Semiparametric Efficiency of GMM under Approximate Constraints

Paul Rochet

Abstract

Generalized empirical likelihood and generalized method of moments are well spread methods of resolution of inverse problems in econometrics. Each method defines a specific semiparametric model for which it is possible to calculate efficiency bounds. By this approach, we provide a new proof of Chamberlain's result on optimal GMM. We also discuss conditions under which GMM estimators remain efficient with approximate moment constraints.

Keywords: GMM; Efficiency Bound; Approximate Constraint.

1   Introduction

We tackle the problem of recovering an unknown probability measure μ based on a sample $X_1, .., X_n$ of i.i.d. realizations with distribution μ, where additional information on μ is available in the form of a set of moments equations

$$\int \Phi(x) d\mu(x) = 0, \tag{1}$$

for some vector valued function Φ. This kind of inverse problems finds many practical applications in econometrics, notably when dealing with instrumental variables, see for instance Donald et al. (2009). In some cases, the function Φ is not known exactly but is assumed to belong to some parametric family $\left\{ \Phi(\theta, .), \theta \in \Theta \subset \mathbb{R}^d \right\}$. We are then interested in the estimation of the true value θ_0 of the parameter, which is, the zero of $\theta \mapsto \int \Phi(\theta, .) d\mu$. The problem of estimating θ_0 in this context has been widely studied in the literature. Two main methods of estimation have been implemented, namely the generalized method of moments (GMM), introduced in Hansen (1982) and the generalized empirical likelihood (GEL), developed in Qin and Lawless

(1994) for this particular context.

Although these two methods aim to estimate the same quantity, we point out that they rely on different descriptions of the statistical model. Hence, each method is related to a specific semiparametric model, for which we can calculate the efficiency bound for estimating θ_0, following van der Vaart (1998). By this approach, we exhibit necessary conditions for efficiency of GMM and we recover some known results of Hansen (1982) and Chamberlain (1987) on optimal GMM.

In many actual situations, the function Φ may have a complicated form that can only be evaluated numerically. Simulation-based methods have been implemented to deal with approximate constraints, see for instance Mcfadden (1989) and Carrasco and Florens (2000). In this paper we extend the GMM framework to situations where only an approximation Φ_m of the true constraint function Φ is available. We provide conditions under which GMM procedures remain efficient asymptotically when replacing Φ by its approximation.

The article falls into the following parts. After exposing the model in Section 2, we make a brief survey on the main methods of estimation in this model, and provide a new proof on the semiparametric efficiency of GMM in Section 3. In Section 4, we discuss the asymptotic efficiency of the method when dealing with an approximate constraint. Proofs are postponed to the Appendix.

2   The model

Let $\mathcal{X}$ be an open subset of $\mathbb{R}^q$, endowed with its Borel field $\mathcal{B}(\mathcal{X})$. We observe an i.i.d. sample $X_1, ..., X_n$ with unknown distribution μ. We are interested in the estimation of a parameter $\theta_0 \in \Theta \subset \mathbb{R}^d$ defined by the moment condition

$$F(\theta_0, \mu) := \int \Phi(\theta_0, x) d\mu(x) = 0, \tag{2}$$

where $\Phi : \Theta \times \mathcal{X} \to \mathbb{R}^k$ ($k \geq d$) is a known map. The question of estimating efficiently θ_0 relies on the amount of information available on μ. Here, the information given by the moment condition (2) is used to determine the set $\mathcal{M}$ of possible values for μ (the model). The true value θ_0 of the parameter being unknown, the distribution of the observations can be any probability measure ν for which the map $\theta \mapsto F(\theta, \nu)$ is null for some value of $\theta = \theta(\nu) \in \Theta$. The model is therefore defined as

$$\mathcal{M} = \left\{ \nu \in \mathcal{P}(\mathcal{X}) : \exists \theta = \theta(\nu) \in \Theta : F(\theta, \nu) = 0 \right\},$$

where $\theta(\nu)$ is the parameter of interest. In these settings, we aim at calculating the efficiency bound for estimating θ, following van der Vaart (1998). We make the following assumptions ($\|.\|$ denotes any norm of an Euclidean space).

- **Assumption 1:** Θ is a compact subset of $\mathbb{R}^d$.

- **Assumption 2:** The map $F(.,\mu)$ is continuous on Θ and has a unique zero θ_0. Moreover, θ_0 lies in the interior of Θ.

- **Assumption 3:** For all $x \in \mathcal{X}$, the map $\theta \mapsto \Phi(\theta, x)$ is continuous on Θ and the map $x \mapsto \sup_{\theta \in \Theta} \|\Phi(\theta, x)\|$ is bounded by some function κ, integrable with respect to μ.

- **Assumption 4:** For all $x \in \mathcal{X}$, $\theta \mapsto \Phi(\theta, x)$ is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0. Moreover $\|\partial \Phi(\theta, x)/\partial \theta\|$ and $\|\partial^2 \Phi(\theta, x)/\partial \theta \partial \theta^t\|$ are continuous and bounded by an integrable function in this neighborhood ($\partial \Phi(\theta, .)/\partial \theta$ will be noted $\nabla \Phi(\theta, .)$ in the sequel).

- **Assumption 5:** The matrices $D := \int \nabla \Phi(\theta_0, x) d\mu(x) \in \mathbb{R}^{d \times k}$ and $V := \int \Phi(\theta_0, x) \Phi^t(\theta_0, x) d\mu(x) \in \mathbb{R}^{k \times k}$ are of full rank.

These assumptions are usual conditions for this problem, see for instance Qin and Lawless (1994). They ensure the unicity of the parameter $\theta(\nu)$ (which, we recall, is defined as the zero of $F(.,\nu)$) when ν is close enough to μ for the total variation topology, and then allow a proper definition of the parameter of interest in the neighborhood of μ.

We can now calculate the efficiency bound for estimating θ_0 in this model. For this, we need the following definitions.

Definition A model $\{\mu_t, t \geq 0\}$ with $\mu_0 = \mu$ is *differentiable in quadratic mean* at μ if there exists $g : \mathcal{X} \to \mathbb{R}$ such that $\int_{\mathcal{X}} g^2 \, d\mu < \infty$ and

$$\lim_{t \to 0} \int_{\mathcal{X}} \left[\frac{1}{t} \left(\sqrt{\frac{d\mu_t}{d\tau_t}} - \sqrt{\frac{d\mu}{d\tau_t}} \right) - \frac{1}{2} g \right]^2 d\tau_t = 0,$$

setting for all $t \geq 0$, $\tau_t = \mu_t + \mu$.

The function g is called the *score* of $\{\mu_t, t \geq 0\}$, it satisfies $\int g d\mu = 0$. In the next definition, for all function $T_n : \mathcal{X}^n \to \Theta$ of the observations, we denote by $\mathcal{L}(T_n | \nu)$ the law of $T_n(X_1, ..., X_n)$ assuming that $X_1, ..., X_n$ are independent with distribution ν.

Definition An estimator $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ of a parameter $\theta : \mathcal{M} \to \Theta$ is *locally Gaussian regular* if for all differentiable submodel $\{\mu_t, t \geq 0\} \subset \mathcal{M}$ with $\mu_0 = \mu$ and for all positive sequence $(t_n)_{n \in \mathbb{N}}$ such that $\sqrt{n} t_n$ is bounded, $\mathcal{L}(\sqrt{n}(\hat{\theta} - \theta(\mu_{t_n}))| \mu_{t_n})$ converges weakly towards a Gaussian distribution as $n \to \infty$.

In a given model, the efficiency bound for estimating a parameter θ_0 is to be understood as a lower bound for the asymptotic variance of locally Gaussian regular estimators of θ_0. An efficiency bound is calculated by considering Fisher Informations of differentiable submodels. We refer to Bickel et al. (1994) and van der Vaart (1998) for further details.

Theorem 2.1 (Theorem 3, Qin and Lawless (1994)) *Suppose that Assumptions 1 to 5 hold. The efficiency bound in this model for estimating θ_0 is*

$$B = \left[DV^{-1}D^t \right]^{-1}.$$

Once we have calculated the efficiency bound in our model, the objective is to build an estimator $\hat{\theta}$ of θ_0 for which the efficiency bound is reached, at least asymptotically in the sense that

$$\lim_{n \to \infty} n \operatorname{var}(\hat{\theta}) = B.$$

In some cases, there may not exist any locally Gaussian regular estimates achieving the bound, see for instance examples in Ritov and Bickel (1990). It may also exist estimators having an asymptotic variance smaller than the efficiency bound, in which case the required regularity conditions are not satisfied, as seen in Chapter 2 in Bickel et al. (1994). Such situations will not occur here, as we assume regularity conditions on the model under which GMM and GEL procedures yield regular estimates.

3   Estimation of the parameter

For this problem, we may adopt two natural, although seemingly different, procedures to estimate θ_0, following Chapter 3 in Bickel et al. (1994). Let $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ denote the empirical distribution, where δ stands for the Dirac measure.

- **Procedure 1:** Find a "smooth" extension $\overline{\theta}$ of θ over a larger set $\mathcal{P} \supseteq \mathcal{M}$ of probability measures containing the empirical distribution μ_n and define the estimator as $\hat{\theta} = \overline{\theta}(\mu_n)$.

- **Procedure 2:** Build an approximation $\hat{\mu}$ of μ lying in the model $\mathcal{M}$ and define the estimator as $\hat{\theta} = \theta(\hat{\mu})$.

In the literature, two main methods have been implemented for this problem, each one providing a good illustration of each procedure.

3.1 Generalized method of moments

The generalized method of moments (GMM) was introduced in Hansen (1982). The method consists in replacing in the moment constraint the true measure μ by its empirical approximation μ_n. Then, find the value of θ for which $F(\theta, \mu_n) = \frac{1}{n} \sum_{i=1}^{n} \Phi(\theta, X_i)$ is as close as possible to 0 according to a given euclidean norm of $\mathbb{R}^k$. Precisely, define for M a symmetric positive definite $k \times k$ matrix and $a \in \mathbb{R}^k$, $\|a\|_M^2 = a^t M a$, the GMM estimator $\hat{\theta}$ of θ_0 associated to the norm $\|.\|_M$ is given by

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \ \|F(\theta, \mu_n)\|_M.$$

In practice, the matrix M may have a dependency in n, in which case it is chosen to converge towards a symmetric positive definite matrix. However, replacing the matrix by its limit leads to the same first order asymptotic properties of the estimate, under regularity conditions, as pointed out in Newey and Smith (2004). Here, we will assume for simplicity that M is fixed, this being sufficient for our purposes.

The generalized method of moments is a good illustration of the first procedure, as the GMM estimator $\hat{\theta}$ can be seen as the image of the empirical distribution μ_n by the function

$$\overline{\theta}_M(\nu) = \underset{\theta \in \Theta}{\operatorname{argmin}} \ \|F(\theta, \nu)\|_M, \ \nu \in \mathcal{P},$$

where $\mathcal{P}$ is an extension of the original model $\mathcal{M}$, containing μ_n. For sake of generality, $\mathcal{P}$ is taken as the set of all probability measures ν for which $F(., \nu)$ can take finite values on Θ. Because Θ is compact, $\mathcal{P}$ does not depend on the scaling matrix M.

This procedure may seem inefficient at first. Indeed, extending the parameter over a larger model $\mathcal{P}$ implicitly increases the size of the model, and thus decreases the information available. To be able to provide an efficient estimation, the extension $\overline{\theta}_M$ must be "smooth" enough so that differentiable submodels in $\mathcal{P}$ carry at least as much information as the original model. Basically, we want the efficiency bound $\overline{B}_M$ for estimating $\overline{\theta}_M$ over $\mathcal{P}$ not to be

higher than the original bound B. Since it obviously can not be lower, the objective is to find an efficient extension, for which $\overline{B}_M = B$.

Theorem 3.1 *Suppose that Assumption 1 to 4 hold. The efficiency bound for estimating $\overline{\theta}_M$ in $\mathcal{P}$ is*

$$\overline{B}_M = [DMD^t]^{-1} [DMVMD^t] [DMD^t]^{-1} .$$

This result was originally shown in Chamberlain (1987), although we propose in the Appendix a different proof, based on modern tools on semiparametric efficiency theory.

As expected, the efficiency bound $\overline{B}_M$ in the extended model $\mathcal{P}$ is larger than in the original model $\mathcal{M}$ (see Lemma 5.1 in the Appendix). The asymptotic variance of the GMM estimator is precisely the lower bound $\overline{B}_M$, as shown in Hansen (1982), which proves the efficiency of the method. The theorem also covers the results of Hansen (1982) and Chamberlain (1987) on optimal GMM for $M = V^{-1}$, leading to an efficiency bound in the extended model that is equal to the original bound B of Theorem 2.1.

Note that the matrix V is generally unknown, since it depends on both μ and θ_0. In this case, it is replaced by a consistent estimate $\tilde{V}$, leading to the same asymptotic properties under regularity conditions. Here again, several approaches are possible.
In the two-step GMM procedure, the estimate $\tilde{V}$ is built using a preliminary estimator $\tilde{\theta}$ of θ_0 obtained by a GMM procedure with known scaling matrix (in general, the identity matrix). As a result, $\tilde{\theta}$ is not in general asymptotically efficient, however, it is $\sqrt{n}$-consistent and enables to construct a consistent estimate of V.
Another solution is to minimize simultaneously over Θ

$$\theta \mapsto F(\theta, \mu_n)^t \hat{V}^{-1}(\theta) F(\theta, \mu_n), \tag{3}$$

where $\hat{V}^{-1}(\theta)$ denotes here an arbitrary consistent estimate of $V^{-1}(\theta)$, for all $\theta \in \Theta$. The latter approach was introduced in Hansen et al. (1996) as the continuous updating estimation (CUE).

3.2 Generalized empirical likelihood

Generalized empirical likelihood (GEL) was first applied to this problem in Qin and Lawless (1994), generalizing an idea of Owen (1991). This method

is an application of the first procedure. An estimate $\hat{\mu}$ of μ is obtained as an entropic projection (in a general sense defined below) of the empirical measure μ_n onto the model $\mathcal{M}$. Hence, the measure $\hat{\mu}$ is the element of the model that minimizes a given f-divergence $\mathcal{D}_f(\mu_n, .)$ with respect to the empirical distribution. Let us recall some definitions.

Definition Let f be a strictly convex function with $f(1) = f'(1) = 0$, and let P, Q be two probability measures on $\mathcal{X}$. The f-divergence of Q with respect to P is defined as

$$\mathcal{D}_f(P, Q) = \int f\left(\frac{dQ}{dP}\right) dP \ \text{ if } Q \ll P, \ \mathcal{D}_f(P, Q) = \infty \text{ otherwise.}$$

A f-divergence measures the "closeness" between two probability measures. It is non negative and is null only if $P = Q$. This definition can be extended to sets of measures by noting for $\mathcal{S}$ a subset of $\mathcal{P}(\mathcal{X})$,

$$\mathcal{D}_f(P, \mathcal{S}) = \inf_{Q \in \mathcal{S}} \mathcal{D}_f(P, Q).$$

Definition We call entropic projection of ν on $\mathcal{S}$ associated to f, an element $\nu^* \in \mathcal{S}$ such that $\mathcal{D}_f(\nu, \mathcal{S}) = \mathcal{D}_f(\nu, \nu^*) < \infty$.

An entropic projection always exists as soon as $\mathcal{S}$ is closed for the total variation topology and $\mathcal{D}_f(\nu, \mathcal{S})$ is finite. Furthermore, it is unique if $\mathcal{S}$ is also convex (see Csiszár (1967)).

Setting for a fixed $\theta \in \Theta$, $\mathcal{M}_\theta := \{\nu \in \mathcal{P}(\mathcal{X}) : F(\theta, \nu) = 0\}$, the model can be written as $\mathcal{M} = \cup_{\theta \in \Theta} \mathcal{M}_\theta$. Thus, the GEL estimator $\hat{\theta} = \theta(\hat{\mu})$ follows by

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{D}_f(\mu_n, \mathcal{M}_\theta).$$

Since $\mathcal{M}_\theta$ is closed and convex, the entropy $\mathcal{D}_f(\mu_n, \mathcal{M}_\theta)$ is reached for a unique measure $\hat{\mu}(\theta)$ in $\mathcal{M}_\theta$, provided that $\mathcal{D}_f(\mu_n, \mathcal{M}_\theta)$ is finite. Then, it appears that computing the GEL estimator involves a two-step procedure. First, build for each $\theta \in \Theta$, the entropic projection $\hat{\mu}(\theta)$ of μ_n onto $\mathcal{M}_\theta$. Then, minimize $\mathcal{D}_f(\mu_n, \hat{\mu}(\theta))$ with respect to θ. Since $\hat{\mu}(\theta)$ is absolutely continuous w.r.t. μ_n by construction, minimizing $\mathcal{D}_f(\mu_n, .)$ reduces to find the proper weights $p_1, ..., p_n$ to allocate to the observations $X_1, ..., X_n$. This turns into a finite dimensional problem, which can be solved by classical convex optimization tools (see for instance Kitamura (2006)). Finally, the GEL estimator $\hat{\theta}$ can be expressed as the solution to the saddle point problem

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \sup_{(\lambda_1, \lambda_2) \in \mathbb{R}^{k+1}} \lambda_1 - \frac{1}{n} \sum_{i=1}^{n} f^*(\lambda_1 + \lambda_2^t \Phi(\theta, X_i)),$$

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

Note that if the choice of the f-divergence plays a key role in the construction of the estimator, it has no influence on its asymptotic efficiency. Indeed, Qin and Lawless (1994) show that all GEL estimators are asymptotically efficient, regardless of the f-divergence used for their computation. Nevertheless, many situations justify the use of specific f-divergences. In its original form, empirical likelihood (EL) estimator in Owen (1991) uses the Kullback entropy $K(.,.)$ as f-divergence, pointing out that minimizing $K(\mu_n,.)$ reduces to maximizing likelihood among multinomial distributions. Newey and Smith (2004) remark that a quadratic f-divergence leads to the CUE estimator of Hansen et al. (1996). Many choices of f-divergence can also be given a Bayesian interpretation, using the maximum entropy on the mean (MEM) approach, as shown in Gamboa and Gassiat (1997).

4 Dealing with an approximate constraint

In many actual applications, only an approximation of the constraint function is available. This may occur if the moment conditions take complicated forms that can only be evaluated numerically or by simulations. Mcfadden (1989) suggested a method dealing with approximate constraint in a similar situation, introducing the method of simulated moments (see also Carrasco and Florens (2000)). In Loubes and Pelletier (2008) and Loubes and Rochet (2009), the authors study a MEM procedure for linear inverse problems with approximate constraints. Here, we propose to extend the GMM framework to a situation with approximate moment conditions. We assume that we observe a sequence $(\Phi_m(\theta,.))_{m\in\mathbb{N}}$ of approximate constraints, independent with the original sample $X_1, ..., X_n$. We are interested in exhibiting sufficient conditions on the sequence $(\Phi_m(\theta,.))_m$ under which estimating θ_0 with GMM procedures remains efficient when the constraint is replaced by its approximation. We discuss the asymptotic properties of the resulting estimates in a framework where both index n and m simultaneously grow to infinity.

In the sequel, we note $W(\theta)$ the inverse of the covariance matrix of $\Phi(\theta, X)$,

$$W(\theta) = \left[\int \Phi(\theta,.)\Phi^t(\theta,.)d\mu - \int \Phi(\theta,.)d\mu \int \Phi^t(\theta,.)d\mu \right]^{-1}, \theta \in \Theta,$$

while $\hat{W}(\theta)$ denotes an arbitrary consistent estimator of $W(\theta)$, built from the observations and the constraint function $\Phi(\theta,.)$. In the same way, $W_m(\theta)$ and

$\hat{W}_m(\theta)$ are defined by replacing Φ by its approximation Φ_m in the expressions of $W(\theta)$ and $\hat{W}(\theta)$ respectively.

For E, an Euclidean space endowed with a norm $\|.\|$, a function $f : \Theta \to E$ and $\mathcal{S} \subseteq \Theta$, note

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

We make the following assumptions, where we recall that $\mathcal{N}$ is a neighborhood of θ_0 defined in Assumption 4.

- **Assumption 6:** $\|\Phi(., x)\|_{\Theta}$, $\|\nabla \Phi(., x)\|_{\mathcal{N}}$ and $\|\partial^2 \Phi(., x)/\partial\theta\partial\theta^t\|_{\mathcal{N}}$ are dominated by a function $\kappa(x)$ such that $\int \kappa^{12}(x)d\mu(x) < \infty$.

- **Assumption 7:** For $(\varphi_m)_{m \in \mathbb{N}}$ a given sequence tending to infinity, the functions $\varphi_m \|\Phi_m(., x) - \Phi(., x)\|_{\Theta}$ and $\varphi_m \|\nabla\Phi_m(., x) - \nabla\Phi(., x)\|_{\mathcal{N}}$ are dominated by a function $\kappa_m(x)$ such that $\sup_m \int \kappa_m^{12}(x)d\mu(x) < \infty$.

- **Assumption 8:** The random map $\theta \mapsto \hat{W}(\theta)$ is differentiable on $\mathcal{N}$ and $\mathbb{E}(\sqrt{n}\|\hat{W} - W\|_{\Theta})^6$, $\mathbb{E}(\sqrt{n}\|\nabla\hat{W} - \nabla W\|_{\mathcal{N}})^6$ and $\mathbb{E}(\varphi_m\|\hat{W}_m - \hat{W}\|_{\Theta})^3$ are bounded as m, n range over $\mathbb{N}$.

Approximate GMM estimation consists in minimizing over Θ

$$\theta \mapsto \hat{\xi}_m(\theta) = \left[\int \Phi_m^t(\theta, .)d\mu_n\right] \tilde{W}_m \left[\int \Phi_m(\theta, .)d\mu_n\right],$$

where $\tilde{W}_m$ is a random matrix with properties to be specified below. It appears that the accuracy of approximate GMM relies on how close the approximate contrast function $\hat{\xi}_m$ is to its true value (i.e. when the constraint function is known). In this purpose, the scaling matrix $\tilde{W}_m$ should be chosen as close as possible to the optimal choice $W_0 = W(\theta_0)$.

As in the situation where the constraint function is known, the two-step GMM procedure provides a natural way to compute the scaling matrix $\tilde{W}_m$. First build a preliminary estimator $\tilde{\theta}_m$, minimizing over Θ

$$\theta \mapsto \tilde{\xi}_m(\theta) = \left[\int \Phi_m^t(\theta, .)d\mu_n\right] \left[\int \Phi_m(\theta, .)d\mu_n\right],$$

which corresponds to a GMM procedure with identity scaling matrix. Then, define $\tilde{W}_m = \hat{W}_m(\tilde{\theta}_m)$ which is used as scaling matrix in the contrast function $\hat{\xi}_m$. The resulting approximate two-step GMM estimator satisfies good asymptotic properties as soon as the approximate function Φ_m converges fast enough towards Φ, as proved in the following theorem.

Theorem 4.1 (Robustness of two-step GMM) *Denote by $\hat{\theta}_m$ and $\hat{\theta}$ the two-step GMM estimators obtained respectively with the constraint functions Φ_m and Φ. If Assumptions 1 to 8 hold,*

$$n\mathbb{E}(\|\hat{\theta}_m - \hat{\theta}\|)^2 = O(n\varphi_m^{-2}) + o(1).$$

In particular, $\hat{\theta}_m$ is $\sqrt{n}$-consistent and asymptotically efficient if n/φ_m^2 tends to zero.

In the same way, the CUE procedure can be adapted to the case with approximate constraint. Although, the robustness of CUE with approximate constraint requires slightly stronger assumptions.

- **Assumption 9:** $\hat{W}(.)$ and $W(.)$ are twice continuously differentiable on $\mathcal{N}$ and $\forall \eta > 0, \mathbb{P}(\|d^2\hat{W}/d\theta d\theta^t - d^2 W/d\theta d\theta^t\|_{\mathcal{N}} > \eta) = o(n^{-1})$. Besides, $\hat{W}_m(.)$ is differentiable on $\mathcal{N}$ and $\mathbb{E}(\varphi_m\|\nabla\hat{W}_m - \nabla\hat{W}\|_{\mathcal{N}})^3$ is bounded as m, n range over $\mathbb{N}$.

Applying the procedure to the approximate constraint, the approximate CUE estimator follows by minimizing over Θ

$$\theta \mapsto \hat{\zeta}_m(\theta) = \left[\int \Phi_m^t(\theta,.)d\mu_n \right] \hat{W}_m(\theta) \left[\int \Phi_m(\theta,.)d\mu_n \right].$$

Corollary 4.2 (Robustness of CUE) *Denote by $\hat{\theta}_m$ and $\hat{\theta}$ the CUE estimators obtained respectively with the constraint functions Φ_m and Φ. If Assumptions 1 to 9 hold,*

$$n\mathbb{E}(\|\hat{\theta}_m - \hat{\theta}\|)^2 = O(n\varphi_m^{-2}) + o(1).$$

In particular, $\hat{\theta}_m$ is $\sqrt{n}$-consistent and asymptotically efficient if n/φ_m^2 tends to zero.

5 Appendix

5.1 Technical lemmas

Lemma 5.1 *For all symmetric positive-definite matrix M,*

$$DMD^t \left[DMVMD^t \right]^{-1} DMD^t \leq DV^{-1}D^t,$$

with equality for $M = V^{-1}$.

Proof. Set $A = V^{1/2}MD^t$, $A[A^tA]^{-1}A^t$ is an orthogonal projection matrix with in particular $A[A^tA]^{-1}A^t \leq Id$. The inequality holds after multiplying each term by $DV^{-1/2}$ on the left and $V^{-1/2}D^t$ on the right, proving the result.

Lemma 5.2 *Let $f : \Theta \to \mathbb{R}$ be a continuous positive function with a unique zero θ_0 lying in the interior of the compact set Θ and with positive definite Hessian matrix at θ_0. Assume that f is twice continuously differentiable on a neighborhood $\mathcal{N}$ of θ_0. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of positive random functions, twice continuously differentiable on $\mathcal{N}$, converging in probability towards f. Note $\mathcal{H} = \partial^2 f/\partial\theta\partial\theta^t$ and $\mathcal{H}_n = \partial^2 f_n/\partial\theta\partial\theta^t$. Moreover, for all $n \in \mathbb{N}$, let $(f_{m,n})_{m\in\mathbb{N}}$ be a sequence of positive random functions converging towards f_n as $m \to \infty$. Denote by $\theta_{m,n}$ and θ_n a minimizer of $f_{m,n}$ and f_n respectively. If the following conditions are met*

i) $\forall \eta > 0$, $\mathbb{P}(\|f_n - f\|_\Theta > \eta) = o(n^{-1})$ and $\mathbb{P}(\|\mathcal{H}_n - \mathcal{H}\|_\mathcal{N} > \eta) = o(n^{-1})$,

ii) the $f_{m,n}$ are differentiable on $\mathcal{N}$ and $C_1 = \sup_{m,n} \mathbb{E}(\varphi_m\|f_{m,n} - f_n\|_\Theta)^p$ and $C_2 = \sup_{m,n} \mathbb{E}(\varphi_m\|\nabla f_{m,n} - \nabla f_n\|_\mathcal{N})^p$ are finite for a $p > 0$ and a sequence $(\varphi_m)_{m\in\mathbb{N}}$ tending to infinity,

then, there is a constant $K > 0$ such that

$$\mathbb{E}\|\theta_{m,n} - \theta_n\|^p \leq K\varphi_m^{-p} + o(n^{-1}).$$

Proof. By continuity of $\mathcal{H}$ around θ_0, we may assume without loss of generality that $\mathcal{N}$ is such that $\mathcal{H}(\theta)$ has all its eigenvalues larger than some constant $2c > 0$ for all $\theta \in \mathcal{N}$. Note ρ_n the smallest eigenvalue of $\mathcal{H}_n(\theta)$ as θ ranges over $\mathcal{N}$. The uniform convergence of $\mathcal{H}_n$ on $\mathcal{N}$ in condition *i)* ensures that $\mathbb{P}(\rho_n < c) = o(n^{-1})$. Besides, since θ_0 is the unique zero of f on the compact set Θ, we can find a constant $\eta_1 > 0$ such that θ_n lies in $\mathcal{N}$ as soon as $\|f_n - f\|_\Theta \leq \eta_1$. Hence, still by condition *i)*, $\mathbb{P}(\theta_n \notin \mathcal{N}) = o(n^{-1})$. In the same way, there is a constant $\eta_2 > 0$ such that $\mathbb{P}(\theta_{m,n} \notin \mathcal{N}) \leq \mathbb{P}(\|f_{m,n} - f\|_\Theta > 2\eta_2)$, with

$$\begin{aligned}
\mathbb{P}(\|f_{m,n} - f\|_\Theta > 2\eta_2) &\leq \mathbb{P}(\|f_{m,n} - f_n\|_\Theta + \|f_n - f\|_\Theta > 2\eta_2) \\
&\leq \mathbb{P}(\|f_{m,n} - f_n\|_\Theta > \eta_2) + \mathbb{P}(\|f_n - f\|_\Theta > \eta_2) \\
&\leq C_1(\varphi_m\eta_2)^{-p} + o(n^{-1}),
\end{aligned}$$

by Chebyshev's inequality. Call Ω the intersection of the three events $\{\theta_n \in \mathcal{N}\}$, $\{\theta_{m,n} \in \mathcal{N}\}$ and $\{\rho_n \geq c\}$, we get $\mathbb{P}(\Omega^c) \leq C_1(\varphi_m\eta_2)^{-p} + o(n^{-1})$, where Ω^c denotes the complementary of Ω. Moreover, on Ω, we have

$$\|\nabla f_{m,n} - \nabla f_n\|_\mathcal{N} \geq \|\nabla f_n(\theta_{m,n})\| \geq c\|\theta_{m,n} - \theta_n\|.$$

11

Let δ be the diameter of Θ, it follows that

$$
\begin{aligned}
\mathbb{E}\|\theta_{m,n} - \theta_n\|^p &\leq c^{-p}\,\mathbb{E}\|\nabla f_{m,n} - \nabla f_n\|_{\mathcal{N}}^p + \delta^p\,\mathbb{P}(\Omega^c) \\
&\leq K\varphi_m^{-p} + o(n^{-1}),
\end{aligned}
$$

for $K = C_1\delta^p/\eta_2^p + C_2/c^p$.

5.2   Proofs

Proof of Theorem 3.1: Note $\mathcal{T}$ the set of bounded functions with zero mean under μ. For any $g \in \mathcal{T}$ and $t > 0$, the measure $\mu_t := (1 + tg)\mu$ lies in $\mathcal{P}$ provided that t is small enough. The path $\{\mu_t, t \geq 0\}$ is thus differentiable with score g.

The uniform convergence of $F(., \mu_t)$ towards $F(., \mu)$ (which follows from Assumptions 1 and 2) ensures the existence of a minimizer $\theta(t)$ of $F(., \mu_t)$ continuously close to θ_0 as $t \to 0$ and satisfying the first order condition $\gamma_M(\theta(t), \mu_t) = 0$ where

$$
\gamma_M(\theta, \nu) = \left[\int (\nabla\Phi(\theta, .))d\nu\right] M \left[\int \Phi(\theta, .)d\nu\right], (\theta, \nu) \in \Theta \times \mathcal{P}.
$$

Under Assumptions 2 to 4, the implicit functions theorem applied to the map $(\theta, t) \mapsto \gamma_M(\theta, \mu_t)$ in a neighborhood of $(\theta_0, 0)$ warrants the unicity of the minimum $\theta(t) = \overline{\theta}_M(\mu_t)$.

Note $\dot{l} = (\dot{l}_1, ..., \dot{l}_d)^t$ the efficient influence function of $\overline{\theta}_M$. By a Taylor expansion of Φ_θ at θ_0 and using that $\gamma_M(\overline{\theta}_M(\mu_t), \mu_t) = 0$, we get

$$
\left[\int \nabla\Phi_{\theta_0}d\mu_t\right] M \left[\left[\int \Phi_{\theta_0}(1 + tg)d\mu\right] + \left[\int \nabla\Phi_{\theta_0}^t d\mu_t\right](\overline{\theta}_M(\mu_t) - \theta_0)\right] = o(t).
$$

Since $\overline{\theta}_M(\mu_t) - \theta_0 = t\int \dot{l}gd\mu + o(t)$ by definition of $\dot{l}$, we obtain after dividing each term by t and making t tend to zero

$$
DM\left[\int \Phi_{\theta_0}gd\mu\right] = -DMD^t(\int \dot{l}gd\mu).
$$

Since this holds for all $g \in \mathcal{T}$, we conclude that

$$
\dot{l}(.) = -\left[DMD^t\right]^{-1}DM\Phi_{\theta_0}(.),
$$

checking beforehand that $\dot{l}$ lies in the closure of $\mathcal{T}$. The efficiency bound is the variance of $\dot{l}(X)$ which proves the result.

Proof of Theorem 4.1: For all $\theta \in \Theta$, let

$$\alpha(\theta) = \int \Phi(\theta, .)d\mu, \ \beta(\theta) = \int \nabla\Phi(\theta, .)d\mu, \ \gamma(\theta) = \int \frac{\partial^2\Phi(\theta, .)}{\partial\theta\partial\theta^t}d\mu.$$

Besides, note $\hat{\alpha}(\theta)$ the empirical estimate of $\alpha(\theta)$ and $\hat{\alpha}_m(\theta)$ the estimate built with Φ_m and define $\hat{\beta}(\theta)$, $\hat{\gamma}(\theta)$, $\hat{\beta}_m(\theta)$ and $\hat{\gamma}_m(\theta)$ analogously.

First, prove that $\mathbb{E}(\|\tilde{\theta}_m - \tilde{\theta}\|^6) = O(\varphi_m^{-6}) + o(n^{-1})$. It suffices to verify the conditions of Lemma 5.2 for $p = 6$, taking $f_n = \tilde{\xi} = \hat{\alpha}\hat{\alpha}^t$, $f_{m,n} = \tilde{\xi}_m = \hat{\alpha}_m\hat{\alpha}_m^t$ and $f = \alpha\alpha^t$. In this particular case, we have $\mathcal{H}_n = \partial^2\tilde{\xi}/\partial\theta\partial\theta^t = 2\hat{\beta}\hat{\beta}^t + 2\hat{\gamma}\hat{\alpha}$ and $\mathcal{H} = 2\beta\beta^t + 2\gamma\alpha$.

First note that $\mathcal{H}(\theta_0) = 2\beta^t(\theta_0)\beta(\theta_0)$ is positive definite by Assumption 5. Furthermore, $\hat{\alpha}(\theta)$ is asymptotically normal and since $\|\Phi(\theta, .)\|$ is dominated by a square integrable function κ on Θ, we have, for all $\eta > 0$,

$$\mathbb{P}(\|\hat{\alpha} - \alpha\|_\Theta \geq \eta) = o(n^{-1}).$$

By assumption, the same argument holds for $\|\hat{\beta} - \beta\|_\mathcal{N}$ and $\|\hat{\gamma} - \gamma\|_\mathcal{N}$. Condition $i)$ in Lemma 5.2 follows directly, noticing that

$$\mathcal{H}_n - \mathcal{H} = 2(\hat{\beta} - \beta)(\hat{\beta} + \beta)^t + 2(\hat{\gamma} - \gamma)\hat{\alpha} + 2\gamma(\hat{\alpha} - \alpha).$$

Moreover, $\|\tilde{\xi}_m - \tilde{\xi}\|_\Theta = \|\hat{\alpha}_m^t\hat{\alpha}_m - \hat{\alpha}^t\hat{\alpha}\|_\Theta \leq \|\hat{\alpha}_m + \hat{\alpha}\|_\Theta\|\hat{\alpha}_m - \hat{\alpha}\|_\Theta$, yielding

$$\mathbb{E}(\varphi_m\|\tilde{\xi}_m - \tilde{\xi}\|_\Theta)^6 \leq \left[\mathbb{E}(\varphi_m\|\hat{\alpha}_m - \hat{\alpha}\|_\Theta)^{12}\right]^{\frac{1}{2}} \left[\mathbb{E}(\|\hat{\alpha}_m + \hat{\alpha}\|_\Theta)^{12}\right]^{\frac{1}{2}}$$

by Cauchy-Schwarz inequality. Thus, $\mathbb{E}(\varphi_m\|\tilde{\xi}_m - \tilde{\xi}\|_\Theta)^6$ is finite by Assumptions 6 and 7. Since $\nabla\tilde{\xi}_m = 2\hat{\beta}_m\hat{\alpha}_m$ and $\nabla\tilde{\xi} = 2\hat{\beta}\hat{\alpha}$, assumptions also warrant that $\mathbb{E}(\varphi_m\|\nabla\tilde{\xi}_m - \nabla\tilde{\xi}\|_\mathcal{N})^6 < \infty$. Lemma 5.2 then gives

$$\|\tilde{\theta}_m - \tilde{\theta}\|^6 = O(\varphi_m^{-6}) + o(n^{-1}).$$

To show the result, we shall now verify that the conditions of Lemma 5.2 hold for $p = 2$ with the functions $f_{m,n} = \hat{\xi}_m$, $f_n = \hat{\xi}$, $f = \xi$. We now consider $\mathcal{H}_n = 2\hat{\beta}\tilde{W}\hat{\beta}^t + 2\hat{\gamma}\tilde{W}\hat{\alpha}$ and $\mathcal{H} = 2\beta W_0\beta^t + 2\gamma W_0\alpha$ where $\tilde{W} = \hat{W}(\tilde{\theta})$ and $W_0 = W(\theta_0)$.

The Hessian matrix $\mathcal{H}(\theta_0) = 2\beta(\theta_0)W_0\beta^t(\theta_0)$ is positive definite by Assumption 5. For condition $i)$ of Lemma 5.2 to be satisfied, we need that for all $\eta > 0$, $\mathbb{P}(\|\tilde{W} - W_0\| > \eta) = o(n^{-1})$. Since $\mathbb{P}(\tilde{\theta} \notin \mathcal{N}) = o(n^{-1})$, we shall only consider the case where $\tilde{\theta} \in \mathcal{N}$. By the triangular inequality, we get $\|\hat{W}(\tilde{\theta}) - W_0\| \leq \|\hat{W}(\tilde{\theta}) - \hat{W}(\theta_0)\| + \|\hat{W}(\theta_0) - W_0\|$ and we use that

$$\mathbb{P}(\|\tilde{W} - W_0\| > \eta) \leq \mathbb{P}(\|\hat{W}(\tilde{\theta}) - \hat{W}(\theta_0)\| > \frac{\eta}{2}) + \mathbb{P}(\|\hat{W}(\theta_0) - W_0\| > \frac{\eta}{2}).$$

Assumption 8 gives $\mathbb{P}(\|\hat{W}(\theta_0) - W_0\| > \eta/2) = o(n^{-1})$, using Chebyshev's inequality. Furthermore, $\|\hat{W}(\tilde{\theta}) - \hat{W}(\theta_0)\| \leq \|\nabla\hat{W}\|_{\mathcal{N}}\|\tilde{\theta} - \theta_0\|$ for a suitable norm in $\mathbb{R}^{d\times k\times k}$ and for $K > \mathbb{E}\|\nabla W\|_{\mathcal{N}}$,

$$\mathbb{P}(\|\nabla\hat{W}\|_{\mathcal{N}}\|\tilde{\theta} - \theta_0\| > \frac{\eta}{2}) \leq \mathbb{P}(\|\tilde{\theta} - \theta_0\| > \frac{\eta}{2K}) + \mathbb{P}(\|\nabla\hat{W}\|_{\mathcal{N}} > K) = o(n^{-1})$$

which ensures condition $i)$ of Lemma 5.2. Write

$$\hat{\xi}_m - \hat{\xi} = (\hat{\alpha}_m - \hat{\alpha})^t \tilde{W}_m \hat{\alpha}_m + \hat{\alpha}_m^t (\tilde{W}_m - \tilde{W})\hat{\alpha} + (\hat{\alpha}_m - \hat{\alpha})^t \tilde{W}\hat{\alpha}$$

where each term can be controlled using Hölder's inequality, as we have for the middle term

$$
\begin{aligned}
\mathbb{E}(\varphi_m\|\hat{\alpha}_m^t(\tilde{W}_m - \tilde{W})\hat{\alpha}\|_\Theta)^2 &\leq \mathbb{E}(\|\hat{\alpha}_m\|_\Theta \varphi_m\|\tilde{W}_m - \tilde{W}\|\|\hat{\alpha}\|_\Theta)^2 \\
&\leq \left[\mathbb{E}(\|\hat{\alpha}_m\|_\Theta\|\hat{\alpha}\|_\Theta)^6\right]^{\frac{1}{3}}\left[\mathbb{E}(\varphi_m\|\tilde{W}_m - \tilde{W}\|)^3\right]^{\frac{2}{3}},
\end{aligned}
$$

for an appropriate norm in $\mathbb{R}^{k\times k}$ for the matrix $\tilde{W}_m - \tilde{W}$. Apply the same procedure for the two other terms, with for instance

$$
\begin{aligned}
\mathbb{E}(\varphi_m\|(\hat{\alpha}_m - \hat{\alpha})^t\tilde{W}_m\hat{\alpha}_m\|_\Theta)^2 &\leq \mathbb{E}(\varphi_m\|\hat{\alpha}_m - \hat{\alpha}\|_\Theta\|\tilde{W}_m\|\|\hat{\alpha}\|_\Theta)^2 \\
&\leq \left[\mathbb{E}(\varphi_m\|\hat{\alpha}_m - \hat{\alpha}\|_\Theta\|\hat{\alpha}\|_\Theta)^6\right]^{\frac{1}{3}}\left[\mathbb{E}\|\tilde{W}_m\|^3\right]^{\frac{2}{3}}.
\end{aligned}
$$

To have $\sup_{m,n}\mathbb{E}(\varphi_m\|\hat{\xi}_m - \hat{\xi}\|_\Theta)^2 < \infty$, it suffices to show $\mathbb{E}(\varphi_m\|\tilde{W}_m - \tilde{W}\|)^3$ is bounded as n and m range over $\mathbb{N}$, since the rest follows from the first part of the proof. This is true as soon as $\tilde{\theta}_m$ and $\tilde{\theta}$ both lie in $\mathcal{N}$ as we have on the event $\Omega = \{\tilde{\theta}, \tilde{\theta}_m \in \mathcal{N}\}$,

$$
\begin{aligned}
\|\tilde{W}_m - \tilde{W}\| &\leq \|\hat{W}_m(\tilde{\theta}_m) - \hat{W}(\tilde{\theta}_m)\| + \|\hat{W}(\tilde{\theta}_m) - \hat{W}(\tilde{\theta})\| \\
&\leq \|\hat{W}_m - \hat{W}\|_\Theta + \|\nabla\hat{W}\|_{\mathcal{N}}\|\tilde{\theta}_m - \tilde{\theta}\|
\end{aligned}
$$

and the result follows from Assumption 8 and by Cauchy-Scharz inequality, since both $\varphi_m\|\tilde{\theta}_m - \tilde{\theta}\|$ and $\|\nabla\hat{W}\|_{\mathcal{N}}$ have finite moments of order 6. Hence,

$$\sup_{n,m\in\mathbb{N}}\mathbb{E}(\varphi_m\|\hat{\xi}_m - \hat{\xi}\|_\Theta\mathbb{1}_\Omega)^2 < \infty.$$

The same reasoning leads to the same conclusion for $\nabla\hat{\xi}_m$ on $\mathcal{N}$, namely

$$\sup_{n,m\in\mathbb{N}}\mathbb{E}(\varphi_m\|\nabla\hat{\xi}_m - \nabla\hat{\xi}\|_{\mathcal{N}}\mathbb{1}_\Omega)^2 < \infty.$$

Following the proof of Lemma 5.2, we show that the complementary of Ω occurs with negligible probability as $\mathbb{P}(\Omega^c) = O(\varphi_m^{-6}) + o(n^{-1})$. Since $\|\hat{\tilde{\theta}}_m - \hat{\theta}\|$

remains bounded on Ω^c, we conclude that $\mathbb{E}(\|\hat{\theta}_m - \hat{\theta}\| \mathbb{1}_{\Omega^c})^2 = o(\varphi_m^{-2}) + o(n^{-1})$, yielding

$$\mathbb{E}(\|\hat{\theta}_m - \hat{\theta}\|)^2 = O(\varphi_m^{-2}) + o(n^{-1}).$$

Proof of Corollary 4.2: The proof is the same as for Theorem 4.1, we show that the conditions of Lemma 5.2 are satisfied for $f_{m,n} = \hat{\zeta}_m = \hat{\alpha}_m^t \hat{W} \hat{\alpha}_m$, $f_n = \hat{\zeta} = \hat{\alpha}^t \hat{W} \hat{\alpha}$ and $f = \alpha W \alpha$. Condition $i)$ follows from Assumptions 6 and 9, and $\mathbb{E}(\varphi_m \|\nabla \hat{\zeta}_m - \nabla \hat{\zeta}\|_{\mathcal{N}})^2$ can be bounded as in the proof of the theorem, using the additional condition that $\mathbb{E}(\varphi_m \|\nabla \hat{W}_m - \nabla \hat{W}\|_{\mathcal{N}})^3$ is bounded.

References

[1] Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD.

[2] J. M. Borwein and A. S. Lewis (1991). Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, 29(2):325–338.

[3] Marine Carrasco and Jean-Pierre Florens (2000). Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 16(6):797–834.

[4] Marine Carrasco and Jean-Pierre Florens (2002). Simulation-based method of moments and efficiency. *Journal of Business & Economic Statistics*, 20(4):482–492. Twentieth anniversary GMM issue.

[5] Gary Chamberlain (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

[6] I. Csiszár (1967). On topology properties of f-divergences. *Studia Scientiarum Mathematicarum Hungarica. A Quarterly of the Hungarian Academy of Sciences*, 2:329–339.

[7] Stephen G. Donald, Guido W. Imbens, and Whitney K. Newey (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics*, 152(1):28–36.

[8] F. Gamboa and E. Gassiat (1997). Bayesian methods and maximum entropy for ill-posed inverse problems. *The Annals of Statistics*, 25(1):328–350.

[9] Lars Peter Hansen, John Heaton, and Amir Yaron (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*.

[10] Lars Peter Hansen and Kenneth J. Singleton (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica. Journal of the Econometric Society*, 50(5):1269–1286.

[11] Lars Peter Hansen (1982). Large sample properties of generalized method of moments estimators. *Econometrica. Journal of the Econometric Society*, 50(4):1029–1054.

[12] Yuichi Kitamura (2006). Empirical likelihood methods in econometrics: Theory and practice. *Cowles foundation discussion paper no. 1569.*

[13] J. M. Loubes and B. Pelletier (2008). Maximum entropy solution to ill-posed inverse problems with approximately known operator. *Journal of Mathematical Analysis and Applications*, 344(1):260–273.

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

[15] Daniel McFadden (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica. Journal of the Econometric Society*, 57(5):995–1026.

[16] Whitney K. Newey and Richard J. Smith (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica. Journal of the Econometric Society*, 72(1):219–255.

[17] Art Owen (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 19(4):1725–1747.

[18] Jing Qin and Jerry Lawless (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325.

[19] Y. Ritov and P. J. Bickel (1990). Achieving information bounds in non and semiparametric models. *The Annals of Statistics*, 18(2):925–938.

[20] Richard J Smith (1997). Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *Economic Journal.*

[21] A. W. van der Vaart (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.