

Maximum entropy method applied to survey sampling

Fabrice Gamboa, Jean-Michel Loubes, Paul Rochet

Institut de Mathématiques de Toulouse, UMR 5219, Université Toulouse 3, 118 route de Narbonne F-31062 Toulouse Cedex 9, France

gamboa@math.univ-toulouse.fr, loubes@math.univ-toulouse.fr,
rochet@math.univ-toulouse.fr

Keywords: survey sampling, inverse problems, maximum entropy method

Introduction

Calibration methods have become increasingly studied in survey sampling over the last decades. By viewing calibration as an inverse problem, this article extends the calibration technique in the presence of complete auxiliary information by using a maximum entropy method (MEM). 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 enables to incorporate prior informations to the problem, giving a Bayesian interpretation to arbitrary settings existing in calibration.

Asymptotic properties of calibrated estimators are studied in a general framework by applying some other predominant methods in survey sampling like generalized calibration or instruments technique. We point out the relation between calibration and linear regression, and extend it to non-linear estimation. Optimality results are obtained using a generalization of the maximum entropy method, yielding a better asymptotic efficiency than classical calibration.

Maximum entropy on the mean method

Finding the solution to a calibration equation involves minimizing an energy with respect to a constraint. Let s be a random sample of size n , y be the variable of interest, x be the auxiliary variable whose information is summarized by a value t_x , and $d \in \mathbb{R}^n$ be the standard sampling weights equal to the inverse of the probabilities of the units being included in the sample. That is, for all $i \in s$, $d_i = 1/\pi_i$ where $\pi_i = p(i \in s) = \sum_{s, i \in s} p(s)$. For a chosen dissimilarity $\mathcal{D}(\cdot, \cdot)$, the weights $w \in \mathbb{R}^n$ are sought minimizing $\mathcal{D}(w, d)$ and such that $w'x = t_x$.

A typical dissimilarity choice 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. So the new estimator is defined as $\hat{t}_y = \frac{1}{N} \sum_{i \in s} \hat{w}_i y_i$, where the weights \hat{w}_i minimizes $\mathcal{D}(\cdot, d)$ under the constraint $\frac{1}{N} \sum_{i \in s} \hat{w}_i x_i = t_x$. For

more generality, dissimilarities of the form $w \mapsto \sum_{i \in s} \phi_i(\pi_i w_i)$, where $\phi_i, i \in s$ are strictly convex functions, are considered in the literature. Although, in the frame of calibration, the choice of the ϕ_i is arbitrary and MEM method aims at giving a Bayesian interpretation to it.

In the MEM point of view, the weights w are seen as expectation of random variables W , drawn from a finite measure μ^* close to a prior μ_0 . This prior distribution conveys the information that w must be close to d . For two finite measures P and Q , let $K(P, Q)$ denote the relative entropy of P with respect to Q , our calibration issue consists in finding the measure μ^* minimizing $K(\cdot, \mu_0)$ under the constraint that the calibration constraint holds in mean:

$$\mathbb{E}_{\mu^*} \left[\frac{1}{N} \sum_{i \in s} W_i x_i \right] = t_x.$$

Extending a result introduced in (Gamboa and Gassiat, 2009), we show the relation between the MEM approach and calibration through the following theorem.

Theorem *Assume that $\pi W = (\pi_i W_i)_{i \in s}$ has distribution $\otimes_{i \in s} \nu_i$ where for all $i \in s$, ν_i is a probability measure with mean 1, the MEM estimator $\hat{w} = \mathbb{E}_{\mu^*}(W)$ minimizes over \mathbb{R}^n the dissimilarity*

$$(w_1, \dots, w_n) \mapsto \sum_{i \in s} \Lambda_{\nu_i}^*(\pi_i w_i)$$

under the constraint $\frac{1}{N} \sum_{i \in s} \hat{w}_i x_i = t_x$, and where Λ_{ν}^ denotes the Cramer transform of ν .*

The MEM approach enables to give a probabilistic interpretation to calibration methods for a large class of dissimilarities, two of the main interpretations being the χ^2 dissimilarity which corresponds to a Gaussian prior and the Kullback dissimilarity $w \mapsto \sum_{i \in s} \log(w_i \pi_i) - w_i \pi_i + 1$, corresponding to an exponential prior.

Asymptotic properties of MEM estimators

The choice of the dissimilarity plays an important role in small sample estimation but is less crucial in an asymptotic framework since the resulting estimators are all asymptotically equivalent to the one obtained by using a χ^2 dissimilarity with a certain choice of smoothing parameters $q_i, i \in s$, as shown in (Deville and Särndal, 1992). More specifically, we have:

$$\hat{t}_y = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})' \hat{B} + o_{\mathbb{P}}(n^{-1/2}),$$

where $\hat{B} = \left[\sum_{i \in s} d_i q_i x_i x_i' \right]^{-1} \sum_{i \in s} d_i q_i y_i x_i$. Note that the vector \hat{B} is a generalized least square estimator for a regression of y in function of x . That is, $\hat{B}'x$ is a linear approximation of y by x . As a result, the variable $y - \hat{B}'x$ tends to have a lower variance than

y , and therefore, may permit to construct a more efficient estimate of t_y . The calibrated estimator \hat{t}_y can be seen as the Horvitz-Thompson estimator of a variable $\tilde{y} = y - \hat{B}'x$ that has a lower variance than y , up to a known additive constant (here $\hat{B}'t_x$) and an asymptotically negligible term.

Hence, we point out that calibration is heavily related to linear regression. When complete auxiliary information is available in the survey, this result can be extended to a non-linear statistic framework. Indeed, complete auxiliary information enables the users of the survey to modify the auxiliary variable x , considering instead an auxiliary variable of the form $u(x)$ for u a known real valued function. The choice of the constraint function u is another important aspect of the optimization problem. Set $t_u = \frac{1}{N} \sum_{i \in U} u(x_i)$, the calibration constraint becomes

$$\frac{1}{N} \sum_{i \in s} w_i u(x_i) = t_u.$$

The data $(y_i, x_i), i \in U$ are assumed to be i.i.d realizations of a random variable (Y, X) with distribution \mathbb{P} such that $\mathbb{E}(Y|X) \neq \mathbb{E}(Y)$. Under some conditions on the sampling design and under a suitable asymptotic framework, we calculate a lower bound for the variance of a calibration estimator of t_y .

Theorem *Let \hat{t}_y be a calibration estimator of t_y built with auxiliary variable $u(x_i), i \in U$. Then,*

$$n\text{var}(\hat{t}_y) \geq V^*(u) + o_{\mathbb{P}}(1).$$

as $n \rightarrow \infty$ and $n/N \rightarrow 0$, and where $V^*(u) = \text{var} \left(Y - \frac{\text{cov}(Y, u(X))}{\text{var}(u(X))} u(X) \right)$.

So, $V^*(u)$ is an asymptotic lower bound for the variance of a calibration estimator in the model with auxiliary variable $u(x)$. Furthermore, the functional $V^*(\cdot)$ reaches its minimum at $u : x \mapsto \mathbb{E}(Y|X = x)$, which conveys that the conditional expectation is the constraint function that provides the most efficient estimators. This points out that a linear relation between the variable of interest and the auxiliary variable is necessary to obtain optimality results.

Approximate maximum entropy on the mean

In practical cases, the conditional expectation function $\Phi(\cdot) = \mathbb{E}(Y|X = \cdot)$ is unknown. So, the last result can not be used directly to compute an efficient estimate of t_y . However, it can occur that we observe a sequence of functions $(\Phi_m)_{m \in \mathbb{N}}$ that converges towards Φ , for instance, non-parametric estimators of Φ derived from the data. The approximate maximum entropy method on the mean (AMEM) consists in applying the MEM approach to the approximate constraint function Φ_m . It was first introduced by

Loubes and Pelletier in a different framework in (Loubes and Pelletier, 2008). Set $\hat{t}_y(\Phi_m)$ the MEM estimator of t_y obtained with calibration constraint $\frac{1}{N} \sum_{i \in s} w_i \Phi_m(x_i) = t_{\Phi_m}$, we show under mild assumptions on the sequence $(\Phi_m)_{m \in \mathbb{N}}$ the following result.

Theorem *Let $\hat{t}_y(\Phi_m)$ be the calibration estimator of t_y built with auxiliary variable $\Phi_m(x_i), i \in U$, and dissimilarity $w \mapsto \sum_{i \in s} d_i(\pi_i w_i - 1)^2$. If $\mathbb{E}(\Phi_m(X) - \Phi(X))^2$ converges towards 0 as $m \rightarrow \infty$, then,*

$$\lim_{\substack{(n, N/n) \rightarrow \infty \\ m \rightarrow \infty}} n \text{var}(\hat{t}_y(\Phi_m)) = \inf_u V^*(u) = \mathbb{E}(Y - \Phi(X))^2.$$

Whilst remaining in the frame of calibration, the AMEM approach enables to increase the efficiency of the estimates. When Φ_m are non-parametric estimators of Φ derived from the observations $y_i, i \in s$ and $x_i, i \in U$, we see some examples where we recover classical non-parametric methods of estimation.

Maximum entropy method on the mean offers several advantages when applied to survey sampling. In regular calibration, the users need to arbitrarily choose a dissimilarity to use in the calibration equations, while the MEM method gives a probabilistic interpretation to it. So, it enables to incorporate prior informations to the problem with a Bayesian approach. Furthermore, in case of large sample estimation, the efficiency can be improved by extending the calibration to non-parametric estimation. This is made by considering approximate constraint functions in the calibration equations, which constitutes a direct application of approximate maximum entropy method on the mean.

References

- [1] Gamboa, F. and Gassiat, E. (1996), Sets of superresolution and the maximum entropy method on the mean, *Journal on Mathematical Analysis*, 27, 1129-1152.
- [2] Deville, J. C. and Särndal, C. E. (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376-382.
- [3] Loubes, J. M. and Pelletier, B. (2008), Maximum entropy solution to ill-posed inverse problems with approximately known operator, *Journal on Mathematical Analysis and Applications*, 344, 260-273.