HAL
open science

Maximum Entropy Estimation for Survey sampling

Fabrice Gamboa, Jean-Michel Loubes, Paul Rochet

To cite this version:

Fabrice Gamboa, Jean-Michel Loubes, Paul Rochet. Maximum Entropy Estimation for Survey sampling. 2009. hal-00419169

HAL Id: hal-00419169

https://hal.science/hal-00419169

Preprint submitted on 22 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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: Survey Sampling, Inverse Problems, Maximum Entropy Method
Subject Class. MSC-2000 : 62F12, 62D05, 94A17

Introduction

Calibration is a well spread method to improve estimation in survey sampling, using extra information from an auxiliary variable. This method provides approximately unbiased estimators with variance smaller than that of the usual Horvitz-Thompson estimator (see for example [15]). Calibration has been introduced by Deville and Särndal in [2], extending an idea of [3]. For general references, we refer to [20], [19] and for an extension to variance estimation to [17].
Finding the solution to a calibration equation involves minimizing an energy under some constraint. More precisely, let s be a random sample of size n drawn from a population U of size N, y is the variable of interest and x is a given auxiliary variable, for which the mean t_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 $\hat{t}_y = N^{-1} \sum_{i \in s} w_i y_i$ of the population mean 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

$$N^{-1} \sum_{i \in s} w_i x_i = t_x. \tag{1}$$

Following [18], 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], [8], [7], [10], [6], [5] or [9].
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 weights in the sample mean in order to get a representative sample for the auxiliary variable x, but still being as close as possible to d, which have the desirable property of yielding an unbiased estimate for the mean. 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, it appears that the relationships between the variable to be estimated and the auxiliary variable are crucial to improve estimation (see for example [13] or [20]). When complete auxiliary information is available, increasing the correlation between the variables is made possible by replacing the auxiliary variable x by some function of it, say $u(x)$. So, 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. Simulations illustrate previous results in Section 4 while the proofs are postponed to Section 5.

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 mean $t_y := N^{-1} \sum_{i \in U} y_i$ when only a random subsample s of the whole population is available. So the observed data are $(y_i)_{i \in s}$.

The sampling design is the probability distribution p defined for each subset $s \subset U$ as the probability $p(s)$ that s is observed. 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 $d_i = 1/\pi_i$ is well defined. A standard estimator of t_y is given by the Horvitz-Thompson estimator:

$$\hat{t}_y^{HT} = N^{-1} \sum_{i \in s} \frac{y_i}{\pi_i} = N^{-1} \sum_{i \in s} d_i y_i.$$

This estimator is unbiased and is widely used for practical cases, see for instance [3] for a complete survey.

Suppose that it exists an auxiliary vector variable $x = (x_1, ..., x_N)$, that is entirely observed and set $t_x = N^{-1} \sum_{i \in U} x_i \in \mathbb{R}^k$. If the Horvitz-Thompson estimator of t_x, $\hat{t}_x^{HT} = N^{-1} \sum_{i \in s} d_i x_i$ is far from the true value t_x, it may imply that the sample does not describe well the behavior of the variable of interest in the total population. So, to prevent biased 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 [2]). The *bad sample effect* is corrected by deriving new weights for the sample mean, but still being close to the d_i's to get a small bias. For this, consider a class of weighted estimators $N^{-1} \sum_{i \in U} w_i y_i$ where the weights $w = (w_i)_{i \in s}$ are selected to be close to $d = (d_i)_{i \in s}$ under the calibration constraint

$$N^{-1} \sum_{i \in s} w_i x_i = t_x.$$

There are two basic components in the construction of calibration estimators, namely a dissimilarity and a set of calibration equations. Let $w \mapsto \mathcal{D}(w, d)$ be a dissimilarity between some weights and the Horvitz-Thompson ones. Assume that this dissimilarity is minimal for $w_i = d_i$. The method consists in choosing weights minimizing $\mathcal{D}(., d)$ under the constraint $N^{-1} \sum_{i \in s} w_i x_i = t_x$.

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 $(q_i)_{i \in s}$ a positive smoothing sequence (see [2]). So the new estimator is defined as $\hat{t}_y = N^{-1} \sum_{i \in s} \hat{w}_i y_i$, 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 $N^{-1} \sum_{i \in s} \hat{w}_i x_i = t_x$. Denote by a^t the transpose of a, the solution of this minimization problem is given by

$$\hat{t}_y = \hat{t}_y^{HT} + (t_x - \hat{t}_x^{HT})^t \hat{B},$$

where $\hat{B} = \left[\sum_{i \in s} q_i d_i x_i x_i^t \right]^{-1} \sum_{i \in s} q_i d_i y_i x_i$. Note that this is a generalized regression estimator. It is natural to consider alternative measures, which are given in [2]. 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 have a probabilistic interpretation. This highlights the properties of the resulting estimators.

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. We observe a random sample $T_1, ..., T_n \sim \mu$. For a given function $\mathrm{x} : \mathcal{X} \to \mathbb{R}^k$ and a known quantity $t_x \in \mathbb{R}^k$, we aim to estimate μ satisfying

$$\int_{\mathcal{X}} \mathrm{x}(t) d\mu(t) = t_x. \tag{2}$$

This issue belongs to the class of generalized moment problems with convex constraints (we refer to [4] for general references), which can be solved using maximum entropy on the mean (MEM). The general idea is to modify the empirical distribution $\mu_n = n^{-1} \sum_{i=1}^n \delta_{T_i}$ in order to take into account the additional information on μ given by the moment equation (2). For this, consider weighted versions of the empirical measure $n^{-1} \sum_{i=1}^n p_i \delta_{T_i}$ for weights p_i properly chosen. The MEM estimator $\hat{\mu}_n$ of μ is a weighted version of μ_n, where the weights are the expectation of a random variable $P = (P_1, ..., P_n)$, drawn from a finite measure ν^* close to a *prior* ν. This prior distribution conveys the information that $\hat{\mu}_n$ must be close 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:

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

where we set $\mathrm{x}_i = \mathrm{x}(T_i)$. We then build the MEM estimator $\hat{\mu}_n = n^{-1}\sum_{i=1}^n \hat{p}_i \delta_{T_i}$, where $\hat{p} = (\hat{p}_1, ..., \hat{p}_n) = \mathbb{E}_{\nu^*}(P)$.

This method provides an efficient way to estimate some linear parameter $t_y = \int_{\mathcal{X}} \mathrm{y} d\mu$ for $\mathrm{y} : \mathcal{X} \to \mathbb{R}$ a given map. The empirical mean $\overline{\mathrm{y}} = \int_{\mathcal{X}} \mathrm{y} d\mu_n$ is an unbiased and consistent estimator of t_y but may not have the smallest variance in this model. We can improve the estimation by considering the MEM estimator $\hat{t}_y = n^{-1}\sum_{i=1}^n \hat{p}_i \mathrm{y}_i$, which has a lower variance than the empirical mean and is asymptotically unbiased (see [7]).

In many actual situations, the function x is unknown and only an approximation to it, say x_m, is available. Under regularity conditions, the efficiency properties of the MEM estimator built with the approximate constraint have been studied in [11] and [12], 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

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

It is shown that, under assumptions on x_m, the AMEM estimator of t_y obtained in this way is consistent as n and m tends 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.

2.2 Maximum entropy method for calibration

Recall that our original problem is to estimate the population mean $t_y = N^{-1}\sum_{i\in U} y_i$ based on the observations $\{y_i, i \in s\}$ and auxiliary information $\{x_i, i \in U\}$. We introduce the following notations:

$$\mathrm{y}_i = nN^{-1} d_i y_i, \ \mathrm{x}_i = nN^{-1} d_i x_i, \ p_i = \pi_i w_i.$$

Note that 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 $\hat{t}_y^{HT} = N^{-1}\sum_{i\in s} d_i y_i = n^{-1}\sum_{i\in s} \mathrm{y}_i$ is the preliminary estimator we aim at improving. The calibration constraint $n^{-1}\sum_{i\in s} p_i \mathrm{x}_i = t_x$ stands for the linear condition satisfied by the discrete measure $(p_i)_{i\in s}$. So, 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

$$\mathbb{E}_{\nu^*}\left[n^{-1}\sum_{i\in s} P_i \mathrm{x}_i\right] = \mathbb{E}_{\nu^*}\left[N^{-1}\sum_{i\in s} W_i x_i\right] = t_x.$$

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

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

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 gives 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 and 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 integrable with mean 1. This last assumption conveys that $\hat{p}_i$ must be close to 1, equivalently, $\hat{w}_i = d_i \hat{p}_i$ must be close to the Horvitz-Thompson weight 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 by Λ_ν the log-Laplace transform of ν:

$$\Lambda_\nu(s) = \log \int e^{sx} d\nu(x),\ s \in \mathbb{R}.$$

Its convex conjugate Λ_ν^* is the Cramer transform of ν. 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 [14]) 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. Notice that, under these assumptions, Λ'_{ν_i} is an increasing bijection between the interior of $D(\nu_i)$ and S_{ν_i}. Moreover, we have the functional equalities $({\Lambda^*_{\nu_i}}')^{-1} = \Lambda'_{\nu_i}$ and $(\Lambda^*_{\nu_i})^* = \Lambda_{\nu_i}$.

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:

$$N^{-1} \sum_{i \in s} \delta_i x_i = t_x.$$

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 $N^{-1} \sum_{i \in s} \hat{w}_i x_i = t_x$.

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.

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. We refer to [9] for a definition. In the MEM method, there are two different kinds of priors for which the resulting dissimilarity may be seen as a Bregman divergence. Let ν be a probability measure with mean 1 and such that Λ_ν is a strictly convex function. Then, Λ^*_ν enables to define a Bregman divergence. It will play the role of the dissimilarity resulting from the MEM procedure in the two following situations.
First, consider priors $\nu_i, i \in s$ all taken equal to ν. It is a simple calculation to see that the assumptions made on ν imply that $\Lambda^*_\nu(1) = \Lambda^{*\prime}_\nu(1) = 0$. The resulting dissimilarity can thus be written as

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

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

$$\forall i \in s, \forall x \in \mathcal{X}, d\nu_i(x) = \exp(\alpha_i x + \beta_i) d\nu(d_i x),$$

where $\beta_i = -\Lambda_\nu(\Lambda^{*\prime}_\nu(d_i))$ and $\alpha_i = d_i \Lambda^{*\prime}_\nu(d_i)$ are properly chosen so that ν_i is a probability measure with mean 1. Here we recover after some computation the following dissimilarity

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

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

2.3 Bayesian interpretation of calibration using MEM

In a classical presentation, calibration methods heavily rely on a distance choice. Here, this choice corresponds to different prior measures $(\nu_i)_{i\in s}$. 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 $q_i, i \in s$, let W_i having a Gaussian distributions

$\mathcal{N}(d_i, d_i q_i)$ which corresponds to $\nu_i \sim \mathcal{N}(1, \pi_i q_i)$. 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 choice of the q_i can be seen as the choice of 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:*

$$\forall i \in s,\ \hat{w}_i = d_i \Lambda'_{\nu_i}(\hat{\lambda}^t d_i x_i) \tag{4}$$

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

We endow y with new weights obtaining the MEM estimator $\hat{t}_y = N^{-1}\sum_{i\in s} \hat{w}_i y_i$. We point out that calibration using maximum entropy framework 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

$$N^{-1}\sum_{i\in s} d_i \Lambda'_{\nu_i}(\hat{\lambda}^t d_i x_i) x_i - t_x = 0.$$

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

2.4 Extension to generalized calibration and instrument estimation

Proposition 2.2 shows that a calibration estimator is defined using a family of functions $\Lambda'_{\nu_i}, i \in s$ satisfying the property that the equation $N^{-1}\sum_{i\in s} d_i \Lambda'_{\nu_i}(\lambda^t d_i x_i) x_i = t_x$ has a unique solution. A natural generalization, known as generalized calibration (GC) (see [16]), consists in replacing the functions $\lambda \mapsto \Lambda'_{\nu_i}(\lambda^t d_i x_i)$ by more general functions $f_i : \mathbb{R}^k \to \mathbb{R},\ i \in s$. Assume that the equation

$$F(\lambda) = N^{-1}\sum_{i\in s} d_i f_i(\lambda) x_i = t_x$$

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 $F(0) = \hat{t}_x^{HT}$. 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 $f_i :\ \lambda \mapsto \Lambda'_{\nu_i}(\lambda^t d_i x_i)$ to recover a calibration problem seen in Section 2.2. Even though the method enables a large choice of functions f_i, most cases can not be given a probabilistic interpretation.

However, an interesting particular choice is given by the functions $\lambda \mapsto 1 + z_i^t\lambda$ for $z_i, i \in s$. This sequence of vectors of $\mathbb{R}^k$ is called instruments (see [16]). If the matrix $X_n := N^{-1}\sum_{i\in s} d_i z_i x_i^t$ is invertible, then, the resulting estimator $\hat{t}_y$, referred to as the instrument estimator obtained with the instruments z_i, is given by:

$$\hat{t}_y = \hat{t}_y^{HT} + (t_x - \hat{t}_x^{HT})^t X_n^{-1} N^{-1} \textstyle\sum_{i\in s} d_i z_i y_i. \tag{5}$$

Remark : (dimension reduction) The estimator $\hat{t}_y$ defined in (5) can be viewed as the instrument estimator obtained with auxiliary variable $\hat{B}^t x$ and instruments $\hat{B}^t z_i, i \in s$ with $\hat{B} = \left[\sum_{i\in s} d_i z_i x_i^t\right]^{-1} \sum_{i\in s} d_i y_i z_i$. Hence, in the frame of instrument estimation, the original k-dimensional calibration constraint can be replaced by a one-dimensional linearly modified one $N^{-1}\sum_{i\in U} w_i \hat{B}^t x_i = \hat{B}^t t_x$, without changing the value of the estimator. This enables to reduce the dimension of the problem. Furthermore, it gives an interesting interpretation of the underlying process of calibration. For instance, take the instruments $z_i = x_i, i \in s$. The corresponding variable $B^t x$ is the quadratic projection of y onto the linear space E_x, spanned by the components of x. In other words, $B^t x$ is a linear approximation of y. As a result, the variable $y - B^t x$ has a lower variance than y, while its mean over the population ξ is known up to t_y. So, the variable $y - B^t x$ can be used to estimate t_y and will provide a more efficient estimator. Since B is unknown, we use $\hat{B}$ to estimate it. Set $\tilde{y} = y - \hat{B}^t x$, we have:

$$\hat{t}_y - \hat{B}^t t_x = N^{-1} \sum_{i\in s} d_i \tilde{y}_i.$$

The calibrated estimator $\hat{t}_y$ appears as the Horvitz-Thompson estimator (up to a known additive constant, here $\hat{B}^t t_x$) of a variable $\tilde{y}$ with a lower variance than y. This points out that calibration relies on linear regression, since an estimator of t_y is computed by first constructing a linear projection $\hat{B}^t x$ of y on a subspace E_x. Reducing the dimension of the problem is made by choosing the proper real-valued auxiliary variable, and therefore, the proper one-dimensional linear subspace on which y is projected.

Note also that the accuracy of the estimator heavily relies on the linear correlation between y and the auxiliary variable. It appears that the accuracy could be improved for some non-linear transformation, say $u(x)$, of the original auxiliary variable x, provided that y is more correlated with $u(x)$ than x. This is discussed in Section 3.

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 $\hat{t}_x^{HT}$ and $\hat{t}_y^{HT}$ converge at a rate of convergence of $\sqrt{n}$, as described in [2] and [19] for instance. This will be our framework here. That is

$$\|\hat{t}_x^{HT} - t_x\| = O_{\mathbb{P}}(n^{-1/2}) \text{ and } (\hat{t}_y^{HT} - t_y) = O_{\mathbb{P}}(n^{-1/2}).$$

In this framework, all GC estimators are $\sqrt{n}$-consistent, as seen in [2].

Definition We say that two GC estimators $\hat{t}_y$ and $\tilde{t}_y$ are asymptotically equivalent if $(\hat{t}_y - \tilde{t}_y) = o_{\mathbb{P}}(n^{-1/2})$.

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 $i \in s$, $\nabla f_i(0) = \nabla g_i(0) = z_i$, and if the matrix $X_n := N^{-1} \sum_{i \in s} d_i z_i x_i^t$ 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 [2]. It states that for all GC estimator, there exists an instrument estimator having the same asymptotic behavior, built by taking as instruments the gradient vectors of the criterion functions at 0: $z_i = \nabla f_i(0), i \in s$. Consequently, a MEM estimator $\hat{t}_y$ built with prior distributions $\nu_i, i \in s$ with mean 1 and respective variances $\pi_i q_i$ for $(q_i)_{i \in s}$ a given positive sequence, satisfies

$$\hat{t}_y = \hat{t}_y^{HT} + (t_x - \hat{t}_x^{HT})^t \hat{B} + o_{\mathbb{P}}(n^{-1/2})$$

where $\hat{B} = \left[\sum_{i \in s} d_i q_i x_i x_i^t\right]^{-1} \sum_{i \in s} d_i q_i x_i y_i$. The negligible term $o_{\mathbb{P}}(n^{-1/2})$ is zero for all n for Gaussian priors $\nu_i \sim \mathcal{N}(1, \pi_i q_i)$, which stresses the important role played by the corresponding χ^2 dissimilarity (see Example 1 in Section 2.3). Note also that the Gaussian equivalent $\tilde{t}_y = \hat{t}_y^{HT} + (t_x - \hat{t}_x^{HT})^t \hat{B}$ is the instrument estimator built with the instruments $z_i = q_i x_i$. This choice of instruments, and in particular the case $q_i = 1$ for all $i \in s$, is often used in practice due to its simplicity and good consistency.

3 Efficiency of calibration estimator with MEM method

By using the auxiliary variable x in the calibration constraint, we implicitly assume that x and y are linearly related. However, other relationships may prevail between the variables and it may be more accurate to consider some other auxiliary variable $u(x)$. Here, we discuss optimal choices of function $u : \mathcal{X} \to \mathbb{R}^d$ to use in the calibration constraint. To do so, we first define a notion of asymptotic efficiency in our model with fixed auxiliary variable $u(x)$. Then, we study the influence of the choice of the constraint function u and find the optimal choice leading to the most efficient estimator. Finally, we propose a method based on the approximate maximum entropy on the mean which enables to compute an asymptotically optimal estimate of t_y, taking into consideration both the choice of the constraint function u and the instruments z_i.

3.1 Asymptotic efficiency

In order to choose between calibration estimators, we now define a notion of asymptotic efficiency for a given calibration constraint. Although a GC estimator is entirely determined by a family $f_i, i \in s$ of functions, only the values $z_i = \nabla f_i(0), i \in s$ matter to study the asymptotic behavior of the estimator, up to a negligible term of order $o_{\mathbb{P}}(n^{-1/2})$. Let $u : \mathcal{X} \to \mathbb{R}^d$ be a given function, and consider:

$$t_u = N^{-1} \sum_{i \in U} u(x_i), \ \hat{t}_{u\pi} = N^{-1} \sum_{i \in s} d_i u(x_i).$$

We make the following assumptions.

A1: $\xi := \{(x_i, y_i), i \in U\}$ are independent realizations of (X, Y), with $\mathbb{E}(Y|X) \neq \mathbb{E}(Y)$ and $\mathbb{E}(|Y^3|) < \infty$. Note respectively P_X and P_{XY} the distributions of X and (X, Y).

A2: The sampling design $p(.)$ does not depend on ξ.

A3: n and N/n tend to infinity. This will be denoted by $(n, N/n) \to \infty$.

Furthermore, u is assumed to be measurable and such that $\mathbb{E}(\|u(X)^3\|) < \infty$. Given the constraint function u and instruments $z_i, i \in s$, we note $\hat{t}_y(u)$ the resulting instrument estimator, the dependency in z_i is dropped for ease of notation. We now study the asymptotic behavior of $\hat{t}_y(u)$ with respect to the instruments $z_i, i \in s$. Here, the weights $\hat{w}$ are adapted to the new calibration constraint $N^{-1} \sum_{i \in U} \hat{w}_i u(x_i) = t_u$, yielding

$$\hat{t}_y(u) = N^{-1} \sum_{i \in U} \hat{w}_i y_i = \hat{t}_y^{HT} + (t_u - \hat{t}_{u\pi})^t \hat{B}_u,$$

where $\hat{B}_u = \left[\sum_{i \in s} d_i z_i u(x_i)^t\right]^{-1} \sum_{i \in s} d_i y_i z_i$ is assumed to be well defined and to converge in probability towards a constant vector B_u as $(n, N/n) \to \infty$.

In order to define a criterion of efficiency, we first need to construct an asymptotic variance lower bound for instrument estimators. Note $\mathbb{E}_\xi(t_y - \hat{t}_y(u))^2$ the quadratic risk of $\hat{t}_y(u)$ under p, the population ξ being fixed, we aim to determine a lower bound for the limit of $n\mathbb{E}_\xi(t_y - \hat{t}_y(u))^2$ as $(n, N/n) \to \infty$ (provided that the limit exists). The value of the limit of course heavily relies on the asymptotic behavior of the sampling design. Without some control on the Horvitz-Thompson weights π_i, we can not derive consistency properties for instrument estimators. Note $\pi_{ij} = \sum_{s:\ i,j \in s} p(s)$ the joint inclusion probability of i and j and let $\Delta_{ij} = \pi_{ij} d_i d_j - 1$, we make the following technical assumptions.

A4: $\sum_{i \in U} \Delta_{ii}^2 = o(N^4 n^{-2})$, $\sum_{i \in U} \sum_{j \neq i} \Delta_{ij}^2 = o(N^3 n^{-2})$.

A5: $\lim\limits_{\substack{n \to \infty \\ N/n \to \infty}} nN^{-2} \sum_{i \in U} \Delta_{ii} = - \lim\limits_{\substack{n \to \infty \\ N/n \to \infty}} nN^{-2} \sum_{i \in U} \sum_{j \neq i} \Delta_{ij} = 1.$

Assumption 4 is sufficient to ensure that the HT estimator of some variable $a(x_i, y_i), i \in U$ is $\sqrt{n}$-consistent provided that $\mathbb{E}(a(X,Y)^2) < \infty$. Furthermore, Assumption 5 ensures the existence of its asymptotic variance. Note that these assumptions do not take into consideration the population ξ, so that it makes them easy to check in practical cases. For example, the assumptions are fulfilled for the uniform sampling design, that is when p is such that every sample $s \subset U$ has the same probability of being observed. In that case, the Horvitz-Thompson weights are $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/N(N-1), \forall i \neq j$, yielding $\Delta_{ii} = N/n - 1$ and $\Delta_{ij} = -(N-n)/n(N-1)$. We can now state our first result.

Lemma 1: *Suppose that Assumptions 1 to 4 hold. Then,*

$$n\mathbb{E}_\xi(t_y - \hat{t}_y(u))^2 \geq \operatorname{var}\left(Y - B_u^t u(X)\right) + o_{\mathbb{P}}(1),$$

with equality if, and only if, Assumption 5 also holds.

We point out that an asymptotic lower bound for the variance can be defined for instrument estimators as soon as Assumptions 1 to 4 hold. The lower bound (denoted by $V^*(u)$) is the minimum of $\operatorname{var}(Y - B^t u(X))$ for B ranging over $\mathbb{R}^d$. It can be computed explicitly if the matrix $\operatorname{var}(u(X))$ is invertible:

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

We say that an estimator $\hat{t}_y(u)$ is asymptotically efficient if its asymptotic variance is $V^*(u)$. Note that this lower bound can not be reached if Assumption 5 is not true. We now come to our second result.

Lemma 2: *Suppose that Assumptions 1 to 5 hold. If* $\operatorname{var}(u(X))$ *is invertible,* $\hat{t}_y(u)$ *built with instrument* $z_i, i \in s$ *is asymptotically efficient if, and only if,*

$$\lim_{(n,N/n)\to+\infty} \left[\sum_{i\in s} d_i z_i u(x_i)^t\right]^{-1} \sum_{i\in s} d_i y_i z_i = \left[\operatorname{var}(u(X))\right]^{-1} \operatorname{cov}(Y, u(X)). \tag{6}$$

In an asymptotic concern and when the calibration function u is fixed, finding the best instruments $z_i, i \in s$ in order to estimate t_y becomes a simple optimization problem which depends only on the limit B_u of $\hat{B}_u = \left[\sum_{i\in s} d_i z_i u(x_i)^t\right]^{-1} \sum_{i\in s} d_i y_i z_i$. Asymptotic efficiency is obtained by choosing instruments minimizing the asymptotic variance. Hence, calculating B_u provides an efficient and easy way to prove the asymptotic efficiency of an instrument estimator. Moreover, this criterion of asymptotic efficiency can be extended to the set of all generalized calibration estimators, as a consequence of Proposition 2.3. A GC estimator defined by the functions $f_i, i \in s$ is asymptotically efficient if and only if the vectors $z_i = \nabla f_i(0), i \in s$ satisfy (6).

Proof of Lemmas 1 and 2: First compute the quadratic risk of $\hat{t}_y(u)$. Due to its non linearity it is a difficult task. We rather consider its linear asymptotic expansion $\hat{t}_{y,\text{lin}}(u) := \hat{t}_y^{HT} + (t_u - \hat{t}_{u\pi})B_u$ where we recall that B_u is the limit (in probability) of $\hat{B}_u$. Note that the random effect is due to the sampling design p, the population ξ is fixed. We obtain after calculation the following expression for the quadratic risk

$$\mathbb{E}_\xi(t_y - \hat{t}_{y,\text{lin}}(u))^2 = N^{-2} \sum_{i,j\in U} \Delta_{ij}\ (y_i - B_u^t u(x_i))(y_j - B_u^t u(x_j)).$$

Then, the results follow directly from Lemma 5.1, given in the Appendix. ∎
We now see some examples of well-used estimators.

Asymptotic variance of some GC estimators

1. Optimal instruments.
   Assume for sake of simplicity that u is real-valued. We denote by B_u^{min} the value of B_u achieving the minimal value of the quadratic risk:

$$B_u^{min} = \frac{\sum_{i,j\in U} \Delta_{ij}\ u(x_j)y_i}{\sum_{i,j\in U} \Delta_{ij}\ u(x_i)u(x_j)} = \frac{\sum_{i\in U} y_i(\sum_{j\in U} \Delta_{ij}\ u(x_j))}{\sum_{i\in U} u(x_i)(\sum_{j\in U} \Delta_{ij}\ u(x_j))}.$$

   The corresponding instruments are $z_i = \sum_{j\in U} \Delta_{ij}\ u(x_j), \forall i$. By Lemma 5.1, we see that B_u^{min} converges toward $\text{cov}(Y, u(X))/\text{var}(u(X))$ as $(n, N/n) \to \infty$, Equation (6) is thus true in that case. If the sampling design is uniform, we obtain after calculation $z_i = \frac{N(N-n)}{n(N-1)}(u(x_i) - t_u)$, and we have:

$$B_u^{min} = \frac{\sum_{i\in U} y_i z_i}{\sum_{i\in U} z_i u(x_i)} = \frac{\text{cov}_e(y, u(x))}{\text{var}_e(u(x))}$$

   where cov_e and var_e denote the empirical covariance and variance for the population ξ given by $\text{cov}_e(y, u(x)) = N^{-1}\sum_{i\in U} y_i(u(x_i) - t_u)$ and $\text{var}_e(u(x)) = \text{cov}_e(u(x), u(x))$. Finally,

$$n\mathbb{E}_\xi(t_y - \hat{t}_{y,\text{lin}})^2 = (1 - nN^{-1})\ \text{var}_e\left(y - \frac{\text{cov}_e(y, u(x))}{\text{var}_e(u(x))}u(x)\right) + o(1).$$

   We have $\lim_{(n,N/n)\to\infty} n\mathbb{E}_\xi(t_y - \hat{t}_{y,\text{lin}})^2 = V^*(u)$, as expected. This estimator is thus asymptotically efficient. Although, instruments used for its computation depend on the whole population $(x_i)_{i\in U}$ and therefore, they may be computationally expensive.

2. MEM estimators.
   Take the instruments $z_i = q_i u(x_i), \forall i \in s$ for $(q_i)_{i\in s}$ a positive sequence. As seen in

Section 2.2, these instruments describe the asymptotic behavior of MEM estimators built using prior distributions ν_i with respective variances $\pi_i q_i$. Even though this choice is often used in practical cases, we see that it does not necessarily lead to an asymptotically efficient estimator $\hat{t}_y(u)$. Indeed, under regularity conditions on q_i which ensure the convergence of $\hat{B}_u$ (basically, the assumptions of Proposition 3.1, which are true for instance if we take $q_i = 1$), we have:

$$\hat{B}_u = \left[\sum_{i\in s} d_i q_i u(x_i) u(x_i)^t\right]^{-1} \sum_{i\in s} d_i q_i y_i u(x_i) \xrightarrow{\mathbb{P}} \left[\mathbb{E}(u(X)u(X)^t)\right]^{-1} \mathbb{E}(Yu(X)).$$

These instruments satisfy Equation (6) only if

$$\left[\mathbb{E}(u(X)u(X)^t)\right]^{-1} \mathbb{E}(Yu(X)) = [\mathrm{var}(u(X)]^{-1} \mathrm{cov}(Y, u(X)).$$

This is true when $u(.) = \mathbb{E}(Y|X = .)$ or for any u such that $\mathbb{E}(u(X)) = 0$, MEM estimators are thus asymptotically efficient in these cases. When this condition is not fulfilled, an easy method to compute an efficient estimator consists in adding the constant variable 1 in the calibration constraint. We then consider the MEM estimator $\hat{t}_y(v)$ where $v = (1, u)^t : \mathcal{X} \to \mathbb{R}^{d+1}$, the calibrated weights now satisfy the constraints

$$N^{-1} \sum_{i\in s} w_i u(x_i) = t_u, \; N^{-1} \sum_{i\in s} w_i = 1.$$

Here, the matrix $\mathrm{var}(v(X))$ is not invertible although we see after a direct calculation that $V^*(v) = V^*(u)$. So, the auxiliary variable is modified but the asymptotic lower bound is unchanged. Furthermore, the MEM estimator $\hat{t}_y(v)$ obtained in this way is asymptotically efficient, as it is proved in the following proposition.

Proposition 3.1 *Suppose that Assumptions 1 to 5 hold. Let $(\nu_i)_{i\in s}$ be a family of probability measures with mean 1 and respective variance $q_i\pi_i$ with $(q_i)_{i\in s}$ a given positive sequence. Assume that there exists $\kappa = \kappa(n, N) \in \mathbb{R}$ such that $\kappa \sum_{i\in s} q_i d_i$ is bounded away from zero and $\kappa^2 \sum_{i\in s} (q_i d_i)^2 \to 0$ as $(n, N/n) \to +\infty$. Let $v = (1, v_1, ..., v_d) : \mathcal{X} \to \mathbb{R}^{d+1}$ be a map, where $1, v_1, ..., v_d$ are linearly independent. Then, the MEM estimator built with prior distribution $\nu = \otimes_{i\in s}\nu_i$ and calibration constraint $N^{-1} \sum_{i\in s} w_i v(x_i) = t_v$ is asymptotically efficient.*

3.2 Approximate Maximum Entropy on the Mean

We now turn on the optimal choice of the auxiliary variable $u(x)$ defining the calibration constraint. For a given constraint function u, we implicitly take asymptotically optimal instruments $z_i, i \in s$, that is, instruments such that the resulting estimator $\hat{t}_y(u)$ has asymptotic variance $V^*(u)$. Hence, minimizing the asymptotic variance of GC estimators with respect to u and $(z_i)_{i\in s}$ reduces to minimizing $V^*(u)$ with respect to u.

In an asymptotic framework, u can be taken with values in $\mathbb{R}$ without loss of generality, as discussed in Section 2.4. So, for a real valued constraint function u, $V^*(u)$ is defined as:

$$V^*(u) = \inf_{B\in\mathbb{R}} \operatorname{var}(Y - Bu(X)) = \operatorname{var}\left(Y - \frac{\operatorname{cov}(Y, u(X))}{\operatorname{var}(u(X))}u(X)\right).$$

A function v for which $V^*(v)$ is minimal over the set σ_X of all real X-measurable functions has the form $v(.) = \alpha\mathbb{E}(Y|X = .) + \beta$ for some $(\alpha, \beta) \in \mathbb{R}^* \times \mathbb{R}$. Hence, the conditional expectation $\Phi(x) = \mathbb{E}(Y|X = x)$ (or any bijective affine transformation of it) turns out to be the best choice for the auxiliary variable in term of asymptotic efficiency. In that case, the asymptotic lower bound is given by:

$$V^* = \min_{u\in\ \sigma_X} V^*(u) = \mathbb{E}(Y - \mathbb{E}(Y|X))^2.$$

For practical applications, this result is useless since the conditional expectation Φ depends on the unknown distribution of (X, Y). If Φ were known, the problem of estimating t_y would be easier since the observed value $t_\Phi = N^{-1}\sum_{i\in U}\Phi(x_i)$ is a $\sqrt{N}$-consistent estimator of t_y and is therefore much more efficient than any calibrated estimator. When the conditional expectation Φ is unknown, 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 variance is equal to the lower bound V^* defined above.

For all measurable function u, we now denote by $\hat{t}_y(u)$ the MEM estimator of t_y obtained with prior distributions $\nu_i \sim \mathcal{N}(1, \pi_i)$ and auxiliary variables $u(x)$ and 1. We recall that $\hat{t}_y(u)$ is $\sqrt{n}$-consistent with asymptotic variance $V^*(u)$, as shown in Proposition 3.1. Moreover, we know that the asymptotic variance of MEM estimators $\hat{t}_y(u)$ is minimal for the unknown value $u = \Phi$. The AMEM procedure consists in replacing Φ by its approximation Φ_m in the calibration constraint. The so-obtained AMEM estimator $\hat{t}_y(\Phi_m)$ is thus quite easily computable but still verifies interesting convergence properties as shown in the next proposition.

Proposition 3.2 *Suppose that Assumptions 1 to 5 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 $n\mathbb{E}_\xi(t_y - \hat{t}_y(\Phi_m))^2$ converges toward V^* as $n, N/n, m \to \infty$.

When applied to this context, approximate maximum entropy on the mean enables to increase 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 weaker conditions.

Corollary 3.3 *Suppose that Assumptions 1 to 5 hold. Let $(\Phi_m)_{m\in\mathbb{N}}$ be a sequence of functions satisfying*

$$i)\ n\mathbb{E}_\xi(\hat{t}_{\Phi\pi} - t_\Phi - (\hat{t}_{\Phi_m\pi} - t_{\Phi_m}))^2 \underset{(n,N/n,m)\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} 0 \ \ \textit{and} \ \ ii)\ \hat{B}_{\Phi_m} \underset{(n,N/n,m)\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} 1.$$

Then, the estimator $\hat{t}_y(\Phi_m)$ is asymptotically efficient.

This corollary does not rule out that the functions Φ_m are estimated using the data, which was not the case in Proposition 3.2. Hence, it becomes possible to compute an asymptotically efficient estimator of t_y without external estimator Φ_m of Φ. A data driven estimator Φ_n provides as well an asymptotically efficient estimator of t_y, as soon as the two conditions of Corollary 3.3 are fulfilled.

Now consider an example of AMEM estimator for which the computation is particularly simple, and that provides interesting interpretations. We assume for simplicity that the sampling design is uniform, here $\hat{t}_y^{HT}$ is simply equal to $N^{-1}\sum_{i\in s} y_i$. Let $(\phi^1, \phi^2, ...)$ be a linearly independent total family of $\mathbb{L}^2(P_X)$. That is, for all measurable function $f : \mathbb{R}^k \to \mathbb{R}$ such that $E(f(X)^2) < \infty$, there exists a unique sequence $(\alpha_n)_{n\in\mathbb{N}}$ such that

$$f(X) = \mathbb{E}(f(X)) + \sum_{i\in\mathbb{N}} \alpha_i[\phi^i(X) - \mathbb{E}(\phi^i(X))].$$

For all m, the projection Φ_m of Φ on vect $\{1, \phi^1, ..., \phi^m\}$ is given by

$$\Phi_m(.) = \mathbb{E}(Y) + \mathrm{cov}(Y, \phi_m(X))^t \left[\mathrm{var}(\phi_m(X))\right]^{-1} \left[\phi_m(.) - \mathbb{E}(\phi_m(X))\right]$$

where $\phi_m = (\phi^1, ..., \phi^m)^t$. When n is large enough in comparison to m, we can define a natural projection estimator $\Phi_{m,n}$ of Φ as

$$\Phi_{m,n}(.) = \hat{t}_y^{HT} + \hat{B}_{\phi_m}^t \left[\phi_m(.) - \hat{t}_{\phi_m\pi}\right]$$

where $\hat{B}_{\phi_m}=\left[\sum_{i\in s} y_i(\phi_m(x_i) - \hat{t}_{\phi_m\pi})\right]^t \left[\sum_{i\in s} \phi_m(x_i)(\phi_m(x_i) - \hat{t}_{\phi_m\pi})^t\right]^{-1}$.
We now consider the AMEM estimator $\hat{t}(\Phi_{m,n})$:

$$\hat{t}_y(\Phi_{m,n}) = \hat{t}_y^{HT} + \frac{\sum_{i\in s} y_i(\Phi_{m,n}(x_i) - \hat{t}_{\Phi_{m,n}\pi})}{\sum_{i\in s} \Phi_{m,n}(x_i)(\Phi_{m,n}(x_i) - \hat{t}_{\Phi_{m,n}\pi})}(t_{\Phi_{m,n}} - \hat{t}_{\Phi_{m,n}\pi})$$

which, after simplification, gives

$$\hat{t}_y(\Phi_{m,n}) = \hat{t}_y^{HT} + \hat{B}_{\phi_m}^t(t_{\phi_m} - \hat{t}_{\phi_m\pi}) = t_{\Phi_{m,n}}.$$

The objective is to find a path $(m(n), n)_{n\in\mathbb{N}}$ for which the estimator $\Phi_n := \Phi_{m(n),n}$ satisfies the conditions of Corollary 3.3. We know that, for all m:

$$\begin{aligned} & n\mathbb{E}_\xi(\hat{t}_{\Phi\pi} - t_\Phi - (\hat{t}_{\Phi_{m,n}\pi} - t_{\Phi_{m,n}}))^2 \\ = \; & n\mathbb{E}_\xi(\hat{t}_{\Phi\pi} - t_\Phi + (t_{\phi_m} - \hat{t}_{\phi_m\pi})^t \hat{B}_{\phi_m})^2 \\ = \; & nN^{-2}\textstyle\sum_{i,j\in U} \Delta_{ij}(\Phi(x_i) - B_{\phi_m}^t \phi_m(x_i))(\Phi(x_j) - B_{\phi_m}^t \phi_m(x_j)) + o_{\mathbb{P}}(1) \end{aligned}$$

where $B_{\phi_m} = \lim_{(n,N/n)\to\infty} \hat{B}_{\phi_m} = \mathrm{cov}(Y, \phi_m(X))^t \left[\mathrm{var}(\phi_m(X))\right]^{-1}$. By Lemma 5.1, we get:

$$\forall m,\; n\mathbb{E}_\xi(\hat{t}_{\Phi\pi} - t_\Phi - (\hat{t}_{\Phi_{m,n}\pi} - t_{\Phi_{m,n}}))^2 \underset{(n,N/n)\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} \mathrm{var}(\Phi(X) - \Phi_m(X)).$$

Since the convergence is true for all m, we can extract a sequence of integers $(m(n))_{n\in\mathbb{N}}$ such that $\Phi_n := \Phi_{m(n),n}$ undergoes the first condition of Corollary 3.3:

$$n\mathbb{E}_\xi(\hat{t}_{\Phi\pi} - t_\Phi - (\hat{t}_{\Phi_n\pi} - t_{\Phi_n}))^2 \underset{(n,N/n)\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} 0.$$

The second condition of Corollary 3.3 is verified for such a sequence $(\Phi_n)_{n\in\mathbb{N}}$ since for all n, $\hat{B}_{\Phi_n} = 1$. So finally we conclude that the AMEM estimator $\hat{t}(\Phi_n)$ is asymptotically optimal.

Remark : The AMEM estimator is obtained by plugging an estimator Φ_n of Φ in the calibration constraint. Note that $\hat{t}_y(\Phi_n)$ is the MEM estimator we obtain with constraint function $(1, \phi_{m(n)}^t)^t$. Indeed, $\hat{t}_y(\Phi_n) = \hat{t}_y^{HT} + \hat{B}_{\phi_{m(n)}}^t(t_{\phi_{m(n)}} - \hat{t}_{\phi_{m(n)}\pi})$. This is a consequence of the dimension reduction property relative to instrument estimators discussed in Section 2.4, Φ_n is an affine approximation of y by the components of $\phi_{m(n)}(x)$. By increasing properly the number of constraints, the projection will converge toward the conditional expectation $\Phi(x)$ yielding an efficient estimator of t_y.
We can also rewrite the estimator as $\hat{t}_y(\Phi_n) = t_{\Phi_n}$. In these settings, we can interpret the AMEM procedure as building an estimator of t_Φ instead of estimating t_y. Because $\Phi(x)$ is not a function of y, it can be estimated by the empirical mean over the whole population U. An estimator of t_Φ will asymptotically yield an estimate of t_y as a consequence of the relation $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$.

4 Numerical simulations

We shall now give some numerical applications of our results. We made a simulation of a population U of size $N = 100000$, where X is a uniform variable on the interval

[1; 2], and we take $Y = \exp(X) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ an independent noise. So, the conditional expectation Φ mentioned in the last section is simply the function exp(.). The sampling design is uniform and the sample s is taken of size 121. We consider six instruments estimators, $\hat{t}_1$ to $\hat{t}_6$, of which we make 50 realizations observed from 50 different samples drawn from the fixed population U, and we give for $i = 1, ..., 6$ an estimator V_i of the variance calculated from the 50 observations. The first estimator considered $\hat{t}_1$ is the Horvitz-Thompson estimator, and the last one $\hat{t}_6$ is the AMEM estimator taken as example in Section 3.2, where we took the family $\{X^i : i \in \mathbb{N}\}$ for the base of $\mathbb{L}^2(P_X)$, and we set the number m of constraint functions to $m = 6$. The construction of the estimators are detailed in the following table. The results are given for two different values of σ^2, namely $\sigma^2 = 1$ and $\sigma^2 = 0.1$.

1. $\varepsilon \sim \mathcal{N}(0, 1)$:

	auxiliary variable	instrument	estimated variance
$\hat{t}_1$ (H-T estimator)	none	none	$V1 = 2.07 \times 10^{-2}$
$\hat{t}_2$	x	$(x_i)_{i \in s}$	$V2 = 7.8 \times 10^{-3}$
$\hat{t}_3$	$\mathrm{x} = (1, x)$	$(\mathrm{x}_i)_{i \in s}$	$V3 = 7.6 \times 10^{-3}$
$\hat{t}_4$	$\exp(x)$	$(\exp(x_i))_{i \in s}$	$V4 = 7.2 \times 10^{-3}$
$\hat{t}_5$	$\mathrm{x} = (1, \exp(x))$	$(\mathrm{x}_i)_{i \in s}$	$V5 = 6.9 \times 10^{-3}$
$\hat{t}_6$ (AMEM estimator)	$\mathrm{x} = (1, x, x^2, x^3, x^4, x^5, x^6)$	$(\mathrm{x}_i)_{i \in s}$	$V6 = 7.2 \times 10^{-3}$

We observe that the calibrated estimators appear to be better than the Horvitz-Thompson estimator. The choice of the auxiliary variable or the instrument does not seem to have a significant effect on the efficiency.

2. $\varepsilon \sim \mathcal{N}(0, 0.1)$:

	auxiliary variable	instrument	estimated variance
$\hat{t}_1$ (H-T estimator)	none	none	$V1 = 1.93 \times 10^{-2}$
$\hat{t}_2$	x	$(x_i)_{i \in s}$	$V2 = 3.1 \times 10^{-3}$
$\hat{t}_3$	$\mathrm{x} = (1, x)$	$(\mathrm{x}_i)_{i \in s}$	$V3 = 8.7 \times 10^{-4}$
$\hat{t}_4$	$\exp(x)$	$(\exp(x_i))_{i \in s}$	$V4 = 6.8 \times 10^{-4}$
$\hat{t}_5$	$\mathrm{x} = (1, \exp(x))$	$(\mathrm{x}_i)_{i \in s}$	$V5 = 6.7 \times 10^{-4}$
$\hat{t}_6$ (AMEM estimator)	$\mathrm{x} = (1, x, x^2, x^3, x^4, x^5, x^6)$	$(\mathrm{x}_i)_{i \in s}$	$V6 = 7.0 \times 10^{-4}$

Here, X explains almost entirely Y since the variance of ε is low ($\sigma^2 = 0.1$). In that case, the choice of the auxiliary variable and instrument appears to play a more important role. We notice a significant difference between $\hat{t}_2$ and $\hat{t}_3$ which points out the importance of the

instrument. More specifically, we see that the instrument $(x_i - \hat{t}_x^{HT})_{i\in s}$ (which is equivalent to adding the constant 1 as an auxiliary variable) provides a better estimator than x_i. Furthermore, also note that using the auxiliary variable $\Phi(x) = \exp(x)$ provides the best estimator in term of minimal variance as we see that $V4$ and $V5$ are the smallest estimated variances. These estimators can be viewed as oracles, since the auxiliary variable used in that case is the optimal choice, but is in general unknown (see Section 3.2). The difference between $\hat{t}_4$ and $\hat{t}_5$ is not significant, as expected, according to the second example of Section 3.1. Finally, the AMEM estimator has its variance lying between that of the standard calibrated estimator $\hat{t}_3$ and that of the oracles, which conveys that it is more efficient than $\hat{t}_3$.

5 Appendix

5.1 Technical lemma

Lemma 5.1 *Let $\mathcal{F}$ be the set of all functions $f : (\mathbb{R}^k \times \mathbb{R}) \to \mathbb{R}$ such that $\mathbb{E}(|f(X,Y)|^3)$ is finite (we set $f_i = f(x_i, y_i)$ for all $i \in U$). Under Assumptions 1, 2 and 4,*

$$\forall f \in \mathcal{F},\ nN^{-2} \sum_{i,j\in U} \Delta_{ij}\ f_i f_j \geq \mathrm{var}(f(X,Y)) + o_{\mathbb{P}}(1)$$

as $(n, N/n) \to \infty$, with equality if and only if Assumption 5 also holds. In that case, the quantity $nN^{-2}\sum_{i,j\in U}\Delta_{ij}\ f_i g_j$ converges in probability toward $\mathrm{cov}(f(X,Y), g(X,Y))$ for all $f, g \in \mathcal{F}$ as $(n, N/n) \to \infty$.

Proof of Lemma 5.1:
Assumptions 1, 2 and 4 yield for all $f \in \mathcal{F}$:

$$\begin{aligned} & nN^{-2}\textstyle\sum_{i,j\in U}\Delta_{ij}\ f_i f_j = nN^{-2}\sum_{i\in U}\Delta_{ii}\ f_i^2 + nN^{-2}\sum_{i\neq j}\Delta_{ij}\ f_i f_j \\ = \ & \left(nN^{-2}\textstyle\sum_{i\in U}\Delta_{ii}\right)\mathbb{E}(f(X,Y)^2) + \left(nN^{-2}\textstyle\sum_{i\neq j}\Delta_{ij}\right)\mathbb{E}(f(X,Y))^2 + o_{\mathbb{P}}(1) \end{aligned}$$

Let $\mathcal{P}_n(U)$ denote the set of all subsample s of U with n elements. By Jensen inequality, we get

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

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

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

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

$$nN^{-2}\textstyle\sum_{i,j\in U}\Delta_{ij}\ f_i f_j \geq \mathrm{var}(f(X,Y)) + o_{\mathbb{P}}(1).$$

as $(n, N/n) \to \infty$. Furthermore, it is not an equality for all $f \in \mathcal{F}$ if Assumption 5 is not true. We show the second part of the lemma using the same pattern as in the beginning of the proof applied to f and g. In particular, it holds when $f = g$.

5.2 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:

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

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:

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

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

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

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:

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

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

$$\mathcal{L}(\lambda, w) = \sum_{i\in s} \Lambda^*_{\nu_i}(w_i\pi_i) - \lambda^t \left(\sum_{i\in s} w_i x_i - N t_x\right)$$

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

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

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 $\hat{\lambda}$:

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

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

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

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

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

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

$$\hat{\lambda} = X_n^{-1}(t_x - \hat{t}^{HT}_x) + o_{\mathbb{P}}(n^{-1/2}).$$

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

$$f_i(\hat{\lambda}) = 1 + z_i^t\hat{\lambda} + o_{\mathbb{P}}(n^{-1/2}) = 1 + z_i^t\tilde{\lambda} + o_{\mathbb{P}}(n^{-1/2}) = g_i(\tilde{\lambda}) + o_{\mathbb{P}}(n^{-1/2}).$$

It follows that $\hat{t}_y$ and $\tilde{t}_y$ are asymptotically equivalent.
We know that MEM estimation reduces to taking $f_i(.) = \Lambda'_{\nu_i}(d_i x_i^t .)$ in a GC procedure. Hence, in that case, $\nabla f_i(0) = d_i\Lambda''_{\nu_i}(0)x_i$. 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 $f_i(\lambda) = \Lambda'_{\nu_i}(d_i x_i^t\lambda) = 1 + q_i x_i^t\lambda$. The resulting MEM estimator is thus the instrument estimator obtained with instruments $z_i = q_i x_i, i \in s$.

Proof of Proposition 3.1:
We set $u = (v_1, ..., v_d)$, the matrix $\mathrm{var}(u(X))$ is invertible. By assumption on $(q_i)_{i\in s}$, we have

$$\kappa \sum_{i\in s} d_i q_i y_i v(x_i) = (\kappa \sum_{i\in s} d_i q_i)\mathbb{E}(Y v(X)) + \kappa\ o_{\mathbb{P}}(1)$$

and

$$\kappa \sum_{i\in s} d_i q_i v(x_i) v(x_i)^t = (\kappa \sum_{i\in s} d_i q_i)\mathbb{E}(v(X)v(X)^t) + \kappa\ o_{\mathbb{P}}(1).$$

Since $(\kappa \sum_{i\in s} d_i q_i)$ is bounded away from zero, it follows that

$$\hat{B}_v = \left[\sum_{i\in s} d_i q_i v(x_i) v(x_i)^t\right]^{-1} \sum_{i\in s} d_i q_i y_i v(x_i) \xrightarrow{\mathbb{P}} \left[\mathbb{E}(v(X)v(X)^t)\right]^{-1} \mathbb{E}(Y v(X)) = B_v.$$

By simple algebra, we show the functional equality $B_v^t v(.) = B_u^t u(.) + K$, where K is constant, and therefore does not modify the value of the variance. More precisely, the asymptotic variance of $\hat{t}_y(v)$ is

$$\mathrm{var}(Y - \mathrm{cov}(Y, u(X))^t \left[\mathrm{var}(u(X)\right]^{-1} u(X) + K) = V^*(u),$$

which proves that the MEM estimator $\hat{t}_y(v)$ is asymptotically efficient.

Proof of Proposition 3.2:
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

$$n\mathbb{E}_\xi(\hat{t}_{\Phi\pi} - t_\Phi - (\hat{t}_{\Phi_m\pi} - t_{\Phi_m}))^2 = O_{\mathbb{P}}(\varphi_m^{-1}) \ \text{ and } \ (\hat{B}_{\Phi_m} - 1) = O_{\mathbb{P}}(\varphi_m^{-1/2})$$

as $n, N/n \to \infty$ and uniformly for all m (see the proof of Lemma 1 in [12]). Hence, 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 $n, N/n, m \to \infty$. We conclude using Result 2 and Lemma 5.1.

Proof of Corollary 3.3:
All conditions are fulfilled so that the proof of Proposition 3.2 remains valid in that case.

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

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

[4] 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.

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] 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.

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

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

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

[15] Carl Erik Särndal. On uniformly minimum variance estimation in finite populations. *Ann. Statist.*, 4(5):993–997, 1976.

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

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

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

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

[20] 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.