Bayesian interpretation of generalized empirical likelihood by maximum entropy

Paul Rochet

Abstract

We study a parametric estimation problem related to moment condition models. As an alternative to the generalized empirical likelihood (GEL) and the generalized method of moments (GMM), a Bayesian approach to the problem can be adopted, extending the MEM procedure to parametric moment conditions. We show in particular that a large number of GEL estimators can be interpreted as a maximum entropy solution. Moreover, we provide a more general field of applications by proving the method to be robust to approximate moment conditions.

1 Introduction

We consider a parametric estimation problem in a moment condition model. Assume we observe an i.i.d. sample $X_1, ..., X_n$ drawn from an unknown probability measure μ_0 defined on a space $\mathcal{X}$. We are interested in recovering a parameter $\theta_0 \in \Theta \subset \mathbb{R}^d$, defined by a set of moment conditions

$$\int_{\mathcal{X}} \Phi(\theta_0, x) d\mu_0(x) = 0, \tag{1}$$

where $\Phi : \Theta \times \mathcal{X} \to \mathbb{R}^k$ is a known map. This model is involved in many problems in Econometry, notably when dealing with instrumental variables. We refer to [Cha87], [Han82], [QL94], [Owe91] and [DIN09]. Two main approaches to the problem have been studied in the literature, namely the generalized method of moments (GMM) and the generalized empirical likelihood (GEL). While the main advantage of GMM relies in its computational feasibility, likelihood-related methods have appeared to be the most efficient in term of small-sample properties. In its original form, the empirical likelihood (EL) of Owen [Owe91] defines an estimator by a maximum likelihood procedure on a discretized version of the model. As an alternative, GEL replaces the Kullback criterion relative to EL by a f-divergence, thus providing a large choice of solutions. A number of estimators corresponding to particular choices of f-divergences have emerged in the literature over the last decades, such as the exponential tilting (ET) of Kitamura and Stutzer [KS97] and the continuous updating estimator (CUE) of Hansen, Yeaton and Yaron [HHY96].

While an attractive feature of GEL is its wide range of solutions, a number of f-divergence used in the computation of the GEL estimators are mainly justified by empirical studies and lack a probabilistic interpretation. This issue can be solved by incorporating some prior information to the problem using a Bayesian point of view, as made for example in [PR94]. In this paper, we investigate a different Bayesian approach to the inverse problem, known as *maximum entropy on the mean* (MEM). Although the method was originally introduced in the frame of exact moment condition models (as opposed here to parametric moment conditions), it appears to provide a natural solution to the problem, expressed as the minimizer of a convex functional on a set of discrete measures and subject to linear constraints. When applied in a particular setting, we show that the MEM approach leads to a GEL solution for which the f-divergence is determined by the choice of the prior. As a result, the method gives an alternate point of view on some widely spread estimators such as EL, ET or CUE, as well as a general Bayesian background to GEL.

In many actual situations, the true moment condition is not exactly known to the statistician and only an approximation is available. It occurs for instance when Φ has a complicated form that must be evaluated numerically. Simulation-based methods have been implemented to deal with approximate constraints in [CF00] and [McF89], in the frame of the generalized method of moments. To our knowledge, the efficiency of GEL in a similar situation has not been studied. In [LP08], the MEM procedure is shown to be robust to approximate moment conditions, introducing the approximate maximum entropy on the mean estimator. Seeing GEL as a particular case of MEM, we extend the model in a situation where only an approximation Φ_m of the true constraint function Φ is available. We provide sufficient conditions under which the GEL method remains efficient asymptotically when replacing Φ by its approximation.

This paper falls into the following parts. Section 2 is devoted to the position of the problem. We introduce the maximum entropy method for parametric moment condition models and discuss its close relationship with generalized empirical likelihood in Section 2.2. In Section 3, we discuss the asymptotic efficiency of the method when dealing with an approximate constraint. Proofs are postponed to the Appendix.

2 Estimation of the parameter

Let $\mathcal{X}$ be an open subset of $\mathbb{R}^q$, endowed with its Borel field $\mathcal{B}(\mathcal{X})$ and let $\mathcal{P}(\mathcal{X})$ denote the set of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. We observe an i.i.d. sample $X_1, ..., X_n$ drawn from the unknown distribution μ_0. We want to estimate the parameter $\theta_0 \in \Theta \subset \mathbb{R}^d$

defined by the moment condition

$$\int_{\mathcal{X}} \Phi(\theta_0, x) d\mu_0(x) = 0, \tag{2}$$

where $\Phi : \Theta \times \mathcal{X} \to \mathbb{R}^k$ ($k \geq d$) is a known map. To avoid a problem of identifiability, we assume that θ_0 is the unique solution of (2). This problem has been widely studied in the literature in Econometry, see for instance [Cha87], [Han82] and [QL94]. The information given by the moment condition (2) can be interpreted to determine the set $\mathcal{M}$ of possible values for μ_0 (the model). The true value of the parameter being unknown, the distribution of the observations can be any probability measure μ for which the map $\theta \mapsto \int \Phi(\theta, .)d\mu$ is null for a unique $\theta = \theta(\mu) \in \Theta$. The model is therefore defined as

$$\mathcal{M} = \left\{ \mu \in \mathcal{P}(\mathcal{X}) : \exists! \ \theta = \theta(\mu) \in \Theta, \int_{\mathcal{X}} \Phi(\theta, .)d\mu = 0 \right\},$$

where the map $\mu \mapsto \theta(\mu)$, defined on $\mathcal{M}$, is the parameter of interest. Let us introduce some notations and assumptions. For μ a measure and g a function, we shall note $\mu[g] = \int g d\mu$. Let E be an Euclidean space and let $\|.\|$ denote an Euclidean norm in E. For a function $f : \Theta \to E$ and a set $\mathcal{S} \subseteq \Theta$, we note

$$\|f\|_{\mathcal{S}} = \sup_{\theta \in \mathcal{S}} \|f(\theta)\|.$$

We assume that the following conditions are fulfilled.

A.1. Θ is a compact subset of $\mathbb{R}^d$.

A.2. The true value θ_0 of the parameter lies in the interior of Θ.

A.3. For all $x \in \mathcal{X}$, $\theta \mapsto \Phi(\theta, x)$ is continuous on Θ and the map $x \mapsto \|\Phi(., x)\|_\Theta$ is dominated by a μ_0-integrable function.

A.4. For all $x \in \mathcal{X}$, $\theta \mapsto \Phi(\theta, x)$ is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 and we note $\nabla\Phi(\theta, x) = \partial\Phi(\theta, x)/\partial\theta \in \mathbb{R}^{d\times k}$ and $\Psi(\theta, x) = \partial^2\Phi(\theta, x)/\partial\theta\partial\theta^t \in \mathbb{R}^{d\times d\times k}$. Moreover, we assume that $x \mapsto \|\nabla\Phi(., x)\|_{\mathcal{N}}$ and $x \mapsto \|\Psi(., x)\|_{\mathcal{N}}$ are dominated by a μ_0-integrable function.

A.5. The matrices

$$D := \int_{\mathcal{X}} \nabla\Phi(\theta_0, x)d\mu_0(x) \in \mathbb{R}^{d\times k} \quad \text{and} \quad V := \int_{\mathcal{X}} \Phi(\theta_0, x)\Phi^t(\theta_0, x)d\mu_0(x) \in \mathbb{R}^{k\times k}$$

are of full rank.

Some issues for estimating θ_0 may be due to the indirect definition of the parameter and these assumptions ensure that the map $\theta(.)$ is sufficiently smooth in a neighborhood of μ_0 for the total variation topology, which will make the asymptotic properties of the GEL estimator easily tractable.

2.1 Generalized empirical likelihood

Generalized empirical likelihood (GEL) was first applied to this problem in [QL94], generalizing an idea of [Owe91]. An estimate $\hat{\mu}$ of μ is obtained as an entropic projection of the empirical measure $\mathbb{P}_n$ onto the model $\mathcal{M}$. Precisely, for two probability measures μ and ν and f a convex function such that $f(1) = f'(1) = 0$, define

$$\mathcal{D}_f(\nu|\mu) = \int_{\mathcal{X}} f\left(\frac{d\nu}{d\mu}\right) d\mu \text{ if } \nu \ll \mu, \quad \mathcal{D}_f(\nu|\mu) = +\infty \text{ otherwise.}$$

Moreover, we define for $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$, $\mathcal{D}_f(\mathcal{A}|\mu) = \inf_{\nu\in\mathcal{A}} \mathcal{D}_f(\nu|\mu)$. The GEL estimator $\hat{\mu}$ of μ_0 is the element of the model that minimizes a given f-divergence $\mathcal{D}_f(.,\mathbb{P}_n)$ with respect to the empirical distribution. Noticing that the model can be written as $\mathcal{M} = \cup_{\theta\in\Theta}\mathcal{M}_\theta$ where $\mathcal{M}_\theta := \{\mu \in \mathcal{P}(\mathcal{X}) : \mu[\Phi(\theta, .)] = 0\}$, the GEL estimator $\hat{\theta} = \theta(\hat{\mu})$ follows by

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \mathcal{D}_f(\mathcal{M}_\theta, \mathbb{P}_n).$$

Since the set of discrete measures in $\mathcal{M}_\theta$ is closed and convex, the entropy $\mathcal{D}_f(\mathcal{M}_\theta, \mathbb{P}_n)$ is reached for a unique measure $\hat{\mu}(\theta)$ in $\mathcal{M}_\theta$, provided that $\mathcal{D}_f(\mathcal{M}_\theta, \mathbb{P}_n)$ is finite. Then, it appears that computing the GEL estimator involves a two-step procedure. First, build for each $\theta \in \Theta$, the entropic projection $\hat{\mu}(\theta)$ of $\mathbb{P}_n$ onto $\mathcal{M}_\theta$. Then, minimize $\mathcal{D}_f(\hat{\mu}(\theta), \mathbb{P}_n)$ with respect to θ. Since $\hat{\mu}(\theta)$ is absolutely continuous w.r.t. $\mathbb{P}_n$ by construction, minimizing $\mathcal{D}_f(.,\mathbb{P}_n)$ reduces to finding the proper weights $w_1, ..., w_n$ to allocate to the observations $X_1, ..., X_n$. This turns into a finite dimensional problem, which can be solved by classical convex optimization tools (see for instance [Kit06]). In fact, the GEL estimator $\hat{\theta}$ can be expressed as the solution to the saddle point problem

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{(\gamma,\lambda)\in\mathbb{R}\times\mathbb{R}^k} \left\{\gamma - \mathbb{P}_n\left[f^*(\gamma + \lambda^t\Phi(\theta, .))\right]\right\},$$

where $f^*(x) = \sup_y \{xy - f(y)\}$ denotes the convex conjugate of f.

Note that if the choice of the f-divergence plays a key role in the construction of the estimator, it has no influence on its asymptotic efficiency. Indeed, it is shown in [QL94] that all GEL estimators are asymptotically efficient, regardless of the f-divergence used for their computation. Nevertheless, some situations justify the use of specific f-divergences. The empirical likelihood estimator introduced by Owen in [Owe91] uses the Kullback entropy $\mathcal{K}(., .)$ as f-divergence, pointing out that minimizing $\mathcal{K}(.,\mathbb{P}_n)$ reduces to maximizing likelihood among multinomial distributions. Newey and Smith [NS04] remark that a quadratic f-divergence leads to the CUE estimator of Hansen Heaton and Yaron [HHY96].

2.2 Maximum entropy on the mean

In this section, we study a Bayesian approach to the inverse problem, known as maximum entropy on the mean (MEM) [GG97]. The method was developed to estimate a measure μ_0 based the observation of some of its moments. In this framework, it turns out that the MEM estimator of μ_0 can be used to estimate efficiently the parameter θ_0. We shall briefly recall the MEM procedure. Consider an estimator of μ_0 in the form of a weighted version of the empirical measure $\mathbb{P}_n$,

$$\mathbb{P}_n(w) = \frac{1}{n}\sum_{i=1}^{n} w_i \ \delta_{X_i},$$

for $w = (w_1, ..., w_n)^t \in \mathbb{R}^n$ a collection of weights. Then, fix a prior distribution ν_0 on the vector of weight w so that each solution $\mathbb{P}_n(w)$ can be viewed as a realization of the random measure $\mathbb{P}_n(W)$, where W is drawn from ν_0. This setting enables to incorporate some prior knowledge on the shape or support of μ_0 through the choice of the prior ν_0, as discussed in [GG97]. Here, the observations $X_1, ..., X_n$ are considered fixed. Actually, it is the moment condition that is used to built the estimator *a posteriori*. In this framework where the true value θ_0 of the parameter is unknown, the information provided by the moment condition reduces to the statement $\mu_0 \in \mathcal{M}$. So, in order to take this information into consideration, the underlying idea of MEM is to build the estimator $\hat{\mu}$ as the expectation of $\mathbb{P}_n(W)$ conditionally to the event $\{\mathbb{P}_n(W) \in \mathcal{M}\}$. However, we may encounter some difficulties if this conditional expectation is not properly defined. To deal with this issue, the MEM method replaces the possibly ill-defined conditional expectation by a well-defined estimator, whose construction is motivated by large deviation principles. Precisely, construct the *posterior* distribution ν^* as the entropic projection of ν_0 onto the set

$$\Pi(\mathcal{M}) = \{\mu \in \mathcal{P}(\mathbb{R}^n), \ \mathbb{E}_\mu \left[\mathbb{P}_n(W)\right] \in \mathcal{M}\},$$

where $\mathbb{E}_\mu \left[\mathbb{P}_n(W)\right]$ denotes the expectation of $\mathbb{P}_n(W)$ when W has distribution μ. The MEM solution to the inverse problem is defined as the expectation of $\mathbb{P}_n(W)$ under the posterior distribution ν^*,

$$\hat{\mu} = \mathbb{E}_{\nu^*} \left[\mathbb{P}_n(W)\right] = \mathbb{P}_n(\mathbb{E}_{\nu^*}(W)).$$

This construction is justified by Theorem 2.3 in [GG97], which establishes the asymptotic equivalence between $\hat{\mu}$ and the conditional expectation $\mathbb{E}_{\nu_0}(\mathbb{P}_n(W)| \ \mathbb{P}_n(W) \in \mathcal{M})$, whenever it is well defined. The existence of the MEM estimator requires the problem to be *feasible* in the sense that there exists at least one solution δ in the interior of the convex hull of the support of ν_0, such that $\mathbb{P}_n(\delta) \in \mathcal{M}$. This assumption warrants that the set $\Pi(\mathcal{M})$ is non-empty and therefore allows the construction of the posterior distribution ν^*.

The MEM estimator $\hat{\mu}$ lies in the model $\mathcal{M}$ by construction. As a result, there exists a solution $\hat{\theta}$ to the moment condition $\hat{\mu}[\Phi(\theta, .)] = 0$. This solution is precisely the MEM estimator of θ_0. In Theorem 2.1 below, we give an explicit expression for the MEM estimator $\hat{\theta}$. We note $\mathbb{1} = (1, ..., 1)^t \in \mathbb{R}^n$, $\Phi(\theta, X) = (\Phi(\theta, X_1), ..., \Phi(\theta, X_n))^t \in \mathbb{R}^{n\times k}$ and as previously, Λ_ν denotes the log-Laplace transform of ν.

Theorem 2.1 *If the problem is feasible, the MEM estimator $\hat{\theta}$ is given by*

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{(\gamma,\lambda)\in\mathbb{R}\times\mathbb{R}^k} \left\{n\gamma - \Lambda_{\nu_0}(\gamma\mathbb{1} + \Phi(\theta, X)\lambda)\right\}.$$

In particular, if ν_0 has equal orthogonal marginals, i.e. $\nu_0 = \nu^{\otimes n}$ for some probability measure ν on $\mathbb{R}$, then

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{(\gamma,\lambda)\in\mathbb{R}\times\mathbb{R}^k} \left\{\gamma - \mathbb{P}_n\left[\Lambda_\nu(\gamma + \lambda^t\Phi(\theta, .))\right]\right\}.$$

The MEM estimator $\hat{\theta}$ can be expressed as the solution to a saddle point problem, specific to generalized empirical likelihood. Actually, this result points out that maximum entropy on the mean with a particular form of prior $\nu_0 = \nu^{\otimes n}$ leads to a GEL procedure, for which the criterion is the log-Laplace transform of ν. This approach provides a general Bayesian interpretation of GEL. Regularity conditions on the criterion Λ_ν in the GEL framework are reflected through conditions on the prior ν. Indeed, the usual normalization conditions $\Lambda'_\nu(0) = \Lambda''_\nu(0) = 1$ corresponds to taking a prior ν with mean and variance equal to one, while the normalization $\Lambda_\nu(0) = 0$ is imposed by the condition $\nu \in \mathcal{P}(\mathbb{R})$.

An interesting choice of prior is the exponential distribution $d\nu(x) = e^{-x}dx$ for $x > 0$. Indeed, observe that if the W_i are i.i.d. with exponential distribution, the likelihood of $\mathbb{P}_n(W)$ is constant over the set of probability discrete measures $\{\mathbb{P}_n(w) : \sum_{i=1}^n w_i = n\}$. Hence, an exponential prior can be roughly interpreted as a non-informative prior in this framework. The discrepancy associated to this prior is $\Lambda_\nu(s) = -\log(1-s),\ s < 1$, which corresponds to the empirical likelihood estimator of Owen [Owe91].

The MEM approach also provides a new probabilistic interpretation of some commonly used specific GEL estimators. The exponential tilting of Kitamura and Stutzer [KS97] is obtained for a Poisson prior of parameter 1, for which we have $\Lambda_\nu(s) = e^s - 1$. Another example is the Gaussian prior $\nu \sim \mathcal{N}(1, 1)$, leading to the continuous updating estimator of Hansen, Yeaton and Yaron [HHY96], as we have in this case $\Lambda_\nu(s) = \frac{1}{2}(s-1)^2$. The Gaussian prior allows the discrete measure $\mathbb{P}_n(W)$ to have negative weights w_i and must be handled with care. Remark however that this is generally not an issue in practice since the solution $\hat{\mu}$ is implicitly chosen close to the empirical distribution $\mathbb{P}_n$ and will have all its weights w_i positive with high probability. There are more examples of priors leading to usual discrepancies which can be found in [GG97].

3 Dealing with an approximate operator

In many actual applications, only an approximation of the constraint function Φ is available to the practitioner. This occurs for instance if the moment condition takes a complicated form that can only be evaluated numerically. In [McF89], McFadden suggested a method dealing with approximate constraint in a similar situation, introducing the method of simulated moments (see also [CF00]). In [LP08] and [LR09], the authors study a MEM procedure for linear inverse problems with approximate constraints. Here, we propose to extend the results of [LP08] and [LR09] to the GEL framework, using the connections between GEL and MEM.

We assume that we observe a sequence $\{\Phi_m\}_{m\in\mathbb{N}}$ of approximate constraints, independent with the original sample $X_1, ..., X_n$ and converging toward the true function Φ at a rate φ_m. We are interested in exhibiting sufficient conditions on the sequence $\{\Phi_m\}_{m\in\mathbb{N}}$ under which estimating θ_0 by the GEL procedure remains efficient when the constraint is replaced by its approximation. We discuss the asymptotic properties of the resulting estimates in a framework where both indices n and m simultaneously grow to infinity.

The approximate estimator is obtained by the GEL methodology, replacing the constraint function Φ by its approximation Φ_m,

$$\hat{\theta}_m = \arg\min_{\theta\in\Theta} \sup_{(\gamma,\lambda)\in\mathbb{R}\times\mathbb{R}^k} \left\{\gamma - \mathbb{P}_n\left[\Lambda(\gamma + \lambda^t\Phi_m(\theta, .))\right]\right\}, \tag{3}$$

where $\Lambda : \mathbb{R} \to \overline{\mathbb{R}}$ is a strictly convex, twice differentiable function such that $\Lambda'(0) = \Lambda''(0) = 1$ and $\Lambda(0) = 0$. As previously, the existence of $\hat{\theta}_m$ requires the feasibility condition that the supremum of $\gamma - \mathbb{P}_n[\Lambda(\gamma + \lambda^t\Phi_m(\theta, .))]$ is reached for a finite value of $(\gamma, \lambda) \in \mathbb{R} \times \mathbb{R}^k$, for at least one value of $\theta \in \Theta$. This condition relies essentially on the domain of Λ being sufficiently widespread. We make the following additional assumptions.

A.6. The functions $x \mapsto \|\Phi(., x)\|_\Theta$, $x \mapsto \|\nabla\Phi(., x)\|_\mathcal{N}$ and $x \mapsto \|\Psi(., x)\|_\mathcal{N}$ are dominated by a function κ such that $\int \kappa^4(x)d\mu_0(x) < \infty$.

A.7. For all $x \in \mathcal{X}$ and for sufficiently large m, the map $\theta \mapsto \Phi_m(\theta, .)$ is twice continuously differentiable in $\mathcal{N}$ and we note $\nabla\Phi_m(\theta, .) = \partial\Phi_m(\theta, .)/\partial\theta$ and $\Psi_m(\theta, .) = \partial^2\Phi_m(\theta, .)/\partial\theta\partial\theta^t$.

A.8. The functions $x \mapsto \|\Phi_m(., x) - \Phi(., x)\|_\Theta$, $x \mapsto \|\nabla\Phi_m(., x) - \nabla\Phi(., x)\|_\mathcal{N}$ and $x \mapsto \|\Psi_m(., x) - \Psi(., x)\|_\mathcal{N}$ are dominated by a function κ_m such that $\int \kappa_m^4(x)d\mu_0(x) = O(\varphi_m^{-4})$.

A.9. The function Λ'' is bounded by a constant $K < \infty$.

Assumptions A.6 to A.8 are made to obtain a uniform control over $\|\hat{\theta}_m - \hat{\theta}\|$ for all $n \in \mathbb{N}$. The condition A.9 implies that Λ is dominated by a quadratic function. In the MEM point of view, this condition is fulfilled for the log-Laplace transform Λ_ν of sub-Gaussian priors ν.

Theorem 3.1 (Robustness of GEL) *If Assumptions 1 to 9 hold,*

$$n\|\hat{\theta}_m - \hat{\theta}\|^2 = O_P(n\varphi_m^{-2}) + o_P(1).$$

Moreover, $\hat{\theta}_m$ is $\sqrt{n}$-consistent and asymptotically equivalent to the GEL estimator computed with exact constraint function Φ whenever $n\varphi_m^{-2}$ tends to zero.

By considering a situation with approximate operator, we extend the GEL model to a more general framework that gives a more realistic formulation of actual problems. The previous theorem gives an upper bound of the error caused by the use of the approximation Φ_m in place of the true function Φ. By this result, we aim to provide an insight on convergence conditions that are necessary for asymptotic efficiency when dealing with an approximate operator.

4 Proofs

4.1 Proof of Theorem 2.1

Let $\mathcal{S}_\theta = \{w \in \mathbb{R}^n : \mathbb{P}_n(w) \in \mathcal{M}_\theta\}$ and $\mathcal{F}_w = \{\mu \in \mathcal{P}(\mathbb{R}^n) : \mathbb{E}_\mu(W) = w\}$. We use that $\inf_{\mu \in \mathcal{F}_w} \mathcal{K}(\mu, \nu_0) = \Lambda_{\nu_0}^*(w)$ (see [GG97]). Let $\Pi(\mathcal{M}_\theta) = \{\mu \in \mathcal{P}(\mathbb{R}^n),\ \mathbb{E}_\mu[\mathbb{P}_n(W)] \in \mathcal{M}_\theta\}$, we have the equality

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \inf_{\mu\in\Pi(\mathcal{M}_\theta)} \mathcal{K}(\mu,\nu_0) = \arg\min_{\theta\in\Theta} \inf_{w\in\mathcal{S}_\theta} \inf_{\mu\in\mathcal{F}_w} \mathcal{K}(\mu,\nu_0),$$

which can be written

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \inf_{w\in\mathcal{S}_\theta} \Lambda_{\nu_0}^*(w) = \arg\min_{\theta\in\Theta} \inf_{w\in\mathcal{S}_\theta} \sup_{\tau\in\mathbb{R}^n} \{\tau^t w - \Lambda_{\nu_0}(\tau)\}.$$

The feasibility assumption warrants that the extrema are reached. Hence, using Sion's minimax Theorem, we find

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{\tau\in\mathbb{R}^n} \inf_{w\in\mathcal{S}_\theta} \{\tau^t w - \Lambda_{\nu_0}(\tau)\},$$

We know that $w = (w_1, ..., w_n)^t \in \mathcal{S}_\theta$ if and only if $\sum_{i=1}^n w_i = n$ and $\sum_{i=1}^n w_i \Phi(\theta, X_i) = 0$. Thus, for a fixed value of τ, the map $w \mapsto \tau^t w - \Lambda_{\nu_0}(\tau)$ can be arbitrarily close to $-\infty$ on $\mathcal{S}_\theta$ whenever τ is not orthogonal to $\mathbb{1}$ and $\Phi(\theta, X)$. As a result, we may assume that $\tau = \gamma\mathbb{1} + \Phi(\theta, X)\lambda$ for some $(\gamma, \lambda) \in \mathbb{R} \times \mathbb{R}^k$ without loss of generality. In this case, the map $w \mapsto \tau^t w - \Lambda_{\nu_0}(\tau)$ is constant over $\mathcal{S}_\theta$, equal to $n\gamma - \Lambda_{\nu_0}(\gamma\mathbb{1} + \Phi(\theta, X)\lambda)$, which ends the proof. If $\nu_0 = \nu^{\otimes n}$, then $\Lambda_{\nu_0}(w) = \sum_{i=1}^n \Lambda_\nu(w_i)$ and we conclude easily.

4.2 Proof of Theorem 3.1

The proof of the results relies mainly on the uniform law of large numbers, using that the set $\{\|\Phi_m(\theta,.)\|, \|\nabla\Phi_m(\theta,.)\|, \|\Psi_m(\theta,.)\|,\ \theta\in\Theta, m\in\mathbb{N}\}$ is a Glivenko-Cantelli class of functions, consequently to A.6 and A.8. For all $\theta\in\Theta$, $v\in\mathbb{R}^k$, $x\in\mathcal{X}$, let

$$\begin{array}{rcl} h_n(\theta,v) &=& \left(\begin{array}{c} \mathbb{P}_n\left[\Phi(\theta,.)\Lambda'(v^t\Phi(\theta,.))\right] \\ \mathbb{P}_n\left[v^t\nabla\Phi^t(\theta,.)\Lambda'(v^t\Phi(\theta,.))\right]\end{array}\right) \\ h_{m,n}(\theta,v) &=& \left(\begin{array}{c} \mathbb{P}_n\left[\Phi_m(\theta,.)\Lambda'(v^t\Phi_m(\theta,.))\right] \\ \mathbb{P}_n\left[v^t\nabla\Phi_m^t(\theta,.)\Lambda'(v^t\Phi_m(\theta,.))\right]\end{array}\right). \end{array}$$

The pair $(\hat{\theta}_m,\hat{v}_m)$ (resp. $(\hat{\theta},\hat{v})$) is defined as the unique zero over $\Theta\times\mathbb{R}^k$ of $h_{m,n}$ (resp. h_n). The condition A.9 implies that there exists a constant $K>0$ such that $\Lambda'(s)\le Ks+1$ for all $s\in\mathbb{R}$. Hence, using successively the mean value theorem and Cauchy-Schwarz's inequality, we show that the contrast function $h_{m,n}$ converges uniformly on every compact set toward h_n as $m\to\infty$, which warrants the convergence of $(\hat{\theta}_m,\hat{v}_m)$ toward $(\hat{\theta},\hat{v})$. For all $v\in\mathbb{R}^k$, the application $\theta\mapsto\nabla h_{m,n}(\theta,v)$ is continuous in a neighborhood on θ_m^* for sufficiently large values of m by the condition A.7, as explicit calculation gives

$$\nabla h_{m,n}(\theta,v)=\left(\begin{array}{cc} A_{m,n}(\theta,v) & D_{m,n}(\theta,v) \\ D_{m,n}^t(\theta,v) & V_{m,n}(\theta,v)\end{array}\right),$$

where

$$\begin{array}{rcl} A_{m,n}(\theta,v) &=& \mathbb{P}_n\left[\Psi_m(\theta,.)v\Lambda'(v^t\Phi_m(\theta,.))+\nabla\Phi_m(\theta,.)v\ v^t\nabla\Phi_m^t(\theta,.)\Lambda''(v^t\Phi_m(\theta,.))\right] \\ D_{m,n}(\theta,v) &=& \mathbb{P}_n\left[\nabla\Phi_m(\theta,.)\Lambda'(v^t\Phi_m(\theta,.))+\nabla\Phi_m(\theta,.)v\Phi_m^t(\theta,.)\Lambda''(v^t\Phi_m(\theta,.))\right] \\ V_{m,n}(\theta,v) &=& \mathbb{P}_n\left[\Phi_m(\theta,.)\Phi_m^t(\theta,.)\Lambda''(v^t\Phi_m(\theta,.))\right]. \end{array}$$

We define in the same way $A_n(\theta,v)$, $D_n(\theta,v)$ and $V_n(\theta,v)$ by replacing Φ_m by Φ in the expressions above. Using Cauchy-Schwarz's inequality, A.8 ensures the uniform convergence of $\nabla h_{m,n}$ toward ∇h_n on every compact set at the rate φ_m. Note ρ_n the smallest eigenvalue of $\nabla h_n(\hat{\theta},\hat{v})$, we know from Theorem 3.2 in [NS04] that $\mathbb{P}(\rho_n>\eta)=o(n^{-1})$ for sufficiently small $\eta>0$, since A.5 ensures that the limit of $\nabla h_n(\hat{\theta},\hat{v})$ as $n\to\infty$ is positive definite. Thus, for $c>0$ sufficiently small, consider the event $\Omega=\{\rho_n>c\}$. Writing the Taylor expansion

$$h_{m,n}(\hat{\theta},\hat{v})=\nabla h_{m,n}(\hat{\theta}_m,\hat{v}_m)\left(\begin{array}{c}\hat{\theta}-\hat{\theta}_m \\ \hat{v}-\hat{v}_m\end{array}\right)+o(\|\hat{\theta}_m-\hat{\theta}\|),$$

we deduce that on Ω,

$$\left(\begin{array}{c}\hat{\theta}_m-\hat{\theta} \\ \hat{v}_m-\hat{v}\end{array}\right)=-\left[\nabla h_n(\hat{\theta},\hat{v})\right]^{-1}h_{m,n}(\hat{\theta},\hat{v})+O_P(\varphi_m^{-1}).$$

The Schur complement formula gives in particular

$$\hat{\theta}_m - \hat{\theta} = -\left[\hat{D}_n \hat{V}_n^{-1} \hat{D}_n^t\right]^{-1} \hat{D}_n \hat{V}_n^{-1} \, \mathbb{P}_n[\Phi_m(\hat{\theta}, .)\Lambda'(\hat{v}^t \Phi_m(\hat{\theta}, .))] + O_P(\varphi_m^{-1}) + o_P(n^{-1}),$$

where $\hat{D}_n = D_n(\hat{\theta}, \hat{v})$ and $\hat{V}_n = V_n(\hat{\theta}, \hat{v})$ and where we used that $\hat{v} = O_P(n^{-1})$ (see for instance Theorem 3.2 in [NS04]). Thus, on the event Ω,

$$\|\hat{\theta}_m - \hat{\theta}\| \leq c \left\|\mathbb{P}_n[\Phi_m(\hat{\theta}, .)\Lambda'(\hat{v}^t \Phi_m(\hat{\theta}, .))]\right\| + O_P(\varphi_m^{-1}) + o_P(n^{-1}).$$

By construction, $\mathbb{P}_n[\Phi(\hat{\theta}, .)\Lambda'(\hat{v}^t \Phi(\hat{\theta}, .))] = 0$, which yields

$$\begin{aligned}
& \left\|\mathbb{P}_n[\Phi_m(\hat{\theta}, .)\Lambda'(\hat{v}^t \Phi_m(\hat{\theta}, .))]\right\| \\
\leq \; & \mathbb{P}_n \left[\|(\Phi_m(\hat{\theta}, .) - \Phi(\hat{\theta}, .))\Lambda'(\hat{v}^t \Phi_m(\hat{\theta}, .))\| + \|\Phi(\hat{\theta}, .)[\Lambda'(\hat{v}^t \Phi_m(\hat{\theta}, .) - \Lambda'(\hat{v}^t \Phi(\hat{\theta}, .))]\|\right] \\
\leq \; & K\|\hat{v}\| \, \mathbb{P}_n \left[\|\Phi_m(\hat{\theta}, .)\| \; \|\Phi_m(\hat{\theta}, .) - \Phi(\hat{\theta}, .)\|\right] + \mathbb{P}_n \|\Phi_m(\hat{\theta}, .) - \Phi(\hat{\theta}, .)\| \\
& + K\|\hat{v}\| \, \mathbb{P}_n \left[\|\Phi(\hat{\theta}, .)\| \; \|\Phi_m(\hat{\theta}, .) - \Phi(\hat{\theta}, .)\|\right],
\end{aligned}$$

as a consequence of A.9. We conclude that $\|\hat{\theta}_m - \hat{\theta}\|^2 \mathbb{1}_\Omega = O_P(\varphi_m^{-2}) + o_P(n^{-1})$ by the condition A.8. On the complement of Ω, $\|\hat{\theta}_m - \hat{\theta}\|$ can be bounded by the diameter δ of Θ, yielding $\|\hat{\theta}_m - \hat{\theta}\| \mathbb{1}_{\Omega^c} = o_P(n^{-1})$, which ends the proof.

References

[CF00] M. Carrasco and J. P. Florens. Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 16(6):797–834, 2000.

[Cha87] G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics*, 34(3):305–334, 1987.

[DIN09] S. G. Donald, G. W. Imbens, and W. K. Newey. Choosing instrumental variables in conditional moment restriction models. *J. Econometrics*, 152(1):28–36, 2009.

[GG97] F. Gamboa and E. Gassiat. Bayesian methods and maximum entropy for ill-posed inverse problems. *Ann. Statist.*, 25(1):328–350, 1997.

[Han82] L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

[HHY96] L. P. Hansen, J. Heaton, and A. Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–80, July 1996.

[Kit06] Y. Kitamura. Empirical likelihood methods in econometrics: Theory and practice. Cowles Foundation Discussion Papers 1569, Cowles Foundation for Research in Economics, Yale University, 2006.

[KS97] Y. Kitamura and M. Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4):861–874, 1997.

[LP08] 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.

[LR09] J. M. Loubes and P. Rochet. Regularization with approximated l^2 maximum entropy method. In *submitted, Electronic version HAL 00389698*. 2009.

[McF89] D. McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5):995–1026, 1989.

[NS04] W. K. Newey and R. J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

[Owe91] A. Owen. Empirical likelihood for linear models. *Ann. Statist.*, 19(4):1725–1747, 1991.

[PR94] Florens J. P. and J.M. Rolin. Bayes, bootstrap, moments. *Discussion Paper 9413, Institut de Statistique, Université catholique de Louvain, Belgium*, 1994.

[QL94] J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *Ann. Statist.*, 22(1):300–325, 1994.