Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function

S. Martínez[a], M. Rueda[b,], H. Martínez[a], A. Arcos[b]

[a] *Department of Mathematics. University of Almería, Spain*
[b] *Department of Statistics and Operational Research. University of Granada, Spain*

Abstract

The calibration technique ([?]) to estimate the finite distribution function have been studied in several papers. Calibration seeks for new weights close enough to sampling weights according to some distance function and that, at the same time, match benchmark constraints on available auxiliary information. The non smooth character of the finite population distribution function causes certain complexities that are resolved by different authors in different ways. One of these is to have consistency at a number of arbitrarily chosen points. This paper deals with the problem of the optimal selection of the number of points and with the optimal selections of these points, when auxiliary information is used by means of calibration.

Keywords: Auxiliary information, calibration technique, distribution function estimates, survey sampling

2010 MSC: 62D05

1. Introduction

Calibration is the principal theme in many recent articles on estimation in survey sampling ([?], [?], [?], [?], [?], [?],...) Calibration has established itself as an important methodological instrument in large scale production of statistics. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources.

Preprint submitted to Journal of Computational and Applied Mathematics October 7, 2015

The calibration approach adapts itself to the estimation of more complex parameters than a population total. Before calibration became popular, several papers considered the estimation of distribution functions, with or without the use of auxiliary information ([?], [?],[?], [?]. As [?] illustrates, there is more than one way to implement the calibration approach in the estimation of the distribution function. Some applications to missing data problems can be seen in [?] and [?].

The non smooth character of the finite population distribution function causes certain complexities; these are resolved by different authors in different ways. Furthermore, in some cases it not possible to find an exact solution of the calibration problem as stated.

The computationally simpler method of [?] is an application of model calibration, in that they calibrate with respect to a population total of predicted y-values. Complete auxiliary information is required. Using the known finite population distribution functions of auxiliary variables, compute first the linear predictions. The calibrated weights are obtained by minimizing the chi-square distance subject to calibration equations stated in terms of the predictions, so as to have consistency at J arbitrarily chosen points. It is suggested that a fairly small number of arbitrarily selected points may suffice.

The idea of to create many benchmarks based on an auxiliary variable was proposed in [?] (Exercise 3.35). This estimator of median can be shown as a special case of how to use 99 known percentiles of an auxiliary variable.

The question of the optimal values in order to obtain the best estimation under simple random sampling without replacement for an arbitrary number of calibration points can be seen in [?]. This paper shows the optimal size of the chosen points (Section 3) and the optimal vector (Section 4). In Section 5 we define the optimum estimator with estimated optimal vector and in Section 6 we include some numerical comparisons.

2. Calibration estimation of the distribution function

Let $U = \{1, 2, \ldots, N\}$ be a finite survey population from which a realized sample $s = \{1, 2, \ldots, n\}$ is drawn with a measurable design d with first and second order inclusion probabilities π_k and π_{kl}. We note by y_k the main variable and by x_k a vector of auxiliary variables at unit k. The values x_k are known for the entire population but y_k is known only if the kth unit is selected on the sample, s. To estimate the distribution function of the study variable y

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

where

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

we consider the calibration approach, which consist in the construction of an estimator $\sum\limits_{k \in s} \omega_k \Delta(t - y_k)$ where the calibration weights ω_k are chosen to minimize their average distance from the basic design weights $d_k = 1/\pi_k$ that are used in the Horvitz-Thompson estimator

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

subject to conditions that use the auxiliary information provided by the auxiliary vector x.

The distance measure is most commonly chosen as

$$\Phi_s = \frac{1}{2} \sum_{k \in s} \frac{(\omega_k - d_k)^2}{d_k q_k} \tag{3}$$

with q_k are known positive constants unrelated to d_k. Following [?] and [?], in the definition of calibration conditions, we consider a pseudo-variable $g_k = \widehat{\beta}' x_k$ for $k = 1, 2, \ldots, N$ where:

$$\widehat{\beta}' = \left(\sum_{k \in s} d_k q_k x_k x_k' \right)^{-1} \cdot \sum_{k \in s} d_k q_k x_k y_k.$$

With the pseudo-variable g, we consider the minimization of (3) subject to the following conditions:

$$\frac{1}{N}\sum_{k\in s}\omega_k\Delta(\mathbf{t_g}-g_k)=F_g(\mathbf{t_g}) \tag{4}$$

with $\mathbf{t_g}=(t_1,\ldots,t_P)'$ is a vector chosen arbitrarily, assuming that

$$t_1<t_2<\ldots<t_P.$$

Note that this estimator can be seen as a particular case under a more general model $g_k=a+bx_k$ with the condition $\sum\omega_k=\sum d_k$. AQUI HABRA QUE AÑADIR UNA REFERENCIA PARA LA CONDICION ESTA The resulting estimator ([**?**]) is given by

$$\widehat{F}_{yc}(t)=\widehat{F}_{YHT}(t)+\Big(F_g(\mathbf{t_g})-\widehat{F}_{GHT}(\mathbf{t_g})\Big)'\cdot\widehat{D} \tag{5}$$

where $\widehat{F}_{GHT}$ is the Horvitz-Thompson estimator of F_g and

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

assuming that the inverse of symmetric matrix T

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

exists.

The asymptotic variance of $\widehat{F}_{yc}(t)$ is studied in [**?**] and is given by:

$$AV(\widehat{F}_{yc}(t))=\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_kE_k)(d_lE_l) \tag{6}$$

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

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

The precision of $\widehat{F}_{yc}(t)$ changes with the selection of $\mathbf{t_g}$. In [**?**], the authors studied, for a fixed P, the problem of selection the optimal vector $\mathbf{t_g}$, under simple random sampling and $q_k=1$ for all $k\in U$, that gives the best estimation of

$F_y(t)$, that is, the problem of determining an auxiliary vector $\mathbf{t_g} = (t_1, \dots, t_P)'$, with $t_1 < t_2 < \dots < t_P$ that minimizes the asymptotic variance of the estimator $\widehat{F}_{yc}(t)$ given a point t for which we want to estimate $F_y(t)$.

Following [**?**], the problem of determining an auxiliary vector $\mathbf{t_g}$ that minimizes the asymptotic variance (6), under simple random sampling, is equivalent to minimizing the following function:

$$G(t_1, t_2, \dots, t_P) = 2NF_y(t) \cdot k_p - \sum_{j=1}^{P} \frac{(k_j - k_{j-1})^2}{(F_g(t_j) - F_g(t_{j-1}))} - k_P^2 \qquad (7)$$

where

$$k_i = \sum_{k \in U} \Delta(t - y_k)\Delta(t_i - g_k) \quad i = 1, 2 \dots, P.$$

The global minimum of the function G ([**?**]) is a vector $\mathbf{t_g} = (t_1, t_2, \dots, t_P)$, with $t_1 < t_2 < \dots < t_P$ and $t_i \in A_t$ or $t_i \in B_t$ for $i = 1, 2, \dots, P$, where the set A_t and B_t are given by

$$A_t = \{g_k : y_k \leq t\} = \{a_1, a_2, \dots, a_M\} \qquad (8)$$

with $a_1 < a_2 < \dots < a_M$ and

$$B_t = \{b_1, b_2, \dots, b_M\} \qquad (9)$$

with

$$b_1 = \max_{l \in U_1}\{g_l\} \text{ where } U_1 = \{l \in U : g_l < a_1\}$$

$$b_h = \max_{l \in U_h}\{g_l\} \text{ where } U_h = \{l \in U : a_{h-1} \leq g_l < a_h\}, \quad h = 2, 3, \dots, M$$

where $b_1 < b_2 < \dots < b_M$. Since the sets A_t and B_t are finite, finding the global minimum is computationally simple. For some h in $1, 2, \dots, M$ the corresponding point b_h may not exist, but in this case, the minimization problem is simpler than the current case ([**?**]).

In the next section, we consider the optimal dimension P of the auxiliary vector $\mathbf{t_g}$, that is, the optimal number of auxiliary points t_i used in the calibration process in order to obtain the best estimation of $F_y(t)$.

3. Optimal dimension P of the auxiliary vector $\mathbf{t_g}$

In this section, given a point t for which we want to estimate $F_y(t)$, we study the problem of the optimal dimension P of the auxiliary vector $\mathbf{t_g}$ used in the calibration process of the estimator $\widehat{F}_{yc}(t)$. The following theorem establishes the optimal dimension of the vector $\mathbf{t_g}$.

Theorem 1. *Suppose that we wish to estimate F_y at point t with the calibration estimator $\widehat{F}_{yc}(t)$, then the optimal dimension of the auxiliary vector $\mathbf{t_g}$ is $P = 2M$, where M is the number of points of the finite set A_t given by (8), provided that b_1 exits and for all $i = 1, \ldots, M-1$, $b_{i+1} \neq a_i$.*

Proof.

The proof of the theorem is developed in two steps. In the first step, we show that the calibration process with $P-1$ auxiliary points is a particular case of P auxiliary calibration points. In the second step, we show that the minimization of the asymptotic variance of $\widehat{F}_{yc}(t)$ with P auxiliary points, where $P > 2M$ is equivalent to the minimization of the asymptotic variance of $\widehat{F}_{yc}(t)$ with $2M$ auxiliary points. In order to develop the two steps, the function $G(t_1, t_2, \ldots, t_P)$ ([?]) is a piecewise function given by: $G(t_1, \ldots, t_P) =$

$$= \begin{cases} 0 & \text{if } t_P \leq b_1 \\ 2NF_y(t)K_{h_P} - \displaystyle\sum_{j=1}^{P} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(t_j) - F_g(t_{j-1}))} - K_{h_P}^2 & \text{if } a_{h_j} \leq t_j \leq b_{h_j+1} \\ & j = 1, 2, \ldots P \\ & h_j = 0, 1, 2, \ldots M \\ (NF_y(t))^2 - \dfrac{(NF_y(t))^2}{F_g(t_1)} & \text{if } a_M \leq t_1 \end{cases} \tag{10}$$

where $a_0 = g_{min}$; $b_{M+1} = g_{max}$; $h_0 = 0$; t_0 any value with $t_0 < a_0$ and

$$g_{min} = \min_{k \in U}\{g_k\}, \quad g_{max} = \max_{k \in U}\{g_k\},$$

and

$$K_{h_j} = \begin{cases} 0 & \text{if } t_j < a_1 \\ & \\ \sum_{k \in U} \Delta(t - y_k)\Delta(a_{h_j} - g_k) & \text{if } a_{h_j} \leq t_j \leq b_{h_j+1} \\ & h_j = 1, 2, \dots M-1, \\ NF_y(t) = K_M & \text{if } a_M \leq t_j. \end{cases} \tag{11}$$

First step. The function G with an auxiliary vector $\mathbf{t_g}$ of dimension $P-1$ is a particular case of function G with an auxiliary vector $\mathbf{t_g}$ of dimension$P$: If we denote by G_{P-1} the function G when the auxiliary vector $\mathbf{t_g} = (t_1, t_2, \dots, t_{P-1})$ have dimension $P-1$ and G_P the function G when the auxiliary vector $\mathbf{t_g}(t_1, t_2, \dots, t_P)$ have dimension P, then it's clear that if $t_1 \geq a_M$, we have:

$$G_{P-1}(t_1, t_2, \dots, t_{P-1}) = G_P(t_1, t_2, \dots, t_P) = (NF_y(t))^2 - \frac{(NF_y(t))^2}{F_g(t_1)}.$$

For $t_{P-1} < b_1$, it is sufficient to consider $t_P < b_1$ and we have

$$G_{P-1}(t_1, t_2, \dots, t_{P-1}) = G_P(t_1, t_2, \dots, t_P) = 0.$$

Finally, for $a_{h_j} \leq t_j \leq b_{h_j+1}$, wih $j = 1, 2, \dots P-1$ and $h_j = 0, 1, 2, \dots M-1$, if we consider $t_P \in [a_{h_{P-1}}, b_{h_{P-1}+1}]$, that is $h_P = h_{P-1}$, it is easy to see that $K_{h_P} = K_{h_{P-1}}$, and consequently

$$G_P(t_1, t_2, \dots, t_P) = 2NF_y(t)K_{h_P} - \sum_{j=1}^{P} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(t_j) - F_g(t_{j-1}))} - K_{h_P}^2 =$$

$$2NF_y(t)K_{h_{P-1}} - \sum_{j=1}^{P-1} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(t_j) - F_g(t_{j-1}))} - K_{h_{P-1}}^2 = G_{P-1}(t_1, t_2, \dots, t_{P-1}).$$

Thus, the function $G_{P-1}(t_1 t_2, \dots, t_{P-1})$ is a particular case of $G_P(t_1, t_2, \dots, t_P)$ and we have

$$\min G_{P-1}(t_1 t_2, \dots, t_{P-1}) \geq \min G_P(t_1, t_2, \dots, t_P).$$

Therefore

$$\min G_1(t_1) \geq \min G_2(t_1, t_2) \geq \ldots \geq \min G_{2M}(t_1, t_2, \ldots, t_{2M}). \tag{12}$$

Second step. Minimization of $G_P(t_1, t_2, \ldots, t_P)$ with $P > 2M$ is equivalent to the minimization of $G_{2M}(t_1, t_2, \ldots, t_{2M})$:

First, to demonstrate the second step, we consider the case where $P = 2M + 1$, that is, the function $G(\mathbf{t_g}) = G_{2M+1}(t_1, t_2, \ldots, t_{2M+1})$ has dimension $2M + 1$. Because the number of different points in the set A_t is M, we have $M + 1$ sets $[a_i, b_{i+1}]$ with $i = 0, 1, \ldots, M$ and it's clear that the auxiliary vector $\mathbf{t_g} = (t_1, t_2, \ldots, t_{2M+1})$ satisfies one of three conditions:

- t_1, t_2 or more points of $\mathbf{t_g}$ are in $[a_0, b_1]$.
- t_{2M+1}, t_{2M} or more points of $\mathbf{t_g}$ are in $[a_M, b_{M+1}]$.
- For some $l = 1, 2, \ldots M-1$, exits $i \in \{1, 2, \ldots 2M-1\}$ such that t_i, t_{i+1}, t_{i+2} or more points of $\mathbf{t_g}$ are in $[a_l, b_{l+1}]$.

Case 1) t_1, t_2 or more points of $\mathbf{t_g}$ are in $[a_0, b_1]$:

If $t_{2M+1} \in [a_0, b_1]$, then

$$G_{2M+1}(t_1, t_2, \ldots, t_{2M+1}) = 0 = G_{2M}(t_2, \ldots, t_{2M+1}) = G_{2M}(T_1, \ldots, T_{2M})$$

where $T_i = t_{i+1}$ and $T_i \in [a_0, b_1]$ for $i = 1, 2, \ldots 2M$, and therefore the minimization of $G_{2M+1}(t_1, t_2, \ldots, t_{2M+1})$ is equivalent to the minimization of $G_{2M}(t_1, t_2, \ldots, t_{2M})$.

If $t_{2M+1} \notin [a_0, b_1]$, the values $K_{h_1} = K_{h_2} = 0$ and the function $G_{2M+1}(t_1, t_2, \ldots, t_{2M+1})$ is given by

$$G_{2M+1}(t_1, t_2, \ldots, t_{2M+1}) = 2NF_y(t)K_{h_{2M+1}} - \sum_{j=3}^{2M+1} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(t_j) - F_g(t_{j-1}))} - K^2_{h_{2M+1}}$$

where for $j = 3, \dots 2M$, $a_{h_j} \leq t_j \leq b_{h_j+1}$ with $h_j = 0, 1, 2, \dots M$, and for $j = 2M + 1$ we have $a_{h_{2M+1}} \leq t_{2M+1} \leq b_{h_{2M+1}+1}$ with $h_{2M+1} = 1, 2, \dots M$.

Thus, if we denote by $T_i = t_{i+1}$ for $i = 2, \dots, 2M$, it is easy to see that

$$G_{2M}(T_1, \dots, T_{2M}) = G_{2M+1}(t_1, \dots, t_{2M+1})$$

and the minimization of $G_{2M+1}(t_1, t_2, \dots, t_{2M+1})$ is equivalent to the minimization of $G_{2M}(t_1, t_2, \dots, t_{2M})$.

Case 2) t_{2M+1}, t_{2M} or more points of $\mathbf{t_g}$ are in $[a_M, b_{M+1}]$:

If $t_1 \in [a_M, b_{M+1}]$ we have:

$$G_{2M+1}(t_1, t_2, \dots, t_{2M+1}) = (NF_y(t))^2 - \frac{(NF_y(t))^2}{F_g(t_1)} = G_{2M}(T_1, \dots, T_{2M})$$

where $T_i = t_i$ and $T_i \in [a_M, b_{M+1}]$ for $i = 1, 2, \dots 2M$, and consequently, the minimization of $G_{2M+1}(t_1, t_2, \dots, t_{2M+1})$ is equivalent to the minimization of $G_{2M}(t_1, t_2, \dots, t_{2M})$.

If $t_1 \notin [a_M, b_{M+1}]$, the values $K_{h_{2M}} = K_{h_{2M+1}} = NF_y(t)$ and we have

$$G_{2M+1}(t_1, t_2, \dots, t_{2M+1}) = \big(NF_y(t)\big)^2 - \sum_{j=1}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(t_j) - F_g(t_{j-1}))}$$

where for $j = 1$ we have $a_{h_1} \leq t_1 \leq b_{h_1+1}$ with $h_1 = 0, 1, \dots M - 1$, and for $j = 2, \dots 2M - 1$; $a_{h_j} \leq t_j \leq b_{h_j+1}$ with $h_j = 0, 1, 2, \dots M$. Then, if we define $T_i = t_i$, we have:

$$G_{2M}(T_1, \dots, T_{2M}) = G_{2M+1}(t_1, \dots, t_{2M+1})$$

and the minimization of $G_{2M+1}(t_1, t_2, \dots, t_{2M+1})$ is equivalent to the minimization of $G_{2M}(t_1, t_2, \dots, t_{2M})$.

Case 3) For some $l = 1, 2, \dots M - 1$, exits $i \in \{1, 2, \dots 2M - 1\}$ such that t_i, t_{i+1}, t_{i+2} or more points of $\mathbf{t_g}$ are in $[a_l, b_{l+1}]$.

In this case, $h_i = h_{i+1} = h_{i+2} = l$ and the values $K_{h_i} = K_{h_{i+1}} = K_{h_{i+2}} = k_l$. Therefore:

$$\begin{aligned} G_{2M+1}(t_1, t_2, \dots, t_{2M+1}) &= 2NF_y(t)K_{h_{2M+1}} - \sum_{j=1}^{i} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(t_j) - F_g(t_{j-1}))} \\ &- \sum_{j=i+3}^{2M+1} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(t_j) - F_g(t_{j-1}))} - K^2_{h_{2M+1}} \end{aligned}$$

and it's clear that the function G_{2M+1} does not depend on t_{i+1}.

Thus, if we consider $T_j = t_j$ for $j = 1, \dots, i$ and $T_j = t_{j+1}$ for $j = i + 1, \dots, 2M$, the minimization of $G_{2M+1}(t_1, t_2, \dots, t_{2M+1})$ is equivalent to the minimization of $G_{2M}(T_1, T_2, \dots, T_{2M})$.

Thus, for all cases, the minimization of $G_{2M+1}(t_1, t_2, \dots, t_{2M+1})$ is equivalent to the minimization of $G_{2M}(T_1, T_2, \dots, T_{2M})$.

If we consider a dimension $P > 2M + 1$ for the auxiliary vector $\mathbf{t_g}$, in a similar way we can establish that the minimization of $G_P(t_1, t_2, \dots, t_P)$ is equivalent to the minimization of $G_{P-1}(t_1, t_2, \dots, t_{P-1})$ and applying L times, with $L = P - 2M$, this property recursively, the minimization of $G_P(t_1, t_2, \dots, t_P)$ is equivalent to the minimization of $G_{P-L}(t_1, t_2, \dots, t_{P-L})$, and it's clear that the minimization of $G_P(t_1, t_2, \dots, t_P)$ is equivalent to the minimization of $G_{2M}(t_1, t_2, \dots, t_{2M})$.

Now, we consider the case where for some $i_1, i_2, \dots i_R \in \{0, 1, \dots, M - 1\}$; $a_{i_1} = b_{i_1} + 1$ with $R \leq M$ and $i_h \neq i_j$ if $h \neq j$. In this case the next theorem establish the optimal dimension of the auxiliary vector $\mathbf{t_g}$.

Theorem 2. *Suppose that we wish to estimate F_y at point t with the calibration estimator $\widehat{F}_{yc}(t)$ and suppose that for some $i_1, i_2, \dots i_R \in \{0, 1, \dots, M - 1\}$; $a_{i_1} = b_{i_1+1}$ with $R \leq M$ and $i_h \neq i_j$ if $h \neq j$, then the optimal dimension of the auxiliary vector $\mathbf{t_g}$ is $P = 2M - R$, where M is the number of points of the finite set A_t given by (8).*

Proof.

It is clear that the function G with an auxiliary vector $\mathbf{t_g}$ of dimension $P - 1$ is a particular case of function G with an auxiliary vector $\mathbf{t_g}$ of dimension P

(previous theorem) and in a similar way, we have

$$\min G_1(t_1) \geq \min G_2(t_1, t_2) \geq \ldots \geq \min G_{2M}(t_1, t_2, \ldots, t_{2M-R}). \tag{13}$$

Now, we consider the case where the dimension of function G is $P = 2M - R + 1$, and $i_h \neq 0$ for all $h = 1, \ldots, R$. Then, we have $M+1$ sets $[a_i, b_{i+1}]$ with $i = 0, 1, \ldots, M$ but for $h = 1, \ldots, R$ the set $[a_{i_h}, b_{i_h+1}] = \{a_{i_h}\}$. Thus, we have R sets $\{a_{i_h}\}$ and $M + 1 - R$ intervals $[a_i, b_{i+1}]$. If we consider an auxiliary vector $\mathbf{t_g} = (t_1, \ldots, t_{2M-R+1})$, the points t_i with $i = 1, \ldots, 2M - R$ satisfies one of four conditions:

- t_1, t_2 or more points of $\mathbf{t_g}$ are in $[a_0, b_1]$.
- t_{2M-R+1}, t_{2M-R} or more points of $\mathbf{t_g}$ are in $[a_M, b_{M+1}]$.
- For some $l = 1, 2, \ldots M - 1$, with $l \neq i_h$ for $h = 1, \ldots, R$, exits $i \in \{1, 2, \ldots 2M-1\}$ such that t_i, t_{i+1}, t_{i+2} or more points of $\mathbf{t_g}$ are in $[a_l, b_{l+1}]$
- For some $l = i_1, \ldots i_R$ exits $i \in \{1, 2, \ldots 2M - 1\}$ such that t_i, t_{i+1} or more points of $\mathbf{t_g}$ are in $\{a_l\}$, that is $t_i = t_{i+1} = a_l$.

Similarly to previous theorem, the minimization of $G_{2M-R+1}(t_1, \ldots, t_{2M-R+1})$ is equivalent to the minimization of $G_{2M-R}(t_1, \ldots, t_{2M-R})$ in the first three cases. For the last case, it is easy to see that:

$$G_{2M-R+1}(t_1, \ldots t_i, t_{i+1}, \ldots, t_{2M-R+1}) = G_{2M-R+1}(t_1, \ldots a_l, a_l, \ldots, t_{2M-R+1}) =$$
$$G_{2M-R}(t_1, \ldots, t_{i-1}, a_l, t_{i+2}, \ldots, t_{2M-R+1}).$$

Therefore, if we consider

$$T_j = t_j \text{ for } j = 1, 2 \ldots, i-1 \text{ and } T_j = t_{j+1} \text{ for } j = i+1, \ldots 2M - R$$

it is clear that the minimization of $G_{2M-R+1}(t_1, \ldots t_i, t_{i+1}, \ldots, t_{2M-R+1})$ is equivalent to the minimization of $G_{2M-R}(T_1, \ldots T_{i-1}, a_l, T_{i+1} \ldots, T_{2M-R})$. Thus, if $i_h \neq 0$ for all $h = 1, \ldots, R$ the optimal dimension of $\mathbf{t_g}$ is $2M - R$.

If for some $h \in \{i_1, \ldots, i_R\}$, $i_h = 0$; then $a_0 = a_1$ and we have M intervals $[a_i, b_{i+1}]$ where R of them are of the form $\{a_{i_h}\}$. In this case, the auxiliary vector $\mathbf{t_g} = (t_1, \dots, t_{2M-R+1})$ fulfills one of the last three above conditions and consequently the optimal dimension of $\mathbf{t_g}$ is $2M - R$.

4. Optimal auxiliary vector $\mathbf{t}_{opt}$

In this section, we will obtain the optimal auxiliary vector $\mathbf{t}_{opt}$ of the optimal dimension, given a point t for which we want to estimate $F_y(t)$, that is, we obtain a vector $\mathbf{t}_{opt}$ of the optimal dimension obtained in the previous section, such that the value of $AV(\widehat{F}_{yc}(t))$ calibrated with $\mathbf{t}_{opt}$ the value is less than the value of $AV(\widehat{F}_{yc}(t))$ calibrated with any vector $\mathbf{t_g}$.

Theorem 3. *Suppose that we wish to estimate F_y at point t with the calibration estimator $\widehat{F}_{yc}(t)$, and suppose that b_1 exits and for all $i = 1, \dots, M-1$; $b_{i+1} \neq a_i$, then the optimal auxiliary vector $\mathbf{t}_{opt} = (t_{O1}, \dots, t_{O2M})$ is a vector of dimension $2M$ given by:*

$$\mathbf{t}_{opt} = (t_{O1}, \dots, t_{O2M}) = (b_1, a_1, b_2, a_2, \dots, b_M, a_M) \tag{14}$$

If for some $i_1, i_2, \dots i_R \in \{0, 1, \dots, M-1\}$; $a_{i_1} = b_{i_1+1}$ with $R \leq M$ and $i_h \neq i_j$ if $h \neq j$ the optimal auxiliary vector $\mathbf{t}_{opt} = (t_{O1}, \dots, t_{O(2M-R)})$ is a vector of dimension $2M - R$ given by:

$$\mathbf{t}_{opt} = (b_1, a_1, b_2, a_2, \dots, b_{i_1}, a_{i_1}, a_{i_1+1}, b_{i_1+2}, \dots, b_{i_h}, a_{i_h}, a_{i_h+1}, b_{i_h+2}, \dots b_M, a_M) \tag{15}$$

Proof.

First, we consider the case where b_1 exits and for all $i = 1, \dots, M-1$; $b_{i+1} \neq a_i$. Because the function $G_{2M}(\mathbf{t_g})$ is a piecewise function, to demonstrate that the auxiliary vector $\mathbf{t}_{opt}$ given by (14) is the vector where the function G_{2M} attains the global minimum, we have to obtain the minimum of the function G_{2M} on each piece and compare the value of the function G_{2M} at the vector $\mathbf{t}_{opt}$ with the minimum obtained in each piece. For it, if we define $K_0 = 0$, the value of function $G_{2M}(t_1,\dots,t_{2M})$ at the vector $\mathbf{t}_{opt}$ given by (14) is:

$$G_{2M}(\mathbf{t}_{opt}) = \Big(NF_y(t)\Big)^2 - \sum_{j=1}^{M} \frac{(K_j - K_{j-1})^2}{(F_g(a_j) - F_g(b_j))}. \tag{16}$$

It is clear that for any $2M$ dimensional auxiliary vector $\mathbf{t_g} = (t_1, \dots t_{2M})$ with

$$t_{2i-1} \in [a_{i-1}, b_i];\;\; t_{2i} \in [a_i, b_{i+1}] \text{ for } i = 1, 2, \dots, M \tag{17}$$

we have:

$$G_{2M}(\mathbf{t_g}) \geq G_{2M}(\mathbf{t}_{opt}).$$

Therefore, the auxiliary vector $\mathbf{t}_{opt}$ is the optimal choice for those vectors which verify (17).

Now, let's analyze the minimum of the function G on each piece. For it, if we consider an auxiliary vector $\mathbf{t_g}$ with $t_{2M} < b_1$, the function G_{2M} is null, then the function G_{2M} has a local minimum at $t_1 = t_2 = t_{2M} = b_1$ and the local minimum value is 0.

Next, if $a_M \leq t_1$, the function G_{2M} attains a local minimum at $t_1 = a_M$ and $t_2, \dots, t_{2M}$ are arbitrary chosen points of the interval $[a_M, b_{M+1})$. Thus, the function G_{2M} attains the local minimum at $t_1 = \dots = t_{2M} = a_M$ and the local minimum value is

$$G_{2M}(t_1, \dots, t_{2M}) = (NF_y(t))^2 - \frac{(NF_y(t))^2}{F_g(a_M)} \leq 0. \tag{18}$$

Therefore, the value 0 in the first case cannot be the global minimum of G_{2M}.

Now, if we consider that

$$a_{h_j} \leq t_j \leq b_{h_j+1};\;\; j = 1, 2, \dots 2M;\;\; h_j = 0, 1, 2, \dots M$$

the function G_{2M} is given by

$$G_{2M}(t_1, \dots, t_{2M}) = 2NF_y(t)K_{h_{2M}} - \sum_{j=1}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(t_j) - F_g(t_{j-1}))} - K^2_{h_{2M}}.$$

The local minimum of the function G_{2M} is attained at $t_j = a_{h_j}$ or $t_j = b_{h_j+1}$ ([**?**]), consequently the local minimum is given by:

$$G_{2M}(A_1, \dots, A_{2M}) = 2NF_y(t)K_{h_{2M}} - \sum_{j=1}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} - K^2_{h_{2M}} \tag{19}$$

with $A_j = a_{h_j}$ or $A_j = b_{h_j+1}$ and where we take A_0 any value such that $A_0 < a_0$.

In order to compare the local minimum of function G_{2M} on each piece with the value of function G_{2M} at $\mathbf{t}_{opt}$, we establish the following inequalities

$$\frac{K_j^2}{F_g(a_j) - F_g(b_1)} \leq \frac{K_{j-1}^2}{F_g(a_{j-1}) - F_g(b_1)} + \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} \text{ for } j = 1, 2, \ldots, M \tag{20}$$

$$\frac{(K_{j+k} - K_j)^2}{F_g(a_{j+k}) - F_g(b_{j+1})} \leq \frac{(K_{j+1} - K_j)^2}{F_g(a_{j+1}) - F_g(b_{j+1})} + \frac{(K_{j+k} - K_{j+1})^2}{F_g(a_{j+k}) - F_g(b_{j+2})} \tag{21}$$

for $j = 0, 1, \ldots, M - 1$ and $k = 1, \ldots, M - j$.

We begin with the proof of the inequality (20). It easy to see that the inequality (20) for $j = 1$ is trivial because $K_0 = 0$ and for $j > 1$, the inequality

$$0 \leq \frac{K_{j-1}^2}{F_g(a_{j-1}) - F_g(b_1)} + \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} - \frac{K_j^2}{F_g(a_j) - F_g(b_1)}$$

is equivalent to

$$0 \leq A \cdot K_{j-1}^2 + B \cdot K_{j-1} + C \tag{22}$$

with

$$A = \frac{1}{F_g(a_{j-1}) - F_g(b_1)} + \frac{1}{F_g(a_j) - F_g(b_j)},$$

$$B = -\frac{2 \cdot K_j}{F_g(a_j) - F_g(b_j)},$$

and

$$C = K_j^2 \cdot \left(\frac{1}{F_g(a_j) - F_g(b_j)} - \frac{1}{F_g(a_j) - F_g(b_1)} \right).$$

The right hand of (22) is a parabola in K_{j-1} with vertex

$$V = \frac{K_j \cdot \Big(F_g(a_{j-1}) - F_g(b_1) \Big)}{\Big(F_g(a_j) - F_g(b_j) + F_g(a_{j-1}) - F_g(b_1) \Big)}.$$

The value of the parabola at the vertex is:

$$
\begin{aligned}
&A \cdot V^2 + B \cdot V + C = \qquad\qquad (23)\\
&\quad K_j^2 \cdot \frac{\Big(F_g(b_j) - F_g(b_1)\Big)}{\Big(F_g(a_j) - F_g(b_j)\Big)\Big(F_g(a_j) - F_g(b_1)\Big)} - \\
&\quad K_j^2 \cdot \frac{\Big(F_g(a_{j-1}) - F_g(b_1)\Big)}{\Big(F_g(a_j) - F_g(b_j)\Big)\Big(F_g(a_j) - F_g(b_j) + F_g(a_{j-1}) - F_g(b_1)\Big)} = \\
&\quad K_j^2 \cdot \frac{\Big(F_g(b_j) - F_g(a_{j-1})\Big)}{\Big(F_g(a_j) - F_g(b_1)\Big)\Big(F_g(a_j) - F_g(b_j) + F_g(a_{j-1}) - F_g(b_1)\Big)}.
\end{aligned}
$$

Because $a_j \geq b_j \geq a_{j-1} \geq b_1$ for all $j > 1$ it is clear that the value of the parabola at the vertex given by (23) is greater than or equal to zero. For the same reason, the value of the coefficient A of the parabola is greater or equal to zero and consequently the inequality (20) holds.

The proof of inequality (21) is similar. For $j = M - 1$ we have $k = 1$ and the inequality is trivial because $K_M = NF_y(t)$. For $j < M - 1$ and $k = 1$, the inequality is trivial too. Therefore, we consider the case $j < M - 1$ and $k = 2, \ldots, M - j$. In this case, the inequality

$$0 \leq \frac{(K_{j+1} - K_j)^2}{F_g(a_{j+1}) - F_g(b_{j+1})} + \frac{(K_{j+k} - K_{j+1})^2}{F_g(a_{j+k}) - F_g(b_{j+2})} - \frac{(K_{j+k} - K_j)^2}{F_g(a_{j+k}) - F_g(b_{j+1})}$$

is equivalent to

$$0 \leq D \cdot K_{j+1}^2 + E \cdot K_{j+1} + F \qquad (24)$$

with

$$
\begin{aligned}
D &= \frac{1}{F_g(a_{j+1}) - F_g(b_{j+1})} + \frac{1}{F_g(a_{j+k}) - F_g(b_{j+2})}\\
E &= -2 \cdot \left(\frac{K_j}{F_g(a_{j+1}) - F_g(b_{j+1})} + \frac{K_{j+k}}{F_g(a_{j+k}) - F_g(b_{j+2})}\right)\\
F &= K_j^2 \cdot \left(\frac{1}{F_g(a_{j+1}) - F_g(b_{j+1})} - \frac{1}{F_g(a_{j+k}) - F_g(b_{j+1})}\right)\\
&+ K_{j+k}^2 \cdot \left(\frac{1}{F_g(a_{j+k}) - F_g(b_{j+2})} - \frac{1}{F_g(a_{j+k}) - F_g(b_{j+1})}\right)\\
&+ \frac{2 \cdot K_j \cdot K_{j+k}}{F_g(a_{j+k}) - F_g(b_{j+1})}.
\end{aligned}
$$

Thus, the right hand of (24) is a parabola and its vertex is

$$V = \frac{K_j \cdot \Big(F_g(a_{j+k}) - F_g(b_{j+2})\Big) + K_{j+k} \cdot \Big(F_g(a_{j+1}) - F_g(b_{j+1})\Big)}{\Big(F_g(a_{j+k}) - F_g(b_{j+2}) + F_g(a_{j+1}) - F_g(b_{j+1})\Big)}$$

If we replace the vertex in the expression of the parabola given by (24), we have:

$$D \cdot V^2 + E \cdot V + F = \frac{\Big(K_{j+k} - K_j\Big)^2 \cdot \Big(F_g(b_{j+2}) - F_g(a_{j+1})\Big)}{\Big(F_g(a_{j+k}) - F_g(b_{j+2}) + F_g(a_{j+1}) - F_g(b_{j+1})\Big) \cdot \Big(F_g(a_{j+k}) - F_g(b_{j+1})\Big)}. \qquad (25)$$

For $j = 0, 1, \ldots, M-1$ and $k = 2, \ldots, M-j$, it is clear that

$$a_{j+k} \geq b_{j+2} \geq a_{j+1} \geq b_{j+1}$$

and consequently the value (25) and the coefficient D are greater or equal to zero. Thus, the inequality (21) holds.

Now, if we compare the value (18) with (16), we have

$$G_{2M}(t_1, \ldots, t_{2M}) - G_{2M}(\mathbf{t}_{opt}) = \sum_{j=1}^{M} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} - \frac{\Big(NF_y(t)\Big)^2}{F_g(a_M)}.$$

Because $K_M = NF_y(t)$, if we apply the inequality (21) recursively for $k = M-j$ and $j = 0, 1, \ldots, M-1$; it is easy to see that

$$\frac{\Big(NF_y(t)\Big)^2}{F_g(a_M)} \leq \frac{\Big(NF_y(t) - K_0\Big)^2}{F_g(a_M) - F_g(b_1)} \leq \sum_{j=1}^{M} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)}$$

Therefore

$$G_{2M}(t_1, \ldots, t_{2M}) - G_{2M}(\mathbf{t}_{opt}) \geq 0.$$

Next, we compare the value (19) with (16). Thus

$$G_{2M}(A_1, \ldots, A_{2M}) - G_{2M}(\mathbf{t}_{opt}) = -\Big(NF_y(t) - K_{h_{2M}}\Big)^2 - \sum_{j=1}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} + \sum_{j=1}^{M} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)}$$

First, we consider the case where $A_{2M} < a_M$, then exits some $l \in \{1, \ldots, M-1\}$ such that $a_l \leq A_{2M} \leq b_{l+1}$ and consequently

$$G_{2M}(A_1, \ldots, A_{2M}) - G_{2M}(\mathbf{t}_{opt}) = \\ -\Big(NF_y(t) - K_l\Big)^2 - \sum_{j=1}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} + \sum_{j=1}^{M} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)}$$

with $a_0 \leq A_j \leq b_{l+1}$ for all $j \in \{1, \ldots, 2M\}$.

Now, it is clear that

$$\Big(NF_y(t) - K_l\Big)^2 \leq \frac{\Big(NF_y(t) - K_l\Big)^2}{F_g(a_M) - F_g(b_{l+1})}$$

and applying the inequality (21) recursively for $j = l, l+1, \ldots, M-1$ and $k = M - j$; we have

$$\frac{\Big(NF_y(t) - K_l\Big)^2}{F_g(a_M) - F_g(b_{l+1})} \leq \sum_{j=l+1}^{M} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)}$$

or equivalently

$$\sum_{j=l+1}^{M} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} - \Big(NF_y(t) - K_l\Big)^2 = H_M \geq 0$$

then

$$G_{2M}(A_1, \ldots, A_{2M}) - G_{2M}(\mathbf{t}_{opt}) = \\ H_M - \sum_{j=1}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} + \sum_{j=1}^{l} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)}$$

with $H_M \geq 0$ and $a_0 \leq A_j \leq b_{l+1}$ for all $j \in \{1, \ldots, 2M\}$.

Next, if we consider that $h_1 = 0$, we have $K_{h_1} = K_0 = 0$. Thus

$$\sum_{j=1}^{2} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} = \frac{0}{F_g(A_1)} + \frac{(K_{h_2} - 0)^2}{F_g(A_2) - F_g(A_1)}$$

with $a_0 \leq A_1 \leq b_1$ and $a_{h_2} \leq A_2 \leq b_{h_2+1}$ and therefore

$$\frac{(K_{h_2})^2}{F_g(A_2) - F_g(A_1)} \leq \frac{(K_{h_2})^2}{F_g(a_{h_2}) - F_g(b_1)}$$

Now, if we apply the inequality (20) recursively for $j = h_2, h_2 - 1, \ldots, 1$; it is easy to see that

$$\frac{(K_{h_2})^2}{F_g(a_{h_2}) - F_g(b_1)} \leq \sum_{j=1}^{h_2} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)}$$

and

$$\sum_{j=1}^{h_2} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} - \sum_{j=1}^{2} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} = H_1 \geq 0$$

Consequently

$$\begin{aligned} &G_{2M}(A_1, \ldots, A_{2M}) - G_{2M}(\mathbf{t}_{opt}) = \\ &\qquad H_M + H_1 - \sum_{j=3}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} + \sum_{j=h_2+1}^{l} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} \end{aligned}$$

with $H_M; H_1 \geq 0$; $a_{h_2} \leq A_j \leq b_{l+1}$ and $h_2 \leq h_j \leq l$ for $j = 3, \ldots, 2M$.

Now, for $j = 3, \ldots, 2M$, we can consider three cases: $h_j = h_{j-1}$; $h_j = h_{j-1} + 1$ or exists k, with $2 \leq k \leq M - h_{j-1}$ such that $h_j = h_{j-1} + k$. If we consider that $h_j = h_{j-1}$ then

$$\frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} = 0$$

If $h_j = h_{j-1} + 1$, then exist some c with $h_2 \leq c \leq l$ such that $h_{j-1} = c$; $h_j = c + 1$; $a_c \leq A_{j-1} \leq b_{c+1}$ and $a_{c+1} \leq A_j \leq b_{c+2}$. Thus

$$\frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} = \frac{(K_{c+1} - K_c)^2}{(F_g(A_j) - F_g(A_{j-1}))} \leq \frac{(K_{c+1} - K_c)^2}{(F_g(a_{c+1}) - F_g(b_{c+1}))}.$$

Finally, if exists k, with $2 \leq k \leq M - h_{j-1}$ such that $h_j = h_{j-1} + k$, then exist some $h_2 \leq c \leq l$ with

$$h_{j-1} = c; \;\; h_j = c + k; \;\; a_c \leq A_{j-1} \leq b_{c+1}; \text{ and } a_{c+k} \leq A_j \leq b_{c+k+1}.$$

Therefore

$$\frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} = \frac{(K_{c+k} - K_c)^2}{(F_g(A_j) - F_g(A_{j-1}))} \leq \frac{(K_{c+k} - K_c)^2}{(F_g(a_{c+k}) - F_g(b_{c+1}))}.$$

Now, applying the inequality (21) recursively for $j = c, c+1, \ldots, c+k-2$ and $k = h_j - h_{j-1}$; we have

$$\frac{(K_{c+k} - K_c)^2}{(F_g(a_{c+k}) - F_g(b_{c+1}))} \leq \sum_{j=c+1}^{c+k} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)}.$$

Thus, in any case, it is easy to see that

$$\sum_{j=h_2+1}^{l} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} - \sum_{j=3}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} \geq 0$$

and consequently

$$G_{2M}(A_1, \ldots, A_{2M}) - G_{2M}(\mathbf{t}_{opt}) \geq 0.$$

On he other hand, if we consider $h_1 > 0$, then $a_{h_1} \leq A_1 \leq b_{h_1+1}$ and

$$\frac{(K_{h_1} - K_{h_0})^2}{(F_g(A_1) - F_g(A_0))} = \frac{(K_{h_1})^2}{F_g(A_1)} \leq \frac{(K_{h_1})^2}{F_g(a_{h_1}) - F_g(b_1)}.$$

Now, if we apply the inequality (20) recursively for $j = h_1, h_1 - 1, \ldots, 1$; it is easy to see that

$$\frac{(K_{h_1})^2}{F_g(a_{h_1}) - F_g(b_1)} \leq \sum_{j=1}^{h_1} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)}.$$

Thus, denoting by

$$H_1 = \sum_{j=1}^{h_1} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} - \frac{(K_{h_1} - K_{h_0})^2}{(F_g(A_1) - F_g(A_0))} \geq 0$$

and replacing H_1 in (26) we have

$$\begin{aligned} & G_{2M}(A_1, \ldots, A_{2M}) - G_{2M}(\mathbf{t}_{opt}) = \\ & \quad H_M + H_1 - \sum_{j=2}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} + \sum_{j=h_1+1}^{l} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} \end{aligned}$$

with $H_M, H_1 \geq 0$.

Now, similarly to the case $h_1 = 0$, it can be shown that

$$G_{2M}(A_1, \ldots, A_{2M}) - G_{2M}(\mathbf{t}_{opt}) \geq 0. \tag{26}$$

Thus, for the case where $A_{2M} < a_M$ we have (26).

Finally, if we consider the case where $A_{2M} \geq a_M$, then $K_{h_{2M}} = NF_y(t)$ and

$$G_{2M}(A_1, \ldots, A_{2M}) - G_{2M}(\mathbf{t}_{opt}) = \sum_{j=1}^{M} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} - \sum_{j=1}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))}$$

In a similar way to the case $A_{2M} < a_M$, it can be shown that

$$\sum_{j=1}^{M} \frac{(K_j - K_{j-1})^2}{F_g(a_j) - F_g(b_j)} - \sum_{j=1}^{2M} \frac{(K_{h_j} - K_{h_{j-1}})^2}{(F_g(A_j) - F_g(A_{j-1}))} \geq 0$$

and the inequality (26) holds in any case. Consequently, the auxiliary vector $\mathbf{t}_{opt}$ given by (14) is the vector where the function G_{2M} attains the global minimum when b_1 exits and for all $i = 1, \ldots, M-1$; $b_{i+1} \neq a_i$.

Now, if for some $i_1, i_2, \ldots i_R \in \{0, 1, \ldots, M-1\}$; $a_{i_1} = b_{i_1+1}$ with $R \leq M$ and $i_h \neq i_j$ if $h \neq j$ the vector $\mathbf{t}_{opt}$ is a vector with dimension $2M - R$ given by (15) and the value of function $G_{2M}(t_1, \ldots, t_{2M})$ at the vector $\mathbf{t}_{opt}$ is:

$$G_{2M}(\mathbf{t}_{opt}) = \Big(NF_y(t)\Big)^2 - \sum_{\substack{j=1 \\ j \neq i_1, \ldots, i_R}}^{M} \frac{(K_j - K_{j-1})^2}{(F_g(a_j) - F_g(b_j))} - \sum_{j=1}^{R} \frac{(K_{i_j+1} - K_{i_j})^2}{(F_g(a_{i_j+1}) - F_g(a_{i_j}))} \tag{27}$$

To prove that the vector $\mathbf{t}_{opt}$ is the vector where the function G_{2M} attains the global minimum, we need inequalities (20) and (21) and we also need the following inequalities that we can establish in a similar way

$$\frac{(K_{i_j+k} - K_{i_j-1})^2}{F_g(a_{i_j+k}) - F_g(b_{i_j})} \leq \frac{(K_{i_j} - K_{i_j-1})^2}{F_g(a_{i_j}) - F_g(b_{i_j})} + \frac{(K_{i_j+k} - K_{i_j})^2}{F_g(a_{i_j+k}) - F_g(a_{i_j})} \tag{28}$$

for $j = 1, \ldots, R$ and $k = 1, \ldots, M - i_j$.

$$\frac{(K_{i_j+k} - K_{i_j})^2}{F_g(a_{i_j+k}) - F_g(a_{i_j})} \leq \frac{(K_{i_j+1} - K_{i_j})^2}{F_g(a_{i_j+1}) - F_g(a_{i_j})} + \frac{(K_{i_j+k} - K_{i_j+1})^2}{F_g(a_{i_j+k}) - F_g(b_{i_j+2})} \tag{29}$$

for $j = 1, \ldots, R$ and $k = 1, \ldots, M - i_j$.

With the inequalities (20); (21);(28) and (29), we can show, as in the previous case, that the value of the function G_{2M} at the vector $\mathbf{t}_{opt}$ given by (27) is less than or equal to the minimum of the function G_{2M} in each piece and therefore the function G_{2M} attains the global minimum at $\mathbf{t}_{opt}$.

5. The optimum estimator with estimated optimal vector

The optimal auxiliary vector $\mathbf{t}_{opt}$ depends on some unknown values, thus a calibration estimator based on $\mathbf{t}_{opt}$ cannot be calculated. In the absence of good a priori knowledge these characteristics, we go to replace the optimal vector $\mathbf{t}_{opt}$ by sample-based estimates. For it, given a point t for which we want to estimate $F_y(t)$, we consider the the following sets based on the sample s

$$A_{st} = \{g_k \in s \ : y_k \leq t\} = \{a_1, a_2, \dots, a_m\} \tag{30}$$

with $a_1 < a_2 < \dots < a_m$ and

$$B_{st} = \{b_1, b_2, \dots, b_m\} \tag{31}$$

with

$$b_1 = \max_{l \in U_{1_s}} \{g_l\} \text{ where } U_{1_s} = \{l \in s : g_l < a_1\}$$

$$b_h = \max_{l \in U_{h_s}} \{g_l\} \text{ where } U_{h_s} = \{l \in s : a_{h-1} \leq g_l < a_h\}, \quad h = 2, 3, \dots, m.$$

If b_1 exits and for all $i = 1, \dots, m-1$; $b_{i+1} \neq a_i$, we consider the auxiliary vector $\widehat{\mathbf{t}}_{\mathbf{OP}}$ given by:

$$\widehat{\mathbf{t}}_{\mathbf{OP}} = (\widehat{t}_{O1}, \dots, \widehat{t}_{O2M}) = (b_1, a_1, b_2, a_2, \dots, b_m, a_m) \tag{32}$$

If for some $i_1, i_2, \dots i_r \in \{0, 1, \dots, m-1\}$; $a_{i_1} = b_{i_1+1}$ with $r \leq m$ and $i_h \neq i_j$ if $h \neq j$ we consider the auxiliary vector $\widehat{\mathbf{t}}_{\mathbf{OP}}$ given by:

$$\mathbf{t}_{opt} = (b_1, a_1, b_2, a_2, \dots, b_{i_1}, a_{i_1}, a_{i_1+1}, b_{i_1+2}, \dots, b_{i_h}, a_{i_h}, a_{i_h+1}, b_{i_h+2}, \dots b_m, a_m). \tag{33}$$

Thus, we can define a new calibration estimator $\widehat{F}_{CALOPT}(t)$ based on the auxiliary vector $\widehat{\mathbf{t}}_{\mathbf{OP}}$ obtained.

6. Numerical comparisons

In this section we present the results of a Monte Carlo comparison of the various estimators of $F(t)$. We compare the precision of the proposed optimal calibration estimator $\widehat{F}_{CALOPT}(t)$ with the following estimators:

- the Horvitz Thompson estimator, $\widehat{F}_{HT}$,
- the Chamber Dunstan estimator, $\widehat{F}_{CD}(t)$ ([?]),
- the ratio estimator, $\widehat{F}_{R}(t)$ ([?]),
- the difference estimator, $\widehat{F}_{D}(t)$ ([?]),
- the Rao, Kovar and Mantel estimator, $\widehat{F}_{RKM}(t)$ ([?]),
- the calibration estimator with $t_1 = Q_g(0.5)$, the population median, as point for calibration, $\widehat{F}_{CAL}(t)$,
- the calibration estimator with one optimal point, $\widehat{F}_{CALMAX}(t)$ ([?]),
- the calibration estimator 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, $\widehat{F}_{CAL.3}(t)$.
- and finally the optimal calibration estimator $\widehat{F}_{CALOPT}(t)$.

Our simulations are programmed in R.

To investigate the efficiency of the estimators under a variety of models for the relationship between y and x, we consider three populations.

The first population is the datasets ToothGrowth included in The R Datasets Package "The Effect of Vitamin C on Tooth Growth in Guinea Pigs". The response, y, is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg), the auxiliary variable x.

The other two populations are also in the library datasets. The DNase data frame contains data obtained during development of an ELISA assay for the recombinant protein DNase in rat serum. As main variable y we select density, the measured optical density (dimensionless) in the assay and as auxiliary variable, x we select the known concentration of the protein. And the Loblolly data frame with records of the growth of Loblolly pine trees. In this case the main variable is the tree heights (ft), and the auxiliary variable the tree ages (yr).

In the simulation study we drawn 1000 samples of several sizes by simple random sampling without replacement. For each sample and for each estimator, estimates of the distribution function $F(t)$ were calculated for 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.

The performance of all the estimators is measured by means of the average relative bias (AVRB) and the average relative efficiency (AVRE), given respectively by

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

where RB and RE are defined as

$$\text{RB}(t) = \frac{1}{B}\sum_{b=1}^{B} \frac{\widehat{F}(t)_b - F_y(t)}{F_y(t)} \quad \text{and} \quad \text{RE}(t) = \frac{MSE[\widehat{F}(t)]}{MSE[\widehat{F}_{HT}(t)]}, \tag{34}$$

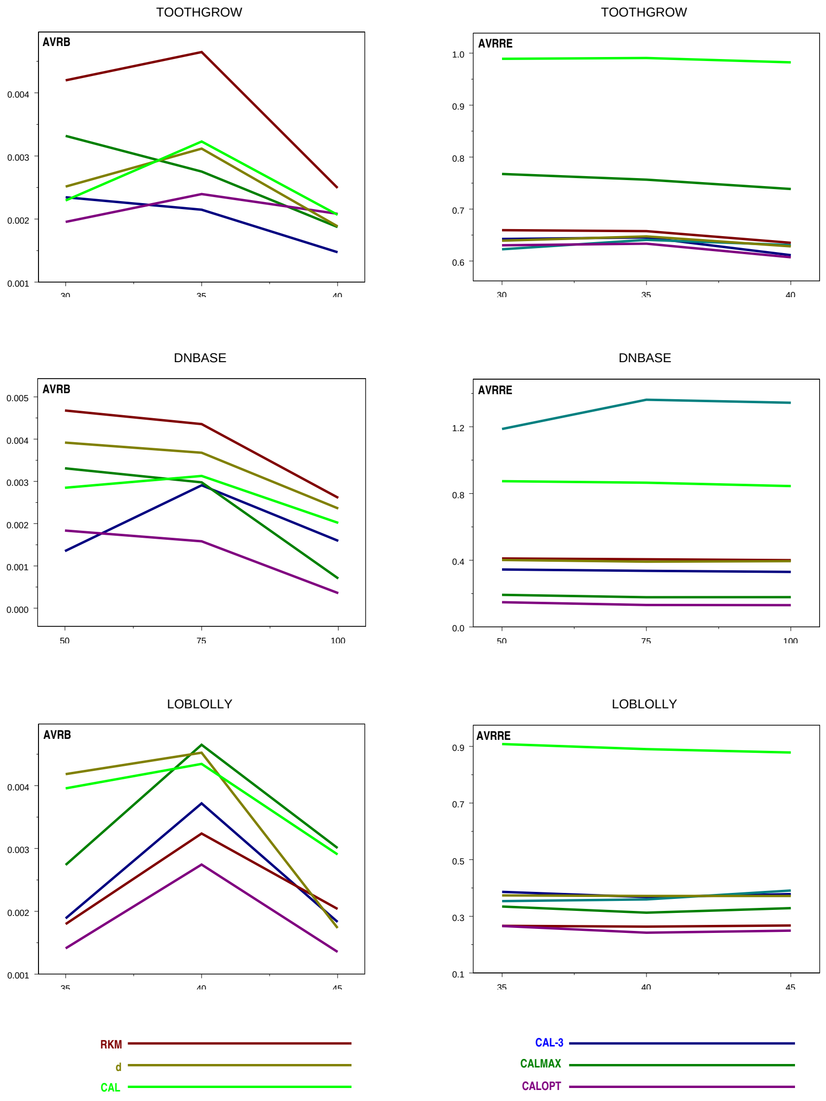
where b indexes the bth simulation run, $\widehat{F}(t)$ is an estimator for the distribution function, $MSE[\widehat{F}(t)] = B^{-1}\sum_{b=1}^{B}[\widehat{F}(t)_b - F_y(t)]^2$ is the empirical mean square error for $\widehat{F}(t)$ and $MSE[\widehat{F}_{HT}(t)]$ is similarly defined for the Horvitz-Thompson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

Figure 1 gives the values of the average relative bias and the average relative efficiency for all populations.

Some observations:

- The $\widehat{F}_{CD}$ estimator has a serious problem of bias, as previously indicated by ([**?**]). This is expected because the CD-estimator is most susceptible to model misspecification. $\widehat{F}_{CD}$ estimator has not been included in Figure 1. By similar raison, the ratio estimator $\widehat{F}_r$ is also excluded from Figure 1.

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



- We found no evidence of any significant bias for the other estimators considered.

- In terms of precision, the best overall performance is achieved by the optimal calibration estimator, $\widehat{F}_{CALOPT}(t)$.

Other simulations studies also show the potential gains from the use of the proposed calibration estimator with optimal points instead of the customary estimators used in the literature. From a computational point of view, the proposed optimal estimator is more efficient that the calibration estimator with one optimal point [**?**], because last theorem yields the calibration points.

In conclusion, we suggest that the study of optimum points for calibration provides a practical approach to estimating distribution functions, and offers useful gains in efficiency.

Acknowledgements

This study was partially supported by Ministerio de Educación y Ciencia (grant MTM2012-35650, Spain) and by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía).

References

Figure 2: Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared.

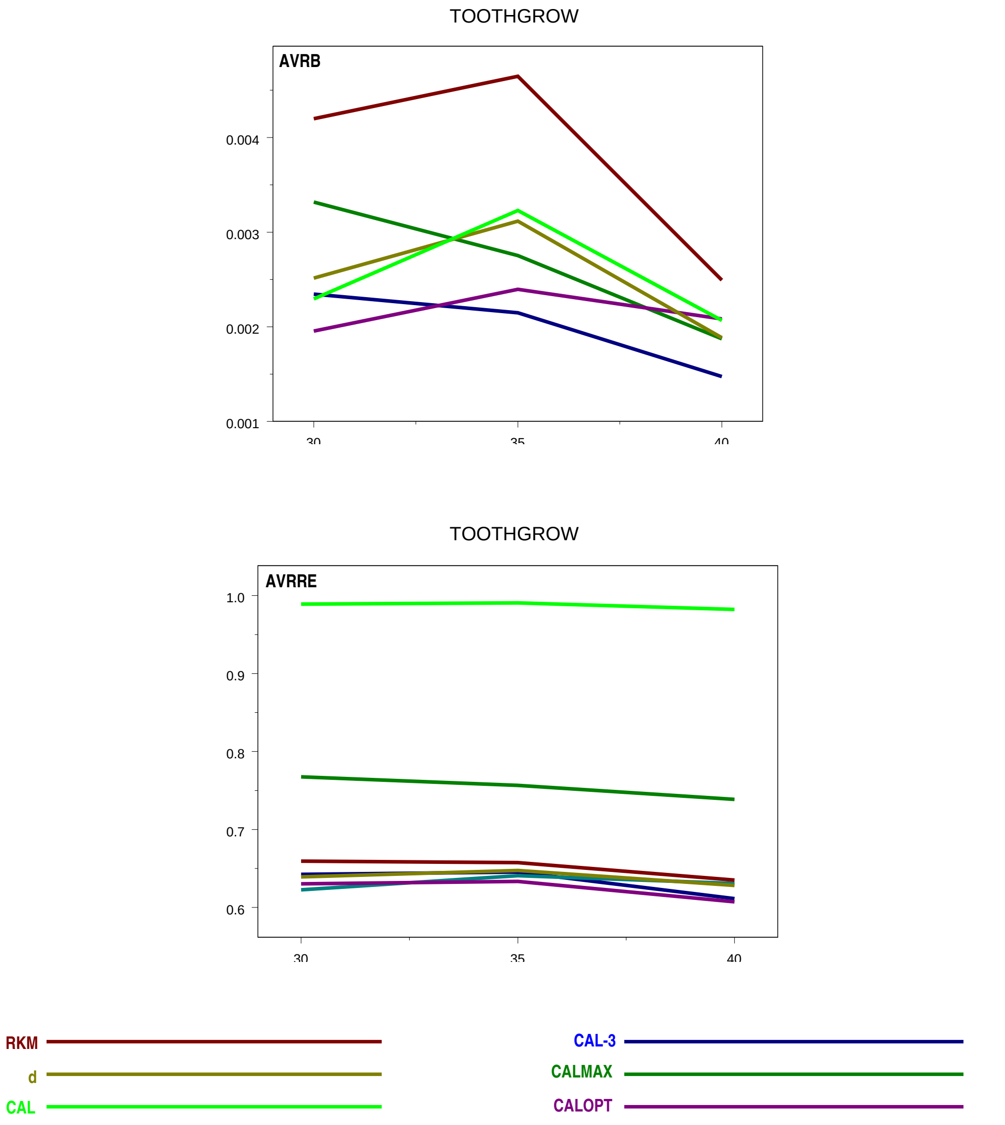


Figure 3: Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared.

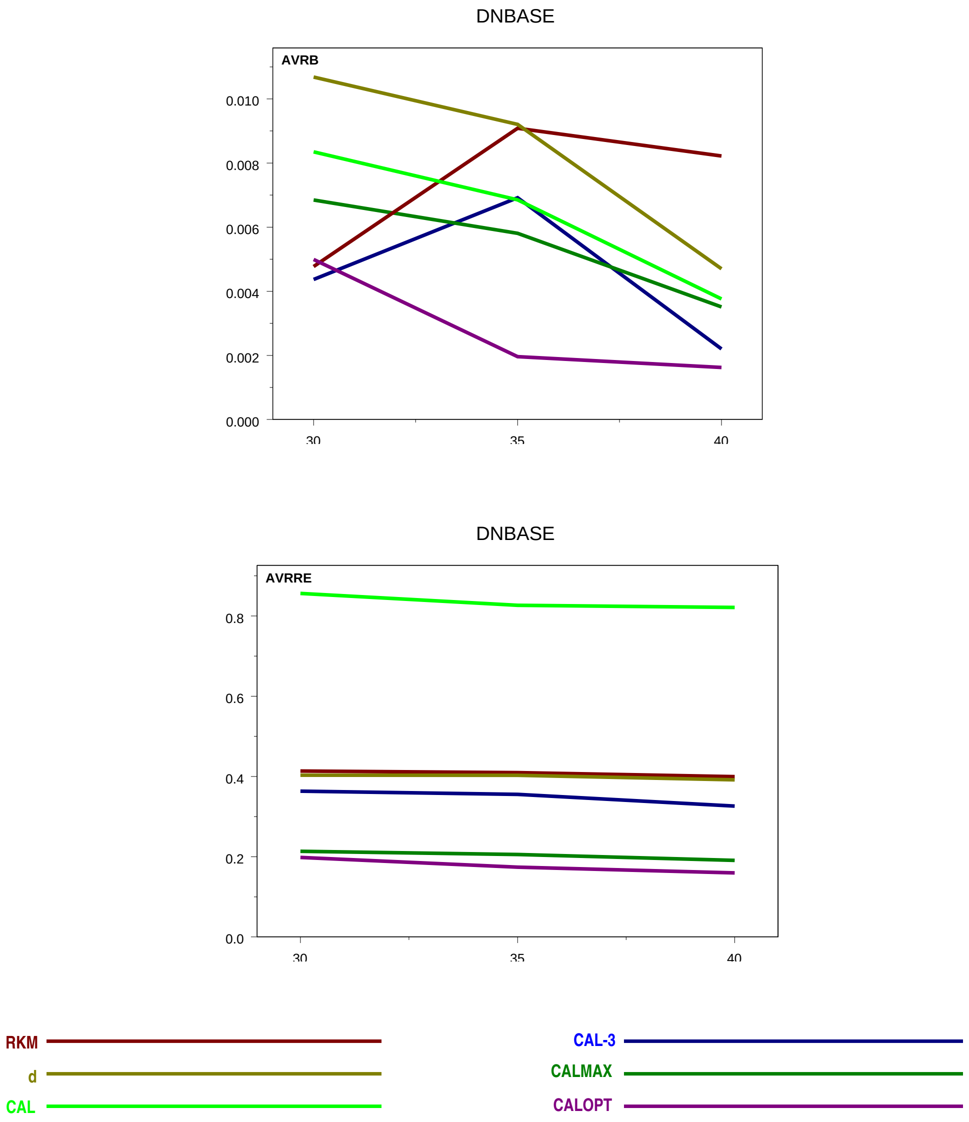
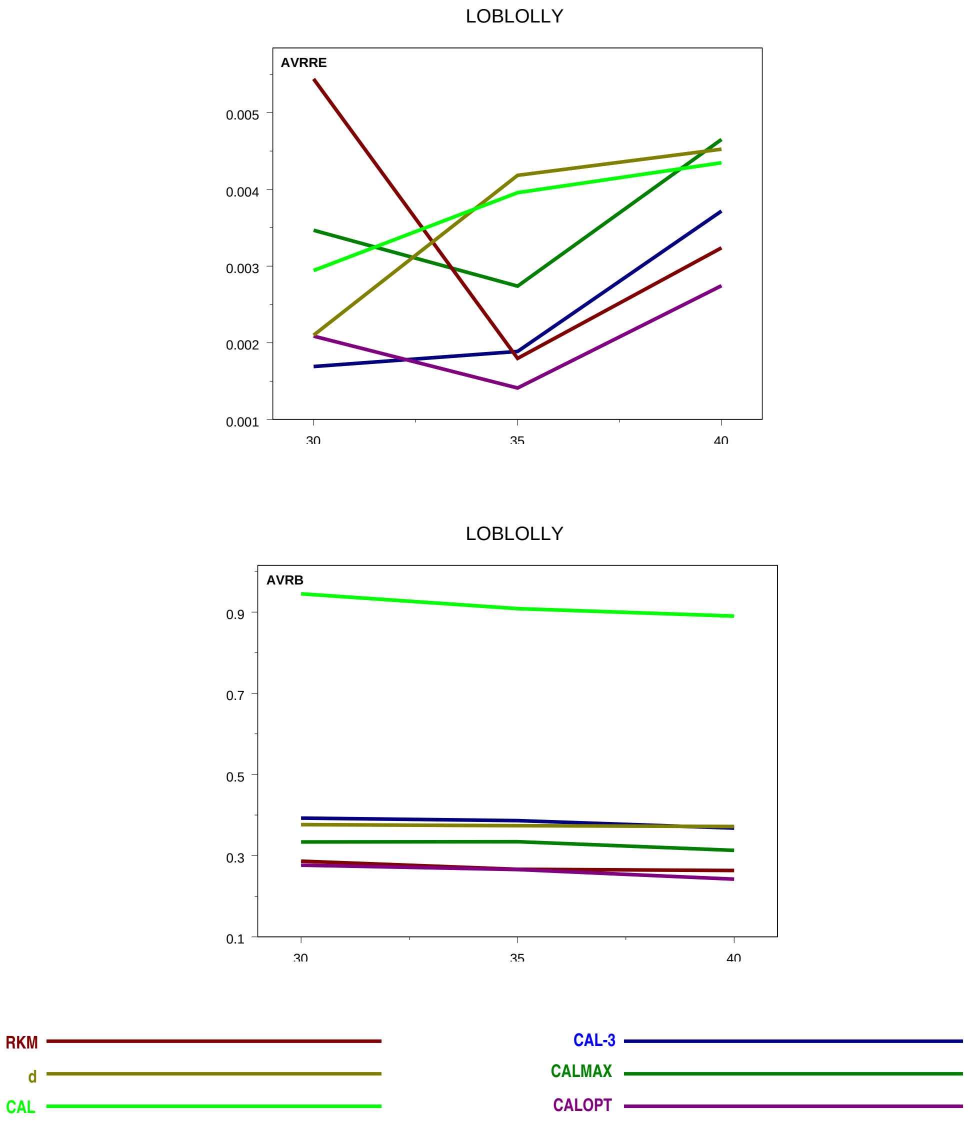


Figure 4: Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared.


LOBLOLLY
AVRRE
LOBLOLLY
AVRB
RKM
d
CAL
CAL-3
CALMAX
CALOPT