EFFICIENT ESTIMATION OF SEMIPARAMETRIC CONDITIONAL MOMENT MODELS WITH POSSIBLY NONSMOOTH RESIDUALS

By

Xiaohong Chen and Demian Pouzo

February 2008

COWLES FOUNDATION DISCUSSION PAPER NO. 1640

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals*

Xiaohong Chen† and Demian Pouzo‡

February 20, 2008

Abstract

For semi/nonparametric conditional moment models containing unknown parametric components (θ) and unknown functions of endogenous variables (h), Newey and Powell (2003) and Ai and Chen (2003) propose sieve minimum distance (SMD) estimation of (θ, h) and derive the large sample properties. This paper greatly extends their results by establishing the followings: (1) The penalized SMD (PSMD) estimator $(\hat{\theta}, \hat{h})$ can simultaneously achieve root-n asymptotic normality of $\hat{\theta}$ and nonparametric optimal convergence rate of $\hat{h}$, allowing for models with possibly nonsmooth residuals and/or noncompact infinite dimensional parameter spaces. (2) A simple weighted bootstrap procedure can consistently estimate the limiting distribution of the PSMD $\hat{\theta}$. (3) The semiparametric efficiency bound results of Ai and Chen (2003) remain valid for conditional models with nonsmooth residuals, and the optimally weighted PSMD estimator achieves the bounds. (4) The profiled optimally weighted PSMD criterion is asymptotically Chi-square distributed, which implies an alternative consistent estimation of confidence region of the efficient PSMD estimator of θ. All the theoretical results are stated in terms of any consistent nonparametric estimator of conditional mean functions. We illustrate our general theories using a partially linear quantile instrumental variables regression, a Monte Carlo study, and an empirical estimation of the shape-invariant quantile Engel curves with endogenous total expenditure.

JEL classification: C14; C22

Keywords: Penalized sieve minimum distance; Nonsmooth generalized residuals; Nonparametric endogeneity; Weighted bootstrap; Semiparametric efficiency; Confidence region; Partially linear quantile IV regression

*First version: September 2005. Revised version: February 2008. We thank R. Blundell for sharing the UK Family Expenditure Survey data set. We acknowledge helpful comments from V. Chernozhukov, J. Horowitz, S. Lee, O. Linton, E. Mammen, W. Newey, J. Powell, P. Robinson, E. Tamer, and participants of August 2006 European ES Summer Meetings, March 2007 Oberwolfach Workshop on Semi/nonparametrics, June 2007 Cemmap Conference on Measurement Matters, and seminars in Northwestern, Vanderbilt, Boston University, Indiana, Yale, Boston College and Toulouse. Chen acknowledges financial support from the National Science Foundation. Any errors are the responsibility of the authors.

†Corresponding author. Department of Economics, Yale University, USA. Tel.: +1 203 432 5852; fax: + 1 203 432 6167. *E-mail address*: xiaohong.chen@yale.edu

‡Department of Economics, New York University, USA. *E-mail address*: dgp219@nyu.edu

1 Introduction

Many semi/nonparametric models are special cases of the following conditional moment models containing unknown functions:

$$E[\rho(Y, X_z; \theta_0, h_{01}(\cdot), ..., h_{0q}(\cdot))|X] = 0, \tag{1.1}$$

in which $Z \equiv (Y', X')'$, Y is a vector of endogenous variables, X_z is a subset of the conditioning variables X, $\rho()$ is a vector of generalized residual functions whose functional forms are known up to the vector of unknown finite dimensional parameters (θ_0) and the vector of unknown real-valued functions ($h_0 \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))$), where the arguments of each real-valued function $h_{0\ell}(\cdot)$ may differ across $\ell = 1, ..., q$, and, in particular, may depend on Y. The conditional distribution, $F_{Y|X}$, of Y given X is not specified; hence the functional form of the conditional expectation, $E[\rho(Z; \theta, h)|X]$, of $\rho(Z; \theta, h)$ given X is unknown.

Assuming that the parameters of interest $\alpha_0 \equiv (\theta_0, h_0)$ are identified by the general conditional moment models (1.1), Newey and Powell (2003) and Ai and Chen (2003) propose Sieve Minimum Distance (hereafter SMD) estimation of (θ_0, h_0). Under the assumptions that the residual function $\rho(Z; \theta, h(\cdot))$ is pointwise Hölder continuous in the parameters $\alpha \equiv (\theta, h) \in \Theta \times \mathcal{H}$, the parameter space $\Theta \times \mathcal{H}$ is compact, and the sieve parameter space $\Theta \times \mathcal{H}_n$ is finite dimensional compact, Newey and Powell (2003) obtain consistency of the SMD estimator of α_0, and Ai and Chen (2003) establish root-n asymptotic normality and efficiency of the SMD estimator of the finite dimensional parameters θ_0.

When some of the $h_{0\ell}(\cdot)$ in the nonparametric conditional moment model $E[\rho(Y, X_z; h_{01}(\cdot), ..., h_{0q}(\cdot))|X] = 0$ depends on the endogenous variables Y, it is difficult to establish convergence rate of any estimator of h_0 under the so-called "strong metric" $||\cdot||_s$, a metric that is not continuous with respect to the quadratic form $E\left[(E[\rho(Z; h(\cdot))|X])' (E[\rho(Z; h(\cdot))|X])\right]$, and the problem becomes a nasty nonlinear ill-posed inverse problem with an unknown operator. In Chen and Pouzo (2007), we propose a penalized SMD (PSMD) estimator, and establish its consistency and convergence rates for h_0 without assuming $||\cdot||_s-$compactness of $\mathcal{H}$ and $\mathcal{H}_n$, and allowing for nonsmooth residual function $\rho(Z; h(\cdot))$ in h.

In this paper, we extend the results of Newey and Powell (2003), Ai and Chen (2003) and Chen and Pouzo (2007) in several directions. First, we show that the PSMD estimator $\widehat{\alpha} \equiv (\widehat{\theta}, \widehat{h})$ can simultaneously

achieve root-n asymptotic normality of $\hat{\theta}$ and optimal convergence rate of $\hat{h}$ (in strong metric $||\cdot||_s$) for the general semiparametric model (1.1), allowing for possibly nonsmooth residuals, and/or possibly noncompact function space ($\mathcal{H}$) and the sieve spaces ($\mathcal{H}_n$) under the strong metric $||\cdot||_s$. It is previously known that sieve M-estimation of semiparametric models (without nonparametric endogeneity) can simultaneously achieve root-n normality of parametric part and optimal convergence rate of nonparametric part; see e.g., Chen and Shen (1998) and Newey et al. (2004). We find that the PSMD estimation of the semiparametric conditional moment model (1.1) (with nonparametric endogeneity) also possesses such a nice property. Second, we show that a simple weighted bootstrap procedure can consistently estimate the limiting distribution of the PSMD $\hat{\theta}$. Previously, Ai and Chen (2003) propose a consistent sieve estimator of the asymptotic variance of $\widehat{\theta}$. Their variance estimator hinges on the differentiability of the residual functions $\rho(Z;\theta,h(\cdot))$ in $\alpha=(\theta,h)$, whereas in our paper $\rho(Z;\theta,h(\cdot))$ could be non-smooth with respect to $\alpha=(\theta,h)$. This is why we propose a weighted bootstrap procedure to consistently estimate the confidence region for any root-n consistent PSMD estimator $\widehat{\theta}$. Third, we show that the semiparametric efficiency bound results of Ai and Chen (2003) remain valid for conditional models with nonsmooth residuals, and establish efficiency of the optimally weighted PSMD procedure. Finally, we show that the profiled optimally weighted PSMD criterion is asymptotically Chi-square distributed. This implies an alternative consistent estimation of confidence region of the efficient PSMD estimator of θ_0 by inverting the profiled optimally weighted criterion function. This alternative confidence region construction avoids the nonparametric estimation of the asymptomatic variance, and it should be easier to compute then the weighted bootstrap procedure. All the general theoretical results are stated in terms of any consistent nonparametric estimator of conditional mean functions $E[\rho(Z;\theta,h)|X=\cdot]$, but we also provide low level sufficient conditions in terms of series least squares (LS) estimator of $E[\rho(Z;\theta,h)|X=\cdot]$. We specialize our theoretical results to an important example of a partially linear quantile instrumental variables (IV) regression: $E[1\{Y_3 \leq \theta_0 Y_1 + h_0(Y_2)\}|X] = \gamma \in (0,1)$. We also present a Monte Carlo study and an empirical estimation of the shape-invariant quantile Engel curves with endogenous total expenditure.

The rest of the paper is organized as follows. Section 2 presents the PSMD estimator $\widehat{\alpha}=(\widehat{\theta},\widehat{h})$, and

its consistency and nonparametric convergence rates. In section 3 we first establish the root-n asymptotic normality of $\widehat{\theta}$. We then show that a weighted bootstrap procedure can consistently estimate the limiting distribution of the $\widehat{\theta}$. In section 4 we first show the validity of the semiparametric efficiency bound, and then the efficiency of the optimally weighted PSMD. In addition, we show that the profile optimally weighted PSMD criterion is asymptotically Chi-squared distributed. Section 5 specializes our general results to a partially linear quantile IV regression example. Section 6 presents a Monte Carlo study and an empirical application. Section 7 briefly concludes. All the proofs and some useful lemmas are gathered in the appendix.

In this paper, we denote $f_{A|B}(a;b)$ ($F_{A|B}(a;b)$) as the conditional probability density (cdf) of random variable A given B evaluated at a and b, and $f_{AB}(a,b)$ ($F_{AB}(a,b)$) the joint density (cdf) of the random variables A and B. Denote $||\cdot||_E$ as the Euclidian norm. Let $L^p(\Omega, d\mu)$ be the space of measurable functions with $||f||_{L^p(\Omega,d\mu)} \equiv \{\int_\Omega |f(t)|^p d\mu(t)\}^{1/p} < \infty$, where Ω is the support of the sigma-finite positive measure $d\mu$ (sometimes $L^p(d\mu)$ and $||f||_{L^p(d\mu)}$ are used for simplicity). For any sequences $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means that there exists two constants $0 < c_1 \leq c_2 < \infty$ such that $c_1 a_n \leq b_n \leq c_2 a_n$, and $a_n = O_P(b_n)$ means that a_n is bounded in probability at rate b_n, i.e., $\Pr(a_n/b_n \geq M) \to 0$ as n and M go to infinity.

2 The Penalized SMD estimator

The semiparametric conditional moment model (1.1) can be equivalently expressed as $m(X, \alpha_0) = 0$ $a.s. - X$, where $m(X,\alpha) \equiv E[\rho(Y, X_z; \alpha)|X] = \int \rho(Y, X_z; \alpha) dF_{Y|X}(y)$ and $\alpha_0 \equiv (\theta_0, h_0) \in \mathcal{A} \equiv \Theta \times \mathcal{H}$. Following Chen and Pouzo (2007), we propose the *penalized* SMD (PSMD) estimator

$$\widehat{\alpha}_n \equiv (\widehat{\theta}_n, \widehat{h}_n) = \arg \inf_{\alpha \in \mathcal{A}_n} \left\{ \frac{1}{n} \sum_{i=1}^n \widehat{m}(X_i, \alpha)' [\widehat{\Sigma}(X_i)]^{-1} \widehat{m}(X_i, \alpha) + \lambda_n P(h) \right\}, \tag{2.1}$$

where $\mathcal{A}_n \equiv \Theta \times \mathcal{H}_n$ is a sieve for $\mathcal{A} \equiv \Theta \times \mathcal{H}$, $\widehat{m}(X,\alpha)$ is any nonparametric consistent estimator of $m(X,\alpha)$, $\widehat{\Sigma}(X)$ is any consistent estimator of a positive definite weighting matrix $\Sigma(X)$, $\lambda_n \geq 0$ is

a penalization tuning parameter such that $\lambda_n = o(1)$, and $P(h) \geq 0$ is a penalization function. See Chen and Pouzo (2007) for a more detailed presentation of the PSMD estimator, and the comparison of a finite dimensional sieve PSMD procedure vs an infinite dimensional sieve PSMD procedure. Here we focus on the finite dimensional sieve PSMD method only.

In this paper, we establish consistency and convergence rate of the PSMD estimator $\widehat{\alpha}_n$, the root-n normality, semiparametric efficiency and confidence region of $\widehat{\theta}_n$ under conditions that are satisfied by any nonparametric estimators $\widehat{m}(X, \alpha)$ and $\widehat{\Sigma}(X)$ of $m(X, \alpha)$ and $\Sigma(X)$ respectively. In addition, we also provide relatively low level sufficient conditions when $\widehat{m}(X, \alpha)$ is a series least squares (LS) estimator, as defined in (2.2):

$$\widehat{m}(X, \alpha) = p^{J_n}(X)'(P'P)^{-} \sum_{i=1}^{n} p^{J_n}(X_i)\rho(Z_i, \alpha), \tag{2.2}$$

where $\{p_j()\}_{j=1}^{\infty}$ is a sequence of known basis functions that can approximate any square integrable functions of X well, $J_n \to \infty$ slowly as $n \to \infty$, $p^{J_n}(X) = (p_1(X), ..., p_{J_n}(X))'$, $P = (p^{J_n}(X_1), ..., p^{J_n}(X_n))'$, and $(P'P)^{-}$ is the generalized inverse of the matrix $P'P$. To simplify presentation, we let $p^{J_n}(X)$ be a tensor-product linear sieve basis, which is the product of univariate linear sieves. For example, let $\{\phi_{i_j} : i_j = 1, ..., J_{j,n}\}$ denote a B-spline (wavelet, Fourier series, power series) basis for $L^2(\mathcal{X}_j, leb.)$, with $\mathcal{X}_j$ a compact interval in $\mathcal{R}$, $1 \leq j \leq d_x$. Then the tensor product $\{\prod_{j=1}^{d_x} \phi_{i_j}(X_j) : i_j = 1, ..., J_{j,n}, j = 1, ..., d_x\}$ is a B-spline (wavelet, Fourier series, power series) basis for $L^2(\mathcal{X}, leb.)$, with $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_{d_x}$. Clearly the number of terms in the tensor-product sieve $p^{J_n}(X)$ is given by $J_n = \prod_{j=1}^{d_x} J_{j,n}$. See Newey (1997), Huang (1998) and Chen (2007) for more details about tensor-product B-splines and other linear sieves.

2.1 Consistency

In this subsection we present some consistency results of the PSMD estimator. We first impose some regularity conditions.

Assumption 2.1. *(i) $\mathcal{A} \equiv \Theta \times \mathcal{H}$, Θ is a compact convex subset of $\mathcal{R}^{d_\theta}$, and $\mathcal{H} \subseteq \mathbf{H}$, $\mathbf{H} \equiv \mathbf{H}^1 \times \cdots \times \mathbf{H}^q$ is a separable Banach space under the metric $\|h\|_c \equiv \sum_{\ell=1}^{q} \|h_\ell\|_{c,\ell}$; (ii) $E[\rho(Z, \alpha_0)|X] = 0$, and $\|\theta_0 - \theta\|_E + \|h_0 - h\|_c = 0$ for any $\alpha = (\theta, h) \in \mathcal{A}$ with $E[\rho(Z, \alpha)|X] = 0$.*

Assumption 2.2. *$\mathcal{A}_k \equiv \Theta \times \mathcal{H}_k$, $k \geq 1$, are the sieve spaces satisfying $\mathcal{H}_k \subseteq \mathcal{H}_{k+1} \subseteq \mathcal{H}$, and there exists $\Pi_n h_0 \in \mathcal{H}_{k(n)}$ such that $||\Pi_{k(n)} h_0 - h_0||_c = o(1)$.*

Denote $\|\alpha\|_c \equiv \|\theta\|_E + \|h\|_c$ on $\mathcal{A} \equiv \Theta \times \mathcal{H}$, and

$$\widehat{Q}_n(\alpha) \equiv \frac{1}{n} \sum_{i=1}^{n} \widehat{m}(X_i, \alpha)' \{\widehat{\Sigma}(X_i)\}^{-1} \widehat{m}(X_i, \alpha) + \lambda_n P(h).$$

Assumption 2.3. *either (a) or (b) holds: (a) $\mathcal{A}_k$ is compact under $||\cdot||_c$, and for any data $\{Z_i\}_{i=1}^n$, $\widehat{Q}_n(\alpha)$ is lower semicontinuous (in $||\cdot||_c$) on $\mathcal{A}_{k(n)}$. (b) $\mathcal{A}_k$ is a bounded, closed and convex subset of a reflexive Banach space $(\Theta \times \mathbf{H}, ||\cdot||_c)$, and for any data $\{Z_i\}_{i=1}^n$, $\widehat{Q}_n(\alpha)$ is weak sequentially lower semicontinuous on $\mathcal{A}_{k(n)}$.*

Assumption 2.4. *(i) $E[m(X, \alpha)' \Sigma(X)^{-1} m(X, \alpha)]$ is continuous at α_0 under $\|\cdot\|_c$; (ii) $\lambda_n P(\cdot) \geq 0$, and is continuous at h_0, and $P(h_0) < \infty$.*

Assumption 2.1(i) defines the parameter space and assumption 2.1(ii) assumes that α_0 is identified (up to an equivalent class under the metric $||\cdot||_c$). The identification condition is a high level assumption and has to be verified in each application. Assumption 2.2 is effectively the definition of a sieve space. Assumption 2.3 provides some sufficient conditions to ensure the PSMD estimator $\widehat{\alpha}_n$ exists and is well defined. The following lemma is a minor modification of Lemma B.1 and Remark B.1 in Chen and Pouzo (2007) hence we omit its proof.

Lemma 2.1. *Let $\widehat{\alpha}_n$ be the PSMD estimator (2.1) with $\lambda_n \geq 0$, $\lambda_n = o(1)$, and $\{(Y_i, X_i)\}_{i=1}^n$ be a strictly stationary sample. Suppose that assumptions 2.1, 2.2, 2.3, 2.4 and the following conditions (2.1.1) and (2.1.2) hold:*

(2.1.1) there are a function $\delta(\lambda, k)$ and a nondecreasing function $g(\varepsilon) \geq 0$ such that for any $k \geq 1$, any $\lambda \geq 0$, and any $\varepsilon > 0$,

$$\inf_{\alpha \in \mathcal{A}_k : ||\alpha - \alpha_0||_c \geq \varepsilon} \left\{ E\left[m(X, \alpha)' \Sigma(X)^{-1} m(X, \alpha)\right] + \lambda [P(h) - P(h_0)] \right\} \geq \delta(\lambda, k) g(\varepsilon) > 0.$$

(2.1.2) (i) $E\left[m(X,\theta_0,\Pi_{k(n)}h_0)'\Sigma(X)^{-1}m(X,\theta_0,\Pi_{k(n)}h_0)\right]+\lambda_n[P(\Pi_{k(n)}h_0)-P(h_0)]=o\left(\delta(\lambda_n,k(n))\right)$*;*

$$(ii)\quad \sup_{\alpha\in\mathcal{A}_{k(n)}}\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{m}(X_i,\alpha)'\widehat{\Sigma}(X_i)^{-1}\widehat{m}(X_i,\alpha)-E\left[m(X,\alpha)'\Sigma(X)^{-1}m(X,\alpha)\right]\right|=o_P(\delta(\lambda_n,k(n))).$$

Then: $||\widehat{\alpha}_n-\alpha_0||_c=o_P(1)$.

Condition (2.1.1) is the so-called "identifiable uniqueness" condition over the sieve space. It allows for both the "well-posed" case (in which $\liminf_{k\to\infty}\delta(\lambda,k)>0$, i.e., $||\alpha-\alpha_0||_c$ is continuous with respect to $E\left[m(X,\alpha)'\Sigma(X)^{-1}m(X,\alpha)\right]$), and the "ill-posed" case (in which $\liminf_{k\to\infty}\delta(\lambda,k)=0$, i.e., $||\alpha-\alpha_0||_c$ is *not* continuous with respect to $E\left[m(X,\alpha)'\Sigma(X)^{-1}m(X,\alpha)\right]$). For the "well-posed" case, we have $\delta(\lambda_n,k(n))=O(1)$, condition (2.1.2)(i) is automatically satisfied under assumption 2.4, and condition(2.1.2)(ii) becomes the standard assumption of uniform convergence over the sieve space. See Chen and Pouzo (2007) for low level sufficient conditions for consistency when the problems could be "well-posed" or "ill-posed".

2.2 Convergence Rates

In the rest of the paper, we let $\|\cdot\|_s$ denote another metric on the infinite-dimensional function space $\mathcal{H}$ that is weaker than the norm $||\cdot||_c$ (i.e., $||h||_s\leq||h||_c$ for all $h\in\mathcal{H}$). In this section we study convergence rate under the metric $||\cdot||_s$. Given the consistency results stated above, we can now restrict our attention to a shrinking $||\cdot||_c-$neighborhood around α_0. Let $\mathcal{A}_{os}\equiv\{\alpha\in\mathcal{A}:||\alpha-\alpha_0||_c=o(1),||h||_c\leq c\}$ and $\mathcal{A}_{osn}\equiv\{\alpha\in\mathcal{A}_n:||\alpha-\Pi_n\alpha_0||_c=o(1),||h||_c\leq c\}$. Then, for the purpose of establishing a rate of convergence under the $||\alpha||_s\equiv\|\theta\|_E+||h||_s$ metric, we can treat $\mathcal{A}_{os}$ as the new parameter space and $\mathcal{A}_{osn}$ as its sieve space.

In order to establish the convergence rate under $||\cdot||_s$ we first establish the rate under a weaker pseudo-metric $||\cdot||$. We define the first pathwise derivative at the direction $[h-h_0]$ evaluated at h_0 as

$$\begin{aligned}\frac{dm(X,\alpha_0)}{d\alpha}[\alpha-\alpha_0] &\equiv \left.\frac{dE[\rho(Z,(1-\tau)\alpha_0+\tau\alpha)|X]}{d\tau}\right|_{\tau=0} \quad a.s.\ \mathcal{X}. \\ &= \frac{dm(X,\alpha_0)}{d\theta'}(\theta-\theta_0)+\frac{dm(X,\alpha_0)}{dh}[h-h_0]\end{aligned} \tag{2.3}$$

Following Ai and Chen (2003), we define the pseudo-metric $||\alpha_1 - \alpha_2||$ for any α_1, $\alpha_2 \in \mathcal{A}_{os}$ as

$$||\alpha_1 - \alpha_2||^2 \equiv E\left[\left(\frac{dm(X,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\right)' \Sigma(X)^{-1} \left(\frac{dm(X,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\right)\right]. \tag{2.4}$$

Likewise we define

$$||h_1 - h_2||^2 \equiv E\left[\left(\frac{dm(X,\alpha_0)}{dh}[h_1 - h_2]\right)' \Sigma(X)^{-1} \left(\frac{dm(X,\alpha_0)}{dh}[h_1 - h_2]\right)\right].$$

We impose the following additional assumptions.

Assumption 2.5. *(i)* $\{(Y_i', X_i')\}_{i=1}^n$ *is an i.i.d. sample; (ii)* $\mathcal{X}$ *is a compact connected subset of* $\mathcal{R}^{d_x}$ *with Lipschitz continuous boundary, and* f_X *is bounded and bounded away from zero over* $\mathcal{X}$*.*

Assumption 2.6. *(i)* $\sup_{x\in\mathcal{X}} \left|\widehat{\Sigma}(x) - \Sigma(x)\right| = o_P(1)$*; (ii)* $\Sigma(X)$ *is positive definite, and its smallest and largest eigenvalues are finite positive uniformly over* $\mathcal{X}$*. (iii) with probability approaching one,* $\widehat{\Sigma}(X)$ *is positive definite, and its smallest and largest eigenvalues are finite positive uniformly over* $\mathcal{X}$*.*

Assumption 2.7. *(i)* $\sup_{\alpha\in\mathcal{A}_n} \sqrt{E\left[||\widehat{m}(X,\alpha) - m(X,\alpha)||_E^2\right]} = O_p(\delta_{m,n}^*) = o_P(1)$*; (ii)* $E\left[||\widehat{m}(X,\alpha)||_E^2\right] \asymp n^{-1}\sum_{i=1}^n ||\widehat{m}(X_i,\alpha)||_E^2$ *uniformly over* $\alpha \in \mathcal{A}_n$*.*

Assumption 2.8. *For any* $\alpha \in \mathcal{A}_{os}$ *(i)* $m(X,\alpha)$ *is continuously pathwise differentiable with respect to* α*; (ii)* $E\left[||m(X,\alpha)||_E^2\right] \asymp ||\alpha - \alpha_0||^2$*, and* $||\alpha - \alpha_0|| \leq K \times ||\alpha - \alpha_0||_s$*; (iii)* $\lambda_n P(h)$ *is continuously pathwise differentiable with respect to* h*.*

Assumption 2.7 is a high level condition imposed on the nonparametric estimator for $m(X,\alpha)$. Nevertheless, it is satisfied when $\widehat{m}(X,\alpha)$ is the series LS estimator (2.2); see Chen and Pouzo (2007). It can be shown to hold for kernel or local linear regression estimator as well. The following lemma is a minor modification of Theorem 4.1 in Chen and Pouzo (2007) hence we omit its proof.

Lemma 2.2. *Let* $\widehat{\alpha}_n$ *be the PSMD estimator (2.1) with* $\lambda_n \geq 0$, $\lambda_n = o(1)$*. Suppose that* $||\widehat{\alpha}_n - \alpha_0||_s =$

$o_P(1)$, *assumptions 2.1, 2.2, 2.4, 2.5, 2.6, 2.7 and 2.8 hold. Then:*

$$\begin{aligned} ||\widehat{\alpha}_n - \Pi_n\alpha_0|| &= O_P\left(\max\left\{\delta^*_{m,n}, o(\sqrt{\lambda_n}), ||h_0 - \Pi_n h_0||\right\}\right) = O_P\left(\delta^*_n\right) = o_P(1), \\ ||\widehat{\alpha}_n - \alpha_0|| &\leq ||\widehat{\alpha}_n - \Pi_n\alpha_0|| + ||h_0 - \Pi_n h_0|| = O_P\left(\delta^*_n\right). \end{aligned}$$

As pointed out in Ai and Chen (2003), to establish root-n asymptotic normality of $\widehat{\theta}$, it suffices to have the nonparametric convergence rate faster than $n^{-1/4}$ under the weaker pseudo-metric, $||\widehat{\alpha}_n - \alpha_0|| = O_P\left(\delta^*_n\right) = o_P(n^{-1/4})$. Nevertheless, in some applications such as the estimation of the system of shape-invariant Engel curves in Blundell et al. (2007), one would like to have the property that an estimator $\widehat{\alpha}_n = (\widehat{\theta}, \widehat{h})$ can achieve the optimal rates for both the parametric part and the unknown functions simultaneously. In the following we shall show that the PSMD estimator possesses such a nice property.

Following Ai and Chen (2003) we define $\overline{\mathbf{V}}$ as the closure of the linear span of $\mathcal{A} - \{\alpha_0\}$ under the metric $||\cdot||$. For any $v_1, v_2 \in \overline{\mathbf{V}}$, we define an inner product corresponding to the metric $||\cdot||$:

$$\langle v_1, v_2\rangle = E\left[\left(\frac{dm(X,\alpha_0)}{d\alpha}[v_1]\right)' \Sigma(X)^{-1}\left(\frac{dm(X,\alpha_0)}{d\alpha}[v_2]\right)\right],$$

thus $(\overline{\mathbf{V}}, \langle\cdot\rangle)$ is a Hilbert space, with $\overline{\mathbf{V}} = \mathcal{R}^{d_\theta} \times \overline{\mathcal{W}}$ and $\overline{\mathcal{W}} \equiv \overline{\mathcal{H}} - \{h_0\}$. Let $h - h_0 = -w(\theta - \theta_0)$, then we write $\frac{dm(X,\alpha_0)}{d\alpha}[\alpha - \alpha_0] \equiv \left(\frac{dm(X,\alpha_0)}{d\theta'} - \frac{dm(X,\alpha_0)}{dh}[w]\right)(\theta - \theta_0) \equiv D_w(X)(\theta - \theta_0)$. For each component θ_j (of θ), $j = 1, ..., d_\theta$, let $w^*_j \in \overline{\mathcal{W}}$ denote the solution to

$$\inf_{w_j \in \overline{\mathcal{W}}} E\left[\left(\frac{dm(X,\alpha_0)}{d\theta_j} - \frac{dm(X,\alpha_0)}{dh}[w_j]\right)' \Sigma(X)^{-1}\left(\frac{dm(X,\alpha_0)}{d\theta_j} - \frac{dm(X,\alpha_0)}{dh}[w_j]\right)\right]. \tag{2.5}$$

Define $w^* = (w^*_1, ..., w^*_{d_\theta})$, $\frac{dm(X,\alpha_0)}{dh}[w^*] = \left(\frac{dm(X,\alpha_0)}{dh}[w^*_1], ..., \frac{dm(X,\alpha_0)}{dh}[w^*_{d_\theta}]\right)$, and

$$D_{w^*}(X) \equiv \frac{dm(X,\alpha_0)}{d\theta'} - \frac{dm(X,\alpha_0)}{dh}[w^*].$$

Assumption 2.9. *(i) $E[D_{w^*}(X)'[\Sigma(X)]^{-1}D_{w^*}(X)]$ is positive-definite; (ii) $\theta_0 \in int(\Theta)$.*

Assumption 2.10. *(i) $\mathcal{H} \subseteq \mathbf{H}$, $(\mathbf{H}, ||\cdot||_s)$ is a Hilbert space with $\langle\cdot,\cdot\rangle_s$ the inner product and $\{q_j\}_{j=1}^{\infty}$ a*

Riesz basis; (ii) $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$.

Assumption 2.10(i) suggests that $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ is a natural sieve for $\mathcal{H}$. For example, if $\mathcal{H} \subseteq W_2^{\gamma_h}([0,1]^d, leb)$ (a Sobolev space), then assumption 2.10 is satisfied with $(\mathbf{H}, ||\cdot||_s) = (L^2([0,1]^d, leb), ||\cdot||_{L^2(leb)})$, and spline or wavelet or power series or Fourier series bases as $\{q_j\}_{j=1}^{\infty}$.

Assumption 2.11. *There are finite constants* c, $C > 0$ *and a non-increasing positive sequence* $\{b_j \asymp \varphi(\nu_j^{-2})\}_{j=1}^{\infty}$ *such that: (i)* $||h||^2 \geq c\sum_{j=1}^{\infty} b_j |\langle h, q_j\rangle_s|^2$ *for all* $h \in \mathcal{H}_{osn}$*; (ii)* $C\sum_j b_j |\langle h_0 - \Pi_n h_0, q_j\rangle_s|^2 \geq ||h_0 - \Pi_n h_0||^2$.

The following lemma is a direct consequence of Theorem 4.2 and Corollary 5.1 of Chen and Pouzo (2007) and Lemma B.1 of Ai and Chen (2003); hence we omit its proof.

Lemma 2.3. *Let* $\widehat{\alpha}_n$ *be the PSMD estimator (2.1) with* $\lambda_n \geq 0$, $\lambda_n = o(1)$. *Suppose that all the assumptions of Lemma 2.2 hold. If assumption 2.9 holds, then:*

(1) $||\widehat{\theta}_n - \theta_0||_E = O_P(||\widehat{\alpha}_n - \alpha_0||) = O_P(\delta_n^*)$.

(2) If $E\left[tr\left\{\left(\frac{dm(X,\alpha_0)}{dh}[w^*]\right)'\left(\frac{dm(X,\alpha_0)}{dh}[w^*]\right)\right\}\right]$ *is finite, then:* $||\widehat{h}_n - h_0|| = O_P(||\widehat{\alpha}_n - \alpha_0||) = O_P(\delta_n^*)$ *and* $||\widehat{h}_n - \Pi_n h_0|| = O_P(\delta_n^*)$

(3) Further, if assumptions 2.10 and 2.11 hold, and $\max\left\{\delta_{m,n}^*, o(\sqrt{\lambda_n})\right\} = \delta_{m,n}^*$*, then:* $\delta_n^* \asymp \delta_{m,n}^*$ *and*

$$||\widehat{h}_n - h_0||_s = O_P\left(||h_0 - \Pi_n h_0||_s + \frac{\delta_{m,n}^*}{\sqrt{b_{k(n)}}}\right), \; ||\widehat{\alpha}_n - \alpha_0||_s = O_P\left(||\widehat{h}_n - h_0||_s\right).$$

2.2.1 Convergence rates when $\widehat{m}$ is a series LS estimator

We now provide some low level sufficient conditions for assumption 2.7 when $\widehat{m}(X, \alpha)$ is the series LS estimator of $m(X, \alpha)$ given in (2.2). In the following we denote $\zeta_n \equiv \sup_x ||p^{J_n}(x)||_E$.

Assumption 2.12. *(i) The smallest and largest eigenvalues of* $E[p^{J_n}(X)p^{J_n}(X)']$ *are bounded and bounded away from zero for all* J_n*; (ii) either* $J_n\zeta_n^2 = o(n)$ *or* $J_n \log(J_n) = o(n)$ *for P-spline sieve* $p^{J_n}(X)$.

Assumption 2.13. *(i)* $\sup_{\alpha\in\mathcal{A}_n} \sup_x Var[\rho(Z,\alpha)|X = x] \leq K < \infty$*; (ii) for any* $g \in \{m(\cdot,\alpha) : \alpha \in \mathcal{A}_n\}$*, there is* $p^{J_n}(X)'\pi$ *such that, uniformly over* $\alpha \in \mathcal{A}_n$*, either (a) or (b) holds: (a)* $\sup_x |g(x) - p^{J_n}(x)'\pi| = O(b_{m,J_n}) = o(1)$*; (b)* $E\{[g(X) - p^{J_n}(X)'\pi]^2\} = O(b_{m,J_n}^2)$ *for* $p^{J_n}(X)$ *sieve with* $\zeta_n = O(J_n^{1/2})$.

Assumption 2.13(ii) is satisfied by typical smooth function classes of $\{m(\cdot,\alpha):\alpha\in\mathcal{A}_n\}$. For example, if $\{m(\cdot,\alpha):\alpha\in\mathcal{A}_n\}$ is a subset of $\Lambda_c^{\gamma_m}(\mathcal{X})$, $\gamma_m > d_x/2$, (or $W_{2,c}^{\gamma_m}(\mathcal{X},leb.)$), then assumption 2.13(ii) (a) (or (b)) holds with $b_{m,J_n} = J_n^{-r_m}$ and $r_m = \gamma_m/d_x$. Denote $||\widehat{\alpha}_n - \alpha_0||_s \equiv O_P(\delta_{s,n}^*)$ and $||\widehat{\alpha}_n - \alpha_0|| \equiv O_P(\delta_n^*)$. The following lemma summaries Lemma B.3 and Corollary 5.1 of Chen and Pouzo (2007); hence we omit its proof.

Lemma 2.4. *(1) Let $\widehat{m}$ be the series LS estimator given in (2.2) with P-splines, cosine/sine or wavelets as the basis $p^{J_n}(X)$. Suppose that assumptions 2.5, 2.12 and 2.13 hold. Then: Assumption 2.7 is satisfied with $\delta_{m,n}^* = \max\{\sqrt{\frac{J_n}{n}}, b_{m,J_n}\}$.*

(2) Let $\widehat{\alpha}_n$ be the PSMD estimator (2.1) with $\lambda_n \geq 0$, $\lambda_n = o(1)$ and $\widehat{m}$ the series LS estimator. Suppose that all the assumptions of Lemma 2.3(3) hold. Let $||h_0 - \Pi_n h_0||_s = O(\{\nu_{k(n)}\}^{-\gamma_h})$ for a finite $\gamma_h > 0$ and an increasing positive sequence $\{\nu_j\}_{j=1}^{\infty}$. If $\delta_{m,n}^* = \max\{\sqrt{\frac{J_n}{n}}, b_{m,J_n}\} = \sqrt{\frac{J_n}{n}} \to 0$ and $\lim_{n\to\infty}\{J_n/k(n)\} = c \in (1,\infty)$, then:

$$\delta_{s,n}^* = O\left(\{\nu_{k(n)}\}^{-\gamma_h} + \sqrt{\frac{k(n)}{n \times \varphi(\nu_{k(n)}^{-2})}}\right) \quad \text{and} \quad \delta_n^* \asymp \delta_{m,n}^* \asymp \sqrt{\frac{k(n)}{n}}.$$

(2.i) Mildly ill-posed case: if $\varphi(\tau) = \tau^a$ for some $a \geq 0$ and $\nu_k \asymp k^{1/d}$, then: $\delta_{s,n}^* = O\left(n^{-\frac{\gamma_h}{2(\gamma_h+a)+d}}\right)$ and $\delta_n^* = O\left(n^{-\frac{\gamma_h+a}{2(\gamma_h+a)+d}}\right)$ provided $k(n) = O\left(n^{\frac{d}{2(\gamma_h+a)+d}}\right)$.

(2.ii) Severely ill-posed case: if $\varphi(\tau) = \exp\{-\tau^{-a/2}\}$ for some $a > 0$ and $\nu_k \asymp k^{1/d}$, then: $\delta_{s,n}^* = O\left([\ln(n)]^{-\gamma_h/a}\right)$ and $\delta_n^* = O\left(\sqrt{\frac{[\ln(n)]^{d/a}}{n}}\right)$ provided $k(n) = O\left([\ln(n)]^{d/a}\right)$.

3 Asymptotic Normality and Weighted Bootstrap

In this section we first establish root-n asymptotic normality of the PSMD estimator $\widehat{\theta}$, which extends the normality result of the SMD estimator of θ_0 derived in Ai and Chen (2003) to allow for nonsmooth generalized residual functions $\rho(Z;\alpha)$, and penalized SMD procedure with any nonparametric estimators $\widehat{m}(X,\alpha)$ and $\widehat{\Sigma}(X)$. We then provide a new weighted bootstrap procedure to consistently approximate the limiting distribution of $\widehat{\theta}$.

3.1 Root-n normality of $\widehat{\theta}$

Under assumption 2.9, for any non-zero $\lambda \in \mathcal{R}^{d_\theta}$, there is a $v^* \in \overline{\mathbf{V}}$ such that $\lambda'(\widehat{\theta}_n - \theta_0) = \langle v^*, \widehat{\alpha}_n - \alpha_0 \rangle$, i.e., $v^* = (v^*_\theta, v^*_h)$ is the Riesz representer of $\lambda'(\widehat{\theta}_n - \theta_0)$, with $v^*_\theta \equiv (E[D_{w^*}(X)'[\Sigma(X)]^{-1}D_{w^*}(X)])^{-1}\lambda$ and $v^*_h = -w^* \times v^*_\theta$. We impose the following extra assumptions to derive root-n asymptotic normality of $\lambda'(\widehat{\theta}_n - \theta_0)$. Denote $\mathcal{N}_{0n} \equiv \left\{\alpha \in \mathcal{A}_{osn} : ||\alpha - \alpha_0|| = O(\delta^*_n), ||\alpha - \alpha_0||_s = O(\delta^*_{s,n})\right\}$.

Assumption 3.1. *(i)* $\sup_{\alpha \in \mathcal{N}_{0n}} n^{-1}\sum_{i=1}^n \|\widehat{m}(X_i,\alpha) - \widehat{m}(X_i,\alpha_0) - m(X_i,\alpha)\|^2_E = o_p(n^{-1})$; *(ii)* $\delta^*_n = o(n^{-1/4})$, $\delta^*_{m,n} = o(n^{-1/4})$; *(iii)* $n^{-1/2}\sum_{i=1}^n \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v^*]\right)' \Sigma(X_i)^{-1}\{\rho(Z_i,\alpha_0) - \widehat{m}(X_i,\alpha_0)\} = o_P(1)$.

Assumption 3.1 is a high level one, but it is satisfied when $m(X,\alpha)$ is estimated by a series LS estimator $\widehat{m}(X,\alpha)$ (see Ai and Chen (2003)) or a kernel estimator (see the 1998 working paper version of Ai and Chen (2003)). The rates $\delta^*_{m,n}$, δ^*_n and $\delta^*_{s,n}$ are obviously linked; see Lemmas 2.3 and 2.4. In particular, under assumption 2.8, we have $\delta^*_n \asymp \delta^*_{m,n}$ and $\delta^*_n = o(\delta^*_{s,n})$, and how fast the difference of these two last rates grows depends on the degree of ill-posedness. In the so-called "mildly ill-posed" case, roughly speaking, the weaker norm is "polynomial order" faster than the strong norm, whereas in the "severely ill-posed" case the difference is exponential. In the following we denote $\sup_x |\widehat{\Sigma}(x) - \Sigma(x)| \equiv O_P(\delta^*_{\Sigma,n})$.

Assumption 3.2. *(i)* $\delta^*_{\Sigma,n} \times \delta^*_n = o(n^{-1/2})$; *(ii)* $\Sigma_0(X) \equiv Var[\rho(Z,\alpha_0)|X]$ *is positive definite for all* $X \in \mathcal{X}$.

Assumption 3.3. *There is* $v^*_n \equiv (v^*_\theta, -\Pi_n w^* \times v^*_\theta) \in \mathcal{A}_n \setminus \{\alpha_0\}$ *such that* $||v^*_n - v^*|| \times \delta^*_n = o_p(n^{-1/2})$.

Assumption 3.4. *(i) The second pathwise derivative of* $m(X,\alpha)$ *wrt* α *exist for all* $\alpha \in \mathcal{N}_{0n}$, *and* $E\left(\sup_{\alpha \in \mathcal{N}_{0n}} \left|\frac{d^2 m(X,\alpha)}{d\alpha d\alpha}[v^*_n, v^*_n]\right|^2\right) < \infty$; *(ii)* $E\left[\left\|\frac{dm(X,\alpha)}{d\alpha}[v^*_n] - \frac{dm(X,\alpha_0)}{d\alpha}[v^*_n]\right\|^2_E\right] = o_p(n^{-1/2})$ *uniformly over* $\alpha \in \mathcal{N}_{0n}$; *(iii)* $\frac{dm(\cdot,\alpha)}{d\alpha}[v^*_n] \in \Lambda^{\gamma'_m}_c(\mathcal{X})$ *with* $r'_m \equiv \gamma'_m/d_x > 1/2$; *(iv)* $\{m(\cdot,\alpha) : \alpha \in \mathcal{N}_0\}$ *is a Donsker class in* $L^2(\mathcal{X})$.

Assumption 3.4(iv) is satisfied by typical smooth function classes, for example,, it is satisfied by the Holder class $\{m(\cdot,\alpha) : \alpha \in \mathcal{N}_{0n}\} \subseteq \Lambda^{\gamma_m}_c(\mathcal{X})$ with $r_m \equiv \gamma_m/d_x > 1/2$.

Assumption 3.5. *For all* $\alpha \in \mathcal{N}_{0n}$, $\overline{\alpha} \in \mathcal{N}_{0n}$,

$$E\left[\left(\frac{dm(X,\alpha_0)}{d\alpha}[v^*]\right)' \Sigma(X)^{-1}\left(\frac{dm(X,\alpha)}{d\alpha}[\overline{\alpha}-\alpha_0]-\frac{dm(X,\alpha_0)}{d\alpha}[\overline{\alpha}-\alpha_0]\right)\right] = o_p(n^{-1/2}).$$

Various low-level sufficient conditions for assumptions 3.4(ii) and 3.5 can be easily obtained in terms of the "strong" norm; hence these assumptions will, in general, be difficult to check for "severely" ill-posed highly nonlinear problems. If $m(X,\alpha)$ is linear in α then these assumptions are redundant.

Assumption 3.6. *Either (a) or (b) holds: (a)* $\lambda_n = 0$*; (b)* $\lambda_n = o_p(n^{-1/2})$, $\frac{d^j P(h)}{dh^j}[v_n^*]$ *exists and* $E\left(\sup_{\alpha\in\mathcal{N}_{0n}}\left|\frac{d^j P(h)}{dh^j}[v_n^*]\right|\right) < \infty$ *for* $j = 1, 2$.

Theorem 3.1. *Let* $\widehat{\alpha}_n$ *be the PSMD estimator (2.1) with* $\lambda_n \geq 0$, $\lambda_n = o(1)$. *Suppose that all the assumptions of Lemma 2.2 hold. If assumptions 2.9, and 3.1 - 3.6 hold. Then:* $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, V^{-1})$, *where*

$$V^{-1} \equiv \begin{bmatrix} \left(E\left[D_{w^*}(X)'\left[\Sigma(X)\right]^{-1} D_{w^*}(X)\right]\right)^{-1} \times \\ \left(E\left[D_{w^*}(X)'\left[\Sigma(X)\right]^{-1}\Sigma_0(X)\left[\Sigma(X)\right]^{-1} D_{w^*}(X)\right]\right) \\ \times \left(E\left[D_{w^*}(X)'\left[\Sigma(X)\right]^{-1} D_{w^*}(X)\right]\right)^{-1} \end{bmatrix}. \tag{3.1}$$

3.1.1 Root-n normality when $\widehat{m}$ is a series LS estimator

In this subsection we provide some low level sufficient conditions for assumption 3.1 when $\widehat{m}(X,\alpha)$ is the series LS estimator of $m(X,\alpha)$ given in (2.2). For this case, assumption 3.1(iii) is trivially satisfied (see corollary C.3(iii) in Ai and Chen (2003)).

Assumption 3.7. *(i) There exists a measurable function* $b(X)$ *with* $E[|b(X)|] < \infty$ *and constant* $\kappa \in (0, 1]$ *and* $r \geq 1$ *such that for all* $\delta > 0$ *and* $\alpha, \alpha' \in \mathcal{N}_{0n}$

$$\sup_{||\alpha-\alpha'||_s \leq \delta} \int |\rho(z,\alpha) - \rho(z,\alpha')|^r dF_{Y|X=x}(y) \leq b(x)^r \delta^{r\kappa};$$

(ii) exists a measurable $C(Z)$ *such that* $|\rho(Z,\alpha)| \leq C(Z)$ *and* $|E[C(Z)^2|X]| \leq M < \infty$.

In the following we denote $\widetilde{m}(X,\alpha) = p^{J_n}(X)'(P'P)^- \sum_{i=1}^n p^{J_n}(X_i) m(X_i,\alpha)$ as the LS projection of

$m(X,\alpha)$ onto $p^{J_n}(X)$.

Proposition 3.1. *Let $\widehat{m}$ be the series LS estimator given in (2.2) with P-splines, cosine/sine or wavelets as the basis $p^{J_n}(X)$. Suppose that assumptions 2.5, 2.12 and 3.7 hold. Then:*

$$(1) \quad \sup_{\alpha\in\mathcal{N}_{0n}} \frac{1}{n}\sum_{i=1}^{n} \|\widehat{m}(X_i,\alpha) - \widehat{m}(X_i,\alpha_0) - \widetilde{m}(X_i,\alpha)\|_E^2 = O_p\left(\frac{J_n}{n}\left(\delta_{s,n}^*\right)^{2\kappa}\right);$$

if further, assumption 2.13 and $\max\{\frac{J_n}{n}\left(\delta_{s,n}^*\right)^{2\kappa}, b_{m,J_n}^2\} = o(n^{-1})$ hold, then assumption 3.1(i) is satisfied.

(2) Let assumptions of Lemma 2.4 and assumptions 3.2 - 3.6 hold. If $\max\{\frac{J_n}{n}\left(\delta_{s,n}^*\right)^{2\kappa}, b_{m,J_n}^2\} = o(n^{-1})$ and $J_n \asymp k(n) = o(n^{1/2})$, then: $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, V^{-1})$ with V^{-1} given in (3.1).

Remark 3.1. *By Lemma 2.4(2), for both the "mildly ill-posed" case and the "severely ill-posed" case, the condition $\frac{J_n}{n}\left(\delta_{s,n}^*\right)^{2\kappa} = o(n^{-1})$ is satisfied provided that $\gamma_h > d/(2\kappa)$. Moreover, the condition $J_n \asymp k(n) = o(n^{1/2})$ is automatically satisfied by the optimal growth order of $J_n \asymp k(n) = O\left([\ln(n)]^{d/a}\right)$ in the "severely ill-posed" case, and it is also satisfied in the "mildly ill-posed" case with the optimal growth order of $J_n \asymp k(n) = O\left(n^{\frac{d}{2(\gamma_h+a)+d}}\right)$ provided that $\gamma_h + a > d/2$.*

3.2 Weighted Bootstrap

To conduct statistical inference on the parametric component we need a way to estimate the confidence region of $\widehat{\theta}$. Previously, Ai and Chen (2003) propose a consistent sieve estimator of the asymptotic variance of $\widehat{\theta}$. Their variance estimator hinges on the differentiability of the residual functions $\rho(Z;\theta,h(\cdot))$ in $\alpha = (\theta,h)$, whereas in our paper $\rho(Z;\theta,h(\cdot))$ could be non-smooth with respect to $\alpha = (\theta,h)$. In this subsection we propose a weighted bootstrap procedure to consistently estimate the confidence region of $\widehat{\theta}$. We establish the validity of a weighted bootstrap by showing that the asymptotic distribution of the weighted bootstrap estimator (centered at $\widehat{\theta}_n$) coincides with the asymptotic distribution of our PSMD estimator (centered at θ_0). In a recent paper Ma and Kosorok (2005) establish a similar result for a semiparametric M-estimation without nonparametric endogeneity. We extend their results to the PSMD estimation of the conditional moment model (1.1) with nonparametric endogeneity.

Assumption 3.8. *$\{W_i\}_{i=1}^n$ is an i.i.d. sample of positive weights satisfying $E[W_i] = 1$ and $Var(W_i) = w_0$, and is independent of $\{(Y_i', X_i')\}_{i=1}^n$.*

In contrast to the nonparametric bootstrap where the weights are draws from a multinomial $(n, n^{-1}, \ldots, n^{-1})$, the weight here must be drawn independently. An example is the so-called Bayesian bootstrap where $W_i = U_i/(n^{-1}\sum_{i=1}^n U_i)$ with $U_i \sim Exp(1)$.

Assumption 3.9. *(i)* $\sup_{\alpha\in\mathcal{N}_{0n}} n^{-1}\sum_{i=1}^n W_i \|\widehat{m}(X_i,\alpha) - \widehat{m}(X_i,\alpha_0) - m(X_i,\alpha)\|_E^2 = o_p(n^{-1})$;

(ii) $n^{-1/2}\sum_{i=1}^n W_i \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v^*]\right)' \Sigma(X_i)^{-1}\{\rho(Z_i,\alpha_0) - \widehat{m}(X_i,\alpha_0)\} = o_P(1)$.

If $\sup_i W_i$ is bounded, then assumptions 3.9(i) and (ii) are directly implied by assumptions 3.1(i) and (iii) respectively. If W_i is not bounded, we can re-define $\rho(Z_i,\alpha)$ as $\rho_W(Z_i,\alpha) \equiv W_i \times \rho(Z_i,\alpha)$ and verify assumption 3.7 when $\widehat{m}(X,\alpha)$ is the series LS estimator of $m(X,\alpha)$.

Theorem 3.2. *Suppose that all the assumptions of Theorem 3.1, assumptions 3.8 and 3.9 hold. Let*

$$\left(\widehat{\theta}_n^*, \widehat{h}_n^*\right) = \widehat{\alpha}_n^* \equiv \arg\inf_{\alpha\in\mathcal{N}_{0n}} \left\{\frac{1}{n}\sum_{i=1}^n W_i \left\{\widehat{m}(X_i,\alpha)'[\widehat{\Sigma}(X_i)]^{-1}\widehat{m}(X_i,\alpha)\right\} + \lambda_n P(h)\right\}.$$

Then: Conditional on the data $\{(Y_i', X_i')\}_{i=1}^n$, $\sqrt{\frac{n}{w_0}}\left(\widehat{\theta}_n^* - \widehat{\theta}_n\right)$ has the same limiting distribution as that of $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$.

The theorem above allow us to construct an estimator for the confidence region in the following way:

1. Draw any i.i.d. sample $\{W_i\}_{i=1}^n$ satisfying assumption 3.8 with $Var(W_i) = 1$.
2. Compute $\widehat{\alpha}_n^*$ for the given sample of weights.
3. Repeat steps 1 and 2 many times (say N numbers of times) and compute the empirical quantiles of $(\widehat{\theta}_{n,q}^*)_{q=1}^N$.

4 Semiparametric Efficiency and Chi-square Approximation

In this section we first show that the semiparametric efficiency bound results of Ai and Chen (2003) remains valid for the models (1.1) with possibly nonsmooth residual functions $\rho(Z,\alpha)$, and that the

optimally weighted PSMD or locally continuously updated PSMD achieves the efficiency bound. We then show that the profiled optimally weighted (or profiled locally continuously updated) PSMD criterion function is asymptotically Chi-square distributed, which suggests another way to construct confidence region.

4.1 Semiparametric efficiency bounds and efficient estimation

Recall that $\Sigma_0(X) \equiv Var(\rho(Z,\alpha_0)|X)$. We define $\overline{\mathbf{V}}_0$ as the closure of the linear span of $\mathcal{A}-\{\alpha_0\}$ under the inner product defined using the optimal weighting $\Sigma_0(X)^{-1}$:

$$\langle v_1, v_2\rangle_0 = E\left[\left(\frac{dm(X,\alpha_0)}{d\alpha}[v_1]\right)' \Sigma_0(X)^{-1}\left(\frac{dm(X,\alpha_0)}{d\alpha}[v_2]\right)\right],$$

thus $(\overline{\mathbf{V}}_0, \langle\cdot\rangle_0)$ is a Hilbert space, with $\overline{\mathbf{V}}_0 = \mathcal{R}^{d_\theta} \times \overline{\mathcal{W}}_0$ and $\overline{\mathcal{W}}_0 = \overline{\mathcal{H}} - \{h_0\}$. Let $h - h_0 = -w(\theta-\theta_0)$, $\frac{dm(X,\alpha_0)}{d\alpha}[\alpha-\alpha_0] \equiv D_w(X)(\theta-\theta_0)$ and $D_w(X) \equiv \frac{dm(X,\alpha_0)}{d\theta'} - \frac{dm(X,\alpha_0)}{dh}[w]$. We define

$$V_0 \equiv \inf_w E\left\{D_w(X)'[\Sigma_0(X)]^{-1}D_w(X)\right\} = E\left\{D_{w_0}(X)'[\Sigma_0(X)]^{-1}D_{w_0}(X)\right\},$$

where $w_0 = (w_{01}, ..., w_{0d_\theta})$ and each $w_{0j} \in \overline{\mathcal{W}}_0$ is the solution to

$$\inf_{w_j \in \overline{\mathcal{W}}_0} E\left[\left(\frac{dm(X,\alpha_0)}{d\theta_j} - \frac{dm(X,\alpha_0)}{dh}[w_j]\right)' \Sigma_0(X)^{-1}\left(\frac{dm(X,\alpha_0)}{d\theta_j} - \frac{dm(X,\alpha_0)}{dh}[w_j]\right)\right].$$

When the residual function $\rho(Z,\alpha)$ is pointwise smooth wrt α, Ai and Chen (2003) establish that V_0 is the semiparametric efficiency bound for θ_0 in the model (1.1). The following theorem shows that their result remains valid when $\rho(Z,\alpha)$ is not pointwise smooth wrt α. We denote $q_0(y,x,\alpha_0)$ as the true joint density of (Y,X). Since $\mathcal{A}$ is convex at α_0 by assumption, for any fixed $h \in \mathcal{H}$, $h_0 + \xi(h-h_0) \in \mathcal{H}$ for small constant $\xi \geq 0$. Let $p(y,x,\theta,\xi) \equiv q_0(y,x,\theta,h_0+\xi(h-h_0))$ denote a parametric submodel passing through $q_0(y,x,\alpha_0)$ at the true values $\theta=\theta_0$ and $\xi=0$.

Assumption 4.1. *(i)* $E\left[\left(\frac{dm(X,\alpha_0)}{d\theta'}\right)' \Sigma_0(X)^{-1}\left(\frac{dm(X,\alpha_0)}{d\theta'}\right)\right]$ *is finite,* $E[D_w(X)'[\Sigma_0(X)]^{-1}D_w(X)]$ *is finite for any* $w = (w_1, ..., w_{d_\theta})$ *with* $w_j \in \overline{\mathcal{W}}_0$*; (ii) for every fixed* $h \in \mathcal{H}$, $p(y,x,\theta,\xi) \equiv q_0(y,x,\theta,h_0+\xi(h-$

h_0)) *is smooth in the sense of Van der Vaart (1991).*

Theorem 4.1. *Let assumptions 2.1, 2.4(i), 2.5, 2.8(i), 2.9(ii), 3.2(ii) and 4.1 hold. Then: (1)* V_0 *is the semiparametric efficiency bound for* θ_0 *in the model (1.1). (2) The positive definiteness of* V_0 *is the necessary condition for* θ_0 *to be estimable at* $\sqrt{n}-$*rate. (3) Suppose that all the assumptions of Theorem 3.1 hold with* $\Sigma(X) = \Sigma_0(X)$*, then the corresponding PSMD estimator of* θ_0 *is efficient with asymptotic variance* V_0^{-1}.

4.2 Chi-square approximation

Previously for sieve MLE, Shen and Shi (2005) provide sufficient conditions to ensure that the sieve likelihood ratio statistic is asymptotically chi-square distributed. Murphy and Van der Vaart (2000) present conditions to ensure that the profiled likelihood of semiparametric M-estimation is asymptotically chi-square distributed. In this subsection we show that the profile optimally weighted PSMD criterion ($\widehat{Q}_n(\theta)$) and the profile continuously updated PSMD criterion ($\widehat{Q}_n^C(\theta)$) also possess such a nice property. As in Ai and Chen (2003), we propose the following locally continuous updated PSMD estimator $\tilde{\alpha}_n \equiv (\tilde{\theta}_n, \tilde{h}_n)$ that solves

$$\min_{\alpha \in \mathcal{N}_{0n}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \widehat{m}(X_i, \alpha)' [\widehat{\Sigma}(X_i, \alpha)]^{-1} \widehat{m}(X_i, \alpha) + \lambda_n P(h) \right\},$$

where $\widehat{\Sigma}(X, \alpha)$ is any nonparametric consistent estimator of $Var[\rho(Z, \alpha)|X]$, and the neighborhood can be centered around $\hat{\alpha}_n$ (the PSMD estimator with $\widehat{\Sigma}(X_i, \alpha) = I$).

We can also define the locally continuous updated profiled SMD estimator:

$$\begin{aligned} Step1 \quad &: \quad \widetilde{h}_\theta = \arg \inf_{h \in \mathcal{H}_{0n}} \frac{1}{n} \sum_{i=1}^{n} \widehat{m}(X_i, \theta, h)' [\widehat{\Sigma}(X_i, \theta, h)]^{-1} \widehat{m}(X_i, \theta, h) + \lambda_n P(h), \\ Step2 \quad &: \quad \tilde{\theta}_n = \arg \inf_{\theta} \frac{1}{n} \sum_{i=1}^{n} \widehat{m}(X_i, \theta, \widetilde{h}_\theta)' [\widehat{\Sigma}(X_i, \theta, \widetilde{h}_\theta)]^{-1} \widehat{m}(X_i, \theta, \widetilde{h}_\theta) + \lambda_n P(\widetilde{h}_\theta), \end{aligned}$$

and $\tilde{h}_n = \widetilde{h}_{\tilde{\theta}_n}$.

Define:

$$\widehat{Q}_n(\theta) \equiv n^{-1} \sum_{i=1}^{n} \widehat{m}(X_i;\theta,\widetilde{h}_\theta)' \left[\widehat{\Sigma}(X_i,\widehat{\alpha}_n)\right]^{-1} \widehat{m}(X_i;\theta,\widetilde{h}_\theta) + o_P(n^{-1}),$$

$$\widehat{Q}_n^C(\theta) \equiv n^{-1} \sum_{i=1}^{n} \widehat{m}(X_i;\theta,\widetilde{h}_\theta)' \left[\widehat{\Sigma}(X_i;\theta,\widetilde{h}_\theta)\right]^{-1} \widehat{m}(X_i;\theta,\widetilde{h}_\theta) + o_P(n^{-1}).$$

Notice that we are not including the penalty term in the above criterion functions. This is due to the fact that the penalty term (after appropriate centering) is of order $o_P(n^{-1})$ and thus negligible. Therefore, for the sake of simplicity we did not include it.

We impose the following additional conditions

Assumption 4.2. *(i) $\widehat{\Sigma}(X,\alpha)$ is finite and positive definite with eigenvalues bounded away from zero uniformly for all $X \in \mathcal{X}$ and $\alpha \in \mathcal{N}_{0n}$; (ii) the same holds for $\Sigma(X,\alpha_0)$.*

Assumption 4.3. *Uniformly over $\alpha_1, \alpha_2 \in \mathcal{N}_{0n}$,*

$$E\left[\left\|\left(\frac{dm(X,\alpha_1)}{d\alpha}[v_n^*]\right)'\Sigma(X,\alpha_2)^{-1} - \left(\frac{dm(X,\alpha_0)}{d\alpha}[v_n^*]\right)'\Sigma(X,\alpha_0)^{-1}\right\|_E^2\right] = o(n^{-1/2})$$

Assumption 4.4. *(i) $\sup_{\alpha\in\mathcal{N}_{0n},x\in\mathcal{X}} |\widehat{\Sigma}(x,\alpha) - \Sigma(x,\alpha)| = O_P(\delta^*_{\Sigma,n})$ with $\delta^*_{\Sigma,n} \times \delta^*_n = o_P(n^{-1/2})$; (ii) $\sup_{\alpha\in\mathcal{N}_{0n},x\in\mathcal{X}} |\Sigma(x,\alpha) - \Sigma(x,\alpha_0)| = o_P(1)$; (iii) $\Sigma(\cdot,\alpha) \in \Lambda_c^{\gamma_\Sigma}(\mathcal{X})$ and $r_\Sigma \equiv \gamma_\Sigma/d_x > 1/2$.*

Assumption 4.3 is a version of assumption 4.4 in Ai and Chen (2003) but in our case, as $\widehat{\Sigma}(X,\widehat{\alpha})$ is estimated in a first stage, we must allow for different arguments in $dm(X,\cdot)/d\alpha$ and $\widehat{\Sigma}(X,\cdot)$. Given the imposed assumptions a sufficient condition for assumption 4.3 to hold is: $E\left[\|\Sigma(X,\alpha) - \Sigma(X,\alpha_0)\|_E^2\right] = o_p(n^{-1/2})$. We can now establish the following theorem.

Theorem 4.2. *Under assumptions 2.5 - 2.8 and 3.1, 2.9- 4.4, it follows that*

$$2n\left(\widehat{Q}_n(\theta_0) - \widehat{Q}_n(\tilde{\theta}_n)\right) \Rightarrow \chi^2_{d_\theta}$$

Next we present an analogous result for the continuous updating estimator.

Assumption 4.5. *(i)* $\sup_{\alpha\in\mathcal{N}_{0n},x\in\mathcal{X}}\left|\widehat{\Sigma}(x,\alpha)-\Sigma(x,\alpha)\right|\times(\delta_n^*+\delta_m^*)^2=o_p(n^{-1})$*; (ii)* $\Sigma(X,\alpha)^{-1}$ *is pathwise twice continuously differentiable with derivatives that are bounded (in the neighborhood* $\mathcal{N}_{on}$*).*

Theorem 4.3. *Under assumptions 2.5 - 2.8 and 3.1, 2.9- 4.3, 4.4(iii) and 4.5, it follows that*

$$2n\left(\widehat{Q}_n^C(\theta_0)-\widehat{Q}_n^C(\tilde{\theta}_n)\right)\Rightarrow\chi^2_{d_\theta}.$$

Constructing an estimator for $\Sigma(X,\alpha)$ that satisfies the required conditions can be a daunting task. In appendix we present a lemma that provides sufficient conditions for assumption 4.4 when $\widehat{\Sigma}$ is a series LS estimator. For alternative nonparametric variance estimators and their properties, see Robinson (1995b), Andrews (1995), Hall and Marron (1990), Brown and Levine (2007) and references therein.

5 A Partially Linear Quantile IV Example

In this section we apply the above general theoretical results to study a partially linear quantile IV regression model. The model is:

$$Y_3=\theta_0Y_1+h_0(Y_2)+U,\quad \Pr(U\leq 0|X)=\gamma, \tag{5.1}$$

where θ_0 is a scalar unknown parameter and $h_0()$ is a real-valued unknown function. The conditional distribution of the error term U given $X=(X_1,X_2')'$ is unspecified, except that $F_{U|X}(0)=\gamma$ for a known fixed $\gamma\in(0,1)$. The support of X is $\mathcal{X}=[0,1]^{d_x}$ with $d_x=1+d_2$, and the support of $Y=(Y_3,Y_1,Y_2')'$ is $\mathcal{Y}\subseteq\mathcal{R}^{2+d_2}$. To map into the general model (1.1), we let $Z=(Y',X')'$, $\alpha=(\theta,h)$, $\rho(Z,\alpha)=1\{Y_3\leq\theta Y_1+h(Y_2)\}-\gamma$ and $m(X,\alpha)=E[F_{Y_3|Y_1,Y_2,X}(\theta Y_1+h(Y_2))|X]-\gamma$.

We estimate α_0 using the PSMD estimator $\widehat{\alpha}_n$, with $\widehat{m}(X,\alpha)$ being a series LS estimator of $m(X,\alpha)$, $P(h)=||\nabla^k h||^j_{L^j(\mathcal{R}^{d_2},w)}$ for a finite $k\geq 0$, $j=1,2$ and a positive continuous weighting function w on $\mathcal{R}^{d_2}$, and $\mathcal{A}_n=[\underline{\theta},\overline{\theta}]\times\mathcal{H}_n$ being a finite dimensional ($\dim(\mathcal{H}_n)\equiv k(n)<\infty$) *linear* sieve. It is easy to check that $\Sigma(X,\alpha_0)=\gamma(1-\gamma)$. Thus we can take $\widehat{\Sigma}(X)=\gamma(1-\gamma)$. Recently Chernozhukov et al. (2007) and Horowitz and Lee (2007) have studied the nonparametric quantile IV regression model $E[1\{Y_3\leq$

$h_0(Y_2)\}|X] = \gamma$. Chen and Pouzo (2007) have illustrated their general convergence rate results using a nonparametric additive quantile IV regression example $E[1\{Y_3 \leq h_{01}(Y_1)+h_{02}(Y_2)\}|X] = \gamma$. Chen et al. (2003) have used an example of partially linear quantile IV regression with an exogenous Y_2 (i.e., $Y_2 = X_2$), and Lee (2003) has studied the partially linear quantile regression with exogenous Y_1 and Y_2 (i.e., $Y_1 = X_1, Y_2 = X_2$). See Koenker (2005) for excellent review on quantile models.

We impose some low level sufficient conditions:

Condition 5.1. *(i)* $\mathcal{H} \subseteq \Lambda^{\gamma_2}(\mathcal{R}^{d_2})$ *with* $r_2 \equiv \gamma_2/d_2 > 1$ *and* $||h||_{L^2}^2 \leq M$*; (ii)* $||\alpha||_s \equiv |\theta| + ||h||_{L^2(\mathcal{R}^{d_2},\omega)}$ *where* ω *is a continuous weighting function whose integral is normalized to one and* $\omega(y_2) \asymp f_{Y_2}(y_2)$ *as* $|y_2| \to \infty$*; (iii) if* $\alpha \in \mathcal{A}$ *and* $m(X,\alpha) = 0$ *then* $||\alpha - \alpha_0||_s = 0$.

Condition 5.2. *(i)* $\mathcal{H}_n = span\{q_1, \ldots, q_{k(n)}\}$ *with* $(q_k)_k$ *being wavelets, P-spline, cosine polynomials or Hermite; (ii)* $k(n) \to \infty$ *and* $k(n)/n = o(1)$.

Condition 5.3. *(i)* $F_{Y_3|Y_1,Y_2,X}$ *is twice continuously differentiable on all its arguments with bounded derivatives; (ii)* $E[F_{Y_3|Y_1,Y_2,X}(\theta Y_1 + h(Y_2))|X = \cdot] \in \Lambda_1^{\gamma_m}(\mathcal{X})$, $E\left\{f_{Y_3|Y_1,Y_2,X}(\theta Y_1 + h(Y_2))[v_n^*]|X = \cdot\right\} \in \Lambda_c^{\gamma'_m}(\mathcal{X})$ *with* $r_m \equiv \gamma_m/d_x$, $r'_m \equiv \gamma'_m/d_x > 1/2$*; (iii)* $E\{(E[|Y_1||X])^2\} \leq M < \infty$.

Condition 5.4. *(i)* $P(h) \equiv ||\nabla^s h||_{L^j}^j$ *with* $0 \leq s < \gamma_2$ *and* $j = 1, 2$*; (ii)* $\lambda_n = o(n^{-1/2})$.

Condition 5.5. *(i) Assumption 2.11 holds with* $b_j \asymp j^{-2a/d_2}$*; (ii)* $\gamma_2 > a + \frac{d_2}{2}$.

Denote $w^* \in L^2(f_{Y_2})$ as the solution to:

$$\inf_{w(y_2)} E\left[\left(E\left\{f_{Y_3|Y_1,Y_2,X}(\theta_0 Y_1 + h_0(Y_2))[Y_1 - w(Y_2)]|X\right\}\right)^2\right].$$

Condition 5.6. *(i)* $E\left\{f_{Y_3|Y_1,Y_2,X}(\theta_0 Y_1 + h_0(Y_2))[g(Y_1,Y_2)]|X\right\} = 0$ *implies* $g(Y_1,Y_2) \equiv 0$ *almost surely, and* Y_1 *is not a measurable function of* Y_2*; (ii)* $E\left[\left(E\left\{[\Pi_n w^*(Y_2) - w^*(Y_2)]|X\right\}\right)^2\right] \times (\delta_n^*)^2 = o_p(n^{-1})$

In Chen and Pouzo (2007) we obtain the nonparametric convergence rate of h_0 for this example. Here we only present the asymptotic normality and efficiency result for the estimation of θ_0.

Proposition 5.1. *Under assumptions 2.5, 2.12, $\theta_0 \in int(\Theta)$ and conditions 5.2 - 5.6, we have:* $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \Rightarrow N\left(0, V_0^{-1}\right)$, *with*

$$V_0 = \frac{E\left[\left(E\left\{f_{Y_3|Y_1,Y_2,X}(\theta_0 Y_1 + h_0(Y_2))[Y_1 - w^*(Y_2)]|X\right\}\right)^2\right]}{\gamma\left(1-\gamma\right)}.$$

Moreover, V_0 is the semiparametric efficiency bound.

Remark 5.1. *(1) Under $Y_j = X_j$ for $j = 1, 2$ (i.e., no endogeneity), V_0 becomes*

$$V_0 = \frac{E\left[\left(f_{U|X}\left(0\right)\right)^2\left(X_1 - w^*\left(X_2\right)\right)^2\right]}{\gamma(1-\gamma)},\ w^*\left(X_2\right) = \frac{E\left[\left(f_{U|X}\left(0\right)\right)^2 X_1|X_2\right]}{E[\left(f_{U|X}\left(0\right)\right)^2|X_2]}.$$

(2) Florens et al. (2006) study the root-n asymptotic normality for the partially linear IV mean regression model: $Y_3 = Y_1\theta_0 + h_0(Y_2) + U$ with $E[U|X] = 0$. When we apply our asymptotic normality result of the PSMD estimation to this example, our Proposition 3.1 allows for severely ill-posed case, i.e., $b_j \asymp \exp\{-j^a\}$. This is due to the fact that the assumptions related to controlling the second order terms (e.g. assumptions 3.4(ii) and 3.5) are trivially satisfied as $m(X, \alpha) = E[Y_3 - Y_1\theta - h(Y_2)|X]$ is linear in $\alpha = (\theta, h)$ in this example. Therefore the rate of convergence under the strong norm is allowed to decay very slowly such as a logarithmic rate. In particular, this generalizes Robinson (1988) to allow for endogenous regressors.

In the following we will establish the confidence intervals for $\widehat{\theta}_n$. Given that $\widehat{\Sigma} = \Sigma(X, \alpha_0) = \gamma(1-\gamma)$ the assumptions for theorem 4.2 are greatly simplified and we can omit the proof of the following proposition.

Proposition 5.2. *Under the same conditions of proposition 5.1 it follows that*

$$2\frac{\sum_{i=1}^n\left\{\left(\widehat{m}(X_i, \theta_0, \widehat{h}_{\theta_0})\right)^2 - \left(\widehat{m}(X_i, \widehat{\theta}_n, \widehat{h}_{\widehat{\theta}_n})\right)^2\right\}}{\gamma(1-\gamma)} \Rightarrow \chi_1^2$$

The previous result allow us to establish confidence interval estimators for $\widehat{\theta}_n$ by computing

$$\left\{\theta \in [\underline{\theta}, \overline{\theta}] : 2\sum_{i=1}^{n}\left\{\left(\widehat{m}(X_i, \theta, \widehat{h}_\theta)\right)^2 - \left(\widehat{m}(X_i, \widehat{\theta}_n, \widehat{h}_{\widehat{\theta}_n})\right)^2\right\} \leq \gamma(1-\gamma)c_p\right\}.$$

Note that by assumption the penalty term is of order $o(n^{-1})$ so it will be negligible for large n. The estimator $\widehat{h}_\theta$ is the profile estimator, obtained by fixing θ and minimizing the criterion function with respect to $h \in \mathcal{H}_n$.

6 Simulation and Empirical Illustration

6.1 A Monte Carlo Study

We assess the finite sample performance of the penalized SMD estimator in a simulation study. We simulate the data from the following partially linear quantile IV model:

$$\begin{aligned} Y_1 &= X_1\theta_0 + h_0(Y_2) + U, \\ U &= \sqrt{0.075}\left(-\Phi^{-1}\left(\frac{E\left[h_0(Y_2)\,|X_2\right] - h_0(Y_2)}{10} + \gamma\right) + \varepsilon\right),\ \varepsilon \sim N(0,1), \end{aligned}$$

where $\theta_0 = 1$, $h_0(y_2) = \Phi\left(\frac{y_2 - \mu_{y_2}}{\sigma_{y_2}}\right)$, $X_1 \sim U[0,1]$ independent of ε and $(Y_2, X_2) \sim f$. Following the way Blundell et al. (2007) conduct their Monte Carlo study, we generate our Monte Carlo experiment from the 1995 British Family Expenditure Survey (FES) data set with subsample of families with no kids. In particular, Y_2 is the endogenous regressor (log-total expenditure) and $\Phi(X_2)$ is its instrument (log-gross earnings). We consider the following specification for the joint density f as a bivariate Gaussian density which first and second moments are estimated from the FES data set. We draw an i.i.d. sample of $(X_1, Y_2, X_2, \varepsilon)$ with sample size $n = 1000$.

We estimate $m(X, \alpha)$ by the series LS estimator $\widehat{m}(X, \alpha)$ given in (2.2) with $p^{J_n}(X)$ being the tensor-product of P-Spline(3,3) and P-Cos(9).[1] We use a linear spline sieve P-Spline(2,6) as $\mathcal{H}_n$. We also add a

[1] The notation P-Spline(p,q) denotes a polynomial spline of order p with q number of knots, and P-Cos(p) stands for cosine series with p number of terms. We have tried other combinations as sieve base for conditional mean function m and all yield very similar results.

penalization term for the L^2 norm of the first derivative of the function with $\lambda_n = \{0.001, 0.01, 0.1\}$.[2] In all the cases we performed 500 Monte Carlo repetitions.[3]

When applying the asymptotic normality theorem to Example 1, we note that it is difficult to verify assumption 3.4(iii) and assumption 3.5 for the severely ill-posed case. In order to shed some light about this case, in table 2 we present for the G-DEN case and for $\gamma = 0.750$ how the variance changes with the different sample sizes $n = \{125, 250, 500, 1000\}$. This will allow us to see how the parametric part of our estimator behaves in different sample sizes, in particular we can check (by eye-balling) if the variance decays at the same order as the sample size. If this is not the case it, then this is evidence that asymptotic normality does not hold in this case. We note that we adjusted the penalization parameter to vary with the sample size, by increasing it by the same proportion as the sample size increased. We can see that the variance decays at approximately the same rate as the sample size decreases, given no evidence that our $\sqrt{n}$-root results do not hold.

Table 1 summarizes the results for the G-DEN case for different quantiles. Notice that all the statistics corresponding to the θ_0 estimate are approximately the same across different quantiles. This is also the case for most of the statistics corresponding to the estimation of the unknown function, h_0. We note that integrated bias squared ($IBIAS^2_{MC}$) is an order of magnitude smaller than the integrated variance ($IVar_{MC}$) for all the quantiles and that the quantile integrated mean square error for $\gamma = 0.50$ is an order of magnitude lower than for the rest of the quantiles. This result is driven by the fact that the variance is much lower for the 0.50 quantile than for any other quantiles. Figure 1 shows the estimated function, the true function h_0 and the 0.95% confidence band, obtained from the Monte Carlo sample. One can see that for all the cases our estimator performs well.

Overall we can conclude that our estimator performs very well and that there is evidence that the parametric part of it behaves asymptotically normal.

[2]The penalization parameter λ_n is chosen to minimize the integrated MSE of $\widehat{h}$ for a small number of Monte Carlo repetitions. This choice of λ_n is adhoc, more complex and appropriate methods, such as Cross Validation, are out of the scope of this paper.

[3]We have also performed 250 and 1000 Monte Carlo iterations but as the results remain almost unchanged throughout the different choices of Monte Carlo repetitions we only report the case of 500 iterations.

6.2 An Empirical Illustration

We apply the penalized SMD to nonparametric quantile IV estimation of Engel curves (or consumer demand functions) using the UK Family Expenditure Survey data. The model is

$$E[1\{Y_{1il} \leq h_{0l}(Y_{2i} - \theta_1 X_{1i}) + \theta_l X_{1i}\}|X_i] = \gamma \in (0,1),\ l = 1, ..., 7,$$

where Y_{1il} is the budget share of household i on good l (in this application, 1 : food-out, 2 : food-in, 3 : alcohol, 4 : fares, 5 : fuel, 6 : leisure goods, and 7 : travel). Y_{2i} is the log-total expenditure of household i that is endogenous, and $X_i \equiv (X_{1i}, X_{2i})$ with X_{1i} being 0 for without kids sample and 1 for with kids sample and X_{2i} is the gross earnings of the head of household, which is the instrumental variable. We work with the whole sample (with and without kids) that consists of 1655 observations. The same data set has been studied in Blundell et al. (2007).

As illustration, we apply the penalized SMD using a finite-dimensional polynomial spline sieve to construct the sieve space $\mathcal{H}_n$ for h, with different types of penalty functions. We have tried $||\nabla^k h||^j_{L^j(d\widehat{\mu})} \equiv n^{-1}\sum_{i=1}^n |\nabla^k h(Y_{2i})|^j$ for $k = 1, 2$ and $j = 1, 2$, and Hermite polynomial sieves, cosine sieves and polynomial splines sieves for the series LS estimator $\hat{m}$. All combinations yielded very similar results; hence we only present figures for one case. Due to the lack of space, in Figure 2 we report the penalized SMD estimated Engel curves only for three different quantiles $\gamma = \{0.25, 0.50, 0.75\}$ and for four selected goods, using P-Spline(2,5) as $\mathcal{H}_n$ and tensor product of P-Spline(2,5)×P-Spline(5,10) for $\hat{m}$.

Table 3 shows the corresponding θ_1 and $(\theta_l)_{l=1}^7$ for the median ($\gamma = 0.50$) and penalization equal to: $\widehat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\widehat{\mu})}$ with $\lambda_n = 0.001$, $\widehat{P}_n(h) = ||\nabla^2 h||_{L^1(d\widehat{\mu})}$ with $\lambda_n = 0.001$, and $\widehat{P}_n(h) = ||\nabla h||^2_{L^2(d\widehat{\mu})}$ with $\lambda_n = 0.003$, respectively. The two last columns in table 3 only presents the median ($\gamma = 0.50$) for $\widehat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\widehat{\mu})}$ with $\lambda_n = 0.0003$ and $\widehat{P}_n(h) = ||\nabla h||^2_{L^2(d\widehat{\mu})}$ with $\lambda_n = 0.0003$, respectively [4]. Figure 2 presents the corresponding curves for each of the five cases and in the last two rows we include the estimator for the partially linear IV mean regression model for comparison.

By inspection we see that the overall estimated function shapes are not very sensitive to the choice of

[4] The values for the rest of the quantiles are available upon request.

λ_n nor the choice of penalization. The parametric values are slightly more sensitive, and thus we observe some changes in the signs. We note that the columns in table 3 (in particular the last two) yield very similar results to the ones in Blundell et al. (2007), except for the sign in fares.

7 Conclusion

In this paper, we study asymptotic properties of the penalized SMD estimator for the conditional moment models containing unknown functions that could depend on endogenous variables. For such models with possibly non-smooth generalized residual functions, and possibly non-compact infinite dimensional parameter spaces, we show that the PSMD estimator of the parametric part is root-n asymptotically normal, and the optimally weighted PSMD reaches the semiparametric efficiency bounds. In addition, we establish the validity of a weighted bootstrap procedure for confidence region construction of possibly inefficient but root-n consistent PSMD estimator. For the optimally weighted efficient PSMD estimator, we show the validity of an alternative confidence region construction method by inverting an efficient profiled criterion function. We illustrate the general theoretic results by a partially linear quantile IV regression example, a simulation study, and an empirical estimation of a shape invariant system of quantile Engel curves with endogenous total expenditure. The weighted bootstrap method could be easily extended to allow for misspecified semoparametric conditional moment models of Ai and Chen (2007).

All the large sample theories obtained in this paper are first-order asymptotics. There is no results on higher order refinement for semiparametric conditional moment models containing functions of endogenous variables yet. There are some second order theories for semiparametric models without nonparametric endogeneity, such as Robinson (1995), Linton (1995), Nishiyama and Robinson (2000), Nishiyama and Robinson (2001) and Nishiyama and Robinson (2005), to name a few. We hope to study the higher order refinement of the weighted bootstrap procedure in another paper.

A Mathematical Appendix

In the following lemma we establish an upper bound for the convergence rate of $|\widehat{\Sigma}(X,\alpha)-\Sigma(X,\alpha_0)|$ for the case where $\widehat{\Sigma}$ is a series LS projection estimator. This provides one kind of sufficient conditions for assumption 4.4 when $\widehat{\Sigma}$ is a series LS estimator. Let $[\widehat{\Sigma}(X,\alpha)]_{jl} \equiv \sum_{i=1}^{n} \rho_j(Z_i,\alpha)\rho_l(Z_i,\alpha)p^{J_n}(X_i)' \left(P'P\right)^{-1} p^{J_n}(X)$. We will denote $\Sigma(X) \equiv \Sigma(X,\alpha_0) = E[\rho(Z,\alpha_0)^2|X]$ and $\widehat{\Sigma}(X) \equiv \widehat{\Sigma}(X,\widehat{\alpha}_n)$. Denote $\xi_n \equiv \xi_{0n} \equiv \sup_{x\in\mathcal{X}} ||p^{J_n}(x)||_E$ and $\xi_{1n} \equiv \sup_{x\in\mathcal{X}} ||\frac{dp^{J_n}(x)}{dx'}||_E$.

Assumption A.1. *(i) Each element of $\rho(X,\alpha)\rho(X,\alpha)'$ satisfies assumption 3.7 in $\mathcal{A}_{os}$; (ii) $\Sigma(\cdot,\alpha) \equiv E\left[\rho(X,\alpha)\rho(X,\alpha)'|X=x\right] \in \Lambda_c^{\gamma_\Sigma}(\mathcal{X})$ with $r_\Sigma = \gamma_\Sigma/d_x > 1/2$; (iii) $\forall g \in \Lambda_c^{r_\Sigma}(\mathcal{X})$ there exists a $p^{J_n}(X)'\pi$ such that $\sup_{x\in\mathcal{X}} \sup_{\alpha\in\mathcal{N}_{0n}} \left|g(X) - p^{J_n}(X)'\pi\right| = O(J_n^{-r_\Sigma})$.*

Assumption A.2. *(i) $\delta^*_{\Sigma,n} \leq K \times n^{-1/2}\xi_n \int_0^{K\xi_n} \sqrt{1+\log(N_{[]}((w/\xi_n)^{1/\kappa}, \mathcal{A}_{osn}, ||\cdot||_s))}dw$; (ii) $\delta^*_{\Sigma,n} = O(\frac{\xi_n^{3/2}}{\sqrt{n}\xi_{1n}})$; (iii) $\delta^*_{\Sigma,n} = O(J_n^{-r_\Sigma})$.*

Lemma A.1. *Under assumptions 2.5, A.1 and A.2, it follows that:*

$$(1) \quad \sup_{x\in\mathcal{X}} \sup_{\alpha\in\mathcal{N}_{0n}} |\widehat{\Sigma}(X,\alpha) - \Sigma(X,\alpha)| = O_p(\delta^*_{\Sigma,n}),$$

where $\delta_{\Sigma,n} = \max\left\{\frac{\xi_n^{3/2}}{\sqrt{n}\xi_{1n}}, n^{-1/2}\xi_n \int_0^{K\xi_n} \sqrt{1+\log\left(N_{[]}((w/\xi_n)^{1/\kappa}, \mathcal{A}_{osn}, ||\cdot||_s)\right)}dw, J_n^{-r_\Sigma}\right\}$.

$$(2) \quad \sup_{x\in\mathcal{X}} \sup_{\alpha\in\mathcal{N}_{0n}} |\Sigma(x,\alpha) - \Sigma(x,\alpha_0)| = K \times ||\alpha-\alpha_0||_s^\kappa.$$

Proof of Lemma A.1: First we will establish the rate $\sup_x \sup_\alpha \left|\widehat{\Sigma}(x,\alpha) - \widetilde{\Sigma}(x,\alpha)\right|$ where

$$\widetilde{\Sigma}(x,\alpha) \equiv \sum_{i=1}^{n} E\left[\rho(Z,\alpha)'\rho(Z,\alpha)|X_i\right] p^{J_n}(X_i)' \left(P'P\right)^{-1} p^{J_n}(x).$$

Define $\epsilon(Z,\alpha) = \rho(Z,\alpha)'\rho(Z,\alpha) - \Sigma(x,\alpha)$. We basically need to study – component by component –

$$p^{J_n}(X)'(P'P)^{-1} \sum_{j=1}^{n} p^{J_n}(X_j)\epsilon(Z_j,\alpha).$$

By invoking maximal inequality arguments it follows

$$\begin{aligned} & E\left[\sup_{X,\alpha}\left|n^{-1}\sum_{j=1}^{n} p^{J_n}(X)'(P'P/n)^{-1}p^{J_n}(X_j)\epsilon(Z_j,\alpha)\right|\right] \\ \leq \; & \frac{\xi_n}{\sqrt{n}} \int_0^{W_2} \sqrt{1+\log\left(N_{[]}(w,\star,||\cdot||_{L^2})\right)}dw, \end{aligned}$$

where W_2 is bounded by $\sup_{x,\alpha}(P'P/n)^{-1}p^{J_n}(X_j)\epsilon(Z_j,\alpha) = O_p(\xi_n)$ and $\star$ stands for the class of functions of the aforementioned form for $(x,\alpha) \in \mathcal{X}\times\mathcal{N}_{0n}$. By arguments similar to the ones in Chen et al. (2003) the term inside the integral is bounded by the entropy of $\mathcal{X}$ and $\mathcal{N}_{0n}$ with appropriately modified radius. Given that $\mathcal{X}$ is compact $\xi_n\int_0^{K\xi_n}\sqrt{1+\log\left(N_{[]}(w,\mathcal{X},||\cdot||_E)\right)}dw \leq K\xi_n\int_0^{K\xi_n}\sqrt{1-d_x\log(w)}dw \leq K\times d_x\times\xi_n^{3/2}$. Given assumptions over $p^{J_n}(X)$ it follows that the appropriate modification of the radius is to scale it by $1/\xi_{1n}$, thus the bound corresponding to this part is of the form $O(d_x\times\xi_n^{3/2}/(\sqrt{n}\xi_{1n}))$. For the entropy in $\mathcal{N}_{0n}$, given assumption A.1(i), we need to modify the radius by $(w/\xi_n)^{1/\kappa}$. It then follows

$$n^{-1/2}\xi_n\int_0^{K\xi_n}\sqrt{1+\log\left(N_{[]}((w/\xi_n)^{1/\kappa},\mathcal{A}_{osn},||\cdot||_s)\right)}dw.$$

By assumptions A.1(ii)(iii) we have:

$$\sup_{x\in\mathcal{X}}\sup_{\alpha\in\mathcal{N}_{0n}}\left|\widetilde{\Sigma}(X,\alpha)-\Sigma(X,\alpha)\right| = O_p(J_n^{-r_\Sigma}).$$

Result (1) follows.

Result (2) is trivially satisfied by assumption A.1(i), and $|\Sigma(X,\alpha)-\Sigma(X,\alpha_0)| \leq b(X)||\alpha-\alpha_0||_s^\kappa (= o_p(1))$. *Q.E.D.*

Lemma A.1 implies assumption 3.2(i). Suppose that $\log(N_{[]}((w/\xi_n)^{1/\kappa},\mathcal{A}_{osn},||\cdot||_s)) \leq K\times k(n)\log\left(k(n)\xi_n/w\right)$, $r_\Sigma = r_m$, $\xi_{qn} = J_n^{1/2+q}$ and $k(n)\asymp J_n \asymp n^{\frac{1}{2r_m+1}}$, and $\delta_n^* = o(n^{-1/4})$. Then assumption A.2(i) implies

$$n^{-1/2}J_n^{1/2}J_n\int_0^{K\xi_n}\sqrt{\log\left(k(n)\xi_n/w\right)}dw \leq n^{-1/2}J_n^{2+1/2}$$

and given that $J_n = O(n^{\frac{1}{2r_m+1}})$ it follows that $n^{\frac{5}{4r_m+2}} \leq n^{\frac{3}{4}}$ thus $2\leq r_m$ will suffice. On the other hand assumption A.2(ii) is $\frac{J_n^{3/4}}{J_n^{3/2}}n^{-1} = \frac{1}{nJ_n^{3/4}}$ and this has to be of order, at least $n^{-1/4}$, which is directly satisfied. Finally assumption A.2(iii) implies that $J_n^{-r_\Sigma} = n^{-\frac{r_\Sigma}{2r_m+1}} \leq n^{-1/4}$ which implies, given that $r_\Sigma = r_m$, $r_m \geq 1/2$. Therefore, if $\Sigma(\cdot,\alpha)$ belongs to a smooth enough class, given that $\delta_n^* = o(n^{-1/4})$ assumption 3.2(i) holds.

Proof of Theorem 3.1: To simplify notation, denote $|||A|||_\Sigma^2 \equiv n^{-1}\sum_{i=1}^n A_i'\left[\Sigma\left(X_i\right)\right]^{-1}A_i$, and define $|||A|||_{\widehat{\Sigma}}^2$ analogously. Note that by assumption 2.6(iii) $|||\cdot|||_{\widehat{\Sigma}}^2 \leq K|||\cdot|||_I^2$. Therefore by assumption 3.1(i) it follows that

$$\sup_{\alpha\in\mathcal{N}_{0n}}|||\widehat{m}(\cdot,\alpha)-\widehat{m}(\cdot,\alpha_0)-m(\cdot,\alpha)|||_{\widehat{\Sigma}}^2 = o_P(n^{-1}).$$

We can then show that $\frac{1}{2}|||\widehat{m}(\cdot,\alpha_0)+m(\cdot,\alpha)|||_{\widehat{\Sigma}}^2 - Z_n \leq |||\widehat{m}(\cdot,\alpha)|||_{\widehat{\Sigma}}^2 \leq 2|||\widehat{m}(\cdot,\alpha_0)+m(\cdot,\alpha)|||_{\widehat{\Sigma}}^2 + Z_n$, with $Z_n\geq 0$ and $Z_n = o_p(n^{-1})$, or $|||\widehat{m}(\cdot,\alpha_0)+m(\cdot,\alpha)|||_{\widehat{\Sigma}} - \sqrt{Z_n} \leq |||\widehat{m}(\cdot,\alpha)|||_{\widehat{\Sigma}} \leq |||\widehat{m}(\cdot,\alpha_0)+m(\cdot,\alpha)|||_{\widehat{\Sigma}} + \sqrt{Z_n}$.

After some algebra $|||\widehat{m}(\cdot,\alpha)|||^2_{\widehat{\Sigma}} = C|||\widehat{m}(\cdot,\alpha_0)+m(\cdot,\alpha)|||^2_{\widehat{\Sigma}} + o_p(n^{-1})$ for a constant $C>0$ and for all $\alpha \in \mathcal{N}_{0n}$.[5] Since $|||\widehat{m}(\widehat{\alpha}_n)|||^2_{\widehat{\Sigma}} + \lambda_n P(\widehat{\alpha}_n) \leq |||\widehat{m}(\alpha)|||^2_{\widehat{\Sigma}} + \lambda_n P(\alpha)$ for all $\alpha \in \mathcal{N}_{0n}$, we have: for all $\alpha \in \mathcal{N}_{0n}$,

$$C|||\widehat{m}(\cdot,\alpha_0)+m(\cdot,\widehat{\alpha}_n)|||^2_{\widehat{\Sigma}} + \lambda_n P(\widehat{\alpha}_n) \leq C|||\widehat{m}(\cdot,\alpha_0)+m(\cdot,\alpha)|||^2_{\widehat{\Sigma}} + \lambda_n P(\alpha) + o_p(n^{-1}).$$

Denote $l(\cdot,\alpha) \equiv \widehat{m}(\cdot,\alpha_0)+m(\cdot,\alpha)$. Then $|||l(\cdot,\alpha)|||^2_{\widehat{\Sigma}} + C^{-1}\lambda_n P(\alpha)$ is a smooth criterion function with $\widehat{\alpha}_n$ as its approximate minimizer.

Given that m is smooth, by assumptions 2.8(i) and 3.4(i), we can now mimic the proof strategy of Ai and Chen (2003) for asymptotic normality of $\widehat{\theta}_n$ using this new criterion function $(|||l(\cdot,\alpha)|||^2_{\widehat{\Sigma}} + C^{-1}\lambda_n P(\alpha))$. The rest of the proof has two discrepancies with that of Ai and Chen (2003). The first one is that we carry an error term $o_p(n^{-1})$, which turns out to be negligible. The second is that we now have sharper convergence rates of under the strong norm $||\cdot||_s$; hence we can relax some of their assumptions.

In what follows we use the fact that, for any vector x of dimension $1\times D$ (some $D \geq 1$) and any matrix A of $D\times D$, $x'Ax = tr(x'Ax) \leq tr(x'x)\sqrt{tr(A'A)} = x'x\sqrt{tr(A'A)}$ (see Newey (1997) equation A.11).

By assumption 3.4(i) and following Ai and Chen (2003) algebra, setting $0 < \epsilon_n = o(n^{-1/2})$ and $u_n^* = \pm v_n^*$ we have

$$0 \geq |||l(\widehat{\alpha}_n)|||^2_{\widehat{\Sigma}} - |||l(\widehat{\alpha}_n+\epsilon_n u_n^*)|||^2_{\widehat{\Sigma}} + C^{-1}\lambda_n\left(P(\widehat{\alpha}_n) - P(\widehat{\alpha}_n+\epsilon_n u_n^*)\right) + o_p(n^{-1}).$$

Doing a second order Taylor expansion to the penalization term, and by assumption 3.6, we have:

$$C^{-1}\lambda_n\left(\frac{dP(\widehat{\alpha}_n)}{d\alpha}[\epsilon_n u_n^*] + \frac{1}{2}\frac{d^2P(\alpha(s))}{d\alpha d\alpha}[\epsilon_n u_n^*, \epsilon_n u_n^*]\right) = \lambda_n\epsilon_n \times O_P(1) = o_p(n^{-1})$$

uniformly over $\alpha(s) = \widehat{\alpha}_n + s\epsilon_n u_n^* \in \mathcal{N}_{0n}$. After the second order Taylor expansion to the term $|||l(\widehat{\alpha}_n)|||^2_{\widehat{\Sigma}} - |||l(\widehat{\alpha}_n+\epsilon_n u_n^*)|||^2_{\widehat{\Sigma}}$, we have:

$$0 \leq \frac{\epsilon_n}{n}\sum_{i=1}^n\left(\frac{dm(X_i,\widehat{\alpha}_n)}{d\alpha}[u_n^*]\right)'\widehat{\Sigma}(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\right) + I_n(\alpha(s)) + II_n(\alpha(s)) + o_p(n^{-1}),$$

with $\alpha(s) = \widehat{\alpha}_n + s\epsilon_n u_n^* \in \mathcal{N}_{0n}$ for some $s \in (0,1)$, and

$$I_n(\alpha(s)) \equiv 2\frac{\epsilon_n^2}{n}\sum_{i=1}^n\left(\frac{d^2m(X_i,\alpha(s))}{d\alpha d\alpha}[u_n^*,u_n^*]\right)'\widehat{\Sigma}(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0)+m(X_i,\alpha(s))\right),$$

[5] In order to show this we can assume that $\sqrt{Z_n}$ goes faster to zero than $|||\widehat{m}(\alpha_0)+m(\alpha)|||_{\widehat{\Sigma}}$, otherwise the cross-product between this two terms if of order $o_p(n^{-1})$ and thus negligible.

$$II_n(\alpha(s)) \equiv 2\frac{\epsilon_n^2}{n}\sum_{i=1}^{n}\left(\frac{dm(X_i,\alpha(s))}{d\alpha}[u_n^*]\right)'\widehat{\Sigma}(X_i)^{-1}\left(\frac{dm(X_i,\alpha(s))}{d\alpha}[u_n^*]\right),$$

with. Applying Cauchy-Schwarz and assumptions 2.6 and 3.4(i), we have:

$$\sup_{\alpha\in\mathcal{N}_{0n}}|I_n(\alpha)| \leq const.\epsilon_n^2\sqrt{\sup_{\alpha\in\mathcal{N}_{0n}}\frac{1}{n}\sum_{i=1}^{n}\|\widehat{m}(X_i,\alpha_0)+m(X_i,\alpha)\|_E^2} = \epsilon_n^2\times O_P(\delta_{m,n}^*+\delta_n^*),$$

where the second equality is due to assumptions 2.7(i)(ii) and 2.8(ii), and the fact

$$\|\widehat{m}(X_i,\alpha_0)+m(X_i,\alpha)\|_E \leq \|\widehat{m}(X_i,\alpha_0)-m(X_i,\alpha_0)\|_E + \|m(X_i,\alpha_0)-m(X_i,\alpha)\|_E,$$

thus $\sup_{\alpha\in\mathcal{N}_{0n}}|I_n(\alpha)| \leq \epsilon_n^2\times o_P(n^{-1/4})$ by assumption 3.1(ii). Next, by assumption 2.6, we have: uniformly over $\alpha\in\mathcal{N}_{0n}$,

$$\begin{aligned}
|II_n(\alpha)| &\leq const.\epsilon_n^2 n^{-1}\sum_{i=1}^{n}\left\|\frac{dm(X_i,\alpha(s))}{d\alpha}[u_n^*]-\frac{dm(X_i,\alpha_0)}{d\alpha}[u_n^*]\right\|_E^2 \\
&\quad + const.\epsilon_n^2 n^{-1}\sum_{i=1}^{n}\left\|\frac{dm(X_i,\alpha_0)}{d\alpha}[u_n^*]\right\|_E^2 \\
&= o_p(n^{-1})+O_p(\epsilon_n^2),
\end{aligned}$$

where the second inequality follows from assumption 3.4(ii)(iii) and corollary C.1(ii) in Ai and Chen (2003) (for which all the needed assumptions are satisfied). Therefore, we have

$$0 \leq \frac{\epsilon_n}{n}\sum_{i=1}^{n}\left(\frac{dm(X_i,\widehat{\alpha}_n)}{d\alpha}[u_n^*]\right)'\widehat{\Sigma}(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\right)+O_p(\epsilon_n^2).$$

Now performing similar algebra to Ai and Chen (2003) we obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{dm(X_i,\widehat{\alpha}_n)}{d\alpha}[v_n^*]\right)'\widehat{\Sigma}(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\right)=o_p(1). \tag{A.1}$$

Note that, by assumption 2.6,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(X_i,\widehat{\alpha}_n)}{d\alpha}[v_n^*]-\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*]\right)'\widehat{\Sigma}(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\right)\right| \\
\leq\ & const.\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{dm(X_i,\widehat{\alpha}_n)}{d\alpha}[v_n^*]-\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*]\right\|_E^2}\times\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\|_E^2} \\
=\ & o_p(n^{-1/4})\times o_p(n^{-1/4})=o_p(n^{-1/2}),
\end{aligned}$$

where the first term is of order $o_p(n^{-1/4})$ by applying assumption 3.4(ii)(iii) and corollary C.1(ii) in Ai and Chen (2003), and we already established that the second term is of the same order $o_p(n^{-1/4})$. Thus, we obtain:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*]\right)'\widehat{\Sigma}(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\right)=o_p(1).$$

Note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*]\right)'\left(\widehat{\Sigma}(X_i)^{-1}-\Sigma(X_i)^{-1}\right)\left(\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\right)\right| \\
\leq\ & O_p(\delta_{\Sigma,n}^*)\times\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*]\right\|_E^2}\times\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\|_E^2} \\
\leq\ & O_p(\delta_{\Sigma,n}^*\times(\delta_n^*+\delta_{m,n}^*))
\end{aligned}$$

where the first inequality is obtained by the fact that, for vectors y, x of $1\times D$ and A of $D\times D$ it follows that $|y'Ax|=\sqrt{tr((y'Ax)^2)}\leq\sqrt{tr(yy')}\sqrt{tr((Ax)(Ax)')}\leq\sqrt{tr(yy')}\sqrt{tr((A'A))}\sqrt{(xx')}$; and the fact that $tr((\widehat{\Sigma}^{-1}(X)-\Sigma(X)^{-1})'(\widehat{\Sigma}^{-1}(X)-\Sigma(X)^{-1}))\leq\sum_{j=1}^{d_\rho}\sum_{k=1}^{d_\rho}(\sup_x|\widehat{\Sigma}^{-1}(x)_{[jk]}-\Sigma(x)^{-1}_{[jk]}|)^2=O_p((\delta_{\Sigma,n}^*)^2)$. The second inequality follows from assumptions 2.7(i)(ii) and 2.8(i). Thereby it follows

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*]\right)'\Sigma(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\right)=o_p(1).$$

Notice that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*-v^*]\right)'\Sigma(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\right)\right| \\
\leq\ & \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*-v^*]\right\|_E^2}\times O_P(\delta_n^*+\delta_{m,n}^*) \\
=\ & O_P\left(||v_n^*-v^*||\right)\times O_P(\delta_n^*+\delta_{m,n}^*)=o_p(n^{-1/2}),
\end{aligned}$$

where the second equality is due to Markov inequality and i.i.d. data, and the last equality is due to assumptions 3.1(ii) and 3.3. Thus,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v^*]\right)'\Sigma(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\right)=o_p(1).$$

Now define $\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v^*]\right)'\Sigma(X_i)^{-1}$ as $g(X_i,v^*)$. Notice that $|g(X,v^*)m(X,\alpha)-g(X,v^*)m(X,\alpha_0)|\leq |g(X,v^*)|\times|m(X,\alpha)-m(X,\alpha_0)|$. Thus given that $E[|g(X,v^*)|^2]<M$ by assumption 2.9(i) and the fact we are in a shrinking neighborhood of α_0 it follows that the entropy under the $L^2(\mathcal{X})$ norm of $\{g(X,v^*)m(X,\alpha):\alpha\in\mathcal{N}_{0n}\}$ is bounded by $\{m(X,\alpha):\alpha\in\mathcal{N}_{0n}\}$ which satisfies Donsker property by assumption 3.4(iv). Therefore it follows

$$n^{-1}\sum_{i=1}^{n}g(X_i,v^*)m(X_i,\alpha)=E\left[g(X,v^*)(m(X,\alpha)-m(X,\alpha_0))\right]+o_p(n^{-1/2}).$$

By applying the mean value theorem to $(m(X,\alpha)-m(X,\alpha_0))$ and assumption 3.5 we obtain:

$$n^{-1}\sum_{i=1}^{n}g(X_i,v^*)m(X_i,\widehat{\alpha}_n)=\langle v^*,\widehat{\alpha}_n-\alpha_0\rangle+o_p(n^{-1/2}),$$

and

$$\sqrt{n}\langle v^*,\widehat{\alpha}_n-\alpha_0\rangle=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v^*]\right)'\Sigma(X_i)^{-1}\widehat{m}(X_i,\alpha_0)+o_p(1)$$

Finally by assumption 3.1(iii), we obtain

$$\sqrt{n}\langle v^*,\widehat{\alpha}_n-\alpha_0\rangle=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v^*]\right)'\Sigma(X_i)^{-1}\rho(Z_i,\alpha_0)+o_p(1) \tag{A.2}$$

and the result follows by applying a standard central limit theorem argument. *Q.E.D*

Proof of Proposition 3.1: (1) Let $\varepsilon(Z,\alpha)\equiv\rho(Z,\alpha)-m(X,\alpha)$, $\Lambda_n(X)\equiv E[(\rho(Z,\alpha)-\rho(Z,\alpha_0))^2|X]$.

Then

$$\begin{aligned}
& \sup_{\alpha\in\mathcal{N}_{0n}} \left(E\left[p^{J_n}(X_i)(P'P)^{-1}P'(\Delta\varepsilon(\alpha))(\Delta\varepsilon(\alpha))'P(P'P)^{-1}p^{J_n}(X_i)'\right]\right)^{1/2} \\
\leq & \sup_{\alpha\in\mathcal{N}_{0n}} \left(E\left[p^{J_n}(X_i)(P'P)^{-1}P'E\left[(\Delta\varepsilon(\alpha))(\Delta\varepsilon(\alpha))'|X_1,\ldots,X_n\right]P(P'P)^{-1}p^{J_n}(X_i)'\right]\right)^{1/2} \\
\leq & \sup_{\alpha\in\mathcal{N}_{0n}} \left(E\left[\Lambda_n \times Tr\left\{n^{-1}p^{J_n}(X_i)'p^{J_n}(X_i)(P'P/n)^{-1}\right\}\right]\right)^{1/2} \\
\leq & K \sup_{\alpha\in\mathcal{N}_{0n}} \left(E\left[E\left[\left(\rho\left(Z_i,\alpha\right)-\rho\left(Z_i,\alpha_0\right)\right)^2|X\right]\right]\times\frac{J_n}{n}\right)^{1/2} \\
\leq & K \sup_{\alpha\in\mathcal{N}_{0n}} \sqrt{\frac{J_n}{n}}\left\|\alpha-\alpha_0\right\|_s^{\kappa} \leq O_p\left(\sqrt{\frac{J_n}{n}}\left(\delta_{s,n}^*\right)^{\kappa}\right),
\end{aligned}$$

where the above inequalities are due to our assumptions 2.5, 3.7(ii), 2.12, 3.7(i), and the definition of $\mathcal{N}_{0n}$. Since

$$\begin{aligned}
& \sup_{\alpha\in\mathcal{N}_{0n}} n^{-1}\sum_{i=1}^n \left\|\widehat{m}(X_i,\alpha)-\widehat{m}(X_i,\alpha_0)-m(X_i,\alpha)\right\|_E^2 \\
\leq & 2\sup_{\alpha\in\mathcal{N}_{0n}} n^{-1}\sum_{i=1}^n \left\|\widehat{m}(X_i,\alpha)-\widehat{m}(X_i,\alpha_0)-\widetilde{m}(X_i,\alpha)\right\|_E^2 + 2\sup_{\alpha\in\mathcal{N}_{0n}} n^{-1}\sum_{i=1}^n \left\|\widetilde{m}(X_i,\alpha)-m(X_i,\alpha)\right\|_E^2 \\
\leq & O_p\left(\frac{J_n}{n}\left(\delta_{s,n}^*\right)^{2\kappa}\right) + O_P\left(E[\sup_{\alpha\in\mathcal{N}_{0n}}\left\|\widetilde{m}(X_i,\alpha)-m(X_i,\alpha)\right\|_E^2]\right) \\
= & O_P\left(\max\left\{\frac{J_n}{n}\left(\delta_{s,n}^*\right)^{2\kappa}, b_{m,J_n}^2\right\}\right),
\end{aligned}$$

where the second inequality is due to Markov inequality and i.i.d. data, and the last equality is due to assumption 2.13. Therefore, assumption 3.1(i) holds provided that $\max\left\{\frac{J_n}{n}\left(\delta_{s,n}^*\right)^{2\kappa}, b_{m,J_n}^2\right\} = o_p(n^{-1})$.

(2) By Lemma 2.4 we obtain assumptions 3.1(ii) and 2.7, with $\delta_n^* \asymp \delta_{m,n}^* = o_P(n^{-1/4})$ provided that $J_n \asymp k(n) = o(n^{1/2})$. Assumption 3.1(iii) is satisfied by applying corollary C.3(iii) in Ai and Chen (2003). Now the asymptotic normality result follows directly from Theorem 3.1. *Q.E.D*

Proof of Theorem 3.2: Recall that under our assumption on penalization function, we have $\lambda_n P(h) - \lambda_n P(\widehat{h}_n) = o_P(n^{-1})$ uniformly over $\alpha \in \mathcal{N}_{0n}$. Thus we could define $\widehat{\alpha}_n^* \equiv \left(\widehat{\theta}_n^*, \widehat{h}_n^*\right)$ as

$$\widehat{\alpha}_n^* = \arg\inf_{\alpha\in\mathcal{N}_{0n}}\left\{\frac{1}{n}\sum_{i=1}^n W_i\left\{\widehat{m}(X_i,\alpha)'[\widehat{\Sigma}(X_i)]^{-1}\widehat{m}(X_i,\alpha)\right\} + o_P(n^{-1})\right\},$$

with $\{W_i\}_{i=1}^n$ being a random sample of positive weights such that $E[W_i] = 1$, $Var(W_i) = w_0$ and are independent from the sample $\{(Y_i, X_i)\}_{i=1}^n$. We establish the conclusion in two steps.

Step 1: We first obtain the asymptotic distribution for $\sqrt{n}(\widehat{\theta}_n^* - \theta_0)$. We will derive this in the

same way as we derive the asymptotic distribution for $\sqrt{n}(\widehat{\theta}_n - \theta_0)$. As in the proof of Theorem 3.1, we first need to establish that $\widehat{\alpha}_n^*$ is an "approximate minimizer" of a smooth criterion function: $n^{-1}\sum_{i=1}^n l_W(X_i,\alpha)'[\widehat{\Sigma}(X_i)^{-1}]l_W(X_i,\alpha)$, with $l_W(X_i,\alpha) \equiv \sqrt{W_i}(m(X_i,\alpha) + \widehat{m}(X_i,\alpha_0))$. Define another norm: $|||\cdot|||_{W,I}^2 \equiv n^{-1}\sum_{i=1}^n W_i||\cdot||_E^2$. By assumption 2.6(i)(ii), we have: $|||\cdot|||_{W,\widehat{\Sigma}}^2 \leq const.|||\cdot|||_{W,I}^2$, and by assumption 3.9(i),

$$\begin{aligned} & \sup_{\alpha\in\mathcal{N}_{0n}} |||\widehat{m}(\cdot,\alpha) - \widehat{m}(\cdot,\alpha_0) - m(\cdot,\alpha)|||_{W,I}^2 \\ = & \sup_{\alpha\in\mathcal{N}_{0n}} \left\{ \frac{1}{n}\sum_{i=1}^n W_i \left\|\widehat{m}(X_i,\alpha) - \widehat{m}(X_i,\alpha_0) - m(X_i,\alpha)\right\|_E^2 \right\} = o_P(n^{-1}). \end{aligned}$$

Thus, performing analogous algebra to the ones in the proof of Theorem 3.1, we can think of $\widehat{\alpha}_n^*$ as the (approximate) minimizer of

$$n^{-1}\sum_{i=1}^n l_W(X_i,\alpha)'\widehat{\Sigma}(X_i)^{-1}l_W(X_i,\alpha).$$

Now using analogous arguments to the ones in the proof of Theorem 3.1, and by assumption 3.8(i) and Cauchy-Schwarz inequality, we obtain:

$$2\frac{\epsilon_n^2}{n}\sum_{i=1}^n W_i \left(\frac{d^2m(X_i,\alpha(s))}{d\alpha d\alpha}[u_n^*,u_n^*]\right)' \widehat{\Sigma}(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0) + m(X_i,\alpha(s))\right) = O_p(\epsilon_n^2),$$

and

$$2\frac{\epsilon_n^2}{n}\sum_{i=1}^n W_i \left(\frac{dm(X_i,\alpha(s))}{d\alpha}[u_n^*]\right)' \widehat{\Sigma}(X_i)^{-1}\left(\frac{dm(X_i,\alpha(s))}{d\alpha}[u_n^*]\right) = O_p(\epsilon_n^2).$$

Thus it follows

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i \left(\frac{dm(X_i,\widehat{\alpha}_n^*)}{d\alpha}[v_n^*]\right)' \widehat{\Sigma}(X_i)^{-1}(\widehat{m}(X_i,\alpha_0) + m(X_i,\widehat{\alpha}_n^*)) = o_p(1).$$

Note that

$$n^{-1}\sum_{i=1}^n W_i^2 \left\|\frac{dm(X_i,\widehat{\alpha}_n^*)}{d\alpha}[v_n^*] - \frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*]\right\|_E^2 = o_p(n^{-1/2})$$

by assumption 3.4(iii) and the fact that W are independent with finite second moment. This and assumption 2.6, following the same steps as in the proof of Theorem 3.1, imply

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*]\right)' \Sigma(X_i)^{-1}(\widehat{m}(X_i,\alpha_0) + m(X_i,\widehat{\alpha}_n^*)) = o_p(1).$$

By Markov inequality, i.i.d. data and $\{W_i\}_{i=1}^n$ is independent of $\{(Y_i, X_i)\}_{i=1}^n$, we have:

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^n W_i^2 \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^* - v^*]\right)' \Sigma(X_i)^{-1} \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^* - v^*]\right) \\
& \leq O_P\left(E\left[W_i^2 \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^* - v^*]\right)' \Sigma(X_i)^{-1} \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^* - v^*]\right)\right]\right) \\
& = O_P\left(E[W^2] \times ||v_n^* - v^*||^2\right).
\end{aligned}
$$

Therefore, by repeating the steps in the proof of Theorem 3.1, we obtain

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v^*]\right)' \Sigma(X_i)^{-1}(\widehat{m}(X_i,\alpha_0) + m(X_i,\widehat{\alpha}_n^*)) = o_p(1).
$$

Defining $g(W, X, v^*) = W\left(\frac{dm(X,\alpha_0)}{d\alpha}[v^*]\right)' \Sigma(X)^{-1}$ it follows that $\{g(W, X, v^*)m(X, \alpha) : \alpha \in \mathcal{N}_0\}$ is a Donsker Class by our assumptions.[6] Thus, with $m(X_i, \alpha_0) = 0$, we have uniformly over $\alpha \in \mathcal{N}_0$,

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^n g(W_i, X_i, v^*)m(X_i, \alpha) \\
& = E\left[g(W, X, v^*)m(X, \alpha)\right] + o_p(n^{-1/2}) = E\left[W\right] E\left[g(X, v^*)m(X, \alpha)\right] + o_p(n^{-1/2}) \\
& = E\left[g(X, v^*)m(X, \alpha)\right] + o_p(n^{-1/2}) = \langle v^*, \alpha - \alpha_0\rangle + o_p(n^{-1/2})
\end{aligned}
$$

where the second equality follows from the fact that W is independent by assumption 3.8 and $E[W] = 1$, and the last equality follows from the same calculations in the proof of Theorem 3.1. Thus

$$
\sqrt{n}\langle v^*, \widehat{\alpha}_n^* - \alpha_0\rangle = -\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i \left(\frac{dm(X,\alpha_0)}{d\alpha}[v^*]\right)' \Sigma(X)^{-1}\widehat{m}(X_i,\alpha_0) + o_p(1),
$$

this and assumption 3.9(ii) imply:

$$
\sqrt{n}\langle v^*, \widehat{\alpha}_n^* - \alpha_0\rangle = -\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i \left(\frac{dm(X,\alpha_0)}{d\alpha}[v^*]\right)' \Sigma(X)^{-1}\rho(Z_i,\alpha_0) + o_p(1), \tag{A.3}
$$

and hence $\sqrt{n}(\widehat{\theta}_n^* - \theta_0)$ is asymptotically normal with zero mean and variance

$$
V_*^{-1} \equiv w_0 V^{-1}.
$$

This follows from the fact that W is an independent random variable.

[6] We already established that $\{g(X, v^*)m(X, \alpha) : \alpha \in \mathcal{N}_{0n}\}$ is a Donsker Class. Given that $E[W^2]$ is finite we can use the same argument as in, say, Ai and Chen (2003) p. 1832.

STEP 2: Subtracting equation (A.2) from (A.3), we obtain:

$$\sqrt{n}\langle v^*, \widehat{\alpha}_n^* - \widehat{\alpha}_n\rangle = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_i - 1)\left(\frac{dm(X,\alpha_0)}{d\alpha}[v^*]\right)' \Sigma(X)^{-1}\rho(Z_i,\alpha_0) + o_p(1).$$

Given that $Var(W-1) = Var(W) = w_0$ and that $\{W_i\}_{i=1}^n$ is independent of $\{(Y_i, X_i)\}_{i=1}^n$, it follows that, conditional on the data $\{(Y_i, X_i)\}_{i=1}^n$, $\sqrt{\frac{n}{w_0}}\left(\widehat{\theta}_n^* - \widehat{\theta}_n\right)$ is asymptotically normal with zero mean and variance V^{-1}, the same limiting distribution as that of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$. *Q.E.D*

PROOF OF THEOREM 4.1: We essentially replicate the proof of theorem 6.1 in Ai and Chen (2003), except that we avoid their use of the differentiability of the generalized residual function $\rho(Z,\alpha)$ with respect to α in a shrinking neighborhood of α_0.

As in their paper first we let $u = \rho(Z,\alpha_0)$ and divide Y into (Y_1, Y_2) with $\dim(Y_1) = \dim(u) = d_\rho$. We will follow Newey (1990) characterizing the tangent set and then defining a projection of the score of θ onto it. The non-parametric parts that will be approximated by parametric submodels can be divided into h, $f_0(u, y_2, x)$ and $g_0(x)$ where f_0 is the true conditional density of (u, Y_2) given $X = x$, and g_0 is the true marginal density of X. For the h part, we define the parametric submodel as $h_0 + \xi_1(h - h_0)$, for the g part, we define the parametric submodel as $g(x,\xi_3) \equiv g_0(x)(1 + \xi_3 \times D_1(x))$ for a bounded $D_1(x)$ with $E[D_1(X)] = 0$, notice that $g(x,\xi_3)$ is a density for a sufficiently small ξ_3. For the $f(u, y_2, x)$ part, we define the parametric submodel as $f(u, y_2, x, \xi_2) \equiv f_0(u, y_2, x)\Delta_f(u, y_2, x, \xi_2)$, where $\Delta_f(u, y_2, x, \xi_2) = 1 + \xi_2 \times D_2(u, y_2, x)$ and $D_2(\cdot)$ has to be such that (i) $E[D_2|X] = 0$; (ii) D_2 is bounded and (iii) $E[uD_2|X] = 0$. With (i)-(ii) we ensure that $f(u, y_2, x, \xi_2)$ is a density for sufficiently small values of ξ_2, and (iii) imposes the model restriction. Define $\Delta(y, x, \delta) = \Delta_f(\rho(z, \theta, h_0 + \xi_1(h - h_0)), y_2, x, \xi_2) \times (1 + \xi_3 \times D_1(x))$ with $\delta \equiv (\theta', \xi_1, \xi_2, \xi_3)'$. Then

$$q(y, x, \delta) \equiv q_0(y, x, \theta, h_0 + \xi_1(h - h_0))\Delta(y, x, \delta)$$

is a parametric submodel passing through the true model.

In order to show that these submodels are "smooth" we are going to replace the requirement in Ai and Chen (2003) p. 1838 with the differentiability notion used in Van der Vaart (1991):

$$\int \left[\frac{\Delta(y, x, \delta_t)^{1/2} - \Delta(y, x, \delta_0)^{1/2}}{t} - \frac{1}{2}g\Delta(y, x, \delta_0)^{1/2}\right]^2 d\mu \to 0$$

where δ_t converges to $\delta_0 \equiv (\theta_0, 0, 0, 0)$ and $g \in L^2(q_0)$. In our case we will define g as $(q_{0\theta}/q_0, q_{0h}/q_0, D_2, D_1)$ (we leave implicit the dependence of (y, x)), and thus is easy to see that under the assumptions over D_1, D_2 and q_0, g belong to $L^2(q_0)$.

The projection of the score function corresponding to θ_j (denoted as $S_{\theta_j} = q_{0\theta_j}(\cdot)/q_0(\cdot)$) onto $D_1(x)$

is naught, as $D_1(x)$ is not informative about θ_j. For D_2, define the tangent space as $\Lambda_2 \equiv \{D_2 :$ (i),(ii),(iii) are satisfied$\}$. In order to compute the projection we solve

$$\min_{D_2 \in \Lambda_2} E\left[\left(S_{\theta_j}(z) - \frac{q_{0h}(z,\alpha_0,[w])}{q_0(z,\alpha_0)} - D_2(u,y_2,x)\right)^2\right].$$

With the solution to this problem in hand (see Ai and Chen (2003) p. 1839) and defining

$$\begin{aligned} E[S_{\theta_j}(y,x)u|x] &\equiv \frac{dE[\rho(Z,\alpha_0)|x]}{d\theta_j}, \\ E\left[\frac{q_{0h}(y,x,\alpha_0,[w])}{q_0(y,x,\alpha_0)}u|x\right] &\equiv \frac{dE[\rho(Z,\alpha_0)|x]}{dh}[w]. \end{aligned}$$

We can now follow the rest of the proof of their theorem 6.1 in Ai and Chen (2003), except replacing their $E\{\frac{d\rho(Z,\alpha_0)}{d\theta_j}|x\}$ and $E\{\frac{d\rho(Z,\alpha_0)}{dh}[w]|x\}$ by our $\frac{dE[\rho(Z,\alpha_0)|x]}{d\theta_j}$ and $\frac{dE[\rho(Z,\alpha_0)|x]}{dh}[w]$ respectively. *Q.E.D*

Proof of Theorem 4.2: For any $\lambda \neq 0$, let $||v_0||^2 = \lambda'(E\left[(D_{w_0}(X))'\Sigma_0(X)^{-1}(D_{w_0}(X))\right])\lambda$ (i.e., we use the optimal weighting $\Sigma_0(X)$ instead of $\Sigma(X)$), and $\alpha^* \equiv \widehat{\alpha}_n - \langle\widehat{\alpha}_n - \alpha_0, v_0\rangle v_0/||v_0||^2$ with $\Pi_n\alpha^* \equiv \widehat{\alpha}_n - \langle\widehat{\alpha}_n - \alpha_0, v_0\rangle\Pi_n v_0/||v_0||^2$. Then $\widehat{\alpha}_n - \Pi_n\alpha^* = \langle\widehat{\alpha}_n - \alpha_0, v_0\rangle\Pi_n v_0/||v_0||^2$.

Note that assumption 3.6 implies $\lambda_n\{P(\widehat{\alpha}) - P(\Pi_n\alpha_n^*)\} = o_P(n^{-1})$. Note also that by assumption 3.4(i) and second order Taylor expansion,

$$|||l(\widehat{\alpha}_n)|||^2_{\widehat{\Sigma}} - |||l(\Pi_n\alpha_n^*)|||^2_{\widehat{\Sigma}} = \frac{d|||l(\Pi_n\alpha^*)|||^2_{\widehat{\Sigma}}}{d\alpha} + \frac{1}{2}\frac{d^2|||l(\overline{\alpha}_n)|||^2_{\widehat{\Sigma}}}{d\alpha d\alpha},$$

with $\overline{\alpha}_n$ a point in between $\widehat{\alpha}_n$ and $\Pi_n\alpha^*$. Note that the second derivative term $\frac{d^2|||l(\overline{\alpha}_n)|||^2_{\widehat{\Sigma}}}{d\alpha d\alpha} = I_n(\overline{\alpha}_n) + II_n(\overline{\alpha}_n)$, with[7]

$$I_n(\overline{\alpha}_n) = n^{-1}\sum_{i=1}^n \left(\frac{d^2m(X_i,\overline{\alpha}_n)}{d\alpha d\alpha}[\widehat{\alpha}_n - \Pi_n\alpha^*, \widehat{\alpha}_n - \Pi_n\alpha^*]\right)' \widehat{\Sigma}_0(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0) + m(X_i,\overline{\alpha}_n)\right),$$

$$II_n(\overline{\alpha}_n) = n^{-1}\sum_{i=1}^n \left(\frac{dm(X_i,\overline{\alpha}_n)}{d\alpha}[\widehat{\alpha}_n - \Pi_n\alpha^*]\right)' \widehat{\Sigma}_0(X_i)^{-1}\left(\frac{dm(X_i,\overline{\alpha}_n)}{d\alpha}[\widehat{\alpha}_n - \Pi_n\alpha^*]\right).$$

Following the same calculations in the proof of Theorem 3.1 and by assumption 4.2, we have:

$$\sup_{\overline{\alpha}_n \in \mathcal{N}_{0n}} |I_n(\overline{\alpha}_n)| = o_p(n^{-1}).$$

[7] We abuse notation by denoting $\widehat{\Sigma}$ as $\widehat{\Sigma}(X,\widetilde{\alpha}_n)$, $\Sigma(X) = \Sigma(X,\alpha)$ and $\Sigma_0(X) = \Sigma(X,\alpha_0)$.

Similarly under assumption 4.4(i)(ii), we have:

$$\begin{aligned} & II_n(\overline{\alpha}_n) \\ = \; & \frac{\langle \widehat{\alpha}_n - \alpha_0, v_0 \rangle}{||v_0||^2} E\left[\left(\frac{dm(X,\alpha_0)}{d\alpha}[v_0] \right)' \Sigma_0(X)^{-1} \left(\frac{dm(X,\alpha_0)}{d\alpha}[v_0] \right) \right] \frac{\langle \widehat{\alpha}_n - \alpha_0, v_0 \rangle}{||v_0||^2} + o_p(n^{-1}). \end{aligned}$$

The first derivative term $\frac{d|||l(\Pi_n\alpha^*)|||^2_{\widehat{\Sigma}}}{d\alpha}$ is given by

$$\frac{d|||l(\Pi_n\alpha^*)|||^2_{\widehat{\Sigma}}}{d\alpha} = n^{-1} \sum_{i=1}^n \left(\frac{dm(X_i,\Pi_n\alpha^*)}{d\alpha}[\widehat{\alpha}_n - \Pi_n\alpha^*] \right)' \widehat{\Sigma}_0(X_i)^{-1} \left(\widehat{m}(X_i,\alpha_0) + m(X_i,\Pi_n\alpha^*) \right).$$

By Cauchy-Schwarz inequality, $\langle v_0, \widehat{\alpha}_n - \alpha_0 \rangle = O_p(n^{-1/2})$, assumptions 4.3 and 4.4, and using the same arguments as the ones in the proof of Theorem 3.1, we obtain:

$$\begin{aligned} & \frac{d|||l(\Pi_n\alpha^*)|||^2_{\widehat{\Sigma}}}{d\alpha} \\ = \; & n^{-1} \sum_{i=1}^n \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[\widehat{\alpha}_n - \Pi_n\alpha^*] \right)' \Sigma_0(X_i)^{-1} \left(\widehat{m}(X_i,\alpha_0) + m(X_i,\Pi_n\alpha^*) \right) + o_p(n^{-1}). \end{aligned}$$

Now, we study

$$\begin{aligned} & n^{-1} \sum_{i=1}^n \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[\widehat{\alpha}_n - \Pi_n\alpha^*] \right)' \Sigma_0(X_i)^{-1} m(X_i,\Pi_n\alpha^*) \\ = \; & O_p(n^{-1/2}) \times \left(\langle \Pi_n\alpha^* - \alpha_0, v_0 \rangle + o_p(n^{-1/2}) \right) \\ = \; & O_p(n^{-1/2}) \times \left(\langle \widehat{\alpha}_n - \alpha_0, v_0 \rangle - \frac{\langle \widehat{\alpha}_n - \alpha_0, v_0 \rangle \langle \Pi_n v_0, v_0 \rangle}{||v_0||^2} + o_p(n^{-1/2}) \right) \\ = \; & O_p(n^{-1/2}) \times \left(\langle \widehat{\alpha}_n - \alpha_0, v_0 \rangle - \langle \widehat{\alpha}_n - \alpha_0, v_0 \rangle + o_p(n^{-1/2}) \right) = o_p(n^{-1}), \end{aligned}$$

where the third equality uses the fact that $\langle \Pi_n v_0 - v_0, v_0 \rangle \leq K \times ||\Pi_n v_0 - v_0|| = o_p(1)$ by assumption

3.3(i). Therefore

$$
\begin{aligned}
& |||l(\widehat{\alpha}_n)|||^2_{\widehat{\Sigma}} - |||l(\Pi_n\alpha^*_n)|||^2_{\widehat{\Sigma}} \\
= \; & n^{-1}\sum_{i=1}^{n}\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[\widehat{\alpha}_n - \Pi_n\alpha^*]\right)'\Sigma_0(X_i)^{-1}\widehat{m}(X_i,\alpha_0) \\
& + \frac{1}{2}\left(\frac{\langle\widehat{\alpha}_n-\alpha_0,v_0\rangle}{||v_0||^2}\right)^2 E\left[\left(\frac{dm(X,\alpha_0)}{d\alpha}[v_0]\right)'\Sigma_0(X)^{-1}\left(\frac{dm(X,\alpha_0)}{d\alpha}[v_0]\right)\right] + o_p(n^{-1}) \\
= \; & -\langle\widehat{\alpha}_n-\alpha_0,v_0\rangle\frac{1}{||v_0||^2}\times\langle\widehat{\alpha}_n-\alpha_0,v_0\rangle + \frac{1}{2}\left(\frac{\langle\widehat{\alpha}_n-\alpha_0,v_0\rangle}{||v_0||^2}\right)^2\times||v_0||^2 + o_p(n^{-1}) \\
= \; & -\frac{1}{2}\langle\widehat{\alpha}_n-\alpha_0,v_0\rangle^2\frac{1}{||v_0||^2} + o_p(n^{-1}),
\end{aligned}
$$

where the first term on the right hand side of the first equality follows from the last result obtained in the asymptotic normality proof.

Next, let $\widetilde{\alpha}_n$ as the minimizer of $|||l(\alpha)|||^2_{\widehat{\Sigma}}$ but subject to $\theta = \theta_0$. Define $\alpha^{**} = \widetilde{\alpha}_n + \langle\widehat{\alpha}_n - \alpha_0, v_0\rangle v_0/||v_0||^2$ and $\Pi_n\alpha^{**} = \widetilde{\alpha}_n + \langle\widehat{\alpha}_n-\alpha_0,v_0\rangle\Pi_n v_0/||v_0||^2$. Note that $\widetilde{\alpha}_n - \Pi_n\alpha^{**} = -(\widehat{\alpha}_n - \Pi_n\alpha^*)$. With these definitions, following the same calculations as before, we obtain:

$$
|||l(\widetilde{\alpha}_n)|||^2_{\widehat{\Sigma}} - |||l(\Pi_n\alpha^{**})|||^2_{\widehat{\Sigma}} = \frac{1}{2}\langle\widehat{\alpha}_n-\alpha_0,v_0\rangle^2\frac{1}{||v_0||^2} + o_p(n^{-1}).
$$

Now following the same steps as in Shen and Shi (2005), we obtain:

$$
2n\left(|||l(\widetilde{\alpha}_n)|||^2_{\widehat{\Sigma}} - |||l(\widehat{\alpha}_n)|||^2_{\widehat{\Sigma}}\right) = \left(\frac{\sqrt{n}\langle\widehat{\alpha}_n-\alpha_0,v_0\rangle}{||v_0||}\right)^2 + o_p(1).
$$

From the asymptotic normality proof of Theorem 3.1, we have: $\sqrt{n}\langle\widehat{\alpha}_n-\alpha_0,v_0\rangle \Rightarrow N(0,||v_0||^2)$. Hence the conclusion follows. *Q.E.D*

Proof of Theorem 4.3: We already shown that the criterion $n^{-1}\sum_{i=1}^n \widehat{m}(X_i,\alpha)'\left[\widehat{\Sigma}(X,\alpha)\right]^{-1}\widehat{m}(X_i,\alpha)$ is equivalent to $n^{-1}\sum_{i=1}^n l(X_i,\alpha)'\left[\widehat{\Sigma}(X,\alpha)\right]^{-1} l(X_i,\alpha) + o_p(n^{-1})$ uniformly over $\alpha\in\mathcal{N}_{0n}$. Define another "smooth criterion function"

$$
\widehat{L}^C_n(\alpha) \equiv n^{-1}\sum_{i=1}^{n} l(X_i,\alpha)'\left[\Sigma(X,\alpha)\right]^{-1} l(X_i,\alpha).
$$

Since

$$\begin{aligned}
&\sup_{\alpha\in\mathcal{N}_{0n}}\left|n^{-1}\sum_{i=1}^{n}l(X_i,\alpha)'\left(\widehat{\Sigma}(X,\alpha)^{-1}-\Sigma(X,\alpha)^{-1}\right)l(X_i,\alpha)\right| \\
\leq\ &\left\{\sup_{\alpha,x}\left|\widehat{\Sigma}(x,\alpha)^{-1}-\Sigma(x,\alpha)^{-1}\right|\right\}\times\sup_{\alpha\in\mathcal{N}_{0n}}n^{-1}\sum_{i=1}^{n}||l(X_i,\alpha)||_E^2 \\
=\ &\left\{\sup_{\alpha,x}\left|\widehat{\Sigma}(x,\alpha)^{-1}-\Sigma(x,\alpha)^{-1}\right|\right\}\times O_p((\delta_n^*+\delta_m^*)^2)=o_p(n^{-1}),
\end{aligned}$$

where the second equality is due to assumptions 4.2 and 4.5(i). We thus establish that our continuously updated criterion function is equivalent (up to $o_p(n^{-1})$) to $\widehat{L}_n^C(\alpha)$.

Following the steps in the proof of theorem 4.2, we perform a second order Taylor expansion of $\widehat{L}_n^C(\widehat{\alpha}_n)$ around $\Pi_n\alpha_n^*$. The same arguments hold in this case, but now we need to control extra terms related to the derivatives of $\Sigma(X,\alpha)$, which are well-defined by assumption 4.5(ii).

For the first derivative of $\widehat{L}_n^C$ the only extra term we have (with respect the proof of theorem 4.2) is

$$n^{-1}\sum_{i=1}^{n}l(X_i,\Pi_n\alpha^*)'\left(\frac{d\Sigma(X,\Pi_n\alpha^*)}{d\alpha}^{-1}[\widehat{\alpha}_n-\Pi_n\alpha^*]\right)l(X_i,\Pi_n\alpha^*). \tag{A.4}$$

By assumption 4.5(ii) the term in the middle is of $O_p(n^{-1/2})$ (recall the definition of $\Pi_n\alpha^*$). Moreover we know that $n^{-1}\sum_{i=1}^{n}||l(X_i,\alpha)||_E^2=O_p((\delta_n^*+\delta_m^*)^2)$, which by assumption 3.1(ii) it implies that the whole expression is of order $o_p(n^{-1})$.

Regarding the second derivative of $\widehat{L}_n^C$ we have two extra terms

$$n^{-1}\sum_{i=1}^{n}\frac{dl(X_i,\overline{\alpha}_n)'}{d\alpha}[\widehat{\alpha}_n-\Pi_n\alpha^*]\left(\frac{d\Sigma(X,\overline{\alpha}_n)}{d\alpha}^{-1}[\widehat{\alpha}_n-\Pi_n\alpha^*]\right)l(X_i,\overline{\alpha}_n)$$

and

$$n^{-1}\sum_{i=1}^{n}l(X_i,\overline{\alpha}_n)'\left(\frac{d^2\Sigma(X,\overline{\alpha}_n)}{d\alpha d\alpha}^{-1}[\widehat{\alpha}_n-\Pi_n\alpha^*]\right)l(X_i,\overline{\alpha}_n).$$

Again by our assumptions is it easy to see that both term are of order $o_p(n^{-1})$. Therefore second order Taylor expansion equals

$$\begin{aligned}
&\widehat{L}_n^C(\widehat{\alpha}_n)-\widehat{L}_n^C(\Pi_n\alpha^*) \\
&=n^{-1}\sum_{i=1}^{n}\frac{dl(X_i,\Pi_n\alpha^*)'}{d\alpha}[\widehat{\alpha}_n-\Pi_n\alpha^*]\left(\Sigma(X,\Pi_n\alpha^*)^{-1}\right)l(X_i,\Pi_n\alpha^*) \\
&+\frac{1}{2}n^{-1}\sum_{i=1}^{n}\frac{dl(X_i,\Pi_n\alpha^*)'}{d\alpha}[\widehat{\alpha}_n-\Pi_n\alpha^*]\left(\Sigma(X,\Pi_n\alpha^*)^{-1}\right)\frac{dl(X_i,\Pi_n\alpha^*)}{d\alpha}[\widehat{\alpha}_n-\Pi_n\alpha^*].
\end{aligned}$$

By assumptions 4.3, 4.4(iii) we can show by taking analogous steps to the ones in asymptotic normality proof that the first term equals

$$n^{-1}\sum_{i=1}^{n}\frac{dl(X_i,\alpha_0)'}{d\alpha}[\widehat{\alpha}_n-\alpha^*]\left(\Sigma(X,\alpha_0)^{-1}\right)l(X_i,\alpha_0).$$

On the other hand the second term, using a similar argument to the one above equals

$$\frac{1}{2}E\left[\frac{dl(X_i,\alpha_0)'}{d\alpha}[\widehat{\alpha}_n-\alpha^*]\left(\Sigma(X,\alpha_0)^{-1}\right)\frac{dl(X_i,\alpha_0)}{d\alpha}[\widehat{\alpha}_n-\alpha^*]\right].$$

We thus arrived to the same result as the one in the proof of theorem 4.2. The desired result now follows. *Q.E.D*

Proof of Proposition 5.1: We obtain the result by applying Proposition 3.1. Assumptions 2.1(i) and (ii) are implied by conditions 5.1(i)(ii) and (iii) respectively. Assumption 2.2 follows from conditions 5.1(i)(ii) and 5.2. Given condition 5.3(i) and the choices of our parameter space $\mathcal{H}$ and sieve space $\mathcal{H}_n$, the PSMD estimator $\widehat{\alpha}_n$ is well-defined with probability approaching one. Conditions 5.1(i)(ii), 5.3(i) and 5.4(i) imply assumption 2.4. Given that $\widehat{\Sigma}=\gamma(1-\gamma)$, assumption 2.6 is redundant. We apply Lemma 2.4(1) to verify assumption 2.7. Assumptions 2.5 and 2.12 are directly imposed. Assumption 2.13(i) is trivially satisfied as $0\leq\rho(Z,\alpha)\leq 1$ for all Z and all α. Condition 5.3(ii) implies that assumption 2.13(ii) holds with $b_{m,J_n}=(J_n)^{-r_m}$. Thus Lemma 2.4(1) applies and assumption 2.7 holds with $\delta^*_{m,n}=\max\{\sqrt{\frac{J_n}{n}},J_n^{-r_m}\}$. For this example, $m(X,\alpha)=E[F_{Y_3|Y_1,Y_2,X}(\theta Y_1+h(Y_2))|X]-\gamma$, thus assumption 2.8(i) is implied by condition 5.3(i). Assumption 2.8(iii) is implied condition 5.4(i). Since

$$\frac{dm(X,\alpha_0)}{d\alpha}[\alpha-\alpha_0]=E\left\{f_{Y_3|Y_1,Y_2,X}(\theta_0Y_1+h_0(Y_2))[Y_1(\theta-\theta_0)+h(Y_2)-h_0(Y_2)]|X\right\},$$

$$||\alpha-\alpha_0||^2=\frac{E\left[\left(E\left\{f_{Y_3|Y_1,Y_2,X}(\theta_0Y_1+h_0(Y_2))[Y_1(\theta-\theta_0)+h(Y_2)-h_0(Y_2)]|X\right\}\right)^2\right]}{\gamma(1-\gamma)}.$$

Notice that

$$\begin{aligned}
& E\left[\left(m(X,\alpha)-m(X,\alpha_0)\right)^2\right]\\
= \;& E\left[\left(E[F_{Y_3|Y_1,Y_2,X}(\theta Y_1+h(Y_2))|X]-E[F_{Y_3|Y_1,Y_2,X}(\theta_0Y_1+h_0(Y_2))|X]\right)^2\right]\\
= \;& E\left[\left(E\left\{f_{Y_3|Y_1,Y_2,X}(\overline{\theta}Y_1+\overline{h}(Y_2))[Y_1(\theta-\theta_0)+h(Y_2)-h_0(Y_2)]|X\right\}\right)^2\right]
\end{aligned}$$

where $\overline{\theta}Y_1 + \overline{h}(Y_2)$ is in between $\theta Y_1 + h(Y_2)$ and $\theta_0 Y_1 + h_0(Y_2)$. By condition 5.3(i). we have

$$E\left[(m(X,\alpha) - m(X,\alpha_0))^2\right] \asymp ||\alpha - \alpha_0||^2.$$

In addition, conditions 5.1(ii) and 5.3(i)(iii) imply $||\alpha - \alpha_0|| \leq const.||\alpha - \alpha_0||_s$. Thus assumption 2.8(ii) holds. Since

$$D_w(X) = \frac{dm(X,\alpha_0)}{d\theta} - \frac{dm(X,\alpha_0)}{dh}[w] = E\left\{f_{Y_3|Y_1,Y_2,X}(\theta_0 Y_1 + h_0(Y_2))[Y_1 - w(Y_2)]|X\right\},$$

assumption 2.9(i) is implied by condition 5.6(i), while assumption 2.9(ii) is directly assumed.

Under condition 5.3(i)(iii) and using the same argument as in Chen et al. (2003), we obtain that assumption 3.7(i) holds with $\kappa = 1/2$, while assumption 3.7(ii) trivially holds as $\rho(Z,\alpha) \in [0,1]$.

Next conditions 5.1 and 5.2 implies that $||\Pi_n h_0 - h_0||_s = O(k(n)^{-r_2})$. Condition 5.5(i) implies assumption 2.11. Thus Lemma 2.4(1)(2.i) is applicable and we obtain: $\delta^*_{s,n} = O\left(n^{-\frac{\gamma_2}{2(\gamma_2+a)+d_2}}\right)$ and $\delta^*_n \asymp \delta^*_{m,n} = O\left(n^{-\frac{\gamma_2+a}{2(\gamma_2+a)+d_2}}\right)$ provided that $J_n \asymp k(n) = O\left(n^{\frac{d_2}{2(\gamma_2+a)+d_2}}\right)$.

Assumption 3.2 follows from the fact that $\widehat{\Sigma} = \Sigma = \gamma(1-\gamma)$.

Regarding assumption 3.3, by condition 5.3(i), we have:

$$\begin{aligned} ||v^*_n - v^*||^2 &= ||w^*_n - w^*||^2 = \frac{E\left[\left(E\left\{f_{Y_3|Y_1,Y_2,X}(\theta_0 Y_1 + h_0(Y_2))[w^*_n(Y_2) - w^*(Y_2)]|X\right\}\right)^2\right]}{\gamma(1-\gamma)} \\ &\leq const \times E\left[\left(E\left\{[w^*_n(Y_2) - w^*(Y_2)]|X\right\}\right)^2\right] \end{aligned}$$

Thus assumption 3.3 follows from condition 5.6(iii).

Assumption 3.4(i) is directly implied by condition 5.3(i). Assumptions 3.4(iii)(iv) are implied by conditions 5.3(ii)(iii). For assumption 3.4(ii), we note that under condition 5.3(i),

$$\begin{aligned} & \frac{dm(X,\alpha)}{d\alpha}[v^*_n] - \frac{dm(X,\alpha_0)}{d\alpha}[v^*_n] \\ = \; & E\left\{\left(f_{Y_3|Y_1,Y_2,X}(\theta Y_1 + h(Y_2)) - f_{Y_3|Y_1,Y_2,X}(\theta_0 Y_1 + h_0(Y_2))\right)[Y_1 - w^*_n(Y_2)]|X\right\} v^*_\theta \\ = \; & E\left\{\frac{df_{Y_3|Y_1,Y_2,X}(\overline{\theta}Y_1 + \overline{h}(Y_2))}{dy_3}[(\theta - \theta_0)Y_1 + h(Y_2) - h_0(Y_2)][Y_1 - w^*_n(Y_2)]|X\right\} v^*_\theta \end{aligned}$$

where $\overline{\theta}Y_1 + \overline{h}(Y_2)$ is in between $\theta Y_1 + h(Y_2)$ and $\theta_0 Y_1 + h_0(Y_2)$. Thus

$$E\left[\left\|\frac{dm(X,\alpha)}{d\alpha}[v^*_n] - \frac{dm(X,\alpha_0)}{d\alpha}[v^*_n]\right\|^2_E\right] \leq const. \times ||\alpha - \alpha_0||^2_s = O\left(n^{-\frac{2\gamma_2}{2(\gamma_2+a)+d_2}}\right),$$

and assumption 3.4(ii) is satisfied under condition 5.5(ii). Similarly, assumption 3.5 follows from conditions 5.3(i) and 5.5(ii). Assumption 3.6 holds by condition 5.4(ii). Thus all the assumptions of Proposition 3.1 holds, and the conclusion follows. *Q.E.D*

References

Ai, C., Chen, X., 2003. Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. Econometrica, 71, 1795-1843

Ai, C., Chen, X., 2007. Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables. Journal of Econometrics, 141, 5-43.

Andrews, D. W. K., 1995. Nonparametric Kernel Estimation for Semiparametric Econometric Models. Econometric Theory, 11, 560-596.

Blundell, R., Chen X., Kristensen, D., 2007. Semi-Nonparametric IV Estimation of Shape Invariant Engel Curves.Econometrica, 75, 1613-1669.

Brown, L., Levine, M., 2007. Variance estimation in nonparametric regression via the difference sequence method. The Annals of Statistics, 35, 2219-2232.

Chen, X., 2007. Large Sample Sieve Estimation of Semi-Nonparametric Models, Chp. 76 in Handbook of Econometrics, Vol. 6B, eds. J.J. Heckman and E.E. Leamer.

Chen, X., Pouzo, D., 2007. Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Moments. Working Paper.

Chen, X., Shen, X., 1998. Sieve Extremum Estimates for Weakly Dependent Data. Econometrica, 66, 289 - 314.

Chen, X., Linton, O., van Keilegom, I., 2003. Estimation of Semiparametric Models with the criterion functions is not smooth. Econometrica, 71, 1591-1608.

Chernozhukov, V., Imbens, G., Newey, W., 2007. Instrumental variable estimation of nonseparable models. Journal of Econometrics, 139, 4-14.

Florens, J. P., Johannes, J., Van Bellegem, S., 2007. Instrumental regression in partially linear models. Working paper.

Hall, P., Marron J. S., 1990. On variance Estimation in nonparametric regression. Biometrika, 77, 415-419.

Horowitz, J., Lee, S., 2007. Nonparametric Instrumental Variables Estimation of Quantile Regression Model. Econometrica, 75, 1191-1209.

Huang, J., 1998. Projection Estimation in Multiple Regression with Application to Functional Anova Models. The Annals of Statistics, 27, 242-272.

Koenker, R., 2005. Quantile Regression. Econometric Society Monograph Series, Cambridge University Press.

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

Linton, O., 1995. Second order approximation in a partially linear regression model. Econometrica (1995) 63, 1079-1113.

Ma, S., Kosorok, M., 2005. Robust semiparametric M-estimation and the weighted bootstrap. Journal of Multivariate Analysis, 96, 190-217.

Murphy S., Van der Vaart, A., 2000. On profile Likelihood. Journal of American Statistical Association, 95, 449-465.

Newey, W., 1990. Semiparametric Efficiency Bounds. Journal of Applied Econometrics, 5, 99-135.

Newey, W., 1997. Convergence rates and asymptotic normality for series estimators. Journal of Econometrics, 79, 147-168.

Newey, W., Powell, J., 2003. Instrumental Variable Estimation of Nonparametric Models. Econometrica, 71, 1565 - 1578.

Newey, W., Hsieh, F., Robins, J., 2004. Twicing Kernels and a Small Bias Property of Semiparametric Estimators. Econometrica, 72, 947-962.

Nishiyama, Y., Robinson, P.M., 2000. Edgeworth Expansions for Semiparametric Averaged Derivatives. Econometrica 68, 931-980.

Nishiyama, Y., Robinson, P.M., 2001. Studentization in Edgeworth Expansions for Estimates of Semiparametric Index Models. Nonlinear Statistical Modeling (Festschrift for Takeshi Amemiya) (C. Hsiao, K. Morimune and J. Powell, eds.), Cambridge University Press, 197-240.

Nishiyama, Y., Robinson, P.M., 2005. The Bootstrap and the Edgeworth Correction for Semiparametric Averaged Derivatives. Econometrica 73, 903-980.

Robinson, P.M., 1988. Root-N-Consistent Semiparametric Regression. Econometrica, 56, 931-954.

Robinson, P.M., 1995. The Normal Approximation for Semiparametric Averaged Derivatives. Econometrica, 63, 667-680.

Robinson, P.M., 1995b. Nearest Neighbour Estimation of Semiparametric Regression Models. Journal of Nonparametric Statistics, 5, 33-41.

Shen, X., Shi, J., 2005. Sieve Likelihood ratio inference on general parameter space. Science in China, 48, 67-78.

Van der Vaart, A., 1991. On Differentiable Functionals. Annals of Statistics, 19, 178-204.

B Figures and Tables

γ	0.125	0.250	0.500	0.750	0.875
$E_{MC}\left[\hat{\theta}_n\right]$	1.0009	0.9981	1.0009	1.0008	0.9991
$Var_{MC}\left[\hat{\theta}_n\right]$	0.0023	0.0018	0.0011	0.0017	0.0028
$BIAS^2_{MC}\left[\hat{\theta}_n\right] \times 10^4$	0.0083	0.0347	0.0084	0.0067	0.0078
$(\theta_{2.5}, \theta_{97.5})_{MC}$	$(0.90, 1.10)$	$(0.91, 1.07)$	$(0.93, 1.07)$	$(0.91, 1.08)$	$(0.89, 1.09)$
$(\theta_{2.5}, \theta_{97.5})_{\chi^2}$	$(0.89, 1.09)$	$(0.91, 1.06)$	$(0.93, 1.05)$	$(0.91, 1.07)$	$(0.88, 1.08)$
$I - BIAS^2_{MC}\left[\widehat{h}_n\right]$	0.0022	0.0015	0.0030	0.0030	0.0044
$I - Var_{MC}\left[\widehat{h}_n\right]$	0.0221	0.0287	0.0056	0.0147	0.0173
$I - MSE^2_{MC}\left[\widehat{h}_n\right]$	0.0244	0.0302	0.0087	0.0177	0.0217

Table 1: Results for G-DEN case of the Monte Carlo experiment.

n	125	250	500	1000
$E_{MC}\left[\hat{\theta}_n\right]$	1.0364	0.9926	1.0028	1.0008
$Var_{MC}\left[\hat{\theta}_n\right]$	0.0278	0.0099	0.0039	0.0017
$BIAS^2_{MC}\left[\hat{\theta}_n\right] \times 10^3$	0.1740	0.0800	0.0810	0.0006

Table 2: Results for G-DEN case of the Monte Carlo experiment for $n = \{125, 250, 500, 1000\}$ and $\gamma = 0.750$.

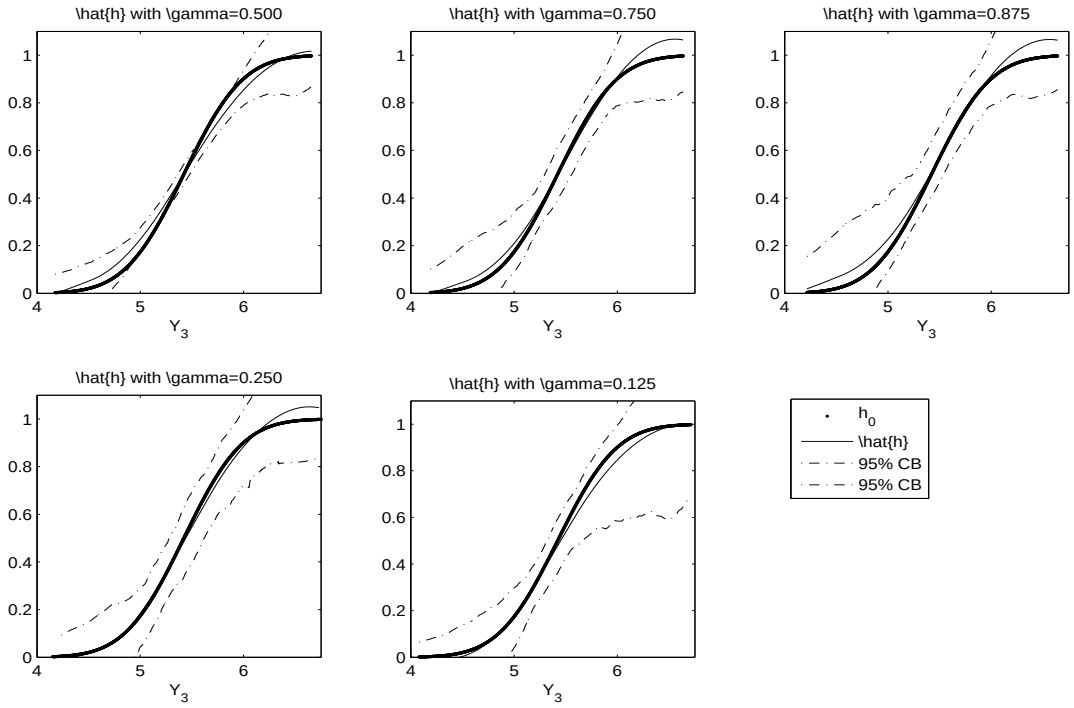

Figure 1: Estimation of h for the G-DEN case of the Monte Carlo experiment.

$\hat{P}_n(h)$	$\|\|\nabla^2 h\|\|^2_{L^2(d\hat{\mu})}$	$\|\|\nabla^2 h\|\|_{L^1(d\hat{\mu})}$	$\|\|\nabla h\|\|^2_{L^2(d\hat{\mu})}$	$\|\|\nabla^2 h\|\|^2_{L^2(d\hat{\mu})}$	$\|\|\nabla h\|\|^2_{L^2(d\hat{\mu})}$
λ_n	0.001	0.001	0.001	0.0003	0.0003
$\hat{\theta}_1$	0.4133	0.3895	0.5479	0.43136	0.36348 (0.3698)
food-in	0.0200	0.0267	-0.0056	0.00989	0.01949 (0.0213)
food-out	0.0010	0.0006	0.0019	0.00033	0.00055 (0.0006)
alcohol	-0.0195	-0.0123	-0.0171	-0.02002	-0.01241 (-0.0216)
fares	0.0106	-0.0031	-0.0001	-0.00009	-0.00173 (-0.0023)
fuel	-0.0027	0.0027	0.0004	-0.00198	-0.00370 (-0.0035)
leisure	0.0208	0.0214	0.0380	0.02582	0.01897 (0.0388)
travel	-0.0207	-0.0218	-0.0084	-0.00622	-0.01536 (-0.0384)

Table 3: θ_l for $l = 1, \ldots, 7$ for the different penalization and $\gamma = 0.50$. In the last column in parenthesis we have the values obtained in Blundell et al. (2007).

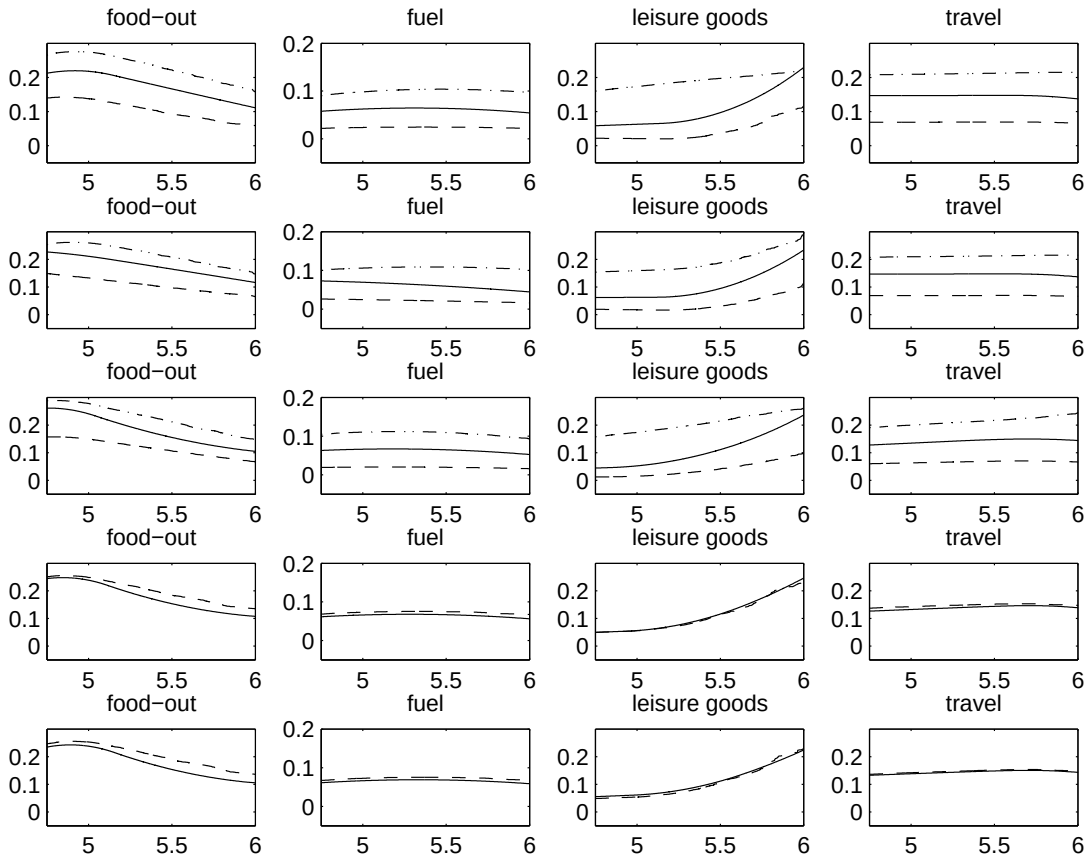

Figure 2: Engel curves for quantiles $\gamma = 0.25$ (dash), 0.50 (solid), 0.75 (dot-dash). $\hat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\hat{\mu})}$ with $\lambda_n = 0.001$ (1st row); $\hat{P}_n(h) = ||\nabla^2 h||_{L^1(d\hat{\mu})}$ with $\lambda_n = 0.001$ (2nd row); $\hat{P}_n(h) = ||\nabla h||^2_{L^2(d\hat{\mu})}$ with $\lambda_n = 0.003$ (3rd row); $\hat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\hat{\mu})}$ with $\lambda_n = 0.0003$ and $\gamma = 0.5$ (solid) and BCK (dash) (4th row); $\hat{P}_n(h) = ||\nabla h||^2_{L^2(d\hat{\mu})}$ with $\lambda_n = 0.0003$ and $\gamma = 0.5$ (solid) and BCK (dash) (5th row).