Chen, Xiaohong; Pouzo, Demian

**Working Paper**

# Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals

cemmap working paper, No. CWP20/09

**Provided in Cooperation with:**
The Institute for Fiscal Studies (IFS), London

*Suggested Citation:* Chen, Xiaohong; Pouzo, Demian (2009) : Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals, cemmap working paper, No. CWP20/09, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2009.2009

This Version is available at:
https://hdl.handle.net/10419/64653

cemmap
centre for microdata methods and practice

# Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals

**Xiaohong Chen**
**Demian Pouzo**

The Institute for Fiscal Studies
Department of Economics, UCL

**cemmap** working paper CWP20/09

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

# Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals[1]

Xiaohong Chen[a,2] , Demian Pouzo[b,3]

[a] *Cowles Foundation for Research in Economics, Yale University, Box 208281, New Haven, CT 06520, USA*

[b] *Department of Economics, New York University, 19 West 4th Street, New York, NY 10012, USA*

First version: September 2005, Final version: March 2009

Forthcoming in *Journal of Econometrics*

---

**Abstract**

This paper considers semiparametric efficient estimation of conditional moment models with possibly nonsmooth residuals in unknown parametric components ($\theta$) and unknown functions ($h$) of endogenous variables. We show that: (1) the penalized sieve minimum distance (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 noncompact function parameter spaces; (2) a simple weighted bootstrap procedure consistently estimates the limiting distribution of the PSMD $\hat{\theta}$; (3) the semiparametric efficiency bound formula of Ai and Chen (2003) remains valid for conditional models with nonsmooth residuals, and the optimally weighted PSMD estimator achieves the bound; (4) the centered, profiled optimally weighted PSMD criterion is asymptotically chi-square distributed. We illustrate our theories using a partially linear quantile instrumental variables (IV) regression, a Monte Carlo study, and an empirical estimation of the shape-invariant quantile IV Engel curves.

*JEL Classification:* C14; C22

*Keywords:* Penalized sieve minimum distance; Nonsmooth generalized residuals; Nonlinear nonparametric endogeneity; Weighted bootstrap; Semiparametric efficiency; Confidence region; Partially linear quantile IV regression; Shape-invariant quantile IV Engel curves

---

[1] This is a much shortened and revised version of Cowles Foundation Discussion Paper 1640.

[2] Corresponding author. Tel.: +1 203 432 5852; Fax: +1 203 432 6167. *E-mail address*: xiaohong.chen@yale.edu.

[3] *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 $X$, and the conditional distribution, $F_{Y|X}$, of $Y$ given $X$ is not specified beyond that $F_{Y|X}$ satisfies (1.1). $\rho()$ is a vector of generalized residual functions whose functional forms are known up to a vector of unknown finite dimensional parameters ($\theta_0$) and a vector of unknown real-valued functions ($h_0 \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))$), where the arguments of each function $h_{0\ell}(\cdot)$ may differ across $\ell = 1, ..., q$, and, in particular, may depend on $Y$. For example, a partially linear quantile Instrumental Variables (IV) regression ($E[1\{Y_3 \leq Y_1'\theta_0 + h_0(Y_2)\}|X] = \gamma$), a single index IV model ($E[Y_1 - h_0(Y_2'\theta_0)|X] = 0$), and a partially additive IV regression with a known link ($g$) model ($E[Y_3 - g(Y_1'\theta_0 + h_{01}(X_1) + h_{02}(Y_2))|X] = 0$) belong to the general framework (1.1).

Newey and Powell (2003) and Ai and Chen (2003) propose Sieve Minimum Distance (SMD) estimation of $\alpha_0 \equiv (\theta_0, h_0)$ for the models (1.1). 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 in a Banach space norm $||\cdot||_s$, Newey and Powell (2003) establish the consistency (with no rate) of the SMD estimator of $(\theta_0, h_0)$ in $||\cdot||_s$. Under the same set of assumptions, Ai and Chen (2003) first derive a faster than $n^{-1/4}$ convergence rate of their SMD estimator $\widehat{h}$ to $h_0()$ in a pseudo metric $||\cdot||$, which is *weaker* than the consistency norm $||\cdot||_s$ when $h()$ depends on $Y$. They then establish root-$n$ asymptotic normality and semiparametric efficiency of the SMD estimator of $\theta_0$. As an illustration, Ai and Chen (2003) present the root-$n$ normality and efficiency of their SMD estimator $\widehat{\theta}$ for a partially linear mean IV regression example $E[Y_1 - X_1'\theta_0 - h_0(Y_2)|X] = 0$, after showing that their SMD estimator $\widehat{h}$ is consistent for $h_0$ in a strong norm $||h||_s = \sqrt{E([h(Y_2)]^2)}$, with a rate faster than $n^{-1/4}$ in a weaker pseudo metric $||h|| = \sqrt{E([E\{h(Y_2)|X\}]^2)}$. *Unfortunately*, when $h_0()$ depends on $Y$ and enters the semiparametric model (1.1) *nonlinearly*, in order to estimate $\theta_0$ at a root-$n$ rate, one also needs some convergence rate of $\widehat{h}$ to $h_0$ in a strong norm $||\cdot||_s$.

For the purely nonparametric conditional moment models $E[\rho(Y, X_z; h_0(\cdot))|X] = 0$ in which $h_0(\cdot)$ may depend on the endogenous variables $Y$, Chen and Pouzo (2008a) propose a Penalized SMD (PSMD) estimator. They establish the consistency and the convergence rates of the PSMD estimator $\widehat{h}$ in a strong metric $||\cdot||_s$ without assuming compactness of $\mathcal{H}$, allowing for nonsmooth residual function $\rho(Z; h(\cdot))$ in $h$. They do not, however, consider the root-$n$ efficient estimation of $\theta_0$ for the more general semiparametric models (1.1), nor any methods of computing tests and

confidence intervals. Finally, none of the existing work investigates whether one can simultaneously estimate $\theta_0$ and $h_0$ for the general semiparametric models (1.1) at their respectively optimal rates.

In this paper, we contribute in several major ways to the existing semiparametric literature allowing for nonparametric endogeneity. **First**, we show that, for the general semiparametric models (1.1), the PSMD estimator $\widehat{\alpha} \equiv (\widehat{\theta}, \widehat{h})$ can simultaneously achieve root-$n$ asymptotic normality of $\widehat{\theta}$ and the optimal convergence rate of $\widehat{h}$ (in the metric $||\cdot||_s$), allowing for possibly nonsmooth residuals, and possibly noncompact (in $||\cdot||_s$) function space ($\mathcal{H}$) and the sieve spaces ($\mathcal{H}_n$). It is previously known that for semiparametric models without nonparametric endogeneity (i.e., the unknown $h()$ does not depend on $Y$), the sieve estimators of $(\theta_0, h_0)$ can simultaneously achieve root-$n$ normality of $\widehat{\theta}$ and the optimal convergence rate of $\widehat{h}$ (in $||\cdot||_s$).[4] From the point of view of empirical estimation of models (1.1) with nonparametric endogeneity (see, e.g., estimation of a system of shape-invariant Engel curves with endogenous total expenditure in Blundell et al. (2007)), it is nice to know that the PSMD estimators still possess such an attractive property. **Second**, under the same sets of sufficient conditions for the root-$n$ normality of the PSMD $\widehat{\theta}$, we show that a simple weighted bootstrap procedure consistently estimates the limiting distribution of the PSMD $\widehat{\theta}$. Previously, Ai and Chen (2003) propose a consistent sieve estimator of the asymptotic variance of $\widehat{\theta}$, which hinges on the pointwise differentiability of the residual functions $\rho(Z; \theta, h(\cdot))$ in $\alpha = (\theta, h)$. In our paper $\rho(Z; \theta, h(\cdot))$ could be pointwise non-smooth with respect to $\alpha = (\theta, h)$, such as in the partially linear quantile IV regression example; therefore we provide a justification of using a weighted bootstrap to construct a confidence region. **Third**, we show that the semiparametric efficiency bound formula of Ai and Chen (2003) remains valid for the conditional models (1.1) with nonsmooth residuals.[5] When the model (1.1) contains several unknown functions $h_0 \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))$ and when some of the $h_{0j}$ depend on $Y$, although the efficiency bound is well-defined and unique, it may not have a closed-form expression and its "least favorable curve" solutions may not be unique. Nevertheless, our optimally weighted PSMD estimator always achieves the efficiency bound for $\theta_0$. **Fourth**, we show that the centered, profiled optimally weighted PSMD criterion is asymptotically chi-square distributed. This leads to an alternative confidence region construction by inverting the centered, profiled optimally weighted criterion. It also avoids the nonparametric estimation of the asymptomatic variance of $\widehat{\theta}$, and is computationally less time-consuming than the weighted bootstrap.

Technically, we are able to achieve the above listed results by first showing that our computable PSMD criterion function with nonsmooth residuals can be approximated well by an infeasible SMD criterion with smooth moments in a shrinking neighborhood of $\alpha_0 = (\theta_0, h_0)$, where the

[4] It is known that the original kernel estimators can not; see e.g., Robinson (1988) and Newey et al. (2004).

[5] We note that the semiparametric efficiency bound theorem of Ai and Chen (2003) (theorem 6.1) and its proof do not rely on the $||\cdot||_s-$compactness of the space $\mathcal{H}$. In fact, the working paper version of Ai and Chen (2003) presents such results without assuming compact parameter space.

neighborhood is defined using the optimal convergence rates in both the strong norm $||\cdot||_s$ and the weaker pseudo metric $||\cdot||$. We then slightly modify the proof strategy in Ai and Chen (2003) (and the references therein) by performing a second order Taylor expansion to the difference of the smoothed infeasible SMD criterion evaluated at two points: the PSMD estimator $\widehat{\alpha} = (\widehat{\theta}, \widehat{h})$ and a deviation from $\widehat{\alpha}$ along an approximately least favorable direction.

In section 2, we present the PSMD estimator $\widehat{\alpha} = (\widehat{\theta}, \widehat{h})$ and its convergence rates in both the strong norm $||\cdot||_s$ and the weaker pseudo metric $||\cdot||$ under the same set of smoothing parameters. In section 3, we establish the root-$n$ asymptotic normality of $\widehat{\theta}$ and the validity of a weighted bootstrap procedure. In section 4, we derive the semiparametric efficiency bound for $\theta$, and show the efficiency of the optimally weighted PSMD of $\theta$. In addition, we show that the profile optimally weighted PSMD criterion is asymptotically chi-squared distributed. Our PSMD estimator and its large sample properties are applicable to all specific models that satisfy the semiparametric conditional models (1.1). Due to the lack of space, we only discuss a partially linear mean IV regression example in section 4, and a partially linear quantile IV regression example in section 5, where the latter example is used to highlight the technical difficulty of estimating $\theta_0$ semiparametrically efficiently when the unknown $h_0(Y)$ enters the generalized residual function $\rho(Z; \theta, h(\cdot))$ nonsmoothly. Although the asymptotic properties of the PSMD estimator are difficult to derive, the estimator is easy to compute and performs well in finite samples. See section 6 for a Monte Carlo study of a partially linear quantile IV example, and a real data study of a system of shape-invariant quantile IV Engel curves. All the proofs are gathered in the appendix.

**Notation**. We assume that all random variables $(Y', X', W)$ are defined on a complete probability space, and for simplicity that $Y, X$ are continuous random variables. Let $f_X$ ($F_X$) be the marginal density (cdf) of $X$, and $f_{Y|X}$ ($F_{Y|X}$) be the conditional density (cdf) of $Y$ given $X$. We often implicitly define a term (such as a notation or an order of convergence rate) using "$\equiv$". For any vector-valued $x$, we denote $||x||_E$ as its Euclidean norm (i.e., $||x||_E \equiv \sqrt{x'x}$, although sometimes we use $|x| = ||x||_E$ for simplicity). Denote $L^p(\Omega, d\mu)$, $1 \leq p < \infty$, as a space of measurable functions with $||g||_{L^p(\Omega,d\mu)} \equiv \{\int_\Omega |g(t)|^p d\mu(t)\}^{1/p} < \infty$, where $\Omega$ is the support of the sigma-finite positive measure $d\mu$ (sometimes $L^p(d\mu)$ and $||g||_{L^p(d\mu)}$ are used for simplicity). For any positive possibly random sequences $\{a_n\}$ and $\{b_n\}$, $a_n = O_P(b_n)$ means that $\Pr(a_n/b_n \geq M) \to 0$ as $n$ and $M$ go to infinity; and $a_n = o_P(b_n)$ means that for all $\varepsilon > 0$, $\Pr(a_n/b_n \geq \varepsilon) \to 0$ as $n$ goes to infinity.

## 2 The Penalized SMD estimator

The semiparametric conditional moment models (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 (2008a), we define the *Penalized* SMD (PSMD) estimator as

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

where $\mathcal{A}_{k(n)} \equiv \Theta \times \mathcal{H}_{k(n)}$ is a sieve for $\mathcal{A} \equiv \Theta \times \mathcal{H}$, $\widehat{m}(X, \alpha)$ and $\widehat{\Sigma}(X)$ are nonparametric estimators of $m(X, \alpha)$ and $\Sigma(X)$ (a positive definite weighting matrix) respectively, $\lambda_n \geq 0$ is a penalization tuning parameter such that $\lambda_n = o(1)$, and $\widehat{P}_n(h) \geq 0$ is a possibly random penalty function. Let $k(n)$ denote the dimension of the sieve $\mathcal{H}_{k(n)}$ for the function space $\mathcal{H}$. In this paper we focus on the PSMD procedure using a *finite* dimensional sieve (i.e., $k(n) < \infty$). See Chen and Pouzo (2008a) for a more detailed presentation of the PSMD procedures with possibly infinite dimensional sieves.

In the working paper version (Chen and Pouzo (2008b)), we establish the asymptotic normality, weighted bootstrap, semiparametric efficiency and chi-square approximation of the PSMD estimator using any nonparametric estimators $\widehat{m}(X, \alpha)$ of $m(X, \alpha)$. In this published version, due to the lack of space, we only present the large sample properties of the PSMD estimator 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 $J_n$ be the dimension of $p^{J_n}(X)$, and $p^{J_n}(X)$ be a tensor-product linear sieve basis, which is the product of univariate linear sieves such as B-splines, polynomial splines (P-splines), wavelets and Fourier series. See Newey (1997), Huang (1998) and Chen (2007) for more details about tensor-product linear sieves.

## 2.1 Review of consistency without compactness

For the purely nonparametric conditional moment models $E[\rho(Y, X_z; h_0(\cdot))|X] = 0$, Chen and Pouzo (2008a) present several consistency results $||\widehat{h}_n - h_0||_s = o_P(1)$ for their PSMD estimator $\widehat{h} = \arg\inf_{h \in \mathcal{H}_{k(n)}} \{\frac{1}{n} \sum_{i=1}^{n} \widehat{m}(X_i, h)' \widehat{m}(X_i, h) + \lambda_n \widehat{P}_n(h)\}$, depending on whether or not the penalty function $\widehat{P}_n(h)$ is lower semicompact (under the metric $\|\cdot\|_s$). All of their consistency theorems can be trivially adapted to establish consistency of our PSMD estimator $\widehat{\alpha}_n \equiv (\widehat{\theta}_n, \widehat{h}_n)$ defined in (2.1). For the sake of easy reference, here we provide one consistency result with lower semicompact penalty. In the following we denote $\|\alpha\|_s \equiv \|\theta\|_E + \|h\|_s$ on $\mathcal{A} \equiv \Theta \times \mathcal{H}$.

**Assumption 2.1.** *(i)* $\{(Y_i', X_i')\}_{i=1}^{n}$ *is an i.i.d. sample; (ii)* $\mathcal{A} \equiv \Theta \times \mathcal{H}$, $\Theta$ *is a compact subset of* $\mathcal{R}^{d_\theta}$, *and* $\mathcal{H} \subseteq \mathbf{H}$, $\mathbf{H}$ *is a separable Banach space under a metric* $\|\cdot\|_s$*; (iii)* $E[\rho(Z, \alpha_0)|X] = 0$, *and* $\|\theta_0 - \theta\|_E + \|h_0 - h\|_s = 0$ *for any* $\alpha = (\theta, h) \in \mathcal{A}$ *with* $E[\rho(Z, \alpha)|X] = 0$.

**Assumption 2.2.** *(i) $\mathcal{A}_k \equiv \Theta \times \mathcal{H}_k$, $k \geq 1$, are the sieve spaces satisfying $\mathcal{H}_k \subseteq \mathcal{H}_{k+1} \subseteq \mathcal{H}$, and there exists a function $\Pi_{k(n)} h_0 \in \mathcal{H}_{k(n)}$ such that $||\Pi_{k(n)} h_0 - h_0||_s = o(1)$; (ii) $E[m(X,\alpha)'\Sigma(X)^{-1}m(X,\alpha)]$ is continuous at $\alpha_0$ under $\|\cdot\|_s$.*

**Assumption 2.3.** *(i) $E[m(X,\alpha)'\Sigma(X)^{-1}m(X,\alpha)]$ is lower semicontinuous (in $||\cdot||_s$) on $\mathcal{A}$; (ii) for each $k \geq 1$, $\mathcal{A}_k$ is closed subspace of $(\mathcal{A}, ||\cdot||_s)$;*

**Assumption 2.4.** *(i) $\lambda_n \sup_{h \in \mathcal{H}_{k(n)}} |\widehat{P}_n(h) - P(h)| = O_P(\lambda_n)$, with $P(\cdot)$ a non-negative real-valued measurable function of $h \in \mathcal{H}$, $P(h_0) < \infty$ and $\lambda_n |P(\Pi_n h_0) - P(h_0)| = O(\lambda_n)$; (ii) the set $\{h \in \mathcal{H} : P(h) \leq M\}$ is compact under $||\cdot||_s$ for all $M \in [0, \infty)$.*

Let $\{\delta_{\Sigma,n}\}_n$ and $\{\delta_{m,n}\}_n$ be real-valued positive sequences decreasing to zero (as $n \to \infty$), denoting the convergence rates of $\widehat{\Sigma} - \Sigma$ and $\widehat{m} - m$ respectively.

**Assumption 2.5.** *(i) $\sup_{x \in \mathcal{X}} |\widehat{\Sigma}(x) - \Sigma(x)| \equiv O_P(\delta_{\Sigma,n})$; (ii) with probability approaching one, $\widehat{\Sigma}(x)$ is positive definite, and its smallest and largest eigenvalues are finite positive uniformly in $x \in \mathcal{X}$; (iii) $\Sigma(x)$ is positive definite, and its smallest and largest eigenvalues are finite positive uniformly in $x \in \mathcal{X}$.*

**Assumption 2.6.** *(i) $\sup_{\alpha \in \mathcal{A}_{k(n)}} E\left[||\widehat{m}(X,\alpha) - m(X,\alpha)||_E^2\right] \equiv O_P(\delta_{m,n}^2)$; (ii) there are finite constants $c, c' > 0$ such that, except on an event whose probability goes to zero as $n \to \infty$, $cE[||\widehat{m}(X,\alpha)||_E^2] \leq n^{-1}\sum_{i=1}^n ||\widehat{m}(X_i,\alpha)||_E^2 \leq c'E[||\widehat{m}(X,\alpha)||_E^2]$ uniformly over $\alpha \in \mathcal{A}_{k(n)}$.*

Assumption 2.6 is a high level condition, and is satisfied when $\widehat{m}(X,\alpha)$ is the series LS estimator (2.2) (see Remark 2.1). Denote $\xi_{0n} \equiv \sup_x ||p^{J_n}(x)||_E$.

**Assumption 2.7.** *(i) $\mathcal{X}$ 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 $\mathcal{X}$; (ii) The smallest and largest eigenvalues of $E[p^{J_n}(X)p^{J_n}(X)']$ are bounded and bounded away from zero for all $J_n$; (iii) either $J_n\xi_{0n}^2 = o(n)$ or $J_n \log(J_n) = o(n)$ for P-spline sieve $p^{J_n}(X)$.*

Let $\{b_{m,J_n}\}_n$ be a real-valued positive sequence decreasing to zero (as $J_n \to \infty$), denoting the bias of approximating $m(\cdot,\alpha)$ by the series basis $p^{J_n}(\cdot)$.

**Assumption 2.8.** *(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) $\sup_x |g(x) - p^{J_n}(x)'\pi| = O(b_{m,J_n})$; (b) $E\{[g(X) - p^{J_n}(X)'\pi]^2\} = O(b_{m,J_n}^2)$ for $p^{J_n}(X)$ sieve with $\xi_{0n} = O(J_n^{1/2})$.*

Assumption 2.8(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 2.8(ii) (a) (or (b)) holds with $b_{m,J_n} = J_n^{-r_m}$ and $r_m \equiv \gamma_m/d_x$.

**Remark 2.1.** *(Lemma B.3 of Chen and Pouzo (2008a)) Let $\widehat{m}$ be the series LS estimator given in (2.2) with B-splines, P-splines, cosine/sine or wavelets as the basis $p^{J_n}(X)$. Suppose that assumptions 2.7 and 2.8 hold. Then: assumption 2.6 is satisfied with $\delta_{m,n} = \max\{\sqrt{\frac{J_n}{n}}, b_{m,J_n}\}$.*

Denote $\Pi_{k(n)}\alpha_0 \equiv (\theta_0, \Pi_{k(n)}h_0) \in \mathcal{A}_{k(n)} \equiv \Theta \times \mathcal{H}_{k(n)}$. The following lemma is a minor modification of Theorem 3.3 of Chen and Pouzo (2008a) hence we omit its proof.

**Lemma 2.1.** *Let $\widehat{\alpha}_n$ be the PSMD estimator (2.1) with $\lambda_n > 0$, $\lambda_n = o(1)$. Let assumptions 2.1 - 2.6 hold. If $\max\{\delta_{m,n}^2, E[m(X, \Pi_{k(n)}\alpha_0)'m(X, \Pi_{k(n)}\alpha_0)]\} = O(\lambda_n)$, then: $||\widehat{\alpha}_n - \alpha_0||_s = o_P(1)$ and $P(\widehat{h}_n) = O_P(1)$.*

## 2.2 Convergence Rates

Denote $\mathcal{A}_{os} \equiv \{\alpha \in \mathcal{A} : ||\alpha - \alpha_0||_s = o(1), P(h) \leq c\}$ and $\mathcal{A}_{osn} \equiv \mathcal{A}_{os} \cap \mathcal{A}_{k(n)}$. For any $\alpha \in \mathcal{A}_{os}$ we define the first pathwise derivative of $m(X, \alpha)$ at the direction $[\alpha - \alpha_0]$ evaluated at $\alpha_0$ 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 $\alpha_1, \alpha_2 \in \mathcal{A}_{os}$ as

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

The next assumption 2.9(i) ensures that the pseudo-metric $||\alpha_1 - \alpha_2||$ is well-defined for any $\alpha_1, \alpha_2 \in \mathcal{A}_{os}$.

**Assumption 2.9.** *(i) $\mathcal{A}_{os}$ and $\mathcal{A}_{osn}$ are convex, $m(X, \alpha)$ is continuously pathwise differentiable with respect to $\alpha \in \mathcal{A}_{os}$; (ii) there are finite constants $c, c' > 0$ such that $c||\alpha - \alpha_0||^2 \leq E\left[||m(X, \alpha)||_E^2\right] \leq c'||\alpha - \alpha_0||^2$ for all $\alpha \in \mathcal{A}_{os}$; (iii) there is a finite constant $K > 0$ such that $K \times ||\alpha - \alpha_0|| \leq ||\alpha - \alpha_0||_s$ for all $\alpha \in \mathcal{A}_{os}$.*

Define $\overline{\mathbf{V}}$ as the closure of the linear span of $\mathcal{A}_{os} - \{\alpha_0\}$ under the metric $||\cdot||$. For any $v_1, v_2 \in \overline{\mathbf{V}}$, we define an inner product corresponding to the metric $||\cdot||$:

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

and for any $v \in \overline{\mathbf{V}}$ we call $v = 0$ if and only if $||v|| = 0$ (i.e., functions in $\overline{\mathbf{V}}$ are defined in an equivalent class sense according to the metric $||\cdot||$). Thus $(\overline{\mathbf{V}}, \langle\cdot\rangle)$ is a Hilbert space. Any $v \equiv (v_\theta, v_h) \in \overline{\mathbf{V}}$ if and only if $v_\theta' E\left[\{\frac{dm(X,\alpha_0)}{d\theta}\}'\Sigma(X)^{-1}\frac{dm(X,\alpha_0)}{d\theta}\right] v_\theta < \infty$ and $E\left[\{\frac{dm(X,\alpha_0)}{dh}[v_h]\}'\Sigma(X)^{-1}\{\frac{dm(X,\alpha_0)}{dh}[v_h]\}\right] < \infty$. We can express $\overline{\mathbf{V}}$ as $\mathcal{R}^{d_\theta} \times \overline{\mathcal{W}}$ with $\overline{\mathcal{W}} \equiv \left\{w : E\left[\left\|\Sigma(X)^{-\frac{1}{2}}\{\frac{dm(X,\alpha_0)}{dh}[w]\}\right\|_E^2\right] < \infty\right\}$. For each

component $\theta_j$ (of $\theta$), $j = 1, ..., d_\theta$, denote $D_{w_j}(X) \equiv \frac{dm(X,\alpha_0)}{d\theta_j} - \frac{dm(X,\alpha_0)}{dh}[w_j]$. Let $w_j^* \in \overline{\mathcal{W}}$ be a solution to

$$\inf_{w_j \in \overline{\mathcal{W}}} E\left[D_{w_j}(X)'\Sigma(X)^{-1}D_{w_j}(X)\right], \tag{2.5}$$

which solves

$$E\left[\left(\frac{dm(X,\alpha_0)}{dh}[w_j]\right)'\Sigma(X)^{-1}\left(\frac{dm(X,\alpha_0)}{d\theta_j} - \frac{dm(X,\alpha_0)}{dh}[w_j^*]\right)\right] = 0 \text{ for all } w_j \in \overline{\mathcal{W}}. \tag{2.6}$$

Denote $\mathbf{w}^* \equiv (w_1^*, ..., w_{d_\theta}^*) \in \overline{\mathcal{W}} \times \cdots \times \overline{\mathcal{W}}$, and

$$D_{\mathbf{w}^*}(X) \equiv \left(D_{w_1^*}(X), ..., D_{w_{d_\theta}^*}(X)\right) = \frac{dm(X,\alpha_0)}{d\theta} - \frac{dm(X,\alpha_0)}{dh}[\mathbf{w}^*].$$

Although the solution $w_j^* \in \overline{\mathcal{W}}$, $j = 1, ..., d_\theta$ to (2.5) (or (2.6)) may not be unique, the minimized value, $E[D_{w_j^*}(X)'\Sigma(X)^{-1}D_{w_j^*}(X)]$, is unique; hence $E[D_{\mathbf{w}^*}(X)'\Sigma(X)^{-1}D_{\mathbf{w}^*}(X)]$ is uniquely defined. If $\Sigma(X) = Var\{\rho(Z,\alpha_0)|X\}$, then $E[D_{\mathbf{w}^*}(X)'\Sigma(X)^{-1}D_{\mathbf{w}^*}(X)]$ becomes the semiparametric efficiency information bound for $\theta_0$. See Section 4 for further details.

If $h_0$ were a parametric function say $h_0(\cdot, \beta_0)$ up to an unknown finite dimensional parameter $\beta_0 \in \mathcal{R}^{d_\beta}$, then $w_j$ becomes a vector in $\mathcal{R}^{d_\beta}$ (instead of a function), and (2.5) (or (2.6)) can be solved in a closed form:

$$E\left(\frac{dm(X,\alpha_0)}{d\beta}'\Sigma(X)^{-1}\frac{dm(X,\alpha_0)}{d\beta}\right) \times w_j^* = E\left(\frac{dm(X,\alpha_0)}{d\beta}'\Sigma(X)^{-1}\frac{dm(X,\alpha_0)}{d\theta_j}\right)$$

for $w_j^* \in \mathcal{R}^{d_\beta}$, $j = 1, ..., d_\theta$; hence $D_{\mathbf{w}^*}(X) = \frac{dm(X,\alpha_0)}{d\theta} - \frac{dm(X,\alpha_0)}{d\beta} \times \mathbf{w}^*$ is simply the weighted least squares regression residual of regressing $\frac{dm(X,\alpha_0)}{d\theta}$ on $\frac{dm(X,\alpha_0)}{d\beta}$ using the weight $\Sigma(X)^{-1}$. (Even for the parametric case, (2.5) (or (2.6)) has a unique solution $w_j^*$ if and only if $E\left(\frac{dm(X,\alpha_0)}{d\beta}'\Sigma(X)^{-1}\frac{dm(X,\alpha_0)}{d\beta}\right)$ is invertible.) We impose the following assumption.

**Assumption 2.10.** *(i)* $E\left[\{\frac{dm(X,\alpha_0)}{d\theta}\}'\Sigma(X)^{-1}\{\frac{dm(X,\alpha_0)}{d\theta}\}\right]$ *is finite; (ii)* $E[D_{\mathbf{w}^*}(X)'[\Sigma(X)]^{-1}D_{\mathbf{w}^*}(X)]$ *is finite, positive-definite.*

Let $\mathcal{H}_{os} \equiv \{h \in \mathcal{H} : ||h - h_0||_s = o(1), P(h) \leq c\}$ and $\mathcal{H}_{osn} \equiv \mathcal{H}_{os} \cap \mathcal{H}_{k(n)}$. For any $h_1, h_2 \in \mathcal{H}_{os}$ we define:

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

**Lemma 2.2.** *Let assumptions 2.9 and 2.10 hold. Then: there are finite positive constants* $c, c'$ *such that for all* $\alpha \in \mathcal{A}_{os}$*, we have:* $c||\theta - \theta_0||_E \leq ||\alpha - \alpha_0||$ *and* $c'||h - h_0|| \leq ||\alpha - \alpha_0||$.

Let $\{\delta_n\}_n$ and $\{\delta_{s,n}\}_n$ be real-valued positive sequences decreasing to zero (as $n \to \infty$), denoting the convergence rates of $||\widehat{\alpha}_n - \alpha_0||$ and $||\widehat{\alpha}_n - \alpha_0||_s$ respectively, i.e., $||\widehat{\alpha}_n - \alpha_0|| \equiv O_P(\delta_n)$ and $||\widehat{\alpha}_n - \alpha_0||_s \equiv O_P(\delta_{s,n})$. Then Lemma 2.2 implies that $||\widehat{\theta}_n - \theta_0||_E = O_P(\delta_n)$ and $||\widehat{h}_n - h_0|| = O_P(\delta_n)$. By definition of the norm $||\cdot||_s$ we also have $||\widehat{h}_n - h_0||_s = O_P(\delta_{s,n})$.

**Assumption 2.11.** *(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 2.11(i) suggests that $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ is a natural sieve for $\mathcal{H}$. For example, if $\mathcal{H} \subseteq W_2^{\varsigma}([0,1]^d, leb)$ (a Sobolev space), then assumption 2.11 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 2.12.** *There are finite constants* $c$, $C > 0$ *and a non-increasing positive sequence* $\{b_j\}_{j=1}^{\infty}$ *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)* $C\sum_j b_j |\langle h_0 - \Pi_{k(n)}h_0, q_j\rangle_s|^2 \geq ||h_0 - \Pi_{k(n)}h_0||^2$.

See Chen and Pouzo (2008a) for interpretation and sufficient conditions for assumption 2.12.

**Lemma 2.3.** *Let* $\widehat{\alpha}_n$ *be the PSMD estimator (2.1) with* $\lambda_n > 0$, $\lambda_n = o(1)$. *Let assumptions of Lemmas 2.1 and 2.2 hold, and* $\sup_{h\in\mathcal{H}_{osn}} |\widehat{P}_n(h) - P(h)| = o_P(1)$. *If* $\max\left\{\delta_{m,n}, \sqrt{\lambda_n}\right\} = \delta_{m,n}$, *then:*

*(1)* $||\widehat{h}_n - h_0|| = O_P(\delta_n) = O_P\left(\max\{\delta_{m,n}, ||h_0 - \Pi_{k(n)}h_0||\}\right)$.

*(2) Further, let assumptions 2.11 and 2.12 hold. Then:*

$$\delta_n = O\left(\delta_{m,n}\right) \quad \text{and} \quad ||\widehat{h}_n - h_0||_s = O_P(\delta_{s,n}) = O_P\left(||h_0 - \Pi_{k(n)}h_0||_s + \frac{\delta_{m,n}}{\sqrt{b_{k(n)}}}\right).$$

**Remark 2.2.** *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 2.3 hold. Let* $h_0 : \mathcal{R}^d \to \mathcal{R}$ *and* $||h_0 - \Pi_{k(n)}h_0||_s = O(\{k(n)\}^{-\varsigma/d})$ *for a finite* $\varsigma > 0$, *and* $\delta_{m,n} = O\left(\sqrt{\frac{k(n)}{n}}\right) = o_P(1)$.

*(i) Mildly ill-posed case: let* $b_k = O(k^{-2a/d})$ *for some* $a \geq 0$. *Then:* $\delta_{s,n} = O\left(n^{-\frac{\varsigma}{2(\varsigma+a)+d}}\right)$ *and* $\delta_n = O\left(n^{-\frac{\varsigma+a}{2(\varsigma+a)+d}}\right)$ *provided* $k(n) = O\left(n^{\frac{d}{2(\varsigma+a)+d}}\right)$*; hence* $\delta_n = o(n^{-1/4})$ *if* $\varsigma + a > d/2$.

*(ii) Severely ill-posed case: let* $b_k = O(\exp\{-k^{a/d}\})$ *for some* $a > 0$. *Then:* $\delta_{s,n} = O\left([\ln(n)]^{-\varsigma/a}\right)$ *and* $\delta_n = O\left(\sqrt{\frac{[\ln(n)]^{d/a}}{n}}\right)$ *provided* $k(n) = O\left([\ln(n)]^{d/a}\right)$*; hence* $\delta_n = o(n^{-1/4})$.

For a nonparametric mean IV regression model $E[Y_1 - h_0(Y_2)|X] = 0$, Chen and Reiss (2007) show that the above convergence rate $||\widehat{h}_n - h_0||_s = O_P(\delta_{s,n})$ in the norm $||h||_s = \sqrt{E([h(Y_2)]^2)}$ achieves the minimax optimal rate. The optimal rate $\delta_{s,n}$ is determined by choosing the smoothing parameters to balance the sieve approximation error rate $O(\{k(n)\}^{-\varsigma/d})$ and the standard derivation part $O\left(\sqrt{\frac{k(n)}{n\times b_{k(n)}}}\right)$, where the term $\{b_{k(n)}\}^{-1/2}$ is called "sieve measure of ill-posedness" (see Blundell et al. (2007) and Chen and Pouzo (2008a)). When $b_k = const$ for all $k$ (or when $a = 0$ in Remark 2.2(i)), the convergence rate $\delta_{s,n}$ becomes the known optimal rate for sieve M-estimation without nonparametric endogeneity; see, e.g., Chen and Shen (1998).

According to Lemma 2.3 and Remark 2.2, the same set of smoothing parameters that achieves the optimal rate for $||\widehat{h}_n - h_0||_s = O_P(\delta_{s,n})$ can also lead to the rate $||\widehat{\alpha}_n - \alpha_0|| = O_P(\delta_n) = O_P(\delta_{m,n}) = o_P(n^{-1/4})$, which is what we need for root-$n$ asymptotic normality of $\widehat{\theta}$; see Theorem 3.1 in Section 3.

# 3 Asymptotic Normality and Weighted Bootstrap

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

In this subsection we establish root-$n$ asymptotic normality of the PSMD estimator $\widehat{\theta}$, which extends the normality result of Ai and Chen (2003) to allow for nonsmooth residuals $\rho(Z;\alpha)$ and any lower semicompact penalty functions. Denote $\mathcal{N}_0 \equiv \{\alpha \in \mathcal{A}_{os} : ||\alpha - \alpha_0|| = O(\delta_n), ||\alpha - \alpha_0||_s = O(\delta_{s,n})\}$ and $\mathcal{N}_{0n} \equiv \mathcal{N}_0 \cap \mathcal{A}_{k(n)}$.

**Assumption 3.1.** *(i) There exist a measurable function $b(X)$ with $E[|b(X)|] < \infty$ and constants $\kappa \in (0,1]$, $r \geq 1$ 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) $\sup_{\alpha \in \mathcal{N}_0} |\rho(Z,\alpha)| \leq C(Z)$ and $E[C(Z)^2|X] \leq const. < \infty$; (iii) $\delta_n^2 \times (\delta_{s,n})^\kappa = o(n^{-1})$.*

By Remark 2.2, for both the "mildly ill-posed" case and the "severely ill-posed" case, assumption 3.1(iii) $\delta_n^2 (\delta_{s,n})^\kappa = o(n^{-1})$ is satisfied if $\varsigma > d/\kappa$.

For any non-zero $\lambda \in \mathcal{R}^{d_\theta}$, assumption 2.10 implies that there is a $v^* \in \overline{\mathbf{V}}$ such that $\lambda'(\widehat{\theta}_n - \theta_0) = \langle v^*, \widehat{\alpha}_n - \alpha_0 \rangle$, that is, $v^* \equiv (v_\theta^*, v_h^*)$ is the Riesz representer of $\lambda'(\widehat{\theta}_n - \theta_0)$, with $v_\theta^* \equiv (E[D_{\mathbf{w}^*}(X)'[\Sigma(X)]^{-1} D_{\mathbf{w}^*}(X)])^{-1}\lambda$ and $v_h^* \equiv -\mathbf{w}^* \times v_\theta^*$.

**Assumption 3.2.** *(i) $\theta_0 \in int(\Theta)$; (ii) $\Sigma_0(X) \equiv Var[\rho(Z,\alpha_0)|X]$ is positive definite for all $X \in \mathcal{X}$; (iii) there is a $v_n^* \equiv (v_\theta^*, -\mathbf{w}_n^* \times v_\theta^*) \in \mathcal{A}_{k(n)} \setminus \{\alpha_0\}$ such that $||v_n^* - v^*|| \times \delta_n = o(n^{-1/2})$.*

**Assumption 3.3.** *(i) $\delta_n = o(n^{-1/4})$; (ii) $\delta_{\Sigma,n} \times \delta_n = o(n^{-1/2})$;*
*(iii) $\lambda_n \sup_{\alpha \in \mathcal{N}_{0n}} |\widehat{P}_n(h \pm \epsilon_n \mathbf{w}_n^* v_\theta^*) - \widehat{P}_n(h)| = o_P(\frac{1}{n})$ with $0 < \epsilon_n = o(n^{-1/2})$.*

Let $\widetilde{m}(X,\alpha) \equiv p^{J_n}(X)'(P'P)^- \sum_{i=1}^n p^{J_n}(X_i) m(X_i,\alpha)$ be the LS projection of $m(X,\alpha)$ onto $p^{J_n}(X)$. Define $g(X,v^*) \equiv \{\frac{dm(X,\alpha_0)}{d\alpha}[v^*]\}'\Sigma(X)^{-1}$ and $\widetilde{g}(X,v^*)$ as its LS projection onto $p^{J_n}(X)$.

**Assumption 3.4.** *(i) $E\left[\left\|\frac{d\widetilde{m}(X,\alpha_0)}{d\alpha}[v^*] - \frac{dm(X,\alpha_0)}{d\alpha}[v^*]\right\|_E^2\right](\delta_n)^2 = o_P(\frac{1}{n})$;*
*(ii) $E\left[\|\widetilde{g}(X,v^*) - g(X,v^*)\|_E^2\right](\delta_n)^2 = o_P(\frac{1}{n})$.*

**Assumption 3.5.** *either (a) or (b) holds: (a) $\left\{(\frac{dm(X,\alpha_0)}{d\alpha}[v^*])'\Sigma(X)^{-1} m(X,\alpha) : \alpha \in \mathcal{N}_{0n}\right\}$ is a Donsker class; (b) $\{m(\cdot,\alpha) : \alpha \in \mathcal{N}_{0n}\} \subseteq \Lambda_c^{\gamma_m}(\mathcal{X})$ with $r_m \equiv \gamma_m/d_x > 1/2$.*

In the proof of Theorem 3.1 we establish that, under assumption 2.10(ii), assumption 3.5(b) implies assumption 3.5(a).

**Assumption 3.6.** *(i)* $m(X,\alpha)$ *is twice pathwise differentiable in* $\alpha \in \mathcal{N}_{0n}$, $E\left(\sup_{\alpha\in\mathcal{N}_{0n}}\left|\frac{d^2m(X,\alpha)}{d\alpha d\alpha}[v_n^*,v_n^*]\right|^2\right) < \infty$; *(ii)* $E\left[\sup_{\alpha\in\mathcal{N}_{0n}}\left\|\frac{dm(X,\alpha)}{d\alpha}[v_n^*]-\frac{dm(X,\alpha_0)}{d\alpha}[v_n^*]\right\|_E^2\right] = o(n^{-1/2})$; *(iii) for all* $\alpha\in\mathcal{N}_{0n}$, $\overline{\alpha}\in\mathcal{N}_0$,

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

Assumption 3.6(ii) can be replaced by *Assumption 3.6(ii)':* $\left\{\frac{dm(X,\alpha)}{d\alpha}[v_n^*]:\alpha\in\mathcal{N}_{0n}\right\}$ *is a Donsker class,* $\sup_{x\in\mathcal{X},\alpha\in\mathcal{N}_{0n}}\left|\frac{dm(x,\alpha)}{d\alpha}[v_n^*]\right| \le const. < \infty$ *and* $E\left[\left\|\frac{dm(X,\alpha)}{d\alpha}[v_n^*]-\frac{dm(X,\alpha_0)}{d\alpha}[v_n^*]\right\|_E^2\right] = o(n^{-1/2})$ *for all* $\alpha\in\mathcal{N}_{0n}$ (see the working paper version Chen and Pouzo (2008b) or Ai and Chen (2003)). Assumption 3.6 is imposed to control the second order remainder term of $m(X,\alpha)$ in a shrinking neighborhood of $\alpha_0$. It is automatically satisfied when $m(X,\alpha)$ is linear in $\alpha$. When $h(Y)$ enters $m(X,\alpha) = E[\rho(Y,X,\theta,h(Y))|X]$ highly nonlinearly, we need some rate of convergence in strong norm ($\delta_{s,n}$) to verify assumption 3.6(ii)(iii); see, e.g., the partially linear quantile IV regression example in Section 5.

**Theorem 3.1.** *Let* $\widehat{\alpha}_n$ *be the PSMD estimator (2.1) with* $\lambda_n \ge 0$, $\lambda_n = o(1)$ *and* $\widehat{m}$ *the series LS estimator. Suppose that all the assumptions of Lemma 2.3 and Remark 2.1 hold. Let assumptions 3.1, and 3.2 - 3.6 hold. Then:* $\sqrt{n}(\hat{\theta}_n-\theta_0)\Rightarrow N(0,V^{-1})$, *where*

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

## 3.2 Weighted Bootstrap

In this subsection we propose a weighted bootstrap procedure, and establish its validity 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 $\theta_0$). 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 models (1.1) with nonparametric endogeneity.

**Assumption 3.7.** $\{W_i\}_{i=1}^n$ *is an i.i.d. sample of positive weights satisfying* $E[W_i]=1$ *and* $Var(W_i)=\omega_0\in(0,\infty)$, *and is independent of* $\{(Y_i',X_i')\}_{i=1}^n$.

**Theorem 3.2.** *Let all the assumptions of Theorem 3.1 and assumption 3.7 hold. Let*[6]

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

[6]We are indebted to Andres Santos for suggesting this weighted bootstrap procedure.

*where* $\widehat{m}_W(X,\alpha) = p^{J_n}(X)'(P'P)^- \sum_{i=1}^n p^{J_n}(X_j)\rho(Z_j,\alpha)W_j$ *Then: Conditional on the data* $\{(Y_i', X_i')\}_{i=1}^n$, $\sqrt{\frac{n}{\omega_0}}(\widehat{\theta}_n^* - \widehat{\theta}_n)$ *has the same limiting distribution as that of* $\sqrt{n}(\widehat{\theta}_n - \theta_0)$.

Chen et al. (2003) establish the validity of a nonparametric bootstrap for a two-step semiparametric GMM estimator of $\theta_0$ under high-level conditions. As Theorem 3.2 indicates, we obtain the validity of the weighted bootstrap under basically the same sets of low-level conditions as those for the root-$n$ asymptotic normality of the original PSMD estimator $\widehat{\theta}$.

# 4 Semiparametric Efficiency and Chi-square Approximation

## 4.1 Semiparametric efficiency bounds and efficient estimation

Recall that $\Sigma_0(X) \equiv Var(\rho(Z,\alpha_0)|X)$. We define $\overline{\mathbf{V}}_0$ and $\overline{\mathcal{W}}_0$ in the same way as $\overline{\mathbf{V}}$ and $\overline{\mathcal{W}}$ defined in subsection 2.2, except using the optimal weighting matrix $\Sigma_0(X)$ instead of $\Sigma(X)$. For example, $\overline{\mathcal{W}}_0 \equiv \left\{ w : E\left[ \left\| \Sigma_0(X)^{-\frac{1}{2}} \{ \frac{dm(X,\alpha_0)}{dh}[w] \} \right\|_E^2 \right] < \infty \right\}$. For any $\mathbf{w} \equiv (w_1, ..., w_{d_\theta})$ with $w_j \in \overline{\mathcal{W}}_0$, let $D_{\mathbf{w}}(X) \equiv \frac{dm(X,\alpha_0)}{d\theta} - \frac{dm(X,\alpha_0)}{dh}[\mathbf{w}]$, and define

$$V_0 \equiv \inf_{\mathbf{w}} E\left\{ D_{\mathbf{w}}(X)'[\Sigma_0(X)]^{-1} D_{\mathbf{w}}(X) \right\} = E\left\{ D_{\mathbf{w}_0}(X)'[\Sigma_0(X)]^{-1} D_{\mathbf{w}_0}(X) \right\}, \tag{4.1}$$

where $\mathbf{w}_0 \equiv (w_{01}, ..., w_{0d_\theta})$, and for $j = 1, ..., d_\theta$, each $w_{0j} \in \overline{\mathcal{W}}_0$ solves:

$$E\left[ \left( \frac{dm(X,\alpha_0)}{dh}[w_j] \right)' \Sigma_0(X)^{-1} \left( \frac{dm(X,\alpha_0)}{d\theta_j} - \frac{dm(X,\alpha_0)}{dh}[w_{0j}] \right) \right] = 0 \text{ for all } w_j \in \overline{\mathcal{W}}_0.$$

When the residual function $\rho(Z,\alpha)$ is pointwise smooth in $\alpha$, Ai and Chen (2003) establish that $V_0$ is the semiparametric efficiency (information) bound for $\theta_0$ identified by the model (1.1). The following theorem shows that their result remains valid when $\rho(Z,\alpha)$ is not pointwise smooth in $\alpha$. We denote $q_0(y,x,\alpha_0)$ as the true joint probability density of $(Y,X)$. Since $\mathcal{A}_{os}$ is convex at $\alpha_0$ by assumption, $h_0 + \xi(h - h_0) \in \{h \in \mathcal{H} : ||h - h_0||_s = o(1)\}$ for any $h \in \{h \in \mathcal{H} : ||h - h_0||_s = o(1)\}$ and any small scalar $\xi \geq 0$. Let $\{q(y,x,\theta,h_0 + \xi(h - h_0);\zeta) : \theta \in int(\Theta), \xi \geq 0, \zeta \geq 0\}$ denote a family of all parametric density submodels that satisfies the conditional moment restriction $\int \rho(y, X_z, \theta, h_0 + \xi(h - h_0)) q(y, X, \theta, h_0 + \xi(h - h_0); \zeta) dy = 0$ a.s.$-X$, and passes through $q_0(y,x,\alpha_0)$ at the true values $\theta = \theta_0$, $\xi = 0$ and $\zeta = 0$.

**Assumption 4.1.** *(i)* $E\left[ \{ \frac{dm(X,\alpha_0)}{d\theta} \}' \Sigma_0(X)^{-1} \{ \frac{dm(X,\alpha_0)}{d\theta} \} \right] < \infty$*; (ii) for any* $h \in \mathcal{H}_{os}$, $\{q(y,x,\theta,h_0 + \xi(h - h_0);\zeta) : \theta \in int(\Theta), \xi \geq 0, \zeta \geq 0\}$ *is smooth in the sense of Van der Vaart (1991).*

Denote $v_0 \equiv (v_\theta^0, -\mathbf{w}_0 \times v_\theta^0)$ with $v_\theta^0 \equiv (V_0)^{-1}\lambda$ for non-zero $\lambda \in \mathcal{R}^{d_\theta}$.

**Theorem 4.1.** *Let assumptions 2.1, 2.2(ii), 2.9(i), 3.2(i)(ii) and 4.1 hold. Then: (1)* $V_0$ *given in (4.1) is the semiparametric efficiency (information) bound for* $\theta_0$ *in the model (1.1). (2) The*

*positive definiteness of $V_0$ is the necessary condition for $\theta_0$ to be estimable at $\sqrt{n}-$rate. (3) Suppose that all the assumptions of Theorem 3.1 hold with $\Sigma(X) = \Sigma_0(X)$ and $v^* = v_0$, then the corresponding PSMD estimator of $\theta_0$ is efficient with asymptotic variance $V_0^{-1}$.*

Under assumption 4.1, the semiparametric information bound $V_0$ given in (4.1) is always well-defined and unique, albeit the solutions $\mathbf{w}_0$ (the "least favorable directions") may not be unique, and may not be solvable in closed-forms. Luckily, our optimally weighted PSMD estimator of $\theta_0$ is automatically efficient, regardless whether there is a closed-form solution $\mathbf{w}_0$.

### 4.2 Chi-square approximation

Previously Murphy and Van der Vaart (2000) show that the profiled likelihood ratio statistics is asymptotically chi-square distributed, and Shen and Shi (2005) establish that the profiled sieve likelihood ratio statistic is asymptotically chi-square distributed. In this subsection we show that the profiled optimally weighted SMD criterion ($\widehat{Q}_n(\theta)$) also possess such a nice property.

In the following we denote $\widehat{\Sigma}(X,\alpha)$ as any nonparametric estimator of $\Sigma(X,\alpha) \equiv Var[\rho(Z,\alpha)|X]$, and $\widehat{\alpha}_n$ as any initial consistent estimator such that $\widehat{\alpha}_n \in \mathcal{N}_{0n}$ with probability approaching one (e.g., the PSMD estimator with $\widehat{\Sigma}(X_i) = I$). The profiled optimally weighted PSMD estimator $\widetilde{\alpha}_n \equiv (\tilde{\theta}_n, \widetilde{h}_n)$ is defined as:

$$\begin{aligned}
\widetilde{h}_\theta &\equiv \arg \inf_{h \in \mathcal{H}_{k(n)}} \{\frac{1}{n}\sum_{i=1}^n \widehat{m}(X_i,\theta,h)'[\widehat{\Sigma}(X_i,\widehat{\alpha}_n)]^{-1}\widehat{m}(X_i,\theta,h) + \lambda_n \widehat{P}_n(h)\} \text{ for any fixed } \theta, \\
\tilde{\theta}_n &\equiv \arg\min_{\Theta}\{\frac{1}{n}\sum_{i=1}^n \widehat{m}(X_i,\theta,\widetilde{h}_\theta)'[\widehat{\Sigma}(X_i,\widehat{\alpha}_n)]^{-1}\widehat{m}(X_i,\theta,\widetilde{h}_\theta) + \lambda_n \widehat{P}_n(h)\}, \text{ and } \widetilde{h}_n \equiv \widetilde{h}_{\tilde{\theta}_n}.
\end{aligned}$$

Define the profiled optimally weighted SMD criterion function as:

$$\widehat{Q}_n(\theta) \equiv \frac{1}{2n}\sum_{i=1}^n \widehat{m}(X_i;\theta,\widetilde{h}_\theta)'[\widehat{\Sigma}(X_i,\widehat{\alpha}_n)]^{-1}\widehat{m}(X_i;\theta,\widetilde{h}_\theta).$$

In the following we define $v_n^0 \equiv (v_\theta^0, -\mathbf{w}_n^0 v_\theta^0) \in \mathcal{A}_{k(n)} \setminus \{\alpha_0\}$ the same way as $v_n^* \equiv (v_\theta^*, -\mathbf{w}_n^* v_\theta^*)$ defined in section 3 except using $\Sigma_0(X)$ instead of $\Sigma(X)$.

**Assumption 4.2.** *(i) $\sup_{x\in\mathcal{X},\alpha\in\mathcal{N}_{0n}} |\widehat{\Sigma}(x,\alpha) - \Sigma_0(x)| = O_P(\delta_{\Sigma,n})$; (ii) $\widehat{\Sigma}(X,\alpha)$ is finite and positive definite with eigenvalues bounded away from zero uniformly for all $X \in \mathcal{X}$ and $\alpha \in \mathcal{N}_{0n}$; (iii) $\lambda_n \sup_{\alpha\in\mathcal{N}_{0n}} |\widehat{P}_n(h \pm \varepsilon_n \mathbf{w}_n^0 v_\theta^0) - \widehat{P}_n(h)| = o_P(\frac{1}{n})$ with $0 < \varepsilon_n = O(n^{-1/2})$.*

Lemma A.1 of the working paper version (Chen and Pouzo (2008b)) provides sufficient conditions for assumption 4.2(i) when $\widehat{\Sigma}(x,\alpha)$ is a series LS estimator. For alternative nonparametric variance estimators and their properties, see Robinson (1995b), Andrews (1995) and references therein.

**Theorem 4.2.** *Suppose that all the assumptions of Theorem 3.1 hold with* $\Sigma(X) = \Sigma_0(X)$. *Let assumptions 4.2 hold. Then:* $\sqrt{n}(\widetilde{\theta}_n - \theta_0) \Rightarrow N(0, V_0^{-1})$ *and* $2n\left(\widehat{Q}_n(\theta_0) - \widehat{Q}_n(\tilde{\theta}_n)\right) \Rightarrow \chi^2_{d_\theta}$.

See the working paper version (Chen and Pouzo (2008b)) for an analogous result for the profiled continuously updated PSMD criterion.

**Remark 4.1.** *For the partially linear IV mean regression model:* $Y_3 = Y_1'\theta_0 + h_0(Y_2) + U$ *with* $E[U|X] = 0$, *Florens et al. (2007) provide identification of* $(\theta_0, h_0)$, *propose a kernel-based Tikhonov regularized estimator for* $\theta_0$, *and obtain its root-n asymptotic normality without assuming compactness of space* $\mathcal{H}$. *For this model, we can compute our optimally weighted PSMD estimator* $\tilde{\theta}_n$, *and Theorem 4.2 immediately implies that:* $\sqrt{n}(\widetilde{\theta}_n - \theta_0) \Rightarrow N(0, V_0^{-1})$ *and* $2n\{\widehat{Q}_n(\theta_0) - \widehat{Q}_n(\tilde{\theta}_n)\} \Rightarrow \chi^2_{d_\theta}$, *where* $\Sigma_0(X) = Var\{U^2|X\}$ *and*

$$V_0 \equiv \inf_{\mathbf{w}} E\left\{E[Y_1' - \mathbf{w}(Y_2)|X]'\Sigma_0(X)^{-1}E[Y_1' - \mathbf{w}(Y_2)|X]\right\}.$$

*Moreover, since* $m(X, \alpha) = E[Y_3 - Y_1'\theta - h(Y_2)|X]$ *is linear in* $\alpha = (\theta, h)$, *assumption 3.6 is trivially satisfied; hence* $\tilde{\theta}_n$ *is root-n asymptotically normal even if* $\widetilde{h}_n$ *converges to* $h_0$ *very slowly in a strong norm, such as* $||\widehat{h}_n - h_0||_{L^2(f_{Y_2})} = O_P\left([\ln(n)]^{-\varsigma/a}\right)$ *in Remark 2.2.*

# 5 A Partially Linear Quantile IV Example

In this section we apply the above general theories to study the following partially linear quantile IV regression model:

$$Y_3 = \theta_0 Y_1 + h_0(Y_2) + U, \quad \Pr(U \leq 0|X) = \gamma \in (0,1), \tag{5.1}$$

where $\theta_0$ is a scalar unknown parameter, $h_0()$ is a real-valued unknown function, and 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 $\gamma$. The support of $X$ is $\mathcal{X} = [0,1]^{d_x}$ with $d_x = 1+d$, and the support of $Y = (Y_3, Y_1, Y_2')'$ is $\mathcal{Y} \subseteq \mathcal{R}^{2+d}$. To map into the general model (1.1), we let $Z = (Y', X')'$, $\alpha = (\theta, h)$ and $\rho(Z, \alpha) = 1\{Y_3 \leq \theta Y_1 + h(Y_2)\} - \gamma$. Recently Chernozhukov et al. (2007) and Horowitz and Lee (2007) study the nonparametric quantile IV regression model $E[1\{Y_3 \leq h_0(Y_2)\}|X] = \gamma$. Chen and Pouzo (2008a) illustrate 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) consider an example of partially linear quantile IV regression with an exogenous $Y_2$ (i.e., $Y_2 = X_2$), and Lee (2003) studies 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 estimate $\alpha_0$ using the PSMD estimator $\widehat{\alpha}_n$, with $\widehat{m}(X, \alpha)$ being a series LS estimator of $m(X, \alpha) = E[F_{Y_3|Y_1,Y_2,X}(\theta Y_1 + h(Y_2))|X] - \gamma$, where $p^{J_n}(X)$ is either B-splines, P-splines, wavelets or cosine series. $\widehat{\Sigma}(X) = \Sigma(X) = \Sigma_0(X) = \gamma(1-\gamma)$, $\widehat{P}_n(h) = P(h)$, and $\mathcal{A}_n = [\underline{b}, \overline{b}] \times \mathcal{H}_{k(n)}$ being

a finite dimensional ($\dim(\mathcal{H}_{k(n)}) \equiv k(n) < \infty$) *linear* sieve. We impose some low level sufficient conditions:

**Condition 5.1.** *(i)* $f_{Y_3|Y_1,Y_2,X}(y_3|y_1,y_2,x)$ *is continuous in* $(y_3,y_1,y_2,x)$, *and* $\sup_{y_3} f_{Y_3|Y_1,Y_2,X}(y_3) \leq$ *const.* $< \infty$ *for almost all* $Y_1, Y_2, X$; *(ii)* $f_{Y_1,Y_2|X=x}(y_1,y_2)$ *is continuous in* $(y_1,y_2,x)$;

*(iii)* $E[F_{Y_3|Y_1,Y_2,X}(\theta Y_1 + h(Y_2))|X = \cdot] \in \Lambda_1^{\gamma_m}(\mathcal{X})$, $r_m \equiv \gamma_m/d_x > 1/2$, *for all* $(\theta,h) \in \Theta \times \mathcal{H}_{k(n)}$.

Let $f_{U|Y_1,Y_2,X}(0) = f_{Y_3|Y_1,Y_2,X}(\theta_0 Y_1 + h_0(Y_2))$. Denote $\varpi(y_2) \equiv \left(1+|y_2|^2\right)^{-\vartheta'/2}$ for some $\vartheta' \geq 0$.

**Condition 5.2.** *(i)* $0 < E\{\left(E[f_{U|Y_1,Y_2,X}(0)Y_1|X]\right)^2\} < \infty$; *(ii)* $E[(1+|Y_2|)^{2\vartheta}] < \infty$ *for some* $\vartheta > \vartheta' \geq 0$; *(iii)* $\Theta = [\underline{b}, \overline{b}] \subset \mathcal{R}$, $\mathcal{H} \subseteq \{h \in L^2(\mathcal{R}^d, f_{Y_2}) : ||\varpi h||_{W_2^\varsigma(leb)} < \infty\}$ *with* $\varsigma > 0$; *(iv)* $\mathcal{H}_{k(n)} =$ *span*$\{q_1, \ldots, q_{k(n)}\}$ *with* $(q_k)_k$ *being wavelet basis for* $W_2^\varsigma(\mathcal{R}^d, \varpi)$; *(v)* $P(h) = ||\nabla^{\varsigma'}(\varpi h)||^2_{L^2(leb)}$ *with* $0 < \varsigma' \leq \varsigma$.

**Condition 5.3.** *(i)* $(\theta_0, h_0) \in int(\Theta) \times \mathcal{H}$ *satisfies the model (5.1); (ii) for all* $\alpha \in \Theta \times \mathcal{H}$ *and all* $\tau \in [0,1]$ *with* $\alpha_\tau \equiv \tau\alpha_0 + (1-\tau)\alpha$, $E\left\{f_{Y_3|Y_1,Y_2,X}(\theta_\tau Y_1 + h_\tau(Y_2))[Y_1(\theta - \theta_0) + h(Y_2) - h_0(Y_2)]|X\right\} = 0$ *implies that* $[Y_1(\theta - \theta_0) + h(Y_2) - h_0(Y_2)] \equiv 0$ *almost surely; (iii)* $E\{|Y_1 - E(Y_1|Y_2)|\} > 0$.

Condition 5.3 is a sufficient condition to ensure that the model (5.1) has a unique solution $\alpha_0 = (\theta_0, h_0) \in \Theta \times \mathcal{H}$.

Let $||h||_s^2 = E\{[h(Y_2)]^2\}$ and $\mathcal{A}_{os} = \{\alpha \in \Theta \times \mathcal{H} : |\theta - \theta_0| + ||h - h_0||_s = o(1), P(h) \leq c\}$. Define linear operators $T_\alpha[g] \equiv E\{f_{Y_3|Y_1,Y_2,X}(\theta Y_1 + h(Y_2))[g(Y_2)]|X = \cdot\}$ and $T_{\alpha_0}[g] \equiv E\{f_{Y_3|Y_1,Y_2,X}(\theta_0 Y_1 + h_0(Y_2))[g(Y_2)]|X = \cdot\}$ that map from $Dom(T_\alpha) \subset L^2(f_{Y_2}) \to L^2(\mathcal{X}, f_X)$. Condition 5.3(ii) also implies that $T_{\alpha_0}$ is invertible (i.e., injective, i.e., $\{g : T_{\alpha_0}[g] = 0\} = \{0\}$). We assume

**Condition 5.4.** *For all* $g \in Dom(T_\alpha)$, *(i) there are finite constants* $c, C > 0$ *such that* $c||T_{\alpha_0}[g]||^2_{L^2(f_X)} \leq ||T_\alpha[g]||^2_{L^2(f_X)} \leq C||T_{\alpha_0}[g]||^2_{L^2(f_X)}$ *for all* $\alpha \in \mathcal{A}_{os}$; *(ii) there are finite positive constants* $a, c, c' > 0$ *such that* $c||T_{\alpha_0}[g]||^2_{L^2(f_X)} \leq \sum_{j=1}^\infty j^{-2a/d}|\langle g, q\rangle_s|^2 \leq c'||T_{\alpha_0}[g]||^2_{L^2(f_X)}$.

**Condition 5.5.** $\max\{\sqrt{\frac{J_n}{n}}, J_n^{-2r_m}, \lambda_n\} = \sqrt{\frac{J_n}{n}} = o(1)$, $J_n > k(n)$, $J_n = O(k(n))$.

Note that $\overline{\mathcal{W}}_0 = \left\{w : E\left[\{E[f_{U|Y_1,Y_2,X}(0)w(Y_2)|X]\}^2\right] < \infty\right\}$. Denote $w_0 \in \overline{\mathcal{W}}_0$ as the solution to:

$$V_0 = \inf_{w \in \overline{\mathcal{W}}_0} \frac{E\left[\left(E\left\{f_{U|Y_1,Y_2,X}(0)[Y_1 - w(Y_2)]|X\right\}\right)^2\right]}{\gamma(1-\gamma)} = \frac{E\left[\left(E\left\{f_{U|Y_1,Y_2,X}(0)[Y_1 - w_0(Y_2)]|X\right\}\right)^2\right]}{\gamma(1-\gamma)}. \tag{5.2}$$

Conditions 5.2(i) and 5.3 imply that $V_0 \in (0, \infty)$.

**Condition 5.6.** *(i) there is an* $w_n^0 \in \mathcal{H}_{k(n)}$ *such that* $E[\left(E\left\{f_{U|Y_1,Y_2,X}(0)[w_n^0(Y_2) - w_0(Y_2)]|X\right\}\right)^2] = o(\{k(n)\}^{-1})$; *(ii) assumption 4.2(iii) holds with* $\widehat{P}_n(h) = P(h)$.

**Condition 5.7.** *(i)* $\varsigma > \max\{a + \frac{d}{2}, 2d\}$; *(ii) for almost all* $Y_1, Y_2, X$, *the partial derivative of* $f_{Y_3|Y_1,Y_2,X}(y_3)$ *with respect to* $y_3$ *exists, is continuous and bounded uniformly in* $y_3$.

In the following, $\widehat{h}_\theta$ denotes the profile PSMD estimator, obtained by fixing $\theta$ and minimizing the PSMD criterion with respect to $h \in \mathcal{H}_n$. By definition $\widehat{h}_n = \widehat{h}_{\widehat{\theta}_n}$.

**Proposition 5.1.** *For the model (5.1), suppose that $\{(Y_i', X_i')\}_{i=1}^n$ is i.i.d., assumption 2.7, and conditions 5.1 - 5.5 hold. Let $k(n) = O\left(n^{\frac{d}{2(\varsigma+a)+d}}\right)$. Then:*

*(1) $||\widehat{h}_n - h_0||_{L^2(f_{Y_2})} = O_P\left(n^{-\frac{\varsigma}{2(\varsigma+a)+d}}\right)$ and $||\widehat{\alpha}_n - \alpha_0|| = O_P\left(n^{-\frac{\varsigma+a}{2(\varsigma+a)+d}}\right)$.*

*(2) If assumptions 5.6 and 5.7 hold, then: $\sqrt{n}(\widehat{\theta}_n - \theta_0) \Rightarrow N\left(0, V_0^{-1}\right)$, where $V_0$ given in equation (5.2) is the semiparametric efficiency bound, and*

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

**Remark 5.1.** ***(1)*** *Proposition 5.1 is directly applicable to the model $E[1\{Y_3 \leq \theta X_1 + h_{02}(X_2)\}|X] = \gamma$ (i.e., $Y_j = X_j$ for $j = 1, 2$, no endogeneity) studied in Lee (2003)), and to the model $E[1\{Y_3 \leq \theta Y_1 + h_{02}(X_2)\}|X] = \gamma$ (i.e., $Y_2 = X_2$, $Y_1 \neq X_1$, endogeneity only in parametric part) considered in Chen et al. (2003)). For both models we have: $a = 0$, and Proposition 5.1 leads to the optimal convergence rate $O_P\left(n^{-\frac{\varsigma}{2\varsigma+d}}\right)$ of $\widehat{h}_n$ to $h_0$ in norm $||\cdot||_{L^2(f_{Y_2})}$, the root-n semiparametric efficient estimation of $\theta_0$, and the chi-square approximation of the PSMD criterion based test statistics of the null hypothesis: $\theta = \theta_0$.*

***(2)*** *One can characterize the semiparametric efficiency bound $V_0$ using the operator formulation. In particular, any solution $w_0 \in \overline{\mathcal{W}}_0$ to the optimization problem (5.2) must satisfy:*

$$\left\langle T_{\alpha_0}[w], E\left\{f_{U|Y_1,Y_2,X}(0) Y_1 | X\right\} - T_{\alpha_0}[w_0] \right\rangle_{L^2(f_X)} = 0 \text{ for all } w \in \overline{\mathcal{W}}_0. \tag{5.3}$$

*Let $T_{0,\theta}(\cdot) \equiv E\left\{f_{U|Y_1,Y_2,X}(0) Y_1 | X = \cdot\right\}$. Define $T_{\alpha_0}^*[r] \equiv E\left\{f_{U|Y_1,Y_2,X}(0) r(X) | Y_2 = \cdot\right\}$ as the adjoint operator of $T_{\alpha_0}$ (i.e., $\langle T_{\alpha_0}[g], r\rangle_{L^2(f_X)} = \left\langle g, T_{\alpha_0}^*[r]\right\rangle_{L^2(f_{Y_2})}$). Then Condition 5.3 implies that $T_{\alpha_0}^* T_{\alpha_0}$ is invertible, and if $|| \left(T_{\alpha_0}^* T_{\alpha_0}\right)^{-1} T_{\alpha_0}^*[T_{0,\theta}]||_{L^2(f_{Y_2})} < \infty$ then*

$$V_0 = \frac{E\left[\left(T_{0,\theta}(X) - T_{\alpha_0}[w_0]\right)^2\right]}{\gamma(1-\gamma)}, \; w_0 = \left(T_{\alpha_0}^* T_{\alpha_0}\right)^{-1} T_{\alpha_0}^*[T_{0,\theta}].$$

*However, this does not imply any explicit expressions for $w_0()$ when $h_0()$ depends on the endogenous variable $Y_2$. When there is no nonparametric endogeneity, then one can solve the efficiency bound in closed-forms. For example, for the partially linear quantile model with parametric endogeneity (i.e., $Y_2 = X_2$, but $Y_1 \neq X_1$) of Chen et al. (2003), we have: $V_0 = [\gamma(1-\gamma)]^{-1} E[\left(E\left\{f_{U|Y_1,X}(0)[Y_1 - w_0(X_2)]|X\right\}\right)^2]$ and*

$$w_0(X_2) = \frac{E\left(E\left[f_{U|Y_1,X}(0) Y_1 | X\right] E\left[f_{U|Y_1,X}(0) | X\right] | X_2\right)}{E[\left(E\{f_{U|Y_1,X}(0) | X\}\right)^2 | X_2]}.$$

Proposition 5.1 only establishes the root-$n$ asymptotic normality and efficiency for the partially linear quantile IV model (5.1) under the mildly ill-posed case ($||\widehat{h}_n - h_0||_{L^2(\mathcal{R}^d, f_{Y_2})} = O_P\left(n^{-\frac{\varsigma}{2(\varsigma+a)+d}}\right)$). This is because the model is *nonlinear* in $h_0(Y_2)$, our sufficient conditions (assumption 3.6(ii)(iii)) for root-$n$ normality rules out the severely ill-posed case ($||\widehat{h}_n - h_0||_{L^2(f_{Y_2})} = O_P\left([\ln(n)]^{-\varsigma/a}\right)$).

# 6 Simulation and Empirical Illustration

## 6.1 A Monte Carlo Study

We assess the finite sample performance of the PSMD 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 $\varepsilon$, and both are independent of $(Y_2, X_2)$. 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), $\Phi(X_2)$ is its instrument (log-gross earnings), and the joint density of $(Y_2, X_2)$ is a bivariate Gaussian density with first and second moments estimated from the FES data set. We draw an i.i.d. sample from the joint density of $(Y_2, X_2, X_1, \varepsilon)$ with sample size $n = 1000$.

We estimate $m(X, \alpha)$ by the series LS estimator $\widehat{m}(X, \alpha)$ given in (2.2), where $p^{J_n}(X)$ consists of P-Spline(3,3), P-Cos(9) and 4 cross-products terms (the second term of P-Spline(3,3) times the first four terms of P-Cos(9)) with $J_n = 20$.[7] We use a linear spline sieve P-Spline(2,6) as $\mathcal{H}_n$. We add a penalization term $\widehat{P}_n(h) = ||\nabla h||^2_{L^2(leb)}$ with $\lambda_n \in [0.001, 0.01]$.[8] In all the cases we performed 500 Monte Carlo repetitions. Table 1 and Figure 1 summarize the simulation results for different quantiles $\gamma = 0.125, 0.25, 0.5, 0.75, 0.875$. One can see that for all the cases our estimator performs well.

## 6.2 An Empirical Illustration

We apply the PSMD to estimate a shape-invariant system of quantile IV Engel curves (or consumer demand functions) using the UK Family Expenditure Survey data. The model is

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

[7] P-Spline(p,q) denotes a polynomial spline of order p with q number of knots, and P-Cos(p) a cosine series with p number of terms. We have tried other combinations as sieve bases for $\widehat{m}$ and all yield very similar results.

[8] The actual $\lambda_n$ is chosen to minimize the integrated MSE of $\widehat{h}$ for a small number of Monte Carlo repetitions.

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 $X_i \equiv (X_{1i}, X_{2i})'$, where $X_{1i}$ is 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. Blundell et al. (2007) have used the same data set in their study of a shape-invariant system of mean IV Engel curves.

As illustration, we apply the PSMD 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 $\widehat{P}_n(h) = ||\nabla^k h||^j_{L^j(d\widehat{\mu})} \equiv n^{-1}\sum_{i=1}^n |\nabla^k h(Y_{2i})|^j$ for $k = 1, 2$ and $j = 1, 2$, and Hermite polynomial sieves, cosine sieves and polynomial spline sieves for the series LS estimator $\hat{m}$. All combinations yield very similar results. Due to the lack of space, in Figure 2 we report the PSMD estimated Engel curves only for three different quantiles $\gamma = \{0.25, 0.5, 0.75\}$ and for four selected goods, using P-Spline(2,5) as $\mathcal{H}_n$, and $p^{J_n}(X)$ for $\hat{m}$ consisting of P-Spline(2,5), P-Spline(5,10) and 4 cross-product terms (the second term of P-Spline(2,5) times the first four terms of P-Spline(5,10)), with $J_n = 27$. Table 2 presents the corresponding PSMD estimates of $\theta_1$ and $(\theta_{2,l})_{l=1}^7$ for the median ($\gamma = 0.5$) case under different combinations of $\widehat{P}_n(h)$ and $\lambda_n$. Figure 2 presents the corresponding estimated curves, and its last two rows include the Engel curve estimates of Blundell et al. (2007) for comparison. Both our $\theta$ estimates and our Engel curve estimates for the $\gamma = 0.5$ quantile are very similar to the estimates reported in Blundell et al. (2007) for the mean IV Engel curve model.

# 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 optimally weighted 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 semiparametric conditional moment models of Ai and Chen (2007).

All the large sample theories obtained in this paper are first-order asymptotics. There are no

results on higher order refinement for semiparametric models (1.1) containing unknown functions of endogenous variables yet. There are some second order theories for semiparametric models without nonparametric endogeneity, such as Robinson (1995a), Linton (1995) 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.

**Acknowledgments**

We thank the guest editor, an anonymous referee and A. Santos whose comments have greatly improved the paper. We acknowledge helpful discussions with R. Blundell, V. Chernozhukov, J. Horowitz, R. Klein, S. Lee, O. Linton, E. Mammen, W. Newey, J. Powell, E. Renault, P. Robinson, N. Swanson, E. Tamer, F. Vella 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, Toulouse, Montreal, Rutgers, Georgetown, UCSD and UCLA. We thank R. Blundell for sharing the UK Family Expenditure Survey data set. Chen acknowledges financial support from National Science Foundation grant SES-0631613. Any errors are the responsibility of the authors.

# A Mathematical Appendix

**Proof of Lemma 2.2**: Assumption 2.9 implies that for any $\alpha = (\theta, h) \in \mathcal{A}_{os}$ with $\theta \neq \theta_0$, we can always rewrite $h - h_0 = -\mathbf{w}(\theta - \theta_0)$ with $\mathbf{w} = (w_1, ..., w_{d_\theta}) \in \overline{\mathcal{W}} \times \cdots \times \overline{\mathcal{W}}$. By definition of $\mathbf{w}^*$ we have for all $\mathbf{w} = (w_1, ..., w_{d_\theta})$ and for each $j = 1, ..., d_\theta$,

$$E\left[D_{w_j^*}(X)'\Sigma(X)^{-1}\left\{D_{w_j}(X) - D_{w_j^*}(X)\right\}\right] = E\left[D_{w_j^*}(X)'\Sigma(X)^{-1}\frac{dm(X,\alpha_0)}{dh}[w_j - w_j^*]\right] = 0.$$

Thus

$$\begin{aligned} & ||\alpha - \alpha_0||^2 = (\theta - \theta_0)'E\left[D_{\mathbf{w}}(X)'\Sigma(X)^{-1}D_{\mathbf{w}}(X)\right](\theta - \theta_0) \\ = \; & (\theta - \theta_0)'\left(\begin{array}{c} E\left[D_{\mathbf{w}^*}(X)'\Sigma(X)^{-1}D_{\mathbf{w}^*}(X)\right] \\ +E\left[\left\{\frac{dm(X,\alpha_0)}{dh}[\mathbf{w}^* - \mathbf{w}]\right\}'\Sigma(X)^{-1}\left\{\frac{dm(X,\alpha_0)}{dh}[\mathbf{w}^* - \mathbf{w}]\right\}\right] \end{array}\right)(\theta - \theta_0). \end{aligned}$$

By assumption 2.10(i) we have:

$$||\alpha - \alpha_0||^2 \geq (\theta - \theta_0)'E\left[D_{\mathbf{w}^*}(X)'\Sigma(X)^{-1}D_{\mathbf{w}^*}(X)\right](\theta - \theta_0) \geq const.||\theta - \theta_0||_E^2.$$

Next,

$$||\alpha - \alpha_0||^2 \geq E\left[\left\|\Sigma(X)^{-\frac{1}{2}}\left\{\frac{dm(X,\alpha_0)}{dh}[h - h_0 + \mathbf{w}^*(\theta - \theta_0)]\right\}\right\|_E^2\right] = ||h - h_0 + \mathbf{w}^*(\theta - \theta_0)||^2.$$

Note that $||h-h_0||^2 \leq 2\left\{||h-h_0+\mathbf{w}^*(\theta-\theta_0)||^2+||\mathbf{w}^*(\theta-\theta_0)||^2\right\}$, and

$$||\mathbf{w}^*(\theta-\theta_0)||^2=(\theta-\theta_0)'E\left[\{\frac{dm(X,\alpha_0)}{dh}[\mathbf{w}^*]\}'\Sigma(X)^{-1}\{\frac{dm(X,\alpha_0)}{dh}[\mathbf{w}^*]\}\right](\theta-\theta_0).$$

Assumption 2.10(i)(ii) and

$$E\left[\left\|\Sigma(X)^{-\frac{1}{2}}\frac{dm(X,\alpha_0)}{d\theta}\right\|_E^2\right]=E\left[D_{\mathbf{w}^*}(X)'\Sigma(X)^{-1}D_{\mathbf{w}^*}(X)\right]+E\left[\left\|\Sigma(X)^{-\frac{1}{2}}\frac{dm(X,\alpha_0)}{dh}[\mathbf{w}^*]\right\|_E^2\right]$$

imply that $||\mathbf{w}^*(\theta-\theta_0)||^2 \leq const.||\theta-\theta_0||_E^2$. Thus $||h-h_0||^2 \leq const.||\alpha-\alpha_0||^2$. *Q.E.D.*

**Proof of Lemma 2.3**: For Result (1), assumption 2.5(ii) implies that there are two finite positive constants $c, c'$ such that

$$c\frac{1}{n}\sum_{i=1}^n\|\widehat{m}(X_i,\alpha)\|_E^2 \leq \frac{1}{n}\sum_{i=1}^n\left\|\widehat{\Sigma}(X_i)^{-\frac{1}{2}}\widehat{m}(X_i,\alpha)\right\|_E^2 \leq c'\frac{1}{n}\sum_{i=1}^n\|\widehat{m}(X_i,\alpha)\|_E^2$$

uniformly over $\alpha\in\mathcal{A}_{k(n)}$. Let $r_n^2=\max\{\delta_{m,n}^2,||\alpha_0-\Pi_{k(n)}\alpha_0||^2,\lambda_n|P(\Pi_{k(n)}h_0)-P(\widehat{h}_n)|\}=o_P(1)$. Since $\widehat{\alpha}_n\in\mathcal{A}_{osn}$ with probability approaching one, we have: for all $M>1$,

$$\begin{aligned}
\Pr\left(\frac{||\widehat{\alpha}_n-\alpha_0||}{r_n}\geq M\right) &\leq \Pr\left(\begin{array}{c}\inf_{\{\alpha\in\mathcal{A}_{osn}:||\alpha-\alpha_0||\geq Mr_n\}}\left[\frac{1}{n}\sum_{i=1}^n\left\|\widehat{\Sigma}(X_i)^{-\frac{1}{2}}\widehat{m}(X_i,\alpha)\right\|_E^2+\lambda_n\widehat{P}_n(h)\right]\\ \leq \frac{1}{n}\sum_{i=1}^n\left\|\widehat{\Sigma}(X_i)^{-\frac{1}{2}}\widehat{m}\left(X_i,\Pi_{k(n)}\alpha_0\right)\right\|_E^2+\lambda_n\widehat{P}_n(\Pi_{k(n)}h_0)\end{array}\right)\\
&\leq \Pr\left(\begin{array}{c}\inf_{\{\alpha\in\mathcal{A}_{osn}:||\alpha-\alpha_0||\geq Mr_n\}}\left[c\frac{1}{n}\sum_{i=1}^n\|\widehat{m}(X_i,\alpha)\|_E^2+\lambda_n\widehat{P}_n(h)\right]\\ \leq c'\frac{1}{n}\sum_{i=1}^n\left\|\widehat{m}\left(X_i,\Pi_{k(n)}\alpha_0\right)\right\|_E^2+\lambda_n\widehat{P}_n(\Pi_{k(n)}h_0)\end{array}\right)\\
&\leq \Pr\left(\begin{array}{c}\inf_{\{\alpha\in\mathcal{A}_{osn}:||\alpha-\alpha_0||\geq Mr_n\}}\left[c\frac{1}{n}\sum_{i=1}^n\|\widehat{m}(X_i,\alpha)\|_E^2+\lambda_nP(h)\right]\\ \leq c'\frac{1}{n}\sum_{i=1}^n\left\|\widehat{m}\left(X_i,\Pi_{k(n)}\alpha_0\right)\right\|_E^2+\lambda_nP(\Pi_{k(n)}h_0)+o_P(\lambda_n)\end{array}\right)
\end{aligned}$$

where the last inequality is due to the assumption that $\sup_{h\in\mathcal{H}_{osn}}|\widehat{P}_n(h)-P(h)|=o_P(1)$. We can now follow the proof of Theorem 4.1(1) of Chen and Pouzo (2008a) (using our $\widehat{\alpha}_n$ instead of their $\widehat{h}_n$), and obtain: $||\widehat{\alpha}_n-\alpha_0||\equiv O_P(\delta_n)=O_P(r_n)=O_P\left(\max\{\delta_{m,n},\sqrt{\lambda_n},||\alpha_0-\Pi_{k(n)}\alpha_0||\}\right)$. Result (1) now follows from our Lemma 2.2 and the fact that $||\alpha_0-\Pi_{k(n)}\alpha_0||=||h_0-\Pi_{k(n)}h_0||$. Result (2) follows from Result (1), Theorem 4.2 and Lemma 5.1 of Chen and Pouzo (2008a). *Q.E.D.*

**Lemma A.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 i.i.d. data, assumptions 2.7, 2.8(i) and 3.1(i)(ii) hold. Then:*

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

$$(2)\ \sup_{\alpha\in\mathcal{N}_{0n}}\frac{1}{n}\sum_{i=1}^n\|\widehat{m}(X_i,\alpha_0)+\widetilde{m}(X_i,\alpha)\|_E^2=O_P\left(\frac{J_n}{n}+\delta_n^2\right).$$

*(3) Let assumptions 2.5(ii) and 3.1(iii) hold, and $\frac{J_n}{n}=O(\delta_n^2)$. Then: Uniformly over $\alpha\in\mathcal{N}_{0n}$,*

$$\frac{1}{n}\sum_{i=1}^n\left\|\widehat{\Sigma}(X_i)^{-\frac{1}{2}}\widehat{m}(X_i,\alpha)\right\|_E^2=\frac{1}{n}\sum_{i=1}^n\left\|\widehat{\Sigma}(X_i)^{-\frac{1}{2}}\{\widehat{m}(X_i,\alpha_0)+\widetilde{m}(X_i,\alpha)\}\right\|_E^2\pm o_P\left(\frac{1}{n}\right).$$

**Proof of Lemma A.1**: For Result (1), let $\varepsilon(Z,\alpha) \equiv \rho(Z,\alpha) - m(X,\alpha)$, $\Delta\varepsilon(\alpha) \equiv \varepsilon(Z,\alpha) - \varepsilon(Z,\alpha_0)$, $\Lambda_n \equiv E[\|\rho(Z,\alpha) - \rho(Z,\alpha_0)\|_E^2 |X]$. Recall that $\widetilde{m}(X,\alpha) = p^{J_n}(X)'(P'P)^- \sum_{i=1}^n p^{J_n}(X_i) m(X_i,\alpha)$ is the LS projection of $m(X,\alpha)$ onto $p^{J_n}(X)$. Then

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

where the first inequality is due to Markov inequality, i.i.d. data, and the subsequent inequalities are due to assumptions 3.1(i)(ii), 2.7, 2.8(i), i.i.d. data and the definition of $\mathcal{N}_{0n}$. Thus Result (1) follows. For Result (2), by triangular inequality, we have: $\|\widehat{m}(X_i,\alpha_0) + \widetilde{m}(X_i,\alpha)\|_E \leq \|\widehat{m}(X_i,\alpha_0)\|_E + \|\widetilde{m}(X_i,\alpha)\|_E$. Following the proof of Lemma B.2 of Chen and Pouzo (2008a) (using our $\alpha$ instead of their $h$), under the i.i.d. data, assumptions 2.7 and 2.8(i) (for $\alpha \in \mathcal{A}_{k(n)}$) and 3.1(ii) (for $\alpha_0$), we obtain: there are finite constants $c, c' > 0$ such that, except on an event whose probability goes to zero as $n \to \infty$, $cE[||\widehat{m}(X,\alpha_0)||_E^2] \leq n^{-1}\sum_{i=1}^n ||\widehat{m}(X_i,\alpha_0)||_E^2 \leq c'E[||\widehat{m}(X,\alpha_0)||_E^2]$, and $cE\left(\|\widetilde{m}(X_i,\alpha)\|_E^2\right) \leq \frac{1}{n}\sum_{i=1}^n \|\widetilde{m}(X_i,\alpha)\|_E^2 \leq c'E\left(\|\widetilde{m}(X_i,\alpha)\|_E^2\right)$ uniformly over $\alpha \in \mathcal{A}_{k(n)}$. By the definition of $\widehat{m}$, the i.i.d. data, and assumption 3.1(ii), we have: $E\left(\|\widehat{m}(X_i,\alpha_0)\|_E^2\right) = O_P\left(\frac{J_n}{n}\right)$. Assumption 2.9(ii) and $\widetilde{m}(X,\alpha_0) = 0$ imply that $E\left(\|\widetilde{m}(X_i,\alpha)\|_E^2\right) \leq const.\|\alpha_0 - \alpha\|^2 = O_P(\delta_n^2)$ uniformly over $\alpha \in \mathcal{N}_{0n}$. Thus Result (2) follows.

For Result (3), denote $|||A(\cdot)|||_{\widehat{\Sigma}}^2 \equiv n^{-1}\sum_{i=1}^n A(X_i)'\widehat{\Sigma}(X_i)^{-1}A(X_i)$ and

$$
\ell(X,\alpha) \equiv \widehat{m}(X,\alpha_0) + \widetilde{m}(X,\alpha), \quad B_n^2 \equiv \sup_{\alpha\in\mathcal{N}_{0n}} \frac{1}{n}\sum_{i=1}^n \left\|\widehat{\Sigma}(X_i)^{-\frac{1}{2}}\left\{\widehat{m}(X_i,\alpha) - \ell(X_i,\alpha)\right\}\right\|_E^2.
$$

By triangle inequality we obtain that uniformly over $\alpha \in \mathcal{N}_{0n}$,

$$
|||\ell(\cdot,\alpha)|||_{\widehat{\Sigma}} - B_n \leq |||\widehat{m}(\cdot,\alpha)|||_{\widehat{\Sigma}} \leq |||\ell(\cdot,\alpha)|||_{\widehat{\Sigma}} + B_n. \tag{A.1}
$$

Results (1) and (2), assumptions 2.5(ii) and 3.1(iii), and $\frac{J_n}{n} = O(\delta_n^2)$ imply that: uniformly over $\alpha \in \mathcal{N}_{0n}$,

$$
\begin{aligned}
B_n^2 &= O_P\left(\frac{J_n}{n}(\delta_{s,n})^{2\kappa}\right) = O_P\left(\delta_n^2(\delta_{s,n})^{2\kappa}\right) = o_P\left(\frac{1}{n}\right), \\
|||\ell(\cdot,\alpha)|||_{\widehat{\Sigma}} \times B_n &= O_P\left(\left\{\sqrt{\frac{J_n}{n}} + \delta_n\right\} \times \sqrt{\frac{J_n}{n}}(\delta_{s,n})^{\kappa}\right) = O_P\left(\delta_n^2(\delta_{s,n})^{\kappa}\right) = o_P\left(\frac{1}{n}\right).
\end{aligned}
$$

These and equation (A.1) now imply that: $|||\widehat{m}(\cdot,\alpha)|||_{\widehat{\Sigma}}^2 = |||\ell(\cdot,\alpha)|||_{\widehat{\Sigma}}^2 \pm o_P\left(\frac{1}{n}\right)$ uniformly over $\alpha \in \mathcal{N}_{0n}$. *Q.E.D*

Denote $|||A(\cdot)|||^2_{\widehat{\Sigma}} \equiv n^{-1}\sum_{i=1}^n A(X_i)'\widehat{\Sigma}(X_i)^{-1}A(X_i)$, and $\ell(X,\alpha) \equiv \widehat{m}(X,\alpha_0) + \widetilde{m}(X,\alpha)$.

**Proof of Theorem 3.1**: Let $0 < \epsilon_n = o(n^{-1/2})$ and $u_n^* = \pm v_n^*$. By the definition of $\widehat{\alpha}_n$ and assumption 3.3(iii), we have: $|||\widehat{m}(\cdot,\widehat{\alpha}_n)|||^2_{\widehat{\Sigma}} - |||\widehat{m}(\cdot,\widehat{\alpha}_n+\epsilon_n u_n^*)|||^2_{\widehat{\Sigma}} + o_P(n^{-1}) \leq 0$. This and Lemma A.1(3) imply that

$$|||\ell(\cdot,\widehat{\alpha}_n)|||^2_{\widehat{\Sigma}} - |||\ell(\cdot,\widehat{\alpha}_n+\epsilon_n u_n^*)|||^2_{\widehat{\Sigma}} + o_P(n^{-1}) \leq 0. \tag{A.2}$$

Under assumption 3.6(i), we can perform second order Taylor expansion to equation (A.2), and obtain:

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

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

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

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

Applying Cauchy-Schwarz, the i.i.d. data, assumptions 2.5(ii) and 3.6(i), and Lemma A.1(2), we have:

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

thus $\sup_{\alpha\in\mathcal{N}_{0n}}|I_n(\alpha)| \leq \epsilon_n^2 \times o_P(n^{-1/4})$ by assumption 3.3(i). Next, by assumption 2.5(ii), we have: uniformly over $\alpha \in \mathcal{N}_{0n}$,

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

where the second inequality follows from the definition of $\widetilde{m}$, the i.i.d. data and assumptions 3.6(ii) and 2.10(ii). Therefore, we have

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

Since $\epsilon_n = o(n^{-1/2})$ and $u_n^* = \pm v_n^*$ we obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\frac{d\widetilde{m}(X_i,\widehat{\alpha}_n)}{d\alpha}[v_n^*]\right)' \widehat{\Sigma}(X_i)^{-1}\left(\widehat{m}(X_i,\alpha_0)+\widetilde{m}(X_i,\widehat{\alpha}_n)\right) = o_P(1). \tag{A.3}$$

Note that, by Cauchy-Schwarz, the i.i.d. data, assumption 2.5(ii) and the definition of $\widetilde{m}$, we have:

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

where the first term is of order $o_P(n^{-1/4})$ by assumptions 3.6(ii) and 2.10(ii) and i.i.d. data, and the second term is $o_P(n^{-1/4})$ by Lemma A.1(2) and assumption 3.3(i). Thus, we obtain:

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

Note that

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

where the first inequality is obtained by assumption 2.5, and the second inequality follows from i.i.d. data, Lemma A.1(2), assumptions 2.10(ii) and 3.3(ii). Thus

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

Notice that

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

where the last inequality is due to Cauchy-Schwarz, the i.i.d. data, Lemma A.1(2), Markov inequality, and assumption 3.2(iii). Thus,

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

This, Cauchy-Schwarz, the i.i.d. data, Lemma A.1(2), assumptions 2.5(iii) and 3.4(i) imply that

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

Recall that $g(X,v^*) \equiv \{\frac{dm(X,\alpha_0)}{d\alpha}[v^*]\}'\Sigma(X)^{-1}$ and $\widetilde{g}(X,v^*)$ is the LS projection of $g(X,v^*)$ onto $p^{J_n}(X)$. Then by the property of LS projection, we have:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i,v^*)\{\widehat{m}(X_i,\alpha_0)+\widetilde{m}(X_i,\widehat{\alpha}_n)\} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{g}(X_i,v^*)\{\rho(Z_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\}$$

$$= \frac{1}{n}\sum_{i=1}^{n} g(X_i,v^*)\{\rho(Z_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\} + o_P(\frac{1}{\sqrt{n}}),$$

where the second equality is due to the i.i.d. data, assumptions 2.5(iii), 3.6(i), 3.2(ii), 2.9(ii), 3.4(ii), $E[\rho(Z_i,\alpha_0)|X_1,...,X_n]=0$ and the Markov inequality. Thus we obtain:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i,v^*)\{\rho(Z_i,\alpha_0)+m(X_i,\widehat{\alpha}_n)\} = o_P(\frac{1}{\sqrt{n}}).$$

Notice that $|g(X,v^*)m(X,\alpha)-g(X,v^*)m(X,\alpha_0)| \leq |g(X,v^*)| \times |m(X,\alpha)-m(X,\alpha_0)|$. Given that $E[|g(X,v^*)|^2] < M$ by assumptions 2.10(ii) and 2.5(iii), 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 the entropy under the $L^\infty(\mathcal{X})$ norm of $\{m(X,\alpha) : \alpha \in \mathcal{N}_{0n}\}$, which is $\epsilon^{d_x/\gamma_m}$ hence a Donsker class by assumption 3.5(b). Therefore, either by assumption 3.5(a) or by assumption 3.5(b), we have: uniformly over $\alpha \in \mathcal{N}_{0n}$,

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

By applying the mean value theorem to $(m(X,\alpha)-m(X,\alpha_0))$ and assumption 3.6(iii), we obtain:

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

Thus, we finally obtain

$$\sqrt{n}\langle v^*, \widehat{\alpha}_n-\alpha_0\rangle = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v^*]\right)'\Sigma(X_i)^{-1}\rho(Z_i,\alpha_0) + o_P(1) \tag{A.4}$$

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

**Proof of Theorem 3.2**: We repeat the proofs of the consistency and the convergence rates of Lemma 2.3, except using $W\rho(Z,\alpha)$ instead of $\rho(Z,\alpha)$. Under assumption 3.7, we can show that the weighted bootstrap estimator, $\widehat{\alpha}_n^* = (\widehat{\theta}_n^*, \widehat{h}_n^*)$, is in $\mathcal{N}_{0n}$ with probability approaching one. We shall establish the limiting distribution in two steps.

STEP 1: We first derive the asymptotic normality of $\sqrt{n}(\widehat{\theta}_n^*-\theta_0)$ by mimicking the proof of Theorem 3.1. Under assumption 3.7, we can repeat the proof of Lemma A.1, and obtain: uniformly over $\alpha \in \mathcal{N}_{0n}$,

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\Sigma}(X_i)^{-\frac{1}{2}}\widehat{m}_W(X_i,\alpha)\right\|_E^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\Sigma}(X_i)^{-\frac{1}{2}}\ell_W(X_i,\alpha)\right\|_E^2 \pm o_P(n^{-1}),$$

where $\ell_W(X_i,\alpha) \equiv \widetilde{m}_W(X_i,\alpha)+\widehat{m}_W(X_i,\alpha_0)$. Moreover, by assumption 3.7, it follows $m_W(X,\alpha) \equiv E[W\rho(Z,\alpha)|X] = E[W]E[\rho(Z,\alpha)|X] = E[\rho(Z,\alpha)|X]$; this property also holds for the projection, $\widetilde{m}_W(X,\alpha) = \widetilde{m}(X,\alpha)$.

Recall that $\widehat{\alpha}_n^*$ solves $\min_{\alpha\in\mathcal{N}_{0n}}\{|||\widehat{m}_W(\cdot,\alpha)|||^2_{\widehat{\Sigma}}+\lambda_n\widehat{P}_n(h)\}$. Under assumption 3.3(iii), we can establish that $\widehat{\alpha}_n^*$ is an "approximate minimizer" of a smooth criterion function: $|||\ell_W(\cdot,\alpha)|||^2_{\widehat{\Sigma}}\equiv n^{-1}\sum_{i=1}^n\ell_W(X_i,\alpha)'[\widehat{\Sigma}(X_i)^{-1}]\ell_W(X_i,\alpha)$ for all $\alpha\in\{\alpha\in\mathcal{N}_{0n}:||\alpha-\widehat{\alpha}_n^*||=O_P(n^{-1/2})\}$. Now we can essentially repeat the proof of Theorem 3.1. Let $0<\epsilon_n=o(n^{-1/2})$ and $u_n^*=\pm v_n^*$. We have:

$$|||\ell_W(\cdot,\widehat{\alpha}_n^*)|||^2_{\widehat{\Sigma}}-|||\ell_W(\cdot,\widehat{\alpha}_n^*+\epsilon_n u_n^*)|||^2_{\widehat{\Sigma}}+o_P(n^{-1})\leq 0. \tag{A.5}$$

By second order Taylor expansion to equation (A.5), following the steps in the proof of Theorem 3.1 and using assumption 3.7, we obtain:

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

By assumption 3.6(ii), we have:

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

This and assumption 2.5(ii) imply that

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

By Markov inequality and i.i.d. data, we have:

$$n^{-1}\sum_{i=1}^n\left(\frac{d\widetilde{m}(X_i,\alpha_0)}{d\alpha}[v_n^*-v^*]\right)'\Sigma(X_i)^{-1}\left(\frac{d\widetilde{m}(X_i,\alpha_0)}{d\alpha}[v_n^*-v^*]\right)=O_P\left(||v_n^*-v^*||^2\right).$$

Therefore, following the steps in the proof of Theorem 3.1, we obtain

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

This, Cauchy-Schwarz, the i.i.d. data, Lemma A.1(2), assumptions 2.5(iii), 3.4(i) and 3.7 imply that

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

Recall that $g(X,v^*)\equiv\{\frac{dm(X,\alpha_0)}{d\alpha}[v^*]\}'\Sigma(X)^{-1}$ and $\widetilde{g}(X,v^*)$ is its LS projection onto $p^{J_n}(X)$. Then we have:

$$\frac{1}{n}\sum_{i=1}^n g(X_i,v^*)\left(\widehat{m}_W(X_i,\alpha_0)+\widetilde{m}(X_i,\widehat{\alpha}_n^*)\right)=\frac{1}{n}\sum_{i=1}^n\widetilde{g}(X_i,v^*)\{W_i\rho(Z_i,\alpha_0)+m(X_i,\widehat{\alpha}_n^*)\}.$$

Following the steps in the proof of Theorem 3.1 and using assumption 3.7, we obtain:

$$\frac{1}{n}\sum_{i=1}^n\{\widetilde{g}(X_i,v^*)-g(X_i,v^*)\}W_i\rho(Z_i,\alpha_0)=o_P(\frac{1}{\sqrt{n}}).$$

$$\frac{1}{n}\sum_{i=1}^{n}\{\widetilde{g}(X_i,v^*)-g(X_i,v^*)\}\{m(X_i,\widehat{\alpha}_n^*)-m(X_i,\alpha_0)\}=o_P(\frac{1}{\sqrt{n}}).$$

Thus we obtain:

$$\frac{1}{n}\sum_{i=1}^{n}g(X_i,v^*)\{W_i\rho(Z_i,\alpha_0)+m(X_i,\widehat{\alpha}_n^*)\}=o_P(\frac{1}{\sqrt{n}}).$$

Recall that $\{g(X,v^*)m(X,\alpha):\alpha\in\mathcal{N}_{0n}\}$ is a Donsker Class. Thus, we have uniformly over $\alpha\in\mathcal{N}_{0n}$,

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

Hence

$$\sqrt{n}\langle v^*,\widehat{\alpha}_n^*-\alpha_0\rangle=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v^*]\right)'\Sigma(X_i)^{-1}\rho(Z_i,\alpha_0)W_i+o_P(1). \tag{A.6}$$

This and assumption 3.7 implies that $\sqrt{n}(\widehat{\theta}_n^*-\theta_0)$ is asymptotically normal with zero mean and variance $V_*^{-1}\equiv\omega_0V^{-1}$.

Step 2: Subtracting equation (A.4) from (A.6), we obtain:

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

Given that $Var(W-1)=Var(W)=\omega_0$ and that $\{W_i\}_{i=1}^n$ is independent of $\{(Y_i,X_i)\}_{i=1}^n$, it follows that, conditional on the data $\{(Y_i,X_i)\}_{i=1}^n$, $\sqrt{\frac{n}{\omega_0}}\left(\widehat{\theta}_n^*-\widehat{\theta}_n\right)$ 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**: The proof essentially replicates that of theorem 6.1 in Ai and Chen (2003), except that we replace their use of the pathwise derivative of the generalized residual function $\rho(Z,\alpha)$ with respect to $\alpha$ by the pathwise derivative of the conditional mean function $E[\rho(Z,\alpha)|X]$ wrt $\alpha$ in a shrinking neighborhood of $\alpha_0$. See the working paper version (Chen and Pouzo (2008b)) for the detailed proof; also see the working paper version of Ai and Chen (2003) for an alternative proof via the empirical likelihood. *Q.E.D*

**Proof of Theorem 4.2**: With the danger of slightly abusing notation, we denote $\widehat{\Sigma}_0(X_i)\equiv\widehat{\Sigma}(X_i,\widehat{\alpha}_n)$. Then we have:

$$\begin{aligned}
\widetilde{\alpha}_n &\equiv (\tilde{\theta}_n,\widetilde{h}_n)=\arg\min_{\theta\in\Theta,h\in\mathcal{H}_{k(n)}}\{|||\widehat{m}(\cdot,\theta,h)|||^2_{\widehat{\Sigma}_0}+\lambda_n\widehat{P}_n(h)\},\quad \widehat{Q}_n(\tilde{\theta}_n)\equiv\frac{1}{2}|||\widehat{m}(\cdot,\widetilde{\alpha}_n)|||^2_{\widehat{\Sigma}_0},\\
\widetilde{\alpha}_n^0 &\equiv (\theta_0,\widetilde{h}_n^0)\equiv\arg\min_{h\in\mathcal{H}_{k(n)}}\{|||\widehat{m}(\cdot,\theta_0,h)|||^2_{\widehat{\Sigma}_0}+\lambda_n\widehat{P}_n(h)\},\quad \widehat{Q}_n(\theta_0)\equiv\frac{1}{2}|||\widehat{m}(\cdot,\widetilde{\alpha}_n^0)|||^2_{\widehat{\Sigma}_0}.
\end{aligned}$$

We shall establish $2n[\widehat{Q}_n(\theta_0)-\widehat{Q}_n(\tilde{\theta}_n)]\Rightarrow\chi^2_{d_\theta}$ by first showing $n\left(|||\ell(\cdot,\widetilde{\alpha}_n^0)|||^2_{\widehat{\Sigma}_0}-|||\ell(\cdot,\widetilde{\alpha}_n)|||^2_{\widehat{\Sigma}_0}\right)\Rightarrow\chi^2_{d_\theta}$ in several steps.

Step 1: Recall that $\widetilde{\alpha}_n\equiv(\tilde{\theta}_n,\widetilde{h}_n)$ is the unconstrained PSMD estimator. Let $\alpha_n^*\equiv\widetilde{\alpha}_n-\langle\widetilde{\alpha}_n-\alpha_0,v_0\rangle v_n^0/||v_0||^2$, where the inner product $\langle\cdot,\cdot\rangle$ is defined using the $\Sigma_0(X)$ instead of $\Sigma(X)$ and $||v_0||^2=\lambda'V_0^{-1}\lambda$. Then $\widetilde{\alpha}_n-\alpha_n^*=\langle\widetilde{\alpha}_n-\alpha_0,v_0\rangle v_n^0/||v_0||^2$. Recall that for any $\lambda\neq0$,

$\lambda'(\tilde{\theta}_n - \theta_0) = \langle \widetilde{\alpha}_n - \alpha_0, v_0 \rangle$. Applying Theorem 3.1 we have: $\sqrt{n}\langle \widetilde{\alpha}_n - \alpha_0, v_0 \rangle \Rightarrow N(0, ||v_0||^2)$. Thus we have $||\widetilde{\alpha}_n - \alpha_n^*||^2 = O_P(n^{-1/2})$. Applying Taylor expansion up to second order, we have:

$$\begin{aligned} & |||\ell(\cdot, \widetilde{\alpha}_n)|||^2_{\widehat{\Sigma}_0} - |||\ell(\cdot, \alpha_n^*)|||^2_{\widehat{\Sigma}_0} \\ = & \frac{2}{n}\sum_{i=1}^n \left( \frac{d\widetilde{m}(X_i, \alpha_n^*)}{d\alpha}[\widetilde{\alpha}_n - \alpha_n^*] \right)' \widehat{\Sigma}_0(X_i)^{-1} \left( \widehat{m}(X_i, \alpha_0) + \widetilde{m}(X_i, \alpha_n^*) \right) + I_n(\overline{\alpha}_n) + II_n(\overline{\alpha}_n) \end{aligned}$$

with $\overline{\alpha}_n \in \mathcal{N}_{0n}$ a point in between $\widetilde{\alpha}_n$ and $\alpha_n^*$, and

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

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

Following the same calculations as those in the proof of Theorem 3.1 and by assumption 4.2(ii), we have: $\sup_{\overline{\alpha}_n \in \mathcal{N}_{0n}} |I_n(\overline{\alpha}_n)| = o_P(n^{-1})$. Similarly under assumption 4.2(i)(ii), we have:

$$II_n(\overline{\alpha}_n) = \left( \frac{\langle \widetilde{\alpha}_n - \alpha_0, v_0 \rangle}{||v_0||^2} \right)^2 E\left[ \{ \frac{dm(X, \alpha_0)}{d\alpha}[v_0] \}' \Sigma_0(X)^{-1} \{ \frac{dm(X, \alpha_0)}{d\alpha}[v_0] \} \right] + o_P(n^{-1}).$$

By Cauchy-Schwarz inequality, $\langle v_0, \widetilde{\alpha}_n - \alpha_0 \rangle = O_P(n^{-1/2})$, assumption 4.2, and using the same arguments as the ones in the proof of Theorem 3.1, we obtain:

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n \left( \frac{d\widetilde{m}(X_i, \alpha_n^*)}{d\alpha}[\widetilde{\alpha}_n - \alpha_n^*] \right)' \widehat{\Sigma}_0(X_i)^{-1} \left( \widehat{m}(X_i, \alpha_0) + \widetilde{m}(X_i, \alpha_n^*) \right) \\ = & n^{-1}\sum_{i=1}^n \left( \frac{dm(X_i, \alpha_0)}{d\alpha}[\widetilde{\alpha}_n - \alpha_n^*] \right)' \Sigma_0(X_i)^{-1} \left( \widehat{m}(X_i, \alpha_0) + m(X_i, \alpha_n^*) \right) + o_P(n^{-1}). \end{aligned}$$

Since $\alpha_n^* - \alpha_0 = \widetilde{\alpha}_n - \alpha_0 - \langle \widetilde{\alpha}_n - \alpha_0, v_0 \rangle v_n^0 / ||v_0||^2$, applying second order Taylor expansion to $m(X_i, \alpha_n^*) - m(X_i, \alpha_0)$, we obtain:

$$\begin{aligned} & n^{-1}\sum_{i=1}^n \left( \frac{dm(X_i, \alpha_0)}{d\alpha}[\widetilde{\alpha}_n - \alpha_n^*] \right)' \Sigma_0(X_i)^{-1} m(X_i, \alpha_n^*) \\ = & O_P(n^{-1/2}) \times \left( \langle \widetilde{\alpha}_n - \alpha_0, v_0 \rangle - \frac{\langle \widetilde{\alpha}_n - \alpha_0, v_0 \rangle \langle v_n^0, v_0 \rangle}{||v_0||^2} + o_P(n^{-1/2}) \right) \\ = & O_P(n^{-1/2}) \times \left( \langle \widetilde{\alpha}_n - \alpha_0, v_0 \rangle - \langle \widetilde{\alpha}_n - \alpha_0, v_0 \rangle + o_P(n^{-1/2}) \right) = o_P(n^{-1}), \end{aligned}$$

where the last equality uses the fact that $\langle v_n^0 - v_0, v_0 \rangle \leq ||v_n^0 - v_0||^2 = o_P(1)$ by assumption 3.2(i). Therefore

$$\begin{aligned} & |||\ell(\cdot, \widetilde{\alpha}_n)|||^2_{\widehat{\Sigma}_0} - |||\ell(\cdot, \alpha_n^*)|||^2_{\widehat{\Sigma}_0} \\ = & \frac{2}{n}\sum_{i=1}^n \left( \frac{dm(X_i, \alpha_0)}{d\alpha}[\widetilde{\alpha}_n - \alpha_n^*] \right)' \Sigma_0(X_i)^{-1} \widehat{m}(X_i, \alpha_0) + \left( \frac{\langle \widetilde{\alpha}_n - \alpha_0, v_0 \rangle}{||v_0||^2} \right)^2 ||v_0||^2 + o_P(\frac{1}{n}) \\ = & -\frac{\langle \widetilde{\alpha}_n - \alpha_0, v_0 \rangle^2}{||v_0||^2} + o_P(n^{-1}). \end{aligned}$$

STEP 2: Recall that $\widetilde{\alpha}_n^0 \equiv (\theta_0, \widetilde{h}_n^0)$ is the constrained PSMD estimator. Define $\alpha_n^{*0} \equiv \widetilde{\alpha}_n^0 + \langle \widetilde{\alpha}_n - \alpha_0, v_0\rangle v_n^0/||v_0||^2$. Note that $\widetilde{\alpha}_n^0 - \alpha_n^{*0} = -(\widetilde{\alpha}_n - \alpha_n^*) = -\langle \widetilde{\alpha}_n - \alpha_0, v_0\rangle v_n^0/||v_0||^2$. Following the same calculations as those in Step 1, we obtain:

$$|||\ell(\cdot, \widetilde{\alpha}_n^0)|||_{\widehat{\Sigma}_0}^2 - |||\ell(\cdot, \alpha_n^{*0})|||_{\widehat{\Sigma}_0}^2 = \langle \widetilde{\alpha}_n - \alpha_0, v_0\rangle^2 \frac{1}{||v_0||^2} + o_P(n^{-1}).$$

STEP 3: Applying Lemma A.1(3), we obtain:

$$|||\ell(\cdot, \widetilde{\alpha}_n)|||_{\widehat{\Sigma}_0}^2 = |||\widehat{m}(\cdot, \widetilde{\alpha}_n)|||_{\widehat{\Sigma}_0}^2 + o_P(\frac{1}{n}),\ |||\ell(\cdot, \alpha_n^{*0})|||_{\widehat{\Sigma}_0}^2 = |||\widehat{m}(\cdot, \alpha_n^{*0})|||_{\widehat{\Sigma}_0}^2 + o_P(\frac{1}{n}). \tag{A.7}$$

By the definitions of $\widetilde{\alpha}_n$, $\alpha_n^{*0}$, and assumption 4.2(iii), we have: $|||\widehat{m}(\cdot, \widetilde{\alpha}_n)|||_{\widehat{\Sigma}_0}^2 \leq |||\widehat{m}(\cdot, \alpha_n^{*0})|||_{\widehat{\Sigma}_0}^2 + o_P(n^{-1})$. This, equation (A.7), and Step 2 imply that

$$|||\ell(\cdot,\widetilde{\alpha}_n^0)|||_{\widehat{\Sigma}_0}^2 - |||\ell(\cdot,\widetilde{\alpha}_n)|||_{\widehat{\Sigma}_0}^2 \geq |||\ell(\cdot,\widetilde{\alpha}_n^0)|||_{\widehat{\Sigma}_0}^2 - |||\ell(\cdot,\alpha_n^{*0})|||_{\widehat{\Sigma}_0}^2 = \frac{\langle \widetilde{\alpha}_n - \alpha_0, v_0\rangle^2}{||v_0||^2} + o_P(\frac{1}{n}) \tag{A.8}$$

STEP 4: Denote $\alpha_n^*(t) \equiv \alpha_n^* + t v_n^0/||v_0||^2$ for a scalar $t \geq 0$. We need to find a point, denoted as $\alpha_n^*(t^*) \equiv \alpha_n^* + t^* v_n^0/||v_0||^2$, that satisfies (a) $\langle \alpha_n^*(t^*), v_0\rangle = \theta_0$ (the constraint), and (b) $|||\ell(\cdot, \alpha_n^*(t^*))|||_{\widehat{\Sigma}_0}^2 - |||\ell(\cdot, \alpha_n^*)|||_{\widehat{\Sigma}_0}^2 = o(n^{-1})$. Suppose such an $\alpha_n^*(t^*)$ exists, then by the definition of $\widetilde{\alpha}_n^0 \equiv (\theta_0, \widetilde{h}_n^0)$, and by Step 1, we obtain:

$$\begin{aligned}
& |||\ell(\cdot, \widetilde{\alpha}_n)|||_{\widehat{\Sigma}_0}^2 - |||\ell(\cdot, \widetilde{\alpha}_n^0)|||_{\widehat{\Sigma}_0}^2 \\
\geq\ & |||\ell(\cdot, \widetilde{\alpha}_n)|||_{\widehat{\Sigma}_0}^2 - |||\ell(\cdot, \alpha_n^*(t^*))|||_{\widehat{\Sigma}_0}^2 = |||\ell(\cdot, \widetilde{\alpha}_n)|||_{\widehat{\Sigma}_0}^2 - |||\ell(\cdot, \alpha_n^*)|||_{\widehat{\Sigma}_0}^2 - o_P(\frac{1}{n}) \\
=\ & -\frac{\langle \widetilde{\alpha}_n - \alpha_0, v_0\rangle^2}{||v_0||^2} + o_P(n^{-1}).
\end{aligned} \tag{A.9}$$

We now show such an $\alpha_n^*(t^*)$ exists. For (a), we want to find a $t^*$ that solves the following equation:

$$0 = \langle \alpha_n^*(t) - \alpha_0, v_0\rangle = \langle \alpha_n^* - \alpha_0, v_0\rangle + t\frac{\langle v_n^0, v_0\rangle}{\langle v_0, v_0\rangle} = \langle \alpha_n^* - \alpha_0, v_0\rangle + t + t\frac{\langle v_n^0 - v_0, v_0\rangle}{\langle v_0, v_0\rangle}.$$

Notice that $\langle \alpha_n^* - \alpha_0, v_0\rangle = \langle \widetilde{\alpha}_n - \alpha_0, v_0\rangle \times \left(-\frac{\langle v_n^0 - v_0, v_0\rangle}{\langle v_0, v_0\rangle}\right) = \langle \alpha_n^* - \alpha^*, v_0\rangle$, since the second term in the middle is $o_p(n^{-1/2})$, it is easy to see that there is a $t^*$ that solves the above equation and such a $t^*$ is of order $o(n^{-1/2})$. For (b), notice that we can approximate $|||\ell(\cdot, \alpha_n^*(t^*))|||_{\widehat{\Sigma}_0}^2$ by $|||\ell(\cdot, \alpha_n^*(0))|||_{\widehat{\Sigma}_0}^2 + \frac{d|||\ell(\cdot, \alpha_n^*(0))|||_{\widehat{\Sigma}_0}^2}{d\alpha}[t^* v_n^0/||v_0||^2] + o(n^{-1})$ (where the last term depends on $t^* = o(n^{-1/2})$ and the second term is also of order $o(n^{-1})$ by following similar calculations as those in Step 1). Thus, (b) holds

STEP 5: Invoking the inequalities (A.8) and (A.9), we obtain

$$n\left(|||\ell(\cdot, \widetilde{\alpha}_n^0)|||_{\widehat{\Sigma}_0}^2 - |||\ell(\cdot, \widetilde{\alpha}_n)|||_{\widehat{\Sigma}_0}^2\right) = \left(\frac{\sqrt{n}\langle \widetilde{\alpha}_n - \alpha_0, v_0\rangle}{||v_0||}\right)^2 + o_P(1) \Rightarrow \chi_{d_\theta}^2$$

where the right hand side chi-square limiting distribution follows from $\sqrt{n}\langle \widetilde{\alpha}_n - \alpha_0, v_0\rangle \Rightarrow N(0, ||v_0||^2)$. Applying Lemma A.1(3), we obtain:

$$\begin{aligned}
2n[\widehat{Q}_n(\theta_0) - \widehat{Q}_n(\tilde{\theta}_n)] &= n\{|||\widehat{m}(\cdot, \widetilde{\alpha}_n^0)|||_{\widehat{\Sigma}_0}^2 - |||\widehat{m}(\cdot, \widetilde{\alpha}_n)|||_{\widehat{\Sigma}_0}^2\} \\
&= n\{|||\ell(\cdot, \widetilde{\alpha}_n^0)|||_{\widehat{\Sigma}_0}^2 - |||\ell(\cdot, \widetilde{\alpha}_n)|||_{\widehat{\Sigma}_0}^2 \pm o_P(\frac{1}{n})\} \Rightarrow \chi_{d_\theta}^2,
\end{aligned}$$

and the conclusion follows. *Q.E.D*

**Proof of Proposition 5.1**: For this model, we have: $\rho(Z,\alpha) = 1\{Y_3 \leq \theta Y_1 + h(Y_2)\} - \gamma$, $m(X,\alpha) = E[F_{Y_3|Y_1,Y_2,X}(\theta Y_1 + h(Y_2))|X] - \gamma$ and $\widehat{\Sigma} = \Sigma = \gamma(1-\gamma)$. For Result (1), it is easy to show that the i.i.d. data, assumption 2.7, and conditions 5.1 - 5.5 imply that all the assumptions of Lemma 2.3 hold. In particular, for assumption 2.1(iii) (identification), suppose that there is $\alpha \equiv (\theta, h)$ satisfying $|\theta - \theta_0| + ||h - h_0||_{L^2(f_{Y_2})} > 0$ and $E[F_{Y_3|Y_1Y_2X}(\alpha)|X] = \gamma$, then by the mean value theorem, there exists $\overline{\alpha} = (\overline{\theta}, \overline{h})$ such that $E[f_{Y_3|Y_1Y_2X}(\overline{\theta}Y_1 + \overline{h}(Y_2))\{Y_1(\theta - \theta_0) + h - h_0\}|X] = 0$, this and condition 5.3 then imply that $|\theta - \theta_0| + ||h - h_0||_{L^2(f_{Y_2})} = 0$; hence a contradiction and assumption 2.1(iii) holds. Condition 5.2(i) implies assumption 2.10(i); and conditions 5.3(ii)(iii) and 5.2(i) imply assumption 2.10(ii). The verifications of the rest of the assumptions of Lemma 2.3 are essentially the same as those in the proof of Proposition 6.4 in Chen and Pouzo (2008a); hence we omit them.

For Result (2), we shall verify that all the assumptions of Theorem 3.1 hold with $\widehat{\Sigma} = \Sigma = \gamma(1-\gamma)$. Condition 5.1(i)(ii) implies that assumption 3.1(i) holds with $\kappa = 1/2$. Since $\rho(Z,\alpha) \in [0,1]$ assumption 3.1(ii) trivially holds. Assumption 3.1(iii) follows from Result (1) and condition 5.7(i). Assumption 3.2(ii) follows from the fact that $\widehat{\Sigma} = \Sigma_0 = \gamma(1-\gamma)$. Regarding assumption 3.2(iii), since $v^* = v_0 \equiv (v_\theta^0, -w_0 v_\theta^0)$, with $v_\theta^0 = V_0^{-1}$ and $v_n^* = v_n^0 \equiv (v_\theta^0, -w_n^0 v_\theta^0)$, by condition 5.1(i), we have:

$$||v_n^* - v^*||^2 = ||w_n^0 - w_0||^2 = \frac{E\left[\left(E\left\{f_{Y_3|Y_1,Y_2,X}(\theta_0 Y_1 + h_0(Y_2))[w_n^0(Y_2) - w_0(Y_2)]|X\right\}\right)^2\right]}{\gamma(1-\gamma)};$$

thus assumption 3.2(iii) follows from condition 5.6(i) and Result (1). Assumption 3.3(i) follows from Result (1) and $\varsigma + a > d/2$ (which is implied by condition 5.7(i)). Assumption 3.3(iii) is implied by condition 5.6(ii). Since $\Sigma = \gamma(1-\gamma)$, and

$$\frac{dm(X,\alpha_0)}{d\alpha}[v^*] = E\left(f_{U|Y_1,Y_2,X}(0)[v_0]|X\right), \ \frac{d\widetilde{m}(X,\alpha_0)}{d\alpha}[v^*] = p^{J_n}(X)'(P'P)^- \sum_{i=1}^n p^{J_n}(X_i)\frac{dm(X_i,\alpha_0)}{d\alpha}[v_0],$$

assumption 3.4 follows from Result (1), assumption 2.7 and conditions 5.1 and 5.7(i). Assumption 3.5(b) follows directly from condition 5.1(iii). Assumption 3.6(i) directly follows from condition 5.7(ii). Also, under condition 5.7(ii), with $v_n^* = v_n^0 \equiv (v_\theta^0, -w_n^0 \times v_\theta^0)$, we have:

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

where $\overline{\theta}Y_1 + \overline{h}(Y_2)$ is in between $\theta Y_1 + h(Y_2)$ and $\theta_0 Y_1 + h_0(Y_2)$. By condition 5.7(ii),

$$E\left[\sup_{\alpha \in \mathcal{N}_{0n}} \left\| \frac{dm(X,\alpha)}{d\alpha}[v_n^*] - \frac{dm(X,\alpha_0)}{d\alpha}[v_n^*] \right\|_E^2\right] \leq const. \times ||\alpha - \alpha_0||_s^2 = O\left(n^{-\frac{2\varsigma}{2(\varsigma+a)+d}}\right),$$

thus assumption 3.6(ii) is satisfied given condition 5.7(i). Similarly, assumption 3.6(iii) follows from condition 5.7(i)(ii). Thus all the assumptions of theorem 3.1 hold, and we obtain: $\sqrt{n}(\widehat{\theta}_n - \theta_0) \Rightarrow$

$N\left(0, V_0^{-1}\right)$. Since $\widehat{\Sigma} = \Sigma = \Sigma_0 = \gamma(1-\gamma)$, the chi-square limiting distribution follows directly from theorem 4.2. *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.

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., 2008a. Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Moments. Cowles Foundation Discussion Paper 1650.

Chen, X., Pouzo, D., 2008b. Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals. Cowles Foundation Discussion Paper 1640 or Cemmap working paper CWP09/08.

Chen, X., Reiss, M., 2007. On Rate Optimality for Nonparametric Ill-posed Inverse Problems in Econometrics. Cowles Foundation Discussion Paper 1626.

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.

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., 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., 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., 1995a. 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 Tables and Figures

| $\gamma$ | 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: Monte Carlo study of a partially linear quantile IV example.

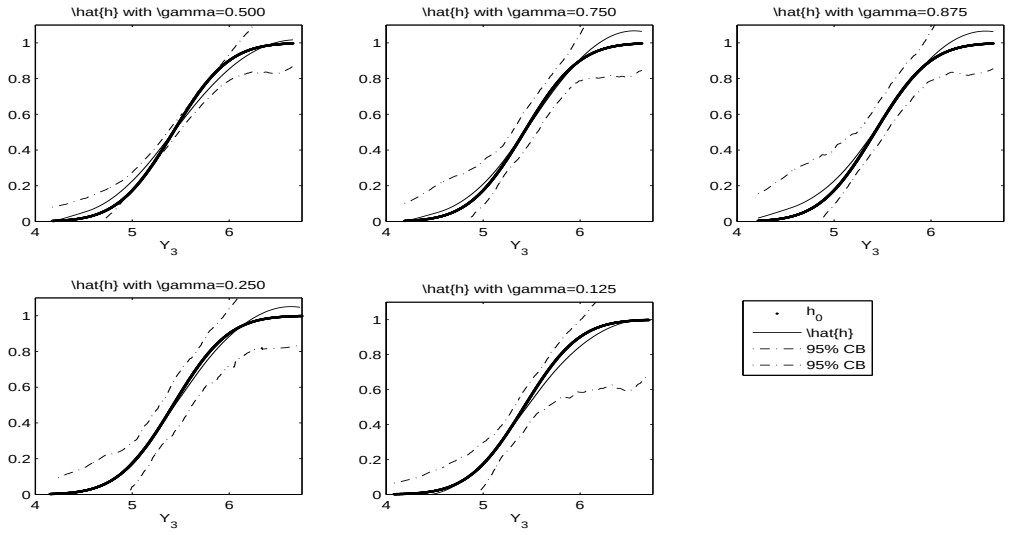


Figure 1: Monte Carlo study: Estimate of $h$.

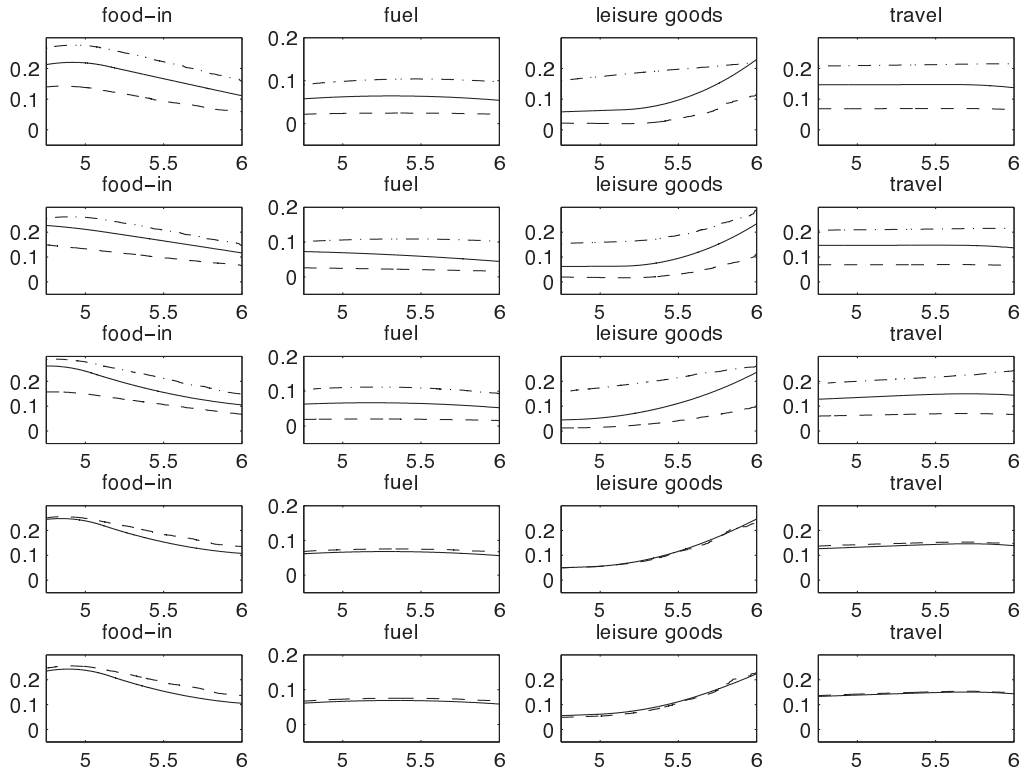


Figure 2: Estimated Engel curves for quantiles $\gamma = 0.25$ (dash), 0.5 (solid), 0.75 (dot-dash). 1st row: $\hat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\hat{\mu})}$, $\lambda_n = 0.001$; 2nd row: $\hat{P}_n(h) = ||\nabla^2 h||_{L^1(d\hat{\mu})}$, $\lambda_n = 0.001$; 3rd row: $\hat{P}_n(h) = ||\nabla h||^2_{L^2(d\hat{\mu})}$,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

| $\hat{P}_n(h)$ | $\|\|\nabla^2 h\|\|^2_{L^2(d\hat{\mu})}$ | $\|\|\nabla^2 h\|\|_{L^1(d\hat{\mu})}$ | $\|\|\nabla h\|\|^2_{L^2(d\hat{\mu})}$ | $\|\|\nabla^2 h\|\|^2_{L^2(d\hat{\mu})}$ | $\|\|\nabla h\|\|^2_{L^2(d\hat{\mu})}$ |
|---|---|---|---|---|---|
| $\lambda_n$ | 0.001 | 0.001 | 0.001 | 0.0003 | 0.0003 (BCK) |
| $\hat{\theta}_1$ | 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 2: Shape-invariant Engel curve quantile IV model with $\gamma = 0.5$: $\theta$ estimates under different penalization. The values in parenthesis are the mean IV estimates of Blundell et al. (2007).