862	0.896	0.304
t_{10}	17.42	0.809	4.127	1.096	0.353	17.526	0.810	4.293	1.031	0.241
t_{11}	19.438	0.641	4.333	1.089	0.262	19.532	0.584	4.564	1.087	0.278
	$n = 175$					$n = 200$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	2.719	0.450	1.719	0.450	0.602	2.763	0.426	1.763	0.426	0.603
t_2	4.507	0.500	1.929	0.699	0.711	4.581	0.494	1.991	0.688	0.613
t_3	5.551	0.498	2.056	0.712	0.450	5.588	0.492	2.167	0.701	0.277
t_4	6.714	0.454	2.288	0.657	0.466	6.776	0.417	2.431	0.634	0.674
t_5	9	0.364	2.858	0.508	0.372	9.022	0.360	2.901	0.485	0.366
t_6	11.062	0.500	3.088	0.685	0.473	11.117	0.494	3.211	0.727	0.328
t_7	13.178	0.617	3.74	0.848	0.440	13.214	0.631	3.846	0.835	0.355
t_8	15.136	0.542	3.765	0.795	0.299	15.191	0.495	3.897	0.754	0.310
t_9	16.474	0.694	4.003	0.859	0.303	16.576	0.667	4.209	0.830	0.326
t_{10}	17.632	0.806	4.514	0.981	0.340	17.726	0.799	4.652	1.021	0.260
t_{11}	19.617	0.524	4.747	1.021	0.268	19.652	0.523	4.891	1.009	0.227
	$n = 250$					$n = 300$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	2.832	0.374	1.832	0.374	0.702	2.894	0.308	1.894	0.308	0.763
t_2	4.659	0.474	2.135	0.691	0.522	4.714	0.452	2.29	0.668	0.519
t_3	5.675	0.469	2.319	0.668	0.537	5.758	0.429	2.517	0.588	0.610
t_4	6.858	0.349	2.596	0.563	0.480	6.93	0.255	2.712	0.479	0.336
t_5	9.069	0.304	3.01	0.387	0.408	9.115	0.346	3.09	0.385	0.369
t_6	11.178	0.465	3.319	0.685	0.453	11.242	0.467	3.392	0.702	0.297
t_7	13.312	0.635	4.061	0.844	0.349	13.405	0.621	4.209	0.817	0.297
t_8	15.268	0.490	4.056	0.702	0.272	15.357	0.506	4.211	0.699	0.325
t_9	16.67	0.649	4.365	0.786	0.330	16.797	0.664	4.582	0.787	0.322
t_{10}	17.897	0.766	4.914	0.980	0.310	18.041	0.735	5.179	0.904	0.304
t_{11}	19.777	0.419	5.157	0.902	0.323	19.821	0.386	5.317	0.892	0.294

TABLE 5 Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared. Population: Simser.

	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 75$		$n = 100$		$n = 125$		$n = 150$	
$\widehat{F}_{HT}$	0.0042	1	0.0049	1	0.0042	1	0.0032	1
$\widehat{F}_{D}$	0.0038	0.9963	0.0039	1.0008	0.0056	1.0022	0.0027	0.9999
$\widehat{F}_{R}$	0.0832	3.0296	0.0687	3.0130	0.0659	2.8977	0.0564	3.0535
$\widehat{F}_{CD}$	0.0305	0.8968	0.0215	0.9234	0.0196	0.9374	0.0117	0.9409
$\widehat{F}_{RKM}$	0.0177	1.0419	0.0106	1.0246	0.0196	1.0530	0.0110	1.0252
$\widehat{F}_{CAL}$	0.0271	1.5805	0.0199	1.5456	0.0274	1.5784	0.0201	1.5548
$\widehat{F}_{CAL3}$	0.0110	0.8375	0.0104	0.7839	0.0090	0.8138	0.0075	0.7860
$\widehat{F}_{CALMAX}$	0.0174	0.6580	0.0146	0.6495	0.0229	0.6857	0.0177	0.6580
$\widehat{F}_{CALOPT}$	0.0732	0.4007	0.0537	0.2894	0.0417	0.2307	0.0327	0.1799
$\widehat{F}_{CALNEWOPT}$	0.0028	0.1516	0.0017	0.1143	0.0015	0.0985	0.0016	0.0787
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 175$		$n = 200$		$n = 250$		$n = 300$	
$\widehat{F}_{HT}$	0.0029	1	0.0026	1	0.0011	1	0.0023	1
$\widehat{F}_{D}$	0.0037	1.0033	0.0022	0.9988	0.0013	1	0.0029	1.0016
$\widehat{F}_{R}$	0.0547	3.0697	0.0461	3.0468	0.0449	3.0664	0.0418	3.0488
$\widehat{F}_{CD}$	0.0094	0.9427	0.0139	0.9404	0.0127	0.9603	0.0121	0.9651
$\widehat{F}_{RKM}$	0.0129	1.0307	0.0072	1.0161	0.0079	1.0221	0.0090	1.0152
$\widehat{F}_{CAL}$	0.0215	1.5860	0.0143	1.4787	0.0160	1.5565	0.0169	1.5472
$\widehat{F}_{CAL3}$	0.0079	0.8379	0.0071	0.8280	0.0051	0.7523	0.0042	0.7902
$\widehat{F}_{CALMAX}$	0.0191	0.6812	0.0138	0.6426	0.0155	0.6591	0.0163	0.6747
$\widehat{F}_{CALOPT}$	0.0266	0.1529	0.0228	0.1248	0.0169	0.0900	0.0128	0.0733
$\widehat{F}_{CALNEWOPT}$	0.0010	0.0714	0.0007	0.0577	0.0007	0.0476	0.0004	0.0424

6 | DISCUSSION AND CONCLUSIONS

In recent years, the calibration technique has attracted significant attention in survey sampling research and survey applications. The calibration method allows obtaining more reliable estimates for a finite population by incorporating auxiliary information available in the population.

In this article, we investigate whether the optimal estimator in the proposal[3] (that can be applied direclty in the estimation of qunatile and poverty measures[21]) based on the calibration method for estimating the distribution function can be improved by reducing the dimension of the optimal vector used in the calibration process. Working with a reduced number of variables may reduce numerical problems related to optimization procedures and also limit the presence of negative, very large and unstable calibration weights. To do this, we have theoretically established the conditions under which a reduction in the dimension of the optimal vector is possible and through an extensive simulation study we have verified how the new estimator $\widehat{F}_{CALNEWOPT}(t)$ can avoid the problems associated with a high-dimensional auxiliary data and allows to improve the execution time maintaining (DNase and Simh) or even improving the efficiency[23] (Simser). Therefore, the new proposal is a more reliable option when carrying out real analyzes where large population sizes can lead to high-dimensional optimal vectors for $\widehat{F}_{CALOPT}(t)$, while $\widehat{F}_{CALNEWOPT}(t)$ can lead to a considerable reduction in this optimal dimension.

Further research is needed regarding the dimension reduction on calibration for the distribution function as our study presents certain limitations. Our paper is restricted to a simple random sampling design. The determination of the optimal vector for calibration (and its dimension) can be extended relatively easily to the case of self-weighted samples (for example, stratified samples

TABLE 6 Average dimension(MD), variance dimension (VD) and comparison of execution time (RT) of the estimators $\widehat{F}_{CALOPT}$ and $\widehat{F}_{CALNEWOPT}$. Population: Simser.

	$n = 75$					$n = 100$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	7.807	2.355	3.036	0.277	1.152	9.846	2.589	3.084	0.376	0.779
t_2	14.767	3.192	3.013	0.122	0.431	18.586	3.430	3.029	0.200	0.516
t_3	18.115	3.395	3.006	0.077	0.449	23.052	3.825	3.022	0.166	0.443
t_4	21.519	3.554	3.007	0.083	0.378	27.492	3.990	3.005	0.071	0.427
t_5	28.379	3.778	3.002	0.045	0.418	36.322	4.225	3.002	0.045	0.39
t_6	35.189	4.032	3.005	0.071	0.4	45.339	4.389	3.013	0.113	0.318
t_7	42.003	4.073	3.015	0.122	0.296	54.139	4.445	3.005	0.071	0.283
t_8	48.955	3.807	3	0	0.305	63.065	4.240	3.006	0.077	0.294
t_9	52.367	3.655	3.003	0.055	0.338	67.537	4.076	3.009	0.094	0.237
t_{10}	55.801	3.565	3.004	0.063	0.246	71.983	3.950	3.002	0.089	0.248
t_{11}	62.753	3.023	2.948	0.363	0.228	80.714	3.531	2.982	0.289	0.268

	$n = 125$					$n = 150$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	11.934	3.145	2.664	0.471	0.946	13.766	2.767	3.141	0.455	0.762
t_2	22.63	3.046	3.697	0.241	0.331	26.001	3.740	3.066	0.303	0.453
t_3	27.879	3.023	3.967	0.169	0.42	32.224	4.073	3.027	0.185	0.337
t_4	33.406	3.014	4.195	0.118	0.348	38.628	4.411	3.013	0.113	0.379
t_5	44.154	3.011	4.626	0.122	0.325	51.218	4.853	3.008	0.090	0.287
t_6	54.934	3.011	4.893	0.104	0.283	63.845	5.074	3.026	0.159	0.259
t_7	65.748	3.01	4.875	0.010	0.253	76.315	5.219	3.019	0.137	0.243
t_8	76.511	3.01	4.815	0.010	0.191	88.878	5.144	3.006	0.077	0.19
t_9	81.862	3.01	4.695	0.010	0.255	95.292	5.014	3.013	0.113	0.212
t_{10}	87.184	3.013	4.555	0.113	0.192	101.589	4.937	3.02	0.140	0.218
t_{11}	98.008	3	4.022	0.219	0.208	114.415	4.415	3.027	0.180	0.195

	$n = 175$					$n = 200$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	15.496	3.036	2.927	0.277	0.855	16.921	2.936	3.258	0.631	0.663
t_2	29.729	3.013	3.853	0.122	0.434	32.618	3.922	3.089	0.351	0.367
t_3	36.67	3.006	4.144	0.077	0.438	40.25	4.137	3.058	0.277	0.26
t_4	43.817	3.007	4.441	0.083	0.335	48.197	4.458	3.037	0.189	0.276
t_5	58.001	3.002	4.757	0.045	0.268	63.872	4.957	3.012	0.109	0.26
t_6	72.284	3.005	5.065	0.071	0.227	79.743	5.293	3.033	0.179	0.234
t_7	86.505	3.015	5.062	0.122	0.25	95.683	5.487	3.061	0.239	0.209
t_8	100.946	3	5.083	0	0.219	111.761	5.488	3.011	0.104	0.191
t_9	108.154	3.003	4.976	0.055	0.194	119.62	5.366	3.022	0.147	0.195
t_{10}	115.308	3.004	4.883	0.063	0.18	127.542	5.398	3.031	0.173	0.171
t_{11}	129.69	2.948	4.693	0.363	0.175	143.597	5.105	3.056	0.230	0.149

	$n = 250$					$n = 300$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	20.073	3.438	3.111	0.787	0.552	22.612	3.533	3.117	0.860	0.646
t_2	38.419	3.173	4.275	0.491	0.294	43.526	3.193	4.048	0.484	0.381
t_3	47.492	3.091	4.653	0.342	0.29	53.736	3.131	4.393	0.387	0.323
t_4	57.048	3.054	5.007	0.226	0.288	64.659	3.072	4.745	0.259	0.301
t_5	75.652	3.015	5.438	0.122	0.25	85.713	3.026	5.212	0.177	0.206
t_6	94.572	3.058	5.786	0.234	0.204	107.232	3.081	5.639	0.273	0.191
t_7	113.331	3.058	5.855	0.234	0.186	128.379	3.094	5.858	0.292	0.177
t_8	132.281	3.024	5.856	0.153	0.174	149.69	3.035	5.987	0.184	0.15
t_9	141.666	3.035	5.778	0.184	0.156	160.366	3.036	6.071	0.186	0.136
t_{10}	150.953	3.043	5.768	0.203	0.14	171.005	3.067	6.076	0.250	0.135
t_{11}	169.593	3.083	5.433	0.276	0.124	192.424	3.096	5.887	0.295	0.132

with proportional allocation). However, the case of sampling with unequal probabilities is more complex and the methodology to be used is not the same. In future research we try to extend the results of this paper from SRSWOR to complex sampling designs.

Another limitation of our work is that the estimator considered is based on a pseudo-variable g_k that assumes a linear relationship between variable y and the covariates. The selection of the optimal auxiliary vector for the estimators based on a non-linear model should be considered in future studies.

Financial disclosure

This study was partially supported by Ministerio de Educación y Ciencia (PID2019-106861RB-I00, Spain), IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033 and FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades (FQM170-UGR20).

Conflict of interest

The authors declare no potential conflict of interests.

How to cite this article: S. Martínez, M. Rueda, M.D. Illescas (2021), Reduction of optimal calibration dimension with a new optimal auxiliary vector for calibrated estimators of the distribution function, *Math. Method. Appl. Sci.*, *2021;00:1–6*.

APPENDIX

A SUPPLEMENTARY CASES FOR SECTION 3

A.1 Dimension reduction of the optimal auxiliary vector for $\mathbf{t = y_{max}}$

If we consider $t = y_{max}$ where

$$y_{max} = \max_{k \in U} y_k$$

the set A_t is given by

$$A_t = \{g_k \, : \, k \in U; y_k \leq t\} = \{a_1, a_2, \dots, a_M\} = A_M \tag{A1}$$

with $a_1 < a_2 < \cdots < a_M$ and M denotes the total number of different values that the pseudo variable g can take in the population U. As a consequence, $B_t = \emptyset$, the optimal dimension $P = M$ is the highest value and the optimal vector is given by:

$$\mathbf{t_{OP}}(t) = (a_1, a_2, \dots, a_M).$$

Our purpose is to analyze the possibility of obtaining the minimum value of (14) by means of a lower-dimensional auxiliary vector. Firstly, if we consider the value $t = y_{max}$, we have:

$$K_t(a_j) = \sum_{k \in U} \Delta(a_j - g_k)\Delta(y_{max} - y_k) = N \cdot F_g(a_j) \quad j = 1, \dots, M$$

and where we set a_0 so that $F_g(a_0) = 0$ and $K_t(a_0)$.

The value of $Q_t(\boldsymbol{\gamma})$ at $\boldsymbol{\gamma} = \mathbf{t_{OP}}(t)$ is given by:

$$Q_t(\mathbf{t_{OP}}(t)) = Q_t(a_1, a_2 \dots, a_M) = 2NF_y(y_{max}) \cdot K_t(a_M) - \sum_{j=1}^{M} \frac{\left(K_t(a_j) - K_t(a_{j-1})\right)^2}{(F_g(a_j) - F_g(a_{j-1}))} - \left(K_t(a_M)\right)^2 =$$

$$= 2NF_y(y_{max}) \cdot N \cdot F_g(a_M) - \sum_{j=1}^{M} \frac{\left(F_g(a_j) - F_g(a_{j-1})\right)^2}{(F_g(a_j) - F_g(a_{j-1}))} - \left(N \cdot F_g(a_M)\right)^2 =$$

$$= 2NF_y(y_{max}) \cdot N \cdot F_g(a_M) - \sum_{j=1}^{M} N^2 \cdot (F_g(a_j) - F_g(a_{j-1})) - \left(N \cdot F_g(a_M)\right)^2 = 2N^2 F_y(y_{max}) \cdot F_g(a_M) - \left(N \cdot F_g(a_M)\right)^2.$$

Since $F_y(y_{max}) = F_g(a_M) = 1$, it is clear that $Q_t(\mathbf{t_{OP}}(t)) = 0$ and consequently the minimum value of $Q_t(\boldsymbol{\gamma})$ for y_{max} is equal to 0.

On the other hand, if we consider $\boldsymbol{\gamma} = (a_M)$ then

$$Q_t(a_M) = 2NF_y(y_{max}) \cdot K_t(a_M) - \frac{\left(K_t(a_M)\right)^2}{(F_g(a_j))} - \left(K_t(a_M)\right)^2 =$$

$$Q_t(a_M) = 2N^2F_y(y_{max}) \cdot F_g(a_M) - 2N^2\left(F_g(a_M)\right)^2 = 0.$$

Thus, with the auxiliary vector $\boldsymbol{\gamma} = (a_M)$ the minimum value of $Q_t(\boldsymbol{\gamma})$ is reached and the optimal dimension can be reducted from M to 1. With the auxiliary vector $\boldsymbol{\gamma} = (a_M)$, the resulting calibration constraint is given by:

$$1 = F_g(a_M) = \frac{1}{N}\sum_{k\in s} \omega_k \Delta(a_M - g_k) = \frac{1}{N}\sum_{k\in s} \omega_k \tag{A2}$$

Under simple random sampling without replacement, the minimization of (5) subject to the condition (A2) results in $d_k = \omega_k$ since the basic weights d_k associated with the simple random sampling without replacement satisfy the condition (A2). Therefore, the minimum value of $Q_t(\boldsymbol{\gamma})$ for y_{max} is equal to 0.

A.2 Dimension reduction of the optimal auxiliary vector when $\mathbf{D_t = \emptyset}$; $\mathbf{Z_t = \emptyset}$ and $\mathbf{F_t = A_t = A_M}$

If we consider the case where $D_t = \emptyset$; $Z_t = \emptyset$ and $F_t = A_t = A_M$ there is not reduction in the optimal auxiliary vector $\mathbf{t_{OP}}(t)$. To see it, it is clear that $q_i^t \neq 0 \ \forall a_i \in A_M$ and consequently:

$$K_t(a_i) = \sum_{jnU} \Delta(a_i - g_k)\Delta(t - y_k) = N \cdot F_g(a_i) - \sum_{j=1}^{i} q_j^t \text{ for } i = 1, 2, \ldots, M.$$

Specifically, for $i = M$ we have:

$$K_t(a_M) = N \cdot F_g(a_M) - \sum_{j=1}^{i} q_j^t = NF_y(t).$$

The minimum value of $Q_t(\gamma)$ is reached at $\gamma = \mathbf{t_{OP}}(t)$ and is given by:

$$Q_t(\mathbf{t_{OP}}(t)) = 2NF_y(t) \cdot K_t(a_M) - \sum_{j=1}^{M} \frac{\left(K_t(a_j) - K_t(a_{j-1})\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \left(K_t(a_M)\right)^2.$$

Since $(K_t(a_j) - K_t(a_{j-1}) = N \cdot F_g(a_j) - N \cdot F_g(a_{j-1}) - q_j^t$, $Q_t(\mathbf{t_{OP}}(t))$ takes the following expression:

$$Q_t(\mathbf{t_{OP}}(t)) = (NF_y(t))^2 - \sum_{j=1}^{M} \frac{\left(N \cdot F_g(a_j) - N \cdot F_g(a_{j-1}) - q_j^t\right)^2}{F_g(a_j) - F_g(a_{j-1})} =$$

$$(NF_y(t))^2 - N^2 \cdot \sum_{j=1}^{M} \left(F_g(a_j) - F_g(a_{j-1})\right) + 2N \cdot \sum_{j=1}^{M} q_j^t - \sum_{j=1}^{M} \frac{(q_j^t)^2}{F_g(a_j) - F_g(a_{j-1})} =$$

$$(NF_y(t))^2 - N^2 \cdot F_g(a_M) + 2N \cdot \sum_{j=1}^{M} q_j^t - \sum_{j=1}^{M} \frac{(q_j^t)^2}{F_g(a_j) - F_g(a_{j-1})} = (NF_y(t))^2 - N^2 + 2N \cdot \sum_{j=1}^{M} q_j^t - \sum_{j=1}^{M} \frac{(q_j^t)^2}{F_g(a_j) - F_g(a_{j-1})}. \tag{A3}$$

If we consider an auxiliary vector where we delete some value $a_i \neq a_M$, i.e $\gamma = (a_1, \ldots a_{i-1}, a_{i+1}, \ldots, a_M)$, we can obtain in a similar way that

$$Q_t(\gamma) = 2NF_y(t) \cdot K_t(a_M) - \sum_{j=1}^{i-1} \frac{\left(K_t(a_j) - K_t(a_{j-1})\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \sum_{j=i+1}^{M} \frac{\left(K_t(a_j) - K_t(a_{j-1})\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \frac{\left(K_t(a_{i+1}) - K_t(a_{i-1})\right)^2}{F_g(a_{i+1}) - F_g(a_{i-1})} - \left(K_t(a_M)\right)^2 =$$

$$= (NF_y(t))^2 - N^2 + 2N \cdot \sum_{\substack{j=1 \\ j\neq i,i+1}}^{M} q_j^t + 2N(q_i^t + q_{i+1}^t) - \sum_{\substack{j=1 \\ j\neq i,i+1}}^{M} \frac{(q_j^t)^2}{F_g(a_j) - F_g(a_{j-1})} - \frac{(q_i^t + q_{i+1}^t)^2}{F_g(a_{i+1}) - F_g(a_{i-1})}.$$

As a consequence, $Q_t(\mathbf{t_{OP}}(t)) - Q_t(\gamma)$ is given by

$$Q_t(\mathbf{t_{OP}}(t)) - Q_t(\gamma) = -\frac{(q_i^t)^2}{F_g(a_i) - F_g(a_{i-1})} - \frac{(q_{i+1}^t)^2}{F_g(a_{i+1}) - F_g(a_i)} + \frac{(q_i^t + q_{i+1}^t)^2}{F_g(a_{i+1}) - F_g(a_{i-1})} =$$

$$= -\frac{(q_i^t)^2\big(F_g(a_{i+1}) - F_g(a_i)\big)}{\big(F_g(a_i) - F_g(a_{i-1})\big)\big((F_g(a_{i+1}) - F_g(a_{i-1})\big)} - \frac{(q_{i+1}^t)^2\big(F_g(a_i) - F_g(a_{i-1})\big)}{\big(F_g(a_{i+1}) - F_g(a_{i-1})\big)\big((F_g(a_{i+1}) - F_g(a_i)\big)} + \frac{2q_i^t \cdot q_{i+1}^t}{F_g(a_{i+1}) - F_g(a_{i-1})} =$$

$$= \Gamma \cdot \Big[-(q_i^t)^2\big(F_g(a_{i+1}) - F_g(a_i)\big)^2 - (q_{i+1}^t)^2\big(F_g(a_i) - F_g(a_{i-1})\big)^2 + 2q_i^t \cdot q_{i+1}^t\big(F_g(a_{i+1}) - F_g(a_i)\big)\big(F_g(a_i) - F_g(a_{i-1})\big)\Big] < 0$$

with

$$\Gamma_= \frac{1}{\big(F_g(a_i) - F_g(a_{i-1})\big)\big(F_g(a_{i+1}) - F_g(a_i)\big)\big(F_g(a_{i+1}) - F_g(a_{i-1})\big)}.$$

Consequently, when deleting some a_i, $Q_t(\mathbf{t_{OP}}(t)) < Q_t(\gamma)$.
If we delete the value a_M, i.e, we consider the auxiliary vector $\gamma = (a_1, \dots, a_{M-1})$, $Q_t(\gamma)$ takes the following expression:

$$Q_t(\gamma) = 2NF_y(t) \cdot K_t(a_{M-1}) - \sum_{j=1}^{M-1} \frac{\big(K_t(a_j - K_t(a_{j-1}))\big)^2}{F_g(a_j) - F_g(a_{j-1})} - (K_t(a_{M-1}))^2$$

On the other hand

$$Q_t(\mathbf{t_{OP}}(t)) = (NF_y(t))^2 - \sum_{j=1}^{M} \frac{\big(K_t(a_j - K_t(a_{j-1}))\big)^2}{F_g(a_j) - F_g(a_{j-1})}$$

and it easy to see that

$$Q_t(\mathbf{t_{OP}}(t)) - Q_t(\gamma) = (NF_y(t) - K_t(a_{M-1}))^2 - \frac{(NF_y(t) - K_t(a_{M-1}))^2}{F_g(a_M) - F_g(a_{M-1})} < 0.$$

Therefore, when we delete a_M; $Q_t(\mathbf{t_{OP}}(t)) < Q_t(\gamma)$.

Thus, if $Z_t = \emptyset$ and $F_t = A_t$ there is not a reduction in the auxiliary vector to reach the minimum of $Q_t(\gamma)$.

A.3 Dimension reduction of the optimal auxiliary vector for p_i; $i \in \{2, \dots, l_t\}$ when $\mathbf{D_t} \neq \emptyset$; $\mathbf{D_t} = \mathbf{A_M}$ and $\mathbf{B_t} \neq \emptyset$

Under the assumptions $\mathbf{D_t} \neq \emptyset$; $\mathbf{D_t} = \mathbf{A_M}$ and $\mathbf{B_t} \neq \emptyset$, if we consider p_i with $i = 2, \dots, l_t$, it is clear that $p_i > p_{(i-1)}$ and $f_{p_i}^t > f_{p_{(i-1)}}^t$. Moreover, because the value $b_{f_{p_i}}^t$ exists, this implies that $f_{p_i}^t > f_{p_{(i-1)}}^t + 1$.

If we suppose that $p_i = p_{(i-1)} + 1$, it is clear that $f_{p_i}^t = f_{(p_{(i-1)}+1)}^t$ and due to the value $b_{f_{p_i}}^t$ exists, the set

$$U_{p_{(i-1)}+1} = U_{p_i} = \{l \in U \,:\, a_{f_{p_{(i-1)}}^t} < g_l < a_{f_{p_{(i-1)}+1}^t}\} = \{l \in U \,:\, a_{f_{p_{(i-1)}}^t} < g_l < a_{f_{p_i}^t}\} \neq \emptyset$$

As a consequence, we have:

$$\{a_{f_{(p_{(i-1)}+1)}^t}, \dots, a_{(f_{p_i}^t - 1)}\} \subseteq D_t$$

and $b_{f_{p_i}}^t = a_{(f_{p_i}^t - 1)}$ and there is not a possible reduction in the dimension.

On the contrary, if we suppose that $p_i > p_{(i-1)} + 1$, then $f_{p_i}^t > f_{(p_{(i-1)}+1)}^t$ and there is a integer $z \geq 1$ such that $p_i = p_{(i-1)} + 1 + z$. For all $j = 1, \dots, z$, the value $b_{f_{p_{(i-1)}+j}^t}$ does not exist and the set $U_{p_{(i-1)}+j} = \emptyset$. As in the previous case (case p_1), we have:

$$a_{f_{(p_{(i-1)}+j)}^t} = a_{(f_{p_{(i-1)}}^t + j)}, \quad j = 1, \dots, z.$$

Thus, if $p_i > p_{(i-1)} + 1$, we have:

$$\{a_{(f_{p_{(i-1)}}^t + 1)}, \dots, a_{(f_{p_{(i-1)}}^t + p_i - p_{(i-1)} - 1)}\} \subseteq A_t.$$

Then, if we define the following sets:

$$A_{p_i} = \{a_{(f_{p_{(i-1)}}^t + 1)}, \dots, a_{(f_{p_{(i-1)}}^t + p_i - p_{(i-1)} - 1)}\}$$

$$Z_{p_i} = \{a_i \in A_{p_i} \,:\, q_i^t = 0\}$$

and

$$F_{p_i} = \{a_i \in A_{p_i} \, : \, 0 < q_i^t < r_i\}$$

we can proof in a similar way to the previous case (case p_1) that if $Z_{p_i} = A_{p_i}$ or $Z_{p_i} \neq A_{p_i}$ but $F_{p_i} \neq A_{p_i}$ there is a possible reduction in the dimension of the auxiliary vector $\mathbf{t_{OP}}(t)$. If $F_{p_i} = A_{p_i}$ there is no possible dimension reduction.

Finally, if we suppose that $p_{l_t} = M_t$ then the value $b_{f^t_{M_t}}$ exists and analogously to previous cases, there is no reduction between the points $b_{f^t_{M_t}}$ and $a_{f^t_{M_t}}$.

On the other hand, if we suppose that $p_{l_t} < M_t$ then for $h = p_{l_t} + 1, p_{l_t} + 2, \ldots, M_t$ the corresponding value $b_{f^t_h}$ does not exist and therefore the sets $U_h = \emptyset$. As a consequence, we have:

$$a_{f^t_{(p_{l_t}+1)}} = a_{(f^t_{p_{l_t}}+1)}, \ldots, a_{f^t_{M_t}} = a_{(f^t_{p_{l_t}}+M_t-p_{l_t})}$$

If we denote by

$$A_{M_t} = \{a_{(f^t_{p_{l_t}}+1)}, \ldots, a_{(f^t_{p_{l_t}}+M_t-p_{l_t})}\} \subseteq A_t$$
$$Z_{M_t} = \{a_i \in A_{M_t} \, : \, q_i^t = 0\}$$
$$F_{M_t} = \{a_i \in A_{M_t} \, : \, 0 < q_i^t < r_i\}$$

then, we can reduce the dimension of the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ if $Z_{M_t} = A_{M_t}$ or if $Z_{M_t} \neq A_{M_t}$ but $F_{M_t} \neq A_{M_t}$. If $F_{M_t} = A_{M_t}$ there is no possible dimension reduction.

References

1. Deville JC, Särndal CE. Calibration estimators in survey sampling. *J Amer Statist Assoc.* 1992;87(418):376–382.
2. Rueda MM, Martínez S, Martínez H, Arcos A. Estimation of the distribution function with calibration methods. *J Stat Plan Infer.* 2007;137(2):435–448.
3. Martínez S, Rueda MM, Martínez H, Arcos A. Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function. *J Comput Appl Math.* 2017;318:444–459.
4. Sedransk N, Sedransk J. Distinguishing among distributions using data from complex sample designs. *J Am Stat Assoc.* 1979;74(368):754–760.
5. Dickens R, Manning A. Has the national minimum wage reduced UK wage inequality?. *J Roy Stat Soc A Sta.* 2004;167(4):613–626.
6. Nickell S. Poverty and worklessness in Britain. *The Economic Journal.* 2004;114(494):C1–C25.
7. Machin S, Manning A, Rahman L. Where the minimum wage bites hard: Introduction of minimum wages to a low wage sector. *Journal of the European Economic Association.* 2003;1(1):154–180.
8. Estevao VM, Särndal CE. Survey estimates by calibration on complex auxiliary information. *Int Stat Rev.* 2006;74(2):127–147.
9. Singh S. Generalized calibration approach for estimating variance in survey sampling. *Ann I Stat Math.* 2001;53(2):404–417.
10. Devaud D, Tillé Y. Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. *Test.* 2019;28(4):1033–1065.
11. Rueda MM. Comments on: Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. *Test.* 2019;28(4):1077–1081.
12. Kovacevic M. Calibration estimation of cumulative distribution and quantile functions from survey data. In: :139–144; 1997.
13. Harms T, Duchesne P. On calibration estimation for quantiles. *Surv Methodol.* 2006;32(1):37–52.
14. Wu C. Optimal calibration estimators in survey sampling. *Biometrika.* 2003;90(4):937–951.

15. Chen J, Wu C. Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica.* 2002;12(4):1223–1239.

16. Breidt FJ, Opsomer JD, Johnson AA, Ranalli MG. Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology.* 2007;33(1):35.

17. Mayor-Gallego JA, Moreno-Rebollo JL, Jiménez-Gamero MD. Estimation of the finite population distribution function using a global penalized calibration method. *Asta-Adv Stat Anal.* 2019;103(1):1–35.

18. Martínez S, Rueda MM, Arcos A, Martínez H. Optimum calibration points estimating distribution functions. *J Comput Appl Math.* 2010;233(9):2265–2277.

19. Martínez S, Rueda MM, Arcos A, Martínez H, Sánchez-Borrego I. Post-stratified calibration method for estimating quantiles. *Comput Stat Data An.* 2011;55(1):838–851.

20. Martínez S, Rueda MM, Martínez H, Arcos A. Determining P optimum calibration points to construct calibration estimators of the distribution function. *J Comput Appl Math.* 2015;275:281–293.

21. Martínez S, Rueda MM, Illescas MD. The optimization problem of quantile and poverty measures estimation based on calibration. *J Comput Appl Math.* 2022;405:113054.

22. Nascimento Silva PLD, Skinner CJ. Variable selection for regression estimation in finite populations. *Survey Methodology.* 1997;23(1):23–32.

23. Chauvet G, Goga C. Asymptotic efficiency of the calibration estimator in a highdimensional data setting. *Journal of Statistical Planning and Inference.* 2022;217:177–187.

24. Chambers RL, Clark RG. Adaptive calibration for prediction of finite population totals. *Survey Methodology.* 2008;34(2):163–172.

25. McConville KS, Breidt FJ, Lee TMC, Moisen GG. Model assisted survey regression estimation with the LASSO. *J. Surv. Stat. Methodol.* 2017;5(2):131–158.

26. Guggemos F, Tillé Y. Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference.* 2010;140(11):3199–3212.

27. Cardot H, Goga C, Shehzad M. Calibration and Partial Calibration on Principal Components when the Number of Auxiliary Variables is Large. *Statistica Sinica.* 2017;27(1):243–260.

28. Rota BJ. Variance estimation in two-step calibration for nonresponse adjustment. *South African Statistical Journal.* 2017;51(2):361–374.

29. Singh HP, Singh S, Kozak MA. A family of estimators of finite-population distribution function using auxiliary information. *Acta Appl Math.* 2008;104(2):115–130.

30. Arcos A, Martínez S, Rueda MM, Martínez H. Distribution function estimates from dual frame context. *J Comput Appl Math.* 2017;318:242–252.

31. Rao JNK, Kovar JG, Mantel HJ. On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika.* 1986;77(2):365–375.

32. Chambers RL, Dunstan R. Estimating distribution functions from survey data. *Biometrika.* 1986;73(3):597–604.

AUTHOR BIOGRAPHY

Sergio Martínez. Sergio Martínez is Ph.D. in Maths from the University of Almería and is currently full time professor at Math department of University of Almería. His current research focuses on parameter estimation in finite population, such as the mean, distribution function and quantiles by calibration techniques. Recently, he has written several articles related to the analysis of large volumes of data through mixed structures of neural networks and articles related to econometric methods to pricing in hospitality firms. His works have been published in journals such as Sociological Methods & Research, International Journal of Hospitality Management, Applied Soft Computing, Journal of Computational and Applied Mathematics.

María del Mar Rueda. María del Mar is Professor of Statistics and I.O. at the University of Granada. His specialty is the inference in finite populations, and the treatment of errors of lack of coverage, lack of response, voluntariness, etc, through modern prediction techniques.

María D. Illescas. María Illescas is a Ph. D student at Department of Economics and Busines of University of Almeria . She received her Master degree of Finance and Accounting and her degree in Business Administration from University of Almería. Her current research focuses on econometric methods to pricing in hospitality firms. Her works have been published in journals such as International Journal of Hospitality Management, International Journal of Computer Mathematics..