ESTIMATION OF SEMIPARAMETRIC MODELS WHEN THE CRITERION FUNCTION IS NOT SMOOTH*

by

Xiaohong Chen
New York University

Oliver Linton
London School of Economics and Political Science

Ingrid Van Keilegom
Université Catholique de Louvain

Contents:

The Suntory Centre
Suntory and Toyota International Centres for
Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 020 7955 6698

Discussion Paper
No.EM/03/450
May 2003

* We would like to thank co-editor Joel Horowitz, two anonymous referees and Whitney Newey for their valuable suggestions that much improved the paper. We also received helpful comments from R Blundell, C Huse, H Hong, B Honore, H Ichimura, C Meghir, R Sherman and E Tamer. The research was completed while Chen was visiting Princeton University from September 2001 to July 2002. Chen is grateful for the great hospitality provided by the Gregory Chow Econometrics Research Program at Princeton University. Chen and Linton would also like to acknowledge financial support from the ESRC. The research of Van Keilegom was supported by the contract 'Projet d'Actions de Recherce Concertées' Nr. 98/03-217, and by the IAP Research Network Nr. P5/24 of the Belgian State.

Abstract

We provide easy to verify sufficient conditions for the consistency and asymptotic normality of a class of semiparametric optimization estimators where the criterion function does not obey standard smoothness conditions and simultaneously depends on some nonparametric estimators that can themselves depend on the parameters to be estimated. Our results extend existing theories like those of Pakes and Pollard (1989), Andrews (1994a) and Newey (1994). We also show that bootstrap provides asymptotically correct confidence regions for the finite dimensional parameters. We apply our results to two examples: a 'hit rate' and a partially linear median regression with some endogenous regressors.

Keywords: Empirical processes; non-smooth criterion; semiparametric estimation; stochastic equicontinuity.

JEL Nos.: C13, C14.

Contact addresses:
Xiaohong Chen, Associate Professor, Department of Economics, New York University, 269 Mercer Street, New York, NY 10003, USA. Email: xiaohong.chen@nyu.edu

Professor Oliver Linton, Department of Economics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. Email: o.linton@lse.ac.uk

Ingrid Van Keilegom, Assistant Professor, Institute of Statistics, Université Catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium. Email: vankeilegom@stat.ucl.ac.be

1 Introduction

In this note we investigate a class of semiparametric estimation problems that involve non-smooth criterion functions that contain both finite dimensional and infinite dimensional unknown parameters. Powell (1994) and Manski (1994) discuss specification of econometric models through quantile, symmetry, mode, and independence restrictions. The corresponding estimation procedures are often non-smooth in the parameters of interest. In practice one may also want to have flexibility in the functional form of the part of the model of interest, an issue also discussed in Powell (1994). We study a number of examples that combine these two features below.

There have been many papers devoted to general theories of estimation, following Huber (1967). The existing theories allow for non-smooth objective functions of finite dimensional parameters (without infinite dimensional parameters) [e.g., Pakes and Pollard (1989) and Newey and McFadden (1994, Section 7)], or smooth objective functions of both finite and infinite dimensional parameters [e.g., Bickel, Klaassen, Ritov, and Wellner (1993), Andrews (1994a), Newey (1994), Newey and McFadden (1994, Section 8), Pakes and Olley (1995), Chen and Shen (1998) and Ai and Chen (2002)]. We are unaware of a general theory on non-smooth objective functions with both finite and infinite dimensional parameters, or rather the existing high level conditions for consistency and asymptotic normality have not been verified in this less regular setting.

A viable approach here is to use the criterion function that has been smoothed over. This then satisfies the usual regularity conditions and the standard distribution theory applies. Horowitz has applied this idea to a number of problems including standard median estimation [Horowitz (1998a)]; he gives some additional justification for this approach in terms of higher order properties. Is 'smoothing over' always the best estimation strategy? The issue here is analogous to whether one should use the smoothed empirical distribution function instead of the usual unsmoothed empirical distribution function. Although there are some statistical reasons for so doing, most applied economists would be content with using the unsmoothed empirical distribution.[1]

We provide sufficient conditions to ensure $\sqrt{n}-$asymptotic normality of the finite dimensional parameters obtained from a non-smooth criterion that depends on a preliminary infinite dimensional parameter estimate. Our results allow for the case where the nonparametric estimator is 'profiled' i.e., is allowed to depend on the parameters. Our approach and results extend those of Pakes and Pollard (1989), Andrews (1994a), Newey (1994) and Pakes and Olley (1995). We also show that the

[1] Horowitz (1992) originally applied this approach to the binary choice model of Manski (1975). He proposed a smoothed maximum score estimator, and showed that his estimator converges faster than the original unsmoothed maximum score and is asymptotically normal. Of course, this is a case where the semiparametric information bound is zero and the problem is correctly viewed as being nonparametric so that the smoothing idea fits in quite naturally. Our theory does not apply to semiparametric models with zero information bounds.

ordinary nonparametric bootstrap provides asymptotically correct confidence regions for the finite dimensional parameters. The theory we present here relies on certain key empirical process results regarding the stochastic equicontinuity properties of the non-smooth objective function, (especially with respect to preliminary nonparametric estimators). We provide these results below by extending the work of Andrews (1994b), and by applying the work of Van der Vaart and Wellner (1996).

Finally, to simplify the presentation we focus on the i.i.d. sample in this note, but the Theorems 1 and 2 and Corollary 1 below actually allow for any dependent, heterogeneous sample as well. Although Theorem B and Theorem 3 both require the i.i.d. structure and should be modified for time series models.

2 A General Class of Estimators

Throughout the paper we assume that the data $\{Z_i\}_{i=1}^n$ is randomly sampled from a distribution P whose support is $\mathbf{Z} \subset \mathbb{R}^{d_z}$. In many applications it is useful to denote a component of Z_i as X_i with $X_i \in \mathbb{R}^{d_x}$ and $1 \leq d_x \leq d_z$. We denote Θ for a finite dimensional parameter set (a compact subset of $\mathbb{R}^k$) and $\mathcal{H}$ for an infinite dimensional parameter set. Suppose there exists a non-random measurable vector-valued function $M : \mathbb{R}^k \times \mathcal{H} \to \mathbb{R}^l$, with $l \geq k$, such that $M(\theta, h_o(\cdot, \theta)) = 0 \quad \text{at } \theta = \theta_o \in \Theta \subset \mathbb{R}^k$. We denote $\theta_o \in \Theta$ and $h_o \in \mathcal{H}$ as the true unknown finite and infinite dimensional parameters. As in Newey (1994), Pakes and Olley (1995) and Ai and Chen (2002), we allow that the function $h_o \in \mathcal{H}$ can depend on the parameters θ and the data Z. We usually suppress the arguments of the function h_o for notational convenience; thus: $(\theta, h) \equiv (\theta, h(\cdot, \theta))$, $(\theta, h_o) \equiv (\theta, h_o(\cdot, \theta))$, and $(\theta_o, h_o) \equiv (\theta_o, h_o(\cdot, \theta_o))$. We assume that $\mathcal{H}$ is a vector space of functions endowed with a pseudo-metric $||\cdot||_{\mathcal{H}}$, which is a sup-norm metric with respect to the θ-argument and a pseudo-metric with respect to all the other arguments. For example when $\mathcal{H}$ is a class of continuous functions mapping from $\mathbf{Z} \times \Theta$ to $\mathbb{R}$ and having finite sup-norms, we can take $||h||_{\mathcal{H}} = \sup_\theta ||h(\cdot, \theta)||_\infty = \sup_\theta \sup_z |h(z, \theta)|$ or $||h||_{\mathcal{H}} = \sup_\theta ||h(\cdot, \theta)||_{L_r(P)} = \sup_\theta \{\int |h(Z, \theta)|^r dP\}^{1/r}$ for $1 \leq r < \infty$. Finally, we denote $||A|| = (tr(A'WA))^{1/2}$ for any matrix A, where for notational ease we suppress the dependence of the norm on the fixed symmetric positive definite matrix W.

Suppose there exists a random vector-valued function $M_n : \mathbb{R}^k \times \mathcal{H} \to \mathbb{R}^l$ depending on the data $\{Z_i : i = 1, \ldots, n\}$, such that $||M_n(\theta, h_o)||$ is close to $||M(\theta, h_o)||$. We allow that $M_n(\theta, h)$ could be non-smooth with respect to (θ, h) but will assume that $M(\theta, h)$ is smooth at (θ_o, h_o) in a sense to be defined later. Suppose that for each θ there is an initial nonparametric estimator $\widehat{h}(\cdot, \theta)$ for $h_o(\cdot, \theta)$. Hence the estimator of h_o can be profiled, see the partially linear median regression with some endogenous regressors example in section 5. Again for notational ease we let $(\theta, \widehat{h}) \equiv (\theta, \widehat{h}(\cdot, \theta))$.

We estimate θ_o by any $\widehat{\theta}$ that approximately solves the sample minimization problem:

$$\min_{\theta\in\Theta} ||M_n(\theta, \widehat{h})||. \tag{1}$$

There are many algorithms available for computing the optimum of general non-smooth functions, e.g., the Nelder-Mead, and the more recent genetic and evolutionary algorithms. Koenker (1997) gives a review of some methods targeted at quantile regression problems.

Our estimation strategy is an extension of the Generalized Method of Moments (GMM) that is popular in econometrics. In the current statistical parlance, we are treating essentially 'Z-estimators' except that we allow for over-identifying restrictions.[2] It is not necessary that $M(\theta, h) = E[M_n(\theta, h)]$ in the general limiting theorems we present below. Nevertheless many econometric applications correspond to this case: $M(\theta, h) = E[m(Z_i, \theta, h)]$ and $M_n(\theta, h) = n^{-1}\sum_{i=1}^n m(Z_i, \theta, h)$, where $m : \mathbb{R}^{d_z} \times \mathbb{R}^k \times \mathcal{H} \to \mathbb{R}^l$ is a measurable vector-valued function such that $E[m(Z_i, \theta, h_o)] = 0$ if and only if $\theta = \theta_o$. Hence the notations $M(\theta, h)$ and $M_n(\theta, h)$ implicitly correspond to population and sample moment conditions. Usually the function h enters m only through $h(Z_i, \theta)$, but there are some cases where $h(Z_1, \theta), \ldots, h(Z_n, \theta)$ enter m; and our Theorems 1 and 2 and Corollary 1 allow this case.

3 The Large Sample Theory

Our large sample theory is a direct extension of the well known theory of Pakes and Pollard (1989). In particular, when $M_n(\theta, h) = M_n(\theta, h_o)$ and $M(\theta, h) = M(\theta, h_o)$ for all h, our Theorem 1 becomes their Corollary 3.2, and our Theorem 2 becomes their Theorem 3.3.

3.1 Consistency

Theorem 1. *Suppose that $\theta_o \in \Theta$ satisfies $M(\theta_o, h_o) = 0$, and that:*
(1.1) $\|M_n(\widehat{\theta}, \widehat{h})\| \leq \inf_{\theta\in\Theta} \|M_n(\theta, \widehat{h})\| + o_p(1)$;
(1.2) *For all $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that $\inf_{\|\theta-\theta_o\|>\delta} \|M(\theta, h_o)\| \geq \epsilon(\delta) > 0$;*
(1.3) *Uniformly for all $\theta \in \Theta$, $M(\theta, h)$ is continuous [with respect to the metric $||\cdot||_{\mathcal{H}}$] in h at $h = h_o$;*
(1.4) $||\widehat{h} - h_o||_{\mathcal{H}} = o_p(1)$;
(1.5) *For all sequences of positive numbers $\{\delta_n\}$ with $\delta_n = o(1)$,*

$$\sup_{\theta\in\Theta, \|h-h_o\|_{\mathcal{H}}\leq\delta_n} \frac{\|M_n(\theta, h) - M(\theta, h)\|}{1 + \|M_n(\theta, h)\| + \|M(\theta, h)\|} = o_p(1).$$

Then, $\widehat{\theta} - \theta_o = o_p(1)$.

[2] Although our approach can easily be modified to treat the 'maximum-likelihood-like (M-) estimators', we do not state theorems for 'M-estimators' due to space limitation.

Remark 1: (i) Condition 1.5 is implied by **condition 1.5'**: *for all positive sequence* $\delta_n = o(1)$,

$$\sup_{\theta\in\Theta, \|h-h_o\|_{\mathcal{H}}\leq\delta_n} \|M_n(\theta,h) - M(\theta,h)\| = o_p(1).$$

(ii) Comparing our Theorem 1 to Newey's (1994) Lemma 5.2 in the case $M(\theta,h) = E[m(Z,\theta,h)]$, the main difference is that while we impose continuity assumption on $E[m(Z,\theta,h)]$ (with respect to θ, h), Newey imposes a continuity assumption directly on $m(Z,\theta,h)$ (with respect to θ, h).

3.2 Asymptotic Normality

In this subsection we assume that $\widehat{\theta}$ is consistent and $\theta_o \in \text{int}(\Theta)$. Therefore, the parameter spaces $\Theta, \mathcal{H}$ can be replaced by small or even shrinking sets. Define $\Theta_\delta \equiv \{\theta \in \Theta : ||\theta - \theta_o|| \leq \delta\}$ and $\mathcal{H}_\delta \equiv \{h \in \mathcal{H} : \|h-h_o\|_{\mathcal{H}} \leq \delta\}$ for some small $\delta > 0$. The pseudo-norm on $\mathcal{H}_\delta$ can be suitably modified to reflect the smaller parameter space Θ_δ, but for notational simplicity we ignore this. For any $(\theta, h) \in \Theta_\delta \times \mathcal{H}_\delta$, we denote the ordinary derivative of $M(\theta,h)$ with respect to θ as $\Gamma_1(\theta,h)$, which satisfies $\Gamma_1(\theta,h)(\overline{\theta}-\theta) = \lim_{\tau\to 0}[M(\theta+\tau(\overline{\theta}-\theta), h(\cdot,\theta+\tau(\overline{\theta}-\theta))) - M(\theta,h(\cdot,\theta))]/\tau$ for all $\overline{\theta} \in \Theta$. We need a notion of functional derivative to capture the effect of the estimation of h_o via $\widehat{h}$ on the variability of $\widehat{\theta}$. For any $\theta \in \Theta_\delta$, we say that $M(\theta,h)$ is pathwise differentiable at $h \in \mathcal{H}_\delta$ in the direction $[\overline{h}-h]$ if $\{h+\tau(\overline{h}-h) : \tau \in [0,1]\} \subset \mathcal{H}$ and $\lim_{\tau\to 0}[M(\theta, h(\cdot,\theta)+\tau(\overline{h}(\cdot,\theta) - h(\cdot,\theta))) - M(\theta,h(\cdot,\theta))]/\tau$ exists; we denote the limit by $\Gamma_2(\theta,h)[\overline{h}-h]$.

Theorem 2. *Suppose that* $\theta_o \in \Theta_\delta$ *satisfies* $M(\theta_o,h_o) = 0$, *that* $\widehat{\theta} - \theta_o = o_p(1)$, *and that:*
(2.1) $\|M_n(\widehat{\theta},\widehat{h})\| = \inf_{\theta\in\Theta_\delta} \|M_n(\theta,\widehat{h})\| + o_p(1/\sqrt{n})$.
(2.2) *(i) The ordinary derivative* $\Gamma_1(\theta,h_o)$ *in* θ *of* $M(\theta,h_o)$ *exists for* $\theta \in \Theta_\delta$, *and is continuous at* $\theta = \theta_o$; *(ii) the matrix* $\Gamma_1 \equiv \Gamma_1(\theta_o,h_o)$ *is of full (column) rank.*
(2.3) *For all* $\theta \in \Theta_\delta$ *the pathwise derivative* $\Gamma_2(\theta,h_o)[h-h_o]$ *of* $M(\theta,h_o)$ *exists in all directions* $[h-h_o] \in \mathcal{H}$; *and for all* $(\theta,h) \in \Theta_{\delta_n} \times \mathcal{H}_{\delta_n}$ *with a positive sequence* $\delta_n = o(1)$*: (i)* $||M(\theta,h) - M(\theta,h_o) - \Gamma_2(\theta,h_o)[h-h_o]|| \leq c||h-h_o||^2_{\mathcal{H}}$ *for a constant* $c \geq 0$; *(ii)* $||\Gamma_2(\theta,h_o)[h-h_o] - \Gamma_2(\theta_o,h_o)[h-h_o]|| \leq o(1)\delta_n$.
(2.4) $\widehat{h} \in \mathcal{H}$ *with probability tending to one; and* $||\widehat{h}-h_o||_{\mathcal{H}} = o_p(n^{-1/4})$.
(2.5) *For all sequences of positive numbers* $\{\delta_n\}$ *with* $\delta_n = o(1)$,

$$\sup_{\|\theta-\theta_o\|\leq\delta_n, \|h-h_o\|_{\mathcal{H}}\leq\delta_n} \frac{\sqrt{n}\|M_n(\theta,h) - M(\theta,h) - M_n(\theta_o,h_o)\|}{1+\sqrt{n}\{||M_n(\theta,h)|| + ||M(\theta,h)||\}} = o_p(1).$$

(2.6) *For some finite matrix* V_1, $\sqrt{n}\{M_n(\theta_o,h_o) + \Gamma_2(\theta_o,h_o)[\widehat{h}-h_o]\} \Longrightarrow \mathcal{N}[0,V_1]$.
Then, $\sqrt{n}(\widehat{\theta}-\theta_o) \Longrightarrow \mathcal{N}[0,\Omega]$, *where* $\Omega \equiv (\Gamma_1'W\Gamma_1)^{-1}\Gamma_1'WV_1W\Gamma_1(\Gamma_1'W\Gamma_1)^{-1}$.

Remark 2: (i) Condition 2.5 is implied by **condition 2.5'**: *for all positive values* $\delta_n = o(1)$,

$$\sup_{\|\theta-\theta_o\|\leq\delta_n, \|h-h_o\|_{\mathcal{H}}\leq\delta_n} \|M_n(\theta,h) - M(\theta,h) - M_n(\theta_o,h_o)\| = o_p(n^{-1/2}).$$

(ii) Comparing our Theorem 2 to Newey's (1994) Lemma 5.3 in the case $M(\theta,h) = E[m(Z,\theta,h)]$, the main difference is that while we impose smoothness assumption on $E[m(Z,\theta,h)]$ (with respect to θ, h), Newey imposes smoothness assumption directly on $m(Z,\theta,h)$ (with respect to θ, h).

(iii) Condition 2.4 is equivalent to the well-known assumption 5.1(ii) in Newey (1994). Note that when $M(\theta,h)$ is linear in h, this assumption is not explicitly required. However, even in this linear case, the convergence rate $||\widehat{h} - h_o||_{\mathcal{H}} = o_p(n^{-1/4})$ is often needed in order to verify condition 2.6, see e.g., Robinson's (1988) partially linear regression example. Of course such a rate is not needed if an asymptotic orthogonality condition is satisfied as in Andrews (1994a, p.49).

Using the arguments of Pakes and Pollard (1989, Lemma 3.5), we obtain:

Corollary 1. *Let* $\widehat{\theta} = \arg\min_{\theta\in\Theta} M_n(\theta,\widehat{h})'\widetilde{W}_n M_n(\theta,\widehat{h})$, *where* $\widetilde{W}_n = W_n(\widetilde{\theta},\widehat{h}(\cdot,\widetilde{\theta}))$, $\widetilde{\theta} - \theta_o = o_p(1)$, *and* $\{W_n(\theta,h) : \theta\in\Theta, h\in\mathcal{H}\}$ *is a family of sequences of random matrices such that* $\sup_{(\theta,h)\in\Theta_{\delta_n}\times\mathcal{H}_{\delta_n}} \|W_n(\theta,h) - W\| = o_p(1)$ *for all positive values* $\delta_n = o(1)$. *If all conditions of Theorem 2 hold, then* $\sqrt{n}(\widehat{\theta} - \theta_o) \Longrightarrow \mathcal{N}[0,\Omega]$.

3.3 The Asymptotic Variance and the Bootstrap

The verification of condition 2.6 is in some cases difficult; it is itself the subject of a long paper by Newey (1994). Condition 2.6 implicitly assumes that the pathwise derivative $\Gamma_2(\theta_o,h_o)[h-h_o]$ is a smooth linear functional of $h - h_o$. In most applications, h is a square integrable function of U (a subset of Z). Denote F_U as the probability measure of U, then by the Riesz representation theorem there is a unique square integrable function γ_2 of U such that $\Gamma_2(\theta_o,h_o)[\widehat{h} - h_o] = \int \gamma_2(u)[\widehat{h}(u) - h_o(u)]dF_U(u)$. When $\widehat{h}$ has a closed form expression [or can be well approximated thereby] such as an empirical c.d.f., a kernel density or regression estimator, one can directly show that for some function $\psi(.)$ with $E[\psi(U_i)] = 0$, $E[||\psi(U_i)||^2] < \infty$, $\int \gamma_2(u)[\widehat{h}(u) - h_o(u)]dF_U(u) = n^{-1}\sum_{i=1}^n \psi(U_i) + o_p(n^{-1/2})$. The function $\psi(.)$ is, under mild conditions, independent of the precise expression of $\widehat{h}$, and can be arrived at by the Riesz representation approach taken in Newey (1994), Chen and Shen (1998), and Ai and Chen (2002). We present two examples in Section 5 on how to check this condition. To summarize, condition 2.6 is implied by **condition 2.6'**:

(i) $\{Z_i\}_{i=1}^n$ is i.i.d., $M_n(\theta_o,h_o) = n^{-1}\sum_{i=1}^n m(Z_i) + o_p(n^{-1/2})$ with $E[m(Z_i)] = 0$, $E[||m(Z_i)||^2] < \infty$;

(ii) $\Gamma_2(\theta_o,h_o)[\widehat{h} - h_o] = \frac{1}{n}\sum_{i=1}^n \psi(U_i) + o_p(n^{-1/2})$ with $E[\psi(U_i)] = 0$, $E[||\psi(U_i)||^2] < \infty$.

Clearly $V_1 = E\left(\{m(Z_i) + \psi(U_i)\}\{m(Z_i) + \psi(U_i)\}'\right)$ under condition 2.6'.

To estimate V_1 one needs to estimate both ψ and m, which both depend on the unknown θ_o, h_o in perhaps a complicated way, and then compute the sample second moment of the estimated quantities. The estimation of Γ_1 is also potentially difficult, especially in the profiled case where Γ_1 contains an additional term from the effect of θ on M indirectly through h_o. A standard approach here is to use numerical derivatives [see Newey and McFadden (1994) and Powell (1994)]. An alternative is to use the bootstrap to form confidence regions.

We next show that the ordinary nonparametric bootstrap can consistently estimate the asymptotic distribution of $\sqrt{n}(\widehat{\theta} - \theta_o)$ in the special case where $M_n(\theta, h) = n^{-1}\sum_{i=1}^n m(Z_i, \theta, h(Z_i, \theta))$ and $M(\theta, h) = E[m(Z_i, \theta, h(Z_i, \theta))]$ with an i.i.d. sample $\{Z_i\}_{i=1}^n$. This result is similar to the bootstrap theorem in Brown and Wegkamp (2002). Let $\{Z_i^*\}_{i=1}^n$ be drawn randomly with replacement from $\{Z_i\}_{i=1}^n$, and let $\widehat{h}^*$ (for each θ) be the same estimator as $\widehat{h}$ but based on the bootstrap data. Let $M_n^*(\theta, h) = n^{-1}\sum_{i=1}^n m(Z_i^*, \theta, h(Z_i^*, \theta))$ for each (θ, h). Following Hall and Horowitz (1996, p897) it is necessary to recenter the moment condition, at least in the overidentified case. Define the recentered moment function $m^c(z, \theta, h(z, \theta)) = m(z, \theta, h(z, \theta)) - m(z, \widehat{\theta}, \widehat{h}(z, \widehat{\theta}))$ and note that $n^{-1}\sum_{i=1}^n m^c(Z_i^*, \theta, h(Z_i^*, \theta)) = M_n^*(\theta, h) - M_n(\widehat{\theta}, \widehat{h})$. Thus, define the bootstrap estimator $\widehat{\theta}^*$ to be any sequence that satisfies

$$\|M_n^*(\widehat{\theta}^*, \widehat{h}^*) - M_n(\widehat{\theta}, \widehat{h})\| = \inf_{\theta \in \Theta} \|M_n^*(\theta, \widehat{h}^*) - M_n(\widehat{\theta}, \widehat{h})\| + o_{p^*}(n^{-1/2}). \tag{2}$$

Here, and subsequently, superscript $*$ denotes a probability or moment computed under the bootstrap distribution conditional on the original data set $\{Z_i\}_{i=1}^n$.

Theorem B. *Suppose that $\{Z_i\}_{i=1}^n$ is i.i.d. and $\theta_o \in int(\Theta)$ satisfies $E[m(Z_i, \theta_o, h_o)] = 0$; that $\widehat{\theta} - \theta_o = o_{a.s.}(1)$; that conditions 2.1, 2.4, 2.5' and 2.6 hold with 'in probability' replaced by 'almost surely'; that condition 2.2 holds with h_o replaced by $h \in \mathcal{H}_\delta$, while condition 2.3 holds with h_o replaced by $h \in \mathcal{H}_{\delta_n}$; and that $\Gamma_1(\theta, h)$ is continuous [with respect to $||\cdot||_{\mathcal{H}}$] in h at $\theta = \theta_o, h = h_o$. Suppose:*

(2.4B) *With P^*-probability tending to one, $\widehat{h}^* \in \mathcal{H}$, and $||\widehat{h}^* - \widehat{h}||_{\mathcal{H}} = o_{p^*}(n^{-1/4})$.*

(2.5'B) $\sup_{(\theta,h)\in\Theta_{\delta_n}\times\mathcal{H}_{\delta_n}} \|M_n^*(\theta, h) - M_n(\theta, h) - \{M_n^*(\theta_o, h_o) - M_n(\theta_o, h_o)\}\| = o_{P^*}(n^{-1/2})$ *for all positive values $\delta_n = o(1)$.*

(2.6B) $\sqrt{n}\{M_n^*(\widehat{\theta}, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h}) + \Gamma_2(\widehat{\theta},\widehat{h})[\widehat{h}^* - \widehat{h}]\} = \mathcal{N}[0, V_1] + o_{P^*}(1)$.

Then, $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ converges in distribution to a $\mathcal{N}(0, \Omega)$ distribution in P^*-probability.

Condition 2.4B can be verified under the same assumptions implying condition 2.4 for a variety of kernel density and regression estimators, see for example Hall (1991). Theorem 3 below will provide primitive sufficient assumptions for conditions 2.5' and 2.5'B. Finally, condition 2.6B holds under only slightly stronger conditions than those imply condition 2.6. Specifically, from Giné and Zinn (1990) we know that the P^*-distribution of $\sqrt{n}\{M_n^*(\widehat{\theta}, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h})\}$ approximates the distribution of $\sqrt{n}\{M_n(\widehat{\theta}, \widehat{h}) - M(\widehat{\theta}, \widehat{h})\}$, which is approximately the same as the distribution of $\sqrt{n}M_n(\theta_o, h_o)$

by condition 2.5'. In the case of kernel estimation, $\widehat{h}^*(z) - \widehat{h}(z)$, and hence $\sqrt{n}\Gamma_2(\widehat{\theta},\widehat{h})[\widehat{h}^* - \widehat{h}]$, is a sum (or approximately so) of mean zero and independent random variables (under P^*) and can be expected to satisfy a central limit theorem, see e.g. the 'hit rate' example.

4 Primitive Conditions for Stochastic Equicontinuity

In this section we provide some primitive sufficient conditions for conditions 1.5', 2.5' and 2.5'B when $M_n(\theta,h) = n^{-1}\sum_{i=1}^n m(Z_i,\theta,h(Z_i,\theta))$, $M(\theta,h) = E[m(Z_i,\theta,h(Z_i,\theta))]$ and $\{Z_i\}_{i=1}^n$ is an i.i.d. sample. Let $\mathcal{F} = \{m(z,\theta,h) : \theta \in \Theta, h \in \mathcal{H}\}$ denote the class of measurable functions indexed by (θ,h). By modern empirical process theory presented in van der Vaart and Wellner (1996) for example, condition 1.5' of Theorem 1 will be satisfied when $\mathcal{F}$ is $P-$ Glivenko-Cantelli, while condition 2.5' of Theorem 2 will be satisfied if the class $\mathcal{F}$ is $P-$ Donsker. Moreover, whether or not $\mathcal{F}$ is a $P-$Glivenko-Cantelli [or $P-$ Donsker] class is closely linked to its $L_1(P)$ [or $L_2(P)$] covering numbers with bracketing. Let $(\mathcal{G}, ||\cdot||_{\mathcal{G}})$ be a subset of a metric space of real-valued functions $g : \mathcal{X} \to \mathbb{R}$ on some set. The *covering number* $N(\varepsilon,\mathcal{G},||\cdot||_{\mathcal{G}})$ is the minimal number of N for which there exist $\varepsilon-$balls $\{\{f : ||f-g_j||_{\mathcal{G}} \leq \varepsilon\}, ||g_j||_{\mathcal{G}} < \infty, j = 1,\ldots,N\}$ to cover $\mathcal{G}$. The *covering number with bracketing* $N_{[]}(\varepsilon,\mathcal{G},||\cdot||_{\mathcal{G}})$ is the minimal number of N for which there exist $\varepsilon-$brackets $\{[l_j,u_j] : ||l_j-u_j||_{\mathcal{G}} \leq \varepsilon, ||l_j||_{\mathcal{G}}, ||u_j||_{\mathcal{G}} < \infty, j = 1,\ldots,N\}$ to cover $\mathcal{G}$ (i.e., for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1,\ldots,N\}$ such that $l_j \leq g \leq u_j$). Therefore, the key to verifying conditions 1.5' and 2.5' is to compute the covering numbers with bracketing for the moment class $\{m(z,\theta,h)\}$ based on the covering numbers with/without bracketing for the parameter class $\{\theta \in \Theta, h \in \mathcal{H} : \|h-h_o\|_{\mathcal{H}} \leq \delta_n\}$. Since Θ is a compact subset of $\mathbb{R}^k$, the covering number of Θ is known. Since in most applications, we estimate h_o by some nonparametric smoothing methods such as kernel and sieve procedures, h_o is often assumed to be in $\mathcal{H}$, a space of smooth functions such as a Sobolev, Hölder or Besov class, or at least lie there with probability tending to one. Therefore, the covering number of the function space $\mathcal{H}$ can be found in many books and papers on approximation theory. When the moment function $m(z,\theta,h)$ is (pointwise) Lipschitz continuous with respect to (θ,h), we can directly bound $N_{[]}(\varepsilon,\mathcal{F},||\cdot||_{L_2(P)})$ from above by the covering number of the parameter class $\{\theta \in \Theta, h \in \mathcal{H} : \|h-h_o\|_{\mathcal{H}} \leq \delta_n\}$, see e.g., Theorem 2.7.11 of van der Vaart and Wellner (1996). This is the approach taken in Chen and Shen (1998) and many others. When the moment function $m(z,\theta,h)$ is (pointwise) Lipschitz continuous with respect to h but not in θ, we can sometimes still apply the results in Andrews (1994b). However we are unaware of general results to handle the case where the moment function $m(z,\theta,h)$ is not Lipschitz continuous with respect to h. In the following, Theorem 3 extends the work of Andrews (1994b) to the case where the moment function $m(z,\theta,h)$ is not (pointwise) continuous with respect to h and θ.

Theorem 3. *Let* $\{Z_i\}_{i=1}^n$ *be i.i.d. with* $E[m(Z_i, \theta_o, h_o)] = 0$. *Suppose that each component* m_j *of* $m = (m_1, \ldots, m_l)'$ *takes the form* $m_j(z, \theta, h) = m_{cj}(z, \theta, h) + m_{lcj}(z, \theta, h)$, *and satisfies:*
(3.1) $m_{cj}(z, \theta, h)$ *is Hölder continuous with respect to* θ, h *in the sense:*

$$|m_{cj}(z, \theta_1, h_1) - m_{cj}(z, \theta_2, h_2)| \leq b_j(z)\{\|\theta_1 - \theta_2\|^{s_{1j}} + \|h_1 - h_2\|_{\mathcal{H}}^{s_j}\}$$

for some constants s_{1j}, $s_j \in (0, 1]$, *a measurable function* $b_j(\cdot)$ *with* $E[b_j(Z)]^r < \infty$ *for some* $r \geq 2$.
(3.2) $m_{lcj}(\cdot, \theta, h)$ *is locally uniformly* $L_r(P), (r \geq 2)-$*continuous with respect to* θ, h *in the sense:*

$$\left(E\left[\sup_{(\theta', h'): ||\theta' - \theta|| < \delta, ||h' - h||_{\mathcal{H}} < \delta} |m_{lcj}(Z, \theta', h') - m_{lcj}(Z, \theta, h)|^r \right] \right)^{1/r} \leq K_j \delta^{s_j}$$

for all $(\theta, h) \in \Theta \times \mathcal{H}$, *all small positive value* $\delta = o(1)$, *and for some constants* $s_j \in (0, 1]$, $K_j > 0$.
(3.3) Θ *is a compact subset of* $\mathbb{R}^k$, *and* $\int_0^\infty \sqrt{\log N(\varepsilon^{1/s_j}, \mathcal{H}, ||\cdot||_{\mathcal{H}})} d\varepsilon < \infty$ *for* $j = 1, \ldots, l$.
Then: condition 2.5' holds, and condition 2.5'B holds.

Remark 3: (i) Condition 3.1 of Theorem 3 is an extension of the "type II class" of Andrews (1994b) from $\theta \in \Theta$ to $(\theta, h) \in \Theta \times \mathcal{H}$; Condition 3.2 is an extension of the "type IV class" of Andrews (1994b) from $\theta \in \Theta$ to $(\theta, h) \in \Theta \times \mathcal{H}$. Condition 3.2 allows for discontinuous moment functions in (θ, h) such as sign and indicator functions of (θ, h).

(ii) Condition 3.3 of Theorem 3 allows for many nonparametric estimators of h_o. As an example, we recall a popular nonparametric class $\mathcal{H}$ stated in van der Vaart and Wellner (1996, p. 154). For any vector $a = (a_1, \ldots, a_{d_x})$ of d_x integers, define the differential operator $D^a = \partial^{|a|} / \partial x_1^{a_1} \ldots \partial x_{d_x}^{a_{d_x}}$, where $|a| = \sum_{i=1}^{d_x} a_i$. Let R_X be a bounded, convex subset of $\mathbb{R}^{d_x}$ with nonempty interior. For any smooth function $h : R_X \to \mathbb{R}$ and some $\alpha > 0$, let $\underline{\alpha}$ be the largest integer smaller than α, and

$$\|h\|_{\infty, \alpha} = \max_{|a| \leq \underline{\alpha}} \sup_x |D^a h(x)| + \max_{|a| = \underline{\alpha}} \sup_{x \neq x'} \frac{|D^a h(x) - D^a h(x')|}{\|x - x'\|^{\alpha - \underline{\alpha}}}.$$

Further, let $C_c^\alpha(R_X)$ be the set of all continuous functions $h : R_X \to \mathbb{R}$ with $\|h\|_{\infty, \alpha} \leq c$. If $\mathcal{H} = C_c^\alpha(R_X)$ with $||\cdot||_{\mathcal{H}} = ||\cdot||_\infty$, then $\log N(\delta, C_c^\alpha(R_X), ||\cdot||_\infty) \leq const. \times \delta^{-d_x/\alpha}$, see e.g., van der Vaart and Wellner (1996, Theorem 2.7.1). Hence $\int_0^\infty \sqrt{\log N(\varepsilon^{1/s}, C_c^\alpha(R_X), ||\cdot||_\infty)} d\varepsilon < \infty$ if $\alpha > d_x/2s$. That is, when the sample moment function $m(Z, \theta, h)$ is less smooth in h (i.e., smaller $s < 1$), we need $h \in \mathcal{H}$ to be a "smaller function space" (i.e., $\alpha > d_x/2s$ or higher smoothness of h) to satisfy the stochastic equicontinuity condition 2.5'.

(iii) In the old version of this paper, we show that the conclusion of Theorem 3 holds under the following more primitive, yet more restrictive, conditions: (3.1') For fixed h, $m_j(z, \theta, h)$ is componentwise monotone with respect to each θ; and for fixed θ, $m_j(z, \theta, h)$ is monotone with respect to h; (3.2') $m_j(z, \theta, h)$ is $L_r(P), (r \geq 2)-$Hölder continuous with respect to θ, h in the sense that

$\|m_j(Z,\theta,h) - m_j(Z,\theta',h')\|_{L_r(P)} \leq K_j\{\|\theta-\theta'\|^{s_{1j}} + \|h-h'\|_{\mathcal{H}}^{s_j}\}$ for some constants s_{1j}, $s_j \in (0,1]$, $K_j > 0$; and (3.3') Θ is a compact subset of $\mathbb{R}^k$, $\int_0^\infty \sqrt{\log N_{[]}(\varepsilon^{1/s_j}, \mathcal{H}, ||\cdot||_{\mathcal{H}})}d\varepsilon < \infty$ for $j = 1,\ldots,l$. This is an extension of Akritas and Van Keilegom's (2001) lemma 1 for indicator functions.

5 Examples

We present two examples in detail and then discuss some extensions. Both our examples arise in practical situations. Although there may be alternative estimation methods for these problems that avoid the technical issues treated in this paper, the methods we propose seem to be the most natural. The estimation procedure for the second example involves profiling the nonparametric estimation, and so is representative of a broad class of problems. Our aim is to demonstrate that the regularity conditions for asymptotic normality in Theorem 2 and for bootstrap consistency in Theorem B are easily verified. We concentrate primarily on the key conditions 2.5', 2.5'B, 2.6 and 2.6B. Condition 2.3 is verified using similar techniques to 2.6; and the verifications of conditions 2.1, 2.2, 2.4 and 2.4B are standard; so we do not discuss them here.[3] In both examples for random variables Y, X we denote the conditional c.d.f. and density functions at evaluation point $X = x$ by $F_{Y|x}, f_{Y|x}$.

Example 1. *Hit Rates.* Suppose that one wants to estimate the parameter $\theta_o = \Pr[h_o(X) \in A(Z)]$, where $A(Z)$ is a random set that depends on the random variable Z, and h_o is an unknown function. Assuming $h_o \in \mathcal{H} = C_1^\alpha(R_X)$ where $\alpha > d_x$ and R_X is a bounded and convex subset of $\mathbb{R}^{d_x}$ with nonempty interior, and let $\|\cdot\|_{\mathcal{H}} = ||\cdot||_\infty$. The natural estimator of θ_o is the sample analogue

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n 1(\widehat{h}(X_i) \in A(Z_i)), \tag{3}$$

where $\widehat{h}(X_i)$ is some nonparametric estimate of $h(X_i)$. The estimator $\widehat{\theta}$ can be interpreted as a member of our class of estimators by taking the sample moment condition to be $M_n(\theta,h) = n^{-1}\sum_{i=1}^n\{1(h(X_i) \in A(Z_i)) - \theta\}$. Bliss (1997) uses a criterion like (3) to evaluate nonparametric yield curve fits. In this case, one observes a bid and an ask quote on a bond, $\{p_{Li}, p_{Ui}\}$, along with maturity and payment information. For convenience, the mid-point of the bid and ask price is taken as a proxy for the actual price p_i. From this one can estimate nonparametrically the discount function and the yield curve [see Linton, Mammen, Nielsen, and Tanggaard (2001)], and hence obtain a predicted price $\widehat{p}_i$ for each bond. One way of evaluating the performance of the procedure is to calculate the so-called hit rate, which is (3) with $A(Z_i) = [p_{Li}, p_{Ui}]$ and $\widehat{h}(X_i) = \widehat{p}_i$. A high hit rate corresponds to a good procedure. Smoothing over the criterion function is not an attractive alternative here, although it would lead to straightforward but messy distribution theory. Another

[3] The old version of this paper contains a complete set of primitive conditions that imply Theorem 2.

potential application of this example is in testing for Revealed Preference as in Blundell, Browning, and Crawford (2003, section 4.3). Their test involves comparing a weighted combination of budget shares [estimated nonparametrically] with some relative prices. They do a test that uses the pointwise confidence interval and then counts violations. Instead one could do a global test based on a count like (3).

We now verify conditions 2.5', 2.5'B, 2.6, and 2.6B in a special case of (3) where $A(Z_i) = (-\infty, Y_i]$ and where $h_o(.)$ is the density of X_i. We first verify that conditions 2.5' and 2.5'B hold by applying Theorem 3. For any $z = (x, y)$, we have $m(z, \theta, h) = 1(h(x) \leq y) - \theta$, and so $|m(z, \theta', h') - m(z, \theta, h)|^2 \leq 2\,|1(h'(x) \leq y) - 1(h(x) \leq y)| + 2\,|\theta' - \theta|^2$. Thus for all small $\delta \in (0, 1]$,

$$\sup_{|\theta'-\theta|\leq\delta, ||h'-h||_{\mathcal{H}}\leq\delta} |m(z, \theta', h') - m(z, \theta, h)|^2 \leq 2\delta^2 + 2 \sup_{||h'-h||_{\mathcal{H}}\leq\delta} |1(h'(x) \leq y) - 1(h(x) \leq y)|\,.$$

Since for all $h' \in \mathcal{H}$ with $||h' - h||_{\mathcal{H}} \leq \delta \leq 1$, we have for all y, x:

$$\begin{aligned} h(x) - \delta &\leq h'(x) \leq h(x) + \delta \text{ hence } 1(h(x) - \delta \leq y) \geq 1(h'(x) \leq y) \geq 1(h(x) + \delta \leq y) \\ h(x) - \delta &\leq h(x) \leq h(x) + \delta \text{ hence } 1(h(x) - \delta \leq y) \geq 1(h(x) \leq y) \geq 1(h(x) + \delta \leq y), \end{aligned}$$

hence $\sup_{||h'-h||_{\mathcal{H}}\leq\delta} |1(h'(x) \leq y) - 1(h(x) \leq y)| \leq 1(h(x) - \delta \leq y) - 1(h(x) + \delta \leq y)$. The right hand side of the preceding term is either one or zero, and its expectation is the probability that $h(X) + \delta > Y \geq h(X) - \delta$ occurs. Then apply the law of iterated expectations to conclude that for small enough $\delta > 0$, $E[\sup_{|\theta'-\theta|\leq\delta, ||h'-h||_{\mathcal{H}}\leq\delta} |m(Z, \theta', h') - m(Z, \theta, h)|^2]$ is bounded above by

$$2\Pr\left[h(X) + \delta > Y \geq h(X) - \delta\right] + 2\delta^2 = 2E\left[F_{Y|X}(h(X) + \delta) - F_{Y|X}(h(X) - \delta)\right] + 2\delta^2 \leq K\delta,$$

for a finite constant $K > 0$, where the last inequality follows provided $F_{Y|x}$ is Lipschitz in Y uniformly in x. Therefore, condition 3.2 of Theorem 3 is satisfied with $r = 2$ and $s = 1/2$, and condition 3.3 of Theorem 3 is satisfied by Remark 3(ii) and the assumption that $h_o \in \mathcal{H} = C_1^\alpha(R_X)$ with $||\cdot||_{\mathcal{H}} = ||\cdot||_\infty$ and $\alpha > d_x$.

We now verify condition 2.6. Note that $M(\theta, h) = \int [1 - F_{Y|x}(h(x))] h_o(x) dx - \theta$. By the law of iterated expectations and interchanging limits we obtain

$$\Gamma_2(\theta, h_o)[h - h_o] = -\int f_{Y|x}(h_o(x))[h(x) - h_o(x)]h_o(x)dx.$$

Now suppose that $\widehat{h}(x)$ is a kernel estimator, i.e., $\widehat{h}(x) = n^{-1}b^{-1}\sum_{i=1}^n K((x - X_i)/b)$ for some kernel K and bandwidth b. Under standard regularity conditions the bias of the nonparametric estimator, $E\widehat{h}(x) - h_o(x)$, can be majorized by some bounded continuous function of x times $o(n^{-1/2})$. Therefore, we just need to examine $\Gamma_2(\theta, h_o)[\widehat{h} - E\widehat{h}]$, which by construction is a sum of mean zero independent

random variables. Using standard change of variables and Taylor expansion arguments we have

$$\begin{aligned}\Gamma_2(\theta,h_o)[\widehat{h}-E\widehat{h}] &= -\frac{1}{nb}\sum_{i=1}^{n}\int f_{Y|x}(h_o(x))h_o(x)\left[K\left(\frac{x-X_i}{b}\right)-EK\left(\frac{x-X_i}{b}\right)\right]dx \\ &= -\frac{1}{n}\sum_{i=1}^{n} f_{Y|X_i}(h_o(X_i))h_o(X_i)-E[f_{Y|X_i}(h_o(X_i))h_o(X_i)]+o_p(n^{-1/2}),\end{aligned}$$

provided the required smoothness and boundedness conditions hold on $f_{Y|x}, h_o$, and the kernel and bandwidth satisfy various conditions; this yields condition 2.6. Regarding condition 2.6B,

$$\begin{aligned}\Gamma_2(\widehat{\theta},\widehat{h})[\widehat{h}^*-\widehat{h}] &= -\frac{1}{nb}\sum_{i=1}^{n}\int f_{Y|x}(\widehat{h}(x))\widehat{h}(x)\left[K\left(\frac{x-X_i^*}{b}\right)-E^*K\left(\frac{x-X_i^*}{b}\right)\right]dx \\ &= -\frac{1}{n}\sum_{i=1}^{n} f_{Y|X_i^*}(\widehat{h}(X_i^*))\widehat{h}(X_i^*)-E^*[f_{Y|X_i^*}(\widehat{h}(X_i^*))\widehat{h}(X_i^*)]+o_{p^*}(n^{-1/2}),\end{aligned}$$

where the approximation follows from the same change of variables and Taylor expansion arguments used above. For this we need that $\widehat{h}(.)$ possesses the same smoothness as $h_o(.)$, which it does by condition 2.4. In the P^*-distribution, $Y_{ni}^* = -f_{Y|X_i^*}(\widehat{h}(X_i^*))\widehat{h}(X_i^*) + E^*[f_{Y|X_i^*}(\widehat{h}(X_i^*))\widehat{h}(X_i^*)]$ are independent and mean zero random variables and $\sum_{i=1}^n Y_{ni}^*/\sqrt{n}$ satisfies a triangular array central limit theorem under weak additional conditions; the asymptotic variance is $E[f^2_{Y|X_i}(h_o(X_i))h_o^2(X_i)] - E^2[f_{Y|X_i}(h_o(X_i))h_o(X_i)]$ under the smoothness conditions and uniform convergence of $\widehat{h}(.)$.

Example 2. *Partially Linear Median Regression with some Endogenous Regressors.* Suppose

$$Y_i = X_{1i}'\theta_o + h_*(X_{2i}) + \varepsilon_i, \quad \text{med}(\varepsilon_i|X_{2i},X_{3i}) = 0 \ a.s., \tag{4}$$

where $h_*(.)$ is an unknown function. We have partitioned $X_i = (X_{1i}, X_{2i}, X_{3i})$. The regressors X_{1i} are endogenous, but we assume that there exist valid instruments X_{3i} whose dimensionality (weakly) exceeds that of X_{1i}. X_{3i} could include some of X_{2i}. The partially linear functional form has been discussed in many places before; it provides a convenient and interpretable intermediate specification between parametric and nonparametric. We can replace $h_*(X_{2i})$ by an index function or an additive function in cases where X_{2i} has high dimensions. In the case with exogenous X_{1i} (say when $X_{1i} = X_{3i}$), Lee (2003) has proposed an estimation method for θ_o that relies on preliminary high-dimensional nonparametric quantile regression function of Y_i given (X_{1i}, X_{2i}). Our method only requires smoothing operations with the dimensions of X_{2i} and permits endogeneity of X_{1i}. For any fixed $\theta \in \Theta$, we denote the function $h_o(X_{2i};\theta) \equiv \text{med}(Y_i - X_{1i}'\theta|X_{2i})$. Clearly, $h_*(X_{2i}) = h_o(X_{2i};\theta_o)$. Assuming for all $\theta \in \Theta$, $h_o(\cdot;\theta) \in \mathcal{H} = C_1^\alpha(R_{X_2})$ where $\alpha > d_2$ and R_{X_2} is a bounded and convex subset of $\mathbb{R}^{d_2}$ with nonempty interior, and let $\|h_o\|_{\mathcal{H}} = \sup_\theta \sup_w |h_o(w;\theta)|$. We first estimate $h_o(\cdot;\theta)$ for each θ by the conditional median of $Y_i - X_{1i}'\theta$ given X_{2i} using some smoothing method like kernels

or series, denoting the estimator $\widehat{h}(\cdot,\theta)$. We next estimate θ by $\widehat{\theta} = \arg\min_\theta ||M_n(\theta,\widehat{h})||$, where

$$M_n(\theta,h) = \frac{1}{n}\sum_{i=1}^n X_{3i}\left[0.5 - 1\{Y_i \le X_{1i}'\theta + h(X_{2i};\theta)\}\right].$$

We now verify conditions 2.5', 2.5'B, 2.6 and 2.6B with $m(z,\theta,h) = x_3[0.5 - 1\{y \le x_1'\theta + h(x_2;\theta)\}]$, and $\dim(X_3) = l \ge k$. We apply Theorem 3 to show conditions 2.5' and 2.5'B. For $j = 1,\ldots,l$,

$$\begin{aligned}|m_j(z,\theta',h') - m_j(z,\theta,h)|^2 \;\le\;& x_{3j}^2\{|1(y \le x_1'\theta' + h'(x_2;\theta')) - 1(y \le x_1'\theta + h'(x_2;\theta'))|\}\\ &+x_{3j}^2\{|1(y \le x_1'\theta + h'(x_2;\theta')) - 1(y \le x_1'\theta + h'(x_2;\theta))|\}\\ &+x_{3j}^2\{|1(y \le x_1'\theta + h'(x_2;\theta)) - 1(y \le x_1'\theta + h(x_2;\theta))|\}.\end{aligned}$$

We consider only the last term of the sum in the above right hand side, since the two other terms can be treated similarly. By the arguments used in the previous example:

$$\begin{aligned}&E\left[\sup_{\|h'-h\|_{\mathcal{H}}<\delta} X_{3j}^2|1(Y \le X_1'\theta + h'(X_2;\theta)) - 1(Y \le X_1'\theta + h(X_2;\theta))|\right]\\ \le\;& E\left[X_{3j}^2|1(Y \le X_1'\theta + h(X_2;\theta) + \delta) - 1(Y \le X_1'\theta + h(X_2;\theta) - \delta)|\right]\\ \le\;& E\left[X_{3j}^2\{F_{Y|X}(X_1'\theta + h(X_2;\theta) + \delta) - F_{Y|X}(X_1'\theta + h(X_2;\theta) - \delta)\}\right] \le K_j\delta\end{aligned}$$

for some $K_j < \infty$, under suitable conditions on $F_{Y|X}$. Hence condition 3.2 is satisfied with $r = 2$ and $s_j = 1/2$ and condition 3.3 holds by Remark 3(ii). We now verify condition 2.6. Let

$$M(\theta,h) = E[m(Z_i,\theta,h)] = E\{X_{3i}[0.5 - F_{Y|X_i}(X_{1i}'\theta + h(X_{2i};\theta))]\},$$

$$\Gamma_1 = \frac{\partial M(\theta,h_o)}{\partial\theta}|_{\theta=\theta_o} = -E\{X_{3i}f_{Y|X_i}(X_{1i}'\theta_o + h_*(X_{2i}))[X_{1i}' + \frac{\partial h_o}{\partial\theta}(X_{2i};\theta_o)]\}.$$

Because $\text{med}(\varepsilon_i|X_{3i}) = 0$ we have $E\left[X_{3i}\{0.5 - 1(\varepsilon_i \le 0)\}\right] = 0$ and hence $M(\theta_o,h_o) = 0$. It follows that θ_o is uniquely identified as long as Γ_1 is non-singular. By similar reasoning as before

$$\Gamma_2(\theta_o,h_o)[h - h_o] = -E\{X_{3i}f_{Y|X_i}(X_{1i}'\theta_o + h_*(X_{2i}))[h(X_{2i};\theta_o) - h_o(X_{2i};\theta_o)]\}.$$

We now substitute in the Bahadur representation for $\widehat{h}(x_2;\theta_o) - h_o(x_2;\theta_o)$ [obtained by Chaudhuri (1991) for local polynomials], interchange integral and summation, and approximate to obtain

$$\Gamma_2(\theta_o,h_o)[\widehat{h} - h_o] = \frac{1}{n}\sum_{i=1}^n[0.5 - 1\{\varepsilon_i \le 0\}]v_*(X_{2i}) + o_p(n^{-1/2}),$$

where $v_*(X_{2i}) = -E[f_{\varepsilon|X_i}(0)X_{3i}|X_{2i}] \div f_{\varepsilon|X_{2i}}(0)$,[4] and we have used the fact that $f_{Y|X_i}(X_{1i}'\theta_o + h_*(X_{2i})) = f_{\varepsilon|X_i}(0)$. Using the definition of $M_n(\theta_o,h_o)$ it follows that

$$M_n(\theta_o,h_o) + \Gamma_2(\theta_o,h_o)[\widehat{h} - h_o] = \frac{1}{n}\sum_{i=1}^n[0.5 - 1\{\varepsilon_i \le 0\}][X_{3i} + v_*(X_{2i})] + o_p(n^{-1/2}),$$

[4]We thank an anonymous referee for finding an error in our formula for v_* in a previous version of this paper.

which is asymptotically normal with mean zero and finite variance under some conditions. Condition 2.6B is satisfied along the same lines as in the previous example using the corresponding Bahadur representation for $\widehat{h}^* - \widehat{h}$.

Finally, the model (4) can be easily generalized to allow for censoring or truncation at least in the absence of endogeneity. Thus suppose that $Y_i = \max\{X_{1i}\theta_o + h_*(X_{2i}) + \varepsilon_i, 0\}$, where $\text{med}(\varepsilon_i|X_{1i}, X_{2i}) = 0$. The usual CLAD estimation method of Powell (1984) can be extended to this case with moment function $m(z, \theta, h) = x_1 1\{x_1'\theta + h(x_2;\theta) > 0\}[0.5 - 1\{y \leq x_1'\theta + h(x_2;\theta)\}]$. The verifications of conditions 2.5' and 2.6 are pretty much the same as in the uncensored case.[5]

6 Concluding Remarks

Finally, we discuss some further areas of application in estimation. An example of current interest appears in Han and Tamer (2002): they considered a linear median regression model $Y_i = X_i'\theta + \varepsilon_i$ in which the covariates can be endogenous and where the dependent variable is subject to a sort of interval censoring, i.e., you only observe a lower and upper bound Y_0, Y_1 on Y. However, you do observe some instruments W for which $\text{med}(\varepsilon_i|W_i) = 0$. Let $m(z, \theta, h) = w\{(0.5 - h_1(z;\theta))\, 1\,(h_1(z;\theta) > 0.5) + (0.5 - h_2(z;\theta))^2\, 1\,(h_2(z;\theta) < 0.5)\}$, where $z = (y_0, y_1, x, w)$ and $h = (h_1, h_2)$ with $h_o = (h_{1o}, h_{2o}) = (F_{Y_0 - X\theta|W}, F_{Y_1 - X\theta|W})$. Then $E[m(Z, \theta, h_o)] = 0$ if and only if $\theta = \theta_o$. Han and Tamer smoothed over the indicator functions and actually worked with the corresponding least squares minimization problem. However, instead one can work with the moment condition m and define an estimator like in (1).

There is a large class of semiparametric models defined through an independence restriction between regressors and error terms. For example, transformation models like in Horowitz (1998b, Chapter 5) and Linton et al. (1997). An appealing estimation procedure here is the minimum distance method [see Koul (2001) for a nice review] which involves minimizing the mean squared distance from independence based on estimated empirical c.d.f.'s. Manski (1983) proposed a version of the minimum distance estimator for models with separable independent error terms. More recently, Brown and Wegkamp (2002) have applied this method to estimating nonlinear 'parametric' simultaneous equations. Our results suggest that the extension to allow for estimated nonparametric components [like for example in Linton et al. (1997)] should hold.

[5] The Lewbel and Linton (2002) or Chen and Khan (2001) procedures can be applied to estimate $h(\cdot;\theta)$.

7 Appendix

Proof of Theorem 1. The proof is similar to that of Corollary 3.2 in Pakes and Pollard (1989). By condition 1.2, for all $\delta > 0$, $\Pr(||\widehat{\theta} - \theta_o|| > \delta) \leq \Pr(||M(\widehat{\theta}, h_o)|| \geq \epsilon(\delta))$, hence it suffices to show that $||M(\widehat{\theta}, h_o)|| = o_p(1)$. Now by the triangle inequality, $||M(\widehat{\theta}, h_o)|| \leq ||M(\widehat{\theta}, h_o) - M(\widehat{\theta}, \widehat{h})|| + ||M(\widehat{\theta}, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h})|| + ||M_n(\widehat{\theta}, \widehat{h})||$. It follows that by conditions 1.3 and 1.4, $||M(\widehat{\theta}, h_o) - M(\widehat{\theta}, \widehat{h})|| = o_p(1)$, and hence by conditions 1.4 and 1.5, $||M(\widehat{\theta}, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h})|| \leq o_p(1) \times \{1 + ||M_n(\widehat{\theta}, \widehat{h})|| + ||M(\widehat{\theta}, h_o)|| + o_p(1)\}$. Hence $||M(\widehat{\theta}, h_o)|| \times \{1 - o_p(1)\} \leq o_p(1) + ||M_n(\widehat{\theta}, \widehat{h})|| \times \{1 + o_p(1)\}$. By condition 1.1, we have $||M_n(\widehat{\theta}, \widehat{h})|| \leq o_p(1) + \inf_{\theta \in \Theta} ||M_n(\theta, \widehat{h})||$. Under conditions 1.3, 1.4, 1.5 and $M(\theta_o, h_o) = 0$, we have:

$$\begin{aligned} ||M_n(\theta, \widehat{h})|| &\leq ||M_n(\theta, \widehat{h}) - M(\theta, \widehat{h})|| + ||M(\theta, \widehat{h}) - M(\theta, h_o)|| + ||M(\theta, h_o) - M(\theta_o, h_o)|| \\ &\leq o_p(1) \times \{1 + ||M_n(\theta, \widehat{h})|| + ||M(\theta, h_o)||\} + o_p(1) + ||M(\theta, h_o) - M(\theta_o, h_o)||, \end{aligned}$$

and hence $||M_n(\theta, \widehat{h})|| \times \{1 - o_p(1)\} \leq o_p(1) + ||M(\theta, h_o) - M(\theta_o, h_o)|| \times \{1 + o_p(1)\}$, where all the $o_p(1)$'s hold uniformly with respect to $\theta \in \Theta$. Now by $M(\theta_o, h_o) = 0$,

$$\inf_{\theta \in \Theta} ||M_n(\theta, \widehat{h})|| \leq \sup_{\theta \in \Theta} o_p(1) + \inf_{\theta \in \Theta} ||M(\theta, h_o) - M(\theta_o, h_o)|| \times \{1 + \sup_{\theta \in \Theta} o_p(1)\} = o_p(1)$$

and the result follows. ■

Proof of Theorem 2. We first establish $\sqrt{n}-$consistency of $\widehat{\theta}$ to θ_o. We choose a positive sequence $\delta_n = o(1)$ such that $\Pr(||\widehat{\theta} - \theta_o|| \geq \delta_n, ||\widehat{h} - h_o||_{\mathcal{H}} \geq \delta_n) \to 0$. Hence we just need to look at $(\theta, h) \in \Theta_\delta \times \mathcal{H}_\delta$. By condition 2.2, there is a constant $C > 0$ such that $||\widehat{\theta} - \theta_o||C$ is bounded by $||M(\widehat{\theta}, h_o)||$ with probability tending to one; this in turn is bounded above by

$$||M(\widehat{\theta}, h_o) - M(\widehat{\theta}, \widehat{h})|| + ||M(\widehat{\theta}, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h}) + M_n(\theta_o, h_o)|| + ||M_n(\widehat{\theta}, \widehat{h})|| + O_p(n^{-1/2}) \tag{5}$$

by the triangle inequality and condition 2.6. By conditions 2.3, 2.4 and 2.6, and the fact that $(\widehat{\theta}, \widehat{h}) \in \Theta_\delta \times \mathcal{H}_\delta$ with probability approaching one, we have

$$\begin{aligned} ||M(\widehat{\theta}, h_o) - M(\widehat{\theta}, \widehat{h})|| &\leq ||M(\widehat{\theta}, \widehat{h}) - M(\widehat{\theta}, h_o) - \Gamma_2(\widehat{\theta}, h_o)[\widehat{h} - h_o]|| \\ &\quad + ||\Gamma_2(\widehat{\theta}, h_o)[\widehat{h} - h_o] - \Gamma_2(\theta_o, h_o)[\widehat{h} - h_o]|| + ||\Gamma_2(\theta_o, h_o)[\widehat{h} - h_o]|| \\ &\leq c\{||\widehat{h} - h_o||_{\mathcal{H}}\}^2 + ||\widehat{\theta} - \theta_o|| \times o_p(1) + ||\Gamma_2(\theta_o, h_o)[\widehat{h} - h_o]|| \\ &= o_p(n^{-1/2}) + ||\widehat{\theta} - \theta_o|| \times o_p(1) + O_p(n^{-1/2}) \\ &\leq \|M(\widehat{\theta}, h_o)\| \times o_p(1) + O_p(n^{-1/2}), \quad \text{by condition 2.2.} \end{aligned} \tag{6}$$

And by condition 2.5,

$$\begin{aligned} &||M(\widehat{\theta}, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h}) + M_n(\theta_o, h_o)|| \leq o_p(1) \times \{n^{-1/2} + ||M_n(\widehat{\theta}, \widehat{h})|| + ||M(\widehat{\theta}, \widehat{h})||\} \\ (\text{by (6)}) \quad &\leq o_p(1) \times \{n^{-1/2} + ||M_n(\widehat{\theta}, \widehat{h})|| + ||M(\widehat{\theta}, h_o)|| + \|M(\widehat{\theta}, h_o)\| \times o_p(1) + O_p(n^{-1/2})\} \\ &= o_p(n^{-1/2}) + o_p(1) \times \{||M_n(\widehat{\theta}, \widehat{h})|| + ||M(\widehat{\theta}, h_o)|| \times (1 + o_p(1))\}. \end{aligned}$$

This and condition 2.1 imply that $||M(\widehat{\theta}, h_o)|| \times \{1 - o_p(1)\}$ is bounded above by

$$o_p(n^{-1/2}) + ||M_n(\widehat{\theta}, \widehat{h})|| \times \{1 + o_p(1)\} + O_p(n^{-1/2}) \leq \inf_{\theta \in \Theta_\delta} ||M_n(\theta, \widehat{h})|| \times \{1 + o_p(1)\} + O_p(n^{-1/2}).$$

Again under conditions 2.3 - 2.6, we have that $||M_n(\theta, \widehat{h})||$ is bounded above by

$$\begin{aligned} & ||M_n(\theta, \widehat{h}) - M(\theta, \widehat{h}) - M_n(\theta_o, h_o)|| + ||M(\theta, \widehat{h}) - M(\theta, h_o)|| + ||M(\theta, h_o)|| + ||M_n(\theta_o, h_o)|| \\ \leq \;& o_p(n^{-1/2}) + o_p(1) \times \{||M_n(\theta, \widehat{h})|| + ||M(\theta, h_o)|| \times (1 + o_p(1))\} + ||M(\theta, h_o)|| + O_p(n^{-1/2}) \end{aligned}$$

with $M(\theta_o, h_o) = 0$ we have: $||M_n(\theta, \widehat{h})|| \times \{1 - o_p(1)\} \leq o_p(1) \times ||M(\theta, h_o) - M(\theta_o, h_o)|| + O_p(n^{-1/2})$, where all the $o_p(1), O_p(n^{-1/2})$'s hold uniformly with respect to $\theta \in \Theta_\delta$. Hence by condition 2.2, $\inf_{\theta \in \Theta_\delta} ||M_n(\theta, \widehat{h})|| = O_p(n^{-1/2})$ and $||\widehat{\theta} - \theta_o|| C \leq ||M(\widehat{\theta}, h_o)|| \leq O_p(n^{-1/2})$.

The rest of the proof is very similar to that of Theorem 3.3 in Pakes and Pollard (1989) for $\sqrt{n}-$normality, hence we just sketch the main steps here. Define the linearization $\mathcal{L}_n(\theta) = M_n(\theta_o, h_o) + \Gamma_1(\theta - \theta_o) + \Gamma_2(\theta_o, h_o)[\widehat{h} - h_o]$. By conditions 2.2 - 2.5 and the root-n rate results above,

$$\begin{aligned} ||M_n(\widehat{\theta}, \widehat{h}) - \mathcal{L}_n(\widehat{\theta})|| &= ||M_n(\theta_o, h_o) + M(\widehat{\theta}, \widehat{h}) + M_n(\widehat{\theta}, \widehat{h}) - M(\widehat{\theta}, \widehat{h}) - M_n(\theta_o, h_o) - \mathcal{L}_n(\widehat{\theta})|| \\ &\leq ||M(\widehat{\theta}, \widehat{h}) - M(\widehat{\theta}, h_o) - \Gamma_2(\theta_o, h_o)[\widehat{h} - h_o]|| + ||M(\widehat{\theta}, h_o) - \Gamma_1(\widehat{\theta} - \theta_o)|| \\ &\quad + ||M_n(\widehat{\theta}, \widehat{h}) - M(\widehat{\theta}, \widehat{h}) - M_n(\theta_o, h_o)|| = o_p(n^{-1/2}). \end{aligned}$$

Similarly, $||M_n(\overline{\theta}, \widehat{h}) - \mathcal{L}_n(\overline{\theta})|| = o_p(n^{-1/2})$, where

$$\sqrt{n}(\overline{\theta} - \theta_o) = -(\Gamma_1' W \Gamma_1)^{-1} \Gamma_1' W \sqrt{n} \left[M_n(\theta_o, h_o) + \Gamma_2(\theta_o, h_o)[\widehat{h} - h_o] \right]$$

is the minimizer of $\mathcal{L}_n(\theta)$. A little more work gives that $\sqrt{n}(\widehat{\theta} - \overline{\theta}) = o_p(1)$, this and condition 2.6 imply that $\sqrt{n}(\widehat{\theta} - \theta_o) \Longrightarrow \mathcal{N}[0, \Omega]$. ■

Proof of Theorem B. Denote $\nu_n^*(\theta, h) = \sqrt{n}(M_n^*(\theta, h) - M_n(\theta, h))$ and $\nu_n(\theta, h) = \sqrt{n}(M_n(\theta, h) - M(\theta, h))$. By the triangle inequality, condition 2.5' (almost sure version) and condition 2.5'B, we have for all positive sequence $\delta_n = o(1)$,

$$\sup_{(\theta,h),(\theta',h') \in \Theta_{\delta_n} \times \mathcal{H}_{\delta_n}} ||\nu_n^*(\theta', h') - \nu_n^*(\theta, h)|| = o_{p*}(1) \text{ a.s. [P]} \tag{7}$$

$$\sup_{(\theta,h),(\theta',h') \in \Theta_{\delta_n} \times \mathcal{H}_{\delta_n}} ||\nu_n(\theta', h') - \nu_n(\theta, h)|| = o_{a.s.}(1) \tag{8}$$

In a similar way as was done in the proof of Theorem 2, it can be shown that $\|\hat{\theta}^* - \hat{\theta}\| = O_{P^*}(n^{-1/2})$ a.s.[P]. Next we approximate $M_n^*(\theta, \widehat{h}^*) - M_n(\widehat{\theta}, \widehat{h})$ with error $o_{p^*}(n^{-1/2})$ by the linear function $\mathcal{L}_n^*(\theta) = M_n^*(\widehat{\theta}, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h}) + \Gamma_2(\widehat{\theta}, \widehat{h})[\widehat{h}^* - \widehat{h}] + \Gamma_1(\widehat{\theta}, \widehat{h})(\theta - \widehat{\theta})$ for θ in a root-n neighborhood

of $\widehat{\theta}$. Here, $\Gamma_1(\widehat{\theta},\widehat{h})$ is the derivative of $M(\theta,h)$ with respect to θ evaluated at $\theta=\widehat{\theta}, h=\widehat{h}$. By adding and subtracting and the triangle inequality

$$\begin{aligned}
& ||M_n^*(\widehat{\theta}^*,\widehat{h}^*) - M_n(\widehat{\theta},\widehat{h}) - \mathcal{L}_n^*(\widehat{\theta}^*)|| \\
= \; & ||M_n^*(\widehat{\theta},\widehat{h}) - M_n(\widehat{\theta},\widehat{h}) + M(\widehat{\theta}^*,\widehat{h}^*) - M(\widehat{\theta},\widehat{h}) + M_n(\widehat{\theta}^*,\widehat{h}^*) - M(\widehat{\theta}^*,\widehat{h}^*) - M_n(\widehat{\theta},\widehat{h}) + M(\widehat{\theta},\widehat{h}) \\
& + M_n^*(\widehat{\theta}^*,\widehat{h}^*) - M_n(\widehat{\theta}^*,\widehat{h}^*) - M_n^*(\widehat{\theta},\widehat{h}) + M_n(\widehat{\theta},\widehat{h}) - \mathcal{L}_n^*(\widehat{\theta}^*)|| \\
\leq \; & ||M(\widehat{\theta}^*,\widehat{h}^*) - M(\widehat{\theta}^*,\widehat{h}) - \Gamma_2(\widehat{\theta},\widehat{h})[\widehat{h}^* - \widehat{h}]|| + ||M(\widehat{\theta}^*,\widehat{h}) - M(\widehat{\theta},\widehat{h}) - \Gamma_1(\widehat{\theta},\widehat{h})(\widehat{\theta}^* - \widehat{\theta})|| \\
& + n^{-1/2}||\nu_n(\widehat{\theta}^*,\widehat{h}^*) - \nu_n(\widehat{\theta},\widehat{h})|| + n^{-1/2}||\nu_n^*(\widehat{\theta}^*,\widehat{h}^*) - \nu_n^*(\widehat{\theta},\widehat{h})|| = o_{p^*}(n^{-1/2}),
\end{aligned}$$

where the $o_{p^*}(n^{-1/2})$ follows from the root-n consistencies, the stochastic equicontinuity properties (7-8), and conditions 2.3 and 2.4B. Similarly $||M_n^*(\overline{\theta}^*,\widehat{h}^*) - M_n(\widehat{\theta},\widehat{h}) - \mathcal{L}_n^*(\overline{\theta}^*)|| = o_{p^*}(n^{-1/2})$, where $\overline{\theta}^*$ is the minimizer of $\mathcal{L}_n^*$. It follows that

$$\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}) = -(\widehat{\Gamma}_1' W \widehat{\Gamma}_1)^{-1}\widehat{\Gamma}_1' W \sqrt{n}\left[M_n^*(\widehat{\theta},\widehat{h}) - M_n(\widehat{\theta},\widehat{h}) + \Gamma_2(\widehat{\theta},\widehat{h})[\widehat{h}^* - \widehat{h}]\right] + o_{p_*}(1),$$

where $\widehat{\Gamma}_1 = \Gamma_1(\widehat{\theta},\widehat{h})$. We can replace $\widehat{\Gamma}_1$ by Γ_1 with probability one and use our assumption 2.6B. ■

Lemma 1. *Let $\{Z_i\}_{i=1}^n$ be i.i.d. with $E[m(Z_i,\theta_o,h_o)] = 0$. Suppose that $\mathcal{F} = \{m(Z,\theta,h) : \theta\in\Theta, h\in\mathcal{H}\}$ is $P-$Donsker (or satisfies $\int_0^\infty \sqrt{\log N_{[]}(\varepsilon,\mathcal{F},||\cdot||_{L_2(P)})}d\varepsilon < \infty$); and that $m(\cdot,\theta,h)$ is $L_2(P)-$continuous at (θ_o,h_o), (i.e., $E[||m(Z,\theta,h)-m(Z,\theta_o,h_o)||^2]\to 0$ as $\|\theta-\theta_o\|\to 0, \|h-h_o\|_{\mathcal{H}}\to 0$). Then: conditions 2.5' and 2.5'B hold.*

Proof of Lemma 1. The fact that condition 2.5' holds is a direct extension of Pakes and Pollard's Lemma 2.17 from their case $m(Z,\theta)$ to our case $m(Z,\theta,h)$; and can be proved the same way as theirs. The fact that condition 2.5'B holds now follows from Giné and Zinn (1990). ■

Proof of Theorem 3. We obtain the result by applying Lemma 1. Since either condition 3.1 or condition 3.2 implies that $m(\cdot,\theta,h)$ is $L_2(P)-$continuous at (θ_o,h_o), by Theorem 6 in Andrews (1994b), it suffices to show that for each $j=1,\ldots,l$, the followings are true:

(1) $\mathcal{F}_{1j} = \{m_{cj}(Z,\theta,h) : \theta\in\Theta, h\in\mathcal{H}\}$ is $P-$Donsker, or $\int_0^\infty \sqrt{\log N_{[]}(\varepsilon,\mathcal{F}_{1j},||\cdot||_{L_r(P)})}d\varepsilon<\infty$;

(2) $\mathcal{F}_{2j} = \{m_{lcj}(Z,\theta,h) : \theta\in\Theta, h\in\mathcal{H}\}$ is $P-$Donsker, or $\int_0^\infty \sqrt{\log N_{[]}(\varepsilon,\mathcal{F}_{2j},||\cdot||_{L_r(P)})}d\varepsilon<\infty$.

For **(1)**. Let $\{\theta_k : k=1,\ldots,N_1\}$ be an $\eta^{1/s_{1j}}-$cover for $(\Theta,||\cdot||)$, and $\{h_k : k=1,\ldots,N_2\}$ be an $\eta^{1/s_j}-$cover for $(\mathcal{H},||\cdot||_{\mathcal{H}})$. Then under condition 3.1, for any $m_{cj}(Z,\theta,h)$, there exist $k_1\in\{1,\ldots,N_1\}$ and $k_2\in\{1,\ldots,N_2\}$ such that $m_{cj}(Z,\theta_{k_1},h_{k_2}) - 2\eta b_j(Z) \leq m_{cj}(Z,\theta,h) \leq m_{cj}(Z,\theta_{k_1},h_{k_2}) + 2\eta b_j(Z)$. Since $E[b_j(Z)]^r<\infty$ for some $r\geq 2$, we have that $\{[m_{cj}(Z,\theta_{k_1},h_{k_2}) - 2\eta b_j(Z), m_{cj}(Z,\theta_{k_1},h_{k_2}) + 2\eta b_j(Z)] : k_1\in\{1,\ldots,N_1\}, k_2\in\{1,\ldots,N_2\}\}$ forms an $\varepsilon = 4\eta||b_j(Z)||_{L_r(P)}-$bracket for $(\mathcal{F}_{1j},||\cdot||_{L_r(P)})$. Hence

$$N_{[]}(\varepsilon,\mathcal{F}_{1j},||\cdot||_{L_r(P)}) \leq N([\frac{\varepsilon}{4||b_j(Z)||_{L_r(P)}}]^{1/s_{1j}},\Theta,||\cdot||)\times N([\frac{\varepsilon}{4||b_j(Z)||_{L_r(P)}}]^{1/s_j},\mathcal{H},||\cdot||_{\mathcal{H}}).$$

This and condition 3.3 imply **(1)**.

For **(2)**. Let $\{\theta_k : k = 1, \ldots, N_1\}$ be an $\delta-$cover for $(\Theta, ||\cdot||)$, and $\{h_k : k = 1, \ldots, N_2\}$ be an $\delta-$cover for $(\mathcal{H}, ||\cdot||_{\mathcal{H}})$. Then under condition 3.2, for any $m_{lcj}(Z, \theta, h)$, there exist $k_1 \in \{1, \ldots, N_1\}$ and $k_2 \in \{1, \ldots, N_2\}$ such that $|m_{lcj}(Z, \theta, h) - m_{lcj}(Z, \theta_{k_1}, h_{k_2})|$ is bounded by

$$\sup_{(\theta', h'): ||\theta' - \theta_{k_1}|| < \delta, ||h' - h_{k_2}||_{\mathcal{H}} < \delta} |m_{lcj}(Z, \theta', h') - m_{lcj}(Z, \theta_{k_1}, h_{k_2})| \equiv b_j(Z, \theta_{k_1}, h_{k_2}, \delta).$$

Hence $m_{lcj}(Z, \theta_{k_1}, h_{k_2}) - b_j(Z, \theta_{k_1}, h_{k_2}, \delta) \leq m_{lcj}(Z, \theta, h) \leq m_{lcj}(Z, \theta_{k_1}, h_{k_2}) + b_j(Z, \theta_{k_1}, h_{k_2}, \delta)$. Again by condition 3.2, $\{E[b_j(Z, \theta_{k_1}, h_{k_2}, \delta)]^r\}^{1/r} \leq K_j \delta^{s_j}$ for all (θ_{k_1}, h_{k_2}) and all positive value $\delta = o(1)$. Therefore, $\{[m_{lcj}(Z, \theta_{k_1}, h_{k_2}) - b_j(Z, \theta_{k_1}, h_{k_2}, \delta), m_{lcj}(Z, \theta_{k_1}, h_{k_2}) + b_j(Z, \theta_{k_1}, h_{k_2}, \delta)] : k_1 \in \{1, \ldots, N_1\}, k_2 \in \{1, \ldots, N_2\}\}$ forms an $\varepsilon = 2K_j \delta^{s_j}-$bracket for $(\mathcal{F}_{2j}, ||\cdot||_{L_r(P)})$. Hence

$$N_{[]}(\varepsilon, \mathcal{F}_{2j}, ||\cdot||_{L_r(P)}) \leq N([\frac{\varepsilon}{2K_j}]^{1/s_j}, \Theta, ||\cdot||) \times N([\frac{\varepsilon}{2K_j}]^{1/s_j}, \mathcal{H}, ||\cdot||_{\mathcal{H}}).$$

This and condition 3.3 imply **(2)**. ■

References

Ai, C., and X. Chen (2002): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," Forthcoming in *Econometrica*.

Akritas, M.G. and I. Van Keilegom (2001): "Nonparametric estimation of the residual distribution," *Scandinavian Journal of Statistics* 28, 549-568.

Andrews, D.W.K. (1994a): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica* 62, 43-72.

Andrews, D.W.K. (1994b): "Empirical process method in econometrics," in *The Handbook of Econometrics*, vol. IV, eds. R.F. Engle and D.L. McFadden. North Holland.

Bickel, P., C. Klaassen, Y. Ritov and J. Wellner (1993): *Efficient and adaptive estimation for semiparametric models*. The John Hopkins University Press, Baltimore and London.

Bliss, R. (1997): "Testing Term Structure Estimation Methods," *Advances in Futures and Options Research* 9, 197-231.

Blundell, R., M. Browning, and I. Crawford (2003): "Nonparametric Engel Curves and Revealed Preference," *Econometrica* 71, 205-240.

Brown, D. and M. Wegkamp (2002): "Weighted Minimum Mean-Square Distance from Independence Estimation," *Econometrica* 70, 2035-2051.

Chaudhuri, P. (1991): "Nonparametric estimates of regression quantiles and their local Bahadur representation," *Annals of Statistics* 19, 760-777.

Chen, S. and S. Khan (2001): "Semiparametric Estimation of a Partially Linear Censored Regression Model," *Econometric Theory* 17, 567-590.

Chen, X. and X. Shen (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica* 66, 289-314.

Giné, E. and J. Zinn (1990): "Bootstrapping General Empirical Measures," *Annals of Probability* 18, 851-869.

Hall, P. (1991): "On convergence rates of suprema," *Probability Theory and Related Fields* 89, 447-455.

Hall, P. and J. Horowitz (1996): "Bootstrap Critical Values for Tests Based on Generalized-Method-Of-Moments Estimators," *Econometrica* 64, 891-916.

Han, H., and E. Tamer (2002): "Inference in Censored Models with Endogenous Regressors," Forthcoming in *Econometrica.*

Horowitz, J. (1992): "A smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica* 60, 505-531.

Horowitz, J. (1998a): "Bootstrap methods for median regression models," *Econometrica* 66, 1327-1352.

Horowitz, J. (1998b): *Semiparametric Methods in Econometrics.* Springer Verlag: Berlin.

Huber, P.J. (1967): "The behavior of maximum likelihood estimates under nonstandard conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 221-233. University of California, Berkeley.

Koenker, R. (1997): "ℓ_1 computation: An interior monologue," in IMS Lecture Notes, Volume 31, pp 15-32.

Koul, H.L. (2001): *Weighted Empirical Processes in Regression and Autoregression Models.* Springer.

Lee, S. (2003): "Efficient Semiparametric Estimation of a Partially Linear Quantile Regression Model," *Econometric Theory* 19, 1-31.

Lewbel, A., and O. Linton (2002): "Nonparametric Censored and Truncated Regression," *Econometrica* 70, 765-780.

Linton, O., R. Chen, N. Wang and W. Härdle (1997): "An analysis of transformations for additive nonparametric regression," *Journal of the American Statistical Association* 92, 1512-1521.

Linton, Mammen, Nielsen, and Tanggaard (2001): "Estimating the Yield Curve by Kernel Smoothing," *Journal of Econometrics* 105/1 185-223.

Manski, C.F. (1975): "The Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics* 3, 205-228.

Manski, C.F. (1983): "Closest Empirical Distribution Estimation," *Econometrica* 51, 305-319.

Manski, C.F. (1994): "Analog Estimation of Econometric Models," in *The Handbook of Econometrics*, vol. IV, eds. R.F. Engle and D.L. McFadden. North Holland.

Newey, W.K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica* 62, 1349-1382.

Newey, W.K., and D.L. McFadden (1994): "Large sample estimation and hypothesis testing," in *The Handbook of Econometrics*, vol. IV, eds. R.F. Engle and D.L. McFadden. North Holland.

Pakes, A., and S. Olley (1995): "A limit theorem for a smooth class of semiparametric estimators," *Journal of Econometrics* 65, 295-332.

Pakes, A., and D. Pollard (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica* 57, 1027-1057.

Powell, J.L. (1984): "Least absolute deviation estimation for the censored regression model," *Journal of Econometrics* 25, 303-325.

Powell, J.L. (1994): "Estimation in semiparametric models," in The *Handbook of Econometrics*, vol. IV. eds R.F. Engle and D.L. McFadden. North Holland.

Robinson, P. (1988): "Root-n-Consistent Semiparametric Regression," *Econometrica*, 56, 931-954.

van der Vaart, A.W. and J.A. Wellner (1996): *Weak convergence and empirical processes.* Springer-Verlag, New York.