2004s-18

On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood

Hélène Bonnal, Eric Renault

Série Scientifique
Scientific Series

Montréal
Mai 2004

CIRANO
Centre interuniversitaire de recherche en analyse des organisations

CIRANO

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du ministère de la Recherche, de la Science et de la Technologie, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Québec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the Ministère de la Recherche, de la Science et de la Technologie, and grants and research mandates obtained by its research teams.

Les organisations-partenaires / The Partner Organizations

PARTENAIRE MAJEUR
. Ministère du développement économique et régional et de la recherche [MDERR]

PARTENAIRES
. Alcan inc.
. Axa Canada
. Banque du Canada
. Banque Laurentienne du Canada
. Banque Nationale du Canada
. Banque Royale du Canada
. Bell Canada
. BMO Groupe Financier
. Bombardier
. Bourse de Montréal
. Caisse de dépôt et placement du Québec
. Développement des ressources humaines Canada [DRHC]
. Fédération des caisses Desjardins du Québec
. GazMétro
. Hydro-Québec
. Industrie Canada
. Ministère des Finances du Québec
. Pratt & Whitney Canada Inc.
. Raymond Chabot Grant Thornton
. Ville de Montréal

. École Polytechnique de Montréal
. HEC Montréal
. Université Concordia
. Université de Montréal
. Université du Québec à Montréal
. Université Laval
. Université McGill
. Université de Sherbrooke

ASSOCIE A :
. Institut de Finance Mathématique de Montréal (IFM^2)
. Laboratoires universitaires Bell Canada
. Réseau de calcul et de modélisation mathématique [RCM^2]
. Réseau de centres d'excellence MITACS (Les mathématiques des technologies de l'information et des systèmes complexes)

ISSN 1198-8177

On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood*

Hélène Bonnal†, *Eric Renault*‡

Résumé / *Abstract*

Plusieurs méthodes alternatives à GMM basées sur un critère d'information ont récemment été proposées. Pour leur utilisation pratique et leur interprétation, le principal défaut de ces alternatives, particulièrement dans le cas de restrictions de moments conditionnels, est de faire appel à des programmes d'optimisation convexe de très grande dimension. La contribution principale de cet article est d'analyser le contenu informatif d'équations estimantes dans le cadre unifié de projections de moindres carrés. L'amélioration de l'inférence par variables de contrôle, le calcul des probabilités impliquées et les interprétations informationnelles des différentes versions de GMM sont discutés dans les deux cadres de moments conditionnels et inconditionnels.

Mots clés : vraisemblance empirique, GMM avec révision continue, information, variables de contrôle, efficacité semi-paramétrique, théorie asymptotique à l'ordre supérieur, chi-deux minimum.

A number of information-theoretic alternatives to GMM have recently been proposed in the literature. For practical use and general interpretation, the main drawback of these alternatives, particularly in the case of conditional moment restrictions, is that they rely on high dimensional convex optimization programs. The main contribution of this paper is to analyze the informational content of estimating equations within the unified framework of least squares projections. Improved inference by control variables, shrinkage of implied probabilities and information-theoretic interpretations of continuously updated GMM are discussed in the two cases of unconditional and conditional moment restrictions.

Keywords: *empirical likelihood, continuously updated GMM, information, control variables, semiparametric efficiency, higher order asymptotics, minimum chi-square.*

* A previous version of this paper has circulated under the title "Minimum Chi-Square Estimation with Conditional Moment Restrictions", W.P. 2001. We thank Xiaohong Chen, Fabrice Gamboa, Christian Gouriéroux, Lars Peter Hansen, Yuichi Kitamura, Esfandiar Maasoumi and Richard Smith for helpful discussions.

† GREMAQ, University of Toulouse 1, France, email: hbonnal@gremaq.univ-tlse1.fr.

‡ Université de Montréal, CIRANO and CIREQ, email: eric.renault@umontreal.ca.

1 Introduction

It has long been appreciated that in some circumstances likelihood functions may not be available and the focus of parametric inference is only on a limited number of structural parameters associated to the data generating process (DGP) by a structural econometric model. Hansen (1982) has fully settled the theory to use efficiently the informational content of such moment conditions about unknown structural parameters while Chamberlain (1987) showed that the semiparametric efficiency bound for conditional moment restriction models is attained by optimal GMM.

However, and somewhat surprisingly, the pre-1990 GMM literature seems to have forgotten that moment restrictions, when they overidentify the structural parameters of interest, may bring useful information about other characteristics of the DGP. To see this, let us consider that we have at our disposal n i.i.d. observations $(X_i, Z_i), i = 1, \cdots, n$ of a random vector (X, Z) on $\mathbb{R}^k \times \mathbb{R}^d$. The focus of our interest in this paper is the information content of either q unconditional moment restrictions:

$$E\left[\Psi(X, \theta^0)\right] = 0 \tag{1.1}$$

or q conditional moment restrictions

$$E\left[\Psi(X, \theta^0)\,|Z\right] = 0 \tag{1.2}$$

which, in both cases, are assumed to define the true unknown value θ^0 of a vector $\theta \in \Theta \subset \mathbb{R}^p$ of p unknown parameters, while $\Psi : \mathbb{R}^k \times \Theta \longrightarrow \mathbb{R}^q$ is a known function. When $q > p$ in case (1.1) or irrespective of the value of q in case (1.2), only one part of the informational content of these moment restrictions is actually used by traditional GMM approaches to estimate θ efficiently. The usefulness of residual information due to overidentification is overlooked.

Actually, following Hansen (1982), efficient estimation of θ^0 from (1.1) goes through a preliminary consistent estimation of a matrix $M\left(\theta^0\right)$ of optimal selection of estimating equations:

$$M(\theta^0) = E\left[\frac{\partial \Psi'}{\partial \theta}(X, \theta^0)\right] Var^{-1}\left[\Psi(X, \theta^0)\right] \tag{1.3}$$

while, as surveyed by Newey (1993), efficient estimation of θ^0 from (1.2) rests upon a preliminary consistent estimation of a matrix $M(Z, \theta^0)$ of optimal instruments:

$$M(Z, \theta^0) = E\left[\frac{\partial \Psi'}{\partial \theta}(X, \theta^0)\,|Z\right] Var^{-1}\left[\Psi(X, \theta^0)\,|Z\right] \tag{1.4}$$

The important idea that such overidentified moment restrictions should also lead us to revise our empirical views about the DGP has first been put forward by the empirical likelihood literature (Owen (1990), (1991), Qin and Lawless (1994)) for a classical approach, and by Zellner's Bayesian Method of Moments (BMOM) for a Bayesian one (Zellner (1991), Zellner and Tobias (2001)).

Typically, as clearly explained in Zellner (2003), the idea is to seek the least informative density function in terms of expected distance subject to the moment conditions. But, while Zellner considers expected distances with respect to priors, we are going to consider distances with respect to empirical probability distributions, that put weights $1/n$ on the n observed values $X_j, j = 1, \cdots n$ in case (1.1) and smoothed kernel weights ω_{ij} on the n observed values $X_j, j = 1, \cdots n$ given the possible conditioning values $Z_i, i = 1, \cdots n$, in case (1.2).

In other words, following Maasoumi (1993), the distance between observed empirical distribution and an hypothetical probability distribution conformable to the moment restrictions will be the unifying tool of this paper. While computing such implied probability distributions should be of interest for a variety of econometric applications like asset pricing, forecasting or simulations, the focus of our interest in this paper is more estimation of the structural parameters θ. However, we show that implied probabilities precisely afford an efficient use of the informational content of estimating equations to learn about any population expectation $Eg(X)$ in case (1.1.) or $E[g(X)|Z]$ in case (1.2) for any test function g.

We actually argue that it is precisely this efficient use which allows us to efficiently estimate the optimal selection matrix (1.3) in case (1.1) as well as the optimal instruments (1.4) in case (1.2). More precisely, we show that implied probabilities provided by some Euclidean empirical likelihood approach, both in the unconditional and conditional cases, define estimators of $E[g(X)]$ and $E[g(X)|Z]$ which make use of the moment conditions $\Psi(X, \theta^0)$ as control variates. In other words, our estimators have less variance than simple empirical counterparts of $E[g(X)]$ or $E[g(X)|Z]$ (empirical mean for the former, kernel estimator for the latter) because covariation between $g(X)$ and moment conditions is exploited.

When applied to estimation of expectations of $\frac{\partial \Psi'}{\partial \theta}(X, \theta^0)$ and $\Psi(X, \theta^0)\Psi'(X, \theta^0)$ (to get rid of (1.3) or (1.4)), this control variates approach precisely addresses an issue pointed out by several authors (see in particular Altonji and Segal (1996)) to explain the poor finite sample performance of standard GMM. This is precisely because we have deleted any perverse correlation between our estimators of $M(\theta^0)$ or $M(Z, \theta^0)$ and moment conditions that we will improve the small sample properties of GMM.

While this control variates improvement is so natural and user-friendly, one may wonder why so much emphasis has recently been put on one-step procedures based on empirical likelihood or Kullback-Leibler information criterion (see Kitamura and Stutzer (1997), Imbens (1997), Imbens, Spady and Johnson (1998), Newey and Smith (2004) for the unconditional case, Kitamura, Tripathi and Ahn (2000), Donald, Imbens, Newey (2001) for the conditional case). We show in this paper that the main advantage of one-step empirical likelihood approaches is to provide estimating equations for θ where the optimal matrices $M(\theta^0)$ or $M(Z, \theta^0)$ are implicitely efficiently estimated. In particular, contrary to what is sometimes said, the issue is not to avoid nonparametric estimation of optimal instruments but just do it simultaneously with estimation of θ.

However, the practical drawback of empirical likelihood is well known. Implied probabilities can be only numerically computed, through a high dimensional convex optimization program. This problem is especially detrimental in the case of conditional implied probabilities since the dimension

of the needed optimization program grows proportionally to the sample size. By contrast, maximization of Euclidean empirical likelihood provides closed form formulas for implied probabilities and natural control variates interpretations of associated estimated expectations. Moreover, we show that the Euclidean empirical likelihood estimator of θ coincides with continuously updated GMM (CUE-GMM) as first proposed by Hansen, Heaton and Yaron (1996). While this result is not really surprising in the unconditional case, it sheds some light on some new conditional versions of CUE-GMM. This interpretation is related to the work of Ai and Chen (2001). They propose a conditional version of efficient two-stage GMM (2S-GMM) by minimizing a well-chosen norm of a kernel estimation of the conditional moments (1.2). By considering the profile criterion for θ of a smoothed version of Euclidean empirical likelihood, we get a similar conditional CUE-GMM.

Finally, we propose an answer to two criticisms often given against Euclidean empirical likelihood by contrast with empirical likelihood.

First, it is known at least in the unconditional case (see Newey and Smith (2004)) that, while a one-step empirical likelihood maximization amounts, in terms of estimation of θ, to an efficient estimation of the optimal selection matrix $M(\theta^0)$ (or, as we show, of the optimal instruments $M(Z, \theta^0)$ in the conditional case), the drawback of one step Euclidean empirical likelihood is to omit the information content of estimation equations to estimate the variance matrix $Var\left[\Psi(X(X, \theta^0)\right]$, while this information is taken into account to estimate the Jacobian matrix $E\left[\frac{\partial \Psi'}{\partial \theta}(X, \theta^0)\right]$ (and similarly in the conditional case). But we argue that nothing prevents us to introduce an additional step of estimation to compute the efficient control variates estimators of these matrices. In order to minimize the computational burden, we then propose a three step estimators: one step to get a consistent estimator, a second step to get an efficient estimation and a third step to get an estimator with similar higher order properties as empirical likelihood, while only quadratic minimizations program are involved.

A second often maintained criticism against Euclidean empirical likelihood is to provide implied probabilities the non-negativity of which is not guaranteed in finite sample. However, we argue that a simple shrinkage towards empirical probabilities may hedge against this risk without any asymptotic efficiency loss.

The paper is organized as follows.

We consider in section 2 the general issue of minimization of power divergence statistics, elements of the Cressie-Read family of divergences. Empirical Likelihood (EL) and Euclidean Empirical Likelihood are particular cases. We show that, when minimized subject to unconditional moment restrictions (1.1), these divergence statistics take implicitly advantage of the overidentifying restrictions to improve estimation of the optimal selection of estimating equations. As a byproduct, such a minimization provide a projection of the empirical probability distribution on the set of probability distributions conformable to the moment restrictions. Among the variety of power divergence statistics, the Euclidean Empirical Likelihood, based on a chi-square distance, is the only one yielding a closed-form formula for projected (or implied) probabilities. As far as estimation of θ is concerned, all the estimators resulting from a minimization of a power diver-

gence statistics are first-order asymptotically equivalent. However, consideration of higher order asymptotics points out better properties for empirical likelihood.

We focus in section 3 on the case of Euclidean Empirical Likelihood. We show that the corresponding implied probabilities amount to estimate population expectations by using the overidentifying restrictions as control variables. Moreover, implied probabilities are asymptotically all positive with probability one. A simple shrinkage procedure solves the negativity problem in finite sample. In terms of estimation of θ, it is shown that Euclidean Empirical Likelihood minimization coincides with CUE-GMM. Moreover we propose a three-step estimator, the computation of which does not involve more than quadratic programming while its higher order asymptotics properties coincide with the ones of Empirical Likelihood.

Section 4 is devoted to extend the results of previous sections to the case of conditional moment restrictions (1.2). We first consider the general issue of minimization of a localized version of power divergence statistics, when the object of interest are now conditional probabilities of the values X_j given Z_i. The idea is a projection of the kernel smoothed version of the conditional empirical probability distribution on the set of conditional probability distributions conformable to the conditional moment restrictions. We show that, in terms of estimation of θ, minimization of such localized divergence subject to the conditional moment restrictions amounts to a nonparametric estimation of optimal instruments. But, by contrast with naïve kernel smoothing, this estimation takes advantage of the informational content of conditional moment restrictions. In the case of conditional Euclidean Empirical Likelihood, the improvement amounts to use the moment restrictions as conditional control variables. Several versions of a conditional extension of CUE-GMM, which are suggested by the profile criterion of conditional Euclidean Empirical Likelihood, are discussed. A three step extension is also proposed to get the same higher order properties as Empirical Likelihood.

Section 5 concludes. The main proofs are gathered in a appendix.

2 Implied probabilities in minimum discrepancy estimators

2.1 The first order conditions

To describe the estimators, let $X_i, (i = 1, \cdots, n)$ be i.i.d. observations on a random vector X. Consider the moment indicator $\Psi(X, \theta) = \left(\Psi^j(X, \theta)\right)_{1 \leq j \leq q}$, a q-vector of functions of the data observation X and the p-vector θ of unknown parameters, with $q \geq p$. It is assumed that the true parameter vector θ^0 satisfies the moment conditions:

$$E\left[\Psi\left(X, \theta^0\right)\right] = 0, \theta^0 \epsilon \Theta \subset \mathbb{R}^q. \tag{2.1}$$

Following Corcoran (1998), let us consider the optimization problem:

$$\begin{aligned} &\underset{\pi_1,\cdots,\pi_n,\theta}{Min} \sum_{i=1}^{n} h\left(\pi_i\right) \\ &\sum_{i=1}^{n} \pi_i \Psi\left(X_i,\theta\right) = 0 \\ &\sum_{i=1}^{n} \pi_i = 1 \end{aligned} \tag{2.2}$$

where $h(\pi)$ is a differentiable convex function of a nonnegative scalar π that measures the discrepancy between π and the empirical probability $1/n$ of a single observation, that can depend on n. Typically, when the optimization problem (2.2) admits a unique solution $\hat{\pi}_1,\cdots,\hat{\pi}_n,\hat{\theta}$ with nonnegative $\hat{\pi}_i s$, these can be interpreted as probabilities that minimize the discrepancy with the empirical measure subject to moment conditions.

The following result will allow us to relate minimum discrepancy estimators to standard theory of estimating equations:

Theorem 2.1 *Assume that (2.2) uniquely defines estimators $\hat{\pi}_1,\cdots,\hat{\pi}_n,\hat{\theta}$ with nonnegative $\hat{\pi}_i s$. Then $\hat{\theta}$ is characterized as solution of the first order conditions:*

$$\left[\sum_{i=1}^{n} \hat{\pi}_i \frac{\partial \Psi_i'}{\partial \theta}(\hat{\theta})\right]\left[\sum_{i=1}^{n} \hat{\pi}_i \Psi_i(\hat{\theta})\Psi_i'(\hat{\theta})\right]^{-1} \sum_{i=1}^{n} h_\pi\left(\hat{\pi}_i\right)\hat{\pi}_i \Psi_i\left(\hat{\theta}\right) = 0 \tag{2.3}$$

where $\Psi_i(\theta)$ denotes $\Psi(X_i,\theta)$ and $h_\pi(\cdot)$ the first derivative of the function h.

Note that the required existence and unicity of a solution of (2.2) is likely to be fulfilled for large n, under standard regularity conditions, insofar as the moment conditions satisfy the following identification assumption which will be maintained hereafter:

Assumption 2.1 *(i)* $E\Psi\left(X,\theta\right) = 0 \Longrightarrow \theta = \theta^0$.
(ii) $\Gamma\left(\theta^0\right) = E\dfrac{\partial \Psi}{\partial \theta'}(X,\theta)_{|\theta=\theta^0}$ *is of rank* $p = \dim\ \theta$.
(iii) $\Omega\left(\theta\right) = E\left[\Psi\left(X,\theta\right)\Psi'(X,\theta)\right]$ *non singular matrix for all* $\theta \epsilon \Theta$.

Another approach to combining estimating functions is to consider p-dimensional vectors of estimating functions $\varphi\left(X,\theta\right) = A(\theta)\Psi(X,\theta)$ (where $A(\theta)$ is a $p \times q$ matrix of real functions of θ) which are linear combinations of the q estimating functions $\Psi^j(X,\theta), j = 1,\cdots,q$. In estimating function theory (e.g. see Godambe and Thompson (1989)), an estimating function $\varphi^*\left(X,\theta\right)$ is

called optimum if the estimator $\hat{\theta}_n$ from $\frac{1}{n}\sum_{i=1}^{n}\varphi^*\left(X_i,\hat{\theta}_n\right)=0$ has minimum asymptotic variance. It is known that the optimal linear combination is given by:

$$\varphi^*(X,\theta)=\Gamma'(\theta)\,\Omega^{-1}(\theta)\Psi(X,\theta). \tag{2.4}$$

The matrices $\Gamma\left(\theta^0\right)$ and $\Omega\left(\theta^0\right)$ are unknown but can be consistently estimated at any value of θ, respectively by:

$$\Gamma_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\Psi_i}{\partial\theta'}(\theta)$$

and

$$\Omega_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\Psi_i(\theta)\,\Psi_i'(\theta).$$

Then, the most common strategy is to evaluate these matrices at any first step consistent estimate $\tilde{\theta}_n$ of θ^0. We then get the consistent asymptotically normal estimator θ_n^* with minimum asymptotic variance by solving:

$$\Gamma_n'\left(\tilde{\theta}_n\right)\Omega_n^{-1}\left(\tilde{\theta}_n\right)\bar{\Psi}_n(\theta_n^*)=0 \tag{2.5}$$

where

$$\bar{\Psi}_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\Psi_i(\theta).$$

More generally, we are going to call in the sequel asymptotically efficient estimator any estimator $\hat{\theta}_n$ of θ^0 first-order asymptotically equivalent to θ_n^*, that is such that

$$\sqrt{n}\left(\hat{\theta}_n-\theta_n^*\right)=o_P(1) \tag{2.6}$$

It is easy to check that a necessary and sufficient condition for (2.6) is:

$$\sqrt{n}\left(\hat{\theta}_n-\theta^0\right)=-\Sigma^{-1}\Gamma'\Omega^{-1}\sqrt{n}\bar{\Psi}_n\left(\theta^0\right)+o_P(1) \tag{2.7}$$

with $\Sigma=(\Gamma'\Omega^{-1}\Gamma)$. In such expressions, the matrices Γ,Ω and Σ should actually be denoted $\Gamma\left(\theta^0\right)$, $\Omega\left(\theta^0\right)$ and $\Sigma\left(\theta^0\right)$. The dependence on the true unknown value θ^0 is omitted for the sake of notational simplicity.

Among the asymptotically efficient estimators of θ^0, one may also consider the one-step estimator $\hat{\theta}_n$ defined by:

$$\Gamma_n'(\hat{\theta}_n)\Omega^{-1}\left(\hat{\theta}_n\right)\bar{\Psi}_n\left(\hat{\theta}_n\right)=0. \tag{2.8}$$

Although in general less computationally convenient than θ_n^*, $\hat{\theta}_n$ defined by (2.9) is worth comparing to the minimum discrepancy estimators $\hat{\theta}$ defined by theorem 2.1. One similarity and two differences are striking: In both cases, a left multiplication by a consistent estimator of the optimal selection matrix $\Gamma'(\theta^0)\Omega^{-1}\left(\theta^0\right)$ is applied to some weighting average of the sample estimating functions. However, both these consistent estimators and these weighted averages are different in general. First, while the common efficient estimation strategies (2.5) or (2.8) make use of the unconstrained estimates $\Gamma_n(\theta)$ and $\Omega_n(\theta)$, the minimum discrepancy estimator resorts to constrained estimates:

$$\Gamma^{\bar{\pi}}(\theta) = \sum_{i=1}^{n} \hat{\pi}_i \frac{\partial \Psi_i}{\partial \theta'}(\theta)$$

and

$$\Omega^{\bar{\pi}}(\theta) = \sum_{i=1}^{n} \hat{\pi}_i \Psi_i(\theta) \Psi_i'(\theta).$$

One may expect that the latter are more accurate than the former since they take advantage of the information provided by the estimating equations. More precisely, $\Gamma^{\bar{\pi}}\left(\hat{\theta}\right)$ and $\Omega^{\bar{\pi}}\left(\hat{\theta}\right)$ are sample counterparts of the population moments $\Gamma\left(\theta^0\right)$ and $\Omega\left(\theta^0\right)$ which are computed with sample weights $\hat{\pi}_i$ that are by definition conformable to the moment restrictions: $\sum_{i=1}^{n} \hat{\pi}_i \Psi_i\left(\hat{\theta}\right) = 0$. By contrast, $\Gamma_n(\theta)$ and $\Omega_n(\theta)$ are computed from the empirical distribution, that is equally weighted observations with weights 1/n, which are in general inconsistent with the moment restrictions:

$$\frac{1}{n}\sum_{i=1}^{n} \Psi_i(\theta) = \bar{\Psi}_n(\theta) \neq 0 \text{ for all } \theta.$$

The second difference between (2.5)/(2.8) and minimum discrepancy estimators is that the consistent estimate of the optimal selection matrix is applied to two different weighted averages of the sample estimating functions. While (2.5)/(2.8) resorts to the common empirical mean $\bar{\Psi}_n(\theta)$, the minimum discrepancy estimator is computed from a more bizarre weighted average, namely $\sum_{i=1}^{n} h_\pi(\hat{\pi}_i)\hat{\pi}_i \Psi_i(\hat{\theta})$. However, the two estimators are going to coincide when $h_\pi(\pi_i)$ is proportional to $1/\pi_i$, that is when the chosen discrepancy function $h(\pi)$ is an affine function of $Log(\pi)$. This particular case corresponds to the so-called *empirical likelihood* (EL) estimator, as first characterized by Qin and Lawless (1994), who already put forward its asymptotic efficiency by reference to the theory of optimal estimating functions.

We are going to focus more generally in all the sequel on homogeneous discrepancy functions: $h_\pi(\pi_i)$ proportional to $\pi_i^{-\lambda}$ for some non-zero real number λ.

2.2 Implied probabilities associated to power–divergence statistics

Cressie and Read (1984) introduced a family of power-divergence statistics as:

$$I_\lambda = \frac{1}{\lambda(\lambda-1)} \sum_{i=1}^{n} \left[(n\pi_i)^{1-\lambda} - 1 \right], \tag{2.9}$$

defined for any real λ, including the two limit cases $\lambda \longrightarrow 0$ and $\lambda \longrightarrow 1$. For $\lambda \notin \{0,1\}$, the minimization of the divergence I_λ with respect to $(\pi_i)_{1\leq i\leq n}$ under the constraints:

$$\begin{cases} \sum\limits_{i=1}^{n} \pi_i = 1 \\ \sum\limits_{i=1}^{n} \pi_i \Psi(X_i, \theta) = 0 \end{cases}$$

is obviously equivalent to the minimum discrepancy optimization problem (2.2) with an homogeneous discrepancy function:

$$h(\pi) = \pi^{1-\lambda} \tag{2.10}$$

Notice that, for $0 < \lambda < 1$, one must actually consider $h(\pi) = -\pi^{1-\lambda}$ to get a convex discrepancy function. This change of sign does not play any role in the first order conditions of interest and will not be made explicit in the sequel. The empirical likelihood case ($h(\pi) = Log\pi$) is also included in this framework by the limit case $\lambda \longrightarrow 1$:

$$\underset{\lambda \longrightarrow 1}{Lim} \frac{1}{\lambda - 1} \left[(n\pi_i)^{1-\lambda} - 1 \right] = -Log(n\pi_i).$$

However, the limit case $\lambda \longrightarrow 0$ $\left(I_0 = -\sum\limits_{i=1}^{n} \pi_i Log\pi_i \right)$ is not included since it does not correspond to any discrepancy function $h(\pi)$ with $h_\pi(\pi_i)$ proportional to $\pi_i^{-\lambda}$. This is the reason why the so-called exponential tilting estimator as studied by Kitamura and Stutzer (1997) will not be considered here.

We are going to focus on all other estimators $\left((\hat{\pi}_{i,\lambda})_{1\leq i\leq n}, \hat{\theta}_{n,\lambda} \right)$ associated to some Cressie and Read power divergence statistics I_λ, for some $\lambda \neq 0$.

By application of theorem 2.1, we know that the estimator $\hat{\theta}_{n,\lambda}$ is characterized by a set of p first order conditions that can be written:

$$\left\{ \hat{E}_{n,\lambda} \left[\frac{\partial \Psi_i'}{\partial \theta} \left(\hat{\theta}_{n,\lambda} \right) \right] \right\} \left\{ \hat{V}_{n,\lambda} \left[\Psi_i(\hat{\theta}_{n,\lambda}) \right] \right\}^{-1} \sum_{i=1}^{n} (\hat{\pi}_{i,\lambda})^{1-\lambda} \Psi_i \left(\hat{\theta}_{n,\lambda} \right) = 0 \tag{2.11}$$

where sample counterparts of population moments are defined by:

$$\hat{E}_{n,\lambda} g(X) = \sum_{i=1}^{n} \hat{\pi}_{i,\lambda} g(X_i) \tag{2.12}$$

$$\hat{V}_{n,\lambda} g(X) = \sum_{i=1}^{n} \hat{\pi}_{i,\lambda} \left[g(X_i) - \hat{E}_{n,\lambda} g(X) \right] g'(X_i).$$

As far as the estimates $\hat{\pi}_{i,\lambda}$ are concerned, they are characterized by the following first order conditions:

Theorem 2.2 *There exist a non-zero real number $\mu_{n,\lambda}$ and a vector $\alpha_{n,\lambda}$ of q reduced Lagrange multipliers such that:*

$$\hat{\pi}_{i,\lambda}^{-\lambda} = \mu_{n,\lambda} \left[1 + \alpha'_{n,\lambda} \Psi_i \left(\hat{\theta}_{n,\lambda} \right) \right].$$

The components of $\alpha_{n,\lambda}$ are termed "reduced Lagrange multipliers" since the product $\mu_{n,\lambda}\alpha_{n,\lambda}$ is actually the vector of Lagrange multipliers associated to the moment restrictions. The real number $\mu_{n,\lambda}$, Lagrange multiplier associated to the constraint $\sum_{i=1}^{n} \pi_i = 1$, is non-zero. The following lemma about the reduced Lagrange multipliers $\alpha_{n,\lambda}$ will often be useful:

Lemma 2.3: $\sqrt{n}\alpha_{n,\lambda} = \lambda \Omega_n^{-1} \left(\hat{\theta}_{n,\lambda} \right) \sqrt{n} \bar{\Psi}_n \left(\hat{\theta}_{n,\lambda} \right) + \mathcal{O}_P(1/\sqrt{n})$

Note that lemma 2.3 means in particular that $\alpha_{n,\lambda} = \mathcal{O}_P(1/\sqrt{n})$. This property is important since it implies that the quantities $\left[1 + \alpha'_{n,\lambda} \Psi_i \left(\hat{\theta}_{n,\lambda} \right) \right]$ are almost surely positive for large n. More precisely:

Theorem 2.4 *For all $\theta \epsilon \Theta$ and $\varepsilon > 0$, $P \left[\underset{1 \leq i \leq n}{Min} \left[1 + \alpha'_{n,\lambda} \Psi_i(\theta) \right] > 1 - \varepsilon \right] \underset{n=\infty}{\longrightarrow} 1$*

Theorem 2.4 leads easily to three useful corollaries:

Corollary 2.5: *Asymptotically almost certainly:*

$$\hat{\pi}_{i,\lambda} = \frac{\left[1 + \alpha'_{n,\lambda} \Psi_i \left(\hat{\theta}_{n,\lambda} \right) \right]^{-1/\lambda}}{\sum_{j=1}^{n} \left[1 + \alpha'_{n,\lambda} \Psi_j \left(\hat{\theta}_{n,\lambda} \right) \right]^{-1/\lambda}}.$$

Then, first order conditions (2.11) can be rewritten:

Corollary 2.6: *The estimator $\hat{\theta}_{n,\lambda}$ is, asymptotically almost certainly, characterized by the p first order conditions:*

$$\left\{ \hat{E}_{n,\lambda} \left[\frac{\partial \Psi'_i}{\partial \theta} (\hat{\theta}_{n,\lambda}) \right] \right\} \left\{ \hat{V}_{n,\lambda} \left[\Psi_i \left(\hat{\theta}_{n,\lambda} \right) \right] \right\}^{-1} \sum_{i=1}^{n} \left[1 + \alpha'_n \Psi_i \left(\hat{\theta}_{n,\lambda} \right) \right]^{\frac{\lambda-1}{\lambda}} \cdot \Psi_i \left(\hat{\theta}_{n,\lambda} \right) = 0$$

To interpret these first order conditions, it is worth noticing that by lemma 2.3:

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[1+\alpha'_{n,\lambda}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\right]^{\frac{\lambda-1}{\lambda}}\cdot\Psi_i\left(\hat{\theta}_{n,\lambda}\right) \\
= & \sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)+\frac{\lambda-1}{\lambda}\cdot\frac{1}{n}\sum_{i=1}^{n}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\Psi'_i\left(\hat{\theta}_{n,\lambda}\right)\alpha_n+\mathcal{O}_P\left(1/\sqrt{n}\right) \\
= & \sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)+(\lambda-1)\sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)+\mathcal{O}_P\left(1/\sqrt{n}\right) \\
= & \lambda\sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)+\mathcal{O}_P\left(1/\sqrt{n}\right).
\end{aligned}$$

This first order expansion shows that the first order conditions of corollary 2.6 can be seen as a consistent estimation of optimal estimating equations (2.4), that is:

$$\Gamma'(\theta^0)\Omega^{-1}\left(\theta^0\right)\sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)=\mathcal{O}_P\left(1/\sqrt{n}\right).$$

By (2.7) and lemma 2.3, this leads to:

Corollary 2.7: *For all* $\lambda\epsilon R^*$,
$\sqrt{n}\left(\hat{\theta}_{n,\lambda}-\theta^0\right)=-\Sigma^{-1}\Gamma'\Omega^{-1}\sqrt{n}\bar{\Psi}_n\left(\theta^0\right)+\mathcal{O}_P\left(1/\sqrt{n}\right)$
$\sqrt{n}\alpha_{n,\lambda}=\lambda P\sqrt{n}\bar{\Psi}_n\left(\theta^0\right)+\mathcal{O}_P\left(1/\sqrt{n}\right)$
$\sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)=\Omega P\sqrt{n}\bar{\Psi}_n(\theta^0)+\mathcal{O}_P\left(1/\sqrt{n}\right)$
with
$\Sigma=\Gamma'\Omega^{-1}\Gamma$
and $P=\Omega^{-1}-\Omega^{-1}\Gamma\Sigma^{-1}\Gamma'\Omega^{-1}$.

Corollary 2.7 implies in particular that for all $\lambda\epsilon\mathbb{R}^*$, the estimator $\hat{\theta}_{n,\lambda}$ associated to the power divergence statistics (2.9) is asymptotically efficient. This property is actually well known, at least since Imbens, Spady and Johnson (1998). In order to better characterize the difference between various estimators associated to various choices of λ, we are going to consider now higher order expansions of first order conditions.

2.3 Stochastic expansions of first order conditions

Theorem 2.8 below provides an higher order expansion of the first order conditions put forward by corollary 2.6:

Theorem 2.8

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[1+\alpha'_{n,\lambda}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\right]^{\frac{\lambda-1}{\lambda}}\Psi_i(\hat{\theta}_{n,\lambda}) \\
= \quad &\lambda\sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)+\frac{\lambda(\lambda-1)}{2}\frac{1}{\sqrt{n}}\sum_{j=1}^{q}\beta_{jn}(\theta^0)e_j+\mathcal{O}_P\left(1/n\right)
\end{aligned}$$

where for $j=1,\cdots,q$, e_j denotes the q-dimensional vector with all coefficients equal to zero, except the j^{th} one equal to one, and:

$$\beta_{jn}(\theta^0)=\sqrt{n}\bar{\Psi}'_n\left(\theta^0\right)P'E\left[\Psi(X,\theta^0)\Psi'\left(X,\theta^0\right)\Psi^j\left(X,\theta^0\right)\right]P\sqrt{n}\bar{\Psi}_n\left(\theta^0\right)=\mathcal{O}_P\left(1\right)$$

In other words, for all $\lambda\neq 0$, $\hat{\theta}_{n,\lambda}$ is defined by first order conditions stochastically expanded as:

$$\left\{\hat{E}_{n,\lambda}\frac{\partial\Psi'_i}{\partial\theta}(\hat{\theta}_{n,\lambda})\right\}\left\{\hat{V}_{n,\lambda}\left[\Psi_i(\hat{\theta}_{n,\lambda})\right]\right\}^{-1}\left\{\sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)+\frac{\lambda-1}{2\sqrt{n}}\sum_{j=1}^{q}\beta_{jn}(\theta^0)e_j\right\}=\mathcal{O}_P\left(1/n\right) \tag{2.13}$$

It means that, up to some $\mathcal{O}_P\left(1/n\right)$, the first order conditions defining $\hat{\theta}_{n,\lambda}$ are different from the ones of empirical likelihood (case $\lambda=1$) when the random variable $\sum_{j=1}^{q}\beta_{jn}(\theta^0)e_j$ is not zero in probability.

Note that each $\beta_{jn}(\theta^0), j=1,\cdots,q$, is asymptotically a quadratic form on a standardized Gaussian vector of dimension q: $T_n\left(\theta^0\right)=\Omega^{-1/2}\left(\theta^0\right)\sqrt{n}\bar{\Psi}_n(\theta^0)$. This quadratic form is zero if and only if:

$$E\left[\Psi(X,\theta^0)\Psi'\left(X,\theta^0\right)\Psi^j\left(X,\theta^0\right)\right]P=0.$$

This condition is fulfilled in particular when third moments of $\Psi\left(X,\theta^0\right)$ are zero:

$$E\left[\Psi^j\left(X,\theta^0\right)\Psi^k(X,\theta^0)\Psi^l(X,\theta^0)\right]=0 \text{ for } j,k,l=1,\cdots,q. \tag{2.14}$$

As noticed by Newey and Smith (2004), this third moment condition will hold in an IV setting, when disturbances are symmetrically distributed. More precisely, Newey and Smith (2004) (see their theorem 4.1 and their comments after corollary 4.4) state that:

First, the higher order bias of $\hat{\theta}_{n,\lambda}$ with respect to empirical likelihood is zero when:

$$E\left[\Psi\left(X,\theta^0\right)\Psi'\left(X,\theta^0\right)P\Psi\left(X,\theta^0\right)\right]=0. \tag{2.15}$$

Second, when the zero third moments condition holds, one can actually show the stronger result that:

$$\hat{\theta}_{n,\lambda}-\hat{\theta}_{n,1}=\mathcal{O}_P\left(n^{-3/2}\right). \tag{2.16}$$

This stronger result will be shown in section 3 below to be a corollary of theorem 2.8 when the $\beta_{jn}\left(\theta^0\right), j = 1, \cdots, q$ are all zero, that is:

$$E\left[\Psi\left(X, \theta^0\right) \Psi'\left(X, \theta^0\right) P\Psi^j\left(X, \theta^0\right)\right] = 0 \text{ for } j = 1, \cdots, q \tag{2.17}$$

Note that this condition, although possibly slightly weaker than the zero third moments condition, is actually stronger than the Newey and Smith's zero-bias condition (2.15).

3 Euclidean Empirical Likelihood

3.1 Continuously updated GMM

As already announced in section 2, higher order properties of the estimators $\hat{\theta}_{n,\lambda}$ should lead to prefer the case $\lambda = 1$ (empirical likelihood) among all the possible Cressie-Read divergence statistics. However, this case may be computationally demanding. To see this, let us just remind that by corollary 2.5, implied probabilities in that case are asymptotically almost certainly proportional to $\left[1 + \alpha'_{n,1}\Psi_i\left(\hat{\theta}_{n,1}\right)\right]^{-1}$. Then, the vector $\alpha_{n,1}$ of Lagrange multipliers should be computed as solution of the following system of q nonlinear equations:

$$\sum_{i=1}^{n} \frac{\Psi_i\left(\hat{\theta}_{n,1}\right)}{1 + \alpha'_{n,1}\Psi_i(\hat{\theta}_{n,1})} = 0. \tag{3.1}$$

Convex duality is useful to solve these nonlinear equations (see Owen (2001) section 3.14 for details) since (3.1) can be seen as first order conditions of the following convex minimization program:

$$\underset{\alpha}{Min} - \sum_{i=1}^{n} Log\left[1 + \alpha'\Psi_i\left(\hat{\theta}_{n,1}\right)\right]. \tag{3.2}$$

The main difficulty is that the dimension of this optimization problem is q, the dimension of Ψ. This may be actually a very high dimensional problem, in particular in the case of conditional moment restrictions. As shown by Kitamura, Tripathi and Ahn (2000) (see also section 4 below), nonparametric smoothing of conditional expectations at each point of the observed sample leads to an effective number q of constraints (dimension of the vector α of Lagrange multipliers) proportional to the sample size.

This gives a strong motivation to look for a less computationally demanding estimator because the vector $\alpha_{n,\lambda}$ of Lagrange multipliers will be easier to recover. The simplest case is $(\lambda = -1)$ since, by corollary 2.5, the Lagrange multipliers $\alpha_{n,-1}$ are determined as solutions of the system of q linear equations:

$$\sum_{i=1}^{n} \left[1 + \alpha'_{n,-1}\Psi_i\left(\hat{\theta}_{n,-1}\right)\right] \Psi_i\left(\hat{\theta}_{n,-1}\right) = 0 \tag{3.3}$$

From (2.9), the case ($\lambda = -1$) corresponds to the so-called Euclidean empirical likelihood:

$$I_{-1} = \frac{1}{2}\sum_{i=1}^{n}\left[(n\pi_i)^2 - 1\right]. \tag{3.4}$$

Notice that I_{-1} is well defined even if some π_i s are negative. For any given θ, its minimization with respect to $(\pi_i)_{1\leq i\leq n}$ under the constraints (2.2) is a quadratic program under linear restrictions which defines profile functions $\pi_i(\theta), i = 1, \cdots, n$. We will denote by $\hat{\theta}_n^Q = \hat{\theta}_{n,-1}$ the Euclidean empirical likelihood estimator of θ.

We first characterize the profile implied probabilities in function of the two alternative estimators of the covariance function $\Omega\left(\theta^0\right) = E\left[\Psi\left(X, \theta^0\right)\Psi'\left(X, \theta^0\right)\right]$:

$$\text{Uncentered second moments:} \qquad \Omega_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\Psi_i(\theta)\Psi_i'(\theta) \tag{3.5}$$

$$\text{Second moments in mean deviation form:} \quad V_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\Psi_i(\theta) - \bar{\Psi}_n(\theta)\right]\Psi_i'\theta \tag{3.6}$$

Theorem 3.1 *For all $\theta\epsilon\Theta$ and $i = 1, \cdots, n$:*
$\pi_i(\theta)$ is proportional to $1 + \alpha_n'(\theta)\Psi_i(\theta)$ and to $1 + \gamma_n'(\theta)\left[\Psi_i(\theta) - \bar{\Psi}_n(\theta)\right]$
with:

$$\alpha_n(\theta) = -\Omega_n^{-1}(\theta)\bar{\Psi}_n(\theta) \text{ and } \gamma_n(\theta) = -V_n^{-1}(\theta)\bar{\Psi}_n(\theta).$$

In particular:

$$\pi_i(\theta) = \frac{1}{n} - \frac{1}{n}\bar{\Psi}_n'(\theta)V_n^{-1}(\theta)\left[\Psi_i(\theta) - \bar{\Psi}_n(\theta)\right].$$

Note that by contrast with any other Cressie-Read divergence statistics, the Euclidean likelihood provides closed form formulas for implied probabilities $\pi_i(\theta), i = 1, \cdots, n$. This convenience rests upon the linearity of equations like (3.3) to determine Lagrange multipliers. Moreover, the almost sure positivity property of theorem 2.4 allows us to state:

Corollary 3.2: *Asymptotically almost certainly:*

$$\hat{\pi}_{i,-1} = \pi_i\left(\hat{\theta}_n^Q\right) \geq 0 \text{ for all } i.$$

The profile functions $\pi_i(\theta), i = 1, \cdots, n$, give us two interesting characterizations of the profile criterion defining the Euclidean empirical likelihood estimator $\hat{\theta}_n^Q$:

Corollary 3.3:

$$\sum_{i=1}^{n} \pi_i^2(\theta) = \frac{1}{n}\left[1 + \bar{\Psi}_n'(\theta) V_n^{-1}(\theta) \bar{\Psi}_n(\theta)\right] = \frac{1}{n}\left[1 - \bar{\Psi}_n'(\theta) \Omega_n^{-1}(\theta) \bar{\Psi}_n(\theta)\right]^{-1}.$$

Corollary 3.3 shows that the continuous updating estimator (CUE) of Hansen, Heaton and Yaron (1996) numerically coincides with the Euclidean empirical likelihood estimator $\hat{\theta}_n^Q$. However, while closed-form formulas for CUE-GMM may be computationally involved when analyzed at the profile level, they are amazingly straightforward in the Euclidean empirical likelihood framework. For instance, the simple fact that implied probabilities $\pi_i(\theta)$ are proportional to both $[1 + \alpha_n'(\theta)\Psi_i(\theta)]$ and $\left[1 + \gamma_n'(\theta)\left(\Psi_i(\theta) - \bar{\Psi}_n(\theta)\right)\right]$ implies the important relationship:

$$1 + Q^V(\theta) = \left[1 - Q^\Omega(\theta)\right]^{-1} \tag{3.7}$$

between the two possible forms of the criterion to minimize for CUE-GMM:

$$Q^V(\theta) = \bar{\Psi}_n'(\theta) V_n^{-1}(\theta) \bar{\Psi}_n(\theta) \tag{3.8}$$

or

$$Q^\Omega(\theta) = \bar{\Psi}_n'(\theta) \Omega_n^{-1}(\theta) \bar{\Psi}_n(\theta). \tag{3.9}$$

Newey and Smith (2004) already mentioned relationship (3.7). Even more importantly, the fact that both forms $\Omega^{-1}(\theta)$ or $V_n^{-1}(\theta)$ of the weighting matrix, in uncentered moments form or in mean deviation form, are valid for CUE-GMM, is true not only at the criterion level:

$$\underset{\theta}{Min} Q^V(\theta) \Longleftrightarrow \underset{\theta}{Min} Q^\Omega(\theta) \tag{3.10}$$

but also in terms of first order conditions. While the latter is far to be an obvious implication of the former, it is stated by corollary 3.4. below:

Corollary 3.4: The Euclidean empirical likelihood estimator (CUE-GMM) $\hat{\theta}_n^Q$ is characterized as solution of any of the two following systems of first order conditions:

i) $$\left[\sum_{i=1}^{n} \pi_i(\hat{\theta}_n^Q) \frac{\partial \bar{\Psi}_i'}{\partial \theta}(\hat{\theta}_n^Q)\right] \Omega_n^{-1}(\hat{\theta}_n^Q) \bar{\Psi}_n(\hat{\theta}_n^Q) = 0$$

ii) $$\left[\sum_{i=1}^{n} \pi_i(\hat{\theta}_n^Q) \frac{\partial \bar{\Psi}_i'}{\partial \theta}(\hat{\theta}_n^Q)\right] V_n^{-1}(\hat{\theta}_n^Q) \bar{\Psi}_n(\hat{\theta}_n^Q) = 0$$

Corollary 3.4 is still a straightforward implication of the two possible forms $\alpha_n(\theta)$ and $\gamma_n(\theta)$ of the vector of reduced Lagrange multipliers. Newey and Smith (2004), theorem 2.3, put forward the first form of these first order conditions. In comparing the empirical likelihood first order conditions ((2.11) with $\lambda = 1$) and CUE-GMM first order conditions, they stress that, while both use the relevant constrained estimator of the Jacobian matrix $\Gamma\left(\theta^0\right)$ by taking into account implied probabilities $\hat{\pi}_{i,\lambda}$, CUE-GMM has the drawback to use an unconstrained estimator of the weighting matrix $\Omega\left(\theta^0\right)$. This criticism may however be mitigated in two respects:

First, as shown by (2.13), the difference between the two estimators can be interpreted in a different way, where the weighting matrix is well-estimated in both cases, but symmetry properties of the moment conditions are at stake.

Second, the form (ii) of first order conditions in corollary 3.4 shows that the weighting matrix estimator may be seen in its mean deviation form. It is important to stress that this is a second advantage, besides the estimation of the Jacobian matrix, of CUE-GMM with respect to common use of two-stage GMM (2S-GMM).

Two-stage GMM is actually defined, since Hansen (1982), as the minimization over θ of:

$$Q^{*\Omega}(\theta) = \bar{\Psi}_n'(\theta)\,\Omega_n^{-1}\left(\tilde{\theta}_n\right)\bar{\Psi}_n(\theta) \tag{3.11}$$

for a given consistent first step estimator $\tilde{\theta}_n$ of θ.

However, Hall (2000) argues that the mean deviation form should be preferred, leading to the minimization over θ of:

$$Q^{*V}(\theta) = \bar{\Psi}_n'(\theta)\,V_n^{-1}\left(\tilde{\theta}_n\right)\bar{\Psi}_n(\theta) \tag{3.12}$$

First order conditions associated to (3.11) define a two-step GMM estimator $\hat{\theta}_n^{2S\Omega}$ as solution of:

$$\frac{\partial\bar{\Psi}_n'}{\partial\theta}\left(\hat{\theta}_n^{2S\Omega}\right)\Omega_n^{-1}\left(\tilde{\theta}_n\right)\bar{\Psi}_n\left(\hat{\theta}_n^{2S\Omega}\right) = 0. \tag{3.13}$$

Another two-step GMM estimator $\hat{\theta}_n^{2SV}$ is defined by the first order conditions of (3.12):

$$\frac{\partial\bar{\Psi}_n'}{\partial\theta}\left(\hat{\theta}_n^{2SV}\right)V_n^{-1}\left(\tilde{\theta}_n\right)\bar{\Psi}_n\left(\hat{\theta}^{2SV}\right) = 0. \tag{3.14}$$

It is worth realizing that, by contrast with corollary 3.4, there is no reason to imagine that equations (3.13) and (3.14) are equivalent. In other words, the common 2S-GMM estimator $\hat{\theta}_n^{2SG}$ should have less nice properties than CUE-GMM not only because it uses more biased estimator of the Jacobian matrix but also because it does not use the estimator of the covariance matrix in its mean deviation form. By contrast, the 2S-GMM estimator in mean deviation form $\hat{\theta}_n^{2SV}$ is expected to have better properties and actually coincides with CUE-GMM in the particular case of separable moment conditions:

$$\Psi(X,\theta) = \varphi(X) - k(\theta). \tag{3.15}$$

Notice that in such a case, there is no issue of estimation of the Jacobian matrix. As far as the bias in the estimation of this matrix is concerned for more general moment conditions, it is worth interpreting the constrained estimator $\sum_{i=1}^{n} \pi_i \left(\hat{\theta}_n^Q\right) \frac{\partial \Psi_i}{\partial \theta'} \left(\hat{\theta}_n^Q\right)$ used by CUE-GMM (see corollary 3.4) in the light of corollary 3.5 below:

Corollary 3.5: *For any integrable real function $g(X)$, $Eg(X)$ can be estimated by:*

$$\begin{aligned} \hat{g}_n\left(\hat{\theta}_n^Q\right) &= \sum_{i=1}^{n} \pi_i \left(\hat{\theta}_n^Q\right) g(X_i) \\ &= \bar{g}_n - Cov_n\left[g(X_i), \Psi_i(\hat{\theta}_n^Q)\right] \left[V_n\left(\hat{\theta}_n^Q\right)\right]^{-1} \bar{\Psi}_n(\hat{\theta}_n^Q) \end{aligned}$$

where:

$$\begin{aligned} \bar{g}_n &= \frac{1}{n}\sum_{i=1}^{n} g\left(X_i\right) \\ Cov_n\left[g(X_i), \Psi_i(\hat{\theta}_n^Q)\right] &= \frac{1}{n}\sum_{i=1}^{n} \left[g\left(X_i\right) - \bar{g}_n\right] \Psi_i'(\hat{\theta}_n^Q) \end{aligned}$$

The intuition behind the estimator $\hat{g}_n\left(\hat{\theta}_n^Q\right)$ is very clear. If we knew the true value θ^0 of θ, we would get an unbiased estimator of $Eg\left(X\right)$ by considering $\bar{g}_n - a'\bar{\Psi}_n\left(\theta^0\right)$ for any given q-dimensional vector a. The minimum variance estimator is obtained for:

$$a = Cov\left[g\left(X\right), \Psi\left(X, \theta^0\right)\right] \left(Var\left[\Psi\left(X, \theta^0\right)\right]\right)^{-1} \tag{3.16}$$

and it will be made feasible by replacing a by its sample counterpart:

$$\hat{a}_n = Cov_n\left[g\left(X_i\right), \Psi_i\left(\theta^0\right)\right] \left[V_n\left(\theta^0\right)\right]^{-1}. \tag{3.17}$$

This is nothing but the well known principle of *control variates* as defined for instance by Fieller and Hartley (1954). Moreover, since we don't know the true value θ^0, the estimator $\hat{g}_n\left(\hat{\theta}_n^Q\right)$ just proposes an extension of the control variates principle where θ^0 is replaced by $\hat{\theta}_n^Q$. In this respect, the choice of Euclidean empirical likelihood to estimate the implied probability distribution appears to be fairly conformable to classical strategies for survey sampling or Monte Carlo experiments. The above control variates interpretation complements the jackknife interpretation of CUE-GMM proposed by Donald and Newey (2000).

3.2 Efficient use of the informational content of estimating equations

The general focus of empirical likelihood kind of approach is the efficient use of the informational content of moment conditions $E\Psi(X,\theta)=0$ about not only the unknown parameters θ but also the unknown probability distribution of X. Our knowledge about the probability distribution of X is actually well encapsulated in our way to estimate $Eg(X)$ for any real function g.

In the particular case of a one-dimensional variable X, Smith (2000) proposes to summarize this knowledge by the *empirical likelihood cumulative distribution function*, that is the estimation of the set of numbers $Eg_a(X)$, $a\in\mathbb{R}$, where g_a denotes the indicator function of the half-line $]-\infty,a]$, that is $g_a(x)=1$ if $x\leq a$, 0 otherwise. More generally, one can for instance use the estimation of $Eg(X)$ for any function g to characterize the probability distribution of X through its Fourier or Laplace transform.

As far as Euclidean empirical likelihood is concerned, we have already shown that it provides a control variates kind of estimator $\hat{g}_n\left(\hat{\theta}_n^Q\right)$ of $Eg(X)$. We are going to show now that this estimator is asymptotically efficient in terms of semiparametric efficiency. Theorem 3.6 characterizes its asymptotic probability distribution:

Theorem 3.6 *For any integrable real function $g(X)$ and $\hat{g}_n\left(\hat{\theta}_n^Q\right)=\sum_{i=1}^{n}\pi_i\left(\hat{\theta}_n^Q\right)g(X_i)$*

we have:

$$\sqrt{n}\left[\hat{g}_n\left(\hat{\theta}_n^Q\right)-Eg(X)\right]\xrightarrow[n=\infty]{d}\mathcal{N}[0,R(g)]$$

where:

$$R(g)=Var\ g(X)-Cov\left[g(X),\Psi\left(X,\theta^0\right)\right]PCov\left[\Psi\left(X,\theta^0\right),g\right]$$

Interpreting theorem 3.6. is straightforward. If θ^0 were known, the control variates estimator of $Eg(X)$, as residual of the affine regression of $g(X)$ on $\Psi\left(X,\theta^0\right)$ would have the asymptotic variance:

$$Var\ g(X)-Cov\left[g(X),\Psi\left(X,\theta^0\right)\right]\Omega\left(\theta^0\right)^{-1}Cov\left[\Psi\left(X,\theta^0\right),g(X)\right].$$

However, since θ^0 is unknown, the efficiency gain with respect to the unconstrained estimator variance $Var\ g(X)$ has to be reduced in proportion of the role of $\hat{\theta}_n^Q$ in the estimation of $\Psi\left(X,\hat{\theta}_n^Q\right)$; this leads to the additional term:

$$\begin{aligned}&Cov\left[g,\Psi\left(X,\theta^0\right)\right]\Omega^{-1}\Gamma\Sigma^{-1}\Gamma'\Omega^{-1}Cov\left[\Psi\left(X,\theta^0\right),g\right]\\ =\ &Cov\left[g,\Psi\left(X,\theta^0\right)\right]\left(\Omega(\theta_0)^{-1}-P\right)Cov\left[\Psi\left(X,\theta^0\right),g\right].\end{aligned}$$

Notice that similar formulas have already been proposed in the empirical likelihood literature (see e.g. Smith (2000) theorem 2 p. 127) but without the control variates interpretation which is specific to Euclidean empirical likelihood. As far as first order asymptotics are concerned, all the Cressie-Read based estimators are actually equivalent. This can be deduced from the following result:

Theorem 3.7 *Let* $\left(\hat{\pi}_{i,\lambda}, i = 1, \cdots, n, \hat{\theta}_{n,\lambda}\right)$ *denote the estimator associated to the power divergence statistics* $I_\lambda, \lambda \neq 0$*, and some moment conditions* $E\Psi(X, \theta) = 0$*. Define the augmented set of moment conditions:* $E\underset{\sim}{\Psi}\left(X, \underset{\sim}{\theta}\right) = 0$*, where:*

$$\begin{aligned} \underset{\sim}{\theta} &= (\theta', \xi)' \\ \underset{\sim}{\Psi}\left(X, \underset{\sim}{\theta}\right) &= \left(\Psi'(X, \theta), g(X) - \xi\right) \end{aligned}$$

for some real integrable function g*.*
Let $\left(\underset{\sim}{\hat{\pi}}_{i,\lambda}, i = 1 \cdots n, \underset{\sim}{\hat{\theta}}_{n,\lambda}\right)$ *the estimator associated to* I_λ *and the augmented set of moment conditions. Then:*

$$\begin{aligned} \underset{\sim}{\hat{\pi}}_{i,\lambda} &= \hat{\pi}_{i,\lambda} \text{ for } i = 1, \cdots n \\ \underset{\sim}{\hat{\theta}}_{n,\lambda} &= \left(\hat{\theta}'_{n,\lambda}, \hat{\xi}_{n,\lambda}\right)' \text{ with} \\ \hat{\xi}_{n,\lambda} &= \sum_{i=1}^{n} \hat{\pi}_{i,\lambda}\, g(X_i) = \hat{E}_{n,\lambda} g(X) \end{aligned}$$

Theorem 3.7 ensures some internal consistency to the estimation approach and, from corollary 2.7, implies the first order asymptotic equivalence of the various estimators of $g(X)$:

Corollary 3.8: *For all* $\lambda \neq 0$

$$\sqrt{n}\left(\hat{E}_{n,\lambda} g(X) - \hat{g}_n(\hat{\theta}_n^Q)\right) = o_P(1)$$

Moreover, the Euclidean likelihood based interpretation of CUE-GMM leads to:

Corollary 3.9: *With the notations of theorem 3.7:*

$$\hat{\xi}_n^Q = \sum_{i=1}^{n} \hat{\pi}_i \left(\hat{\theta}_n^Q\right) g(X_i) = \hat{g}_n(\hat{\theta}_n^Q)$$

corresponds to the CUE-GMM estimator $\underset{\sim}{\hat{\theta}}_n^Q = \left(\hat{\theta}_n^{Q\prime}, \hat{\xi}_n^Q\right)'$ defined by the augmented set of moment conditions $E\underset{\sim}{\Psi}\left(X, \underset{\sim}{\theta}\right) = 0$.

It is worth reminding that Back and Brown (1993) had already derived a similar result in the context of 2S-GMM. However, the framework of Euclidean empirical likelihood makes the

argument even more straightforward. Moreover, the GMM kind of interpretation allows us to refer to the GMM literature (see e.g. Chamberlain (1987)) to conclude that the informational content of estimating equations has been used in an efficient way to estimate not only the parameters θ but also the probability distribution of X:

Corollary 3.10: $\hat{\xi}_n^Q = \sum_{i=1}^n \hat{\pi}_i(\hat{\theta}_n^Q) g(X_i)$, *and more generally* $\hat{E}_{n,\lambda} g(X)$ *for any* $\lambda \neq 0$, *are consistent estimators of* $\xi = Eg(X)$ *which are semiparametrically asymptotically efficient with respect to the information* $E\Psi(X,\theta) = 0$.

Among the various asymptotically equivalent efficient estimators of the probability distribution of X through expectations $Eg(X)$, the main drawback of Euclidean empirical likelihood is that it allows negativity of some implied probabilities $\pi(\hat{\theta}_n^Q)$. However, we know from the asymptotic almost sure positivity property of theorem 2.4 that positivity is not really an issue. More precisely, since $\underset{1\leq i\leq n}{Min}\, \pi_i(\hat{\theta}_n^Q)$ is asymptotically nonnegative with probability one, some well tuned shrinkage may restore nonnegativity in finite sample without introducing any asymptotic efficiency loss. Let us consider the following shrinkage:

$$\pi_i^*(\hat{\theta}_n^Q) = \frac{1}{1+\varepsilon_n(\hat{\theta}_n^Q)} \pi_i(\hat{\theta}_n^Q) + \frac{\varepsilon_n(\hat{\theta}_n^Q)}{1+\varepsilon_n(\hat{\theta}_n^Q)} \cdot \frac{1}{n}, \tag{3.18}$$

with:

$$\varepsilon_n(\theta) = -nMin\left[\underset{1\leq i\leq n}{Min}\, \pi_i(\theta), 0\right] \tag{3.19}$$

Then : $\pi_i^*(\hat{\theta}_n^Q) \geq 0$ for all i and:

Corollary 3.11: *Let*

$$\begin{aligned} g_n^*(\hat{\theta}_n^Q) &= \sum_{i=1}^n \pi_i^*(\hat{\theta}_n^Q) g(X_i) \\ &= \frac{1}{1+\varepsilon_n(\hat{\theta}_n^Q)} \hat{g}_n(\hat{\theta}_n^Q) + \frac{\varepsilon_n(\hat{\theta}_n^Q)}{1+\varepsilon_n(\hat{\theta}_n^Q)} \bar{g}_n. \end{aligned}$$

Then $g_n^*(\hat{\theta}_n^Q)$ and $\hat{g}_n(\hat{\theta}_n^Q)$ are asymptotically equivalent efficient estimators of $Eg(X)$:

$$\sqrt{n}\left[g_n^*(\hat{\theta}_n^Q) - \hat{g}_n(\hat{\theta}_n^Q)\right] = o_P(1).$$

Such a shrinkage of implied probabilities is going to appear particularly relevant in finite sample when they are actually used to estimate some covariance matrix. In this respect, our approach is similar in spirit to the one of Ledoit and Wolf (2001) who propose such a shrinkage to restore positivity of constrained estimates of covariance matrices. This issue will be at stake in subsection 3.3 below.

3.3 A three step Euclidean likelihood

The main message of stochastic expansions of first order conditions (theorem 2.8 and formula (2.13)) is that, except in the case of some zero third moments, Cressie-Read divergences other than empirical likelihood introduce a parasite term in first order conditions. A similar parasite term has been put forward by Newey and Smith (2004) and stressed as responsible for unambiguous better higher properties of empirical likelihood with respect to Cressie-Read contenders. However, as already explained, empirical likelihood may be involved, in computational grounds. This is the reason why Newey and Smith (2004) also noticed that, similarly to Robinson (1988), after three iterations that start at an initial root-n consistent estimator, numerical procedures for solving empirical likelihood first order conditions will produce an estimator with the same leading terms in the stochastic expansions.

We also propose in this subsection to use Robinson (1988) to characterize a three-step estimator with the same leading terms as genuine EL. But our approach does not go through empirical likelihood optimization, even through numerical iterations. We argue instead that, since all the Taylor expansions of Cressie-Read type of first order conditions are based on quadratic terms corresponding to Euclidean Empirical likelihood, it is even more convenient to remain true to quadratic programming, all along the three steps.

Our first two steps are actually devoted to get an asymptotically efficient estimator $\tilde{\theta}_n$ of θ^0, that is $\tilde{\theta}_n$ conformable to (2.7). Note that a common 2S-GMM can do the job with two consecutive quadratic optimizations. While iterated GMM would consist in applying for a third time this GMM optimization device without any well-documented finite sample improvement (see Hansen, Heaton and Yaron (1996)) or higher order advantage, we propose here another third step which affords a genuine improvement and is even easier to perform.

Since the drawback of 2S-GMM and iterated GMM as well is to solve first order conditions where the Jacobian matrix Γ and the covariance matrix Ω are just replaced by their unconstrained inefficient estimators, that suggests to use the efficient estimator $\tilde{\theta}_n$ to efficiently estimate these matrices with a control variables kind of principle, according to corollary 3.5 and corollary 3.10. In other words, the unconstrained estimators:

$$\begin{aligned} \Gamma_n\left(\tilde{\theta}_n\right) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \Psi_i}{\partial \theta'}\left(\tilde{\theta}_n\right) \\ \Omega_n\left(\tilde{\theta}_n\right) &= \frac{1}{n}\sum_{i=1}^{n}\Psi_i\left(\tilde{\theta}_n\right)\Psi_i'\left(\tilde{\theta}_n\right) \end{aligned} \tag{3.20}$$

are improved as:

$$\begin{aligned} \hat{\Gamma}_n^Q\left(\tilde{\theta}_n\right) &= \Gamma_n\left(\tilde{\theta}_n\right) - K_n'\left(\tilde{\theta}_n\right) \\ \hat{\Omega}_n^Q\left(\tilde{\theta}_n\right) &= \Omega_n\left(\tilde{\theta}_n\right) - H_n\left(\tilde{\theta}_n\right) \end{aligned} \tag{3.21}$$

where the j^{th} columns, $j = 1, \cdots, q$ of the matrices $K_n\left(\tilde{\theta}_n\right)$ and $H_n\left(\tilde{\theta}_n\right)$ are respectively defined by:

$$K_n^j\left(\tilde{\theta}_n\right) = Cov_n\left[\frac{\partial \Psi_i^j}{\partial \theta}\left(\tilde{\theta}_n\right), \Psi_i\left(\tilde{\theta}_n\right)\right]\left[V_n\left(\tilde{\theta}_n\right)\right]^{-1}\bar{\Psi}_n\left(\tilde{\theta}_n\right) \tag{3.22}$$

and

$$H_n^j\left(\tilde{\theta}_n\right) = Cov_n\left[\Psi_i\left(\tilde{\theta}_n\right)\Psi_i^j\left(\tilde{\theta}_n\right), \Psi_i\left(\tilde{\theta}_n\right)\right]\left[V_n\left(\tilde{\theta}_n\right)\right]^{-1}\bar{\Psi}_n\left(\tilde{\theta}_n\right).$$

Notice that, as confirmed by theorem 3.12 below, this improvement in the estimation of the j^{th} column of the matrix $\Omega\left(\theta^0\right)$ would be useless if:

$$Cov\left[\Psi_i\left(\theta^0\right)\Psi_i^j\left(\theta^0\right), \Psi_i\left(\theta^0\right)\right] = E\left[\Psi_i\left(\theta^0\right)\Psi_i'\left(\theta^0\right)\Psi_i^j\left(\theta^0\right)\right] = 0 \tag{3.23}$$

This actually confirms the intuition provided by stochastic expansions of first order conditions in theorem 2.8 and formula (2.13). While CUE-GMM shares with 2S-GMM the drawback of a biased estimator of the covariance matrix used in first order conditions (see corollary 3.4), this bias vanishes when third moments are zero as in (3.23). The deep reason for this higher order equivalence between CUE-GMM and empirical likelihood with kind of symmetric errors is that, in that case, the control variables principle based on the information "$E\Psi\left(X, \theta^0\right) = 0$" does not allow to improve the estimation of the covariance matrix; the cross-product terms $\Psi^j\left(X, \theta^0\right)\Psi^l\left(X, \theta^0\right)$ the expectation of which defines the covariance matrix are actually uncorellated with the random vector $\Psi\left(X, \theta^0\right)$.

On the contrary, when the zero third moment conditions is not fulfilled, our control variates improvement of $\Omega_n(\tilde{\theta}_n)$ by $\hat{\Omega}_n^Q(\tilde{\theta}_n)$ is exactly what is needed to protect against the bias of 2S-GMM well-documented in small samples. For example, in a simulation study, Altonji and Segal (1996) demonstrated that "the bias arises because sampling errors in the moments are correlated with sampling errors in the estimate of the covariance matrix of the sample moments". Since $\hat{\Omega}_n^Q$ and $\hat{\Gamma}_n^Q$ are defined from residuals of affine regressions on the moments of interest, such perverse correlations have precisely been deleted. As far as higher order equivalence between empirical likelihood and a suitably corrected quadratic procedure is concerned, it will be obtained thanks to the following result:

Theorem 3.12 *Let $\tilde{\theta}_n$ be an asymptotically efficient estimator of θ^0:*

$$\tilde{\theta}_n - \tilde{\theta}_{n,\lambda} = \mathcal{O}_p\left(1/n\right) \text{ for all } \lambda \neq 0.$$

Let $\hat{\theta}_n$ defined as solution of p equations:

$$\hat{\Gamma}_n^{Q\prime}\left(\tilde{\theta}_n\right)\left[\hat{\Omega}_n^Q\left(\tilde{\theta}_n\right)\right]^{-1}\bar{\Psi}_n\left(\hat{\theta}_n\right) = \mathcal{O}_P\left(1/n\sqrt{n}\right).$$

Then

$$\hat{\theta}_n - \hat{\theta}_{n,1} = \mathcal{O}_P\left(1/n\sqrt{n}\right).$$

In other words, the third step estimator $\hat{\theta}_n$ is higher-order asymptotically equivalent to the empirical likelihood estimator $\hat{\theta}_{n,1}$. While this result corresponds to Taylor expansions of first-order conditions, it would allow under quite general conditions (see e.g. Bhattacharya and Ghosh (1978)) to conclude on higher order identity of Edgeworth expansions. Following Rothenberg (1984), Newey and Smith (2004) even argue that conclusions can be drawn in terms of higher order bias and variance of the estimators.

Note that the three step estimator that we put forward in this section is actually defined by the equality to zero:

$$\hat{\Gamma}_n^{Q\prime}\left(\tilde{\theta}_n\right)\left[\hat{\Omega}_n^Q\left(\tilde{\theta}_n\right)\right]^{-1}\bar{\Psi}_n\left(\hat{\theta}_n\right)=0 \tag{3.24}$$

where $\tilde{\theta}_n$ is a 2S-GMM estimator. However, it is worth realizing that the right hand side of (3.24) may be $\mathcal{O}_P\left(1/n\sqrt{n}\right)$ instead of exactly zero without modifying the conclusion. This is for instance useful to deduce from theorem 2.8 and (2.13) that, when the third moments (2.14) are all zero, any Cressie-Read estimator $\hat{\theta}_{n,\lambda}, \lambda \neq 0$, is higher order equivalent to empirical likelihood $\hat{\theta}_{n,1}$.

The reason why it only matters to have $\mathcal{O}_P\left(1/n\sqrt{n}\right)$ on the right hand side of the defining equation of $\hat{\theta}_n$ is that the order of magnitude of $\hat{\theta}_n - \hat{\theta}_{n,1}$ is actually deduced, by application of Robinson (1988), theorem 1 p. 533, from the order of magnitude of $g_n\left(\hat{\theta}_{n,1}\right)$, when g_n defines $\hat{\theta}_n$ by the p equations:$g_n\left(\hat{\theta}_n\right)=0$.

4 Conditional implied probabilities

4.1 Smoothed power divergence statistics

Let $(X_i, Z_i), (i = 1, \cdots, n)$ be i.i.d observations on a random vector (X, Z) on $\mathbb{R}^l \times \mathbb{R}^d$. We consider as in previous sections $\Psi(X, \theta) = \left(\Psi^j(X, \theta)\right)_{1 \leq j \leq q}$, a q-vector of functions of the data observation X and the p-vector θ of unknown parameters. But it is now assumed that the true parameter vector θ^0 satisfies the conditional moment restrictions:

$$E\left[\Psi\left(X, \theta^0\right) | Z\right] = 0, \theta^0 \in \Theta \subset \mathbb{R}^p \tag{4.1}$$

Of course, any choice of a vector $g(Z)$ of instruments would allow to apply the results of previous sections to unconditional moment restrictions:

$$E\left[g(Z) \otimes \Psi\left(X, \theta^0\right)\right] = 0$$

However, efficient estimation of θ^0 from (4.1) would then rest upon a selection of optimal instruments (see e.g. Newey (1993)). Moreover, we are also interested in estimating conditional implied probabilities of X given Z taking advantage of the informational content of conditional restrictions (4.1). For these two reasons, we propose in this section alternative estimation techniques which avoid estimating optimal instruments in a preliminary step, while allowing one step efficient estimation of both θ and the conditional distribution of X given Z.

While estimation of optimal instruments would involve nonparametric estimation of conditional expectations given Z of $\frac{\partial \Psi}{\partial \theta'}\left(X, \theta^0\right)$ and $\Psi\left(X, \theta^0\right)\Psi'(X, \theta^0)$, kernel smoothing of probabilities given Z will be introduced here from the beginning through implied probabilities and corresponding discrepancy statistics. The starting point is a localized version of the Cressie and Read (1984) power divergence family of statistics. While it involves in (2.9) the relative differences between the perceived probabilities (see Bera and Bilias (2002) for more intuition about this) as:

$$\left(\frac{\pi_j}{w_j}\right)^{1-\lambda} - 1$$

where π_j and $w_j = \frac{1}{n}$ denote respectively the implied and the empirical probabilities for the possible (that is observed) values X_j of X, it will involve now the relative differences:

$$\left(\frac{\pi_{ij}}{w_{ij}}\right)^{1-\lambda} - 1 \tag{4.2}$$

where π_{ij} and w_{ij} denote respectively the implied and the empirical conditional probabilities for the possible (observed) values X_j of X, given $Z = Z_i$.

Of course, when the conditioning variable Z is continuous, the so-called empirical conditional probabilities must be defined through smoothing. In all the sequel, kernel smoothing will be performed with a Rosenblatt-Parzen kernel K which is a probability density function on $\mathbb{R}^d$, symmetric

about the origin and continuously differentiable. Under standard regularity conditions not detailed here (see e.g. Ai and Chen (2003) and Kitamura, Tripathi and Ahn (2001)), well-suited asymptotic theory of kernel estimators including uniform convergence will be valid. Then, localization is carried out through the positive weights:

$$w_{ij} = \frac{K_{ij}}{\sum_{l=1}^{n} K_{il}} \tag{4.3}$$

where:

$$K_{ij} = K\left(\frac{Z_i - Z_j}{b_n}\right)$$

and b_n is a bandwidth sequence of positive numbers such that $b_n \underset{n=\infty}{\longrightarrow} 0$ and $nb_n^d \underset{n=\infty}{\longrightarrow} \infty$. For the sake of notational simplicity, the dependence of w_{ij} and K_{ij} upon n is suppressed.

Then, the localized version of the Cressie and Read divergence statistic I_λ defined in (2.9) is:

$$I_\lambda = \frac{1}{\lambda(\lambda-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left[\left(\frac{\pi_{ij}}{w_{ij}}\right)^{1-\lambda} - 1\right] \tag{4.4}$$

To interpret (4.4), it is worth seeing the i^{th} term of the first summation operator as corresponding to conditioning by $Z = Z_i$. Then the relative differences (4.2) between perceived conditional probabilities π_{ij} and w_{ij} of possible values X_j, $j = 1, \cdots n$, given $Z = Z_i$, are weighted by kernel weights w_{ij} which assign smaller weights to those X_j's which are farther away from X_i. Of course, if the weight were by chance all identical, they would be all equal to $(1/n)$ and (4.4) would become exactly similar to (2.9).

For $\lambda \notin \{0, 1\}$, the minimization of the divergence I_λ with respect to $(\pi_{ij})_{1 \leq i,j \leq n}$ is equivalent to the optimization of a discrepancy statistic:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^{\lambda} h^{(\lambda)}(\pi_{ij}) \tag{4.5}$$

where $h^{(\lambda)}(\pi) = \pi^{1-\lambda}$ as in (2.10). As in section 2, we do not make explicit that $h^{(\lambda)}(\pi) = -\pi^{1-\lambda}$ should rather be considered for minimization in the case $0 < \lambda < 1$. Moreover, by contrast with (2.10), the weights w_{ij}^{λ} are not all equal and thus, depend explicitly upon λ.

The smoothed empirical likelihood case $\left(h^{(1)}(\pi) = \log \pi\right)$, as studied by Kitamura, Tripathi and Ahn (2001), is also nested in this framework by the limit case $\lambda \longrightarrow 1$ since:

$$\underset{\lambda \longrightarrow 1}{Lim} \frac{1}{\lambda - 1} \left[\left(\frac{\pi_{ij}}{w_{ij}}\right)^{1-\lambda} - 1\right] = -\log \frac{\pi_{ij}}{w_{ij}}$$

and (4.5) becomes:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} Log\left(\pi_{ij}\right).$$

In summary, for all $\lambda \neq 0$, we consider the following family of smoothed discrepancy statistics:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{\lambda} h^{(\lambda)}\left(\pi_{ij}\right) \tag{4.6}$$

with a first derivative $h_{\pi}^{(\lambda)}(\pi) = \pi^{-\lambda}$, that is

$$h^{(\lambda)}(\pi) = \begin{cases} \pi^{1-\lambda} & \text{if } \lambda \notin \{0,1\} \\ Log\pi & \text{if } \lambda = 1 \end{cases} \tag{4.7}$$

Similarly to (2.2), we are interested in the optimization of (4.6) under the constraints:

$$\begin{cases} \sum\limits_{j=1}^{n} \pi_{ij} \Psi\left(X_j, \theta\right) = 0 \text{ for } i = 1, \cdots, n \\ \sum\limits_{j=1}^{n} \pi_{ij} = 1 \text{ for } i = 1, \cdots, n \end{cases} \tag{4.8}$$

When the optimization problem of (4.6) under (4.8) admits a unique solution $\left(\hat{\pi}_{i,j,\lambda}\right)_{1 \leq i,j \leq n}, \hat{\theta}_{n,\lambda}$ with nonnegative $\hat{\pi}_{i,j,\lambda}$s, the n numbers $\hat{\pi}_{i,j,\lambda}, j = 1, \cdots, n$ (for any given $i = 1, \cdots n$) can be interpreted as conditional probabilities of the values X_j given $Z = Z_i$. The constraints (4.8) simply mean that the implied probabilities sum to one and meet the conditional moment restrictions. More generally, for any integrable function $g(X)$, $\sum\limits_{j=1}^{n} \hat{\pi}_{i,j,\lambda} g\left(X_j\right)$ defines an estimator of the conditional expectation $E\left[g(X) \left| Z = Z_i \right.\right]$ that takes advantage of the information carried out by the conditional moment restrictions. By analogy with previous sections, we will denote by

$$\hat{E}_{n,\lambda}\left[g(X) \left| Z_i \right.\right] = \sum_{j=1}^{n} \hat{\pi}_{i,j,\lambda} g(X_j) \tag{4.9}$$

this *constrained* estimator while the *unconstrained* estimator is nothing but the Nadaraya-Watson kernel estimator:

$$\hat{E}_{n}\left[g(X) \left| Z_i \right.\right] = \bar{g}_{i,n} = \sum_{j=1}^{n} w_{i,j} g(X_j) \tag{4.10}$$

The following result is the exact analog of theorem 2.1:

Theorem 4.1 *Assume that the optimization of (4.6) under (4.8) uniquely defines estimators* $\hat{\pi}_{i,j,\lambda}, 1 \leq i,j \leq n, \hat{\theta}_{n,\lambda}$ *with nonnegative* $\hat{\pi}_{i,j,\lambda}$*s. Then* $\hat{\theta}_{n,\lambda}$ *is characterized as solution of the first order conditions:*

$$\sum_{i=1}^{n} \hat{E}_{n,\lambda}\left[\frac{\partial \Psi'\left(X,\hat{\theta}_{n,\lambda}\right)}{\partial \theta}|Z_i\right]\left(\hat{E}_{n,\lambda}\left[\Psi\left(X,\hat{\theta}_{n,\lambda}\right)\Psi'\left(X,\hat{\theta}_{n,\lambda}\right)|Z_i\right]\right)^{-1}\sum_{j=1}^{n} w_{i,j}^{\lambda}\hat{\pi}_{i,j,\lambda}^{1-\lambda}\Psi\left(X_j,\hat{\theta}_{n,\lambda}\right)=0 \tag{4.11}$$

Note that, since by (4.9) and (4.10), $\hat{\pi}_{i,j,\lambda}$ and w_{ij} are both localization weights allowing us to estimate conditional expectations given $Z = Z_i$, the last term of (4.11) can also be interpreted as a nonparametric estimator of the conditional moment restriction of interest. We will denote it:

$$\tilde{E}_{n,\lambda}\left[\Psi\left(X,\hat{\theta}_{n,\lambda}\right)|Z_i\right]=\sum_{j=1}^{n} w_{ij}^{\lambda}\hat{\pi}_{i,j,\lambda}^{1-\lambda}\Psi\left(X_j,\hat{\theta}_{n,\lambda}\right) \tag{4.12}$$

It coincides with the kernel estimator (4.10) in the particular case of smoothed empirical likelihood that is $\lambda = 1$. In any case, it is worth noticing that the first order conditions (4.11) to compute the estimator $\hat{\theta}_{n,\lambda}$ can be interpreted, with shortened notations, as:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{E}_{n,\lambda}\left[\frac{\partial \Psi'}{\partial \theta}\left(\hat{\theta}_{n,\lambda}\right)|Z_i\right]\hat{V}_{n,\lambda}^{-1}\left[\Psi\left(\hat{\theta}_{n,\lambda}\right)|Z_i\right]\tilde{E}_{n,\lambda}\left[\Psi\left(\hat{\theta}_{n,\lambda}\right)|Z_i\right]\right\}=0 \tag{4.13}$$

This interpretation has several interesting consequences. First, it points out the identification assumptions which are relevant to extend assumption 2.1 to the conditional framework:

Assumption 4.1

(i) $E\left[\Psi\left(X,\theta\right)|Z\right]=0 \Longleftrightarrow \theta=\theta^0$

(ii) $\Omega_Z(\theta)=E\left[\Psi\left(X,\theta\right)\Psi'\left(X,\theta\right)|Z\right]$

is, for all $\theta \in \Theta$*, a nonsingular matrix with probability one.*

(iii) $\Gamma_Z(\theta^0)=E\left[\dfrac{\partial \Psi}{\partial \theta'}\left(X,\theta\right)_{|\theta=\theta^0}|Z\right]$

is such that:

$$I\left(\theta^0\right)=E\left[\Gamma'_Z\left(\theta^0\right)\Omega_Z^{-1}(\theta^0)\Gamma_Z(\theta^0)\right]$$

is a nonsingular matrix.

Moreover, (4.13) can be seen as an empirical counterpart of the moment restrictions:

$$E\left[\Gamma'_Z\left(\theta^0\right)\Omega_Z^{-1}\left(\theta^0\right)\Psi\left(X,\theta\right)\right]=0, \tag{4.14}$$

since, by the law of iterated expectations, these moment restrictions can also be written:

$$E\left[\Gamma'_Z\left(\theta^0\right)\Omega_Z^{-1}\left(\theta^0\right)E\left[\Psi\left(X,\theta\right)|Z\right]\right]=0. \tag{4.15}$$

With respect to the standard efficient treatment of conditional moment restrictions (see e.g. Newey (1993)), estimating equations (4.13) have important similarities and differences which are secondary, in terms of first order asymptotics. The important similarity is that the optimal matrix of instruments $\Gamma'_Z\left(\theta^0\right)\Omega_Z^{-1}\left(\theta^0\right)$ has been replaced by a nonparametric estimator which will be consistent under standard regularity conditions. Actually, following Newey (1993), an efficient estimator $\hat{\theta}_n$ of θ would be obtained by solving the p equations:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{E}_n\left[\frac{\partial\Psi'}{\partial\theta}\left(\tilde{\theta}_n\right)|Z_i\right]\hat{V}_n^{-1}\left[\Psi\left(\tilde{\theta}_n\right)|Z_i\right]\Psi\left(X_i,\hat{\theta}_n\right)\right\}=0 \tag{4.16}$$

where $\hat{E}_n,\hat{V}_n$ denote standard kernel estimators and $\tilde{\theta}_n$ is a first step consistent estimator of θ.

It is then quite clear that, under standard regularity conditions, the differences between (4.13) and (4.16) will not matter as far as first order asymptotics are concerned. Both these equations will provide an asymptotically efficient estimator of θ, that is an estimator such that:

$$\sqrt{n}\left(\hat{\theta}_n-\theta^0\right)=-I\left(\theta^0\right)^{-1}\sqrt{n}\bar{\varphi}_n\left(\theta^0\right)+o_p(1) \tag{4.17}$$

where:

$$\varphi\left(X_i,Z_i,\theta\right)=\Gamma'_{Z_i}(\theta)\Omega_{Z_i}^{-1}(\theta)\Psi\left(X_i,\theta\right).$$

For sake of simplicity, sufficient regularity conditions to ensure (4.17) for all the estimators of interest are not discussed here in details. Convenient smoothness and moment existence conditions can be found in Newey (1993). In any case, a maintained assumption is weak consistency of the kernel estimators of interest:

Assumption 4.2 *The kernel estimators* $\hat{E}_n\left[\dfrac{\partial\Psi'}{\partial\theta}(X,\theta^0)|Z\right]$, $\hat{E}_n\left[\Psi\left(X,\theta^0|Z\right)\right]$ *and* $\hat{E}_n\left[\left(X,\theta^0\right)\Psi'\left(X,\theta^0\right)|Z\right]$ *are weakly consistent estimators of corresponding conditional expectations.*

By definition, all asymptotically efficient estimators $\hat{\theta}_n$ are such that the asymptotic probability distribution of $\sqrt{n}\left(\hat{\theta}_n-\theta^0\right)$ is $\mathcal{N}\left[0,I\left(\theta^0\right)^{-1}\right]$. However, by analogy with the arguments put forward in the unconditional case, one may expect that higher order asymptotic properties and finite sample properties as well are better for estimators $\hat{\theta}_{n,\lambda}$ deduced from equations like (4.13) than

for more standard estimators $\hat{\theta}_n$ like in (4.16). The reason for that is that the optimal instrumental matrix $\Gamma_Z'\left(\theta^0\right)\Omega_Z^{-1}\left(\theta^0\right)$, although consistently estimated in both cases, is better estimated in (4.13) since its constrained estimator

$$\hat{E}_{n,\lambda}\left[\frac{\partial\Psi'}{\partial\theta}\left(\hat{\theta}_{n,\lambda}\right)|Z\right]\hat{V}_{n,\lambda}^{-1}\left[\Psi\left(\hat{\theta}_{n,\lambda}\right)|Z\right]$$

takes into account the informational content of conditional moment restrictions by using the implied probabilities $\hat{\Pi}_{i,j,\lambda}$. The actual computation of these probabilities will be based on the following expression of first order conditions:

Theorem 4.2 *For $i=1,\cdots,n$ there exist a non-zero real number $\mu_{i,n,\lambda}$ and a vector $\alpha_{i,n,\lambda}$ of q reduced Lagrange multipliers such that:*

$$\hat{\pi}_{i,j,\lambda}^{-\lambda}=\mu_{i,n,\lambda}w_{i,j}^{-\lambda}\left[1+\alpha_{i,n,\lambda}'\Psi_j\left(\hat{\theta}_{n,\lambda}\right)\right]$$

From arguments of asymptotic almost sure nonnegativity similar to the ones put forward in the unconditional case, one can deduce from theorem 4.2 that asymptotically almost certainly:

$$\hat{\pi}_{i,j,\lambda}=\frac{w_{ij}\left[1+\alpha_{i,n,\lambda}'\Psi_j\left(\hat{\theta}_{n,\lambda}\right)\right]^{-1/\lambda}}{\sum_{l=1}^{n}w_{il}\left[1+\alpha_{i,n,\lambda}'\Psi_l\left(\hat{\theta}_{n,\lambda}\right)\right]^{-1/\lambda}} \tag{4.18}$$

As already pointed out, the computation of reduced Lagrange multipliers $\alpha_{i,n,\lambda}$ from (4.18) and conditional moment restrictions will in general be involved, except in the case of Euclidean empirical likelihood ($\lambda=-1$) where the equations to solve appear to be linear. As far as empirical likelihood ($\lambda=1$) is concerned, it amounts to the resolution of n convex minimization programs of size q according to:

$$\underset{\alpha_{i,n}}{Min}-\sum_{j=1}^{n}w_{i,j}Log\left[1+\alpha_{i,n}'\Psi_j\left(\hat{\theta}_{n,1}\right)\right]\text{ for }i=1,\cdots,n.$$

In other words, the actual size of the computational problem is nq. This is the reason why we choose to focus below on the simplest case of Euclidean empirical likelihood.

4.2 Two conditional versions of continuously updated GMM

We focus here on the quadratic version of the minimization problem (4.6) under constraints (4.8), that is the case $\lambda=-1$:

$$\begin{cases}\underset{\pi_{i,j,},\theta}{Min}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{ij}^2}{w_{ij}}\\ \sum_{j=1}^{n}\pi_{ij}\Psi_j(\theta)=0 \qquad \forall i=1,\cdots,n\\ \sum_{j=1}^{n}\pi_{ij}=1 \qquad \forall i=1,\cdots,n\end{cases} \tag{4.19}$$

Similarly to what has been done in the unconditional case, we first consider the minimization problem (4.19) with respect to the $\pi_{i,j}$ s, for a given value of θ. We so characterize profile functions $\pi_{i,j}(\theta)$ by using the two alternative estimators of the conditional covariance function $\Omega_Z(\theta^0) = E\left[\Psi\left(X,\theta^0\right)\Psi'\left(X,\theta^0\right)|Z\right]$:
Uncentered second moments:

$$\Omega_n\left(\theta\,|Z_i\right) = \sum_{j=1}^{n} w_{ij}\Psi_j(\theta)\Psi_j'(\theta) \tag{4.20}$$

Second moments in mean deviation form:

$$V_n\left(\theta\,|Z_i\right) = \sum_{j=1}^{n} w_{ij}\Psi_j(\theta)\left[\Psi_j(\theta) - \bar{\Psi}_i(\theta)\right]' \tag{4.21}$$

with:

$$\bar{\Psi}_i(\theta) = \sum_{j=1}^{n} w_{ij}\Psi_j(\theta). \tag{4.22}$$

Note that $\Omega_n\left(\theta\,|Z_i\right), V_n\left(\theta\,|Z_i\right)$ and $\bar{\Psi}_i\left(\theta\right)$ are nothing but Nadaraya-Watson kernel estimators of conditional expectations of interest.

Theorem 4.3 *For all $\theta \in \Theta$ and $i = 1, \cdots n$:*
$\pi_{i,j}(\theta), j = 1, \cdots, n$, is proportional to $1 + \alpha_{i,n}'(\theta)\Psi_j(\theta)$ and to
$1 + \gamma_{in}'(\theta)\left[\Psi_j(\theta) - \bar{\Psi}_i(\theta)\right]$ with:
$\alpha_{in}(\theta) = -\Omega_n^{-1}\left(\theta\,|Z_i\right)\bar{\Psi}_i\left(\theta\right)$ and $\gamma_{in}(\theta) = -V_n^{-1}\left(\theta\,|Z_i\right)\bar{\Psi}_i(\theta)$.
In particular:

$$\pi_{i,j}(\theta) = w_{ij} - w_{ij}\bar{\Psi}_i'(\theta)V_n^{-1}\left(\theta\,|Z_i\right)\left[\Psi_j(\theta) - \bar{\Psi}_i(\theta)\right].$$

The profile functions $\pi_{i,j}(\theta), i, j = 1, \cdots, n$ give us two interesting characterizations of the profile criterion defining the Euclidean empirical likelihood estimator $\hat{\theta}_n^Q = \hat{\theta}_{n,-1}$:

Corollary 4.4:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{i,j}^2(\theta)}{w_{ij}} = \sum_{i=1}^{n}\left[1 - \bar{\Psi}_i'(\theta)\Omega_n^{-1}\left(\theta\,|Z_i\right)\bar{\Psi}_i\left(\theta\right)\right]^{-1} = \sum_{i=1}^{n}\left[1 + \bar{\Psi}_i'(\theta)V_n^{-1}\left(\theta\,|Z_i\right)\bar{\Psi}_i\left(\theta\right)\right].$$

Corollary 4.4 shows that the Euclidean conditional empirical likelihood estimator $\hat{\theta}_n^Q$ numerically coincides with a conditional generalization of the CUE-GMM estimator of Hansen, Heaton and Yaron (1996):

$$\hat{\theta}_n^Q = \underset{\theta}{ArgMin}\sum_{i=1}^{n}\bar{\Psi}_i'(\theta)V_n^{-1}\left(\theta\,|Z_i\right)\bar{\Psi}_i\left(\theta\right) \tag{4.23}$$

It is worth noticing that (4.23) fully coincides in spirit with CUE-GMM since the kernel estimators $\bar{\Psi}_i'(\theta), i = 1, 2 \cdots n$ are known to be asymptotically independent. In other words, by seeing the nq-dimensional vector $\left(\bar{\Psi}_i(\theta)\right)_{1 \leq i \leq n}$ as the sample counterpart of conditional moment restrictions, this vector has a block diagonal asymptotic covariance matrix which a posteriori justifies the additively separable form of its squared norm minimized in (4.23).

Note that the two summation forms of the profile function provided by corollary 4.4 actually coincide term by term since:

$$\left[1 + \bar{\Psi}_i'(\theta) V_n^{-1}(\theta \,|Z_i\,)\, \bar{\Psi}_i(\theta)\right]\left[1 - \bar{\Psi}_i'(\theta) \Omega_n^{-1}(\theta \,|Z_i\,)\, \bar{\Psi}_i(\theta)\right] = 1 \tag{4.24}$$

as it can be seen by developing the product in a sum of four terms and noticing that:

$$\bar{\Psi}_i(\theta)\, \bar{\Psi}_i'(\theta) = \Omega_n(\theta \,|Z_i\,) - V_n(\theta \,|Z_i\,).$$

(4.24) can actually be seen as a generalization of (3.7) to the case of weighted averages. However, by contrast with the results of section 3, the two possible ways to perform CUE-GMM in a conditional setting do not numerically coincide, even though they are asymptotically equivalent:

$$\hat{\theta}_n^Q \neq Arg\underset{\theta}{Min} \sum_{i=1}^{n} \bar{\Psi}_i'(\theta) \Omega_n^{-1}(\theta \,|Z_i\,)\, \bar{\Psi}_i(\theta) \tag{4.25}$$

We can just say from (4.24) that:

$$\underset{\theta}{Min} \sum_{i=1}^{n} \bar{\Psi}_i'(\theta) \Omega_n^{-1}(\theta \,|Z_i\,)\, \bar{\Psi}_i(\theta)$$

$$\Longleftrightarrow$$

$$\underset{\theta}{Min} \sum_{i=1}^{n} \frac{\bar{\Psi}_i'(\theta) V_n^{-1}(\theta \,|Z_i\,)\, \bar{\Psi}_i(\theta)}{1 + \bar{\Psi}_i'(\theta) V_n^{-1}(\theta \,|Z_i\,)\, \bar{\Psi}_i(\theta)}$$

While (4.23) and (4.25) define two natural extensions of CUE-GMM in a conditional setting, a 2S-GMM version of (4.25) had already been proposed by Ai and Chen (2001) as:

$$\underset{\theta}{Min} \sum_{i=1}^{n} \bar{\Psi}_i'(\theta) \Omega_n^{-1}\left(\tilde{\theta}_n \,|Z_i\right) \bar{\Psi}_i(\theta) \tag{4.26}$$

where $\tilde{\theta}_n$ is a first step consistent estimator of θ.

As in the unconditional case, we argue that, when computed in its Euclidean conditional empirical likelihood form (4.23), conditional CUE-GMM should have better properties than its competitors (4.25) and (4.26) since it makes a more efficient use of the informational content of estimating equations to estimating optimal instruments. Therefore, the terminology conditional CUE-GMM will be used in the following only for $\hat{\theta}_n^Q$ as defined by (4.23):

Corollary 4.5:

The Euclidean conditional empirical likelihood estimator (conditional CUE-GMM) $\hat{\theta}_n^Q$ is characterized as solution of following system of first order conditions:

$$\sum_{i=1}^{n} \hat{E}_n^Q \left[\frac{\partial \Psi'}{\partial \theta} \left(X, \hat{\theta}_n^Q \right) | Z_i \right] V_n^{-1} \left(\hat{\theta}_n^Q | Z_i \right) \bar{\Psi}_i \left(\hat{\theta}_n^Q \right) = 0$$

where, for any integrable function $g(X)$, $\hat{E}_n^Q \left[g(X) | Z_i \right] = \hat{E}_{n,-1} \left[g(X) | Z_i \right]$ denotes the estimation of $E \left[g(X) | Z = Z_i \right]$ deduced from quadratic implied probabilities $\hat{\pi}_{i,j} \left(\hat{\theta}_n^Q \right)$ as defined by theorem 4.3.

By comparison with theorem 4.1 in the case $\lambda = 1$, corollary 4.5 shows that the drawback of Euclidean empirical likelihood ($\lambda = -1$) with respect to empirical likelihood ($\lambda = 1$) is that implied probabilities are not used to improve the estimation of the covariance matrix. This is going to motivate the introduction of a three step Euclidean likelihood estimator.

4.3 Efficient use of the informational content of estimating equations

The focus of interest of this section is to assess the informational content of the conditional moment restrictions:

$$E \left[\Psi \left(X, \theta^0 \right) | Z \right] = 0,$$

not only about the true unknown value θ^0 of the parameters but also about the conditional probability distribution of X given Z.

Similarly to what has been done in section 3.2 in the unconditional case, our knowledge about the conditional distribution of X given Z is summarized by the way to estimate conditional expectations $E \left[g(X) | Z \right]$, for any real test function g.

Even though all Cressie-Read power divergence statistics provide first-order asymptotically equivalent estimators, the advantages of the Euclidean empirical likelihood approach are even more striking in the conditional case. As already noticed, theorem 4.3 provides closed-form formulas for implied probabilities in the Euclidean case, while such formulas are not available in other cases. But, even more importantly, we can apply the same formulas for conditional probabilities given any possible value z of Z, without being limited to the observed values $Z_i, i = 1, \cdots, n$. More precisely, a straightforward extension of theorem 4.3 suggests to define the conditional implied probabilities of observed values X_j, $j = 1, \cdots, n$ given $Z = z$ by:

$$\hat{\pi}_j(z) = w_j(z) - w_j(z) \bar{\Psi}_z' \left(\hat{\theta}_n^Q \right) V_n^{-1} \left(\hat{\theta}_n^Q | z \right) \left[\Psi_j \left(\hat{\theta}_n^Q \right) - \bar{\Psi}_z \left(\hat{\theta}_n^Q \right) \right]$$

where:

$$w_j(z) = \frac{K_j(z)}{\sum_{l=1}^{n} K_l(z)}, K_j(z) = K \left(\frac{z - Z_j}{b_n} \right),$$

$$\bar{\Psi}_z\left(\hat{\theta}_n^Q\right) = \sum_{j=1}^{n} w_j(z)\Psi_j\left(\hat{\theta}_n^Q\right) \tag{4.27}$$

$$V_n\left(\hat{\theta}_n^Q\,|z\right) = \sum_{j=1}^{n} w_j(z)\Psi_j\left(\hat{\theta}_n^Q\right)\left[\Psi_j\left(\hat{\theta}_n^Q\right) - \bar{\Psi}_z\left(\hat{\theta}_n^Q\right)\right]'.$$

With a similar definition of $Cov_n\left[g(X_j), \Psi_j\left(\hat{\theta}_n^Q\right)|z\right]$, we then deduce easily:

Theorem 4.6 *For any integrable real function $g(X)$, $E\left[g(X)\,|Z=z\right]$ can be estimated by:*

$$\begin{aligned}\hat{E}_n^Q\left[g(X)\,|Z=z\right] &= \sum_{j=1}^{n}\hat{\pi}_j(z)g(X_j)\\ &= \bar{g}_z - Cov_n\left[g(X_j), \Psi_j\left(\hat{\theta}_n^Q\right)|z\right] V_n^{-1}\left(\hat{\theta}_n^Q\,|z\right)\bar{\Psi}_z\left(\hat{\theta}_n^Q\right)\end{aligned}$$

where:

$$\bar{g}_z = \sum_{j=1}^{n} w_j(z)g(X_j)$$

is the Nadaraya-Watson kernel estimator.

The intuition behind the proposed estimator is very clear. We improve, through a control variates strategy, the naive kernel estimator $\bar{g}_z$ by taking into account the information content of the conditional moment restrictions $E\left[\Psi(X,\theta^0)\,|Z\right]=0$.

More precisely, if we knew the true unknown value θ^0, we would replace the estimation problem of $E\left[g(X)\,|Z\right]$ by the more favourable problem of estimation of $E\left[g(X)-a'\Psi(X,\theta^0)\,|Z\right]$. In order to minimize the conditional variance of the resulting estimator, the optimal value of the coefficient a is given by conditional affine regression:

$$a'(Z,\theta^0) = Cov\left[g(X),\Psi(X,\theta^0)\,|Z\right]\left(Var\left[\Psi(X,\theta^0)\,|Z\right]\right)^{-1} \tag{4.28}$$

The estimator put forward by theorem 4.6 is nothing but:

$$E\left[g(X)\,|Z=z\right] - a'(z,\theta^0)E\left[\Psi(X,\theta^0)\,|Z=z\right]$$

after replacement of population conditional expectations by their kernel counterpart and of θ^0 by $\hat{\theta}_n^Q$. It is worth noticing that, by contrast with the empirical likelihood case, the availability of closed-form formulas allows us to take advantage of the information content of conditional moment restrictions even for conditioning values not observed in sample. In particular, it is easy to check that for any z:

$$\hat{E}_n^Q\left[\Psi(X,\theta)\,|Z=z\right]=0 \text{ for } \theta=\hat{\theta}_n^Q.$$

In this respect, we can claim that we have made an efficient use of the informational content of conditional moment restrictions about the conditional probability distribution of X given Z. To make this claim more precise, theorem 4.5 below makes explicit the efficiency gain in terms of asymptotic variance with respect to the naive kernel estimator of $E\left[g(X)\left|Z\right.\right]$.

This theorem is valid under any set of assumptions which ensures asymptotic normality with zero bias for kernel estimators of $E\left[g(X)\left|Z=z\right.\right]$ and $E\left[\Psi\left(X,\theta^0\right)\left|Z=z\right.\right]$:

Assumption 4.3
(i) Z is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ with a density function f which is twice continuously differentiable in a neighborhood of z, interior point of the support of Z.
(ii) K is a Parzen-Rosenblatt kernel with in particular:

$$\begin{aligned} \int K(u)du \quad &= 1 \quad , \qquad \int u\ K(u)du = 0, \\ \int \left|K(u)\right|^{2+\delta} du \quad &< \quad +\infty \qquad \text{for some } \delta > 0. \end{aligned}$$

(iii) b_n is bandwith sequence such that:

$$nb_n^d \underset{n=\infty}{\longrightarrow} \infty \text{ and } nb_n^{d+4} \underset{n=\infty}{\longrightarrow} 0$$

(iv) $\sqrt{nb_n^d}\left[\bar{g}_z - E\left[g(X)\left|Z=z\right.\right]\right]$ is asymptotically distributed as a normal with zero mean and variance:

$$\frac{\sigma^2(z)}{f(z)}\int K^2(u)du.$$

(v) $\sqrt{nb_n^d}\left[\bar{\Psi}_z(\theta^0) - E\left[\Psi\left(X,\theta^0\right)\left|Z=z\right.\right]\right]$ is asymptotically distributed as a normal with zero mean and variance:

$$\frac{\Omega_z(\theta^0)}{f(z)}\int K^2(u)du.$$

Then, we have:

Theorem 4.7 *Under assumptions 4.1, 4.2, 4.3 and standard regularity assumptions:*

$$\sqrt{nb_n^d}\left(\hat{E}_n^Q\left[g(X)\left|Z=z\right.\right] - E\left[g(X)\left|Z=z\right.\right]\right)$$

is asymptotically distributed as a normal with zero mean and variance:

$$\frac{1}{f(z)}\left(\int K^2(u)du\right)\left(\sigma^2(z) - \eta^2(z)\right)$$

with:

$$\eta^2(z) = Cov\left[g(X), \Psi\left(X,\theta^0\right)\left|Z=z\right.\right]\Omega_z^{-1}(\theta^0)Cov\left[\Psi(X,\theta^0), g(X)\left|Z=z\right.\right]$$

Note that if θ^0 were known, theorem 4.5 would be a straightforward consequence of the affine regression argument put forward by theorem 4.6. Of course, θ^0 must actually be replaced by its consistent estimator $\hat{\theta}_n^Q$ to compute $\hat{E}_n^Q[g(X)|Z=z]$. But, since $\hat{\theta}_n^Q$ is root-n consistent, this estimation error does not play any role with respect to the main estimation error in theorem 4.7 which goes to zero at the slower rate $\sqrt{nb_n^d}$. This is the reason why theorem 4.7 is even simpler that its analog theorem 3.6 in the unconditional case. The efficiency gain with respect to the unconstrained estimator $\bar{g}_z$, as measured by $\frac{\eta^2(z)}{f(z)}\left(\int K^2(u)du\right)$ has not to be reduced in proportion of the estimation error $\left(\hat{\theta}_n^Q - \theta^0\right)$.

In this respect, one can argue that, in the same way a conditional expectation $E[g(X)|Z]$ can be seen as efficiently estimated by its kernel counterpart (see e.g. Severini and Tripathi (2001), section 7), we have efficiently estimated this conditional expectation when taking into account the additional information:

$$E\left[\Psi(X,\theta^0)|Z\right]=0.$$

As explained above, the kernel estimator of $E\left[g(X)-a'(Z,\theta^0)\Psi\left(X,\theta^0\right)|Z\right]$ is, in some sense, the best that we can do. However, as far as semiparametric efficiency is concerned, for the purpose of estimation of the conditional expectation functional $m(z)=E[g(X)|Z=z]$, we know (see Severini and Tripathi (2001) for several illustrations of this approach) that we should focus on estimation of associated parameters:

$$\beta=\int_D m(z)\omega(z)dz \tag{4.29}$$

where D is some compact region of integration and $\omega(z)$ is an arbitrary weighting function such that:

$$\int_D \omega^2(z)dz<+\infty.$$

The idea is that, since β is a one-dimensional parameter, it admits root-n consistent estimators such that it makes sense to compare estimators through their asymptotic variances.

In other words, we must assess the accuracy of our estimator $\hat{m}_n(z)=\hat{E}_n^Q[g(X)|Z=z]$ by considering the associated plug-in estimator of β:

$$\hat{\beta}_n=\int_D \hat{m}_n(z)\omega(z)dz \tag{4.30}$$

Of course, since $\hat{\beta}_n$ has a root-n rate of convergence, the role of the estimation error in θ is restored and we get the following result:

Theorem 4.8 *Under assumptions of theorem 4.7,* $\sqrt{n}\left(\hat{\beta}_n-\beta\right)\underset{n=\infty}{\overset{d}{\rightarrow}}\mathcal{N}[0,W_D+H_D]$ *with:*

$$W_D=E\left[\frac{\sigma^2(Z)-\eta^2(Z)}{f^2(z)}\omega^2(Z)\mathbf{1}_D(Z)\right]$$

and

$$H_D = E\left[a'(Z,\theta^0)\Gamma_Z(\theta^0)\frac{\omega(Z)\mathbf{1}_D(Z)}{f(Z)}\right] I^{-1}(\theta^0) E\left[\Gamma'_Z(\theta^0)a(Z,\theta^0)\frac{\omega(Z)\mathbf{1}_D(Z)}{f(Z)}\right]$$

Theorem 4.8 is a natural extension of Severini and Tripathi (2001) efficiency bound for estimation of a conditional expectation functional as summarized by the formula they give at the bottom of page 43. More precisely, the asymptotic variance $W_D + H_D$ can be seen as the result of three steps of reasoning. First, $E\left[\frac{\sigma^2(Z)}{f^2(z)}\omega^2(Z)\mathbf{1}_D(Z)\right]$ is, as shown by Severini and Tripathi (2001), the semiparametric efficiency bound associated to the kernel estimator $\bar{g}_z$ of $m(z)$. Second, as shown by theorem 4.7, we improve this kernel estimator by computing instead the kernel estimator of $E\left[g(X) - a'(Z,\theta^0)\Psi(X,\theta^0)\,|Z\right]$.

In other words, the Severini and Tripathi (2001) efficiency bound becomes

$$E\left[\frac{\sigma^2(Z) - \eta^2(Z)}{f^2(z)}\omega^2(Z)\mathbf{1}_D(Z)\right]$$

where $\eta^2(z)$ represents the gain in variance obtained thanks to the use of $\Psi(X,\theta^0)$ as conditional control variates. Finally, the additional positive term H_D is the price we must pay for not knowing θ^0 and plugging in the estimator $\hat{\theta}_n^Q$. This is the reason why this term H_D is the exact analog, with a fairly similar expression, of the additional term put forward in the comments of theorem 3.6.

With these three steps of reasoning, the proof of theorem 4.8 appears to be fairly straightforward and the efficiency claim is fully warranted. Of course, all the efficiency arguments put forward in this section are only about first order asymptotics and one may wonder whether empirical likelihood should not be preferred for higher order asymptotics. However, nothing prevents us to propose a three-step Euclidean likelihood approach in order, as it has been done in the unconditional case, to mimic the properties of empirical likelihood, without involving the same computational burden. In other words, from a two-step efficient estimator $\tilde{\theta}_n$ of θ, we compute a third-step estimator $\hat{\theta}_n$ of θ by solving:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{E}_n^Q\left[\frac{\partial\Psi'}{\partial\theta}(\tilde{\theta}_n)\,|Z\right]\left[\hat{V}_n^Q\left[\Psi(\tilde{\theta}_n)\,|Z_i\right]\right]^{-1}\Psi\left(X_i,\hat{\theta}_n\right) = 0 \tag{4.31}$$

where $\hat{E}_n^Q\left[\frac{\partial\Psi'}{\partial\theta}(\tilde{\theta}_n)\,|Z_i\right]$ and $\hat{V}_n^Q\left[\Psi(\tilde{\theta}_n)\,|Z_i\right]$ denote here conditional control variate estimators of $E\left[\frac{\partial\Psi'}{\partial\theta}(X,\theta^0)\,|Z_i\right]$ and $V\left[\Psi\left(X,\theta^0\right)|Z_i\right]$ obtained by application of the formula of theorem 4.6, either with $\tilde{\theta}_n = \hat{\theta}_n^Q$ or, if preferred, with another efficient estimator $\tilde{\theta}_n$.

To figure out the advantages of this three-step estimator $\hat{\theta}_n$, several remarks are in order.

First, equations (4.31) mimic equations (4.16) which define the efficient 2S-GMM estimators with optimal instruments. But, by contrast with (4.16), optimal instruments have been computed

by taking advantage of the informational content of conditional moment restrictions. In particular, through the control variates principle, any perverse covariation between estimated optimal instruments and moment conditions has been deleted. Second, even though empirical likelihood first order conditions (see theorem 4.1 with $\lambda = 1$) share similar advantages, they involve two important computational drawbacks. Not only implied probabilities are much more difficult to compute, but, in addition, (4.11) introduces with respect to (4.31) an additional smoothing of moment conditions which does not appear to be necessary.

Finally, while the only finite-sample drawback of Euclidean empirical likelihood implied probabilities is that their positivity is not guaranteed, a shrinkage similar to the one proposed in the unconditional case may be introduced to ensure positivity. This appears to be especially important for the estimation of the covariance matrix.

5 Conclusion

In this paper we have presented a unified framework for learning from i.i.d. data when the only available prior information about the DGP is encapsulated in some moment conditions either conditional or unconditional. We have put a special emphasis on the usefulness of this learning process to estimate the unknown structural parameters θ defined by the moment conditions. The main message is that the widely documented poor finite-sample performance of two-step GMM is due likely to an inefficient use of the information contained in the moment restrictions in the first step.

Indeed, both θ and implied probabilities should be the focus of the first step, to efficiently estimate the optimal selection matrix or the optimal instruments. Moreover, we argue that chi-square distances and associated control variables estimation of expectations may be much more user-friendly than contenders, like empirical likelihood or Kullback-Leibler information criterion, without any efficiency loss, even at higher orders.

As far as other applications of the proposed implied probabilities are concerned, we expect that they should work remarkably well in practice to perform constrained Monte-Carlo simulations, to compute asset prices conformable to some pricing kernel model and to forecast out of sample. In particular, by contrast with their contenders, the proposed implied probabilities admit closed form formulas that can be used even with out-of-sample conditioning values.

APPENDIX

Proof of Theorem 2.1

The Lagrangian of problem (2.2) can be written:

$$\mathcal{L} = \sum_{i=1}^{n} h\left(\pi_i\right) - \beta' \sum_{i=1}^{n} \pi_i \Psi_i(\theta) - \mu \left(\sum_{i=1}^{n} \pi_i - 1 \right).$$

Then, the estimator $\hat{\pi}_1, \cdots, \hat{\pi}_n, \hat{\theta}$ are characterized, for well-suited values β_n and μ_n of the Lagrange multipliers, by the following first order conditions:

$$h_\pi\left(\hat{\pi}_i\right) = \beta_n' \Psi_i\left(\hat{\theta}\right) + \mu, \quad i = 1, \cdots, n \tag{A.1}$$

$$\beta_n' \sum_{i=1}^{n} \hat{\pi}_i \frac{\partial \Psi_i}{\partial \theta'}\left(\hat{\theta}\right) = 0 \tag{A.2}$$

When multiplying equation i of (A.1) by $\hat{\pi}_i \Psi_i'\left(\hat{\theta}\right)$ and summing over $i = 1, \cdots, n$ one gets:

$$\sum_{i=1}^{n} h_\pi\left(\hat{\pi}_i\right) \hat{\pi}_i \Psi_i'\left(\hat{\theta}\right) = \beta_n' \sum_{i=1}^{n} \hat{\pi}_i \Psi_i\left(\hat{\theta}\right) \Psi_i'\left(\hat{\theta}\right)$$

since, by definition:

$$\sum_{i=1}^{n} \hat{\pi}_i \Psi_i'\left(\hat{\theta}\right) = 0.$$

Therefore, the q-vector of Lagrange multipliers associated to the moment restrictions is:

$$\beta_n = \left[\sum_{i=1}^{n} \hat{\pi}_i \Psi_i\left(\hat{\theta}\right) \Psi_i'\left(\hat{\theta}\right) \right]^{-1} \sum_{i=1}^{n} h_\pi\left(\hat{\pi}_i\right) \hat{\pi}_i \Psi_i\left(\hat{\theta}\right)$$

By virtue of (A.2), this gives the announced result.

Proof of theorem 2.2:

With $h_\pi\left(\pi_i\right)$ proportional to $\pi_i^{-\lambda}$, (A.1) can be rewritten:

$$\hat{\pi}_{i,\lambda}^{-\lambda} = \mu_{n,\lambda} + \beta_{n,\lambda}' \Psi_i\left(\hat{\theta}_{n,\lambda}\right), \quad i = 1, \cdots n. \tag{A.3}$$

When multiplying equation i of (A.3) by $\hat{\pi}_{i,\lambda}$ and summing over $i = 1, \cdots, n$ one gets:

$$\sum_{i=1}^{n} \hat{\pi}_{i,\lambda}^{1-\lambda} = \mu_{n,\lambda}.$$

Therefore, $\mu_{n,\lambda} \neq 0$ and, by denoting $\alpha_{n,\lambda} = \beta_n/\mu_{n,\lambda}$, we rewrite (A.3) as:

$$\hat{\pi}_{i,\lambda}^{-\lambda} = \mu_{n,\lambda} \left[1 + \alpha'_{n,\lambda} \Psi_i \left(\hat{\theta}_{n,\lambda}\right)\right], i = 1, \cdots, n. \tag{A.4}$$

Proof of lemma 2.3 and theorem 2.4 and corollaries:

Proofs of this subsection can be seen as extensions of the proof of the so-called "empirical likelihood theorem" (Owen (2001), Theorem 2.2, p. 16). We start from a generalization of Owen (2001), lemma 11.2, p. 218:

Lemma A.1: Let Y_i be real-valued random variables with a common distribution and finite variance. Then,

$$\frac{1}{\sqrt{n}} \underset{1 \leq i \leq n}{Max} |Y_i| = o_P(1)$$

Proof of lemma A.1 is based on the following textbook formula (see e.g. Durrett (1996) p. 43) for computing expectations of positive random variables:

$$EY_1^2 = \int_0^{\infty} P\left[Y_1^2 > u\right] du \geq \sum_{n=0}^{\infty} P\left[Y_1^2 > n\right].$$

Then, if we consider the sequence of events $A_n = \left[Y_n^2 > n\right]$, we get:

$$\sum_{n=0}^{\infty} P(A_n) = \sum_{n=0}^{\infty} P\left[Y_1^2 > n\right] < +\infty.$$

Thus, by application of Borel-Cantelli lemma:

$$P\left[\lim Sup\ A_n\right] = 0.$$

It is worth noticing that, for any $\varepsilon > 0$, the same result applies to $A_n^{\varepsilon} = \left[Y_n^2 > n\varepsilon^2\right]$:

$$P\left[\lim SupA_n^{\varepsilon}\right] = 0.$$

To see this, it suffices to apply the previous argument to the sequence of random variables (Y_i/ε) which have a common distribution and finite variance. This will allow us to get the convergence result of lemma A.1, that is:

$$P\left[\underset{1 \leq i \leq n}{Max} |Y_i| > \varepsilon\sqrt{n}\right] \xrightarrow{n=\infty} 0$$

for all $\varepsilon > 0$.

Let $\eta > 0$. We want to show that there exists some n_0 such that:

$$n \geq n_0 \Longrightarrow P\left[\exists i = 1, \cdots n, \quad Y_i^2 > n\varepsilon^2\right] < \eta$$

For any given $N < n$, we can decompose:

$$\begin{aligned}
& P\left[\exists i = 1, \cdots n \qquad Y_i^2 > n\varepsilon^2\right] \\
\leq\ & P\left[\exists i = 1, \cdots, N \ , \quad Y_i^2 > n\varepsilon^2\right] \\
& + P\left[\exists i = N+1, \cdots, n \ , \quad Y_i^2 > n\varepsilon^2\right] \\
\leq\ & P\left[\underset{1 \leq i \leq N}{Max} Y_i^2 > n\varepsilon^2\right] + P\left[\exists i = N+1, \cdots, n \ , \quad Y_i^2 > i\varepsilon^2\right] \\
\leq\ & P\left[\underset{1 \leq i \leq N}{Max} Y_i^2 > n\varepsilon^2\right] + P\left[\underset{i > N}{\cup} A_i^{\varepsilon}\right].
\end{aligned}$$

From $P\left[Lim\ Sup\ A_n^{\varepsilon}\right] = 0$, we deduce that for some $N_0(\eta)$:

$$P\left[\underset{i > N_0(\eta)}{\cup A_i^{\varepsilon}}\right] < \frac{\eta}{2}.$$

Then, since for given $N_0(\eta)$, $\underset{1 \leq i \leq N_0(\eta)}{Max} Y_i^2$ is bounded in probability, we can find some $n_1(\eta)$ such that:

$$n > n_1(\eta) \Longrightarrow P\left[\underset{1 \leq i \leq N_0(\eta)}{Max} Y_i^2 > n\varepsilon^2\right] < \frac{\eta}{2}.$$

Thus:

$$\begin{aligned}
n \quad & > \quad Max\left[N_0(\eta), n_1(\eta)\right] \\
& \Longrightarrow \quad P\left[\exists i = 1, \cdots, n \ , \quad Y_i^2 > n\varepsilon^2\right] < \eta.
\end{aligned}$$

This completes the proof of lemma A.1. We deduce from this lemma that for all $\theta \epsilon \Theta$ and $\varepsilon > 0$,

$$\begin{aligned}
& P\left[\underset{1 \leq i \leq n}{Min}\left[1 + \alpha'_{n,\lambda}\Psi_i(\theta)\right] > 1 - \varepsilon\right] \\
=\ & 1 - P\left[\exists i = 1, \cdots n, \quad \alpha'_{n,\lambda}\Psi_i(\theta) < -\varepsilon\right] \\
\geq\ & 1 - P\left[\left\|\sqrt{n}\alpha_{n,\lambda}\right\| \frac{1}{\sqrt{n}} \underset{1 \leq i \leq n}{Max} \left\|\Psi_i(\theta)\right\| > \varepsilon\right].
\end{aligned}$$

This shows that theorem 2.4 will be implied by lemma A.1 applied to $Y_i = \left\|\Psi_i(\theta)\right\|$ insofar as $\sqrt{n}\alpha_{n,\lambda}$ is bounded in probability.

Regularity properties ensuring that $\alpha_{n,\lambda}$ and $\hat{\theta}_{n,\lambda} - \theta^0$ are $O_P(1/\sqrt{n})$ are well-known (see e.g. Imbens, Spady and Johnson (1998)) and will not be made explicit here. Under these maintained assumptions, the required positivity is thus guaranteed asymptotically almost certainly. Then, corollary 2.5 is directly implied by theorem 2.2 and implies in turn corollary 2.6 by application of (2.11).

As far as lemma 2.3 is concerned, note that since the Lagrange multipliers $\alpha_{n,\lambda}$ are characterized by the moment restrictions:

$$\sum_{i=1}^{n} \hat{\pi}_{i,\lambda} \Psi_i \left(\hat{\theta}_{n,\lambda} \right) = 0$$

with $\hat{\pi}_{i,\lambda}$ given by corollary 2.5, we have asymptotically almost certainly:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[1 + \alpha'_{n,\lambda} \Psi_i \left(\hat{\theta}_{n,\lambda} \right) \right]^{-1/\lambda} \Psi_i \left(\hat{\theta}_{n,\lambda} \right) = 0$$

By a Taylor expansion:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_i \left(\hat{\theta}_{n,\lambda} \right) - \frac{1}{\lambda} \left[\frac{1}{n} \sum_{i=1}^{n} \Psi_i \left(\hat{\theta}_{n,\lambda} \right) \Psi'_i \left(\hat{\theta}_{n,\lambda} \right) \right] \sqrt{n} \hat{\alpha}_{n,\lambda} = \mathcal{O}_P \left(1/\sqrt{n} \right)$$

Thus:

$$\sqrt{n} \hat{\alpha}_{n,\lambda} = \lambda \Omega_n^{-1} \left(\hat{\theta}_{n,\lambda} \right) \sqrt{n} \bar{\Psi}_n \left(\hat{\theta}_{n,\lambda} \right) + \mathcal{O}_P \left(1/\sqrt{n} \right),$$

which, ex post, confirms that $\hat{\alpha}_{n,\lambda}$ is $\mathcal{O}_P \left(1/\sqrt{n} \right)$ and thus, from (2.7):

$$\sqrt{n} \left(\hat{\theta}_{n,\lambda} - \theta^0 \right) = -\Sigma^{-1} \Gamma' \Omega^{-1} \sqrt{n} \bar{\Psi}_n \left(\theta^0 \right) + \mathcal{O}_P \left(1/\sqrt{n} \right).$$

Then, the above expansion of $\sqrt{n} \hat{\alpha}_{n,\lambda}$ can be rewritten:

$$\begin{aligned} \sqrt{n} \hat{\alpha}_{n,\lambda} &= \lambda \Omega_n^{-1} \left(\hat{\theta}_{n,\lambda} \right) \sqrt{n} \bar{\Psi}_n \left(\theta^0 \right) \\ &\quad + \lambda \Omega_n^{-1} \left(\hat{\theta}_{n,\lambda} \right) \frac{\partial \bar{\Psi}_n}{\partial \theta'} \left(\theta^0 \right) \cdot \sqrt{n} \left(\hat{\theta}_{n,\lambda} - \theta^0 \right) + \mathcal{O}_P \left(1/\sqrt{n} \right) \end{aligned}$$

Or:

$$\begin{aligned} \sqrt{n} \hat{\alpha}_{n,\lambda} &= \lambda \Omega^{-1} \sqrt{n} \bar{\Psi}_n \left(\theta^0 \right) \\ &\quad + \lambda \Omega^{-1} \Gamma \sqrt{n} \left(\hat{\theta}_{n,\lambda} - \theta^0 \right) + \mathcal{O}_P \left(1/\sqrt{n} \right). \end{aligned}$$

By plugging into the above expansion of $\sqrt{n} \left(\hat{\theta}_{n,\lambda} - \theta^0 \right)$ we get:

$$\sqrt{n} \hat{\alpha}_{n,\lambda} = \lambda \left[\Omega^{-1} - \Omega^{-1} \Gamma \Sigma^{-1} \Gamma' \Omega^{-1} \right] \sqrt{n} \bar{\Psi}_n \left(\theta^0 \right) + \mathcal{O}_P \left(1/\sqrt{n} \right)$$

which completes the proof of corollary 2.7. **Q.E.D**

Proof of theorem 2.8: The proof is based on the following lemma:

Lemma A.2:

$$\begin{aligned} & \sum_{i=1}^{n}\left(\alpha'_{n,\lambda}\Psi_i(\hat{\theta}_{n,\lambda})\right)^2\Psi_i(\hat{\theta}_{n,\lambda}) \\ = \quad & \lambda^2\sum_{i=1}^{q}\beta_{jn}\left(\theta^0\right)e_j+\mathcal{O}_P\left(1/\sqrt{n}\right) \end{aligned}$$

To see this, just note that by plugging into:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{n}\alpha'_{n,\lambda}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\right)^2\Psi_i\left(\hat{\theta}_{n,\lambda}\right)$$

the expansion of $\sqrt{n}\alpha_{n,\lambda}$ given by corollary 2.7, we get:

$$\begin{aligned} & \frac{\lambda^2}{n}\sum_{i=1}^{n}\sqrt{n}\bar{\Psi}'_n\left(\theta^0\right)P'\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\Psi'_i\left(\hat{\theta}_{n,\lambda}\right)P\sqrt{n}\bar{\Psi}_n\left(\theta^0\right)\Psi_i\left(\hat{\theta}_{n,\lambda}\right)+\mathcal{O}_P\left(1/\sqrt{n}\right) \\ = \quad & \lambda^2\sum_{j=1}^{q}\sqrt{n}\bar{\Psi}'_n\left(\theta^0\right)P'\left[\frac{1}{n}\sum_{i=1}^{n}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\Psi'_i\left(\hat{\theta}_{n,\lambda}\right)\Psi_i^j\left(\hat{\theta}_{n,\lambda}\right)\right]P\sqrt{n}\Psi_n\left(\theta^0\right)e_j+\mathcal{O}_P\left(1/\sqrt{n}\right) \end{aligned}$$

since: $\Psi_i\left(\hat{\theta}_{n,\lambda}\right)=\sum_{j=1}^{q}\Psi_i^j\left(\hat{\theta}_{n,\lambda}\right)e_j$.

This completes the proof of lemma A.2 by noting that:

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\Psi'_i\left(\hat{\theta}_{n,\lambda}\right)\Psi_i^j\left(\hat{\theta}_{n,\lambda}\right)=E\left[\Psi\left(X,\theta^0\right)\Psi'\left(X,\theta^0\right)\Psi^j\left(X,\theta^0\right)\right]+\mathcal{O}_P\left(1/\sqrt{n}\right)$$

Lemma A.2 allows us to push the expansion of lemma 2.3 one step further. Since $\hat{\alpha}_{n,\lambda}$ is defined asymptotically almost certainly by the q equations:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[1+\alpha'_{n,\lambda}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\right]^{-1/\lambda}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)=0$$

We get:

$$\begin{aligned} & \sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)-\frac{1}{\lambda}\left[\frac{1}{n}\sum_{i=1}^{n}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\Psi'_i\left(\hat{\theta}_{n,\lambda}\right)\right]\sqrt{n}\alpha_{n,\lambda} \\ + & \frac{1}{2\sqrt{n}}\left(-\frac{1}{\lambda}\right)\left(-\frac{1}{\lambda}-1\right)\sum_{i=1}^{n}\left(\alpha'_{n,\lambda}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\right)^2\Psi_i\left(\hat{\theta}_{n,\lambda}\right)=\mathcal{O}_P\left(1/n\right). \end{aligned}$$

Therefore, by lemma A.2:

$$\begin{aligned} & \Omega_n\left(\hat{\theta}_{n,\lambda}\right)\sqrt{n}\alpha_{n,\lambda} \\ = \;& \lambda\sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)+\frac{1}{2\sqrt{n}}\frac{\lambda+1}{\lambda}\sum_{j=1}^{q}\beta_{jn}\left(\theta^0\right)e_j+\mathcal{O}_P\left(1/n\right). \end{aligned} \tag{A.5}$$

Let us then consider:

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[1+\alpha'_{n,\lambda}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\right]^{\frac{\lambda-1}{\lambda}}\Psi_i\left(\hat{\theta}_{n,\lambda}\right) \\ = \;& \sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)+\frac{\lambda-1}{\lambda}\left[\frac{1}{n}\sum_{i=1}^{n}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\Psi'_i\left(\hat{\theta}_{n,\lambda}\right)\right]\sqrt{n}\alpha_{n,\lambda} \\ & -\frac{1}{2\sqrt{n}}\frac{\lambda-1}{\lambda^2}\sum_{i=1}^{n}\left(\alpha'_{n,\lambda}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\right)^2\Psi_i\left(\hat{\theta}_{n,\lambda}\right)+\mathcal{O}_P\left(1/n\right). \end{aligned}$$

By plugging in the above formula the stochastic expansions given by lemma A.2 and by (5.1), we get after regrouping similar terms:

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[1+\alpha'_{n,\lambda i}\Psi_i\left(\hat{\theta}_{n,\lambda}\right)\right]^{\frac{\lambda-1}{\lambda}}\Psi_i\left(\hat{\theta}_{n,\lambda}\right) \\ = \;& \lambda\sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,\lambda}\right)+\frac{\lambda\left(\lambda-1\right)}{2}\frac{1}{\sqrt{n}}\sum_{j=1}^{q}\beta_{jn}\left(\theta^0\right)e_j+\mathcal{O}_P\left(1/n\right). \end{aligned}$$

Proof of theorem 3.1 and corollaries

Since we can minimize (3.4) without taking care of positivity constraints, we get from (A.4) with $\lambda=-1$:

$$\pi_i\left(\theta\right)\text{ proportional to }\left[1+\alpha'_n\left(\theta\right)\Psi_i\left(\theta\right)\right].$$

Moreover, notice that:

$$\begin{aligned} \pi_i\left(\theta\right) &= \mu\left[1+\alpha'_n\left(\theta\right)\Psi_i\left(\theta\right)\right] \\ &\Longrightarrow\; 1=n\mu+n\mu\alpha'_n\left(\theta\right)\bar{\Psi}_n\left(\theta\right) \end{aligned} \tag{A.6}$$

Thus, with $\gamma_n\left(\theta\right)=n\mu\alpha_n\left(\theta\right)$:

$$\begin{aligned} \pi_i\left(\theta\right) &= \mu+\frac{\gamma'_n\left(\theta\right)}{n}\Psi_i\left(\theta\right) \\ &= \frac{1}{n}\left[n\mu+\gamma'_n\left(\theta\right)\bar{\Psi}_n\left(\theta\right)+\gamma'_n\left(\theta\right)\left(\Psi_i\left(\theta\right)-\bar{\Psi}_n\left(\theta\right)\right)\right] \\ &= \frac{1}{n}\left[1+\gamma'_n\left(\theta\right)\left(\Psi_i\left(\theta\right)-\bar{\Psi}_n\left(\theta\right)\right)\right]. \end{aligned}$$

To summarize:

$$\begin{aligned} \pi_i(\theta) &= \mu\left[1+\alpha_n'(\theta)\,\Psi_i(\theta)\right] \\ &= \frac{1}{n}\left[1+\gamma_n'(\theta)\left(\Psi_i(\theta)-\bar{\Psi}_n(\theta)\right)\right]. \end{aligned} \tag{A.7}$$

From the moment conditions:

$$\sum_{i=1}^{n}\pi_i(\theta)\,\Psi_i(\theta)=0$$

we then deduce:

$$\alpha_n(\theta)=-\Omega_n^{-1}(\theta)\,\bar{\Psi}_n(\theta)$$

and

$$\gamma_n(\theta)=-V_n^{-1}(\theta)\,\bar{\Psi}_n(\theta).$$

Thus, by plugging this value of $\gamma_n(\theta)$ into (5.1):

$$\pi_i(\theta)=\frac{1}{n}-\frac{1}{n}\bar{\Psi}_n'(\theta)\,V_n^{-1}(\theta)\left[\Psi_i(\theta)-\bar{\Psi}_n(\theta)\right].$$

To get corollary 3.3, we compute $\sum_{i=1}^{n}\pi_i^2(\theta)$ from the two alternative expressions (5.1) of $\pi_i(\theta)$:

$$\begin{aligned} &\sum_{i=1}^{n}\pi_i^2(\theta) \\ =\;& n\mu^2\left[1+2\alpha_n'(\theta)\,\bar{\Psi}_n(\theta)+\alpha_n'(\theta)\,\Omega_n(\theta)\,\alpha_n(\theta)\right] \\ =\;& \frac{1}{n}\left[1+\gamma_n'(\theta)\,V_n(\theta)\,\gamma_n(\theta)\right] \end{aligned}$$

that is, by plugging into the above expressions of $\alpha_n(\theta)$ and $\gamma_n(\theta)$:

$$\begin{aligned} \sum_{i=1}^{n}\pi_i^2(\theta) &= n\mu^2\left[1-\bar{\Psi}_n'(\theta)\,\Omega_n^{-1}(\theta)\,\bar{\Psi}_n(\theta)\right] \\ &= \frac{1}{n}\left[1+\bar{\Psi}_n(\theta)\,V_n^{-1}(\theta)\,\bar{\Psi}_n(\theta)\right]. \end{aligned}$$

This gives the two announced formulas when taking into account that from (5.1):

$$\begin{aligned} n\mu &= \left[1+\alpha_n'\bar{\Psi}_n(\theta)\right]^{-1} \\ &= \left[1-\bar{\Psi}_n'(\theta)\,\Omega_n^{-1}(\theta)\,\bar{\Psi}_n(\theta)\right]^{-1}. \end{aligned}$$

Moreover, since both $\alpha_n\left(\hat{\theta}_n^Q\right)$ and $\gamma_n\left(\hat{\theta}_n^Q\right)$ are proportional to the vector of Lagrange multipliers, we know from (A.2) that the first order conditions for θ can be written:

$$\alpha_n'\left(\hat{\theta}_n^Q\right)\sum_{i=1}^{n}\pi_i\left(\hat{\theta}_n^Q\right)\frac{\partial\Psi_i}{\partial\theta'}\left(\hat{\theta}_n^Q\right)=0$$

or

$$\gamma_n'\left(\hat{\theta}_n^Q\right)\sum_{i=1}^{n}\pi_i\left(\hat{\theta}_n^Q\right)\frac{\partial\Psi_i}{\partial\theta'}\left(\hat{\theta}_n^Q\right)=0,$$

which gives the formulas of corollary 3.4 when replacing $\alpha_n(\theta)$ and $\gamma_n(\theta)$ by their above expression.

Corollary 3.5 is a straightforward implication of the expression of $\pi_i\left(\hat{\theta}_n^Q\right)$ given by theorem 3.1.

Proof of theorem 3.6:

With the notations of (A.4) and (5.1):

$$\begin{aligned}&\sqrt{n}\left(\hat{g}_n(\hat{\theta}_n^Q)-Eg(X)\right)\\ =\;&\sqrt{n}\left(\bar{g}_n-Eg(X)\right)-\hat{a}_n'\sqrt{n}\bar{\Psi}_n(\hat{\theta}_n^Q)+o_P(1)\\ =\;&\sqrt{n}\left(\bar{g}_n-Eg(X)\right)-a'\Omega P\sqrt{n}\bar{\Psi}_n(\theta^0)+o_P(1)\end{aligned}$$

where the last equality is deduced from corollary 2.7. Thus, since:

$$\Omega P=Id-\Gamma\Sigma^{-1}\Gamma'\Omega^{-1},$$

we get:

$$\begin{aligned}&\sqrt{n}\left(\hat{g}_n(\hat{\theta}_n^Q)-Eg(X)\right)\\ &\sqrt{n}\left(\bar{g}_n-a'\bar{\Psi}_n\left(\theta^0\right)-Eg(X)\right)+a'\Omega\Sigma^{-1}\Gamma'\Omega^{-1}\sqrt{n}\bar{\Psi}_n(\theta^0)+o_P(1)\end{aligned}$$

Moreover, by definition of a, the two asymptotically normal variables $\sqrt{n}\left(\hat{g}_n-a'\Psi_n(\theta^0)\right)$ and $\sqrt{n}\bar{\Psi}_n(\theta^0)$ are asymptotically independent. We can then conclude that $\sqrt{n}\left(\hat{g}_n(\hat{\theta}_n^Q)-Eg(X)\right)$ converges in distribution toward a zero-mean normal distribution with variance $\Sigma_1+\Sigma_2$ where:

$$\Sigma_1=Var\;g(X)-Cov\left[g(X),\Psi(X,\theta^0)\right]\Omega^{-1}Cov\left[\Psi\left(X,\theta^0\right),g(X)\right]$$

is the asymptotic variance of $\sqrt{n}\left(\bar{g}_n-a'\bar{\Psi}_n\left(\theta^0\right)\right)$

and

$$\Sigma_2=a'\Gamma\Sigma^{-1}\Gamma'\Omega^{-1}\Gamma\Sigma^{-1}\Gamma'a$$

is the asymptotic variance of

$$a'\Gamma\Sigma^{-1}\Gamma'\Omega^{-1}\sqrt{n}\bar{\Psi}_n(\theta^0).$$

Therefore, we will get the announced formula of theorem 3.6 if we check that:

$$\Omega^{-1}\Gamma\Sigma^{-1}\Gamma'\Omega^{-1}\Gamma\Sigma^{-1}\Gamma'\Omega^{-1} = \Omega^{-1} - P$$

that is:

$$\Omega^{-1} - P = \left(\Omega^{-1} - P\right)\Omega\left(\Omega^{-1} - P\right)$$

or:

$$Id - \Omega P = (Id - \Omega P)(Id - \Omega P).$$

This equality is fulfilled since $(Id - \Omega P)$ is a projection matrix:

$$Id - \Omega P = \Gamma\left(\Gamma'\Omega^{-1}\Gamma\right)^{-1}\Gamma'\Omega^{-1}$$

is a projection on the vectorial space spanned by the columns of Γ.

Proof of theorem 3.7 and corollaries:

From (A.1) and (A.2), the augmented set of moment conditions defines estimators $\left(\underset{\sim}{\hat{\pi}}_{i,,\lambda}, \underset{\sim}{\hat{\theta}}_{n,\lambda}\right)$ and Lagrange multipliers $\left(\underset{\sim}{\beta}_n, \underset{\sim}{\mu}_n\right)$ such that:

$$\begin{aligned}
\underset{\sim}{\hat{\pi}}_{i,\lambda} &= \underset{\sim}{\beta}'_n \underset{\sim}{\Psi}_i\left(\underset{\sim}{\hat{\theta}}_{n,\lambda}\right) + \underset{\sim}{\mu}_n \\
\underset{\sim}{\beta}'_n \cdot \sum_{i=1}^{n} \underset{\sim}{\hat{\pi}}_{i,\lambda} \frac{\partial \underset{\sim}{\Psi}_i}{\partial \underset{\sim}{\theta}'}\left(\underset{\sim}{\hat{\theta}}_{n,\lambda}\right) &= 0
\end{aligned}$$

With $\underset{\sim}{\beta}_n = (\beta'_n, \delta_n)'$, the second set of equations can be decomposed in:

$$\begin{cases}
\underset{\sim}{\beta}'_n \sum_{i=1}^{n} \underset{\sim}{\hat{\pi}}_{i,\lambda} \dfrac{\partial \Psi_i}{\partial \theta'}\left(\hat{\theta}_{n,\lambda}\right) = 0 \\
\delta_n \sum_{i=1}^{n} \underset{\sim}{\hat{\pi}}_{i,\lambda} = 0
\end{cases}$$

Therefore $\delta_n = 0$ and the above first order conditions can be rewritten:

$$\begin{cases} \underset{\sim}{\hat{\pi}}_{i,\lambda} = \beta_n' \Psi_i \left(\hat{\theta}_{n,\lambda}\right) + \mu_n \\ \beta_n' \cdot \sum_{i=1}^{n} \underset{\sim}{\hat{\pi}}_{i,\lambda} \dfrac{\partial \Psi_i}{\partial \theta'}(\hat{\theta}_{n,\lambda}) = 0 \end{cases}$$

They coincide with (A.1) and (A.2), and jointly with the corresponding constraints, define the same estimators $\left(\hat{\pi}_{i,\lambda}, \hat{\theta}_{n,\lambda}\right)$ and multipliers (β_n, μ_n). Finally, $\hat{\xi}_{n,\lambda}$ is defined by the additional constraint

$$\sum_{i=1}^{n} \hat{\pi}_{i,\lambda} \left[g\left(X_i\right) - \hat{\xi}_{n,\lambda}\right] = 0$$

which completes the proof of theorem 3.7. Corollaries 3.8, 3.19 and 3.10 are obvious implications. Finally:

$$\begin{aligned} & \sqrt{n}\left[g_n^*(\hat{\theta}_n^Q) - \hat{g}_n(\hat{\theta}_n^Q)\right] \\ = & \frac{\varepsilon_n(\hat{\theta}_n^Q)}{1+\varepsilon_n(\hat{\theta}_n^Q)} \sqrt{n}\left[\bar{g}_n - \hat{g}_n(\hat{\theta}_n^Q)\right]. \end{aligned}$$

But, since all the $\pi_i(\hat{\theta}_n^Q), i = 1, \cdots, n$ are asymptotically nonnegative with probability 1:

$$\varepsilon_n(\hat{\theta}_n^Q) = o_P\left(1\right)$$

Since:

$$\sqrt{n}\left[\bar{g}_n - \hat{g}_n(\hat{\theta}_n^Q)\right] = \mathcal{O}_P\left(1\right)$$

We can conclude that:

$$\sqrt{n}\left[g_n^*(\hat{\theta}_n^Q) - \hat{g}_n(\hat{\theta}_n^Q)\right] = o_P\left(1\right).$$

Proof of theorem 3.12:

We want to compare the two estimators $\hat{\theta}_n$ and $\hat{\theta}_{n,1}$ defined respectively by:

$$g_n\left(\hat{\theta}_n\right) = 0 \text{ and } f_n\left(\hat{\theta}_{n,1}\right) = 0$$

where:

$$f_n\left(\theta\right) = \left[\hat{E}_{n,1} \frac{\partial \Psi_i'}{\partial \theta}\left(\theta\right)\right] \left[\hat{V}_{n,1} \Psi_i\left(\theta\right)\right]^{-1} \bar{\Psi}_n\left(\theta\right)$$

and

$$g_n\left(\theta\right) = \left[\hat{\Gamma}_n^{Q\prime}\left(\tilde{\theta}_n\right)\right] \left[\hat{\Omega}_n^{Q\prime}\left(\tilde{\theta}_n\right)\right]^{-1} \bar{\Psi}_n\left(\theta\right) + \mathcal{O}_P(1/n\sqrt{n}).$$

In order to apply Robinson (1988), theorem 1, we need to check that the two properties of his assumption (A.2) are fulfilled. A first condition is:

$$\frac{\partial g_n}{\partial \theta'}\left(\theta^0\right) = G + o_P(1)$$

with G nonsingular matrix. This condition is fulfilled here with:

$$G = 2\Gamma'\left(\theta^0\right)\left[\Omega\left(\theta^0\right)\right]^{-1}\Gamma\left(\theta^0\right).$$

A second condition is akin to a kind of asymptotic continuity of $\frac{\partial g_n}{\partial \theta'}(\theta)$ at the point θ^0. This assumption will be maintained here.

Then, we can conclude from Robinson (1988) that:

$$\begin{aligned}\hat{\theta}_n - \hat{\theta}_{n,1} &= \mathcal{O}_P\left(\left\|g_n\left(\hat{\theta}_{n,1}\right)\right\|\right) \\ &= \mathcal{O}_P\left(\left\|g_n\left(\hat{\theta}_{n,1}\right) - f_n\left(\hat{\theta}_{n,1}\right)\right\|\right).\end{aligned}$$

Since $\sqrt{n}\bar{\Psi}_n\left(\hat{\theta}_{n,1}\right) = \mathcal{O}_P(1)$, we will then get the announced result if we show that:

$$\left\|\hat{\Gamma}_n^Q\left(\tilde{\theta}_n\right) - \hat{E}_{n,1}\frac{\partial \Psi_i}{\partial \theta'}\left(\hat{\theta}_{n,1}\right)\right\| = \mathcal{O}_P(1/n)$$

and

$$\left\|\hat{\Omega}_n^Q\left(\tilde{\theta}_n\right) - \hat{V}_{n,1}\Psi_i\left(\hat{\theta}_{n,1}\right)\right\| = \mathcal{O}_P(1/n).$$

Since $\tilde{\theta}_n - \tilde{\theta}_n^Q = \mathcal{O}_P(1/n)$, a simple Taylor expansion argument actually gives:

$$\left\|\hat{\Gamma}_n^Q\left(\tilde{\theta}_n\right) - \hat{\Gamma}_n^Q(\hat{\theta}_n^Q)\right\| = \mathcal{O}_P(1/n)$$

and

$$\left\|\hat{\Omega}_n^Q\left(\tilde{\theta}_n\right) - \hat{\Omega}_n^Q(\hat{\theta}_n^Q)\right\| = \mathcal{O}_P(1/n).$$

Therefore, by a triangle inequality argument, we conclude that we just need to show that:

$$\left\|\hat{\Gamma}_n^Q(\hat{\theta}_n^Q) - \hat{E}_{n,1}\frac{\partial \Psi_i}{\partial \theta'}(\hat{\theta}_{n,1})\right\| = \mathcal{O}_P(1/n)$$

and

$$\left\|\hat{\Omega}_n^Q\left(\hat{\theta}_n^Q\right) - \hat{V}_{n,1}\Psi_i\left(\hat{\theta}_{n,1}\right)\right\| = \mathcal{O}_P(1/n).$$

This is a straightforward consequence of corollary 3.8.

Proof of Theorem 4.1:

When optimizing (4.6) under the constraints (4.8), the Lagrangian can be written:

$$\begin{aligned}\ell &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{\lambda} h^{(\lambda)}\left(\pi_{i,j}\right) \\ &\quad - \sum_{i=1}^{n} \beta_i' \sum_{j=1}^{n} \pi_{ij} \Psi_j\left(\theta\right) \\ &\quad - \sum_{i=1}^{n} \mu_i \left(\sum_{j=1}^{n} \pi_{ij} - 1 \right).\end{aligned}$$

Then, the estimators $\hat{\pi}_{i,j,\lambda}, \hat{\theta}_{n,\lambda}$ are characterized, for well-suited values β_{in} and μ_{in} of the Lagrange multipliers, by the following first order conditions:

$$w_{ij}^{\lambda} \hat{\pi}_{i,j,\lambda}^{-\lambda} = \beta_{in}' \Psi_j\left(\hat{\theta}_{n,\lambda}\right) + \mu_{in}, i, j = 1, \cdots, n, \tag{A.8}$$

$$\sum_{i=1}^{n} \beta_{in}' \sum_{j=1}^{n} \hat{\pi}_{i,j,\lambda} \frac{\partial \Psi_j}{\partial \theta'}\left(\hat{\theta}_{n,\lambda}\right) = 0 \tag{A.9}$$

When multiplying equation (i,j) of (A.8) by $\hat{\pi}_{i,j,\lambda} \Psi_j'\left(\hat{\theta}_{n,\lambda}\right)$ and summing over $j = 1, \cdots, n$ one gets, for $i = 1, \cdots, n$:

$$\sum_{j=1}^{n} w_{ij}^{\lambda} \hat{\pi}_{i,j,\lambda}^{1-\lambda} \Psi_j'\left(\hat{\theta}_{n,\lambda}\right) = \beta_{in}' \sum_{j=1}^{n} \hat{\pi}_{i,j,\lambda} \Psi_j\left(\hat{\theta}_{n,\lambda}\right) \Psi_j'\left(\hat{\theta}_{n,\lambda}\right),$$

since, by definition:

$$\sum_{j=1}^{n} \hat{\pi}_{i,j,\lambda} \Psi_j'\left(\hat{\theta}_{n,\lambda}\right) = 0.$$

Therefore, the q-vector of Lagrange multipliers associated to the conditional moment restrictions given $Z = Z_i$ is:

$$\beta_{in} = \left[\sum_{j=1}^{n} \hat{\pi}_{i,j,\lambda} \Psi_j\left(\hat{\theta}_{n,\lambda}\right) \Psi_j'\left(\hat{\theta}_{n,\lambda}\right) \right]^{-1} \sum_{j=1}^{n} w_{ij}^{\lambda} \hat{\pi}_{i,j,\lambda}^{1-\lambda} \Psi_j\left(\hat{\theta}_{n,\lambda}\right).$$

By virtue of (A.9), this gives the announced result.

Proof of theorem 4.2:

From (A.8), we have:

$$\hat{\pi}_{i,j,\lambda}^{-\lambda} = w_{i,j}^{-\lambda}\left[\mu_{in} + \beta_{in}'\Psi_j\left(\hat{\theta}_{n,\lambda}\right)\right] \tag{A.10}$$

when multiplying equation (i,j) of (A.10) by $w_{i,j}^{\lambda}\hat{\pi}_{i,j,\lambda}$ and summing over $j = 1, \cdots, n$, one gets:

$$\sum_{j=1}^{n} w_{i,j}^{\lambda}\hat{\pi}_{i,j,\lambda}^{1-\lambda} = \mu_{i,n}.$$

Therefore, $\mu_{i,n} \neq 0$ and, by denoting $\mu_{i,n} = \mu_{i,n,\lambda}$ and $\alpha_{i,n,\lambda} = \beta_{i,n}/\mu_{i,n}$, we rewrite (A.10) as:

$$\hat{\pi}_{i,j,\lambda}^{-\lambda} = \mu_{i,n,\lambda}w_{i,j}^{-\lambda}\left[1 + \alpha_{i,n,\lambda}'\Psi_j\left(\hat{\theta}_{n,\lambda}\right)\right]. \tag{A.11}$$

Proof of theorem 4.3 and corollaries:

Since we can minimize (4.19) without taking care of positivity constraints, we get from theorem 4.2 with $\lambda = -1$: For $i = 1, \cdots, n$ the $\pi_{i,j}(\theta), j = 1, \cdots, n$, are proportional to:

$$w_{ij}\left[1 + \alpha_{i,n}'(\theta)\Psi_j(\theta)\right].$$

Moreover, notice that:

$$\begin{array}{rcl} \pi_{i,j}(\theta) & = & \mu_i w_{ij}\left[1 + \alpha_{i,n}'(\theta)\Psi_j(\theta)\right] \\ & \Longrightarrow & 1 = \mu_i + \mu_i\alpha_{i,n}'(\theta)\bar{\Psi}_j(\theta). \end{array} \tag{A.12}$$

Thus, with $\gamma_{i,n}(\theta) = \mu_i\alpha_{i,n}(\theta)$:

$$\begin{array}{rcl} \pi_{i,j}(\theta) & = & \mu_i w_{ij} + w_{ij}\gamma_{i,n}'(\theta)\bar{\Psi}_i(\theta) + w_{ij}\gamma_{i,n}'(\theta)\left[\Psi_j(\theta) - \bar{\Psi}_i(\theta)\right] \\ & & w_{ij} + w_{ij}\gamma_{i,n}'(\theta)\left[\Psi_j(\theta) - \bar{\Psi}_i(\theta)\right]. \end{array}$$

From the moment conditions

$$\sum_{j=1}^{n}\pi_{i,j}(\theta)\Psi_j(\theta) = 0$$

We then deduce:

$$\begin{array}{rcl} \alpha_{i,n}(\theta) & = & -\Omega_n^{-1}(\theta\,|Z_i)\bar{\Psi}^i(\theta) \\ & & \text{and} \\ \gamma_{i,n}(\theta) & = & -V_n^{-1}(\theta\,|Z_i)\bar{\Psi}^i(\theta). \end{array}$$

Thus, by plugging this value of $\gamma_{i,n}(\theta)$ into the above expression of $\pi_{i,j}(\theta)$ we get:

$$\pi_{i,j}(\theta) = w_{ij} - w_{ij}\bar{\Psi}_j'(\theta)V_n^{-1}(\theta\,|Z_i)\left[\Psi_j(\theta) - \bar{\Psi}^i(\theta)\right].$$

To get corollary 4.4, we compute $\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{i,j}^2(\theta)}{w_{ij}}$ from the two alternative expressions of $\pi_{i,j}(\theta)$:

$$\begin{aligned}\sum_{i,j}\frac{\pi_{i,j}^2(\theta)}{w_{ij}} &= \sum_{i=1}^{n}\mu_i^2\sum_{j=1}^{n}w_{ij}\left[1-\bar{\Psi}_i'(\theta)\Omega_n^{-1}(\theta\,|Z_i)\bar{\Psi}_j(\theta)\right]^2\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left[1-\bar{\Psi}_i'(\theta)V_n^{-1}(\theta\,|Z_i)\left(\Psi_j(\theta)-\bar{\Psi}^i(\theta)\right)\right]^2.\end{aligned}$$

Therefore:

$$\begin{aligned}\sum_{i,j}\frac{\pi_{i,j}^2(\theta)}{w_{ij}} &= \sum_{i=1}^{n}\mu_i^2\left[1-\bar{\Psi}_i'(\theta)\Omega_n^{-1}(\theta\,|Z_i)\bar{\Psi}_j(\theta)\right]\\ &= \sum_{i=1}^{n}\left[1+\bar{\Psi}_i'(\theta)V_n^{-1}(\theta\,|Z_i)\bar{\Psi}_i(\theta)\right].\end{aligned}$$

Since by (A.12):

$$\mu_i=\left[1+\alpha_{i,n}'(\theta)\bar{\Psi}_i(\theta)\right]^{-1}=\left[1-\bar{\Psi}_i'(\theta)\Omega_n^{-1}(\theta\,|Z_i)\bar{\Psi}_j(\theta)\right]^{-1}$$

we conclude that:

$$\sum_{i=1}^{n}\mu_i^2\left[1-\bar{\Psi}_i'(\theta)\Omega_n^{-1}(\theta\,|Z_i)\bar{\Psi}_i(\theta)\right]=\sum_{i=1}^{n}\left[1-\bar{\Psi}_i'(\theta)\Omega_n^{-1}(\theta\,|Z_i)\bar{\Psi}_i(\theta)\right]^{-1}$$

By comparing (A.8) and (A.12), we see that $\beta_{in}=\gamma_{in}$, so that plugging the value of γ_{in} in first order conditions (A.9) provides the characterization of corollary 4.5.

References

[1] Ai., C. and Chen, X. (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions", *Econometrica* **71**, 6, 1795-1843.

[2] Ai, C. and Chen, X. (2001), "Efficient Sieve Minimum Distance Estimation of Semiparametric Conditional Moment Models", W.P.

[3] Altonji, J. G. and Segal, L. M. (1996), "Small Sample Bias in GMM Estimation of Covariance Structures", *Journal of Economic and Business Statistics* **14**, 353-366.

[4] Andrews, D. W. K. (1999), "Moment Selection Procedures for Generalized Method of Moments Estimation", *Econometrica* **67**, 543-564.

[5] Back, K. and Brown, D. P. (1993) "Implied probabilities in GMM estimators", *Econometrica* **61**, No.4, 971-975.

[6] Baggerly, K. A. (1998) "Empirical Likelihood as a goodness-of-fit Measure", *Biometrika* **85**, No.3, 535-547.

[7] Bera, A. K. and Bilias Y. (2002), "The MM, ME, ML, EL, EF and GMM approaches to estimation: a synthesis", *Journal of Econometrics* **107**, 51-86.

[8] Bonnal, H. (2003), "Estimation d'une distribution de probabilité sous contraintes de moments", forthcoming, Ph.D. thesis, Université de Rennes 1.

[9] Chamberlain, G. (1987) "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics* **34**, 305-334.

[10] Corcoran, S. (1998), "Bartlett Adjustment of Empirical Discrepancy Statistics", *Biometrika* **85**, No.4, 967-972.

[11] Donald, S. G., Imbens, G. W. and Newey, W.K. (2001), "Empirical Likelihood Estimation and Consistent Tests with Conditional Moment Restrictions", W.P.

[12] Donald, S. G. and Newey, W.K. (2000), "A Jackknife Interpretation of the Continuous Updating Estimator", *Economics Letters* **67**, 239-243.

[13] Durrett, R. (1996), "Probability: Theory and Examples, Second Edition", Duxbury Press, ITP.

[14] Fieller, F.C. and H.O. Hartley (1954), "Sampling with Control Variables", *Biometrika* **41**, 3/4, 494-501.

[15] Godambe, V. P. and Thompson (1989), "An Extension of Quasi-likelihood Estimation" (with discussion), *Journal of Statistical Planning and Inference* **22**, 137-172.

[16] Hall, A. (2000), "Covariance Matrix Estimation and the Power of the Overidentifying Restrictions Test", *Econometrica* **68**, 1517-1527.

[17] Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators", 1982, *Econometrica* **50**, No.4, 1029-1054.

[18] Hansen, L. P., Heaton, J. and Yaron, A. (1996), "Finite-Sample Properties of Some Alternative GMM Estimators", *Journal of Business & Economic Statistics* **14**, No.3, 262-280.

[19] Imbens, G. W. (1997), "One-Step Estimators for Over-identified Generalized Method of Moments Models", *Review of Economic Studies* **64**, No.1, 359-383.

[20] Imbens, G. W., Spady, R. H. and Johnson P. (1998), "Information Theoretic Approaches to Inference in Moment Conditions Models", *Econometrica* **66**, No.2, 333-357.

[21] Kitamura, Y. and Stutzer, M. (1997),. "An Information Theoretic Alternative to Generalized Method of Moments Estimation", *Econometrica* **65**, No.4, 861-874.

[22] Kitamura, Y. Tripathi, G. and Ahn, H. (2000), "Empirical Likelihood-based Inference in Conditional Moment Restriction Models", W.P.

[23] Ledoit, O. and M. Wolf (2001), "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection", *Journal of Empirical Finance*, forthcoming.

[24] Maasoumi E. (1993), "A Compendium to Information Theory in Economics and Econometrics", *Econometric Reviews*, 12(2), 137-181.

[25] Newey, W. K. (1993), "Efficient Estimation of Models with Conditional Moment Restrictions" in G. S. Maddala, C. R. Rao and H. D. Vinod Eds, *Handbook of Statistics*, volume 11: *Econometrics*, North Holland.

[26] Newey, W. and R. J. Smith (2004), "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators", *Econometrica* **72**, 1, 219-255.

[27] Owen, A. (1990) "Empirical Likelihood Ratio Confidence Regions", *The Annals of Statistics* **18**, No.1, 90-120.

[28] Owen, A. (1991) "Empirical Likelihood for Linear Models", *The Annals of Statistics* **19**, No.4, 1725-1747.

[29] Owen, A. (2001) *Empirical Likelihood*, Chapman & Hall.

[30] Pakes, A. and Pollard, D. (1989), "Simulation and the Asymptotics of Optimization Estimators", *Econometrica* **57**, No. 5, 1027-1057.

[31] Qin, J. and Lawless, J. (1994), "Empirical Likelihood and General Estimating Equations", *The Annals of Statistics* **22**, No.1, 300-325.

[32] Read, T. R., C. and Cressie, N. (1988) *Goodness-of-Fit Statistics for Discrete Multivariate Data*, 1988, Springer-Verlag, New York.

[33] Robinson, P. M. (1988), "The Stochastic Difference Between Econometric Statistics", *Econometrica*, 56, 531-548.

[34] Rothenberg, T. J. (1984), "Approximating the Distributions of Econometric Estimators and Test Statistics", in *Handbook of Econometrics*, Vol. 2, ed. by Z. Griliches and M. D. Intriligator, North-Holland.

[35] Severini, T.A. and G. Tripathi (2001), "A Simplified Approach to Computing Efficiency Bounds in Semiparametric Models"*, Journal of Econometrics* ***102,*** *23-66.*

[36] Smith, R. J. (2000), *Empirical Likelihood Estimation and Inference,* in "Applications of Differential Geometry to Econometrics", P. Marriott and M. Salmon Eds, Cambridge, University Press.

[37] Zellner, A. (1991), "Bayesian Methods and Entropy in Economics and Econometrics," in Grandy, W. T. and Schick, L. H. (eds.)., Maximum Entropy and Bayesian Methods, Dordrecht/Boston/London: Kluwer Academic Publishers, 17-31.

[38] Zellner, A. (2003), "Some Aspects of the History of Bayesian Information Processing", W.P. Chicago GSB.

[39] Zellner, A. and Tobias, J. (2001), "Further Results on Bayesian Method of Moments Analysis of the Multiple Regression Model", *International Economic Review* **42**, No. 1, 121-140.