Maximum Entropy Estimation for Survey Sampling

Fabrice Gamboa, Jean-Michel Loubes and Paul Rochet

Abstract

Calibration methods have been widely studied in survey sampling over the last decades. Viewing calibration as an inverse problem, we extend the calibration technique by using a maximum entropy method. Finding the optimal weights is achieved by considering random weights and looking for a discrete distribution which maximizes an entropy under the calibration constraint. This method points a new frame for the computation of such estimates and the investigation of its statistical properties.

Keywords: Inverse Problems; Calibration; Bayesian Estimation
Subject Class. MSC-2000 : 62F12, 62D05, 94A17

Introduction

Calibration is a well spread method to improve estimation in survey sampling, using extra information from auxiliary variables. This method provides approximately unbiased estimators with variance smaller than that of the usual Horvitz-Thompson estimator. Calibration has been introduced by Deville and Särndal in [3], extending an idea of [4]. For general references, we refer to [20] and for an extension to variance estimation to [22]. Finding the solution to a calibration equation involves minimizing a distance under some constraint. More precisely, let s be a random sample of size n drawn from a population U of size N, y be the variable of interest and $\mathbf{x}$ be a given auxiliary vector variable, for which the total $t_{\mathbf{x}}$ over the population is known. Further, let $d \in \mathbb{R}^n$ be the standard sampling weights (that is the Horvitz-Thompson ones). Calibration derives an estimator $\hat{t}_y = \sum_{i \in s} w_i y_i$ of the population total t_y of y. The weights w_i are chosen to minimize a dissimilarity (or distance) $\mathcal{D}(., d)$ on $\mathbb{R}^n$ with respect to the Horvitz-Thompson weights d_i and under the constraint

$$\sum_{i \in s} w_i \mathbf{x}_i = t_{\mathbf{x}}. \tag{1}$$

Following [23], we will view here calibration as a linear inverse problem. In this paper, we use Maximum Entropy Method on the Mean (MEM) to build the calibration weights.

Indeed, MEM is a strong machinery for solving linear inverse problems. It tackles a linear inverse problem by finding a measure maximizing an entropy under some suitable constraint. It has been extensively studied and used in many applications, see for example [1], [12], [10], [14], [9], [7] or [13].
Let us roughly explain how MEM works in our context. First we fix a *prior* probability measure ν on $\mathbb{R}^n$ with mean value equal to d. Then, the idea is to modify the standard weights d in order to get a representative sample for the auxiliary variable $\mathbf{x}$, but still being as close as possible to d, which have the desirable property of yielding an unbiased estimate for the population total. So, we will look for a *posterior* probability measure minimizing the entropy (or Kullback information) with respect to ν and satisfying a constraint related to (1). It appears that the MEM estimator is in fact a specific calibration estimator for which the corresponding dissimilarity $\mathcal{D}(.,d)$ is determined by the choice of the prior distribution ν. Hence, the MEM methodology provides a general Bayesian frame to fully understand calibration procedures in survey sampling where the different choices of dissimilarities appear as different choices of prior distributions.
An important problem when studying calibration methods is to understand the amount of information contained in the auxiliary variable. Indeed, the relationships between the variable to be estimated and the auxiliary variable are crucial to improve estimation (see for example [18], [26] or [25]). When complete auxiliary information is available, *model calibration* proposed by Wu and Sitter [26] aims to increase the correlation between the variables by replacing the auxiliary variable $\mathbf{x}$ by some function of it, say $u(\mathbf{x})$. We consider efficiency issues for a collection of calibration estimators, depending on both the choice of the auxiliary variable and the dissimilarity. Finally, we provide an optimal way of building an efficient estimator using the MEM methodology.

The article falls into the following parts. The first section recalls the calibration method in survey sampling, while the second exposes the MEM methodology in a general framework, and its application to calibration and instrument estimation. Section 3 is devoted to the choice of a data driven calibration constraint in order to build an efficient calibration estimator. It is shown to be optimal under strong asymptotic assumptions on the sampling design. Proofs are postponed to Section 4.

1 Calibration Estimation of a linear parameter

Consider a large population $U = \{1, ..., N\}$ and an unknown characteristic $y = (y_1, ..., y_N) \subset \mathbb{R}^N$. Our aim is to estimate its total $t_y := \sum_{i \in U} y_i$ when only a random subsample s of the whole population is available. So the observed data are $(y_i)_{i \in s}$. Each sample s has a probability $p(s)$ of being observed. The distribution $p(.)$ is called sampling design. We assume that $\pi_i := p(i \in s) = \sum_{s,\ i \in s} p(s)$ is strictly positive for all $i \in U$ so that $d_i = 1/\pi_i$ is well defined. A standard estimator of t_y is given by the

Horvitz-Thompson estimator:

$$\hat{t}_y^{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i.$$

This estimator is unbiased and is widely used for practical cases, see for instance [11].

Suppose that it exists an auxiliary vector variable $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_N)$ entirely observed and set $t_{\mathbf{x}} = \sum_{i \in U} \mathbf{x}_i \in \mathbb{R}^k$. If the Horvitz-Thompson estimator of $t_{\mathbf{x}}$, $\hat{t}_{\mathbf{x}}^{HT} = \sum_{i \in s} d_i \mathbf{x}_i$ is far from the true value $t_{\mathbf{x}}$, we may reasonably assume that the sample will not adequately reflect the behavior of the variable of interest in the whole population. So, to prevent inefficient estimation due to bad sample selection, inference on the sample can be achieved by considering a modification of the weights of the individuals chosen in the sample.

One of the main methodology used to correct this effect is the calibration method, (see [3]). The *bad sample effect* is corrected by replacing the Horvitz-Thompson weights d_i by new weights w_i close to d_i. Let $w \mapsto \mathcal{D}(w, d)$ be a dissimilarity between w and the Horvitz-Thompson weights that is minimal for $w_i = d_i$. The method consists in choosing weights $\hat{w}_i$ minimizing $\mathcal{D}(., d)$ under the constraint

$$\sum_{i \in s} \hat{w}_i \mathbf{x}_i = t_{\mathbf{x}}.$$

Then, consider the new weighted estimators $\hat{t}_y = \sum_{i \in s} \hat{w}_i y_i$.

A typical dissimilarity is the χ^2 distance $w \mapsto \sum_{i \in s} (\pi_i w_i - 1)^2 / (q_i \pi_i)$ for $(q_i)_{i \in s}$ some known positive sequence. In most applications, the q_i's are taken equal to 1 which generally warrants a consistent estimator. Nevertheless unequal weights can be used as treated in Example 1 in [3], in order to lay more or less stress on the distance between some of the the weights and the original Horvitz-Thompson ones. The new estimator is defined as $\hat{t}_y = \sum_{i \in s} \hat{w}_i y_i$, where the weights $\hat{w}_i$ minimizes $\mathcal{D}(w, d) = \sum_{i \in s} (\pi_i w_i - 1)^2 / q_i \pi_i$ under the constraint $\sum_{i \in s} \hat{w}_i \mathbf{x}_i = t_{\mathbf{x}}$. Denote by a^t the transpose of a, the solution of this minimization problem is given by

$$\hat{t}_y = \hat{t}_y^{HT} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}}^{HT})^t \hat{B},$$

where $\hat{B} = \left[\sum_{i \in s} q_i d_i \mathbf{x}_i \mathbf{x}_i^t\right]^{-1} \sum_{i \in s} q_i d_i y_i \mathbf{x}_i$. Note that this is a generalized regression estimator. It is natural to consider alternative dissimilarities, see for instance [3]. We first point out that the existence of a solution to the constrained minimization problem depends on the choice of the dissimilarities. Then, different choices can lead to weights with different behaviors, different ranges of values for the weights that may be found unacceptable by the users. We propose an approach where dissimilarities are given a probabilistic interpretation.

2 Maximum Entropy for Survey Sampling

2.1 MEM methodology

Consider the problem of recovering an unknown measure μ on a measurable space $\mathcal{X}$ under moment conditions. This issue belongs to the class of generalized moment problems with convex constraints (we refer to [5] for general references). This inverse problem has been widely studied and in particular it can be solved using the maximum entropy on the mean (MEM).
Here, we aim at estimating μ from random observations $T_1, ..., T_n \sim \mu$ and knowing that, there exists a given function $\tilde{\mathbf{x}} : \mathcal{X} \to \mathbb{R}^k$ and a known quantity $t_{\mathbf{x}} \in \mathbb{R}^k$, such that

$$\int_{\mathcal{X}} \tilde{\mathbf{x}} d\mu = t_{\mathbf{x}}. \tag{2}$$

Solving this problem using the MEM framework amounts to approximate the inverse problem (2) by a sequence of finite dimensional problems which are obtained by a discretization of the space $\mathcal{X}$ using the random sample $T_1, \ldots, T_n$. For this, first consider the empirical distribution $\mu_n = n^{-1} \sum_{i=1}^n \delta_{T_i}$, δ standing for the Dirac mass. The general idea is to modify μ_n in order to take into account the additional information on μ given by the moment equation (2). For this, we associate to each observation T_i a random weight P_i and consider the corresponding random weighted version of the empirical measure

$$\tilde{\mu}_n = \frac{1}{n} \sum_{i=1}^n P_i \delta_{T_i}.$$

Choosing properly the weights is the second step of the MEM procedure. The underlying idea is to incorporate some prior information by choosing $P = (P_1, ..., P_n)$, drawn from a finite measure ν^* close to a *prior* ν, and looking at the weighted measures satisfying the constraint in mean. This prior distribution conveys the information that $\hat{\mu}_n$ must be close, in a given sense, to the empirical distribution μ_n. More precisely, let first define the relative entropy or Kullback information between two finite measures Q, R on a space $(\Omega, \mathcal{A})$ by setting

$$K(Q, R) = \begin{cases} \int_\Omega \log\left(\frac{dQ}{dR}\right) dQ - Q(\Omega) + 1 & \text{if } Q \ll R \\ +\infty & \text{otherwise.} \end{cases}$$

Since this quantity is not symmetric, we will call it the relative entropy of Q with respect to R. Note also that, among the literature in optimization, the relative entropy is often defined as the opposite of the entropy defined above, which explains the name of maximum entropy method, while with our notations, we consider the minimum of the entropy.

Given our prior ν, we now define ν^* as the measure minimizing $K(.,\nu)$ under the constraint that the linear constraint holds in mean:

$$\mathbb{E}_{\nu^*}\left[n^{-1}\sum_{i=1}^n P_i\tilde{\mathbf{x}}_i\right] = \frac{1}{\nu^*(\mathbb{R}^n)}\int_{\mathbb{R}^n}\left[n^{-1}\sum_{i=1}^n p_i\tilde{\mathbf{x}}_i\right]d\nu^*(p_1,...,p_n) = t_{\mathbf{x}},$$

where we set $\tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}(T_i)$. We then build the MEM estimator as

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \hat{p}_i\delta_{T_i} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\nu^*}(P_i)\delta_{T_i}.$$

So, for a fixed n, $\hat{\mu}_n$ is the maximum entropy reconstruction of μ with reference $\nu^\star$. This method provides an efficient way to estimate some linear parameter $t_y = \int_{\mathcal{X}}\tilde{y}d\mu$ for $\tilde{y}:\mathcal{X}\to\mathbb{R}$ a given map. The empirical mean $\int_{\mathcal{X}}\tilde{y}d\mu_n$ is an unbiased and consistent estimator of t_y but may not have the smallest variance in this model. However, integrating $\tilde{y}$ against $\hat{\mu}_n$ provides an asymptotically unbiased estimate of t_y with a lower variance than the empirical mean (see [10]).

In many actual situations, the function $\tilde{\mathbf{x}}$ is unknown and only an approximation to it, say $\tilde{\mathbf{x}}_m$, is available. Under regularity conditions, the efficiency properties of the MEM estimator built with the approximate constraint have been studied in [15] and [16], introducing the approximate maximum entropy on the mean method (AMEM). More precisely, the AMEM estimate of the weights is defined as the expectation of the variable P under the distribution ν_m^* minimizing $K(.,\nu)$ under the approximate constraint

$$\mathbb{E}_{\nu_m^*}\left[n^{-1}\sum_{i=1}^n P_i\,\tilde{\mathbf{x}}_m(T_i)\right] = t_{\mathbf{x}}. \tag{3}$$

It is shown that, under assumptions on $\tilde{\mathbf{x}}_m$, the AMEM estimator of t_y obtained in this way is consistent as n and m tend to infinity. This procedure enables to increase the efficiency of a calibration estimator while remaining in a Bayesian framework, as shown in Section 3.2. This situation occurs for instance, when dealing with inverse problem with unknown operators which still can be approximated either using another sample or directly from the data. For instance, in econometric, when dealing with instrumental variables the operator which corresponds here to the function $\tilde{x}$ is unknown but can be estimated, see [2]. The practical case of aerosol remote sensing is tackled in [15].

2.2 Maximum entropy method for calibration

Recall that our original problem is to estimate the population total $t_y = \sum_{i\in U} y_i$ based on the observations $\{y_i, i\in s\}$ and auxiliary information $\{\mathbf{x}_i, i\in U\}$. We introduce the following notations:

$$\tilde{y}_i = nd_iy_i,\ \tilde{\mathbf{x}}_i = nd_i\mathbf{x}_i,\ p_i = \pi_iw_i.$$

The variables of interest are rescaled to match the MEM framework. The weights $(p_i)_{i\in s}$ are now identified with a discrete measure on the sample s. The Horvitz-Thompson estimator $\hat{t}_y^{HT} = \sum_{i\in s} d_i y_i = n^{-1}\sum_{i\in s}\tilde{y}_i$ is the preliminary estimator we aim at improving. The calibration constraint $n^{-1}\sum_{i\in s} p_i\tilde{\mathbf{x}}_i = t_{\mathbf{x}}$ stands for the linear condition satisfied by the discrete measure $(p_i)_{i\in s}$. In these settings, it appears that the calibration problem follows the pattern of maximum entropy on the mean. Let ν be a prior distribution on the vector of the weights $(p_i)_{i\in s}$. The solution $\hat{p} = (\hat{p}_i)_{i\in s}$ is the expectation of the random vector $P = (\pi_i W_i)_{i\in s}$ drawn from a *posterior* distribution ν^*, defined as the minimizer of the Kullback information $K(.,\nu)$ under the condition that the calibration constraint holds in mean

$$\mathbb{E}_{\nu^*}\left[n^{-1}\sum_{i\in s} P_i\tilde{\mathbf{x}}_i\right] = \mathbb{E}_{\nu^*}\left[\sum_{i\in s} W_i\mathbf{x}_i\right] = t_{\mathbf{x}}. \tag{4}$$

We take the solution $\hat{p} = \mathbb{E}_{\nu^*}(P)$ and define the corresponding MEM estimator $\hat{t}_y$ as

$$\hat{t}_y = n^{-1}\sum_{i\in s}\hat{p}_i\tilde{y}_i = \sum_{i\in s}\hat{w}_i y_i,$$

where we set $\hat{w}_i = d_i\hat{p}_i$ for all $i \in s$. Under the following assumptions, we will show in Theorem 2.1 that maximum entropy on the mean provides a Bayesian interpretation of calibration methods.

The random weights $P_i, i \in s$ (and therefore the $W_i, i \in s$) are taken independent. We denote by ν_i the prior distribution of P_i. It follows that $\nu = \otimes_{i\in s}\nu_i$. Moreover, all prior distributions ν_i are probability measures with mean 1. This last assumption conveys the information that $\hat{p}_i$ must be close to 1, equivalently, $\hat{w}_i = d_i\hat{p}_i$ must be close to d_i.
Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a closed convex map, the convex conjugate φ^* of φ is defined as

$$\forall s \in \mathbb{R},\ \varphi^*(s) = \sup_{t\in\mathbb{R}}(st - \varphi(t)).$$

For ν a probability measure on $\mathbb{R}$, we denote respectively by Λ_ν and Λ_ν^* the log-Laplace transform and Cramer transform of ν:

$$\begin{aligned}\Lambda_\nu(s) &= \log\int e^{sx}d\nu(x),\\ \Lambda_\nu^*(s) &= \sup_{t\in\mathbb{R}}(st - \Lambda_\nu(t)),\ s \in \mathbb{R}.\end{aligned}$$

Moreover, denote by S_ν the interior of the convex hull of the support of ν and let $D(\nu) = \{s \in \mathbb{R} :\ \Lambda_\nu(s) < \infty\}$. In the sequel, we will always assume that Λ_{ν_i} is essentially smooth (see [19]) for all i, strictly convex and that ν_i is not concentrated on a single point. The last assumption means that if $D(\nu_i) = (-\infty;\alpha_i)$, $(\alpha_i \leq +\infty)$, then $\Lambda'_{\nu_i}(s)$ goes to $+\infty$ whenever $\alpha_i < +\infty$ and s goes to α_i. Under these assumptions, Λ'_{ν_i} is an increasing bijection between the interior of $D(\nu_i)$ and S_{ν_i}. So, denote by $\psi_i = {\Lambda'_{\nu_i}}^{-1}$ its inverse

function, the Cramer transform $\Lambda^*_{\nu_i}$ of ν_i, which is defined as the convex conjugate of Λ_{ν_i}, satisfies

$$\Lambda^*_{\nu_i}(s) = s\psi_i(s) - \Lambda_{\nu_i}(\psi_i(s)).$$

Classical choices of priors ν_i lead to easily computable functions $\Lambda^*_{\nu_i}$ in most cases. Some examples are given at the end of the section.

Definition : We say that the optimization problem is feasible if there exists a vector $\delta = (\delta_i)_{i\in s} \in \otimes_{i\in s} S_{\nu_i}$ such that:

$$\sum_{i\in s} \delta_i \mathbf{x}_i = t_{\mathbf{x}}. \tag{5}$$

Under the last assumptions, the following proposition claims that the solutions $(\hat{w}_i)_{i\in s}$ are easily tractable.

Theorem 2.1 (survey sampling as MEM procedure) *Assume that the optimization problem is feasible. The MEM estimator $\hat{w} = (\hat{w}_1, ..., \hat{w}_n)$ minimizes over $\mathbb{R}^n$*

$$(w_1, ..., w_n) \mapsto \sum_{i\in s} \Lambda^*_{\nu_i}(\pi_i w_i)$$

under the constraint $\sum_{i\in s} \hat{w}_i \mathbf{x}_i = t_{\mathbf{x}}$.

Hence, we point out that maximum entropy on the mean method leads to calibration estimation, where the dissimilarity is determined by the Cramer transforms $\Lambda^*_{\nu_i}, i \in s$ of the prior distributions ν_i. Conditions we require on the priors in the MEM procedure correspond to regularity conditions on the dissimilarity. Indeed, taking priors ν_i with mean 1 yields $\Lambda^*_{\nu_i}(1) = \Lambda^{*\prime}_{\nu_i}(1) = 0$, which is a classical condition in calibration, see for instance [3] or Theorem 2.7.1 in [8]. To see that, apply Jensen inequality to $\Lambda_\nu(t) = \log \int e^{tx} d\nu(x)$ to show that $\Lambda_\nu(t) \geq t$, which implies $\Lambda^*_\nu(1) = 0$. Since Λ^*_ν is smooth, non negative and strictly convex by construction, we also get $\Lambda^{*\prime}_\nu(1) = 0$.

Note that we require feasibility condition (5) since we only consider here exact constraints in (4). An alternative would have been to consider a weakened constraint of the form

$$\|\mathbb{E}_{\nu^*_m}\left[n^{-1}\sum_{i=1}^n P_i\, \tilde{\mathbf{x}}_m(T_i)\right] - t_x\| \leq \epsilon$$

for a well chosen ϵ.

Remark : (relationship with Bregman divergences) Taking the priors ν_i in a certain class of measures may lead to specific dissimilarities known as Bregman divergences (see [13]). The definition of a Bregman divergence requires a strictly convex function, which in our situation, is given by the Cramer transform Λ^*_ν of some probability measure

ν. Since we know that $\Lambda_\nu^*(1) = \Lambda_\nu^{*\prime}(1) = 0$, taking equal priors $\nu_i = \nu$ for all $i \in s$ leads to a dissimilarity that can be written

$$\mathcal{D}(w,d) = \sum_{i\in s} \Lambda_\nu^*(\pi_i w_i) = \sum_{i\in s} \left[\Lambda_\nu^*(\pi_i w_i) - \Lambda_\nu^*(1) - \Lambda_\nu^{*\prime}(1)(\pi_i w_i - 1)\right].$$

Here, we recognize the expression of the Bregman divergence between the weights $\{\pi_i w_i,\ i \in s\}$ and 1 associated to the convex function Λ_ν^*.
Another possibility is to take prior distributions ν_i lying in some suitable exponential family. More precisely, define the prior distributions as

$$d\nu_i(x) = \exp(\alpha_i x + \beta_i) d\nu(d_i x), i \in s,$$

where $\beta_i = -\Lambda_\nu(\Lambda_\nu^{*\prime}(d_i))$ and $\alpha_i = d_i \Lambda_\nu^{*\prime}(d_i)$ are taken so that ν_i is a probability measure with mean 1. We recover after calculation the following dissimilarity

$$\mathcal{D}(w,d) = \sum_{i\in s} \left[\Lambda_\nu^*(w_i) - \Lambda_\nu^*(d_i) - \Lambda_\nu^{*\prime}(d_i)(w_i - d_i)\right],$$

which is the Bregman divergence between w and d associated to Λ_ν^*.

2.3 Bayesian interpretation of calibration using MEM

The two basic components of calibration are the set of constraint equations and the choice of a dissimilarity. Here, the latter is justified by prior measures $(\nu_i)_{i\in s}$ on the weights. We now see the probabilistic interpretation of some commonly used distances.

Stochastic interpretation of some usual calibrated survey sampling estimators

1. Generalized Gaussian prior.
   For a given positive sequence $q_i, i \in s$, take $\nu_i \sim \mathcal{N}(1, \pi_i q_i)$. We get

   $$\forall t \in \mathbb{R},\ \Lambda_{\nu_i}(t) = \frac{q_i \pi_i t^2}{2} + t\ ;\ \Lambda_{\nu_i}^*(t) = \frac{(t-1)^2}{2\pi_i q_i}$$

   The calibrated weights in that cases minimize the criterion

   $$\mathcal{D}_1(w,d) = \sum_{i\in s} \frac{(\pi_i w_i - 1)^2}{q_i \pi_i}.$$

   So, we recover the χ^2 distance discussed in Section 1. This is one of the main distance used in survey sampling. The q_i's can be seen as a smoothing sequence determined by the variance of the Gaussian prior. The larger the variance, the less stress is laid on the distance between the weights and the original Horvitz-Thompson weights.

2. Exponential prior.
   We take a unique prior ν with an exponential distribution with parameter 1. That is, $\nu = \nu^{\otimes n}$. We have in that case

$$\forall t \in \mathbb{R}_+^*,\ \Lambda_\nu^*(t) = -\log t + t - 1.$$

   This corresponds to the following dissimilarity

$$\mathcal{D}_2(w, d) = \sum_{i \in s} -\log(\pi_i w_i) + \pi_i w_i.$$

   We here recognize the Bregman divergence between $(\pi_i w_i)_{i \in s}$ and 1 associated to Λ_ν^*, as explained in the previous remark. A direct calculation shows that this is also the Bregman divergence between w and d associated to Λ_ν^*. The two distances are the same in that case.

3. Poisson prior.
   If we choose for prior $\nu_i = \nu, \forall i \in s$, where ν is the Poisson distribution with parameter 1, then we obtain

$$\forall t \in \mathbb{R}_+^*,\ \Lambda_\nu^*(t) = t \log t - t + 1.$$

   So we have the following contrast

$$\mathcal{D}_3(w, d) = \sum_{i \in s} \pi_i w_i \log(\pi_i w_i) - \pi_i w_i.$$

   So we recover the Kullback information where $(\pi_i w_i)_{i \in s}$ is identified with a discrete measures on s.

MEM leads to a classical calibration problem where the solution is defined as a minimizer of a convex function subject to linear constraints. The following result gives another expression of the solution for which the computation may be easier in practical cases.

Proposition 2.2 *Assume that the optimization problem is feasible, the MEM estimator $\hat{w}$ is given by:*

$$\forall i \in s,\ \hat{w}_i = d_i \Lambda'_{\nu_i}(\hat{\lambda}^t d_i \mathbf{x}_i)$$

where $\hat{\lambda}$ minimizes over $\mathbb{R}^k$ $\lambda \mapsto \sum_{i \in s} \Lambda_{\nu_i}(\lambda^t d_i \mathbf{x}_i) - \lambda^t t_{\mathbf{x}}$.

We endow y with new weights obtaining the MEM estimator $\hat{t}_y = \sum_{i \in s} \hat{w}_i y_i$.
Note that the function $t \mapsto \Lambda'_{\nu_i}(d_i t)$ corresponds to Deville and Särndal's function F_i in [3], while taking identical priors $\nu_i = \nu$ for all $i \in s$ recovers the particular case $\Lambda'_\nu = F$ with $q_i = d_i$ according to their notations.

Calibration using maximum entropy turns into a general convex optimization program which can be easily solved. Indeed, computing the new weights $w_i, i \in s$, only involves a two step procedure. First, we find the unique $\hat{\lambda} \in \mathbb{R}^k$ such that

$$\sum_{i \in s} d_i \Lambda'_{\nu_i}(d_i \hat{\lambda}^t \mathbf{x}_i)\mathbf{x}_i - t_{\mathbf{x}} = 0. \tag{6}$$

This is achieved by optimizing a scalar convex function. Then, compute the new weights $\hat{w}_i = d_i \Lambda'_{\nu_i}(d_i \hat{\lambda}^t \mathbf{x}_i)$.

2.4 Extension to generalized calibration and instrument estimation

Computing a calibration estimator requires that (6) has a unique solution. This condition follows from the convexity of the functions $\Lambda_{\nu_i}, i \in s$. Aiming to provide wider possibilities of estimation, the method of generalized calibration (GC) considered in [21] consists in replacing the functions $\lambda \mapsto \Lambda'_{\nu_i}(d_i \lambda^t \mathbf{x}_i)$ by more general functions $f_i : \mathbb{R}^k \to \mathbb{R}$. Assume that the equation

$$F(\lambda) = \sum_{i \in s} d_i f_i(\lambda)\mathbf{x}_i = t_{\mathbf{x}} \tag{7}$$

has a unique solution $\hat{\lambda}$. Assume also that the f_i are continuously differentiable at 0, and are such that $f_i(0) = 1$ so that $F(0) = \hat{t}_{\mathbf{x}}^{HT}$. Then, take as the solution to the generalized calibration procedure, the weights:

$$\forall i \in s,\ \hat{w}_i = d_i f_i(\hat{\lambda}).$$

Calibration is of course a particular example of generalized calibration where we set $f_i :\ \lambda \mapsto \Lambda'_{\nu_i}(d_i \lambda^t \mathbf{x}_i)$ to recover a calibration problem seen in Section 2.2. An interesting example of GC is to take affine functions $\lambda \mapsto 1 + \mathbf{z}_i^t \lambda$, where $(\mathbf{z}_i)_{i \in s}$ is a sequence of vectors of $\mathbb{R}^k$. The $\mathbf{z}_i$'s are called instruments (see [21]). If the matrix $X_n := N^{-1} \sum_{i \in s} d_i \mathbf{z}_i \mathbf{x}_i^t$ is invertible, the resulting estimator $\hat{t}_y$, referred to as the instrument estimator obtained with the instruments $\mathbf{z}_i$, is given by:

$$\hat{t}_y = \hat{t}_y^{HT} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}}^{HT})^t X_n^{-1} \textstyle\sum_{i \in s} d_i \mathbf{z}_i y_i. \tag{8}$$

Remark : (dimension reduction) As remarked in [24] in the case $\mathbf{z}_i = \mathbf{x}_i$, the estimator of the population total is identical to the one obtained with one-dimensional auxiliary variable $\hat{B}^t \mathbf{x}$, where $\hat{B}$ is estimated by least squares. More generally, reducing the dimension of the auxiliary variable to one is always possible when using instruments. The new auxiliary variable and instruments are linear transformations $\hat{B}^t \mathbf{x}$ and $\hat{B}^t \mathbf{z}$ of the original variables $\mathbf{x}$ and $\mathbf{z}$, where

$$\hat{B} = \left[\textstyle\sum_{i \in s} d_i \mathbf{z}_i \mathbf{x}_i^t\right]^{-1} \textstyle\sum_{i \in s} d_i y_i \mathbf{z}_i.$$

This points out the relationship between calibration and linear regression discussed in [3]. The method implicitly aims at constructing a variable $\tilde{y} = y - \hat{B}^t\mathbf{x}$ with a lower variance than that of y (at least for sufficiently large samples), and for which the population total is known up to t_y. The calibrated estimator $\hat{t}_y$ can be written

$$\hat{t}_y = \sum_{i\in s} d_i\tilde{y}_i + \hat{B}^t t_{\mathbf{x}},$$

that is, $\hat{t}_y$ is the Horvitz-Thompson estimator (up to a known additive constant, here $\hat{B}^t t_{\mathbf{x}}$) of the variable $\tilde{y}$.

Instrument estimators play a crucial role when studying the asymptotic properties of generalized calibration estimation. A classical asymptotic framework in calibration is to consider that n and N simultaneously go to infinity while the Horvitz-Thompson estimators of the mean converge at a rate of convergence of $\sqrt{n}$, as described in [3] and [24] for instance. Hence, we assume that

$$N^{-1}\|\hat{t}_{\mathbf{x}}^{HT} - t_{\mathbf{x}}\| = O(n^{-1/2}) \quad \text{and} \quad N^{-1}(\hat{t}_y^{HT} - t_y) = O(n^{-1/2}),$$

further assumptions on our asymptotic framework are made in Section 3.

Definition We say that two GC estimators $\hat{t}_y$ and $\tilde{t}_y$ are asymptotically equivalent if

$$N^{-1}(\hat{t}_y - \tilde{t}_y) = o(n^{-1/2}).$$

Proposition 2.3 *Let $\hat{t}_y$ and $\tilde{t}_y$ be the GC estimators obtained respectively with the functions $f_i, i \in s$ and $g_i, i \in s$. If for all $i \in s$, $\nabla f_i(0) = \nabla g_i(0) = \mathbf{z}_i$, and if the matrix $X_n := N^{-1}\sum_{i\in s} d_i\mathbf{z}_i\mathbf{x}_i^t$ converges toward an invertible matrix X, then $\hat{t}_y$ and $\tilde{t}_y$ are asymptotically equivalent. In particular, two MEM estimators are asymptotically equivalent as soon as their prior distributions have the same respective variances.*

This proposition is a consequence of Result 3 in [3]. It states that first order asymptotic behavior of GC estimators in only determined by the gradient vectors $\mathbf{z}_i = \nabla f_i(0), i \in s$, where the f_i's are the functions used in (7). As a result, all GC estimator can be shown to have an asymptotically equivalent instrument estimator.
The frame of calibration and MEM estimation corresponds to instruments of the form $\mathbf{z}_i = q_i\mathbf{x}_i$. This particular case is discussed in [3] where the authors prove that a calibration estimator can always be approximated by a regression estimator under regularity conditions. A different proof of this result is also given in Theorem 2.7.1 in [8]. Thus, a MEM estimator $\hat{t}_y$ obtained with prior distributions $\nu_i, i \in s$ with respective variances $\pi_i q_i$ satisfies

$$\hat{t}_y = \hat{t}_y^{HT} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}}^{HT})^t\hat{B} + o(Nn^{-1/2})$$

where $\hat{B} = \left[\sum_{i\in s} d_i q_i \mathbf{x}_i \mathbf{x}_i^t\right]^{-1} \sum_{i\in s} d_i q_i \mathbf{x}_i y_i$. The negligible term is null for Gaussian priors, leading to a χ^2 dissimilarity in the frame of calibration (see Example 1 in Section 2.3) and to the instrument estimator built with instruments $\mathbf{z}_i = q_i \mathbf{x}_i$. This choice of instruments, and in particular the case $q_i = 1$ for all $i \in s$, is often used in practice since it provides a consistent estimate which can be easily computed.

3 Efficiency of calibration estimator with MEM method

The accuracy of the estimator heavily relies on the linear correlation between y and the auxiliary variable $\mathbf{x}$. If a relationship other than linear prevails, $\mathbf{x}$ may not be an efficient choice of calibration variable. When complete information is available, *model calibration* proposed by Wu and Sitter aims to generalize the calibration procedure by considering an auxiliary variable of the form $u(\mathbf{x})$ for $u : \mathbb{R}^k \to \mathbb{R}^d$ a given function. Their objective is to increase the linear correlation between the variables, leading to a better efficiency of the estimation. In [26], Wu and Sitter assume that the optimal calibration function u belongs to a known parametric class of functions for which the true value of the parameter is estimated from the data. Montanari and Ranalli [18] discuss the estimation of the optimal choice for the function u in a non parametric model.
With complete information, the choice of the calibration function u and the instruments are the two main aspects of the estimation of t_y in an asymptotic framework. In this section, we first study the influence of the instruments $\mathbf{z}$ when the calibration function u is fixed. Then, we discuss ways of improving the estimation by allowing both the instruments and the calibration variable to vary with the observations.

3.1 Asymptotic efficiency

We consider the usual asymptotic framework in survey sampling where there is a sequence of sampling designs and finite populations, indexed by r. The population size and the sample size, denoted respectively by N_r and n_r, both grow to infinity as $r \to +\infty$. The asymptotic framework is to be understood in the sense that $r \to +\infty$, but, in the following, the index r will be suppressed to simplify notation. We consider the population measurements $\{(\mathbf{x}_i, y_i), i = 1, ..., N\}$ as independent realizations of a random variable (X, Y) from a superpopulation model ξ.
For $u : \mathbb{R}^k \to \mathbb{R}^d$ a given function, we note $u_i = u(\mathbf{x}_i)$ and

$$t_u = \sum_{i\in U} u_i, \ \hat{t}_{u\pi} = \sum_{i\in s} d_i u_i.$$

In the sequel, we assume that $\mathbb{E}(|Y|^3) < \infty$ and $\mathbb{E}(\|u(X)^3\|) < \infty$, where $\mathbb{E}$ denotes the expectation with respect to the distribution of (X, Y). In a general setting where the

auxiliary variable takes the form $u(\mathbf{x})$, instrument estimators have the following expression

$$\hat{t}_y = \hat{t}_y(u) = \hat{t}_y^{HT} + (t_u - \hat{t}_{u\pi})^t \hat{B}_u,$$

where $\hat{B}_u = \left[\sum_{i\in s} d_i \mathbf{z}_i u_i^t\right]^{-1} \sum_{i\in s} d_i y_i \mathbf{z}_i$ is assumed to be well defined. Furthermore, define the joint inclusion probabilities $\pi_{ij} = \sum_{s:\ i,j\in s} p(s)$ and set $\Delta_{ij} := \pi_{ij} d_i d_j - 1$.
The nonlinearity of $\hat{t}_y$ makes it difficult to evaluate its quadratic risk. Following [6], easier is to consider its linear asymptotic expansion $\hat{t}_{y,\text{lin}}(u) := \hat{t}_y^{HT} + (t_u - \hat{t}_{u\pi})B_u$ where B_u is a vector, independent from the sample s, such that $\|\hat{B}_u - B_u\| = o(1)$. The linear expansion $\hat{t}_{y,\text{lin}}(u)$ is design unbiased and is asymptotically equivalent to $\hat{t}_y$. As proved in [17], the variance of $\hat{t}_{y,\text{lin}}(u)$, which is given by

$$\text{var}_p(\hat{t}_{y,\text{lin}}(u)) = \sum_{i,j\in U} \Delta_{ij}\ (y_i - B_u^t u_i)(y_j - B_u^t u_j),$$

depends on the instruments only through the value of B_u and is minimal for $B_u = B_u^*$ given by

$$B_u^* = \left[\text{var}_p(\hat{t}_{u\pi})\right]^{-1} \text{cov}_p(\hat{t}_{u\pi}, \hat{t}_y^{HT}) = \left[\sum_{i,j\in U} \Delta_{ij}\ u_i u_j\right]^{-1} \sum_{i,j\in U} \Delta_{ij}\ u_j y_i,$$

where var_p and cov_p denote respectively the variance and covariance under the sampling design p. We make the following assumptions

A1: The sampling design $p(.)$ is weakly dependent of ξ in the sense that for any sequence $\{a(x_i, y_i)\}_{i\in U} = \{a_i\}_{i\in U}$ such that $N^{-1}\sum_{i\in U} |a_i|^3 = O(1)$,

$$\mathbb{E}(\textstyle\sum_{i,j\in U} \Delta_{ij}\ a_i a_j) = \sum_{i,j\in U} \Delta_{ij}\ \mathbb{E}(a_i a_j) + o(N^2 n^{-1}).$$

A2: There exists $0 \leq \pi < 1$, such that $\limsup_{r\to\infty} nN^{-1} = \pi$.

A3: $\lim nN^{-2}\sum_{i\in U} \Delta_{ii} = -\lim nN^{-2}\sum_{i\in U}\sum_{j\neq i} \Delta_{ij} = 1 - \pi$.

The first assumption conveys the information that no design weight is disproportionately large compared to the others. It holds for instance if p and ξ are independent and if $\sum_{i\in U} \Delta_{ii}^2 = o(N^4 n^{-2})$ and $\sum_{i\in U}\sum_{j\neq i} \Delta_{ij}^2 = o(N^3 n^{-2})$.
Assumption 2 is not restrictive, it simply states that the number of unobserved data never becomes negligible compared to the population size. This is a classical assumption in survey sampling, see for instance [18].
The last assumption is essentially made to ensure the existence of efficient estimators as shown further. It is fulfilled for the uniform sampling design, that is when every sample $s \subset U$ has the same probability of being observed, provided that the sample size and the population size remain of the same order. In that case, the Horvitz-Thompson

weights are $\pi_i = n/N$, $\pi_{ij} = n(n-1)/N(N-1), \forall i \neq j$, yielding $\Delta_{ii} = N/n - 1$ and $\Delta_{ij} = -(N-n)/n(N-1)$.

Lemma: *Suppose that Assumptions 1 and 2 hold. Then,*

$$nN^{-2}\ \mathbb{E}(\mathbb{E}_p(t_y - \hat{t}_y(u))^2) \geq (1-\pi)\text{var}\left(Y - B_u^t u(X)\right) + o(1),$$

with equality if, and only if, Assumption 3 also holds.

The proof is a direct consequence of Lemma 4.1 in the Appendix.
This result provides a natural criterion of asymptotic efficiency. Indeed, finding instruments for which the right term of the inequality is minimal appears as a natural objective, whether the sampling design satisfies Assumption 3 or not. So, the variance lower bound is defined as the minimum $V^*(u)$ of $(1-\pi)\text{var}(Y - B^t u(X))$ as B ranges over $\mathbb{R}^d$. We say that an estimator $\hat{t}_y(u)$ is asymptotically efficient if the expectation of its design quadratic risk converges towards $V^*(u)$. This is an analog of optimal calibration in [17], where in our framework, optimality requires that

$$\lim \hat{B}_u = [\text{var}(u(X))]^{-1}\,\text{cov}(Y, u(X)), \tag{9}$$

assuming that $\text{var}(u(X))$ is invertible. In this case, we get

$$V^*(u) = (1-\pi)\text{var}\left(Y - \text{cov}(Y, u(X))^t\,[\text{var}(u(X))]^{-1}\,u(X)\right). \tag{10}$$

Note that this lower bound can not be reached if Assumption 3 is not fulfilled.

Estevao and Särndal [6] propose the instruments $\mathbf{z}_i^* = \sum_{j\in U} \Delta_{ij}\ u_j$ as a natural choice by identification, noticing that the optimal value B_u^* for a fixed N verifies

$$B_u^* = \left[\sum_{i,j\in U} \Delta_{ij}\ u_i u_j\right]^{-1} \sum_{i,j\in U} \Delta_{ij}\ u_j y_i = \left[\sum_{i\in U} u_i \mathbf{z}_i^*\right]^{-1} \sum_{i\in U} y_i \mathbf{z}_i^*.$$

In our framework, these instruments satisfy condition (9), as a consequence of Lemma 4.1. However, a noticeable drawback is that the calculation of each instrument $\mathbf{z}_i^*$ involves the whole population $(\mathbf{x}_i)_{i\in U}$, yielding a computationally expensive estimate.

The simple choice $\mathbf{z}_i = u_i, i \in s$ provides a good alternative. As shown in Proposition 2.3, the resulting estimator is asymptotically equivalent to MEM estimators built using prior distributions ν_i with variance π_i. The consistency of the Horvitz-Thomspon estimates leads to

$$\hat{B}_u = \left[\sum_{i\in s} d_i u_i u_i^t\right]^{-1} \sum_{i\in s} d_i y_i u_i \longrightarrow [\mathbb{E}(u(X)u(X)^t)]^{-1}\,\mathbb{E}(Y u(X)).$$

Although condition (9) for optimality is not fulfilled for most functions u, the problem can easily be solved by adding the constant variable 1 in the calibration constraint. We

then consider the MEM estimator $\hat{t}_y(v)$ where $v = (1, u)^t : \mathbb{R}^k \to \mathbb{R}^{d+1}$, the calibrated weights now satisfy the constraints

$$\sum_{i \in s} w_i u_i = t_u, \ \sum_{i \in s} w_i = N.$$

Here, the matrix $\mathrm{var}(v(X))$ is not invertible although it is simple algebra to see that $V^*(v) = V^*(u)$. So, the auxiliary variable is modified but the asymptotic lower bound is unchanged. As a result of the dimension reduction property of calibration, adding the constant in the calibration constraint reduces to use the instruments $\mathbf{z}_i = u_i - \hat{t}_{u\pi}$ up the a negligible term. A direct calculation shows that these instruments now satisfy condition (9).

3.2 Approximate Maximum Entropy on the Mean

We now turn to the question of the optimal auxiliary variable. By minimizing the asymptotic variance lower bound $V^*(u)$ with respect to u, the conditional expectation $\Phi(\mathbf{x}_i) = \mathbb{E}(Y|X = \mathbf{x}_i)$ appears as the optimal choice since $\Phi(.)$ is the unique (up to affine transformations) minimizer of the functional $u \mapsto V^*(u)$ in Equation (10) (u is taken real-valued without loss of generality). This confirms the result stated in Theorem 1 in [24], where Wu proves the variable $\Phi(\mathbf{x}_i), i \in U$ to be optimal. In that case, the asymptotic lower bound is given by:

$$V^* = (1 - \pi)\mathbb{E}(Y - \mathbb{E}(Y|X))^2.$$

Note that, since this optimal choice depends on the unknown distribution of (X, Y), this result does not provide a tractable estimator. A natural solution is to replace Φ by an estimate Φ_m, and then plug it into the calibration constraint. Under regularity conditions that will be made precise later, we show that this approach enables to compute an asymptotically optimal estimator of t_y, in the sense that its asymptotic expected design variance is equal to the lower bound V^* defined above.

In this section, $\hat{t}_y(u)$ will denote a MEM estimator of t_y obtained with auxiliary variable $(u(\mathbf{x}), 1)^t$ and prior distributions ν_i with variance π_i. We recall that for any measurable function u, $\hat{t}_y(u)$ is $\sqrt{n}$-consistent with asymptotic variance $V^*(u)$.
Assume that we observe approximations $(\Phi_m)_{m \in \mathbb{N}}$ of Φ, we define the AMEM estimator as $\hat{t}_y(\Phi_m)$, i.e., the MEM estimator calibrated with the variable $(\Phi_m(\mathbf{x}), 1)^t$.

Theorem 3.1 *Suppose that Assumptions 1 to 3 hold. Let $(\Phi_m)_{m \in \mathbb{N}}$ be a sequence of functions independent with ξ and such that*

$$\mathbb{E}(\Phi(X) - \Phi_m(X))^2 = O(\varphi_m^{-1}) \text{ with } \lim_{m \to \infty} \varphi_m = +\infty.$$

Then, the AMEM estimator $\hat{t}_y(\Phi_m)$ is asymptotically optimal among all GC estimators in the sense that $nN^{-2}\mathbb{E}(\mathbb{E}_p(t_y - \hat{t}_y(\Phi_m))^2)$ converges toward V^* as $(r, m) \to \infty$.

When applied to this context, approximate maximum entropy on the mean increases the efficiency of calibration estimators when an additional information is available, namely, an external estimate of the conditional expectation function Φ is observed. Nevertheless, in our model, it is possible to obtain similar properties under weakened conditions.

Corollary 3.2 *Suppose that Assumptions 1 to 3 hold. Let $(\Phi_m)_{m\in\mathbb{N}}$ be a sequence of functions satisfying*

$$\begin{aligned} &i) \quad nN^{-2}\mathbb{E}(\mathbb{E}_p(\hat{t}_{\Phi\pi} - t_\Phi - (\hat{t}_{\Phi_m\pi} - t_{\Phi_m}))^2) \longrightarrow 0 \\ &ii) \quad \hat{B}_{\Phi_m} = \left[\sum_{i\in s} d_i \Phi_m(\mathbf{x}_i)^2\right]^{-1} \sum_{i\in s} d_i y_i \Phi_m(\mathbf{x}_i) \longrightarrow 1, \end{aligned}$$

as $(r, m) \to \infty$. Then, the estimator $\hat{t}_y(\Phi_m)$ is asymptotically optimal.

This corollary does not rule out that the functions Φ_m are estimated using the data, which was not the case in Theorem 3.1. Hence, it becomes possible to compute an asymptotically efficient estimator of t_y with a single sample. A data driven estimator Φ_n provides as well an asymptotically efficient estimator of t_y, as soon as the two conditions of Corollary 3.2 are fulfilled.
Remark that although this natural way to extend calibration to non parametric procedures can be claimed to be asymptotically optimal, the resulting estimator may still be highly unstable for relatively small samples or under irregular sampling designs.

Many non parametric methods could be used in the frame of calibration, see for instance [18]. Here, we study an approach where the conditional expectation Φ is estimated by projection onto finite dimensional subspaces. Let $\phi = (1, \phi_1, \phi_2, ...)$ be a sequence of linearly independent functions, total in the space of square integrable functions. This sequence is referred to as a projection basis. Typically, it can be polynomials if X takes values in a compact subset of $\mathbb{R}^k$ or wavelets but other forms may be chosen, depending on the situation.
Denote by $\phi^m = (1, \phi_1, ..., \phi_m)$ the vector of the first $m + 1$ components of ϕ, we build a projection estimator Φ_m of Φ by considering a suitable linear combination $\hat{B}_m^t \phi^m$ of the functions, the vector $\hat{B}_m$ being generally obtained by least squares on the variables y and $\phi^m(\mathbf{x})$. In the context of calibration, it is natural to consider design based estimates Φ_m. As a result of a reciprocal effect of the dimension reduction property, taking $\Phi_m = \hat{B}_m^t \phi^m$ with

$$\hat{B}_m = \hat{B}_{\phi^m} = \left[\sum_{i\in s} d_i \phi_i^m \phi_i^{mt}\right]^{-1} \sum_{i\in s} d_i y_i \phi_i^m,$$

leads to the estimator calibrated with the vector variable $\phi^m(\mathbf{x})$ up to a negligible term. Indeed, the auxiliary variable $\Phi_m(\mathbf{x}) = \hat{B}_m^t \phi^m(\mathbf{x})$ is obtained as the one dimensional equivalent of $\phi^m(\mathbf{x})$ discussed in Section 2.4. So, we point out that calibration is here extended to non parametric procedures by simply increasing the number of auxiliary variables. The estimator calibrated with a χ^2 dissimilarity can be expressed as

$$\hat{t}_y(\Phi_m) = \hat{t}_y^{HT} + (t_{\Phi_m} - \hat{t}_{\Phi_m\pi}) = \hat{t}_y^{HT} + (t_{\phi^m} - \hat{t}_{\phi^m\pi})^t \hat{B}_{\phi^m},$$

which illustrates the equivalence between auxiliary variables $\Phi_m(\mathbf{x})$ and $\phi^m(\mathbf{x})$. With the notations of Corollary 3.2, the vector $\hat{B}_{\Phi_m}$ corresponding to $\Phi_m(\mathbf{x})$ is equal to 1 for all m and therefore satisfies the condition $ii)$ in the corollary. The condition $i)$ can also be fulfilled with this method, although, a proper number of constraints must be chosen. If m is fixed, we know that $\hat{t}_y(\Phi_m)$ converges towards t_y with an asymptotic variance $V^*(\phi^m)$. The convergence of $(V^*(\phi^m))_{m\in\mathbb{N}}$ towards V^* warrants the existence of a sequence of integers $(m(n))_{n\in\mathbb{N}}$ such that $\Phi_{m(n)}$ undergoes the first condition of Corollary 3.2. Note however that finding such a sequence is a difficult task and belongs to the class of model selection issues.

Asymptotic results of non parametric methods are to be taken with care since it may require a large number of observations before the method becomes really effective. Here we assumed strong regularity conditions on the sampling design, allowing good consistency of the non parametric estimation with relatively small samples. AMEM procedures in survey sampling have the advantage to enable to implement non parametric procedures while remaining in a Bayesian framework.

4 Appendix

4.1 Technical lemma

Lemma 4.1 *Under Assumptions 1 and 2, for any sequence* $\{a(x_i,y_i)\}_{i\in U}=\{a_i\}_{i\in U}$ *such that* $N^{-1}\sum_{i\in U}|a_i|^3=O(1)$,

$$nN^{-2}\mathbb{E}(\textstyle\sum_{i,j\in U}(\Delta_{ij}\ a_ia_j))\geq(1-\pi)\mathrm{var}(a(X,Y))+o(1)$$

with equality if and only if Assumption 3 also holds. Moreover, under Assumptions 1 to 3, the quantity $nN^{-2}\sum_{i,j\in U}\Delta_{ij}\ a_ib_j$ *converges in probability towards* $\mathrm{cov}(a(X,Y),b(X,Y))$ *for all sequence* $\{b_i\}_{i\in U}$ *satisfying the same conditions as* $\{a_i\}_{i\in U}$.

Proof of Lemma 4.1:
For such a sequence $a=\{a_i\}_{i\in U}$, Assumptions 1 and 2 yield:

$$\begin{aligned} & nN^{-2}\textstyle\sum_{i,j\in U}\Delta_{ij}\ a_ia_j = nN^{-2}\sum_{i\in U}\Delta_{ii}\ a_i^2+nN^{-2}\sum_{i\neq j}\Delta_{ij}\ a_ia_j \\ = \ & \left(nN^{-2}\textstyle\sum_{i\in U}\Delta_{ii}\right)\mathbb{E}(a(X,Y)^2)+\left(nN^{-2}\sum_{i\neq j}\Delta_{ij}\right)\mathbb{E}(a(X,Y))^2+o(1)\end{aligned}$$

Denote by $\mathcal{P}_n(U)$ the set of all subsamples s of U with n elements. By Jensen inequality,

$$\textstyle\sum_{i,j\in U}\Delta_{ij}=\sum_{s\in\mathcal{P}_n(U)}\left(\sum_{i\in s}d_i\right)^2p(s)-N^2\geq\left[\sum_{s\in\mathcal{P}_n(U)}\left(\sum_{i\in s}d_i\right)p(s)\right]^2-N^2\geq 0$$

which implies that $\sum_{i\neq j}\Delta_{ij}\geq-\sum_{i\in U}\Delta_{ii}$. Thus:

$$nN^{-2}\textstyle\sum_{i,j\in U}\Delta_{ij}\ f_if_j\geq\left(nN^{-2}\sum_{i\in U}\Delta_{ii}\right)\mathrm{var}(f(X,Y))+o(1).$$

Since $\sum_{i\in U}\pi_i = n$, we know that $nN^{-2}\sum_{i\in U}\Delta_{ii} \geq 1 - nN^{-1}$ by convexity of $x \mapsto 1/x$ on $\mathbb{R}_+^*$. Hence

$$nN^{-2}\sum_{i,j\in U}\Delta_{ij}\ a_i a_j \geq (1-\pi)\mathrm{var}(a(X,Y)) + o(1)$$

without equality for all sequence a if Assumption 3 is not true. The end of the lemma follows directly by using the same guideline applied to a and b. In particular, it holds when $a = b$.

4.2 Proofs

Proof of Theorem 2.1:
For all $w \in \mathbb{R}^n$, let $f_w : \mathbb{R}^n \to \mathbb{R}_+$ be the unique minimizer of the functional $f \mapsto K(f\nu, \nu)$ on the set $\mathcal{F}_w = \left\{f : \int_{\mathbb{R}^n}(\tau - \pi w)f(\tau)d\nu(\tau) = 0\right\}$. We have:

$$f_w = \underset{f\in\mathcal{F}_w}{\mathrm{argmin}}\ \int_{\mathbb{R}^n} f(\log(f) - 1)d\nu.$$

We calculate the Lagrangian $\mathcal{L}(\lambda, f)$ associated to the problem:

$$\mathcal{L}(\lambda, f) = \int_{\mathbb{R}^n}[f(\tau)\log(f(\tau)) - f(\tau)]d\nu(\tau) - \lambda^t\int_{\mathbb{R}^n}(\tau - \pi w)f(\tau)d\nu(\tau)$$

where $\lambda \in \mathbb{R}^n$ is the Lagrange multiplier. The first order conditions are:

$$\forall\tau \in \mathbb{R}^n,\ \log(f(\tau)) = \lambda^t(\tau - \pi w).$$

Hence, $\forall\tau,\ f_w(\tau) = e^{\lambda_w^t(\tau-\pi w)}$ where λ_w verifies:

$$\int_{\mathbb{R}^n}(\tau - \pi w)e^{\lambda^t(\tau-\pi w)}d\nu(\tau) = 0 \Longleftrightarrow \lambda_w = \underset{\lambda\in\mathbb{R}^n}{\mathrm{argmin}}\int_{\mathbb{R}^n} e^{\lambda^t(\tau-\pi w)}d\nu(\tau)$$

Let $S = \left\{(w_i)_{i\in s} :\ N^{-1}\sum_{i\in s}\mathbf{x}_i w_i = t_{\mathbf{x}}\right\}$, we notice that

$$\begin{aligned}
\hat{w} = \mathbb{E}_{\nu^*}(W) &= \underset{w\in S}{\mathrm{argmin}}\left\{\min_{f\in\mathcal{F}_w}\ \int_{\mathbb{R}^n} f(\log(f) - 1)d\nu\right\} \\
&= \underset{w\in S}{\mathrm{argmin}}\left\{\int_{\mathbb{R}^n} f_w(\log(f_w) - 1)d\nu\right\} \\
&= \underset{w\in S}{\mathrm{argmin}}\left\{\lambda_w^t\int_{\mathbb{R}^n}(\tau - \pi w)e^{\lambda_w^t(\tau-\pi w)}d\nu(\tau) - \int_{\mathbb{R}^n} e^{\lambda_w^t(\tau-\pi w)}d\nu(\tau)\right\} \\
&= \underset{w\in S}{\mathrm{argmin}}\left\{-\min_{\lambda\in\mathbb{R}^n}\ e^{-\lambda^t\pi w}\int_{\mathbb{R}^n} e^{\lambda^t\tau}d\nu(\tau)\right\}.
\end{aligned}$$

by definition of λ_w. Recall that $\nu = \otimes_{i\in s}\nu_i$. Since the function $t \mapsto -\log t$ is decreasing, we have that

$$\min_{\lambda\in\mathbb{R}^n}\left\{e^{-\lambda^t\pi w}\int_{\mathbb{R}^n} e^{\lambda^t\tau}d\nu(\tau)\right\} = \exp -\sup_{\lambda\in\mathbb{R}^n}\left\{\sum_{i\in s}[\lambda_i\pi_i w_i - \log\int_{\mathbb{R}} e^{\lambda_i\tau_i}d\nu_i(\tau_i)]\right\}$$

The supremum being taken for $\lambda \in \mathbb{R}^n$, we see that

$$\sup_{\lambda\in\mathbb{R}^n} \left\{\sum_{i\in s}[\lambda_i\pi_i w_i - \log\int_{\mathbb{R}} e^{\lambda_i\tau_i}d\nu_i(\tau_i)]\right\} = \sum_{i\in s}\sup_{\lambda_i\in\mathbb{R}} \left\{\lambda_i\pi_i w_i - \log\int_{\mathbb{R}} e^{\lambda_i\tau_i}d\nu_i(\tau_i)\right\}$$

Finally we obtain:

$$\hat{w} = \operatorname*{argmin}_{w\in S} -\exp\left(-\sum_{i\in s}\Lambda^*_{\nu_i}(\pi_i w_i)\right) = \operatorname*{argmin}_{w\in S} \sum_{i\in s}\Lambda^*_{\nu_i}(\pi_i w_i).$$

Proof of Proposition 2.2:
This is a classic convex optimization problem. Let $\mathcal{L}$ be the Lagrangian associated to the problem:

$$\mathcal{L}(\lambda, w) = \sum_{i\in s}\Lambda^*_{\nu_i}(w_i\pi_i) - \lambda^t\left(\sum_{i\in s} w_i\mathbf{x}_i - Nt_{\mathbf{x}}\right)$$

where $\lambda \in \mathbb{R}^k$ is the Lagrange multiplier. The solutions to the first order conditions satisfy for all $i \in s$,

$$w_i = d_i(\Lambda^{*\,\prime}_{\nu_i})^{-1}(\lambda^t d_i\mathbf{x}_i),$$

where we recall that the functions $\Lambda^*_{\nu_i}$ are assumed to be strictly convex, so that $(\Lambda^{*\,\prime}_{\nu_i})^{-1}$ exists for all i, and is equal to Λ'_{ν_i}. Now it suffices to apply the solutions of the first order conditions to the constraint to obtain an expression of the solution $\hat{\lambda}$:

$$N^{-1}\sum_{i\in s} d_i\Lambda'_{\nu_i}(\hat{\lambda}^t d_i\mathbf{x}_i)\mathbf{x}_i - t_{\mathbf{x}} = 0 \Longleftrightarrow \hat{\lambda} = \operatorname*{argmin}_{\lambda\in\mathbb{R}^k}\sum_{i\in s}\Lambda_{\nu_i}(\lambda^t d_i\mathbf{x}_i) - \lambda^t t_{\mathbf{x}}.$$

The equivalence is justified by the fact that Λ_{ν_i} is strictly convex, and therefore, so is $\lambda \mapsto \sum_{i\in s}\Lambda_{\nu_i}(\lambda^t d_i\mathbf{x}_i) - \lambda^t t_{\mathbf{x}}$. For that reason, $\hat{\lambda}$ is uniquely defined. We finally obtain an expression of the calibrated weights

$$\forall i \in s,\ \hat{w}_i = d_i\Lambda'_{\nu_i}(\hat{\lambda}^t d_i\mathbf{x}_i).$$

Proof of Proposition 2.3:
Let $F : \lambda \mapsto N^{-1}\sum_{i\in s} d_i f_i(\lambda)\mathbf{x}_i$, and $G : \lambda \mapsto N^{-1}\sum_{i\in s} d_i g_i(\lambda)\mathbf{x}_i$. We call respectively $\hat{\lambda}$ and $\tilde{\lambda}$ the solutions to $F(\lambda) = t_{\mathbf{x}}$ and $G(\lambda) = t_{\mathbf{x}}$. We have

$$F(\hat{\lambda}) = F(0) + X_n\hat{\lambda} + o(\|\hat{\lambda}\|)$$

and then $(t_{\mathbf{x}} - \hat{t}^{HT}_{\mathbf{x}}) = X_n\hat{\lambda} + o(\|\hat{\lambda}\|)$. By assumption, X_n is invertible for large values of n since it converges towards an invertible matrix X. Thus, whenever $\hat{t}^{HT}_{\mathbf{x}}$ is close enough to $t_{\mathbf{x}}$, there exists λ_0 in a neighborhood of 0 such that $F(\lambda_0) = t_{\mathbf{x}}$. By uniqueness of the solution, we conclude that $\lambda_0 = \hat{\lambda}$. Hence, for large values of n,

$$\hat{\lambda} = X_n^{-1}(t_{\mathbf{x}} - \hat{t}^{HT}_{\mathbf{x}}) + o(n^{-1/2}).$$

A similar reasoning for $\tilde{\lambda}$ yields $\|\tilde{\lambda} - \hat{\lambda}\| = o(n^{-1/2})$. Thus, $\hat{\lambda}$ and $\tilde{\lambda}$ converge toward 0 and by Taylor formula:

$$f_i(\hat{\lambda}) = 1 + \mathbf{z}_i^t\hat{\lambda} + o(n^{-1/2}) = 1 + \mathbf{z}_i^t\tilde{\lambda} + o(n^{-1/2}) = g_i(\tilde{\lambda}) + o(n^{-1/2}).$$

It follows that $\hat{t}_y$ and $\tilde{t}_y$ are asymptotically equivalent.
We know that MEM estimation reduces to taking $f_i(.) = \Lambda'_{\nu_i}(d_i\mathbf{x}_i^t.)$ in a GC procedure. Hence, in that case, $\nabla f_i(0) = d_i\Lambda''_{\nu_i}(0)\mathbf{x}_i$. Since the variance of a probability measure ν_i is given by $\Lambda''_{\nu_i}(0)$, two MEM estimators with prior distributions having the same respective variances are asymptotically equivalent. Furthermore, a Gaussian prior $\nu_i \sim \mathcal{N}(1, q_i\pi_i)$ has a log-Laplace transform $\Lambda_{\nu_i} : t \mapsto \pi_i q_i t^2/2 + t$ which corresponds to $f_i(\lambda) = \Lambda'_{\nu_i}(d_i\mathbf{x}_i^t\lambda) = 1 + q_i\mathbf{x}_i^t\lambda$. The resulting MEM estimator is thus the instrument estimator obtained with instruments $\mathbf{z}_i = q_i\mathbf{x}_i, i \in s$.

Proof of Theorem 3.1:
We decompose the AMEM estimator as follow

$$\hat{t}_y(\Phi_m) = \hat{t}_y^{HT} + (t_\Phi - \hat{t}_{\Phi\pi}) + (\hat{t}_{\Phi\pi} - t_\Phi - (\hat{t}_{\Phi_m\pi} - t_{\Phi_m})) + (\hat{B}_{\Phi_m} - 1)(t_{\Phi_m} - \hat{t}_{\Phi_m\pi}).$$

We have by assumption

$$nN^{-2}\mathbb{E}(\mathbb{E}_p(\hat{t}_{\Phi\pi} - t_\Phi - (\hat{t}_{\Phi_m\pi} - t_{\Phi_m}))^2) = O(\varphi_m^{-1}) \quad \text{and} \quad (\hat{B}_{\Phi_m} - 1) = O(\varphi_m^{-1/2})$$

uniformly for all m (see the proof of Lemma 1 in [16]). Therefore, the terms $(\hat{t}_{\Phi\pi} - t_\Phi - (\hat{t}_{\Phi_m\pi} - t_{\Phi_m}))$ and $(\hat{B}_{\Phi_m} - 1)(t_{\Phi_m} - \hat{t}_{\Phi_m\pi})$ are asymptotically negligible in comparison to $(t_\Phi - \hat{t}_{\Phi\pi})$ as $m \to \infty$. We conclude using Lemma 4.1.

Proof of Corollary 3.2:
Follows directly from the proof of Theorem 3.1.

We are very grateful to an anonymous referee for its helpful comments on survey sampling.

References

[1] J. M. Borwein, A. S. Lewis, and D. Noll. Maximum entropy reconstruction using derivative information. I. Fisher information and convex duality. *Math. Oper. Res.*, 21(2):442–468, 1996.

[2] M. Carrasco, J-P. Florens, and E. Renault. *Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization*, volume 6. North Holland, 2006.

[3] J. C. Deville and C. E. Särndal. Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, 87(418):376–382, 1992.

[4] Jean-Claude Deville. Estimation linéaire et redressement sur informations auxiliaires d'enquête par sondages. *Economica*, 1988.

[5] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[6] V. M. Estevao and C. E. Särndal. Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2):127–147, 2006.

[7] A. K. Fermín, J. M. Loubes, and C. Ludeña. Bayesian methods for a particular inverse problem: seismic tomography. *Int. J. Tomogr. Stat.*, 4(W06):1–19, 2006.

[8] W. A. Fuller. *Sampling Statistics*. Wiley Series in Survey Methodology. John Wiley & Sons Inc, NJ, 2009.

[9] F. Gamboa. New Bayesian methods for ill posed problems. *Statist. Decisions*, 17(4):315–337, 1999.

[10] F. Gamboa and E. Gassiat. Bayesian methods and maximum entropy for ill-posed inverse problems. *Ann. Statist.*, 25(1):328–350, 1997.

[11] Robert M. Groves, Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey methodology*. Wiley Series in Survey Methodology. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2004.

[12] H. Gzyl. *The method of maximum entropy*, volume 29 of *Series on Advances in Mathematics for Applied Sciences*. World Scientific Publishing Co. Inc., River Edge, NJ, 1995. Sections (6.19)–(6.21) by Aldo Tagliani.

[13] A. Kaplan and R. Tichatschke. Extended auxiliary problem principle using Bregman distances. *Optimization*, 53(5-6):603–623, 2004.

[14] Y. Kitamura and M. Stutzer. Connections between entropic and linear projections in asset pricing estimation. *J. Econometrics*, 107(1-2):159–174, 2002. Information and entropy econometrics.

[15] J. M. Loubes and B. Pelletier. Maximum entropy solution to ill-posed inverse problems with approximately known operator. *J. Math. Anal. Appl.*, 344(1):260–273, 2008.

[16] J.-M. Loubes and P. Rochet. Regularization with approximated l^2 maximum entropy method. In *submitted, Electronic version HAL 00389698*. 2009.

[17] G. E. Montanari. Post-sampling efficient QR-prediction in large-sample surveys. *Internat. Statist. Rev.*, 55(2):191–202, 1987.

[18] G. E. Montanari and M. G. Ranalli. Nonparametric model calibration estimation in survey sampling. *J. Amer. Statist. Assoc.*, 100(472):1429–1442, 2005.

[19] R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

[20] C. E. Särndal. The calibration approach in survey theory and practice. *Statistics Canada*, 33(2):33–119, 2007.

[21] O. Sautory. A new version for the calmar calibration adjustment program. In *Statistics Canada International Symposium Series*.

[22] S. Singh. Generalized calibration approach for estimating variance in survey sampling. *Ann. Inst. Statist. Math.*, 53(2):404–417, 2001.

[23] A. Théberge. Extensions of calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, 94(446):635–644, 1999.

[24] Chang Wu. Optimal calibration estimators in survey sampling. *Biometrika*, 90(4):937–951, 2003.

[25] Chang-chun Wu and Run-chu Zhang. A model-calibration approach to using complete auxiliary information from stratified sampling survey data. *Chinese Quart. J. Math.*, 21(2):309–316, 2006.

[26] Changbao Wu and Randy R. Sitter. A model-calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.*, 96(453):185–193, 2001.