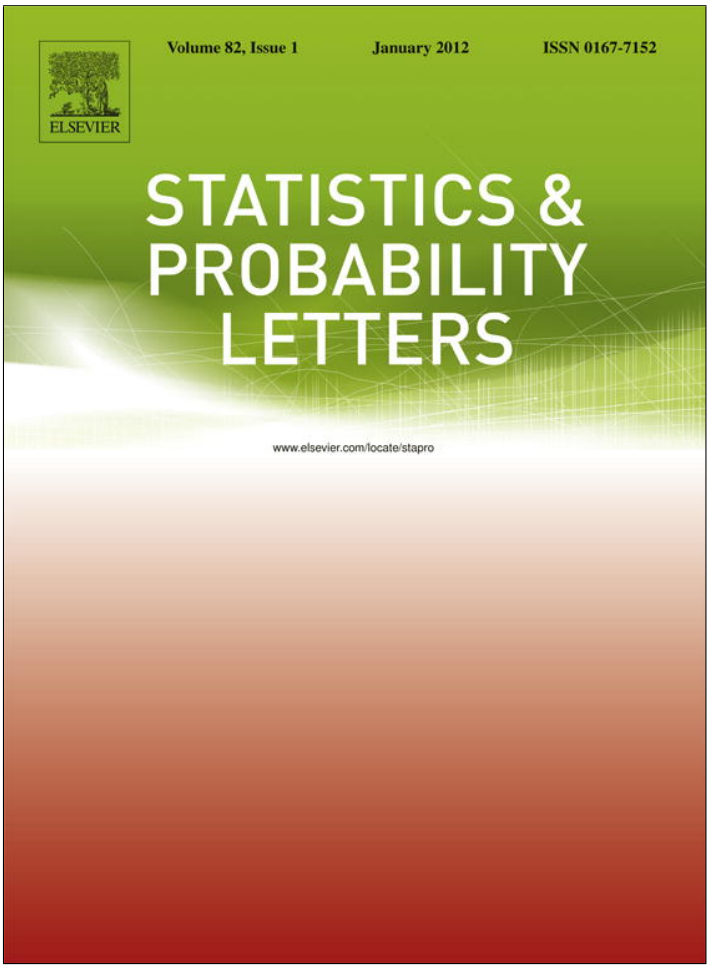

(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Statistics and Probability Letters 82 (2012) 972–978

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Approximate maximum entropy on the mean for instrumental variable regression

Jean-Michel Loubes [1], Paul Rochet *

Université Paul Sabatier Toulouse III, 118 Route de Narbonne, 31068 Toulouse, France

ARTICLE INFO

Article history:
Received 7 June 2011
Received in revised form 8 February 2012
Accepted 8 February 2012
Available online 18 February 2012

Keywords:
Approximate maximum entropy
Inverse problem

ABSTRACT

We want to estimate an unknown finite measure μ_X from a noisy observation of generalized moments of μ_X, defined as the integral of a continuous function Φ with respect to μ_X. Assuming that only a quadratic approximation Φ_m is available, we define an approximate maximum entropy solution as a minimizer of a convex functional subject to a sequence of convex constraints. We establish asymptotic properties of the approximate solution under regularity assumptions on the convex functional, and we study an application of this result to instrumental variable estimation.

1. Introduction

We tackle the inverse problems of reconstructing an unknown finite measure μ_X on a set $\mathcal{X} \subset \mathbb{R}^d$, from observations of generalized moments of μ_X,

$$y = \int_{\mathcal{X}} \Phi(x) d\mu_X(x),$$

where $\Phi : \mathcal{X} \to \mathbb{R}^k$ is a given map. This problem has been notably studied in Econometrics models involving endogenous variables, that can be stated in the question of recovering a function g from observations

$$Y = g(X) + U,$$

where the centered noise U is correlated with the explanatory variable X, i.e. $\mathbb{E}(U|X) \neq 0$. Such problems can be solved using an auxiliary variable W that is correlated with X but uncorrelated with the noise U. Indeed, the observation of a variable W satisfying $\mathbb{E}(U|W) = 0$ provides some information on the function g in the form of linear constraints that can be used to estimate g. Some of the main references on this topic are Chamberlain (1987), Hansen (1982) and Owen (1991).

The problem of recovering the unknown measure μ_X is said to be *ill-posed*, in particular, because a solution to the equation $y = \int \Phi \, d\mu_X$ is not unique. For inverse problems with known operator Φ, regularization techniques have been implemented in order to turn the problem into a convex optimization program for which a solution is uniquely defined. Precisely, a solution is obtained as the minimizer of a convex functional $\nu \mapsto J(\nu)$ subject to the linear constraint $\int \Phi d\nu = y$ when y is observed, or more generally, subject to a convex constraint of the form $\int \Phi d\nu \in K_Y$ in presence of noise, for some convex set K_Y. Several types of regularizing functionals have been introduced in the literature. In this general setting, the inversion procedure is deterministic, i.e. the noise distribution is not used in the definition of the regularized solution.

* Corresponding author. Tel.: +33 561 55 63 71; fax: +33 561 55 60 89.
E-mail addresses: loubes@math.univ-toulouse.fr (J.-M. Loubes), rochet@math.univ-toulouse.fr (P. Rochet).
[1] Tel.: +33 561 55 85 73; fax: +33 561 55 60 89.

0167-7152/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2012.02.006

Bayesian approaches to inverse problems allow one to handle the noise distribution, provided it is known, yet in general, a distribution like the normal distribution is postulated (see Evans and Stark (2002) for a survey). However in many real-world inverse problems, the noise distribution is unknown, and only the output y is easily observable, contrary to the input to the operator. Consequently very few paired data are available to reliably estimate the noise distribution, thereby causing robustness deficiencies on the retrieved parameters. Nonetheless, even if the noise distribution is unavailable to the practitioner, she often knows the noise level, i.e. the maximal magnitude of the disturbance term, say $\eta > 0$, and this information may be reflected by taking a constraint set K_Y of diameter 2η.

We focus on a regularization functional with grounding in information theory, leading to maximum entropy solutions to the inverse problem. The method, known as *maximum entropy on the mean* (MEM), provides a very simple and natural manner to incorporate constraints on the support and the range of the solution, as discussed in Gamboa and Gassiat (1997). In a deterministic framework, maximum entropy solutions have been studied in Borwein et al. (2003), Borwein and Lewis (1991), while some other studies exist in a Bayesian setting (Gamboa, 1999; Gamboa and Gassiat, 1997), in seismic tomography (Fermin et al., 2006), in image analysis (Gzyl and Zeev, 2002) and in survey sampling (Gamboa et al., 2011).

In many actual situations, the map Φ is unknown and only an approximation Φ_m is available. In this paper, we introduce an approximate maximum entropy on the mean (AMEM) estimate $\hat{\mu}_{m,n}$ of the measure μ_X to be reconstructed. This estimate is expressed in the form of a discrete measure concentrated on n points of $\mathcal{X}$. In our main result, we prove that the convergence in $\mathbb{L}^2$-norm of the sequence $\{\Phi_m\}_{m\in\mathbb{N}}$ toward Φ is sufficient to ensure the weak convergence at an explicit rate of the estimator $\hat{\mu}_{m,n}$ to the solution of the initial inverse problem as $m \to \infty$ and $n \to \infty$. Moreover, this approximate framework can be encountered when dealing with instrumental variables in Econometrics and we will provide a new estimation procedure in this setting.

The paper is organized as follows. Section 2 introduces some notations and the definition of the AMEM estimate. We state our main result (Theorem 3.1) in Section 3 and an application to instrumental variables is studied in Section 4. The Appendix is devoted to the proofs of our results.

2. The AMEM estimate

Let Φ be a continuous and bounded map defined on a subset $\mathcal{X}$ of $\mathbb{R}^d$ and taking values in $\mathbb{R}^k$. We note $\mathcal{B}(\mathcal{X})$ the Borel σ-field of $\mathcal{X}$ and $\mathcal{M}(\mathcal{X})$ the set of finite measures on $\mathcal{X}$. Let $\mu_X \in \mathcal{M}(\mathcal{X})$ be an unknown measure satisfying the constraint $y = \int \Phi d\mu_X$. Assume we observe a perturbed version y^{obs} of y:

$$y^{obs} = \int_{\mathcal{X}} \Phi(x) d\mu_X(x) + \varepsilon, \tag{1}$$

where ε is an error term supposed bounded in norm from above by some positive constant η, representing the maximal noise level. Such problems are encountered in various fields of sciences, like medical imaging, time-series analysis, speech processing, image denoising, spectroscopy, geophysical sciences, crystallography, and tomography, see for example Decarreau et al. (1992), Hermann and Noll (2000), and Skilling (1986). This problem has also been extensively studied in the literature in Econometrics and more specifically in instrumental variable models that can be formalized using linear moment constraints as in Chamberlain (1987), Hansen (1982) and Owen (1991).

As an alternative to standard regularization methods such as Tikhonov and Galerkin (see for instance Engl et al. (1996)), we aim at reconstructing the measure μ_X with a maximum entropy procedure. In image analysis this measure may be viewed as the intensity at each pixel of the image, blurred by an unknown filter. Other applications in seismic tomography can be found in Fermin et al. (2006), while we discuss an application to Econometrics in Section 4.

Let us introduce some definitions and notations. For two probability measures ν, μ, we define the relative entropy of ν with respect to μ by

$$H(\nu|\mu) = \int \log\left(\frac{d\nu}{d\mu}\right) d\nu \quad \text{if } \nu \ll \mu, \qquad H(\nu|\mu) = +\infty \quad \text{otherwise.}$$

Although the relative entropy defines a notion of proximity between measures, this quantity is not a distance (in particular it is not symmetric). We denote by K_Y the closed ball of $\mathbb{R}^k$ centered at the observation y^{obs} and of radius η. For P a measure and g a function, we shall use the notation $Pg = \int g dP$. The true measure μ_X is known to satisfy the moment condition $\mu_X\Phi \in K_Y$, however, the map Φ being unknown, we consider the approximate moment condition

$$\int_{\mathcal{X}} \Phi_m(x) d\mu_X(x) \in K_Y. \tag{2}$$

Let us now explain the construction of the AMEM estimator. Let $X_1, \ldots, X_n$ be a discretization of the space $\mathcal{X}$, for which the associated empirical measure $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is assumed to converge weakly to some distribution $\mathbb{P}_X$ having full support on $\mathcal{X}$. The X_i's may be i.i.d. realizations of a random variable X with distribution $\mathbb{P}_X$, or a deterministic design, in which case

$\mathbb{P}_X$ is known by the statistician. We search for an estimator of μ_X which can be written as a weighted version of the empirical measure $\mathbb{P}_n$:

$$L_n(z) := \frac{1}{n}\sum_{i=1}^{n} z_i \delta_{X_i},$$

for some vector $z = (z_1, \ldots, z_n)' \in \mathbb{R}^n$. The objective is to find a suitable vector of weights z for which the associated discrete measure $L_n(z)$ is a good estimation of μ_X. The problem of estimating μ_X is then turned into a parametric problem where the parameter of interest z is of dimension n.

Let $Z = (Z_1, \ldots, Z_n)'$ be a vector of n i.i.d. realizations drawn from a measure ν_Z and consider the random measure $L_n(Z)$. From a Bayesian point of view, the measure $\nu_Z^{\otimes n}$ can be interpreted as a *prior* distribution on the parameter z. Define ν^* as the probability measure minimizing the relative entropy $H(\cdot|\nu_Z^{\otimes n})$ under the constraint that the approximate moment condition (2) holds in mean,

$$\mathbb{E}_{\nu^*}\left[\int_{\mathcal{X}} \Phi_m dL_n(Z)\right] = \frac{1}{n}\sum_{i=1}^{n} \Phi_m(X_i)\mathbb{E}_{\nu^*}(Z_i) \in K_Y.$$

The measure ν^* may be seen as an *a posteriori* distribution from a Bayesian point of view. The estimator $\hat{\mu}_{m,n}$ is obtained as the expectation of $L_n(Z)$ under ν^*,

$$\hat{\mu}_{m,n} = \mathbb{E}_{\nu^*}[L_n(Z)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\nu^*}(Z_i)\delta_{X_i}.$$

The existence of ν^* requires the existence of a vector z_0 in the convex hull of the support of $\nu_Z^{\otimes n}$ such that $\int \Phi_m dL_n(z_0) \in K_Y$. It is shown in Loubes and Pelletier (2008) that under the Assumptions of Theorem 3.1, this condition tends to be verified with probability 1 as $m \to \infty$ and $n \to \infty$. Hence for m and n large enough, the AMEM estimate $\hat{\mu}_{m,n}$ is well defined with high probability, and asymptotically with probability 1.

3. Convergence of the AMEM estimate

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

$$\Lambda_\nu(s) = \log \int_{\mathbb{R}} e^{sx} d\nu(x) \quad \text{and} \quad \Lambda_\nu^*(s) = \sup_{u\in\mathbb{R}}\{su - \Lambda_\nu(u)\}, \quad s \in \mathbb{R}.$$

The log-Laplace transform Λ_ν is a twice differentiable convex function and Λ_ν^* is called its *convex conjugate*. Define the functional

$$I_{\nu_Z}(\mu|\mathbb{P}_X) = \int_{\mathcal{X}} \Lambda_{\nu_Z}^*\left(\frac{d\mu}{d\mathbb{P}_X}\right) d\mathbb{P}_X \quad \text{if } \mu \ll \mathbb{P}_X, \qquad I_{\nu_Z}(\mu|\mathbb{P}_X) = +\infty \quad \text{otherwise.}$$

The quantity $I_{\nu_Z}(\mu|\mathbb{P}_X)$ is called the *f-divergence* of μ with respect to $\mathbb{P}_X$ associated to the convex function $\Lambda_{\nu_Z}^*$. The notion of f-divergence was introduced by Csiszár as a generalization of the relative entropy. We refer to Csiszár (1967) for more details.

We note by $\mathcal{C}_b$ the set of continuous bounded functions on $\mathcal{X}$. For all $g \in \mathcal{C}_b$, we denote by $|\cdot|_g$ the semi-norm defined for $\mu \in \mathcal{M}(\mathcal{X})$ by $|\mu|_g = \left|\int g d\mu\right|$. We recall that the family of semi-norms $\{|\cdot|_g, g \in \mathcal{C}_b\}$ defines the weak topology: a sequence $\{\mu_n\}_{n\in\mathbb{N}}$ converges weakly toward μ if, and only if, $\lim_{n\to\infty} |\mu_n - \mu|_g = 0$, for all $g \in \mathcal{C}_b$.

We make the following assumptions.

A1. The minimization problem admits at least one solution, i.e. there exists a continuous function g_0 taking values in the convex hull of the support of ν_Z such that $\int \Phi g_0 \, d\mathbb{P}_X \in K_Y$.

A2. The function Λ_{ν_Z}'' is bounded.

A3. The approximating sequence Φ_m converges to Φ in $\mathbb{L}^2(\mathbb{P}_X)$. Its rate of convergence is given by

$$\|\Phi_m - \Phi\|_{\mathbb{L}^2} := \sqrt{\mathbb{E}\|\Phi_m(X) - \Phi(X)\|^2} = O(\varphi_m^{-1}),$$

for some growing sequence $\{\varphi_m\}_{m\in\mathbb{N}}$.

A4. The function $G : x \mapsto \sup_{m\in\mathbb{N}} \|\Phi_m(x)\|$ is square integrable: $\int G^2 d\mathbb{P}_X < \infty$.

A5. For all $m \in \mathbb{N}$, the components of Φ_m are linearly independent.

We are now in a position to state our main result.

Theorem 3.1 (*Convergence of the AMEM Estimate*). *Suppose that* A1 *and* A2 *hold and let* μ_X^* *be the minimizer of the functional* $\mu \mapsto I_{\nu_Z}(\mu|\mathbb{P}_X)$ *subject to the constraint* $\int \Phi d\mu \in K_Y$.

- *The AMEM estimate* $\hat{\mu}_{m,n}$ *is given by*

$$d\hat{\mu}_{m,n}(x) = \Lambda'_{\nu_Z}(\langle \hat{v}_{m,n}, \Phi_m(x)\rangle) d\mathbb{P}_n(x),$$

where $\hat{v}_{m,n}$ *minimizes over* $\mathbb{R}^k$, $H_{m,n}(v) = \mathbb{P}_n \Lambda_{\nu_Z}(\langle v, \Phi_m\rangle) - \inf_{y\in K_Y}\langle v, y\rangle$.
- *If* A3–A5 *also hold,* $\hat{\mu}_{m,n}$ *converges weakly to* μ_X^* *as* $m, n \to \infty$ *and its rate of convergence is expressed as follows,*

$$\forall g \in \mathcal{C}_b, \quad |\hat{\mu}_{m,n} - \mu_X^*|_g = O(\varphi_m^{-1}) + \kappa_{m,n},$$

with $\sup_{m\in\mathbb{N}} \kappa_{m,n} = O_P(n^{-1/2})$.

The condition A2 is a rather strong requirement on the choice of the prior ν_Z. It is equivalent to assuming that Λ_{ν_Z} is dominated by a quadratic function. This condition is satisfied, for instance, for Gaussian priors or if ν_Z has compact support. As a result, the function $H : v \mapsto \mathbb{P}_X \Lambda_{\nu_Z}(\langle v, \Phi\rangle) - \inf_{y\in K_Y}\langle v, y\rangle$ attains its minimum at a unique point v^* belonging to the interior of its domain $\mathbb{R}$. If this assumption is not met, it is shown in Borwein and Lewis (1993) and Gamboa and Gassiat (1997) that the minimizers of $I_{\nu_Z}(\cdot|\mathbb{P}_X)$ over the set of finite measures satisfying the moment constraint may have a singular part with respect to $\mathbb{P}_X$.

The construction of the AMEM estimate relies on a discretization of the space $\mathcal{X}$ according to the probability $\mathbb{P}_X$. Therefore by varying the support of $\mathbb{P}_X$, the practitioner may easily incorporate some *a priori* knowledge concerning the support of the solution. Similarly, the AMEM estimate also depends on the measure ν_Z, which determines the domain of $\Lambda^*_{\nu_Z}$, and so the range of the solution.

4. Application to instrumental variable estimation

A natural field of application is given by nonparametric regression models involving instrumental variables. This kind of problem has been extensively studied in the literature in Econometrics, we refer, for instance, to Florens (2003), Hansen and Singleton (1982) and Newey (1990). In some cases, the instrumental variable estimation framework can be viewed as an inverse problem with unknown operator that can be solved using the AMEM procedure.

Let $X_1, \ldots, X_n$ be here a discretization of the space $\mathcal{X}$ such that the associated empirical distribution $\mathbb{P}_n$ converges weakly toward a known distribution $\mathbb{P}_X$ having full support on $\mathcal{X}$. Let $g : \mathcal{X} \to \mathbb{R}_+$ be an unknown function for which we observe a noisy evaluation at each point X_i,

$$Y_i = g(X_i) + U_i, \quad i = 1, \ldots, n,$$

where the U_i's are centered real valued random variables. Contrary to the classical regression framework, we suppose here that the noises U_i are correlated with the X_i's (i.e. $\mathbb{E}(U_i|X_i) \neq 0$), which causes identification issues. This kind of model is used, for instance, to deal with simultaneous causality between supply and demand in economic markets. Assume that we want to estimate nonparametrically the price Y of a good with respect to its production X, the noise U in the corresponding model turns out to be correlated with X due to the mutual influence between the price and the production. To overcome this difficulty, econometricians assume there exist *instrumental variables*, that affect the price only through the produced quantity (for example, the amount of rain in the case of an agricultural product). Hence, we assume we observe simultaneously with (X_i, Y_i), an additional variable $W_i \in \mathbb{R}^k$ such that $\mathbb{E}(W_i|X_i) \neq 0$ and $\mathbb{E}(U_i|W_i) = 0$. In particular, we have the relation

$$y := \mathbb{E}(WY) = \mathbb{E}(Wg(X)). \tag{3}$$

In most cases, using the instrumental variable W is not sufficient to solve the identification issue, but it still provides some information that may be rendered in the form of linear constraints on g. Indeed, setting $\Phi : x \mapsto \mathbb{E}(W|X = x)$ and $d\mu_X(x) = g d\mathbb{P}_X(x)$, $x \in \mathcal{X}$, the Eq. (3) can be written as

$$y = \int \Phi(x) d\mu_X(x).$$

Here, y is unknown but we observe a noisy version $y^{obs} = n^{-1}\sum_{i=1}^n W_i X_i$ that is close to y with high probability and asymptotically with probability one. The conditional expectation Φ is also unknown but can be estimated from the data by nonparametric procedures, yielding a converging sequence $\{\Phi_n\}$. As a result, it is possible to estimate the measure μ_X by the AMEM procedure, considering an approximate moment condition of the form $\int \Phi_n d\mu \in K_Y$. We obtain a sequence of estimators $\hat{\mu}_n$, which is shown in Theorem 3.1 to converge weakly toward the minimizer μ_X^* of the convex functional $I_{\nu_Z}(\cdot|\mathbb{P}_X)$ subject to the moment constraint. Equivalently, the method ensures the convergence in a weak sense of the density $\hat{g} = d\hat{\mu}_n/d\mathbb{P}_n$ of the AMEM estimator toward the function $g^* := d\mu_X^*/d\mathbb{P}_X$. In particular, the identification issue on g is solved by incorporating some *a priori* knowledge on μ_X through the choice of the design $X_1, \ldots, X_n$ and the limit distribution $\mathbb{P}_X$.

Acknowledgment

We are very grateful to an anonymous referee for his helpful remarks.

Appendix

A.1. Technical lemmas

We use the following notations

$$v_m^* = \operatorname*{argmin}_{v\in\mathbb{R}^k} H_m(v) = \operatorname*{argmin}_{v\in\mathbb{R}^k} \left\{ \mathbb{P}_X \Lambda_{\nu_Z}(\langle \Phi_m, v\rangle) - \inf_{y\in K_Y} \langle v, y\rangle \right\},$$

$$\hat{v}_{m,n} = \operatorname*{argmin}_{v\in\mathbb{R}^k} H_{m,n}(v) = \operatorname*{argmin}_{v\in\mathbb{R}^k} \left\{ \mathbb{P}_n \Lambda_{\nu_Z}(\langle \Phi_m, v\rangle) - \inf_{y\in K_Y} \langle v, y\rangle \right\},$$

$$v^* = \operatorname*{argmin}_{v\in\mathbb{R}^k} H(v) = \operatorname*{argmin}_{v\in\mathbb{R}^k} \left\{ \mathbb{P}_X \Lambda_{\nu_Z}(\langle \Phi, v\rangle) - \inf_{y\in K_Y} \langle v, y\rangle \right\}.$$

Lemma A.1. *If Assumptions* 1–5 *hold,*

$$\sup_{m\in\mathbb{N}} \|\hat{v}_{m,n} - v_m^*\| = O_P(n^{-1/2}).$$

Proof. For all $x \in \mathcal{X}$, $v \in \mathbb{R}^k$, set

$$h_m(v, x) = \Lambda_{\nu_Z}(\langle \Phi_m(x), v\rangle) - \inf_{y\in K_Y} \langle v, y\rangle.$$

The parameter $\hat{v}_{m,n}$ is defined as the minimizer of the empirical contrast function $v \mapsto H_{m,n}(v) = \mathbb{P}_n h_m(v, \cdot)$. To prove the result, we need to show that $h_m(v, x)$ satisfies the conditions of Corollary 5.53 in van der Vaart (1998). First remark that $H_{m,n}$ is convex, which ensures the convergence in probability of its minimizer $\hat{v}_{m,n}$ toward v_m^*. Since K_Y is the ball centered in y^{obs} and of radius η, we may write

$$h_m(v, x) = \Lambda_{\nu_Z}(\langle \Phi_m(x), v\rangle) - \langle v, y^{obs}\rangle + \eta\|v\|.$$

By A2, we know there exists a $K > 0$ such that $\Lambda'_{\nu_Z}(s) \le Ks + 1$ for all $s \in \mathbb{R}$. For all v_1, v_2 in a neighborhood $\mathcal{N}$ of v_m^*, we have by the triangular inequality and the mean value theorem

$$\begin{aligned} |h_m(v_1, \cdot) - h_m(v_2, \cdot)| &\le \left|\Lambda_{\nu_Z}(\langle \Phi_m, v_1\rangle) - \Lambda_{\nu_Z}(\langle \Phi_m, v_2\rangle)\right| + \left|\langle v_1 - v_2, y^{obs}\rangle + \eta\,|\|v_1\| - \|v_2\||\right| \\ &\le \left[K\|v_2\|\,\|\Phi_m\| + 1 + \|y^{obs}\| + \eta\right]\|v_1 - v_2\| \\ &\le \left[K\delta\, G + 1 + \|y^{obs}\| + \eta\right]\|v_1 - v_2\|, \end{aligned}$$

where G is the function defined in A4 and where we set $\delta = \sup_{v\in\mathcal{N}} \|v\|$. Since v_m^* converges toward v^*, we may assume, without loss of generality, that $\mathcal{N}$ and δ are fixed for m sufficiently large. Hence the function h_m satisfies the first condition of Corollary 5.53 in van der Vaart (1998),

$$|h_m(v_1, \cdot) - h_m(v_2, \cdot)| \le \dot{h}\|v_1 - v_2\|,$$

where $\dot{h} : x \mapsto K\delta\, G(x) + 1 + \|y^{obs}\| + \eta$ does not depend and m and is such that $P_X\dot{h}^2 < \infty$. For all $v \in \mathbb{R}^k$, let $V_m(v)$ be the Hessian matrix of H_m at point v, which is well defined for all $v \neq 0$. Assume that $v_m^* \neq 0$, we need to prove that $V_m(v_m^*)$ is non-negative definite. The case $v_m^* = 0$ can be treated separately without difficulty using Theorem 5.52 in van der Vaart (1998), by considering the derivative at 0^+ of the functions $t \mapsto V_m(tv)$, $v \in \mathbb{R}^k$. Let ∂_i be the derivative with respect to the i-th component. For $v \neq 0$, we have

$$\begin{aligned} [V_m(v)]_{ij} = \partial_i\partial_j H_m(v) &= \mathbb{P}_X[\partial_i\partial_j h_m(v, \cdot)] \\ &= \mathbb{P}_X[\Phi_m^i \Phi_m^j \Lambda''_{\nu_Z}(\langle \Phi_m, v\rangle)] + \eta\ \partial_i\partial_j N(v) \end{aligned}$$

where we set $N : v \mapsto \|v\|$. Thus, $V_m(v_m^*)$ can be split into the sum $A_m + \eta B_m$, with

$$(A_m)_{ij} = P_X[\Phi_m^i \Phi_m^j \Lambda''_{\nu_Z}(\langle \Phi_m, v_m^*\rangle)], \qquad (B_m)_{ij} = \partial_i\partial_j N(v_m^*).$$

A_m is a Gram matrix, therefore it is positive definite, by A5. Moreover, since the A_m converge toward the positive-definite matrix $A = (P_X[\Phi^i\Phi^j\Lambda''_{\nu_Z}(\langle \Phi, v^*\rangle)])_{1\le i,j\le k}$, we conclude there exist an integer M and a constant $c > 0$ such that, for all $a \in \mathbb{R}^k$,

$$\inf_{m\ge M} a^T A_m a \ge c\|a\|^2.$$

By convexity of the map $N(\cdot)$ on $\mathbb{R}^k \setminus \{0\}$, the matrix B_m is non-negative definite and so is $V_m(v_m^*) = A_m + \eta B_m$. Hence, H_m undergoes the assumptions of Corollary 5.53 in van der Vaart (1998), uniformly for $m \in \mathbb{N}$, which proves the result. □

Lemma A.2. *If Asssumptions* 1–5 *hold,*

$$\|v_m^* - v^*\| = O(\varphi_m^{-1}).$$

Proof. Using successively the mean value theorem and Cauchy–Schwarz's inequality, we find

$$\begin{aligned}|H_m(v) - H(v)| &= |\mathbb{P}_X[\Lambda_{\nu_Z}(\langle\Phi_m, v\rangle) - \Lambda_{\nu_Z}(\langle\Phi(x), v\rangle)]| \\ &\leq (K\|v\|^2\,\|G\|_{\mathbb{L}^2} + \|v\|)\|\Phi_m - \Phi\|_{\mathbb{L}^2}.\end{aligned}$$

We deduce that H_m converges uniformly on every compact set toward H as $m \to \infty$. By convexity of H_m, this warrants the convergence of v_m^* toward v^*. Moreover,

$$\begin{aligned}\nabla H_m(v) - \nabla H(v) &= \mathbb{P}_X\left[\Phi_m \Lambda'_{\nu_Z}(\langle\Phi_m, v\rangle) - \Phi\,\Lambda'_{\nu_Z}(\langle\Phi, v\rangle)\right] \\ &= \mathbb{P}_X\left[(\Phi_m - \Phi)\Lambda'_{\nu_Z}(\langle\Phi_m, v\rangle) + \Phi[\Lambda'_{\nu_Z}(\langle\Phi_m, v\rangle) - \Lambda'_{\nu_Z}(\langle\Phi, v\rangle)]\right].\end{aligned}$$

In the same way as previously, we find

$$\|\nabla H_m(v) - \nabla H(v)\| \leq \|\Phi_m - \Phi\|_{\mathbb{L}^2}\|v\|\left(K\|\Phi_m\|_{\mathbb{L}^2} + 1 + \|\Phi\|_{\mathbb{L}^2}\|\Lambda''_{\nu_Z}\|_\infty\right),$$

which proves that ∇H_m converges toward ∇H, uniformly on every compact set. Noticing that $\nabla H(v_m^*) = \nabla H(v_m^*) - \nabla H_m(v_m^*)$, it follows that $|\nabla H(v_m^*)| = O(\varphi_m^{-1})$. Note $V(v^*)$ the Hessian matrix of H at v^*. We know it is positive definite by a similar reasoning as in the proof of Lemma A.1. Writing the Taylor expansion

$$\nabla H(v_m^*) = V(v^*)(v^* - v_m^*) + o(\|v^* - v_m^*\|),$$

we conclude $\|v^* - v_m^*\| = O(\varphi_m^{-1})$. □

A.2. Proof of Theorem 3.1

The first part of the theorem is proved in Theorem 3.1 in Loubes and Pelletier (2008). We here focus on the proof of the second part. We use the following notations

$$\hat{\mu}_{m,n} = \Lambda'_{\nu_Z}(\langle\hat{v}_{m,n}, \Phi_m\rangle)\mathbb{P}_n \quad \text{and} \quad \mu_m^* = \Lambda'_{\nu_Z}(\langle v_m^*, \Phi_m\rangle)\mathbb{P}_X.$$

For $g \in \mathcal{C}_b$, write $\left|\hat{\mu}_{m,n} - \mu_X^*\right|_g \leq \left|\hat{\mu}_{m,n} - \mu_m^*\right|_g + \left|\mu_m^* - \mu_X^*\right|_g$. We shall bound each term separately. We have

$$\begin{aligned}\left|\hat{\mu}_{m,n} - \mu_m^*\right|_g &= \left|\Lambda'_{\nu_Z}(\langle\Phi_m, \hat{v}_{m,n}\rangle)\mathbb{P}_n - \Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_X\right|_g \\ &\leq \left|\Lambda'_{\nu_Z}(\langle\Phi_m, \hat{v}_{m,n}\rangle)\mathbb{P}_n - \Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_n\right|_g + \left|\Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_n - \Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_X\right|_g.\end{aligned}$$

We obtain for all $x \in \mathcal{X}$,

$$|\Lambda'_{\nu_Z}(\langle\Phi_m(x), \hat{v}_{m,n}\rangle) - \Lambda'_{\nu_Z}(\langle\Phi_m(x), v_m^*\rangle)| \leq \|\Lambda''_{\nu_Z}\|_\infty\|\Phi_m(x)\|\,\|\hat{v}_{m,n} - v_m^*\|,$$

by Cauchy–Schwarz's inequality. We get

$$\left|\Lambda'_{\nu_Z}(\langle\Phi_m, \hat{v}_{m,n}\rangle)\mathbb{P}_n - \Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_n\right|_g \leq \|g\|_\infty\|\Lambda''_{\nu_Z}\|_\infty\|\hat{v}_{m,n} - v_m^*\|\,\mathbb{P}_n G.$$

Using Slutsky's lemma and Lemma A.1, we conclude

$$\sup_{m\in\mathbb{N}}\left|\Lambda'_{\nu_Z}(\langle\Phi_m, \hat{v}_{m,n}\rangle)\mathbb{P}_n - \Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_n\right|_g = O_P(n^{-1/2}).$$

The rate of convergence of the term $\left|\Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_n - \Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_X\right|_g$ follows directly from the uniform law of large numbers. We obtain

$$\sup_{m\in\mathbb{N}}\left|\hat{\mu}_{m,n} - \mu_m^*\right|_g = O_P(n^{-1/2}).$$

The second step is to bound the term $\left|\mu_m^* - \mu_X^*\right|_g$. We follow the same guidelines,

$$\begin{aligned}\left|\mu_m^* - \mu_X^*\right|_g &= \left|\Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_X - \Lambda'_{\nu_Z}(\langle\Phi, v^*\rangle)\mathbb{P}_X\right|_g \\ &\leq \left|\Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_X - \Lambda'_{\nu_Z}(\langle\Phi_m, v^*\rangle)\mathbb{P}_X\right|_g + \left|\Lambda'_{\nu_Z}(\langle\Phi_m, v^*\rangle)\mathbb{P}_X - \Lambda'_{\nu_Z}(\langle\Phi, v^*\rangle)\mathbb{P}_X\right|_g.\end{aligned}$$

In the same way as previously, the first term is bounded as follows

$$\left|\Lambda'_{\nu_Z}(\langle\Phi_m, v_m^*\rangle)\mathbb{P}_X - \Lambda'_{\nu_Z}(\langle\Phi_m, v^*\rangle)\mathbb{P}_X\right|_g \leq \|\Lambda''_{\nu_Z}\|_\infty\|g\|_\infty\,\mathbb{E}\|\Phi_m(X)\|\,\|v_m^* - v^*\|,$$

which is shown to be of order $O(\varphi_m^{-1})$ in Lemma A.2. For the last term, we have in the same way

$$\left|\Lambda'_{\nu_Z}(\langle\Phi_m, v^*\rangle)\mathbb{P}_X - \Lambda'_{\nu_Z}(\langle\Phi, v^*\rangle)\mathbb{P}_X\right|_g \leq \|v^*\|\,\|\Lambda''_{\nu_Z}\|_\infty\,\|g\|_\infty\,\mathbb{E}\|\Phi_m(X) - \Phi(X)\|.$$

Regrouping all the terms, we get

$$\left|\hat{\mu}_{m,n} - \mu_X^*\right|_g = \kappa_{m,n} + O(\varphi_m^{-1}),$$

where $\kappa_{m,n} \leq \left|\hat{\mu}_{m,n} - \mu_m^*\right|_g$ satisfies $\sup_{m\in\mathbb{N}} \kappa_{m,n} = O_P(n^{-1/2})$.

References

Borwein, J.M., Lewis, A.S., 1991. Duality relationships for entropy-like minimization problems. SIAM J. Control Optim. 29, 325–338.
Borwein, J.M., Lewis, A.S., 1993. Partially-finite programming in L_1 and the existence of maximum entropy estimates. SIAM J. Optim. 3, 248–267.
Borwein, J.M., Lewis, A.S., Noll, D., 2003. Maximum entropy reconstruction using derivative information, I. Fisher information and convex duality. Math. Oper. Res. 21, 442–468.
Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. J. Econometrics 34, 305–334.
Csiszár, I., 1967. On topology properties of f-divergences. Studia Sci. Math. Hungar. 2, 329–339.
Decarreau, A., Hilhorst, D., Lemaréchal, C., Navaza, J., 1992. Dual methods in entropy maximization. Application to some problems in crystallography. SIAM J. Optim. 2, 173–197.
Engl, H.W., Hanke, M., Neubauer, A., 1996. Regularization of Inverse Problems. Kluwer Academic Publishers Group, Dordrecht.
Evans, S.N., Stark, P.B., 2002. Inverse problems as statistics. Inverse Problems 18, R55–R97.
Fermin, A.K., Loubes, J.M., Ludeña, C., 2006. Bayesian methods for a particular inverse problem seismic tomography. Int. J. Tomogr. Stat. 4, 1–19.
Florens, J.P., 2003. Inverse Problems and Structural Econometrics: The Example of Instrumental Variables, vol. 36. Cambridge University Press.
Gamboa, F., 1999. New Bayesian methods for ill posed problems. Statist. Decisions 17, 315–337.
Gamboa, F., Gassiat, E., 1997. Bayesian methods and maximum entropy for ill-posed inverse problems. Ann. Statist. 25, 328–350.
Gamboa, F., Loubes, J.M., Rochet, P., 2011. Maximum entropy estimation for survey sampling. J. Statist. Plann. Inference 141, 305–317.
Gzyl, H., Zeev, N., 2002. Probabilistic approach to an image reconstruction problem. Methodol. Comput. Appl. Probab. 4, 279–290.
Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.
Hansen, L.P., Singleton, K.J., 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. Econometrica 50, 1269–1286.
Hermann, U., Noll, D., 2000. Adaptive image reconstruction using information measures. SIAM J. Control Optim. 38, 1223–1240.
Loubes, J.M., Pelletier, B., 2008. Maximum entropy solution to ill-posed inverse problems with approximately known operator. J. Math. Anal. Appl. 344, 260–273.
Newey, W.K., 1990. Efficient instrumental variables estimation of nonlinear models. Econometrica 58, 809–837.
Owen, A., 1991. Empirical likelihood for linear models. Ann. Statist. 19, 1725–1747.
Skilling, J., 1986. Maximum-Entropy and Bayesian Methods in Science and Engineering. Kluwer Acad. Publ., Dordrecht.
van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press, Cambridge.