ESTIMATION OF NONPARAMETRIC CONDITIONAL MOMENT MODELS WITH POSSIBLY NONSMOOTH MOMENTS

By

Xiaohong Chen and Demian Pouzo

April 2008
Revised October 2008

COWLES FOUNDATION DISCUSSION PAPER NO. 1650

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

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

Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Moments[1]

Xiaohong Chen[2] and Demian Pouzo[3]

First version: August 2005, Current version: April 2008

Abstract

This paper studies nonparametric estimation of conditional moment models in which the residual functions could be nonsmooth with respect to the unknown functions of endogenous variables. It is a problem of nonparametric nonlinear instrumental variables (IV) estimation, and a difficult nonlinear ill-posed inverse problem with an unknown operator. We first propose a penalized sieve minimum distance (SMD) estimator of the unknown functions that are identified via the conditional moment models. We then establish its consistency and convergence rate (in strong metric), allowing for possibly non-compact function parameter spaces, possibly non-compact finite or infinite dimensional sieves with flexible lower semicompact or convex penalty, or finite dimensional linear sieves without penalty. Under relatively low-level sufficient conditions, and for both mildly and severely ill-posed problems, we show that the convergence rates for the nonlinear ill-posed inverse problems coincide with the known minimax optimal rates for the nonparametric mean IV regression. We illustrate the theory by two important applications: root-n asymptotic normality of the plug-in penalized SMD estimator of a weighted average derivative of a nonparametric nonlinear IV regression, and the convergence rate of a nonparametric additive quantile IV regression. We also present a simulation study and an empirical estimation of a system of nonparametric quantile IV Engel curves.

KEYWORDS: Nonsmooth residuals, nonlinear ill-posed inverse, penalized sieve minimum distance, modulus of continuity, average derivative of a nonparametric nonlinear IV regression, nonparametric additive quantile IV regression.

JEL Classification: C13, C14, D12.

1 Introduction

Many semi/nonparametric structural 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 $Z \equiv (Y', X_z')'$, Y is a vector of endogenous (or dependent) variables, X_z is a subset of the conditioning (or instrumental) variables X, $\rho()$ is a vector of generalized residual functions whose functional forms are known up to the unknown vector of finite dimensional parameters (θ_0) and the

unknown functions ($h_0 \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))$), where each real-valued function $h_{0\ell}(\cdot), \ell = 1, ..., q$, may depend on different arguments. The conditional distribution, $F_{Y|X}$, of Y given X is not specified; hence the functional form of the conditional expectation, $E[\rho(Z, \theta_0, h_0)|X]$, of $\rho(Z, \theta_0, h_0)$ given X is unknown.

Assuming that the parameters of interest (θ_0, h_0) are identified by the general conditional moment models (1.1), Newey and Powell (hereafter NP, 2003) and Ai and Chen (hereafter AC, 2003) propose Sieve Minimum Distance (hereafter SMD) estimation of (θ_0, h_0). Under the assumptions that the residual function $\rho(Z, \theta, h(\cdot))$ is pointwise Hölder continuous in the parameters $(\theta, h) \in \Theta \times \mathcal{H}$, the parameter space $\Theta \times \mathcal{H}$ is compact, and the sieve parameter space $\Theta \times \mathcal{H}_n$ is finite dimensional compact, NP (2003) obtain consistency of the SMD estimator of (θ_0, h_0), and AC (2003) establish root-n asymptotic normality and efficiency of the SMD estimator of the finite dimensional parameters θ_0. However, neither paper studies the optimal rates of convergence for the SMD estimator of h_0.

When $h_0(\cdot)$ in the general framework (1.1) depends on the endogenous variables Y, it is difficult to establish point identification of h_0, consistency and convergence rate of any estimator of h_0 under the so-called "strong metric" $||\cdot||_s$, which is a metric that is not continuous with respect to the quadratic form $E\left[(E[\rho(Z, \theta, h(\cdot))|X])' (E[\rho(Z, \theta, h(\cdot))|X])\right]$, and the problem becomes a nasty nonlinear ill-posed inverse problem with an unknown operator.

There are some recent papers on identification and consistent estimation of a real-valued $h_0(Y)$ for two important special cases of (1.1). The first case is the nonparametric mean instrumental variables (NPIV) regression model:

$$E[Y_1 - h_0(Y_2))|X] = 0. \tag{1.2}$$

See NP (2003), Darolles, Florens and Renault (hereafter DFR, 2006), Blundell, Chen and Kristensen (hereafter BCK, 2007), Carrasco, Florens and Renault (2007), Severini and Tripathi (2006) and Florens, Johannes and van Bellegem (FJvB, 2007) for identification; NP (2003) for consistency, Hall and Horowitz (hereafter HH, 2005), DFR (2006), BCK (2007), Chen and Reiss (2007) and Gagliardini and Scaillet (GS, 2007) for convergence rates of their respective estimators of the NPIV model (1.2). The second important case is the nonparametric quantile instrumental variables (NPQIV) regression model:

$$E[1\{Y_1 \leq h_0(Y_2)\}|X] = \gamma \in (0, 1), \tag{1.3}$$

where $1\{\cdot\}$ denotes the indicator function. See Chernozhukov and Hansen (2005) and Chernozhukov, Imbens and Newey (hereafter CIN, 2007) for identification;[4] CIN (2007) for consistency, and Horowitz and Lee (hereafter HL, 2007) for consistency and convergence rate of their respective estimators of the NPQIV model (1.3). Recently, Chernozhukov, Gagliardini and Scaillet (CGS,

---

[4]See Chesher (2003) and Matzkin (2007) for additional identification results on nonsmooth nonseparable models.

2008) send us their unpublished manuscript about the convergence rate and pointwise limiting distribution of their penalized estimator for the NPQIV model.

To the best of our knowledge, except for the NPIV and the NPQIV models, there is no published work that establishes convergence rate of any estimator of $h_0 \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))$ for the general conditional moment models (1.1) when some of the $h_{0\ell}(\cdot), \ell = 1, ..., q$ depend on Y. Moreover, even for the NPIV and the NPQIV models, the above mentioned papers establish convergence rates for their respective estimators under different sets of regularity conditions that are very difficult to compare.

In this paper, we first propose a general class of penalized SMD estimators for $h_0 \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))$ satisfying the following nonparametric conditional moment models:[5]

$$E[\rho(Y, X_z; h_{01}(\cdot), ..., h_{0q}(\cdot))|X] = 0, \tag{1.4}$$

in which some of the $h_{0\ell}(\cdot), \ell = 1, ..., q$ depend on Y. Our penalized SMD procedure is very flexible. It allows for (i) nonlinear and possibly nonsmooth residual function $\rho()$; (ii) mildly ill-posed or severely ill-posed problems; (iii) possibly non-compact (under $||\cdot||_s$) infinite dimensional parameter space, (iv) possibly non-compact (under $||\cdot||_s$) finite or infinite dimensional sieve spaces, and (v) any "lower semicompact" (see Section 3 for its definition) or any convex penalization. Our penalized SMD procedure using finite dimensional linear sieves and "lower semicompact" or convex penalty is essentially the same as the original SMD procedure using finite dimensional compact sieves that has been previously studied in NP (2003), AC (2003), BCK (2007), CIN (2007) and Ai and Chen (2007), except that we establish consistency and convergence rates under the "strong metric" $||\cdot||_s$ without assuming the $||\cdot||_s$−compactness of the entire parameter space. This result is of great interest to those who like to implement the original SMD procedure using finite dimensional sieves, as they no longer need to worry about whether the entire parameter space is compact under $||\cdot||_s$. Our penalized SMD procedure using infinite dimensional linear sieves and "lower semicompact" or convex penalty extends the current Tikhonov regularization procedures of DFR (2006), HH (2005), GS (2007), FJvB (2007) and others for the NPIV model, and HL (2007) and CGS (2008) for the NPQIV model to allow for any model belonging to the class (1.1), and more flexible penalty functions (not restricted to Tikhonov regularization using square integrable norm of h or square integrable norm of first or higher order derivatives of h).

Secondly and more importantly, we establish consistency and convergence rates (in "strong metric" $||\cdot||_s$) of the penalized SMD estimator for $h_0()$ of the nonparametric conditional moment models (1.4), allowing for the above (i) - (v). Our large sample results are derived under any

---

[5]In Chen and Pouzo (2008), we obtain the semiparametric efficiency and the root-n asymptotic normality of the penalized SMD estimator $\widehat{\theta}_n$ of θ_0 for the general semiparametric conditional moment models (1.1) when $\rho(Z, \theta, h(\cdot))$ is not pointwise smooth in (θ, h). The results in Chen and Pouzo (2008) depend crucially on the consistency and convergence rates of the penalized SMD estimator $\widehat{h}_n$ of h_0, which are the main focuses of our this paper.

nonparametric consistent estimator of the conditional mean functions $E[\rho(Y, X_z; h)|X = \cdot]$. Some of the results allow for possibly non-uniqueness of $h_0()$ satisfying the general model (1.4), although we present sufficient conditions such that the model (1.4) and the penalty jointly identify $h_0()$. We show that for both mildly and severely ill-posed problems, the convergence rates are closely related to the notion of "modulus of continuity" (see Section 4 for its definition). More precisely, for the penalized SMD estimator using infinite dimensional sieves, the convergence rate is given by the "modulus of continuity". For the penalized SMD estimator using finite dimensional sieves, the convergence rate is determined by balancing the sieve approximation error rate of $h_0()$ and the "sieve modulus of continuity", which is a natural generalization of the "sieve measure of ill-posedness" introduced in BCK (2007) for the NPIV model (1.2). We also provide low-level sufficient conditions to bound the sieve modulus of continuity and the modulus of continuity. When we specialize our convergence rate results to the NPIV model (1.2), our rates coincide with the known minimax optimal rates derived in HH (2005) and Chen and Reiss (2007).[6] In addition, our rates for the general problems of nonlinear and nonsmooth residual functions ρ also coincide with the optimal ones for the linear ill-posed inverse problems.

Although we establish consistency and convergence rates for the penalized SMD estimator allowing for both finite dimensional sieves and infinite dimensional sieves, our sufficient conditions for the ones using finite dimensional sieves are slightly weaker than those for the penalized SMD estimators using infinite dimensional sieves. In addition, based on our simulation studies and those reported in BCK (2007) and Chen and Pouzo (2008), the penalized SMD estimator using a finite dimensional linear sieve and a flexible penalty is not only easy to compute but also performing well in finite samples. When h_0 enters the residual function $\rho()$ linearly such as in the NPIV model (1.2), the infinite dimensional Tikhonov regularized estimators can be computed in closed-forms, and their asymptotic properties are relatively easy to analyze; see, e.g., DFR (2006) and HH (2005). However, when h_0 enters the residual function $\rho()$ nonlinearly and non-smoothly, such as in the NPQIV model (1.3), the infinite dimensional regularized estimators are impossible to compute. In fact, in their simulation study of the NPQIV model (1.3), HL (2007) actually approximate the unknown function $h_0(\cdot)$ by a Fourier series with lots of terms; hence they could ignore the Fourier series approximation error, and view their implemented procedure as the one of infinite dimensional sieve Tikhonov regularization. Similarly, GS (2007) and CGS (2008) use finite many spline and polynomial series terms to approximate unknown h in their simulation and empirical implementations of their Tikhonov first derivative regularized estimators for the NPIV and the NPQIV models.

Finally, we illustrate the usefulness of the consistency and convergence rate (in strong metric

---

[6]The rates also coincide with those in Efromovich and Koltchinskii (2001) and Hoffmann and Reiss (2008) for statistical linear ill-posed inverse problems with unknown operators, and in Cavalier et al. (2002) and the references therein for known operators.

$||\cdot||_s$) results by deriving the root-n asymptotic normality of the plug-in penalized SMD estimator of a weighted average derivative of $h_0(Y)$ identified through the general model $E[\rho(Y, X_z, h_0(Y))|X] = 0$, in which $\rho()$ could be nonlinear and non-pointwise smooth in h_0. This result is very important in its own right, as the weighted average derivatives are widely used in testing various economic hypothesis of h_0 when $h_0(Y)$ may enter $\rho()$ nonlinearly; see, e.g., Chen and Ludvigson (2004). Previously, Ai and Chen (2007) establish root-n asymptotic normality of the plug-in SMD estimator of a weighted average derivative of h_0 for the NPIV model (1.2): $E[Y_1 - h_0(Y_2))|X] = 0$. Thanks to the linearity of the NPIV model in $h(Y)$, they obtain the normality result without requiring convergence rate of their SMD estimator $\widehat{h}_n$ under the strong norm ($||h||_s = \sqrt{E[\{h(Y_2)\}^2]}$ for the NPIV model). Here we extend their results to allow for possibly nonlinear and non-pointwise smooth $\rho()$ in $h(Y)$, without imposing $||\cdot||_s$−compactness of the entire function parameter space. Unfortunately, when $h(Y)$ enters $\rho()$ nonlinearly such as in the NPQIV model (1.3), in order to achieve root-n asymptotic normality of a plug-in estimate of a weighted average derivative of $h_0(Y)$, we now need certain convergence rate of our penalized SMD estimator $\widehat{h}_n$ under the strong norm $||\cdot||_s$.

The rest of the paper is organized as follows. Section 2 presents the penalized SMD procedures, a small Monte Carlo study of the NPQIV model (1.3), and an empirical illustration of the NPQIV estimation of system of Engel curves using British Family Expenditure Survey (FES) data. Section 3 establishes consistency, and discusses the implications of regularity conditions for the original SMD with finite dimensional sieve without explicit penalty, the penalized SMD with lower semicompact penalty, and the penalized SMD with general convex penalty. Section 4 derives convergence rates in terms of sieve modulus of continuity, and Section 5 presents sufficient conditions to bound sieve modulus of continuity. Section 6 provides two important applications of the general results. The first application obtains the consistency and convergence rate for the nonparametric additive quantile IV model: $E[1\{Y_3 \leq h_{01}(Y_1) + h_{02}(Y_2)\}|X] = \gamma \in (0, 1)$ where $h_0 = (h_{01}, h_{02})$. The second application establishes the root-n asymptotic normality of the plug-in penalized SMD estimator of a weighted average derivative of $h_0()$ for the general nonlinear model $E[\rho(Y, X_z, h_0(Y))|X] = 0$. Section 7 briefly concludes. Appendix A presents a brief review of some functional spaces and sieve bases, and the rest of the appendices contain the proofs.

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

5

that for all $\varepsilon > 0$, $\Pr(a_n/b_n \geq \varepsilon) \to 0$ as n goes to infinity. For any vector-valued x, we use $||x||_E$ denote its Euclidean norm (i.e., $||x||_E \equiv \sqrt{x'x}$, although sometimes we also use $|x| = ||x||_E$ without too much confusion).

2   Penalized SMD Estimators and Empirical Illustration

Suppose that the observations $\{(Y_i, X_i) : i = 1, 2, \ldots, n\}$ are drawn independently from the distribution of (Y, X) with support $\mathcal{Y} \times \mathcal{X}$, where $\mathcal{Y}$ is a subset of $\mathcal{R}^{d_y}$ and $\mathcal{X}$ is a compact subset of $\mathcal{R}^{d_x}$. Denote $Z \equiv (Y', X_z')' \in \mathcal{Z} \equiv \mathcal{Y} \times \mathcal{X}_z$ and $\mathcal{X}_z \subseteq \mathcal{X}$. Suppose that the unknown distribution of (Y, X) satisfies the conditional moment restriction given by (1.4), where $\rho : \mathcal{Z} \times \mathcal{H} \to \mathcal{R}^{d_\rho}$ is a known mapping, up to an unknown vector of parameters, $h_0 \in \mathcal{H}$. We assume that $\mathcal{H} \equiv \mathcal{H}^1 \times \cdots \times \mathcal{H}^q$, with each $\mathcal{H}^j, j = 1, ..., q$, being a space of real-valued measurable functions whose arguments vary across different applications.

Denote $m(x, h) \equiv \int \rho(y, x_z, h(\cdot)) dF_{Y|X=x}(y)$ as the conditional mean function of $\rho(Y, X_z, h(\cdot))$ given X. Under the assumption that model (1.4) identifies h_0, we have

$$h_0 = \arg \inf_{h \in \mathcal{H}} E\left[m(X, h)' m(X, h)\right]. \tag{2.1}$$

Since the functional forms of $F_{Y|X}$ and $m(X, h)$ are not specified, NP (2003) and AC (2003) propose to estimate h_0 by the SMD procedure:

$$\widehat{h}_n = \arg \inf_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \widehat{m}(X_i, h)' \widehat{m}(X_i, h), \tag{2.2}$$

where $\widehat{m}(X, h)$ is any nonparametric consistent estimator of $m(X, h)$, and $\mathcal{H}_n \equiv \mathcal{H}_n^1 \times \cdots \times \mathcal{H}_n^q$ is a *finite* dimensional sieve parameter space whose complexity grows with sample size and becomes dense in the original functional space $\mathcal{H} \equiv \mathcal{H}^1 \times \cdots \times \mathcal{H}^q$; see, e.g., Grenander (1981), Shen and Wong (1994), Van de Geer (2000) and Chen (2007).

Let $||\cdot||_c$ denote a metric on $\mathcal{H}$, which is a metric that may or may not be continuous with respect to the quadratic form $E[m(X, h)' m(X, h)]$ on $\mathcal{H}$. To obtain consistency of the SMD estimator $\widehat{h}_n$ under the metric $|| \cdot ||_c$, NP (2003), AC (2003), CIN (2007) and BCK (2007) assume that the original space $\mathcal{H}$ and the sieve spaces $\mathcal{H}_n$ are compact under the metric $|| \cdot ||_c$. Although the compact function space $\mathcal{H}$ is a reasonable assumption for some applications (see BCK (2007) for such an example), it would be nice to relax this assumption in other applications when the criterion function $E[m(X, h)' m(X, h)]$ is convex with respect to $h \in \mathcal{H}$.

2.1   Penalized SMD Estimators

In this paper we consider the following *penalized SMD*:

$$\widehat{h}_n = \arg \inf_{h \in \mathcal{H}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \widehat{m}(X_i, h)' \widehat{m}(X_i, h) + \lambda_n \widehat{P}_n(h) \right\}, \tag{2.3}$$

where the penalization parameter $\lambda_n \geq 0$ and $\lambda_n \to 0$ as $n \to \infty$, the penalization function $\widehat{P}_n(h)$ is typically a non-negative convex function in h given the data. Here the sieve space $\mathcal{H}_n \equiv \mathcal{H}_n^1 \times \cdots \times \mathcal{H}_n^q$ could be finite-dimensional, infinite-dimensional, compact or non-compact. When the original space $\mathcal{H}$ is infinite-dimensional compact (under $||\cdot||_c$), and the sieve space $\mathcal{H}_n$ is finite-dimensional or infinite-dimensional compact (under $||\cdot||_c$), then one can set $\lambda_n = 0$ and (2.3) reduces to (2.2). When the original space $\mathcal{H}$ is a closed (but not compact under $||\cdot||_c$), infinite-dimensional subset of a separable Banach space (under the metric $||\cdot||_c$), if one chooses the sieve space $\mathcal{H}_n$ to be some finite-dimensional compact sets, then one could still set $\lambda_n = 0$; however, if one set the sieve space $\mathcal{H}_n$ to be some infinite-dimensional non-compact sets, such as $\mathcal{H}_n = \mathcal{H}$, then one needs $\lambda_n \widehat{P}_n(h) > 0$. When $\mathcal{H}_n = \mathcal{H}$ and $\widehat{P}_n(h) = ||h - h^\star||_c^2$ with $h^\star \in \mathcal{H}$ being an initial guess, this procedure (2.3) becomes the minimum distance estimation with nonlinear Tikhonov regularization. For example, when $\mathcal{H}$ is the space of square integrable functions against a sigma-finite measure $d\mu$, $L^2(d\mu)$, we can let $||\cdot||_c$ be the $L^2(d\mu)-$norm, and $\widehat{P}_n(h) = ||h - h^\star||_{L^2(d\mu)}^2$ for a known measure $d\mu$, or $\widehat{P}_n(h) = ||h - h^\star||_{L^2(d\widehat{\mu})}^2$ for an empirical measure $d\widehat{\mu}$ when $d\mu$ is unknown. When $\mathcal{H}$ is a mixed weighted Sobolev space $\{h : ||h||_{L^2(d\mu)}^2 + ||\nabla^k h||_{L^2(leb)}^2 < \infty\}$, where $\nabla^k h$ is the k-th derivative of h for some integer $k \geq 1$, we can again let $||\cdot||_c$ be the $L^2(d\mu)-$norm and $\widehat{P}_n(h) = ||h - h^\star||_{L^2(d\mu)}^2 + ||\nabla^k h - \nabla^k h^\star||_{L^2(leb)}^2$. As we shall illustrate later on, one could also take $\widehat{P}_n(h) = ||\nabla^k h - \nabla^k h^\star||_{L^2(leb)}^2$ in some applications.

When $n^{-1} \sum_{i=1}^n \widehat{m}(X_i, h)' \widehat{m}(X_i, h)$ is convex in $h \in \mathcal{H}$ and $\mathcal{H}$ is a closed convex (but not compact under $||\cdot||_c$) space, it is computationally attractive to choose the penalization function $\widehat{P}_n(h)$ as a convex function, and to choose the sieve space $\mathcal{H}_n$ as a convex set, say $\mathcal{H}_n^c = \{h \in \mathcal{H} : \widehat{P}_n^1(h) \leq B_n\}$ with $B_n \to \infty$ slowly as $n \to \infty$. Then the penalized SMD procedure (2.3) is equivalent to compute

$$\inf_{h \in \mathcal{H}_n^c} n^{-1} \sum_{i=1}^n \widehat{m}(X_i, h)' \widehat{m}(X_i, h) + \lambda_n \widehat{P}_n(h). \tag{2.4}$$

Let $clsp(\mathcal{H}_n^c)$ denote the closed linear span of $\mathcal{H}_n^c$ under the metric $||\cdot||_c$. Then the optimization problem (2.4) is equivalent to

$$\inf_{h \in clsp(\mathcal{H}_n^c)} n^{-1} \sum_{i=1}^n \widehat{m}(X_i, h)' \widehat{m}(X_i, h) + \lambda_n \widehat{P}_n(h) + \lambda_{1n} \widehat{P}_n^1(h), \tag{2.5}$$

where λ_{1n} is chosen such that $\widehat{P}_n^1(\widehat{h}_n) = B_n$; see Eggermont and LaRiccia (2001). For most applications, it suffices to have either $\lambda_n > 0$ or $\lambda_{1n} > 0$.

Remark 2.1. *To compute the penalized SMD estimator $\widehat{h}_n$ for h_0, one could use any nonparametric estimator $\widehat{m}(X, h)$ for $m(X, h) \equiv E[\rho(X, h)|X]$, such as the series least squares (LS) estimator $\widehat{m}(X, h)$:*

$$\widehat{m}(X, h) = p^{J_n}(X)'(P'P)^- \sum_{i=1}^n p^{J_n}(X_i)\rho(Z_i, h), \tag{2.6}$$

where $\{p_j()\}_{j=1}^\infty$ is a sequence of known basis functions that can approximate any square integrable functions of X well, $J_n \to \infty$ slowly as $n \to \infty$, $p^{J_n}(X) = (p_1(X), ..., p_{J_n}(X))'$, $P =$

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

Remark 2.2. *The penalized SMD procedures introduced in this section can be trivially extended to estimate all the parameters of interest $\alpha_0 \equiv (\theta_0, h_0)$ of the general semi/nonparametric conditional moment models (1.1). For instance, we can extend the procedure (2.3) to*

$$\widehat{\alpha}_n = \arg \inf_{\alpha = (\theta, h) \in \Theta \times \mathcal{H}_n} \left\{ \frac{1}{n} \sum_{i=1}^n \widehat{m}(X_i, \alpha)'[\widehat{\Sigma}(X_i)]^{-1} \widehat{m}(X_i, \alpha) + \lambda_n \widehat{P}_n(h) \right\}, \qquad (2.7)$$

where $\alpha \equiv (\theta, h) \in \Theta \times \mathcal{H}$ and Θ is a compact subset of $\mathcal{R}^{d_\theta}$ with fixed $d_\theta < \infty$, $\widehat{m}(X, \alpha)$ is any nonparametric estimator of $m(X, \alpha) \equiv \int \rho(y, X_z, \alpha) dF_{Y|X}(y)$, and $\widehat{\Sigma}(X)$ is any nonparametric estimator of a positive definite weighting matrix $\Sigma(X)$ that is used for the purpose of semiparametric efficient estimation of $\theta_0 \in \Theta$. In Appendix B, we provide a general consistency theorem (Lemma B.1) that is also applicable to this penalized SMD estimator $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{h}_n)$. In the main text of the paper, however, we focus on nonparametric convergence rate of various penalized SMD estimators of the unknown functions h_0. To avoid tedious notations, we shall state properties of the penalized SMD estimators $\widehat{h}_n$ given in (2.3) only. See Chen and Pouzo (2008) for root-n asymptotic normality and semiparametric efficiency of smooth functionals of the penalized SMD estimator $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{h}_n)$ given in (2.7).

2.2   Monte Carlo Simulation

We report a small Monte Carlo (MC) study of penalized SMD estimation for the NPQIV model (1.3):

$$Y_1 = h_0(Y_2) + U, \quad \Pr(U \leq 0|X) = \gamma \in \{0.25, 0.5, 0.75\}.$$

The MC is designed to mimic the real data application in the next subsection as well as that in BCK (2007). First, we simulate $(Y_2, \widetilde{X})$ according to a bivariate Gaussian density whose mean and covariance are set to the ones estimated from the UK Family Expenditure Survey Engel curve data set (see BCK (2007) for more details). Second, we let $X = \Phi^{-1}\left(\frac{\widetilde{X} - \mu_x}{\sigma_x}\right)$ and $h_0(y_2) = \Phi\left(\frac{y_2 - \mu_2}{\sigma_2}\right)$ where Φ denotes the standard normal cdf, and the means μ_x, μ_2 and the variances σ_x, σ_2 are the estimated ones. Third, we generate Y_1 from $Y_1 = h_0(Y_2) + U$, where $U = C_2[V - \Phi^{-1}\left(\gamma + C_1\{E[h_0|\widetilde{X}] - h_0(Y_2)\}\right)]$, with $V \sim N(0, 1)$, $C_2 = \sqrt{0.075}$ and $C_1 = 0.01$. The number of

8

observation is set to $n = 500$. We have also tried to draw $(Y_2, \widetilde{X})$ from the kernel density estimator using the BCK data set, and to draw U from other distributions such as Pareto distribution. The simulation results are very similar to the ones reported here.

In this MC study and for the sake of concreteness, we estimate $h_0()$ using the penalized SMD estimator $\widehat{h}_n$ given in (2.3), with $\widehat{m}(X, h)$ being the series LS estimator (2.6) of $m(X, h)$, and $\mathcal{H}_n$ being a finite dimensional $(\dim(\mathcal{H}_n) \equiv k(n) < \infty)$ *linear* sieve. An example of a typical finite dimensional sieve of dimension $k(n)$ is a polynomial spline sieve, denoted as P-spline(q,r) with q being the order and r being the number of knots, then $k(n) = q(n) + r(n) + 1$. See Appendix A for other sieves such as wavelets and Hermite polynomials sieves.

There are three kinds of smoothing parameters in the penalized SMD procedure (2.3): one $(k(n))$ for the sieve approximation of $\mathcal{H}$ by $\mathcal{H}_n$, one (λ_n) for the penalization, and one (say J_n) for the nonparametric estimation $\widehat{m}(X, h)$. In the subsequent theoretical sections, we show that we could obtain optimal rate in either the "sieve dominating case" (the case of choosing $k(n) \asymp J_n$, $k(n) < J_n$ properly and letting $\lambda_n = 0$ or $\lambda_n \searrow 0$ fast), or the "sieve penalization balance case" (the case of choosing $k(n) \asymp J_n$, $k(n) \leq J_n$ and $\lambda_n \asymp \frac{J_n}{n}$ properly), or the "penalization dominating case" (the case of choosing $\lambda_n \geq \frac{J_n}{n}$ properly and letting $k(n) = \infty$ or $k(n) >> J_n$ and $k(n) \nearrow \infty$ fast). In this MC study, we compare the finite sample performance of the "sieve dominating case" and the "sieve penalization balance case".

Figure 2.1 summarizes the results for three quantiles $\gamma \in \{0.25, 0.5, 0.75\}$, each with 500 Monte Carlo repetitions. The first row corresponds to the "sieve dominating case" and the second row the "sieve penalization balance case". To compute the estimator $\widehat{h}$, we use P-Spline(2,5) (hence $k(n) = 8$) for $\mathcal{H}_n$ and $\lambda_n = 0.003$ in the "sieve dominating case", and P-Spline(5,10) (hence $k(n) = 16$) for $\mathcal{H}_n$ and $\lambda_n = 0.006$ in the "sieve penalization balance case", and in both cases, we use P-Spline(5,10) (hence $J_n = 16$) for $\hat{m}$ and $\widehat{P}_n(h) = ||\nabla h||^2_{L^2(leb)}$. We have also tried other sieve bases such as Hermite polynomials for $\widehat{h}$, Fourier basis, B-spline basis and Hermite basis for $\hat{m}$, and L^1 norm (against *leb.* or against empirical measure $d\widehat{\mu}$) of first or second derivative penalty $\widehat{P}_n(h)$. As long as the choice of $k(n)$, λ_n and J_n are similar to the ones reported here, the simulation results are similar; hence we do not report them due to the lack of space. In Figure 2.1, each panel shows the true function (solid thick line), the corresponding estimator (solid thin), the Monte Carlo 95% confidence bands, and a sample realization of Y_1 (that is arbitrarily picked from the last MC iteration). Both estimators perform very well for all the quantiles, with the "sieve dominating case" estimator performing slightly better. Nevertheless, we note that it is much faster to compute the "sieve dominating case" procedure. For example, using a AMD Athlon 64 processor with 2.41 GHz and 384 MB of RAM, the MC experiment written in FORTRAN took (approximately) 105 minutes to finish for the "sieve dominating case", whereas it took (approximately) 390 minutes to finish for the "sieve penalization balance or penalization dominating case".

9

Table 2.1 shows the integrated square bias $(I-BIAS^2)$, the integrated variance $(I-VAR)$ and the integrated mean square error $(I-MSE)$, which are computed using numerical integration over a grid ranging from 2.5% and 97.5%. Figure 2.2 shows corresponding estimated curves and MC confidence bands. Here for simplicity we have only reported the estimated quantile with $\gamma = 0.5$ and 250 MC replications. The first three rows belong to the "sieve dominating case"; the rest of the rows deal with "sieve penalization balance or penalization dominating cases". For this MC study, the "sieve dominating cases" (the first three rows) perform well in terms of $I-BIAS^2$ and $I-VAR$ (hence $I-MSE$), and are much more economical in terms of computational time. Secondly, within the latter two cases derivative-based penalization perform better than "level-based" penalization.
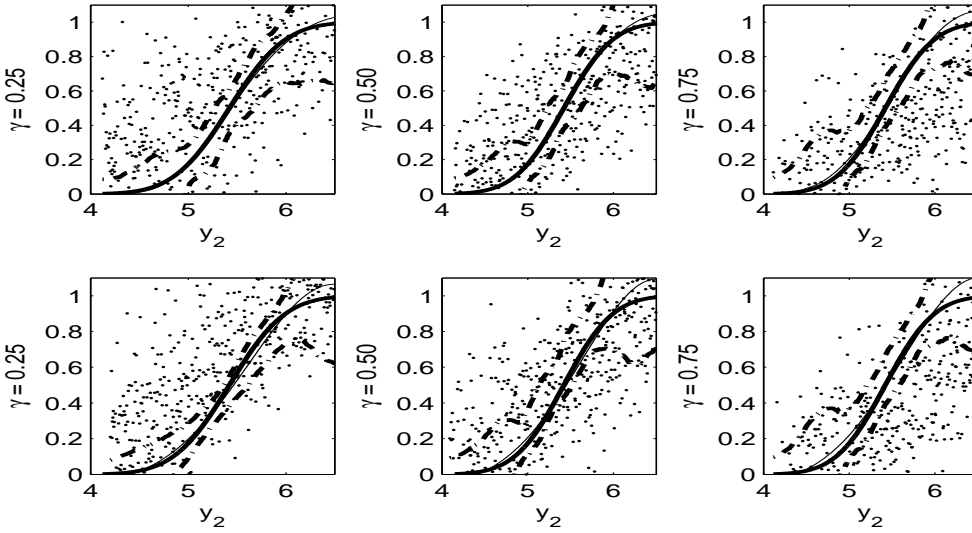

Figure 2.1: h_0 (solid thick line), $\widehat{h}_n$ (solid thin), MC 95% confidence bands (dashed), a sample of Y_1 (dots), sieve dominating (1st row), sieve penalization balance (2nd row)

Table 2.1: MC simulation Results

$(k(n), J_n)$	$I-BIAS^2$	$I-VAR$	$I-MSE$	Pen	λ_n	$time$ (in min.)
$(6, 16)$	0.00259	0.00349	0.00609	$\|\cdot\|^2_{L^2}$	0.00001	23
$(6, 16)$	0.00256	0.00423	0.00680	$\|\nabla^2 \cdot\|_{L^1}$	0.00001	25
$(6, 16)$	0.00272	0.00401	0.00674	$\|\nabla^2 \cdot\|^2_{L^2}$	0.00001	25
$(8, 16)$	0.00108	0.02626	0.02731	$\|\cdot\|^2_{L^2}$	0.00010	43
$(8, 16)$	0.00131	0.01820	0.01954	$\|\nabla^2 \cdot\|_{L^1}$	0.00010	48
$(8, 16)$	0.00030	0.01853	0.01855	$\|\nabla^2 \cdot\|^2_{L^2}$	0.00010	40
$(16, 16)$	0.00170	0.05464	0.05631	$\|\cdot\|^2_{L^2}$	0.00050	82
$(16, 16)$	0.00015	0.03704	0.03714	$\|\nabla^2 \cdot\|^2_{L^2}$	0.00050	84
$(16, 31)$	0.00011	0.02801	0.02813	$\|\nabla^2 \cdot\|^2_{L^2}$	0.00100	235

Finally, we need to point out that our theoretical results in the subsequent sections allow for $k(n) = \infty$ and/or $k(n) >> J_n$, and $\widehat{P}_n(h) = \|h\|^2_{L^2(leb)}$ or $\|h\|^2_{L^2(d\widehat{\mu})}$ in the "penalization dominating
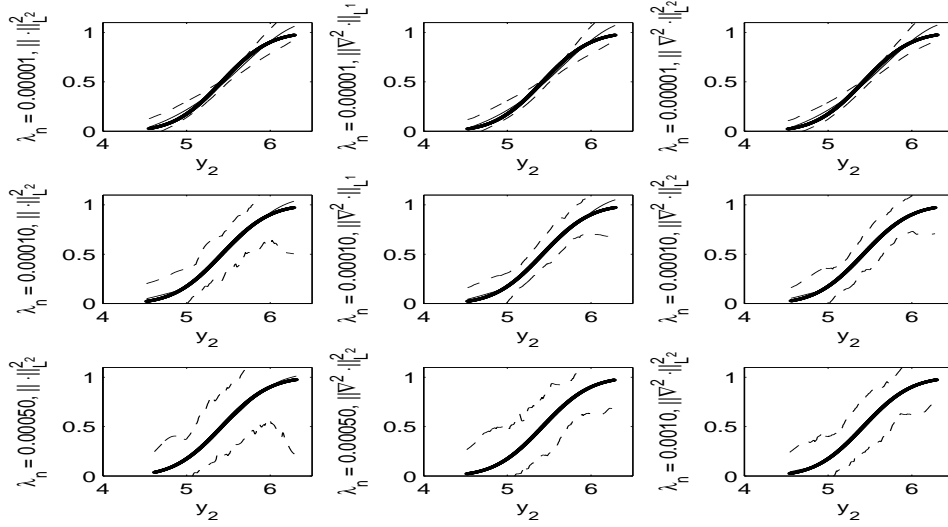
Figure 2.2: Table 2.1 experiments. "Sieve dominating" (1st row), sieve penalization balance (2nd row), "Penalization dominating" (3rd row).

case". However, at least in our MC study, the numerical implementations of such choices are too unstable to report.

2.3 Empirical Illustration

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

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

where Y_{1il} is the budget share of household i on good l (in this application, 1 : food-out, 2 : food-in, 3 : alcohol, 4 : fares, 5 : fuel, 6 : leisure goods, and 7 : travel). Y_{2i} is the log-total expenditure of household i that is endogenous, and X_i is the gross earnings of the head of household, which is the instrumental variable. We work with the no kids sample that consists of 628 observations. The same data set has been studied in BCK (2007) for the NPIV model (1.2). See Koenker (2005) for the linear quantile regression and nonparametric quantile regression ($E[1\{Y_{1il} \leq h_{0l}(X_i)\}|X_i] = \gamma$) of Engel curves when the total expenditure is exogenous (i.e., $Y_2 = X$).

As illustration, we apply the penalized SMD using a finite-dimensional polynomial spline sieve to construct the sieve space $\mathcal{H}_n$ for h, with different types of penalty functions. We have tried $||\nabla^k h||^j_{L^j(d\hat{\mu})} \equiv n^{-1} \sum_{i=1}^n |\nabla^k h(Y_{2i})|^j$ for $k = 1, 2$ and $j = 1, 2$, and Hermite polynomial sieves, cosine sieves and polynomial splines sieves for the series LS estimator $\hat{m}$. All combinations yielded very similar results; hence we only present figures for one "sieve dominating case". Due to the lack of space, in Figure 2.3 we report the estimated Engel curves only for three different quantiles

11

$\gamma = \{0.25, 0.50, 0.75\}$ and for four selected goods, using P-Spline(2,5) as $\mathcal{H}_n$ and P-Spline(5,10) for $\hat{m}$. Figure 2.3 presents the estimated Engel curves using $\widehat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\hat{\mu})}$ with $\lambda_n = 0.001$ and $\widehat{P}_n(h) = ||\nabla^2 h||_{L^1(d\hat{\mu})}$ with $\lambda_n = 0.001$ in the first and second rows; $\widehat{P}_n(h) = ||\nabla h||^2_{L^2(d\hat{\mu})}$ with $\lambda_n = 0.001$ (third row), and $\lambda_n = 0.003$ (fourth row); and $\widehat{P}_n(h) = ||\nabla h||^2_{L^2(leb)}$ with $\lambda_n = 0.005$ (fifth row). By inspection, we see that the overall estimated function shapes are not very sensitive to the choices of λ_n and $\widehat{P}_n(h)$, which is again consistent with the theoretical results for the penalized SMD estimator in the "sieve dominating case".
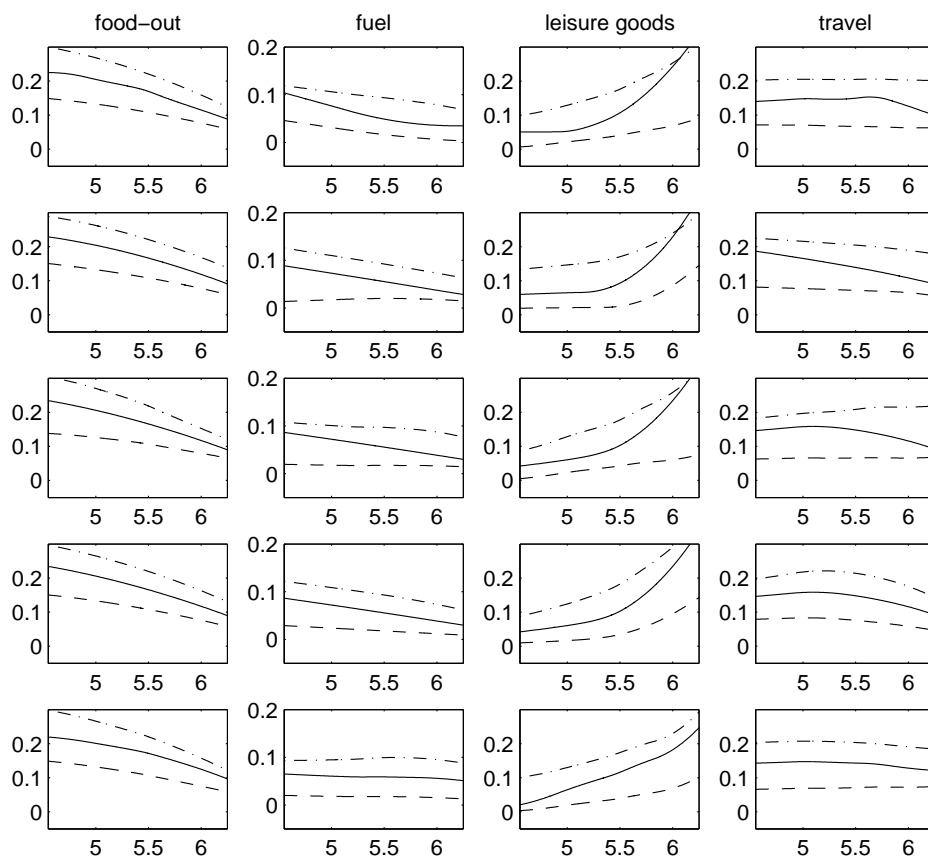
Figure 2.3: Engel curves for quantiles $\gamma = 0.25$ (dash), 0.50 (solid), 0.75 (dot-dash). $\hat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\hat{\mu})}$ with $\lambda_n = 0.001$ (1st row); $\hat{P}_n(h) = ||\nabla^2 h||_{L^1(d\hat{\mu})}$ with $\lambda_n = 0.001$ (2nd row); $\hat{P}_n(h) = ||\nabla h||^2_{L^2(d\hat{\mu})}$ with $\lambda_n = 0.001$ (3rd row), $\lambda_n = 0.003$ (4th row); $\hat{P}_n(h) = ||\nabla h||^2_{L^2(leb)}$ with $\lambda_n = 0.005$ (5th row).

3   Consistency

Lemma B.1 in Appendix B provides a general consistency lemma for an approximate penalized sieve extremum estimator that applies to both well-posed and ill-posed problems. Here in the main

text we provide some concrete sufficient conditions for consistency of the penalized SMD estimators (2.3).

ASSUMPTION 3.1. *(i)* $\{(Y_i, X_i)\}_{i=1}^n$ *is a random sample; (ii)* $\mathcal{H} \subseteq \mathbf{H}$, *and* $\mathbf{H} \equiv \mathbf{H}^1 \times \cdots \times \mathbf{H}^q$ *is a separable Banach space under the metric* $\|h\|_c \equiv \sum_{\ell=1}^q \|h_\ell\|_{c,\ell}$; *(iii)* $E[\rho(Z, h_0)|X] = 0$, *and* $\|h_0 - h\|_c = 0$ *for any* $h \in \mathcal{H}$ *with* $E[\rho(Z, h)|X] = 0$.

ASSUMPTION 3.2. *(i)* $\{\mathcal{H}_k : k \geq 1\}$ *are the sieve spaces satisfying* $\mathcal{H}_k \subseteq \mathcal{H}_{k+1} \subseteq \mathcal{H}$, *and there exists* $\Pi_n h_0 \in \mathcal{H}_{k(n)}$ *such that* $\|\Pi_n h_0 - h_0\|_c = o(1)$; *(ii)* $E[m(X, \Pi_n h_0)' m(X, \Pi_n h_0)] = o(1)$.

Given $m(X, h_0) = 0$ and assumption 3.2(i), assumption 3.2(ii) is implied by *assumption 3.2(ii)'*: $E[m(X, h)' m(X, h)]$ *is continuous at* h_0 *under* $\|\cdot\|_c$.

ASSUMPTION 3.3. *(i)* $\widehat{m}(x, h)$ *and* $\widehat{P}_n(h)$ *are measurable functions of the data* $\{(Y_i, X_i)\}_{i=1}^n$ *for almost all* $x \in \mathcal{X}$ *and all* $h \in \mathcal{H}_{k(n)}$; *(ii)* $\widehat{h}_n \in \mathcal{H}_{k(n)}$ *is well-defined with probability approaching one.*

See Remark B.1 in Appendix B for general sufficient conditions for assumption 3.3.

ASSUMPTION 3.4. *either (a) or (b) holds: (a)* $\lambda_n = 0$; *(b)* $\lambda_n > 0$, $\lambda_n \sup_{h \in \mathcal{H}_n} |\widehat{P}_n(h) - P(h)| = O_P(\lambda_n) = o_P(1)$, *with* $P(\cdot)$ *a non-negative real-valued measurable function of* $h \in \mathcal{H}$, $P(h_0) < \infty$ *and* $\lambda_n |P(\Pi_n h_0) - P(h_0)| = O(\lambda_n) = o(1)$.

3.1 Well-posed case

Although our main focus is on the ill-posed problems, for the sake of comparison, we first present a consistency theorem for the well-posed cases (i.e., the $\|h - h_0\|_c$ metric is continuous with respect to the criterion $E[m(X, h)' m(X, h)]$ near zero). Let $N(\delta, \mathcal{H}_n, \|\cdot\|_c)$ denote the minimal number of radius δ covering balls of $\mathcal{H}_n$ under the $\|\cdot\|_c$ metric.

ASSUMPTION 3.5. *(i) There are a measurable function* $b(X)$ *with* $E[b(X)] < \infty$ *and a finite constant* $\kappa > 0$ *such that for all* $\delta > 0$ *we have*

$$\sup_{\{h, h' \in \mathcal{H}_n: \|h - h'\|_c \leq \delta\}} \|m(x, h) - m(x, h')\|_E \leq b(x) \delta^\kappa;$$

(ii) either (a) $\sup_{h \in \mathcal{H}_n} \|m(X, h)\|_E \leq K < \infty$ *almost all* X; *or (b)* $E[|b(X)|^2]$ *and* $E\left[\sup_{h \in \mathcal{H}_n} \|m(X, h)\|_E^2\right]$ *are bounded; (iii)* $\sup_{h \in \mathcal{H}_n} n^{-1} \sum_{i=1}^n \|\widehat{m}(X_i, h) - m(X_i, h)\|_E^2 = o_P(1)$; *(iv)* $\log\left(N(\epsilon^{1/\kappa}, \mathcal{H}_n, \|\cdot\|_c)\right) = o(n)$ *for all* $\epsilon > 0$.

We note that assumptions 3.1(i) and 3.5 imply condition (3.1.2) in the next theorem.

THEOREM 3.1. *Let* $\hat{h}_n$ *be the penalized SMD estimator with* $\lambda_n \geq 0$, $\lambda_n = o(1)$, *and* $\widehat{m}(X, h)$ *any consistent estimator of* $m(X, h)$. *Let assumptions 3.1(ii)(iii), 3.2, 3.3 and 3.4 hold. Suppose that the following condition (3.1.1) and condition (3.1.2) (or assumptions 3.1(i) and 3.5) hold:*

13

(3.1.1) for any sequence $\{h_k \in \mathcal{H}_k\}$ with $\liminf_{k \to \infty} E[m(X, h_k)'m(X, h_k)] = 0$, it holds $||h_k - h_0||_c \to 0$ as $k \to \infty$; (3.1.2) $\sup_{h \in \mathcal{H}_n} |n^{-1} \sum_{i=1}^{n} ||\widehat{m}(X_i, h)||_E^2 - E[||m(X, h)||_E^2]| = o_P(1)$.

Then: $||\hat{h}_n - h_0||_c = o_P(1)$.

3.1.1 Compact parameter space case

We now specialize Theorem 3.1 to the case when the original parameter space $\mathcal{H}$ is compact under $|| \cdot ||_c$. We impose the following sufficient conditions for assumptions 3.3 and 3.5(iv), conditions (3.1.1) and (3.1.2):

ASSUMPTION 3.6. *(i) the sieve spaces $\mathcal{H}_n$ are compact under $|| \cdot ||_c$; (ii) $\lambda_n \sup_{h \in \mathcal{H}_n} P(h) = o(1)$.*

ASSUMPTION 3.7. *(i) the parameter space $\mathcal{H}$ is compact under $|| \cdot ||_c$; (ii) $E[m(X, h)'m(X, h)]$ is lower semicontinuous on $\mathcal{H}$ under $||\cdot||_c$.*

COROLLARY 3.1. *Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n \geq 0$, $\lambda_n = o(1)$, and $\widehat{m}(X, h)$ any consistent estimator of $m(X, h)$. Let assumptions 3.1, 3.2, 3.4, 3.5(i)(ii)(iii), 3.6 and 3.7 hold. Then: $||\hat{h}_n - h_0||_c = o_P(1)$.*

Under assumptions 3.6(i) and 3.7(i), assumption 3.5(iii) is satisfied by commonly used nonparametric regression estimators of $m(X, h)$, such as the kernel estimator, the local linear regression, and the series least square (LS) estimator $\widehat{m}(X, h)$ defined in (2.6). See, e.g., Newey (1991), Andrews (1995), NP (2003), AC (2003) and CIN (2007) for details.

3.2 Ill-posed case

We next present consistency results that allow for ill-posed problems (i.e., the $||h - h_0||_c$ metric is not continuous with respect to the criterion $E[m(X, h)'m(X, h)]$ near zero) and without assuming $|| \cdot ||_c$−compactness of $\mathcal{H}$. We first strengthen the speed of convergence of $\widehat{m}(X, h)$ to $m(X, h)$:

ASSUMPTION 3.8. *(i) $\sup_{h \in \mathcal{H}_n} E[||\widehat{m}(X, h) - m(X, h)||_E^2] = O_p(\delta_{m,n}^2) = o_P(1)$; (ii) $E[||\widehat{m}(X, h)||_E^2] \asymp n^{-1} \sum_{i=1}^{n} ||\widehat{m}(X_i, h)||_E^2$ uniformly over $h \in \mathcal{H}_n$.*

Many commonly used nonparametric estimator of the conditional mean function $m(X, h)$ can be shown to satisfy assumption 3.8. For example, under the following two mild assumptions 3.9 and 3.10, the series LS estimator $\widehat{m}(X, h)$ satisfies assumption 3.8 with $\delta_{m,n}^2 = \max\{\frac{J_n}{n}, b_{m,J_n}^2\}$, and $\sup_{h \in \mathcal{H}_n} n^{-1} \sum_{i=1}^{n} ||\widehat{m}(X_i, h) - m(X_i, h)||_E^2 = O_P(\delta_{m,n}^2)$; see Lemmas B.2 and B.3 in the Appendix.

ASSUMPTION 3.9. *(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) Denote $\xi_n \equiv \sup_{X \in \mathcal{X}} ||p^{J_n}(X)||_E$. Either $\xi_n^2 J_n = o(n)$ or $J_n \log(J_n) = o(n)$ for polynomial spline $p^{J_n}(X)$ sieve.*

If $p^{J_n}(X)$ is the spline or cosine/sine or wavelet sieves, then $\xi_n \asymp J_n^{1/2}$; see e.g. Newey (1997) or Huang (1998).

ASSUMPTION 3.10. *(i)* $\sup_{h \in \mathcal{H}_n} \sup_x Var[\rho(Z, h)|X = x] \leq K < \infty$; *(ii) for any* $g \in \{m(\cdot, h) : h \in \mathcal{H}_n\}$, *there is* $p^{J_n}(X)'\pi$ *such that, uniformly over* $h \in \mathcal{H}_n$, *either (a) or (b) holds: (a)* $\sup_x |g(x) - p^{J_n}(x)'\pi| = O(b_{m,J_n}) = o(1)$; *(b)* $E\{[g(X) - p^{J_n}(X)'\pi]^2\} = O(b_{m,J_n}^2)$ *for* $p^{J_n}(X)$ *sieve with* $\xi_n = O(J_n^{1/2})$.

Assumption 3.10(ii) is satisfied by typical smooth function classes of $\{m(\cdot, h) : h \in \mathcal{H}_n\}$ and typical linear sieves $p^{J_n}(X)$. For example, if $\{m(\cdot, h) : h \in \mathcal{H}_n\}$ is a subset of $\Lambda_c^{\gamma_m}(\mathcal{X})$ (or $W_{2,c}^{\gamma_m}(\mathcal{X}, leb.)$) with $\gamma_m > 0$, then assumption 3.10(ii) (a) and (b) hold with $b_{m,J_n} = J_n^{-r_m}$ where $r_m = \gamma_m/d_x$ and tensor product polynomial splines, wavelets or Fourier series sieves.

Before we present consistency results for the ill-posed case, we state a general lemma about the property of penalty function $\lambda_n P(\widehat{h}_n)$. The following assumption is a stronger version of assumption 3.4(b):

ASSUMPTION 3.11. $\lambda_n > 0$, $\lambda_n \sup_{h \in \mathcal{H}_n} |\widehat{P}_n(h) - P(h)| = o_P(\lambda_n)$, *with* $P(\cdot)$ *a non-negative real-valued measurable function of* $h \in \mathcal{H}$, $P(h_0) < \infty$ *and* $\lambda_n |P(\Pi_n h_0) - P(h_0)| = o(\lambda_n)$.

Under assumption 3.2(i), $\lambda_n > 0$ and $P(h_0) < \infty$, a sufficient condition for $\lambda_n |P(\Pi_n h_0) - P(h_0)| = o(\lambda_n)$ is that $P(\cdot)$ *is continuous at* h_0 *under* $\|\cdot\|_c$. Note that assumptions 3.4(b) and 3.11 are trivially satisfied when $\mathcal{H}_n = \mathcal{H}$ and $\widehat{P}_n = P$.

LEMMA 3.1. *Let* $\hat{h}_n$ *be the penalized SMD estimator satisfy assumption 3.3 with* $\lambda_n > 0$, $\lambda_n = o_P(1)$, *and* $\widehat{m}(X, h)$ *any consistent estimator of* $m(X, h)$ *satisfying assumption 3.8 at* $h = \Pi_n h_0$.

(1) Under assumption 3.4(b) and $\max\{\delta_{m,n}^2, E[\|m(X, \Pi_n h_0)\|_E^2]\} = O(\lambda_n)$, $P(\widehat{h}_n) = O_P(1)$.

(2) Under assumption 3.11 and $\max\{\delta_{m,n}^2, E[\|m(X, \Pi_n h_0)\|_E^2]\} = o(\lambda_n)$, $P(\widehat{h}_n) \leq P(h_0) + o_P(1)$.

3.2.1 Finite dimensional sieve dominating case

ASSUMPTION 3.12. *There are a positive non-increasing function* $B(k)$ *and a non-decreasing lower semicontinuous function* $g_m()$ *with* $g_m(0) = 0, g_m(\varepsilon) > 0$ *for* $\varepsilon > 0$, *such that*

$$E[m(X, h)'m(X, h)] \geq B(k)g_m(\|h - h_0\|_c) \quad \text{for all } h \in \mathcal{H}_k, \text{ all } k \geq 1.$$

THEOREM 3.2. *Let* $\hat{h}_n$ *be the penalized SMD estimator with* $\lambda_n \geq 0$, $\lambda_n = o(1)$, *and* $\widehat{m}(X, h)$ *any consistent estimator of* $m(X, h)$ *satisfying assumption 3.8 (or assumptions 3.9 - 3.10 for series LS estimator* $\widehat{m}(X, h)$). *Suppose that assumptions 3.1, 3.2, 3.3, 3.4, and 3.12 hold. If*

$$\max\left\{\delta_{m,n}^2, E(\|m(X, \Pi_n h_0)\|_E^2), \lambda_n\right\} = o\left(B(k(n))\right),$$

then $\|\widehat{h}_n - h_0\|_c = o_P(1)$, *and if* $\lambda_n > 0$ *then* $P(\widehat{h}_n) = O_P(1)$.

Theorem 3.2 allows for $\lambda_n = 0$; hence it establishes consistency for the original SMD estimator proposed in NP (2003) and AC (2003) without assuming $|| \cdot ||_c-$compactness of $\mathcal{H}$ and $\mathcal{H}_{k(n)}$. This theorem also allows for any kinds of penalty function $P(h)$ with $\lambda_n > 0$, $\lambda_n = O(\max\{\delta_{m,n}^2, E(\|m(X, \Pi_n h_0)\|_E^2)\}) = o\left(B(k(n))\right)$ as long as it satisfies the mild assumption 3.4(b).

3.2.2   Lower semicompact penalty

In this subsection we present a consistency result when the original parameter space $\mathcal{H}$ is not known *a priori* compact under $|| \cdot ||_c$ but the penalty function is *lower semicompact* (i.e., the set $\{h \in \mathcal{H} : P(h) \leq M\}$ is compact under $|| \cdot ||_c$ for all $M < \infty$).

ASSUMPTION 3.13. *(i) The set $\{h \in \mathcal{H} : P(h) \leq M\}$ is compact under $|| \cdot ||_c$ for all $0 \leq M < \infty$; (ii) the sieve spaces $\mathcal{H}_n$ are closed under $|| \cdot ||_c$.*

The next consistency result indicates that the lower semicompact penalty converts an ill-posed problem to a well-posed one.[7] It also demonstrates that one can replace assumption 3.12 by some stronger condition on the penalty function. To present an easy-to-verify consistency result we also replace assumption 3.3 by some simply sufficient conditions.

THEOREM 3.3. *Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n > 0$, $\lambda_n = o(1)$, and $\widehat{m}(X, h)$ any consistent estimator of $m(X, h)$ satisfying assumption 3.8 (or assumptions 3.9 - 3.10 for series LS estimator $\widehat{m}(X, h)$). Suppose that assumptions 3.1, 3.2, 3.4(b), 3.7(ii), and 3.13 hold. If*

$$\max\left\{\delta_{m,n}^2, E\{\|m(X, \Pi_n h_0)\|_E^2\}\right\} = O\left(\lambda_n\right) = o(1),$$

then: $||\hat{h}_n - h_0||_c = o_P(1)$ and $P(\hat{h}_n) = O_P(1)$.

Remark 3.1. *When $P(h)$ is convex, under assumptions 3.1(iii) and 3.13, the penalized SMD estimator $\hat{h}_n$ using a closed finite dimensional linear sieve $\mathcal{H}_{k(n)}$ is equivalent to the original SMD estimator using a finite dimensional compact sieve:*

$$\hat{h}_n = \arg \inf_{h \in \mathcal{H}_{k(n)}:\widehat{P}_n(h) \leq M_n} \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(X_i, h)'\widehat{m}(X_i, h), \quad with \quad M_n \to \infty \text{ slowly.}$$

Therefore, Theorem 3.3 also establishes the consistency of the original SMD estimator using finite dimensional compact sieves of the type $\{h \in \mathcal{H}_{k(n)} : \widehat{P}_n(h) \leq M_n\}$ without assuming the $|| \cdot ||_c-$compactness of the original parameter space $\mathcal{H}$. In particular, this immediately implies consistency of the SMD estimators of the NPIV model (1.2) $E[Y_1 - h_0(Y_2)|X] = 0$ studied in NP (2003), AC (2003) and BCK (2007), and the NPQIV model studied in CIN (2007), without requiring that $\mathcal{H}$ is a compact subset of the space $L^2(f_{Y_2})$.

---

[7]We are grateful to Victor Chernozhukov for pointing out the nice property of lower semicompact penalty. See Remark 3.3 for further discussion.

3.2.3 Penalization dominating case with general penalty

In this and the next two subsections we present consistency results for general penalty functions that may not be lower semicompact, but they satisfy assumption 3.11 (which is a stronger version of assumption 3.4(b)).

For a Banach space $\mathbf{H}$ we denote $\mathbf{H}^*$ as the dual of $\mathbf{H}$, and $\langle \cdot, \cdot \rangle_{\mathbf{H}^*, \mathbf{H}}$ as the inner product that links the space $\mathbf{H}$ with its dual $\mathbf{H}^*$.

ASSUMPTION 3.14. *There are a $t_0 \in \mathbf{H}^*$ with $\langle t_0, \cdot \rangle_{\mathbf{H}^*, \mathbf{H}}$ a bounded linear functional with respect to $|| \cdot ||_c$, and a non-decreasing lower semicontinuous function $g()$ with $g(0) = 0, g(\varepsilon) > 0$ for $\varepsilon > 0$, such that*

$$P(h) - P(h_0) - \langle t_0, h - h_0 \rangle_{\mathbf{H}^*, \mathbf{H}} \geq g(||h - h_0||_c) \quad \text{for all } h \in \mathcal{H}_k, \text{ all } k \geq 1.$$

When $\mathcal{H}$ is convex, Assumption 3.14 is satisfied if $P(h)$ is *strongly convex* at h_0 under $|| \cdot ||_c$, that is, there exists a $c' > 0$ such that

$$P(h) - P(h_0) - \langle DP(h_0), h - h_0 \rangle_{\mathbf{H}^*, \mathbf{H}} \geq c' ||h - h_0||_c^2 \quad \text{for all } h \in \mathcal{H},$$

where $DP(h_0) \in \mathbf{H}^*$ is the Gateaux derivative of $P()$ at h_0. We note that the strong convexity is satisfied by commonly used penalization function $P(h)$, and it obviously implies that $P(h)$ is *strictly* convex at h_0; see, e.g., Eggermont and LaRiccia (2001).

ASSUMPTION 3.15. *For all $\{h_k \in \mathcal{H}_k\}$ with $\liminf_{k\to\infty} E[m(X, h_k)'m(X, h_k)] = 0$, it holds that $\liminf_{k\to\infty} \langle t_0, h_k - h_0 \rangle_{\mathbf{H}^*, \mathbf{H}} = c$ for some $c \geq 0$.*

THEOREM 3.4. *Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n > 0$, $\lambda_n = o(1)$, and $\widehat{m}(X, h)$ any consistent estimator of $m(X, h)$ satisfying assumption 3.8 (or assumptions 3.9 - 3.10 for series LS estimator $\widehat{m}(X, h)$). Let assumptions 3.1, 3.2, 3.3, 3.11, 3.14 and 3.15 hold. Suppose that either (a) or (b) holds:*

(a) $\max\{\delta_{m,n}^2, E(||m(X, \Pi_n h_0)||_E^2)\} = o(\lambda_n)$;

(b) assumption 3.12 holds with $\max\{\delta_{m,n}^2, E(||m(X, \Pi_n h_0)||_E^2)\} = o(\max\{B(k(n)), \lambda_n\})$.

Then: $||\widehat{h}_n - h_0||_c = o_P(1)$, and $P(\widehat{h}_n) = P(h_0) + o_P(1)$.

3.2.4 Closed, convex and bounded parameter space case

We now present some sufficient conditions for assumptions 3.3 and 3.15 without requiring that the infinite dimensional parameter space $\mathcal{H}$ is compact under $|| \cdot ||_c$. We first recall some standard definitions. A sequence $\{h_j\}$ in a Banach space $\mathbf{H}$ converges *weakly* to h iff $\lim_{j\to\infty} \langle v, h_j - h_0 \rangle_{\mathbf{H}^*, \mathbf{H}} = \langle v, h - h_0 \rangle_{\mathbf{H}^*, \mathbf{H}}$ for all $v \in \mathbf{H}^*$. A functional $F : \mathcal{H} \subseteq \mathbf{H} \to [-\infty, +\infty]$ is said to be *weak sequentially lower semicontinuous* at $h \in \mathcal{H}$ iff $F(h) \leq \liminf_{j\to\infty} F(h_j)$ for each sequence $\{h_j\}$ in $\mathcal{H}$ that

converges weakly to h. A Banach space $\mathbf{H}$ is *reflexive* iff $(\mathbf{H}^*)^* = \mathbf{H}$. For example, the spaces L^p for $1 < p < \infty$, and the Sobolev spaces W_p^γ for $1 < p < \infty$ are reflexive and separable Banach spaces.

ASSUMPTION 3.16. *(i)* $(\mathbf{H}, || \cdot ||_c)$ *is a reflexive Banach space; (ii)* $\mathcal{H}$ *is a closed and convex subset in* $(\mathbf{H}, || \cdot ||_c)$*; (iii)* $\mathcal{H}$ *is bounded in* $|| \cdot ||_c$ *(i.e.,* $\sup_{h \in \mathcal{H}} ||h||_c \leq K < \infty$*)*.

Assumption 3.16(iii) is implied by the so-called *coercive* condition, denoted as *Assumption 3.16(iii)'*: $E[m(X,h)'m(X,h)] + \lambda P(h) \to +\infty$ *as* $||h||_c \to +\infty$ *for* $h \in \mathcal{H}$ *and* $\lambda \in (0,1]$.

ASSUMPTION 3.17. $E[m(X,h)'m(X,h)]$ *is weak sequentially lower semicontinuous on* $\mathcal{H}$.

Remark 3.2. *Under assumption 3.16, assumption 3.17 is implied by either 3.17' or 3.17":*

*Assumption 3.17': * $m(\cdot, h) : \mathcal{H} \subseteq \mathbf{H} \to L^2(f_X)$ *is compact (i.e., continuous and maps bounded sets in* $\mathcal{H}$ *into relatively compact sets in* $L^2(f_X)$*).*

*Assumption 3.17": * $E[m(X,h)'m(X,h)]$ *is convex and lower semicontinuous on* $\mathcal{H}$ *(in* $\|\cdot\|_c$*).*

ASSUMPTION 3.18. *Either (a) or (b) holds: (a)* $\mathcal{H}_k$ *are compact under* $|| \cdot ||_c$*, and* $P(h)$ *is lower semicontinuous on* $\mathcal{H}_k$ *(in* $\|\cdot\|_c$*); (b)* $\mathcal{H}_k$ *are closed and convex subsets of* $\mathcal{H}$*, and* $P(h)$ *is convex and lower semicontinuous on* $\mathcal{H}_k$ *(in* $\|\cdot\|_c$*).*

COROLLARY 3.2. *Let* $\hat{h}_n$ *be the penalized SMD estimator with* $\lambda_n > 0$*,* $\lambda_n = o(1)$*, and* $\widehat{m}(X,h)$ *any consistent estimator of* $m(X,h)$ *satisfying assumption 3.8 (or assumptions 3.9 - 3.10 for series LS estimator* $\widehat{m}(X,h)$*). Let assumptions 3.1, 3.2, 3.11, 3.14, 3.16, 3.17 (or 3.17' or 3.17") and 3.18 hold. Then the conclusion of Theorem 3.4 holds.*

Remark 3.3. *Comparing Theorem 3.4 and Corollary 3.2 to Theorem 3.3, all consistency results allow for non-compact (in* $||\cdot||_c$*) parameter space* $\mathcal{H}$ *and without assumption 3.12. Nevertheless, the condition* $\max\left\{\delta_{m,n}^2, E\{\|m(X, \Pi_n h_0)\|_E^2\}\right\} = o(\lambda_n)$ *imposed in Theorem 3.4 and Corollary 3.2 for a general penalty can be improved to the condition* $\max\left\{\delta_{m,n}^2, E\{\|m(X, \Pi_n h_0)\|_E^2\}\right\} = O(\lambda_n) = o(1)$ *in Theorem 3.3 for a lower semicompact penalty. In addition, using a lower semicompact penalty, Theorem 3.3 leads to consistency without imposing assumptions 3.14 and 3.16. This means that by applying Theorem 3.3, one can obtain sup-norm consistency of the penalized SMD estimator using a lower semicompact penalty.*

3.2.5   Point identification induced by convex penalty

In this subsection we consider an important class of problems in which $E[m(X,h)'m(X,h)]$ is convex. We shall replace the old uniqueness assumption 3.1(iii) by including a prior information of $P(h) \leq M_0$ for a known constant $M_0 < \infty$. As already mentioned in Section 2, from the well-known

results on convex optimization (see, e.g., Eggermont and LaRiccia, 2001), when $E\left[m(X,h)'m(X,h)\right]$ and $P(h)$ are convex, and $\mathcal{H}$ is closed and convex, the constraint optimization problem

$$h_0 \in \mathcal{M}_0 \equiv \left\{ h : \arg \inf_{h \in \mathcal{H}: P(h) \leq M_0} E\left[m(X,h)'m(X,h)\right] \right\} \tag{3.1}$$

is equivalent to an unconstrained optimization problem

$$h_0 \in \mathcal{M}_0 = \left\{ h : \arg \inf_{h \in \mathcal{H}} \{E\left[m(X,h)'m(X,h)\right] + \lambda_0 P(h)\} \right\},$$

with $\lambda_0 \geq 0$ such that $\lambda_0[P(h_0) - M_0] = 0$ (more precisely, $\lambda_0 > 0$ for $P(h_0) = M_0$ and $\lambda_0 = 0$ for $P(h_0) < M_0$).

ASSUMPTION 3.19. (i) For all h_0, h' belonging to $\mathcal{M}_0$ defined in (3.1), it follows that $||h_0 - h'||_c = 0$; (ii) $P(\cdot)$ is convex and lower semicontinuous on $\mathcal{H}$ (in $\|\cdot\|_c$).

Assumption 3.19(i) implicitly assumes that the set $\mathcal{M}_0$ is not empty and explicitly imposes that it is a singleton $\{h_0\}$ (up to an equivalent class in $||\cdot||_c$).

Remark 3.4. If $E[||m(X,h)||_E^2] + \lambda_0 P(h)$ is strictly convex on $\mathcal{H}$, then assumption 3.19(i) is automatically satisfied. For instance, the condition that $E[||m(X,h)||_E^2]$ is convex and $P(h)$ is strictly convex, will suffice. For the class of problems that $m(X,h)$ is linear in $h \in \mathcal{H} = L^p$ (such as in the NPIV model), assumption 3.19 is trivially satisfied if $P(h) = ||h||_{L^p}^p$ with $p > 1$. Although the strict convexity of $E[||m(X,h)||_E^2]$ alone does imply assumption 3.1(iii), it might be too strong. For example, in the NPIV model (1.2) we have $m(X,h) = E[Y_1 - h(Y_2)|X]$, and hence $E[||m(X,h)||_E^2]$ is strictly convex in $h \in \mathcal{H} = L^2(f_{Y_2})$ iff the conditional density of Y_2 given X is complete; see, e.g., NP (2003), DFR (2006) and CFR (2007).

The next consistency result says that we can replace the original identification assumption 3.1(iii) by this new assumption 3.19(i), and that we can compute minimization over unconstraint, closed and convex sieves $\mathcal{H}_{k(n)}$.

THEOREM 3.5. Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n = \lambda_0 + o(1) > 0$, $\lambda_0 \geq 0$, and $\widehat{m}(X,h)$ any consistent estimator of $m(X,h)$ satisfying assumption 3.8 (or assumptions 3.9 - 3.10 for series LS estimator $\widehat{m}(X,h)$). Let assumptions 3.1(i)(ii), 3.2, 3.11, 3.14, 3.16(i)(ii), 3.16(iii)', 3.17", 3.18(b) and 3.19 hold. Suppose that either (a) or (b) holds:

(a) $\max\{\delta_{m,n}^2, E(\|m(X,\Pi_n h_0)\|_E^2)\} = o(\lambda_n)$;

(b) assumption 3.12 holds with $\max\{\delta_{m,n}^2, E(\|m(X,\Pi_n h_0)\|_E^2)\} = o(\max\{B(k(n)), \lambda_n\})$.

Then: $||\widehat{h}_n - h_0||_c = o_P(1)$ and $P(\widehat{h}_n) = P(h_0) + o_P(1)$, where $\{h_0\} = \mathcal{M}_0$.

If there exists $h_0 \in \mathcal{H}$ such that $E[||m(X,h_0)||_E^2] = 0$ *but* is not unique, then assumption 3.19(i) and Theorem 3.5 imply that the penalized SMD estimator will converge to a $h_0 = \arg \inf_{h \in \mathcal{H}} \{P(h) :$

$E[||m(X,h)||_E^2] = 0\}$. If $P(h)$ is a norm such as $||h||_{L^2}^2$ then this becomes the so-called minimum norm solution in the literature on ill-posed inverse problems; see, e.g., Engl, Hanke and Neubauer (1996).

In Theorem 3.5, $\max\left\{\delta_{m,n}^2, E\{||m(X, \Pi_n h_0)||_E^2\}\right\} = o(\lambda_n)$ and $\lambda_n = \lambda_0 + o(1) > 0$, with $\lambda_0 > 0$ for $P(h_0) = M_0$ and $\lambda_0 = 0$ for $P(h_0) < M_0$. Thus, when the constraint is binding $P(h_0) = M_0$, it suffices that $\max\left\{\delta_{m,n}^2, E\{||m(X, \Pi_n h_0)||_E^2\}\right\} = o(1)$ and λ_n can be chosen such that $\widehat{P}_n(\hat{h}_n) = M_0$.

Remark 3.5. *When $E[||m(X,h)||_E^2]$ and $P(h)$ are convex, for the minimization problem (3.1), the penalized SMD estimator $\hat{h}_n = \arg\inf_{h\in\mathcal{H}_{k(n)}}\left\{\frac{1}{n}\sum_{i=1}^n \widehat{m}(X_i,h)'\widehat{m}(X_i,h) + \lambda_n\widehat{P}_n(h)\right\}$ using a closed finite dimensional linear sieve $\mathcal{H}_{k(n)}$ (hence compact) is equivalent to the original SMD estimator using a finite dimensional compact sieve:*

$$\hat{h}_n = \arg\inf_{h\in\mathcal{H}_{k(n)}:\widehat{P}_n(h)\le M_0}\frac{1}{n}\sum_{i=1}^n \widehat{m}(X_i,h)'\widehat{m}(X_i,h).$$

Therefore, Theorem 3.5 also establishes consistency of the original SMD estimator using finite dimensional compact sieves of the type $\{h \in \mathcal{H}_{k(n)} : \widehat{P}_n(h) \le M_0\}$ without assuming the $||\cdot||_c$-compactness of the original parameter space $\mathcal{H}$.

3.3 Consistency of the weighted penalized SMD estimator

In Remark 2.2 we presented a semi/nonparametric weighted version of our penalized SMD estimator. In this section we point out that all previous consistency results remain valid for the following nonparametric weighted penalized SMD estimator:

$$\widehat{h}_n \equiv \arg\inf_{h\in\mathcal{H}_n}\left\{\frac{1}{n}\sum_{i=1}^n \widehat{m}(X_i,h)'[\widehat{\Sigma}(X_i)]^{-1}\widehat{m}(X_i,h) + \lambda_n\widehat{P}_n(h)\right\}. \tag{3.2}$$

ASSUMPTION 3.20. *(i) $\sup_{x\in\mathcal{X}}\left|\widehat{\Sigma}(x) - \Sigma(x)\right| = o_P(1)$; (ii) $\Sigma(X)$ is finite positive definite, and its smallest and largest eigenvalues are positive and bounded uniformly over X.*

The next theorem can be trivially established by following all the consistency proofs for the penalized SMD estimator without $\widehat{\Sigma}(x)$; hence we omit the proof due to the length of the paper.

THEOREM 3.6. *Let $\hat{h}_n$ be the weighted penalized SMD estimator (3.2) with $\lambda_n \ge 0$, $\lambda_n = o(1)$, and $\widehat{m}(X,h)$ any consistent estimator of $m(X,h)$. Let assumption 3.20 hold. Then: Theorem 3.1, Corollary 3.1, Theorem 3.2, Theorem 3.3, Theorem 3.4, Corollary 3.2, and Theorem 3.5 remain true.*

Theorem 3.6 can also be trivially extended to establish consistency for the semi/nonparametric weighted penalized SMD estimator $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{h}_n)$ defined in (2.7). See Chen and Pouzo (2008) for details.

4 Convergence Rates

In the rest of the paper, we let $\|\cdot\|_s$ denote another metric on the infinite-dimensional function space $\mathcal{H}$ that is weaker than the norm $||\cdot||_c$ (i.e., $||h||_s \leq ||h||_c$ for all $h \in \mathcal{H}$). In this section we study convergence rate under the metric $||\cdot||_s$. Given the consistency results stated in Section 3, we can now restrict our attention to a shrinking $||\cdot||_c$−neighborhood around h_0. Let $\mathcal{H}_{os} \equiv \{h \in \mathcal{H} : ||h - h_0||_c = o(1), ||h||_c \leq c, P(h) \leq M_0\}$ and $\mathcal{H}_{osn} \equiv \{h \in \mathcal{H}_n : ||h - \Pi_n h_0||_c = o(1), ||h||_c \leq c, P(h) \leq M_0\}$. Then, for the purpose of establishing a rate of convergence under the $||\cdot||_s$ metric, we can treat $\mathcal{H}_{os}$ as the new parameter space and $\mathcal{H}_{osn}$ as its sieve space.

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

$$\frac{dm(X, h_0)}{dh}[h - h_0] \equiv \frac{dE[\rho(Z, (1-\tau)h_0 + \tau h)|X]}{d\tau}\bigg|_{\tau=0} \quad a.s.\ \mathcal{X}. \tag{4.1}$$

Following AC (2003), we define the pseudo-metric $||h_1 - h_2||$ for any $h_1, h_2 \in \mathcal{H}_{os}$ as

$$||h_1 - h_2|| \equiv \sqrt{E\left[\left(\frac{dm(X, h_0)}{dh}[h_1 - h_2]\right)'\left(\frac{dm(X, h_0)}{dh}[h_1 - h_2]\right)\right]}. \tag{4.2}$$

ASSUMPTION 4.1. *(i) $\mathcal{H}_{os}$ and $\mathcal{H}_{osn}$ are convex, $m(X, h)$ is continuously pathwise differentiable with respect to $h \in \mathcal{H}_{os}$. There is a finite constant $C > 0$ such that $||h - h_0|| \leq C||h - h_0||_s$ for all $h \in \mathcal{H}_{os}$; (ii) there are finite constants $c_1, c_2 > 0$ such that $||h - h_0||^2 \leq c_1 E[m(X, h)'m(X, h)]$ hold for all $h \in \mathcal{H}_{osn}$; and $c_2 E[m(X, h)'m(X, h)] \leq ||h - h_0||^2$ holds for all $h \in \mathcal{H}_{os}$.*

Assumption 4.1 implies that the weak metric $||h - h_0||$ is well-defined in $\mathcal{H}_{os}$ and is continuous with respect to the criterion function $E[m(X, h)'m(X, h)]$.

ASSUMPTION 4.2. *There is a $t_0 \in \mathbf{H}^*$ with $\langle t_0, \cdot \rangle_{\mathbf{H}^*, \mathbf{H}}$ a bounded linear functional with respect to $||\cdot||_s$ such that $\lambda_n \{P(h) - P(\Pi_n h_0) - \langle t_0, h - \Pi_n h_0 \rangle_{\mathbf{H}^*, \mathbf{H}}\} \geq 0$ for all $h \in \mathcal{H}_{osn}$.*

THEOREM 4.1. *Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n \geq 0$, $\lambda_n = o(1)$, $\widehat{P}_n(h) = P(h)$, and $\widehat{m}(X, h)$ any consistent estimator of $m(X, h)$ satisfying assumption 3.8 (or assumptions 3.9 - 3.10 for series LS estimator $\widehat{m}(X, h)$). Let $h_0 \in \mathcal{H}_{os}$ and $\hat{h}_n \in \mathcal{H}_{osn}$ with probability approaching one. Suppose that assumptions 3.1(i)(ii), 3.2, 3.4 and 4.1 hold. Then:*

$$(1)\quad ||\hat{h}_n - \Pi_n h_0|| = O_P\left(\max\left\{\delta_{m,n}, \sqrt{\lambda_n |P(\widehat{h}_n) - P(\Pi_n h_0)|}, ||\Pi_n h_0 - h_0||\right\}\right);$$

$= O_P\left(\max\{\delta_{m,n}, \sqrt{\lambda_n}, ||\Pi_n h_0 - h_0||\}\right)$ *under assumption 3.4;* $= O_P\left(\max\{\delta_{m,n}, o(\sqrt{\lambda_n}), ||\Pi_n h_0 - h_0||\}\right)$ *under assumption 3.11.*

(2) If assumption 4.2 holds, then:

$$
\begin{aligned}
||\hat{h}_n - \Pi_n h_0|| &= O_P\left(\max\left\{\delta_{m,n}, \sqrt{\lambda_n ||\hat{h}_n - \Pi_n h_0||_s}, ||\Pi_n h_0 - h_0||\right\}\right) \\
&= O_P\left(\max\left\{\delta_{m,n}, o(\sqrt{\lambda_n}), ||\Pi_n h_0 - h_0||\right\}\right).
\end{aligned}
$$

According to Theorem 4.1, one can obtain the convergence rate under the weak metric $||\cdot||$ by balancing the three parts: (1) $\delta_{m,n}$ (the estimation error rate of $m(X,h)$, which is $\sqrt{\frac{J_n}{n}} + b_{m,J_n}$ for the series LS estimator $\widehat{m}(X,h)$); (2) $||\Pi_n h_0 - h_0||$ (the sieve bias error rate under the weak metric $||\cdot||$); (3) λ_n (the penalization bias error rate).

Before we establish the convergence rate under the strong norm $||\hat{h}_n - h_0||_s$, we introduce two measures of ill-posedness in a shrinking neighborhood of h_0: the *sieve modulus of continuity*, $\omega_n(\delta, \mathcal{H}_{osn})$, and the *modulus of continuity*, $\omega(\delta, \mathcal{H}_{os})$, which are defined as[8]

$$\omega_n(\delta, \mathcal{H}_{osn}) \equiv \sup_{h \in \mathcal{H}_{osn}:||h-\Pi_n h_0|| \leq \delta} ||h - \Pi_n h_0||_s, \; \omega(\delta, \mathcal{H}_{os}) \equiv \sup_{h \in \mathcal{H}_{os}:||h-h_0|| \leq \delta} ||h - h_0||_s.$$

The definition of modulus of continuity, $\omega(\delta, \mathcal{H}_{os})$, does not depend on the choice of any estimation method. Therefore, when $\frac{\omega(\delta, \mathcal{H}_{os})}{\delta}$ goes to infinity as δ goes to zero, we say the problem of estimating h_0 under $||\cdot||_s$ is *ill-posed*.

In the following we present two theorems of convergence rates under the strong metric $||\cdot||_s$.

ASSUMPTION 4.3. $\omega_n(||\Pi_n h_0 - h_0||, \mathcal{H}_{osn}) \leq c||\Pi_n h_0 - h_0||_s$.

THEOREM 4.2. *(Sieve dominating case) Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n \geq 0$, $\lambda_n = o(1)$. Suppose that assumption 4.3 and all the assumptions of Theorem 4.1(1) hold. If $\max\{\delta_{m,n}, \sqrt{\lambda_n}\} = \delta_{m,n}$, then: $||\hat{h}_n - h_0||_s = O_P\left(||h_0 - \Pi_n h_0||_s + \omega_n(\delta_{m,n}, \mathcal{H}_{osn})\right)$.*

Theorem 4.2 allows for finite dimensional sieves with or without penalization, although $\sqrt{\lambda_n} \to 0$ faster than $\max\{\delta_{m,n}, ||\Pi_n h_0 - h_0||\} \to 0$ ("sieve dominating case"). It generalizes theorem 2 of BCK (2007) on sieve nonparametric IV regression to allow for nonlinear ill-posed inverse problems and possibly non-zero λ_n. In particular, when $\lambda_n = 0$ and $\mathcal{H}$ is compact under $||\cdot||_s$, Theorem 4.2 provides convergence rate in $||\cdot||_s$ for the original SMD estimators proposed in NP (2003) and AC (2003) for general nonlinear semi/nonparametric conditional moment models. To apply this theorem one, needs to compute the sieve modulus of continuity $\omega_n(\delta, \mathcal{H}_{osn})$; see subsection 5.1 for sufficient conditions to bound this term.

THEOREM 4.3. *(Penalization dominating case) Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n > 0$, $\lambda_n = o(1)$. Suppose that assumption 4.3 and all the assumptions of Theorem 4.1(1) hold. Let either assumption 3.13(i) holds with $\max\{\delta_{m,n}, \sqrt{\lambda_n}\} = \delta_{m,n} = O(\sqrt{\lambda_n})$, or assumption 4.2 holds with $\max\left\{\delta_{m,n}, \sqrt{\lambda_n||\hat{h}_n - \Pi_n h_0||_s}\right\} = \delta_{m,n}$. Then:*

$$||\hat{h}_n - h_0||_s = O_P\left(||h_0 - \Pi_n h_0||_s + \omega_n(\delta_{m,n}, \mathcal{H}_{osn})\right).$$

If $||h_0 - \Pi_n h_0||_s = 0$ then $||\hat{h}_n - h_0||_s = O_P\left(\omega(\delta_{m,n}, \mathcal{H}_{os})\right)$.

[8]Our definitions are inspired by the approach of Daubechies, Defrise and de Mol (2004) in their convergence analysis for the linear ill-posed inverse problem with a deterministic noise and a known operator.

Theorem 4.3 allows for both finite and infinite dimensional sieves with penalization, although now either $\max\{\delta_{m,n}, ||\Pi_n h_0 - h_0||\} = O(\sqrt{\lambda_n})$ as in the case with a lower semicompact penalty, or $\max\{\delta_{m,n}, ||\Pi_n h_0 - h_0||\} = o(\sqrt{\lambda_n})$ as in the case with a non-lower semicompact penalty. To apply this theorem, one needs to compute either the sieve modulus of continuity $\omega_n(\delta, \mathcal{H}_{osn})$ or the modulus of continuity $\omega(\delta, \mathcal{H}_{os})$; see subsection 5.2 for sufficient conditions to bound these terms.

The following corollary establishes the convergence rates for the penalized SMD estimator defined with $\widehat{\lambda}_n \widehat{P}_n(h)$ instead of $\lambda_n P(h)$.

COROLLARY 4.1. *Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n = o(1)$ and $\widehat{m}(X, h)$ any consistent estimator of $m(X, h)$ satisfying assumption 3.8 (or assumptions 3.9 - 3.10 for series LS estimator $\widehat{m}(X, h)$). If $\sup_{h \in \mathcal{H}_{osn}} \left| \frac{\widehat{\lambda}_n \widehat{P}_n(h) - \lambda_n P(h)}{\lambda_n P(h)} \right| = o_P(1)$ for $\lambda_n > 0$, then Theorems 4.1, 4.2 and 4.3 remain true.*

5 Sieve Modulus of Continuity and Optimal Rates

In this section we shall present some sufficient conditions to bound the sieve modulus of continuity and modulus of continuity. Throughout this section, we assume that $\mathcal{H}_{os}$ is a subset of a separable Hilbert space $\mathbf{H}$ with an inner product $\langle \cdot, \cdot \rangle_s$. Let $\{q_j\}_{j=1}^\infty$ be a Riesz basis associated with the Hilbert space $(\mathbf{H}, || \cdot ||_s)$, that is, any $h \in \mathbf{H}$ can be expressed as $h = \sum_j \langle h, q_j \rangle_s q_j$, and there are two finite constants $c_1, c_2 > 0$ such that $c_1 ||h||_s^2 \leq \sum_j |\langle h, q_j \rangle_s|^2 \leq c_2 ||h||_s^2$ for all $h \in \mathbf{H}$. See Appendix A for examples of commonly used function spaces and Riesz bases. For instance, if $\mathcal{H}_{os}$ is a subset of a Besov space, then the wavelet basis is a Riesz basis $\{q_j\}_{j=1}^\infty$.

5.1 Sufficient conditions

We first provide some sufficient conditions for the sieve modulus of continuity $\omega_n(\delta, \mathcal{H}_{osn})$ and assumption 4.3.

ASSUMPTION 5.1. *(i) $\{q_j\}_{j=1}^\infty$ is a Riesz basis for a real-valued separable Hilbert space $(\mathbf{H}, || \cdot ||_s)$, and $\mathcal{H}_{os}$ is a subset of $\mathbf{H}$; (ii) $||h_0 - \sum_{j=1}^{k(n)} \langle h_0, q_j \rangle_s q_j||_s = O(\{\nu_{k(n)}\}^{-\gamma_h})$ for a finite $\gamma_h > 0$ and an increasing positive sequence $\{\nu_j\}_{j=1}^\infty$.*

Assumption 5.1 suggests that $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ is a natural sieve for the estimation of h_0. For example, if $h_0 \in W_2^{\gamma_h}([0,1]^d, leb))$, then assumption 5.1(i) is satisfied with spline or wavelet or power series or Fourier series bases with $(\mathbf{H}, || \cdot ||_s) = (L^2([0,1]^d, leb), || \cdot ||_{L^2(leb)})$, and assumption 5.1(ii) is satisfied with $\nu_{k(n)} = \{k(n)\}^{1/d}$.

ASSUMPTION 5.2. *There are finite constants $c, C > 0$ and a non-increasing positive sequence $\{b_j \asymp \varphi(\nu_j^{-2})\}_{j=1}^\infty$ such that: (i) $||h||^2 \geq c \sum_{j=1}^\infty b_j |\langle h, q_j \rangle_s|^2$ for all $h \in \mathcal{H}_{osn}$; (ii) $C \sum_j b_j |\langle h_0 - \Pi_n h_0, q_j \rangle_s|^2 \geq ||h_0 - \Pi_n h_0||^2$.*

23

Assumption 5.2(i) is a low-level sufficient condition that links the weak-norm $||h||$ to its strong norm in a sieve shrinking neighborhood $\mathcal{H}_{osn}$ (of h_0). Assumption 5.2(ii) is so-called "stability condition" that is only required to hold in terms of the sieve approximation error $h_0 - \Pi_n h_0$ (of h_0).

LEMMA 5.1. Let $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ and assumption 5.1(i) hold.

   (1) If assumption 5.2(i) holds, then: $\omega_n(\delta, \mathcal{H}_{osn}) \leq const. \times \delta/\sqrt{b_{k(n)}}$.

   (2) If assumption 5.2(ii) holds, then: $||h_0 - \Pi_n h_0|| \leq const.\sqrt{b_{k(n)}}||h_0 - \Pi_n h_0||_s$.

   (3) If assumption 5.2(i)(ii) holds, then: assumption 4.3 is satisfied.

In order to bound the modulus of continuity $\omega(\delta, \mathcal{H}_{os})$ we need to strengthen both assumption 5.1 (on sieve approximation rate) and assumption 5.2(i) that links the weak metric $||h||$ to its strong metric $||h||_s$.

ASSUMPTION 5.3. There exist finite constants $M > 0$, $\gamma_h > 0$ and an increasing positive sequence $\{\nu_j\}_{j=1}^{\infty}$ such that $||h - \sum_{j=1}^{k}\langle h, q_j\rangle_s q_j||_s \leq M(\nu_{k+1})^{-\gamma_h}$ for all $h \in \mathcal{H}_{os}$.

Remark 5.1. Under assumption 5.1(i), assumption 5.3 is satisfied if there are finite constants $M > 0$, $\gamma_h > 0$ and an increasing positive sequence $\{\nu_j\}_{j=1}^{\infty}$ such that either (a) or (b) holds: (a) (ellipsoid) $\sum_{j=1}^{\infty}\nu_j^{2\gamma_h}|\langle h, q_j\rangle_s|^2 \leq M^2$ for all $h \in \mathcal{H}_{os}$; or (b) (hyperrectangle) $|\langle h, q_j\rangle_s| \leq \nu_j^{-\gamma_h}$ and $\sum_{j=1}^{\infty}\nu_j^{-2\gamma_h} < \infty$ for all $h \in \mathcal{H}_{os}$.

ASSUMPTION 5.4. There are finite constants $c, C > 0$ and a non-increasing positive sequence $\{b_j \asymp \varphi(\nu_j^{-2})\}_{j=1}^{\infty}$ such that: (i) $||h||^2 \geq c\sum_{j=1}^{\infty}b_j|\langle h, q_j\rangle_s|^2$ for all $h \in \mathcal{H}_{os}$; (ii) $||h||^2 \leq C\sum_{j=1}^{\infty}b_j|\langle h, q_j\rangle_s|^2$ for all $h \in \mathcal{H}_{os}$.

It is obvious that assumption 5.4(i) and (ii) implies assumption 5.2(i) and (ii) respectively.

LEMMA 5.2. Let assumptions 5.1, 5.4(i) and 5.3 hold. Then: there is an integer $k^* \in (1, \infty)$ such that $\delta^2/b_{k^*-1} < M^2(\nu_{k^*})^{-2\gamma_h}$ and $\delta^2/b_{k^*} \geq M^2(\nu_{k^*})^{-2\gamma_h}$; hence $\omega(\delta, \mathcal{H}_{os}) \leq const. \times \delta/\sqrt{b_{k^*}}$.

   (1) If $b_j \asymp \varphi(\nu_j^{-2}) = \nu_j^{-2a}$ then $\omega(\delta, \mathcal{H}_{os}) \leq const.(\delta^{\gamma_h/(a+\gamma_h)})$.

   (2) If $b_j \asymp \varphi(\nu_j^{-2}) = \exp\{-\nu_j^a\}$ then $\omega(\delta, \mathcal{H}_{os}) \leq const. \left([-\ln(\delta)]^{-\gamma_h/a}\right)$.

Assumptions 5.4(i) and 5.3 also yield the following better bound on the sieve modulus of continuity $\omega_n(\delta, \mathcal{H}_{osn})$.

LEMMA 5.3. Let $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ and assumptions 5.1(i), 5.4(i) and 5.3 hold. Let k^* be given in Lemma 5.2. Then:

   (1) $\omega_n(\delta, \mathcal{H}_{osn}) \leq const. \times \delta/\sqrt{b_{\overline{k}}}$, where $\overline{k} \equiv \min\{k(n), k^*\} \in (1, \infty)$.

   (2) If $k(n) \geq k^*$, then $||h - \Pi_n h||_s \leq \delta/\sqrt{b_{k^*}}$ for all $h \in \mathcal{H}_{os}$.

5.2 Optimal convergence rates

Theorem 4.2, Theorem 4.3 and Lemma 5.1 together imply the following corollary for the convergence rate for the penalized SMD estimator using a finite-dimensional sieve:

COROLLARY 5.1. *(Sieve dominating case) Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n \geq 0$, $\lambda_n = o(1)$, and all the assumptions of Theorem 4.1(1) hold with $\widehat{m}(X, h)$ being the series LS estimator. Let assumptions 5.1 and 5.2 hold, with $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ and $b_j \asymp \varphi(\nu_j^{-2})$ for a continuous non-decreasing function φ. If $\max\left\{\frac{J_n}{n}, b_{m,J_n}^2, \lambda_n\right\} = \frac{J_n}{n} \to 0$ and $\lim_{n \to \infty}\{J_n/k(n)\} = c \in (1, \infty)$, then:*

$$||\hat{h}_n - h_0||_s = O_P\left(\{\nu_{k(n)}\}^{-\gamma_h} + \sqrt{\frac{k(n)}{n \times \varphi(\nu_{k(n)}^{-2})}}\right).$$

(1) Mildly ill-posed case: if $\varphi(\tau) = \tau^a$ for some $a \geq 0$ and $\nu_k \asymp k^{1/d}$, then: $||\hat{h}_n - h_0||_s = O_P\left(n^{-\frac{\gamma_h}{2(\gamma_h + a) + d}}\right)$ provided $k(n) = O\left(n^{\frac{d}{2(\gamma_h + a) + d}}\right)$.

(2) Severely ill-posed case: if $\varphi(\tau) = \exp\{-\tau^{-a/2}\}$ for some $a > 0$ and $\nu_k \asymp k^{1/d}$, then: $||\hat{h}_n - h_0||_s = O_P\left([\ln(n)]^{-\gamma_h/a}\right)$ provided $k(n) = O\left([\ln(n)]^{d/a}\right)$.

Corollary 5.1 allows for $\lambda_n = 0$. The next corollary allows for all the three smoothing parameters $(J_n, k(n), \lambda_n)$ to balance one another.

COROLLARY 5.2. *(Sieve penalization balance case) Under all the conditions of Corollary 5.1, if either assumption 3.13(i) holds with $\lambda_n = O(\frac{J_n}{n})$, or assumption 4.2 holds with $\lambda_n = O\left(\sqrt{\frac{J_n}{n}}\sqrt{\varphi(\nu_{k(n)}^{-2})}\right)$, then: all the conclusions of Corollary 5.1 remain true.*

Theorem 4.3, Lemma 5.2 and Lemma 5.3 immediately imply the following corollary for the convergence rate for the penalized SMD estimator using either a finite dimensional sieve with lots of sieve terms or an infinite dimensional sieve.

COROLLARY 5.3. *(Penalization dominating case) Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n > 0$, $\lambda_n = o_P(1)$, and all the assumptions of Theorem 4.1(1) hold with $\widehat{m}(X, h)$ being the series LS estimator. Let assumptions 5.1(i), 5.2(ii), 5.4(i) and 5.3 hold. Let $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ with $k(n) > k^*$ given in Lemma 5.2, and $b_j \asymp \varphi(\nu_j^{-2})$ for a continuous non-decreasing function φ. Let $\frac{J_n}{n} \asymp b_{m,J_n}^2 = O(J_n^{-2r_m})$ for some $r_m > 0$. If either assumption 3.13(i) holds with $\lambda_n = O(\frac{J_n}{n})$, or assumption 4.2 holds with $\lambda_n = O\left(\sqrt{\frac{J_n}{n}}\sqrt{\varphi(\nu_{k^*}^{-2})}\right)$, then:*

$$||\hat{h}_n - h_0||_s = O_P\left(\frac{n^{-r_m/(2r_m + 1)}}{\sqrt{\varphi(\nu_{k^*}^{-2})}}\right).$$

(1) Mildly ill-posed case: if $\varphi(\tau) = \tau^a$ for some $a \geq 0$ and $\nu_k \asymp k^{1/d}$, then: (1.i) $||\hat{h}_n - h_0||_s = O_P\left(n^{-\frac{r_m}{2r_m + 1}\frac{\gamma_h}{(\gamma_h + a)}}\right)$; (1.ii) if 5.4(ii) holds, then $r_m \geq (\gamma_h + a)/d$ and $||\hat{h}_n - h_0||_s = O_P\left(n^{-\frac{\gamma_h}{2(\gamma_h + a) + d}}\right)$.

(2) Severely ill-posed case: if $\varphi(\tau) = \exp\{-\tau^{-a/2}\}$ for some $a > 0$, then: $||\widehat{h}_n - h_0||_s = O_P\left([\ln(n)]^{-\gamma_h/a}\right)$.

(3) Further, if $\mathcal{H}_n = \mathcal{H}$ (or $k(n) = \infty$), then assumption 5.2(ii) holds, and all the above conclusions remain true.

We note that for the mildly ill-posed case (i.e., when $b_j \asymp \varphi(\nu_j^{-2}) = \nu_j^{-2a}$ for a finite $a \geq 0$), assumptions 4.1, 5.4(i)(ii) and 5.3 imply the restriction $r_m \geq (\gamma_h + a)/d$. See Yang and Barron (1999). In the following we denote $\frac{dm(X,h_0)}{dh}[a]$ as $T_{h_0}[a]$, where $T_{h_0} : \mathcal{H}_{os} \subset \mathbf{H} \to L^2(f_X)$ and $T_{h_0}^*$ as its adjoint (under the inner product, $\langle \cdot, \cdot \rangle$ associated with the weak metric $|| \cdot ||$). Then for all $h \in \mathcal{H}_{os}$, we have $||h||^2 \equiv ||T_{h_0}h||^2_{L^2(f_X)} \asymp ||(T_{h_0}^* T_{h_0})^{1/2}h||^2_s$ by assumption 4.1.

Remark 5.2. *Under assumptions 5.1, 5.4(ii) and 5.3, Chen and Reiss (2007) establish the minimax lower bound for estimation of the NPIV model (1.2): $E[Y_1 - h_0(Y_2)|X] = 0$. When we specialize Corollaries 5.1, 5.2 and 5.3 to the NPIV model (1.2), our rates coincide with their minimax lower bound under the metric $||h||_s = ||h||_{L^2(f_{Y_2})}$. In particular, our assumptions 5.1 and 5.3 correspond to their approximation condition. Let B be a self-adjoint unbounded operator defined as: $Bh = \sum_{j=1}^{\infty} \nu_j \langle h, q_j \rangle_{L^2(f_{Y_2})} q_j$ with $Dom(B) = \{h \in L^2(f_{Y_2}) : \sum_{j=1}^{\infty} \nu_j^2 \langle h, q_j \rangle^2_{L^2(f_{Y_2})} < \infty\}$. Then our assumption 5.3 implies $\mathcal{H}_{os} \subseteq Dom(B^{\gamma_h}) = \{h \in L^2(f_{Y_2}) : B^{\gamma_h} h = \sum_{j=1}^{\infty} \nu_j^{\gamma_h} \langle h, q_j \rangle_{L^2(f_{Y_2})} q_j \in L^2(f_{Y_2})\}$. Our assumption 5.4(ii) becomes their link condition: $||T_{h_0}h||^2_{L^2(f_X)} \leq C \sum_{j=1}^{\infty} b_j |\langle h, q_j \rangle_{L^2(f_{Y_2})}|^2$ with $b_j \asymp \varphi(\nu_j^{-2})$ for a continuous increasing function φ. Then the rates obtained in Corollaries 5.1, 5.2 and 5.3 reach the minimax lower bound for the NPIV model (1.2) in Chen and Reiss (2007).*

5.3 Relation to source condition

Under assumption 4.1, we have $||h||^2 \asymp ||(T_{h_0}^* T_{h_0})^{1/2}h||^2_s$ for all $h \in \mathcal{H}_{os}$; hence assumption 5.4 can be restated in terms of the operator $T_{h_0}^* T_{h_0}$. Assuming that T_{h_0} is a compact operator (this is a mild condition, for example, T_{h_0} is compact if $m(\cdot, h) : \mathcal{H} \subseteq \mathbf{H} \to L^2(f_X)$ is compact and is Frechet differentiable at $h_0 \in \mathcal{H}_{os}$; see Zeidler (1985, proposition 7.33)).[9] Then T_{h_0} has a singular value decomposition $\{\mu_k; \phi_{1k}, \phi_{0k}\}_{k=1}^{\infty}$, where $\{\mu_k\}_{k=1}^{\infty}$ are the singular numbers arranged in non-increasing order ($\mu_k \geq \mu_{k+1} \searrow 0$), $\{\phi_{1k}()\}_{k=1}^{\infty}$ and $\{\phi_{0k}(x)\}_{k=1}^{\infty}$ are eigenfunctions of the operators $(T_{h_0}^* T_{h_0})^{1/2}$ and $(T_{h_0} T_{h_0}^*)^{1/2}$ respectively. It is obvious that $\{\phi_{1k}()\}_{k=1}^{\infty}$ is an orthonormal basis for $\mathbf{H}$ hence a Riesz basis, and $||(T_{h_0}^* T_{h_0})^{1/2}h||^2_s = \sum_{k=1}^{\infty} \mu_k^2 |\langle h, \phi_{1k} \rangle_s|^2$ for all $h \in \mathbf{H}$. In the numerical analysis literature on ill-posed inverse problems with known operators, it is common to measure the smoothness of the function class $\mathcal{H}_{os}$ in terms of the spectral representation of $T_{h_0}^* T_{h_0}$. The so-called "*general source condition*" assumes that there is a continuous function ψ with $\psi(0) = 0$

---

[9]See Bissantz, et al (2007) for convergence rates of statistical linear ill-posed inverse problems via the Hilbert scale (or general source condition) approach for possibly non-compact but known operators.

and $\lambda^{-1/2}\psi(\lambda)$ non-decreasing such that

$$\mathcal{H}_{source} \equiv \{h = \psi(T_{h_0}^* T_{h_0})v : \quad v \in \mathbf{H}, ||v||_s^2 \leq M\} \tag{5.1}$$

$$= \left\{h = \sum_{j=1}^{\infty} \langle h, \phi_{1j}\rangle_s \phi_{1j} : \quad \sum_{j=1}^{\infty} \frac{\langle h, \phi_{1j}\rangle_s^2}{\psi^2(\mu_j^2)} \leq M\right\}, \tag{5.2}$$

for a finite constant M, and the original "source condition" corresponds to the choice $\psi(\lambda) = \lambda^{1/2}$ (see Engl, Hanke and Neubauer (1996)). Therefore the general source condition implies our assumptions 5.1(i), 5.4 and 5.3 by setting $q_j = \phi_{1j}$, $b_j \asymp \varphi(\nu_j^{-2}) = \mu_j^2$ and $\psi(\mu_j^2) = \nu_j^{-\gamma_h}$ for all $j \geq 1$. Then $\varphi(\tau) = \tau^a$ is equivalent to $\psi(\lambda) = \lambda^{\gamma_h/(2a)}$ and $\lambda^{-1/2}\psi(\lambda)$ non-decreasing iff $\gamma_h \geq a$; $\varphi(\tau) = \exp\{-\tau^{-a/2}\}$ is equivalent to $\psi(\lambda) = [-\log(\lambda)]^{-\gamma_h/a}$ We gather these simple results into the next lemma:

LEMMA 5.4. *Let T_{h_0} be a compact operator with a singular value decomposition $\{\mu_k; \phi_{1k}, \phi_{0k}\}_{k=1}^{\infty}$. Then: (1) assumptions 5.1(i) and 5.4 hold with $q_j = \phi_{1j}$ and $b_j \asymp \varphi(\nu_j^{-2}) = \mu_j^2$ for all j. In addition, if $\mathcal{H}_{os} \subseteq \mathcal{H}_{source}$, then assumption 5.3 holds with $\nu_j^{-\gamma_h} \geq \psi(\mu_j^2)$ for all j.*

6 Applications

In this section we present two important applications to illustrate the general results obtained in the previous sections. We first provide sufficient conditions for consistency and convergence rate of the penalized SMD estimators for a nonparametric additive quantile IV regression model. We then obtain root-n asymptotic normality of a plug-in penalized SMD estimator of a weighted average derivative of $h_0(Y_2)$ satisfying the conditional moment model $E[\rho(Y, X_z; h_0(Y_2))|X] = 0$, in which the residual function $\rho()$ could be non-pointwise smooth with respect to $h(Y_2)$.

6.1 Nonparametric Additive Quantile IV Regression Model

The model is:

$$Y_3 = h_{01}(Y_1) + h_{02}(Y_2) + U, \quad \Pr(U \leq 0|X) = \gamma, \tag{6.1}$$

where h_{01}, h_{02} are the unknown functions of interest, the conditional distribution of the error term U given X is unspecified, except that $F_{U|X}(0) = \gamma$ for a known fixed $\gamma \in (0,1)$. The support of $Y = (Y_1', Y_2', Y_3)'$ is $\mathcal{Y} = [0,1]^d \times \mathcal{R}^d \times \mathcal{Y}_3$ with $\mathcal{Y}_3 \subseteq \mathcal{R}$, and the support of X is $\mathcal{X} = [0,1]^{d_x}$ with $d_x \geq d \geq 1$. To map into the general model (1.4), we let $Z = (Y', X')'$, $h = (h_1, h_2)$, $\rho(Z, h) = 1\{Y_3 \leq h_1(Y_1) + h_2(Y_2)\} - \gamma$ and $m(X, h) = E[F_{Y_3|Y_1, Y_2, X}(h_1(Y_1) + h_2(Y_2))|X] - \gamma$.

For the sake of concreteness and illustration, we estimate $h_0()$ using the penalized SMD estimator $\widehat{h}_n$ given in (2.3), with $\widehat{m}(X, h)$ being the series LS estimator of $m(X, h)$, $\mathcal{H}_n = \mathcal{H}_n^1 \times \mathcal{H}_n^2$ being either a finite dimensional $(\dim(\mathcal{H}_n) \equiv k(n) = k_1(n) + k_2(n) < \infty)$ or an infinite dimensional $(k(n) = \infty)$ *linear* sieve, and $\widehat{P}_n(h) = P(h_2) \geq 0$.

27

We present three propositions on consistency. The first one assumes that the function space $\mathcal{H}^2$ is compact under a weighted sup norm $||h_2||_{w,\infty} = \sup_{y\in\mathcal{R}^d} |h_2(y)\,w(y)|$ for a positive continuous weight w. The second and the third do not assume compactness of $\mathcal{H}^2$; while the second one considers a lower semicompact penalty and the third uses a convex (but not lower semicompact) penalty. For all three results we assume:

CONDITION 6.1. *(i)* $\{(Y_i', X_i')\}_{i=1}^n$ *is i.i.d.; (ii)* $f_{Y_3|Y_1,Y_2,X}(y_3|y_1,y_2,x)$ *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; *(iii)* $E[(1+|Y_2|)^\theta] < \infty$ *for a finite* $\theta > 0$; *(iv)* $f_{Y_1,Y_2|X=x}(y_1, y_2)$ *is continuous in* (y_1, y_2, x)

CONDITION 6.2. $h_0 = (h_{01}, h_{02}) \in \mathcal{H} = \mathcal{H}^1 \times \mathcal{H}^2$, *(i)* $\mathcal{H}^1 = \{h_1 \in \Lambda_1^{\gamma_1}([0,1]^d) : h_1(y_1^*) = 0\}$ *for* $\gamma_1 > 0$; *(ii)* $E[1\{Y_3 \leq h_1(Y_1) + h_2(Y_2)\}|X] = \gamma$ *for* $h = (h_1, h_2) \in \mathcal{H}$ *implies* $h_1(Y_1) + h_2(Y_2) = h_{01}(Y_1) + h_{02}(Y_2)$ *almost surely; (iii)* $\mathcal{H}^2 \subset L^2(\mathcal{R}^d, f_{Y_2})$.

Condition 6.2(i)(ii) is a global identification condition. Instead of condition 6.2(i), one could assume condition 6.2(i)': $\mathcal{H}^1 = \Lambda_1^{\gamma_1}([0,1]^d)$ *for* $\gamma_1 > 0$, *and the conditional expectation operator,* $E[h_2(Y_2)|Y_1]$, *mapping from* $\mathcal{H}^2$ *to* $L^2([0,1]^d, f_{Y_1})$ *is compact*. Without unknown h_{01}, condition 6.2(ii) is the global identification condition for h_{02} in the NPQIV model (1.3) that is imposed in CIN (2007) and HL (2007). See CIN (2007) and Chernozhukov and Hansen (2005) for further discussion and sufficient conditions for identification of NPQIV model (1.3).

CONDITION 6.3. *(i) assumption 3.9 holds with* $p^{J_n}(X)$ *being a tensor product P-spline or B-spline or wavelet or cosine linear sieves; (ii)* $\mathcal{H}_n = \mathcal{H}_n^1 \times \mathcal{H}_n^2$, *where* $\mathcal{H}_n^1$ *is a tensor product P-spline or B-spline or wavelet or cosine or power series closed linear subspace of* $\mathcal{H}^1$, *and* $\mathcal{H}_n^2$ *is a tensor product wavelet closed linear subspace of* $\mathcal{H}^2$.

In the following we denote $\varpi(y_2) \equiv (1 + |y_2|^2)^{-\vartheta/2}$ for some $\vartheta \geq 0$ and $w(y_2) \equiv (1 + |y_2|^2)^{-\theta/2}$ for some $\theta > 0$. We let $||\cdot||_{\mathcal{T}_{p,q}^\gamma}$ denote the norm of a Banach space $\mathcal{T}_{p,q}^\gamma(\mathcal{R}^d, leb)$, which is either a Besov space $\mathcal{B}_{p,q}^\gamma(\mathcal{R}^d, leb)$ for $p, q \in [1, \infty]$ or a F-space $\mathcal{F}_{p,q}^\gamma(\mathcal{R}^d, leb)$ for $p \in [1, \infty), q \in [1, \infty]$; see Appendix A for their definitions and properties.

PROPOSITION 6.1. *For the model (6.1), let* $\hat{h}_n$ *be the penalized SMD estimator with* $\lambda_n \geq 0$, $\lambda_n = o(1)$ *and* $\widehat{m}(X, h)$ *be the series LS estimator. Let conditions 6.1, 6.2 and 6.3 hold. Let* $\mathcal{H}^2 = \{h_2 \in L^2(\mathcal{R}^d, f_{Y_2}) : ||\varpi h_2||_{\mathcal{T}_{p,q}^{\gamma_2}} \leq M_0\}$ *for* $\gamma_2 > 0, p, q \in [1, \infty]$ *(and* $p < \infty$ *for* $\mathcal{T}_{p,q}^{\gamma_2} = \mathcal{F}_{p,q}^{\gamma_2}$), *a known constant* $M_0 < \infty$. *Let* $\lambda_n P(h_2) = \lambda_n ||\varpi h_2||_{\mathcal{T}_{p_2,q_2}^{s_2}}$ *for* $s_2 \in [0, \gamma_2 - d(p^{-1} - p_2^{-1})], p_2 \in [1, \infty)$ *and* $q \leq q_2 \leq \infty$. *Let* $J_n/n = o(1)$ *and* $J_n, k_1(n), k_2(n) \to \infty$ *as* $n \to \infty$.

(1) If $\gamma_2 > d/p$ *and* $\theta > \vartheta \geq 0$, *then:*

$$\sup_{y_1 \in [0,1]^d} \left|\widehat{h}_{1,n}(y_1) - h_{01}(y_1)\right| + \sup_{y_2 \in \mathcal{R}^d} \left|w(y_2)\left[\widehat{h}_{2,n}(y_2) - h_{02}(y_2)\right]\right| = o_P(1);$$

hence if $E[(1+|Y_2|)^{2\theta}] < \infty$ *then:* $||\widehat{h}_{1,n} - h_{01}||_{L^2(f_{Y_1})} + ||\widehat{h}_{2,n} - h_{02}||_{L^2(f_{Y_2})} = o_P(1)$.

(2) If $\gamma_2 + d/2 > d/p$, $p^{-1} + (\theta - \vartheta)/d > 1/2$, and $E[(1 + |Y_2|)^{2\theta}] < \infty$, then:

$$\sup_{y_1 \in [0,1]^d} \left| \widehat{h}_{1,n}(y_1) - h_{01}(y_1) \right| + ||w[\widehat{h}_{2,n} - h_{02}]||_{L^2(\mathcal{R}^d, leb)} = o_P(1);$$

hence $||\widehat{h}_{1,n} - h_{01}||_{L^2(f_{Y_1})} + ||\widehat{h}_{2,n} - h_{02}||_{L^2(f_{Y_2})} = o_P(1)$.

We now present a second consistency result in which the parameter space $\mathcal{H}^2$ is not compact but the penalty is lower semicompact. We assume:

CONDITION 6.4. *$E[F_{Y_3|Y_1,Y_2,X}(h_1(Y_1) + h_2(Y_2))|X = \cdot] \in W_{2,c}^{\gamma_m}([0,1]^{d_x}, leb)$ with $\gamma_m > 0$ for all $h \in \mathcal{H}_n$.*

In the following we denote $r_m \equiv \gamma_m/d_x$, $r_1 \equiv \gamma_1/d$ and $r_2 \equiv \gamma_2/d$.

PROPOSITION 6.2. *For the model (6.1), let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n > 0$, $\lambda_n = o(1)$ and $\widehat{m}(X, h)$ be the series LS estimator. Let conditions 6.1, 6.2, 6.3 and 6.4 hold. Let $\mathcal{H}^2 = \{h_2 \in L^2(\mathcal{R}^d, f_{Y_2}) : ||\varpi h_2||_{\mathcal{T}_{p,q}^{\gamma_2}} < \infty\}$ for $\gamma_2 > 0, p, q \in [1, \infty]$ (and $p < \infty$ for $\mathcal{T}_{p,q}^{\gamma_2} = \mathcal{F}_{p,q}^{\gamma_2}$). Let $P(h_2) = ||\varpi h_2||_{\mathcal{T}_{p,q}^{\gamma_2}}$. Let $\max\{[k_1(n)]^{-2r_1}, [k_2(n)]^{-2r_2}\} = O(\lambda_n)$ and $\frac{J_n}{n} + J_n^{-2r_m} = O(\lambda_n)$. Then: (1) Results (1) and (2) of Proposition 6.1 remain true; (2) $P(\widehat{h}_{2,n}) = O_P(1)$.*

We next present a third consistency result in which the parameter space $\mathcal{H}^2$ is not compact but the penalty is convex. We assume:

CONDITION 6.5. *Condition 6.4 holds for all $h \in \mathcal{H}$.*

PROPOSITION 6.3. *For the model (6.1), let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n > 0$, $\lambda_n = o(1)$ and $\widehat{m}(X, h)$ be the series LS estimator. Let conditions 6.1, 6.2, 6.3 and 6.5 hold. Let $\mathcal{H}^2 = \{wh_2 \in W_2^{\gamma_2}(\mathcal{R}^d, leb) : ||wh_2||_{L^2(leb)} \le M\}$ for $\gamma_2 > 0$, $P(h) = ||(wh_2)||_{L^2(\mathcal{R}^d, leb)}^2$, and $E[(1 + |Y_2|)^{2\theta}] < \infty$. Let $\max\{[k_1(n)]^{-2r_1}, [k_2(n)]^{-2r_2}\} = o(\lambda_n)$ and $\frac{J_n}{n} + J_n^{-2r_m} = o(\lambda_n)$. Then:*

$$\sup_{y_1 \in [0,1]^d} \left| \widehat{h}_{1,n}(y_1) - h_{01}(y_1) \right| + ||w[\widehat{h}_{2,n} - h_{02}]||_{L^2(\mathcal{R}^d, leb)} = o_P(1),$$

and $||\widehat{h}_{1,n} - h_{01}||_{L^2(f_{Y_1})} + ||\widehat{h}_{2,n} - h_{02}||_{L^2(f_{Y_2})} = o_P(1)$.

Without unknown h_1, Proposition 6.1 is essentially the same as theorem 4.1 of CIN (2007) for the SMD estimator (2.2) of the NPQIV model (1.3); while Proposition 6.3 is very similar to that of HL (2007) except that we allow for sieve approximation and unbounded support of Y_2.

For the model (6.1), we have $||h||_c = ||h_1||_{L^2(f_{Y_1})} + ||h_2||_{L^2(f_{Y_2})}$. Let $||h||_s^2 = E\{[h_1(Y_1) + h_2(Y_2)]^2\}$, then $||h||_s \le ||h||_c$ for all $h \in \mathcal{H}$. Recall that $\mathcal{H}_{os} \equiv \{h = (h_1, h_2) \in \mathcal{H} : ||h - h_0||_c = o(1), ||h||_c \le c, P(h) \le c\}$. For any $h \in \mathcal{H}_{os}$ define the linear integral operator $T_h[g_1 + g_2] \equiv E\{f_{Y_3|Y_1,Y_2,X}(h_1(Y_1) + h_2(Y_2))[g_1(Y_1) + g_2(Y_2)]|X = \cdot\}$ that maps from $Dom(T_h) \subset L^2(f_{Y_1}) \oplus L^2(f_{Y_2}) \to L^2([0,1]^{d_x}, f_X)$. Let $\mathcal{B}(Dom(T_h), L^2([0,1]^{d_x}, f_X))$ denote the class of all bounded linear

operators from $Dom(T_h)$ to $L^2([0,1]^{d_x}, f_X)$. The j-th approximation number $a_j(T_h)$ of T_h is defined as (see Edmunds and Triebel (1996)):

$$a_j(T_h) \equiv \inf \left\{ \sup_{g \in Dom(T_h)} \frac{||T_h[g] - L[g]||_{L^2(f_X)}}{||g||_s} : L \in \mathcal{B}\left(Dom(T_h), L^2([0,1]^{d_x}, f_X)\right), \dim(Range(L)) < j \right\}.$$

We assume

CONDITION 6.6. *(i) Condition 6.4 holds for all $h \in \mathcal{H}_{os}$; (ii) $f_{Y_3|Y_1,Y_2,X}(y_3|y_1,y_2,x)$ has continuous derivative $f'_{Y_3|Y_1,Y_2,X}(y_3|y_1,y_2,x)$ with respect to y_3, and $\sup_{y_3,y_1,y_2,x} |f'_{Y_3|Y_1,Y_2,X}(y_3|y_1,y_2,x)| \leq$ const. $< \infty$; (iii) there are finite constants $c, C > 0$ such that $ca_j(T_{h_0}) \leq a_j(T_h) \leq Ca_j(T_{h_0})$ for all $j \geq 1$ and for all $h \in \mathcal{H}_{os}$.*

CONDITION 6.7. *If $E\{f_{Y_3|Y_1,Y_2,X}(h_{01}(Y_1) + h_{02}(Y_2))[\Delta_1(Y_1) + \Delta_2(Y_2)]|X\} = 0$ then $\Delta_1(Y_1) + \Delta_2(Y_2) = 0$ for all $\Delta_1(Y_1)$, $\Delta_2(Y_2)$ such that $\Delta + h_0 \in \mathcal{H}_{os}$.*

CONDITION 6.8. *(i) Y_1 and Y_2 are independent; (ii) there is a non-negative, continuous increasing function φ such that $||T_{h_0}[g_1 + g_2]||^2_{L^2(f_X)} \asymp \sum_{j=1}^\infty \varphi(\nu_j^{-2})|\langle g_1 + g_2, q_{1,j} + q_{2,j}\rangle_s|^2$ for all $g_1 + g_2 \in Dom(T_{h_0}) \cap \mathcal{H}_{os}$.*

In the following we denote $\gamma_h \equiv \min\{\gamma_1, \gamma_2\}$. We first present a convergence rate result for the finite dimensional sieve dominating case.

PROPOSITION 6.4. *(Sieve dominating case) For the model (6.1), suppose that conditions 6.6 - 6.8 hold. Let either conditions of Proposition 6.2 hold with $\max\left\{\frac{J_n}{n}, J_n^{-2r_m}, \lambda_n\right\} = \frac{J_n}{n}$, or conditions of Proposition 6.3 hold with $\max\left\{\frac{J_n}{n}, J_n^{-2r_m}, o(\lambda_n)\right\} = \frac{J_n}{n}$. Let $\lim_{n\to\infty}\{J_n/k(n)\} = c \in [1,\infty)$, $k(n) = k_1(n) + k_2(n) \to \infty$, $k_1(n) \asymp k_2(n) \asymp k(n)$ and $\nu_{k(n)} = \{k(n)\}^{1/d}$. Then:*

$$||\hat{h}_n - h_0||_s = O_P\left(\{\nu_{k(n)}\}^{-\gamma_h} + \sqrt{\frac{k(n)}{n \times \varphi(\nu_{k(n)}^{-2})}}\right).$$

(1) If $\varphi(\tau) = \tau^a$ for some $a \geq 0$, then: $||\hat{h}_n - h_0||_s = O_P\left(n^{-\frac{\gamma_h}{2(\gamma_h+a)+d}}\right)$ provided $k(n) = O\left(n^{\frac{d}{2(\gamma_h+a)+d}}\right)$.

(2) If $\varphi(\tau) = \exp\{-\tau^{-a/2}\}$ for some $a > 0$, then: $||\hat{h}_n - h_0||_s = O_P\left([\ln(n)]^{-\gamma_h/a}\right)$ provided $k(n) = O\left([\ln(n)]^{d/a}\right)$.

Next we provide convergence rate for the penalization dominating case.

PROPOSITION 6.5. *(Penalization dominating case) For the model (6.1), suppose that conditions 6.6 - 6.8 hold. Let either conditions of Proposition 6.2 hold with $\lambda_n \asymp \frac{J_n}{n}$, or conditions of Proposition 6.3 hold with $\lambda_n = O\left(\sqrt{\frac{J_n}{n}}\sqrt{\varphi(\nu_{J_n}^{-2})}\right)$ and $\mathcal{H}_{os}^2 = \{wh_2 \in W_2^{\gamma_2}(\mathcal{R}^d, leb) : ||\varpi h_2||_{W_2^{\gamma_2}(\mathcal{R}^d,leb)} \leq M\}$ for $\theta > \vartheta \geq 0$. Let $\frac{J_n}{n} \asymp J_n^{-2r_m}$, $\min\{k_1(n), k_2(n)\} \geq J_n$ and $\nu_j = j^{1/d}$.*

(1) If $\varphi(\tau) = \tau^a$ for some $a \geq 0$, then: $||\hat{h}_n - h_0||_s = O_P\left(n^{-\frac{\gamma_h}{2(\gamma_h+a)+d}}\right)$.

(2) If $\varphi(\tau) = \exp\{-\tau^{-a/2}\}$ for some $a > 0$, then: $||\hat{h}_n - h_0||_s = O_P\left([\ln(n)]^{-\gamma_h/a}\right)$.

30

When Y_1 and Y_2 are measurable with respect to X, we have $a = 0$ in Propositions 6.4(1) and 6.5(1). The resulting convergence rates $||\widehat{h}_n - h_0||_s = O_P\left(n^{-\frac{\gamma_h}{2\gamma_h + d}}\right)$ coincide with the known rates for the additive quantile regression model: $Y_3 = h_{01}(X_1) + h_{02}(X_2) + U$, $\Pr(U \leq 0|X_1, X_2) = \gamma$; see, e.g., Horowitz and Mammen (2007).

6.2 Plug-in Penalized SMD of Weighted Average Derivatives of $h_0(Y)$.

The model is:

$$E[\rho(Y, X_z; h_0(Y_2))|X] = 0 \text{ and } \beta_0 = E[a(Y_2)\nabla^k h_0(Y_2)], \tag{6.2}$$

where $Y_2 \subseteq Y$, $k \geq 1$ is a known integer, and $a(Y_2)$ is a known positive weight function. For simplicity we assume that $h()$ and $\rho_2()$ are scalar valued, and Y_2 has support $\mathcal{R}$. Results presented in this subsection can be directly extended to vector valued $h()$ and $\rho_2()$ as well as multivariate Y_2; see, e.g., Chen and Pouzo (2007). Let $\widehat{h}_n()$ be a penalized SMD estimator of $h_0()$ based on the conditional moment restriction $E[\rho(Y, X_z; h_0(Y_2))|X] = 0$. Then the plug-in penalized SMD estimator of β_0 is simply defined as

$$\widehat{\beta}_n = n^{-1} \sum_{i=1}^{n} a(Y_{2i})\nabla^k \widehat{h}_n(Y_{2i}).$$

When $\rho(Y, X_z; h(Y_2))$ is pointwise Hölder continuous with respect to h, this model (6.2) fits into the general setup of Ai and Chen (2007). In fact, Ai and Chen (2007) already present sufficient conditions to ensure root-n asymptotic normality of their plug-in SMD estimator of β_0 when $\rho(Y, X_z; h_0(Y_2)) = Y_1 - h_0(Y_2)$ (i.e., the NPIV model). However, the general results in Ai and Chen (2007) rule out the cases when $\rho(Y, X_z; h(Y_2))$ is not pointwise Hölder continuous in h. In this subsection, we obtain a theorem that allows for nonlinear and non-pointwise smooth $\rho(Y, X_z; h(Y_2))$ in h, as well as non $||\cdot||_s-$compactness of the function space $\mathcal{H}$ for h_0.

For the sake of concreteness we will only consider the penalized SMD estimator $\widehat{h}_n$ using a finite dimensional linear sieves $(\dim(\mathcal{H}_n) \equiv k(n) < \infty)$, $P(h) = ||\nabla^{\overline{k}}(\varpi h)||_{L^2(\mathcal{R}, leb)}^2$ for some $\overline{k} > k$, and the series LS estimator $\widehat{m}(X, h)$ of $m(X, h) = E[\rho(Y, X_z; h(Y_2))|X]$. We shall only present some relatively low level sufficient conditions for the consistency and asymptotic normality of $\widehat{\beta}_n$, and refer readers to Chen and Pouzo (2007, 2008) for more general conditions. Given the results presented in previous sections, we can assume that we already established consistency and convergence rate for $\widehat{h}_n$ under $||\cdot||_s = ||\cdot||_{L^2(f_{Y_2})}$. In this subsection we focus on the consistency and asymptotic normality of $\widehat{\beta}_n$.

CONDITION 6.9. *(i)* $h_0 \in \mathcal{H} \subseteq \{h \in L^2(\mathcal{R}, f_{Y_2}) : ||\varpi h||_{W_2^{\gamma_h}(leb)} < \infty\}$, $P(h) = ||\nabla^{\overline{k}}(\varpi h)||_{L^2(\mathcal{R}, leb)}^2$ *for some* $\gamma_h \geq \overline{k} > k$; *(ii)* $E[(1 + |Y_2|)^{2\theta}] < \infty$ *for some* $\theta > \vartheta \geq 0$. *(iii)* $\mathcal{H}_n$ *is a wavelet linear sieve of* $\mathcal{H}$; *(iv)* $E\{m(X, h)^2\}$ *is continuous at* h_0 *under the norm* $||\cdot||_c = ||\cdot||_{L^2(f_{Y_2})}$.

CONDITION 6.10. (i) $\beta_0 \in B$, where B is a closed bounded interval of $\mathcal{R}$; (ii) $\sup_{y_2} |a(y_2)| \leq$ const. $< \infty$.

PROPOSITION 6.6. For the model (6.2), let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n > 0$, $\lambda_n = o(1)$ and $\widehat{m}(X,h)$ be the series LS estimator. Suppose that assumptions 3.9 - 3.10, 3.1(i)(iii), 3.7(ii), and condition 6.9 hold. Let $\max\left\{\frac{J_n}{n} + b_{m,J_n}^2, E\{m(X,\Pi_n h_0)^2\}\right\} = O(\lambda_n) = o(1)$. Then:

(1) (i) $||w[\hat{h}_n - h_0]||_{L^2(\mathcal{R},leb)} = o_P(1)$, $||\hat{h}_n - h_0||_{L^2(f_{Y_2})} = o_P(1)$, $P(\hat{h}_n) = O_P(1)$; and (ii) $||\nabla^k(w[\hat{h}_n - h_0])||_{L^2(\mathcal{R},leb)} = o_P(1)$, $||\nabla^k[\hat{h} - h_0]||_{L^2(f_{Y_2})}^2 = o_P(1)$.

(2) If condition 6.10 holds, then $|\hat{\beta}_n - \beta_0| = o_P(1)$.

We need some extra notations before we state normality result. Define $\mathcal{H}_{os} \equiv \{h \in \mathcal{H} : ||h - h_0||_s = o(1), ||\nabla^k[h - h_0]||_{L^2(f_{Y_2})} = o(1), ||h||_s \leq c, P(h) \leq c\}$. Then $h_0 \in \mathcal{H}_{os}$ and $\mathcal{H}_{os} \subset \mathcal{H} \subset L^2(\mathcal{R}, f_{Y_2})$, and $||h - h_0||^2 \equiv E\left[\left(\frac{dm(X,h_0)}{dh}[h - h_0]\right)^2\right]$. Denote $\frac{dm(X,h_0)}{dh}[g]$ as $T_{h_0}[g]$, where $T_{h_0} : \mathcal{H}_{os} \subset L^2(\mathcal{R}, f_{Y_2}) \to L^2([0,1], f_X)$ and $T_{h_0}^*$ as its adjoint. Then for all $h \in \mathcal{H}_{os}$, we have $||h - h_0||^2 = ||T_{h_0}[h - h_0]||_{L^2(f_X)}^2$. If T_{h_0} is a compact operator, then it has a singular value decomposition $\{\mu_j; \phi_{1j}(y_2), \phi_{0j}(x)\}_{j=1}^\infty$ (see section 5.3 for details).

CONDITION 6.11. (i) $[a(y_2)f_{Y_2}(y_2)]$ is $k-$times continuously differentiable, and $\nabla^k[af_{Y_2}]$ goes to zero continuously as $|y_2| \to \infty$; (ii) $E[(l^{(k)}(Y_2))^2] < \infty$ where $l^{(k)} \equiv \frac{\nabla^k[af_{Y_2}]}{f_{Y_2}}$; (iii) $\beta_0 \in int(B)$.

CONDITION 6.12. (i) $\left\|(T_{h_0}^* T_{h_0})^{-\frac{1}{2}}\left[l^{(k)}\right]\right\|_{L^2(f_{Y_2})}^2 < \infty$ (or T_{h_0} is compact and $\sum_{j=1}^\infty \left(\frac{E[l^{(k)}(Y_2)\phi_{1j}(Y_2)]}{\mu_j}\right)^2 < \infty$); (ii) $\left\|(T_{h_0}^* T_{h_0})^{-1}\left[l^{(k)}\right]\right\|_{L^2(f_{Y_2})}^2 < \infty$; (or T_{h_0} is compact and $\sum_{j=1}^\infty \left(\frac{E[l^{(k)}(Y_2)\phi_{1j}(Y_2)]}{\mu_j^2}\right)^2 < \infty$).

Condition 6.12(i) is necessary (but not sufficient) for $\beta_0 = E[a(Y_2)\nabla^k h_0(Y_2)] = (-1)^k E[l^{(k)}(Y_2)h_0(Y_2)]$ to be estimable at a $\sqrt{n}-$rate. It ensures the existence of a Riesz representor v^* for $E\{l^{(k)}(Y_2)(h(Y_2) - h_0(Y_2))\}$ as it ensures $||v^*||^2 = ||T_{h_0}[v^*]||_{L^2(f_X)}^2 = \left\|(T_{h_0}^* T_{h_0})^{-\frac{1}{2}}\left[l^{(k)}\right]\right\|_{L^2(f_{Y_2})}^2 < \infty$. Condition 6.12(ii) is stronger than condition 6.12(i), and ensures that we can solve the Riesz representor v^* in a closed form: $v^* = (T_{h_0}^* T_{h_0})^{-1}\left[l^{(k)}\right] \in L^2(f_{Y_2})$.

CONDITION 6.13. there is a $v_n^* \in \mathcal{H}_n$ such that $||v_n^* - v^*|| \times ||\hat{h}_n - h_0|| = o_P(n^{-1/2})$.

CONDITION 6.14. (i) $||\hat{h}_n - h_0|| = O_P(\delta_n^*)$, with $\delta_n^* = O_p(\max\{\sqrt{\frac{J_n}{n}}, b_{m,J_n}\}) = o_P(n^{-1/4})$; (ii) $||\hat{h}_n - h_0||_s = O_P(\delta_{s,n}^*)$; (iii) $\lambda_n\{P(\hat{h}_n \pm \varepsilon_n v_n^*) - P(\hat{h}_n)\} = o_P(n^{-1})$ with $0 < \varepsilon_n = o(n^{-1/2})$.

Denote $\mathcal{N}_o \equiv \left\{h \in \mathcal{H}_{os} : ||h - h_0|| = O(\delta_n^*), ||h - h_0||_s = O(\delta_{s,n}^*)\right\}$ and $\mathcal{N}_{on} \equiv \{h \in \mathcal{N}_o \cap \mathcal{H}_n\}$.

CONDITION 6.15. (i) There are constants $\kappa \in (0,1]$, $r \geq 1$, and a measurable function $b(X)$ with $E[|b(X)|] < \infty$ such that for all $\delta > 0$ and all $h, h' \in \mathcal{N}_{on}$,

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

(ii) $\sup_{h \in \mathcal{N}_o} |\rho(Z,h)| \leq C(Z)$ and $E[C(Z)^2|X] \leq$ const. $< \infty$; (iii) $\frac{J_n}{n}(\delta_{s,n}^*)^{2\kappa} = o(n^{-1})$.

Condition 6.15 (iii) is satisfied in both "severely" and "mildly" ill-posed case provided that γ_h is big enough. In the following we denote $\widetilde{g}^*(X)$ as the LS projection of $\frac{dm(X,h_0)}{dh}[v^*]$ onto the linear sieve basis $p^{J_n}(X)$.

CONDITION 6.16. *(i)* $\{m(\cdot, h) : h \in \mathcal{N}_{on}\}$ *is a Donsker class in* $L^2([0,1], f_X)$; *(ii)* $\sup_{x \in [0,1]} \left| \frac{dm(x,h_0)}{dh}[v^*] \right| \leq$ *const.* $< \infty$; *(iii)* $\|\widetilde{g}^*(\cdot) - \frac{dm(\cdot, h_0)}{dh}[v^*]\|_{L^2(f_X)} \times O(\delta_n^*) = o(n^{-1/2})$.

CONDITION 6.17. $m(X, h)$ *is twice continuously differentiable in* $h \in \mathcal{N}_{on}$, *(i)* $E \left[\sup_{h \in \mathcal{N}_{on}} \left| \frac{d^2 m(X,h)}{dh^2}[v_n^*, v_n^*] \right|^2 \right] <$ ∞; *(ii)* $E \left[\sup_{h \in \mathcal{N}_{on}} \left\| \frac{dm(X,h)}{dh}[v_n^*] - \frac{dm(X,h_0)}{dh}[v_n^*] \right\|_E^2 \right] = o(n^{-1/2})$; *(iii) for all* $h \in \mathcal{N}_{on}$ *and all* $\overline{h} \in \mathcal{N}_o$,

$$E \left[\left(\frac{dm(X,h_0)}{dh}[v^*] \right)' \left(\frac{dm(X,\overline{h})}{dh}[h - h_0] - \frac{dm(X,h_0)}{dh}[h - h_0] \right) \right] = o(n^{-1/2}).$$

Condition 6.17(i)(ii)(iii) are imposed to control the second order remainder term of $m(X, h) = E[\rho(Y, X_z; h(Y_2))|X]$ in the shrinking neighborhood of h_0. These conditions are automatically satisfied when $m(X, h)$ is linear in h, such as when $m(X, h) = E[Y_1 - h(Y_2)|X]$ in the NPIV model. However, when $m(X, h)$ is nonlinear in h, such as when $m(X, h) = E[1\{Y_1 \leq h(Y_2)\} - \gamma | X]$ in the NPQIV model, condition 6.17(ii)(iii) may be difficult to verify when the problem is "severely" ill-posed. See Chen and Pouzo (2008) for further discussions.

PROPOSITION 6.7. *For the model (6.2), suppose that all the conditions of Proposition 6.6 hold. Let assumption 4.1 and conditions 6.11 - 6.17 hold and* $\overline{k} > k + 0.5$. *Then:* $\sqrt{n}(\widehat{\beta}_n - \beta_0) \Rightarrow N(0, V^{-1})$, *with*

$$V^{-1} = Var \left\{ (\beta_0 - a(Y_2)\nabla^k h_0(Y_2)) + (-1)^k \left(T_{h_0}[v^*] \right) \rho(Z_i, h_0) \right\}.$$

Remark 6.1. *(1) The semiparametric efficiency bound for* β_0 *identified through the model (6.2) can be derived by applying the results of Ai and Chen (2005). When* Y_2 *is endogenous, the plug-in penalized SMD estimator* $\widehat{\beta}_n$ *fails to reach this efficiency bound. Nevertheless, by combining the rate results of our this paper and the semiparametric efficient estimation procedures proposed in Ai and Chen (2005) and Chen and Pouzo (2008), one can obtain efficient estimation of* β_0. *(2) When we specialize Proposition 6.7 to the NPQIV model (1.3) with* $\rho(Y, X_z; h_0(Y_2)) = 1\{Y_1 \leq h_0(Y_2)\} - \gamma$, *we immediate obtain* $\sqrt{n}$−*asymptotic normality of the plug-in penalized SMD estimator of the weighted average derivative of the quantile IV function* $\beta_0 = E[a(Y_2)\nabla^k h_0(Y_2)]$. *See Chen and Pouzo (2007) for semiparametric efficient estimation of the weighted average derivative of the quantile IV function.*

7 Conclusion

In this paper, we propose penalized SMD estimation of conditional moment models containing unknown functions of endogenous variables. The estimation problem is a difficult nonlinear ill-

posed inverse problem with an unknown operator. We establish consistency and convergence rate of the penalized SMD estimator, allowing for (i) possibly non-compact original parameter space; (ii) possibly non-compact finite or infinite dimensional sieve spaces with flexible penalty; (iii) possibly nonsmooth generalized residual functions; (iv) any lower semicompact or convex penalty, or SMD with finite dimensional linear sieves without penalty; and (v) mildly or severely ill-posed inverse problems. Under relatively low-level sufficient conditions, we show that the convergence rates coincide with the known minimax optimal rates for the NPIV model (1.2). We illustrate the general theory by two important applications: the consistency and convergence rate of a nonparametric additive quantile IV regression, and the root-n asymptotic normality of the plug-in penalized SMD estimator of a weighted average derivative of $h_0(Y)$ in the nonlinear model $E[\rho(Y, X_z, h_0(Y))|X] = 0$. We also present a simulation study and an estimation of a system of nonparametric quantile IV Engel curves using the UK Family Expenditure Survey.

In Chen and Pouzo (2008), for the general conditional moment models (1.1), we show that the semiparametric efficiency bounds and the root-n asymptotic normality of θ_0 are still valid even when $\rho(Y, X_z, \theta, h(\cdot))$ is not pointwise smooth in (θ, h). We establish that a weighted bootstrap procedure consistently estimate the confidence region of the penalized SMD estimator $\widehat{\theta}_n$. We also derive that the scaled and centered profiled optimally weighted penalized SMD criterion function is asymptotically Chi-square distributed.

A   Some Function Spaces and Sieves

Let $\mathcal{S}(\mathcal{R}^d)$ be the Schwartz space of all complex-valued, rapidly decreasing, infinitely differentiable functions on $\mathcal{R}^d$. Let $\mathcal{S}^*(\mathcal{R}^d)$ be the space of all tempered distributions on $\mathcal{R}^d$, which is the topological dual of $\mathcal{S}(\mathcal{R}^d)$. For $h \in \mathcal{S}(\mathcal{R}^d)$ we let $\widehat{h}$ denote the Fourier transform of h (i.e., $\widehat{h}(\xi) = (2\pi)^{-d/2} \int_{\mathcal{R}^d} \exp\{-iy'\xi\} h(y) dy$), and $(g)^\vee$ the inverse Fourier transform of g (i.e., $(g)^\vee(y) = (2\pi)^{-d/2} \int_{\mathcal{R}^d} \exp\{iy'\xi\} g(\xi) d\xi$). Let $\varphi_0 \in \mathcal{S}(\mathcal{R}^d)$ be such that $\varphi_0(x) = 1$ if $|x| \leq 1$ and $\varphi_0(x) = 0$ if $|x| \geq 3/2$. Let $\varphi_1(x) = \varphi_0(x/2) - \varphi_0(x)$ and $\varphi_k(x) = \varphi_1(2^{-k+1}x)$ for all integer $k \geq 1$. Then the sequence $\{\varphi_k : k \geq 0\}$ forms a dyadic resolution of unity (i.e., $1 = \sum_{k=0}^\infty \varphi_k(x)$ for all $x \in \mathcal{R}^d$). Let $\nu \in \mathcal{R}$ and $p, q \in (0, \infty]$, the *Besov space* $\mathcal{B}_{p,q}^\nu(\mathcal{R}^d)$ is the collection of all functions $h \in \mathcal{S}^*(\mathcal{R}^d)$ such that $\|h\|_{\mathcal{B}_{p,q}^\nu}$ is finite:

$$\|h\|_{\mathcal{B}_{p,q}^\nu} \equiv \left(\sum_{j=0}^\infty \left\{ 2^{j\nu} \left\| \left(\varphi_j \widehat{h} \right)^\vee \right\|_{L^p(leb)} \right\}^q \right)^{1/q} < \infty$$

(with the usual modification if $q = \infty$). Let $\nu \in \mathcal{R}$ and $p \in (0, \infty), q \in (0, \infty]$, the *F-space* $\mathcal{F}_{p,q}^\nu(\mathcal{R}^d)$ is the collection of all functions $h \in \mathcal{S}^*(\mathcal{R}^d)$ such that $\|h\|_{\mathcal{F}_{p,q}^\nu}$ is finite:

$$\|h\|_{\mathcal{F}_{p,q}^\nu} \equiv \left\| \left(\sum_{j=0}^\infty \left\{ 2^{j\nu} \left| \left(\varphi_j \widehat{h} \right)^\vee (\cdot) \right| \right\}^q \right)^{1/q} \right\|_{L^p(leb)} < \infty$$

(with the usual modification if $q = \infty$). For $\nu > 0$, $p, q \geq 1$, it is known that $\mathcal{F}_{p',q'}^{-\nu}\left(\mathcal{R}^d\right)$ $\left(\mathcal{B}_{p',q'}^{-\nu}\left(\mathcal{R}^d\right)\right)$ is the dual space of $\mathcal{F}_{p,q}^{\nu}\left(\mathcal{R}^d\right)$ $\left(\mathcal{B}_{p,q}^{\nu}\left(\mathcal{R}^d\right)\right)$ with $1/p' + 1/p = 1$ and $1/q' + 1/q = 1$.

Let $\mathcal{T}_{p,q}^{\nu}\left(\mathcal{R}^d\right)$ denote either $\mathcal{B}_{p,q}^{\nu}\left(\mathcal{R}^d\right)$ or $\mathcal{F}_{p,q}^{\nu}\left(\mathcal{R}^d\right)$. If $p, q \geq 1$ then $\mathcal{T}_{p,q}^{\nu}\left(\mathcal{R}^d\right)$ is a Banach space. Moreover, $\mathcal{T}_{p,q}^{\nu}\left(\mathcal{R}^d\right)$ gets larger with increasing q ($\mathcal{T}_{p,q_1}^{\nu}\left(\mathcal{R}^d\right) \subseteq \mathcal{T}_{p,q_2}^{\nu}\left(\mathcal{R}^d\right)$ for $q_1 \leq q_2$), gets larger with decreasing p ($\mathcal{T}_{p_1,q}^{\nu}\left(\mathcal{R}^d\right) \subseteq \mathcal{T}_{p_2,q}^{\nu}\left(\mathcal{R}^d\right)$ for $p_1 \geq p_2$), and gets larger with decreasing ν ($\mathcal{T}_{pq}^{\nu_1}\left(\mathcal{R}^d\right) \subseteq \mathcal{T}_{p,q}^{\nu_2}\left(\mathcal{R}^d\right)$ for $\nu_1 \geq \nu_2$). The spaces $\mathcal{T}_{p,q}^{\nu}\left(\mathcal{R}^d\right)$ include many well-known function spaces as special cases. For example, $L^p(\mathcal{R}^d, leb) = \mathcal{F}_{p,2}^0\left(\mathcal{R}^d\right)$ for $p \in (1, \infty)$; the Hölder space $\Lambda^r(\mathcal{R}^d) = \mathcal{B}_{\infty,\infty}^r\left(\mathcal{R}^d\right)$ for any real-valued $r > 0$; the Hilbert-Sobolev space $W_2^k(\mathcal{R}^d) = \mathcal{B}_{2,2}^k\left(\mathcal{R}^d\right)$ for integer $k > 0$; and the (fractional) Sobolev space $W_p^\nu(\mathcal{R}^d) = \mathcal{F}_{p,2}^\nu\left(\mathcal{R}^d\right)$ for any $\nu \in \mathcal{R}$ and $p \in (1, \infty)$, which has the equivalent norm $||h||_{W_p^\nu} \equiv \left\| \left((1 + |\cdot|^2)^{\nu/2} \widehat{h}(\cdot) \right)^{\vee} \right\|_{L^p(leb)} < \infty$ (note that for $\nu > 0$, the norm $||h||_{W_p^{-\nu}}$ is a shrinkage in the Fourier domain).

We can also define "weighted" versions of the afore-mentioned spaces as follows. Let $w(\cdot) = (1 + |\cdot|^2)^{\zeta/2}$, $\zeta \in \mathcal{R}$ be the weight function and define $||h||_{\mathcal{T}_{p,q}^\nu(\mathcal{R}^d, w)} = ||wh||_{\mathcal{T}_{p,q}^\nu(\mathcal{R}^d)}$, that is, $\mathcal{T}_{p,q}^\nu(\mathcal{R}^d, w) = \{h : ||wh||_{\mathcal{T}_{p,q}^\nu(\mathcal{R}^d)} < \infty\}$. See Edmunds and Triebel (1996) for additional properties of the general Besov spaces and the F-spaces, especially the properties of continuous embeddings and compact embeddings between any two spaces $\mathcal{T}_{p_1,q_1}^{\nu_1}(\mathcal{R}^d, w_1)$ and $\mathcal{T}_{p_2,q_2}^{\nu_2}(\mathcal{R}^d, w_2)$.

If $\mathcal{H} \subseteq H$ with is a Besov space then a wavelet basis $\{\psi_j\}$ is a natural choice of $\{q_j\}_j$ to satisfy assumption 5.1 in Section 4. A real-valued function ψ is called a "mother wavelet" of degree γ if it satisfies: (a) $\int_{\mathcal{R}} y^k \psi(y) dy = 0$ for $0 \leq k \leq \gamma$; (b) ψ and all its derivatives up to order γ decrease rapidly as $|y| \to \infty$; (c) $\{2^{k/2}\psi(2^k y - j) : k, j \in \mathbb{Z}\}$ forms a Riesz basis of $L^2(leb)$, that is, the linear span of $\{2^{k/2}\psi(2^k y - j) : k, j \in \mathbb{Z}\}$ is dense in $L^2(leb)$ and

$$\left\| \sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} a_{kj} 2^{k/2} \psi(2^k y - j) \right\|_{L^2(\mathcal{R})}^2 \asymp \sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} |a_{kj}|^2$$

for all doubly bi-infinite square-summable sequence $\{a_{kj} : k, j \in \mathbb{Z}\}$. A scaling function φ is called a "father wavelet" of degree γ if it satisfies: (a') $\int_{\mathcal{R}} \varphi(y) dy = 1$; (b') φ and all its derivatives up to order γ decrease rapidly as $|y| \to \infty$; (c') $\{\varphi(y - j) : j \in \mathbb{Z}\}$ forms a Riesz basis for a closed subspace of $L^2(leb)$.

Some examples:

Orthogonal wavelets. Given an integer $\gamma > 0$, there exist a father wavelet φ of degree γ and a mother wavelet ψ of degree γ, both compactly supported, such that for any integer $k_0 \geq 0$, any function h in $L^2(leb)$ has the following wavelet $\gamma-$ regular multiresolution expansion:

$$h(y) = \sum_{j=-\infty}^{\infty} a_{k_0 j} \varphi_{k_0 j}(y) + \sum_{k=k_0}^{\infty} \sum_{j=-\infty}^{\infty} b_{kj} \psi_{kj}(y), \quad y \in \mathcal{R},$$

where

$$a_{kj} = \int_{\mathcal{R}} h(y) \varphi_{kj}(y) dy, \quad \varphi_{kj}(y) = 2^{k/2} \varphi(2^k y - j), \quad y \in \mathcal{R},$$

$$b_{kj} = \int_{\mathcal{R}} g(y) \psi_{kj}(y) dy, \quad \psi_{jk}(y) = 2^{k/2} \psi(2^k y - j), \quad y \in \mathcal{R},$$

and $\{\varphi_{k_0 j}, j \in \mathbb{Z}; \psi_{kj}, k \geq k_0, j \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(leb)$; see Meyer (1992, theorem 3.3). For an integer $K_n > k_0$, we consider the finite-dimensional linear space spanned by this wavelet basis of order γ:

$$h_n(y) = \psi^{k_n}(y)'\Pi = \sum_{j=0}^{2^{K_n}-1} h_{K_n,j}\varphi_{K_n,j}(y), \quad k(n) = 2^{K_n}.$$

Cardinal B-spline wavelets of order γ:

$$h_n(y) = \psi^{k_n}(y)'\Pi = \sum_{k=0}^{K_n}\sum_{j\in\mathcal{K}_n}\pi_{kj}2^{k/2}B_\gamma(2^k y - j), \quad k(n) = 2^{K_n} + 1, \tag{A.1}$$

where $B_\gamma(\cdot)$ is the cardinal B-spline of order γ,

$$B_\gamma(y) = \frac{1}{(\gamma-1)!}\sum_{i=0}^{\gamma}(-1)^i \begin{pmatrix} \gamma \\ i \end{pmatrix}[\max(0, y-i)]^{\gamma-1},$$

which is $\gamma - 1$ times differentiable and has support on $[0, \gamma]$. For any fixed integer $k = 0, 1, ..., K_n$, $\mathcal{K}_n$ is the set consisting of those j's such that the support of $z \rightarrow B_\gamma(2^k z - j)$ overlaps with the empirical support of the data, $j = \pm 1, \pm 2, ...$ The compact support of $B_\gamma(\cdot)$ ensures that $\#\mathcal{K}_n$ is finite for any fixed k.

In the empirical illustration and simulation study in Section 2, we also applied polynomial spines (P-splines) and Hermite polynomial sieves:

Polynomial splines of order q_n:

$$h_n(y) = \psi^{k_n}(y)'\Pi = \sum_{j=0}^{q_n}\pi_j(y)^j + \sum_{k=1}^{r_n}\pi_{q_n+k}(y - \nu_k)_+^{q_n}, \quad k(n) = q_n + r_n + 1, \tag{A.2}$$

where $(y - \nu)_+^q = \max\{(y-\nu)^q, 0\}$ and $\{\nu_k\}_{k=1,...,r_n}$ are the knots. In the empirical application, for any given number of knots value r_n, the knots $\{\nu_k\}_{k=1,...,r_n}$ are simply chosen as the empirical quantiles of the data.

Hermite polynomials of order $k(n) - 1$:

$$h_n(y) = \psi^{k_n}(y)'\Pi = \sum_{j=0}^{k_n-1}\pi_j(y - \nu_1)^j \exp\left\{-\frac{(y - \nu_1)^2}{2\nu_2^2}\right\}, \tag{A.3}$$

where ν_1 and ν_2^2 can be chosen as the sample mean and variance of the data.

B   Consistency

We first present a general consistency lemma that is applicable to any approximate penalized sieve extremum estimation problems, be them well-posed or ill-posed.

LEMMA B.1. *Let $\widehat{\alpha}_n$ be such that $\widehat{Q}_n(\widehat{\alpha}_n) \leq \inf_{\alpha\in\mathcal{A}_{k(n)}}\widehat{Q}_n(\alpha) + O_P(\eta_n)$ with $\eta_n = o(1)$. Suppose there are real-valued functions $\overline{Q}(\alpha), \overline{Q}_n(\alpha)$ such that the following conditions (B.1.1) - (B.1.4) hold:*

(B.1.1) (i) $\overline{Q}(\alpha_0) \leq \overline{Q}_n(\alpha_0) < \infty$, *and* $\overline{Q}_n(\alpha_0) - \overline{Q}(\alpha_0) = o(1)$; *(ii) there is a positive function* $g_0(n, k, \varepsilon)$ *such that:*

$$\inf_{\alpha \in \mathcal{A}_k : ||\alpha - \alpha_0||_c \geq \varepsilon} \overline{Q}_n(\alpha) - \overline{Q}(\alpha_0) \geq g_0(n, k, \varepsilon) > 0 \quad \text{for each } n \geq 1, k \geq 1, \varepsilon > 0,$$

and uniformly in $\varepsilon > 0$

$$\liminf_{n \to \infty} g_0(n, k(n), \varepsilon) \geq 0, \ \lim_n \frac{\inf_{\alpha \in \mathcal{A}_{k(n)} : ||\alpha - \alpha_0||_c \geq \varepsilon} \overline{Q}_n(\alpha) - \overline{Q}_n(\alpha_0)}{g_0(n, k(n), \varepsilon)} > 0.$$

(B.1.2) (i) $\mathcal{A} \subseteq \mathbf{A}$ *and* $(\mathbf{A}, || \cdot ||_c)$ *is a metric space; (ii)* $\mathcal{A}_k \subseteq \mathcal{A}_{k+1} \subseteq \mathcal{A}$ *for all* $k \geq 1$, *and there exists a sequence* $\Pi_n \alpha_0 \in \mathcal{A}_{k(n)}$ *such that* $||\Pi_n \alpha_0 - \alpha_0||_c \to 0$ *as* $n \to \infty$.

(B.1.3) (i) $\widehat{Q}_n(\alpha)$ *is a measurable function of the data* $\{(Y_i, X_i)\}_{i=1}^n$ *for all* $\alpha \in \mathcal{A}_{k(n)}$; *(ii)* $\widehat{\alpha}_n$ *is well-defined and measurable.*

(B.1.4) Let $\hat{c}^Q(k(n)) \equiv \sup_{\alpha \in \mathcal{A}_{k(n)}} \left| \widehat{Q}_n(\alpha) - \overline{Q}_n(\alpha) \right| = o_P(1)$. *Uniformly over* $\varepsilon > 0$,

$$\frac{\max \left\{ \hat{c}^Q(k(n)), \eta_n, \left| \overline{Q}_n(\Pi_n \alpha_0) - \overline{Q}_n(\alpha_0) \right| \right\}}{g_0(n, k(n), \varepsilon)} = o(1).$$

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

PROOF OF LEMMA B.1: Under condition (B.1.3)(ii) $\widehat{\alpha}_n$ is well-defined and measurable. It follows that for any $\varepsilon > 0$,

$$\Pr \left(||\widehat{\alpha}_n - \alpha_0||_c > \varepsilon \right)$$
$$\leq \Pr \left(\inf_{\alpha \in \mathcal{A}_{k(n)} : ||\alpha - \alpha_0||_c \geq \varepsilon} \widehat{Q}_n(\alpha) \leq \widehat{Q}_n(\Pi_n \alpha_0) + O(\eta_n) \right)$$
$$\leq \Pr \left(\inf_{\alpha \in \mathcal{A}_{k(n)} : ||\alpha - \alpha_0||_c \geq \varepsilon} \left\{ \overline{Q}_n(\alpha) - \left| \widehat{Q}_n(\alpha) - \overline{Q}_n(\alpha) \right| \right\} \leq \widehat{Q}_n(\Pi_n \alpha_0) + O(\eta_n) \right)$$
$$\leq \Pr \left(\inf_{\alpha \in \mathcal{A}_{k(n)} : ||\alpha - \alpha_0||_c \geq \varepsilon} \overline{Q}_n(\alpha) \leq 2\hat{c}^Q(k(n)) + \overline{Q}_n(\Pi_n \alpha_0) + O(\eta_n) \right)$$
$$\leq \Pr \left(\inf_{\alpha \in \mathcal{A}_{k(n)} : ||\alpha - \alpha_0||_c \geq \varepsilon} \overline{Q}_n(\alpha) - \overline{Q}_n(\alpha_0) \leq 2\hat{c}^Q(k(n)) + \overline{Q}_n(\Pi_n \alpha_0) - \overline{Q}_n(\alpha_0) + O(\eta_n) \right)$$
$$\leq \Pr \left(\frac{\inf_{\alpha \in \mathcal{A}_{k(n)} : ||\alpha - \alpha_0||_c \geq \varepsilon} \overline{Q}_n(\alpha) - \overline{Q}_n(\alpha_0)}{g_0(n, k(n), \varepsilon)} \leq \frac{2\hat{c}^Q(k(n)) + \left| \overline{Q}_n(\Pi_n \alpha_0) - \overline{Q}_n(\alpha_0) \right| + O(\eta_n)}{g_0(n, k(n), \varepsilon)} \right)$$

which goes to 0 by conditions (B.1.1)(ii) and (B.1.4). Q.E.D.

Remark B.1. *(1) Let* $(\mathbf{A}, \mathcal{T})$ *be a topological space. Condition (B.1.3) is satisfied if one of the following two conditions holds: (a) for each* $k \geq 1$, $\mathcal{A}_k$ *is a compact subset of* $(\mathbf{A}, \mathcal{T})$, *and for any data* $\{Z_i\}_{i=1}^n$, $\widehat{Q}_n(\alpha)$ *is lower semicontinuous (in the topology* $\mathcal{T}$) *on* $\mathcal{A}_k$. *(b) for any data* $\{Z_i\}_{i=1}^n$, *the sets* $\{\alpha \in \mathcal{A}_k : \widehat{Q}_n(\alpha) \leq r\}$ *is compact in* $(\mathbf{A}, \mathcal{T})$ *for all* $r \in (-\infty, +\infty)$.

(2) Let $(\mathbf{A}, || \cdot ||_c)$ *be a Banach space. Condition (B.1.3) is satisfied if one of the following three conditions holds: (a)* $\mathcal{A}_k$ *is compact under* $|| \cdot ||_c$, *and for any data* $\{Z_i\}_{i=1}^n$, $\widehat{Q}_n(\alpha)$ *is lower semicontinuous (in* $|| \cdot ||_c$) *on* $\mathcal{A}_{k(n)}$. *(b)* $\mathcal{A}_k$ *is a bounded, and weak sequentially closed (i.e., for each weakly convergent sequence in* $\mathcal{A}_k$, *its limit belongs to* $\mathcal{A}_k$) *subset of a reflexive Banach space*

$(\mathbf{A}, ||\cdot||_c)$, and for any data $\{Z_i\}_{i=1}^n$, $\widehat{Q}_n(\alpha)$ is weak sequentially lower semicontinuous on $\mathcal{A}_{k(n)}$.
(c) $\mathcal{A}_k$ is a bounded, closed and convex subset of a reflexive Banach space $(\mathbf{A}, ||\cdot||_c)$, and for any
data $\{Z_i\}_{i=1}^n$, $\widehat{Q}_n(\alpha)$ is convex and lower semicontinuous on $\mathcal{A}_{k(n)}$. Moreover, (c) implies (b). See
Zeidler (1985, proposition 38.7, theorem 38.A, corollary 38.8 and theorem 38.B).

PROOF OF THEOREM 3.1: We verify that all the conditions of Lemma B.1 are satisfied with
$\alpha = h$ and $\eta_n = 0$. Let $\widehat{Q}_n(h) = n^{-1}\sum_{i=1}^n \widehat{m}(X_i, h)'\widehat{m}(X_i, h) + \lambda_n \widehat{P}_n(h)$, $\overline{Q}_n(h) = \overline{Q}(h) + \lambda_n P(h)$
and $\overline{Q}(h) = E\left[m(X, h)'m(X, h)\right]$. Then $\widehat{h}_n = \arg\inf_{h \in \mathcal{H}_n} \widehat{Q}_n(h)$ and conditions (B.1.2) and (B.1.3)
are directly assumed. Condition (B.1.1) is satisfied given assumptions 3.1(iii), 3.4, as $\overline{Q}(h_0) = 0 \leq$
$\overline{Q}_n(h_0) = \lambda_n P(h_0) < \infty$, and for each $\varepsilon > 0$, $k \geq 1$, $\lambda_n \geq 0$, we have

$$\inf_{h \in \mathcal{H}_k : ||h - h_0||_c \geq \varepsilon} \left\{ E\left[m(X, h)'m(X, h)\right] + \lambda_n P(h) \right\} \geq \inf_{h \in \mathcal{H}_k : ||h - h_0||_c \geq \varepsilon} \left\{ E\left[m(X, h)'m(X, h)\right] \right\} = g_0(n, k, \varepsilon) > 0.$$

Condition (3.1.1) implies that uniformly over $\varepsilon > 0$,

$$\liminf_n g_0(n, k(n), \varepsilon) > 0.$$

This and assumption 3.4 and $\lambda_n \geq 0, \lambda_n = o(1)$ then imply uniformly over $\varepsilon > 0$,

$$\lim_n \frac{\inf_{h \in \mathcal{H}_{k(n)} : ||h - h_0||_c \geq \varepsilon} \left\{ E\left[m(X, h)'m(X, h)\right] + \lambda_n P(h) \right\} - \lambda_n P(h_0)}{g_0(n, k(n), \varepsilon)} = 1.$$

Also, assumptions 3.2 and 3.4 and $\lambda_n = o(1)$ imply $\overline{Q}_n(h_0) - \overline{Q}_n(\Pi_n h_0) = o(1)$. It remains to
check condition (B.1.4) by establishing $\sup_{h \in \mathcal{H}_n} |\widehat{Q}_n(h) - \overline{Q}_n(h)| = o_P(1)$, which is satisfied given
condition (3.1.2) and assumption 3.4, as

$$\left| \widehat{Q}_n(h) - \overline{Q}_n(h) \right|$$
$$\leq \left| \frac{1}{n}\sum_{i=1}^n \widehat{m}(X_i, h)'\widehat{m}(X_i, h) - E\left[m(X, h)'m(X, h)\right] \right| + \left| \lambda_n \widehat{P}_n(h) - \lambda_n P(h) \right|$$
$$= o_P(1) \text{ uniformly over } \mathcal{H}_n.$$

Finally, we show that assumptions 3.1(i) and 3.5 imply condition (3.1.2). Notice that

$$\left| \frac{1}{n}\sum_{i=1}^n \widehat{m}(X_i, h)'\widehat{m}(X_i, h) - E\left[m(X, h)'m(X, h)\right] \right|$$
$$\leq \left| \frac{1}{n}\sum_{i=1}^n \left[\widehat{m}(X_i, h)'\widehat{m}(X_i, h) - m(X_i, h)'m(X_i, h)\right] \right| + \left| \frac{1}{n}\sum_{i=1}^n m(X_i, h)'m(X_i, h) - E\left[m(X_i, h)'m(X_i, h)\right] \right|.$$

$$\left| \frac{1}{n}\sum_{i=1}^n \left[\widehat{m}(X_i, h)'\widehat{m}(X_i, h) - m(X_i, h)'m(X_i, h)\right] \right|$$
$$\leq \left| \frac{1}{n}\sum_{i=1}^n \left(\widehat{m}(X_i, h) - m(X_i, h)\right)'\left(\widehat{m}(X_i, h) - m(X_i, h)\right) \right| + \left| \frac{2}{n}\sum_{i=1}^n m(X_i, h)'\left(\widehat{m}(X_i, h) - m(X_i, h)\right) \right|$$
$$= o_P(1) \text{ uniformly over } \mathcal{H}_n,$$

where the last equality is due to assumption 3.5(ii)(iii) and Cauchy-Schwarz inequality. Finally, applying the Glivenko-Cantelli theorem, assumptions 3.1(i) and 3.5(i)(ii)(iv) imply

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i, h)' m(X_i, h) - E\left[m(X_i, h)' m(X_i, h)\right] = o_P(1) \text{ uniformly over } \mathcal{H}_n.$$

Thus condition (3.1.2) is satisfied, and the result follows from Lemma B.1. *Q.E.D.*

PROOF OF COROLLARY 3.1: This can be trivially obtained by applying Lemma B.1 with $\overline{Q}_n(\alpha) = E\left[m(X, h)' m(X, h)\right] = \overline{Q}(\alpha)$ and $\widehat{Q}_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(X_i, h)' \widehat{m}(X_i, h) + O_P(\eta_n)$, where $O_P(\eta_n) = \lambda_n\{\widehat{P}_n(h) - P(h)\} + \lambda_n P(h)$; hence $O_P(\eta_n) = o_P(1)$ uniformly over $h \in \mathcal{H}_n$ by assumptions 3.1(i), 3.4 and 3.6(ii). Assumptions 3.1(iii) and 3.7 imply condition (B.1.1) holds with $\liminf_n g_0(n, k(n), \varepsilon) > 0$ uniformly in $\varepsilon > 0$. Assumptions 3.1(i)(ii), 3.4, 3.5, 3.6(i) and 3.7, and Remark B.1 imply assumption 3.3 (i.e., condition (B.1.3)). Condition (B.1.2) is directly assumed. Finally condition (B.1.4) follows from condition (3.1.2), which is implied by assumptions 3.1(i) and 3.5 (see the proof of Theorem 3.1), while assumption 3.5(iv) is implied by assumption 3.6(i). *Q.E.D.*

In the following we denote $||g||_X^2 \equiv E[g^2(X)]$, $||g||_{n,X}^2 \equiv \frac{1}{n}\sum_{i=1}^{n} g^2(X_i)$ and $\langle g, \overline{g}\rangle_{n,X} \equiv \frac{1}{n}\sum_{i=1}^{n} g(X_i)\overline{g}(X_i)$.

LEMMA B.2. *Let assumptions 3.9 and 3.10(i) hold with an i.i.d. sample $\{(Y_i, X_i)\}_{i=1}^n$. Let $G_n \equiv \{g : g(x) = \sum_{k=1}^{J_n} \langle g_h, p_k\rangle_{n,X} p_k(x); h \in \mathcal{H}_n, \sup_x |g(x)| < \infty\}$ where g_h is a square integrable function of X indexed by $h \in \mathcal{H}_n$, and $\{p_k\}_{k=1}^{J_n}$ is some linear sieve basis functions (e.g., B-Splines). Then*

$$\sup_{h \in \mathcal{H}_n} \left| \frac{||g_h||_{n,X}^2}{||g_h||_X^2} - 1 \right| = o_P(1).$$

Consequently, there are finite constants $K, K' > 0$ such that, except on an event whose probability goes to zero as $n \to \infty$,

$$K'||\widehat{m}(\cdot, h)||_X^2 \le ||\widehat{m}(\cdot, h)||_{n,X}^2 \le K||\widehat{m}(\cdot, h)||_X^2 \quad \text{uniformly on } \mathcal{H}_n.$$

PROOF OF LEMMA B.2: Note that for functions in G_n we have that

$$\sup_{h \in \mathcal{H}_n} ||g_h||_{n,X} = \sup_{h \in \mathcal{H}_n} \left\| \sum_{k=1}^{J_n} \langle g_h, p_k\rangle_{n,X} p_k \right\|_{n,X} \equiv \sup_{g \in G_n} ||g||_{n,X}.$$

Define $A_n \equiv \sup_{g \in G_n} \frac{\sup_x |g(x)|}{||g||_X}$. Then under assumption 3.9 and the definition of G_n, we have $A_n \asymp \xi_n$. Thus, by assumption 3.9(iii), the result follows from Lemma 4 of Huang (1998) for general linear sieves $\{p_k\}_{k=1}^{J_n}$ and Corollary 3 of Huang (2003) for polynomial spline sieves. *Q.E.D.*

Let $\widetilde{m}(X, h) \equiv p^{J_n}(X)' (P'P)^{-1} P' m(h)$ and $m(h) = (m(X_1, h), \ldots, m(X_n, h))'$.

LEMMA B.3. *(1) Let assumptions 3.9 and 3.10(i) hold with an i.i.d. sample $\{(Y_i, X_i)\}_{i=1}^n$. Then:*

$$\sup_{h \in \mathcal{H}_n} ||\widehat{m}(\cdot, h) - \widetilde{m}(\cdot, h)||_{n,X}^2 \asymp \sup_{h \in \mathcal{H}_n} ||\widehat{m}(\cdot, h) - \widetilde{m}(\cdot, h)||_X^2 = O_P\left(\frac{J_n}{n}\right);$$

(2) If, further, assumption 3.10(ii) holds, then:

$$\sup_{h \in \mathcal{H}_n} ||\widehat{m}(\cdot, h) - m(\cdot, h)||_X^2 = O_P\left(\frac{J_n}{n} + b_{m, J_n}^2\right).$$

PROOF OF LEMMA B.3: For result (1), by Lemma B.2 and definitions of $\widetilde{m}(X, h)$, we have:

$$\sup_{h \in \mathcal{H}_n} n^{-1} \sum_{i=1}^{n} \|\widehat{m}(X_i, h) - \widetilde{m}(X_i, h)\|_E^2 \asymp \sup_{h \in \mathcal{H}_n} E\left[n^{-1} \sum_{i=1}^{n} \|\widehat{m}(X_i, h) - \widetilde{m}(X_i, h)\|_E^2\right].$$

$$
\begin{aligned}
&E\left[n^{-1} \sum_{i=1}^{n} \|\widehat{m}(X_i, h) - \widetilde{m}(X_i, h)\|_E^2\right] \\
\leq\ &E\left[Tr\left\{n^{-1} \sum_{i=1}^{n} p^{J_n}(X_i)'(P'P)^{-1} P' \varepsilon(h) \varepsilon(h)' P(P'P)^{-1} p^{J_n}(X_i)\right\}\right] \\
\leq\ &E\left[n^{-1} Tr\left\{P(P'P)^{-1} P' \varepsilon(h) \varepsilon(h)'\right\}\right] \leq E\left[n^{-1} Tr\left\{P(P'P)^{-1} P' E\left[\varepsilon(h) \varepsilon(h)' | X\right]\right\}\right] \\
\leq\ &K E\left[n^{-1} Tr\left\{P(P'P)^{-1} P'\right\}\right] \leq K J_n/n
\end{aligned}
$$

where $\varepsilon(h) = (\varepsilon(Z_1, h), \ldots, \varepsilon(Z_n, h))'$, $\varepsilon(Z, h) = \rho(Z, h) - m(X, h)$ and K is a finite constant independent of $h \in \mathcal{H}_n$, the fourth inequality follows from assumption 3.10(i), and the last inequality follows from assumption 3.9(ii).

For Result (2), given Result (1), assumption 3.10(ii) and the following inequality

$$\|\widehat{m}(\cdot, h) - m(\cdot, h)\|_X \leq \|\widehat{m}(\cdot, h) - \widetilde{m}(\cdot, h)\|_X + \|\widetilde{m}(\cdot, h) - m(\cdot, h)\|_X,$$

Result (2) follows trivially. *Q.E.D.*

LEMMA B.4. *Let $\hat{h}_n$ be the penalized SMD estimator with $\lambda_n \geq 0$, $\lambda_n = o_P(1)$, and $\widehat{m}(X, h)$ any consistent estimator of $m(X, h)$ satisfying assumption 3.8 (or assumptions 3.9 - 3.10 for series LS estimator $\widehat{m}(X, h)$). Let assumptions 3.1(i)(ii), 3.2(i) and 3.3 hold. Then: (1) under assumption 3.4, for all $\varepsilon > 0$,*

$$
\begin{aligned}
&\Pr\left(\|\hat{h}_n - h_0\|_c > \varepsilon\right) \\
\leq\ &\Pr\left(\begin{array}{c} \inf_{h \in \mathcal{H}_{k(n)}: \|h - h_0\|_c \geq \varepsilon} \{E[m(X, h)'m(X, h)] + \lambda_n P(h)\} \\ \leq K\{E[m(X, \Pi_n h_0)'m(X, \Pi_n h_0)] + O_P(\delta_{m,n}^2) + \lambda_n P(h_0) + O_P(\lambda_n)\}\end{array}\right);
\end{aligned}
$$

(2) under assumption 3.11, for all $\varepsilon > 0$,

$$
\begin{aligned}
&\Pr\left(\|\hat{h}_n - h_0\|_c > \varepsilon\right) \\
\leq\ &\Pr\left(\begin{array}{c} \inf_{h \in \mathcal{H}_{k(n)}: \|h - h_0\|_c \geq \varepsilon} \{E[m(X, h)'m(X, h)] + \lambda_n P(h)\} \\ \leq K\{E[m(X, \Pi_n h_0)'m(X, \Pi_n h_0)] + O_P(\delta_{m,n}^2) + \lambda_n P(h_0) + o_P(\lambda_n)\}\end{array}\right).
\end{aligned}
$$

PROOF OF LEMMA B.4: By definition of $\hat{h}_n$ and $\Pi_n h_0$ and assumptions 3.1(i)(ii), 3.2(i) and 3.3, we have: for any $\varepsilon > 0$,

$$
\begin{aligned}
&\Pr\left(\|\hat{h}_n - h_0\|_c > \varepsilon\right) \\
\leq\ &\Pr\left(\begin{array}{c} \inf_{h \in \mathcal{H}_{k(n)}: \|h - h_0\|_c \geq \varepsilon}\{n^{-1} \sum_{i=1}^{n} \widehat{m}(X_i, h)'\widehat{m}(X_i, h) + \lambda_n \widehat{P}(h)\} \\ \leq n^{-1} \sum_{i=1}^{n} \widehat{m}(X_i, \Pi_n h_0)'\widehat{m}(X_i, \Pi_n h_0) + \lambda_n \widehat{P}(\Pi_n h_0)\end{array}\right).
\end{aligned}
$$

By the i.i.d. sample, and assumption 3.8(ii) for any consistent estimator $\widehat{m}$, (or assumptions 3.9 - 3.10(i) and Lemma B.2 for the series LS estimator $\widehat{m}$), there are finite positive constants K and K'

such that for all $h \in \mathcal{H}_n$, we have:

$$K' E\left[\widehat{m}(X, h)' \widehat{m}(X, h)\right] \geq n^{-1} \sum_{i=1}^{n} \widehat{m}(X_i, h)' \widehat{m}(X_i, h) \geq K E\left[\widehat{m}(X, h)' \widehat{m}(X, h)\right].$$

Moreover, using the fact that $(a-b)^2 + b^2 \geq \frac{1}{2} a^2$ we have:

$$E\left[\widehat{m}(X, h)' \widehat{m}(X, h)\right] + E\left[(\widehat{m}(X, h) - m(X, h))'(\widehat{m}(X, h) - m(X, h))\right] \geq \frac{1}{2} E\left[m(X, h)' m(X, h)\right],$$

thus

$$n^{-1} \sum_{i=1}^{n} \widehat{m}(X_i, h)' \widehat{m}(X_i, h) \geq K \left\{ \frac{1}{2} E\left[m(X, h)' m(X, h)\right] - E\left[||\widehat{m}(X, h) - m(X, h)||_E^2\right] \right\}.$$

Again by the i.i.d. sample and assumption 3.8(ii), and using the fact that $(a+b)^2 \leq 2a^2 + 2b^2$, we have:

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{m}(X_i, \Pi_n h_0)' \widehat{m}(X_i, \Pi_n h_0) \leq 2K' \left\{ ||\widehat{m}(\cdot, \Pi_n h_0) - m(\cdot, \Pi_n h_0)||_X^2 + E[m(X, \Pi_n h_0)' m(X, \Pi_n h_0)] \right\}.$$

By assumption 3.8(i) for any consistent estimator $\widehat{m}$, (or assumptions 3.9 - 3.10 and Lemma B.3 for the series LS estimator $\widehat{m}$ with $\delta_{m,n}^2 = \frac{J_n}{n} + b_{m,J_n}^2$), we have:

$$\inf_{h \in \mathcal{H}_n} \left\{ -||\widehat{m}(\cdot, h) - m(\cdot, h)||_X^2 \right\} = - \sup_{h \in \mathcal{H}_n} ||\widehat{m}(\cdot, h) - m(\cdot, h)||_X^2 = O_P\left(\delta_{m,n}^2\right)$$

and

$$||\widehat{m}(\cdot, \Pi_n h_0) - m(\cdot, \Pi_n h_0)||_X^2 = O_P\left(\delta_{m,n}^2\right).$$

By assumption 3.4, we have: $\lambda_n \sup_{h \in \mathcal{H}_n} |\widehat{P}(h) - P(h)| = O_P(\lambda_n)$ and $\lambda_n |P(\Pi_n h_0) - P(h_0)| = O(\lambda_n)$. Thus, for all $\varepsilon > 0$,

$$\Pr\left(||\widehat{h}_n - h_0||_c > \varepsilon \right)$$
$$\leq \Pr\left(\begin{array}{c} \inf_{h \in \mathcal{H}_{k(n)} : ||h - h_0||_c \geq \varepsilon} \left\{ E\left[m(X, h)' m(X, h)\right] + \lambda_n P(h) \right\} \\ \leq K \left\{ E[m(X, \Pi_n h_0)' m(X, \Pi_n h_0)] + O_P\left(\delta_{m,n}^2\right) + \lambda_n P(\Pi_n h_0) + O_P(\lambda_n) \right\} \end{array} \right)$$
$$\leq \Pr\left(\begin{array}{c} \inf_{h \in \mathcal{H}_{k(n)} : ||h - h_0||_c \geq \varepsilon} \left\{ E\left[m(X, h)' m(X, h)\right] + \lambda_n P(h) \right\} \\ \leq K \left\{ E[m(X, \Pi_n h_0)' m(X, \Pi_n h_0)] + O_P\left(\delta_{m,n}^2\right) + \lambda_n P(h_0) + O_P(\lambda_n) \right\} \end{array} \right).$$

By assumption 3.11, we have: $\lambda_n \sup_{h \in \mathcal{H}_n} |\widehat{P}(h) - P(h)| = o_P(\lambda_n)$ and $\lambda_n |P(\Pi_n h_0) - P(h_0)| = o(\lambda_n)$. Thus, for all $\varepsilon > 0$,

$$\Pr\left(||\widehat{h}_n - h_0||_c > \varepsilon \right)$$
$$\leq \Pr\left(\begin{array}{c} \inf_{h \in \mathcal{H}_{k(n)} : ||h - h_0||_c \geq \varepsilon} \left\{ E\left[m(X, h)' m(X, h)\right] + \lambda_n P(h) \right\} \\ \leq K \left\{ E[m(X, \Pi_n h_0)' m(X, \Pi_n h_0)] + O_P\left(\delta_{m,n}^2\right) + \lambda_n P(\Pi_n h_0) + o_P(\lambda_n) \right\} \end{array} \right)$$
$$\leq \Pr\left(\begin{array}{c} \inf_{h \in \mathcal{H}_{k(n)} : ||h - h_0||_c \geq \varepsilon} \left\{ E\left[m(X, h)' m(X, h)\right] + \lambda_n P(h) \right\} \\ \leq K \left\{ E[m(X, \Pi_n h_0)' m(X, \Pi_n h_0)] + O_P\left(\delta_{m,n}^2\right) + \lambda_n P(h_0) + o_P(\lambda_n) \right\} \end{array} \right).$$

Thus we obtain results (1) and (2). Q.E.D.

PROOF OF LEMMA 3.1: By definition of $\widehat{h}_n$, we have for any $\lambda_n > 0$,

$$\lambda_n \widehat{P}_n(\widehat{h}_n) \leq \frac{1}{n} \sum_{i=1}^{n} ||\widehat{m}(X_i, \widehat{h}_n)||_E^2 + \lambda_n \widehat{P}_n(\widehat{h}_n) \leq \frac{1}{n} \sum_{i=1}^{n} ||\widehat{m}(X_i, \Pi_n h_0)||_E^2 + \lambda_n \widehat{P}_n(\Pi_n h_0),$$

and

$$\lambda_n\{P(\widehat{h}_n) - P(h_0)\} + \lambda_n\{\widehat{P}_n(\widehat{h}_n) - P(\widehat{h}_n)\}$$
$$\leq \quad \frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i, \Pi_n h_0)||_E^2 + \lambda_n\{\widehat{P}_n(\Pi_n h_0) - P(\Pi_n h_0)\} + \lambda_n\{P(\Pi_n h_0) - P(h_0)\}.$$

Thus

$$\lambda_n\{P(\widehat{h}_n) - P(h_0)\}$$
$$\leq \quad \frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i, \Pi_n h_0)||_E^2 + 2\lambda_n \sup_{h\in\mathcal{H}_n}\left|\widehat{P}_n(h) - P(h)\right| + \lambda_n\left|P(\Pi_n h_0) - P(h_0)\right|$$
$$\leq \quad O_P\left(\delta_{m,n}^2 + E[||m(X, \Pi_n h_0)||_E^2]\right) + 2\lambda_n \sup_{h\in\mathcal{H}_n}\left|\widehat{P}_n(h) - P(h)\right| + \lambda_n\left|P(\Pi_n h_0) - P(h_0)\right|$$

where the last inequality is due to assumption 3.8 holds for $h = \Pi_n h_0$.

Therefore, for all $M > 0$,

$$\Pr\left(P(\widehat{h}_n) - P(h_0) > M\right)$$
$$= \quad \Pr\left(\lambda_n\{P(\widehat{h}_n) - P(h_0)\} > \lambda_n M\right)$$
$$\leq \quad \Pr\left(O_P\left(\delta_{m,n}^2 + E[||m(X, \Pi_n h_0)||_E^2]\right) + 2\lambda_n \sup_{h\in\mathcal{H}_n}\left|\widehat{P}_n(h) - P(h)\right| + \lambda_n\left|P(\Pi_n h_0) - P(h_0)\right| > \lambda_n M\right).$$

(1) Under assumption 3.4, $\lambda_n \sup_{h\in\mathcal{H}_n}\left|\widehat{P}_n(h) - P(h)\right| + \lambda_n\left|P(\Pi_n h_0) - P(h_0)\right| = O_P(\lambda_n)$, therefore

$$\Pr\left(P(\widehat{h}_n) - P(h_0) > M\right) \leq \Pr\left(O_P(\frac{\delta_{m,n}^2 + E[||m(X, \Pi_n h_0)||_E^2]}{\lambda_n}) + O_P(1) > M\right)$$

which, under $\max\{\delta_{m,n}^2, E[||m(X, \Pi_n h_0)||_E^2]\} = O(\lambda_n)$, goes to zero as $M \to \infty$. Thus $P(\widehat{h}_n) - P(h_0) = O_P(1)$. Since $0 \leq P(h_0) < \infty$ we have: $P(\widehat{h}_n) = O_P(1)$.

(2) Under assumption 3.11, $\lambda_n \sup_{h\in\mathcal{H}_n}\left|\widehat{P}_n(h) - P(h)\right| + \lambda_n\left|P(\Pi_n h_0) - P(h_0)\right| = o_P(\lambda_n)$, therefore

$$\Pr\left(P(\widehat{h}_n) - P(h_0) > M\right) \leq \Pr\left(O_P(\frac{\delta_{m,n}^2 + E[||m(X, \Pi_n h_0)||_E^2]}{\lambda_n}) + o_P(1) > M\right)$$

which, under $\max\{\delta_{m,n}^2, E[||m(X, \Pi_n h_0)||_E^2]\} = o(\lambda_n)$, goes to zero for all $M > 0$. Thus $P(\widehat{h}_n) - P(h_0) \leq o_P(1)$. Q.E.D.

PROOF OF THEOREM 3.2: By Lemma B.4(1), assumption 3.12 and $\lambda_n P(h) \geq 0$, we have: for all $\varepsilon > 0$,

$$\Pr\left(||\widehat{h}_n - h_0||_c > \varepsilon\right)$$
$$\leq \quad \Pr\left(\begin{array}{c}\inf_{h\in\mathcal{H}_{k(n)}:||h-h_0||_c\geq\varepsilon}\{B(k(n))g_m(||h - h_0||_c) + \lambda_n P(h)\} \\ \leq K\{E[m(X, \Pi_n h_0)'m(X, \Pi_n h_0)] + O_P\left(\delta_{m,n}^2\right) + \lambda_n P(h_0) + O_P(\lambda_n)\}\end{array}\right)$$
$$\leq \quad \Pr\left(g_m(\varepsilon) \leq \frac{K\{E[m(X, \Pi_n h_0)'m(X, \Pi_n h_0)] + O_P\left(\delta_{m,n}^2\right) + \lambda_n P(h_0) + O_P(\lambda_n)\}}{B(k(n))}\right)$$

which goes to zero under $\max\{\delta_{m,n}^2, E(||m(X, \Pi_n h_0)||_E^2), \lambda_n\} = o(B(k(n)))$. Q.E.D.

PROOF OF THEOREM 3.3: By Lemma B.4(1), for all $\varepsilon > 0$,

$$\Pr\left(||\widehat{h}_n - h_0||_c > \varepsilon\right) \leq \Pr\left(\begin{array}{c} \inf_{h \in \mathcal{H}_n : ||h - h_0||_c \geq \varepsilon}\{E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h)\} \\ \leq K\left\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)] + O_P\left(\delta_{m,n}^2\right) + \lambda_n P(h_0) + O_P(\lambda_n)\right\} \end{array}\right).$$

Recall that $E[m(X,h)'m(X,h)]$ is lower semicontinuous on $\mathcal{H}$ under $||\cdot||_c$ (assumption 3.7(ii)) iff the set $\{h \in \mathcal{H} : E[m(X,h)'m(X,h)] \leq M\}$ is closed in $||\cdot||_c$ relatively to $\mathcal{H}$ for all $-\infty < M < \infty$. Given assumption 3.13, the set $\{h \in \mathcal{H}_n : ||h - h_0||_c \geq \varepsilon, E[m(X,h)'m(X,h)] \leq M, P(h) \leq \overline{M}\}$ is compact under $||\cdot||_c$ for all $M, \overline{M} < \infty$. Theorem 38.B of Zeidler (1985) now implies that the minimum problem,

$$\min_{h \in \mathcal{H}_n : ||h - h_0||_c \geq \varepsilon}\{E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h)\}$$

has a solution, h_n, which belongs to the set:

$$\left\{\begin{array}{c} h \in \mathcal{H}_n : ||h - h_0||_c \geq \varepsilon, \\ E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h) \leq K\left\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)] + \delta_{m,n}^2 + \lambda_n P(h_0) + O(\lambda_n)\right\} \end{array}\right\}.$$

Since $E\left[m(X,h)'m(X,h)\right] \geq 0$, $\lambda_n > 0$ and $\mathcal{H}_n \subseteq \mathcal{H}$, we have that the sequence $\{h_n\}$ belongs to the set

$$\left\{h \in \mathcal{H} : ||h - h_0||_c \geq \varepsilon, P(h) \leq K\left(\frac{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)] + \delta_{m,n}^2}{\lambda_n} + O(1)\right)\right\},$$

which is compact under $||\cdot||_c$ by assumption 3.13(i), the fact that $\{h \in \mathcal{H} : ||h - h_0||_c \geq \varepsilon\}$ is closed, and that $\max\{\delta_{m,n}^2, E(||m(X,\Pi_n h_0)||_E^2)\} = O(\lambda_n)$. Therefore, the sequence $\{h_n\}$ must have a further subsequence, denoted as $\{h_{n_k}\}$, that converges to a limit h_∞ in $||\cdot||_c$ and $h_\infty \in \{h \in \mathcal{H} : ||h - h_0||_c \geq \varepsilon, P(h) \leq \overline{M}\}$ for some $\overline{M} \in [0, +\infty)$. Moreover, by assumption 3.7(ii) and $P(h) \geq 0$, we have:

$$0 \leq E\left[m(X,h_\infty)'m(X,h_\infty)\right] \leq \liminf_n E\left[m(X,h_n)'m(X,h_n)\right]$$
$$\leq \liminf_n K\left\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)] + \delta_{m,n}^2 + O(\lambda_n) + \lambda_n P(h_0)\right\} = 0.$$

This and assumption 3.1(iii) together imply that $||h_\infty - h_0||_c = 0$, which contradicts to $h_\infty \in \{h \in \mathcal{H} : ||h - h_0||_c \geq \varepsilon, P(h) \leq \overline{M}\}$. Thus $||\widehat{h}_n - h_0||_c = o_P(1)$. Lemma 3.1 (1) implies $P(\widehat{h}_n) = O_P(1)$. Q.E.D.

PROOF OF THEOREM 3.4: By Lemma B.4(2) and assumption 3.14, we have: for all $\varepsilon > 0$,

$$\Pr\left(||\widehat{h}_n - h_0||_c > \varepsilon\right)$$
$$\leq \Pr\left(\begin{array}{c} \inf_{h \in \mathcal{H}_{k(n)} : ||h - h_0||_c \geq \varepsilon}\{E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h)\} \\ \leq K\left\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)] + O_P\left(\delta_{m,n}^2\right) + \lambda_n P(h_0) + o_P(\lambda_n)\right\} \end{array}\right).$$

Case 1: If $\liminf_{h \in \mathcal{H}_{k(n)} : ||h - h_0||_c \geq \varepsilon}\{E\left[m(X,h)'m(X,h)\right]\} \geq const. > 0$, then since $\lambda_n P(h) \geq 0$,

$$\liminf_{h \in \mathcal{H}_{k(n)} : ||h - h_0||_c \geq \varepsilon}\{E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h)\} \geq const. > 0$$

we have $\Pr\left(||\widehat{h}_n - h_0||_c > \varepsilon\right) \to 0$ as long as $\max\{\delta_{m,n}^2, E(||m(X,\Pi_n h_0)||_E^2)\} = o(1)$ and $\lambda_n = o(1)$.

43

Case 2: If $\liminf_{h\in\mathcal{H}_{k(n)}:||h-h_0||_c\geq\varepsilon}\{E[m(X,h)'m(X,h)]\}=0$, by assumption 3.15, for all the sequences $\{h_n\in\mathcal{H}_{k(n)}:||h_n-h_0||_c\geq\varepsilon\}$ with $\liminf_{h\in\mathcal{H}_{k(n)}:||h-h_0||_c\geq\varepsilon}\{E[m(X,h)'m(X,h)]\}=0$, we have $\liminf_n\langle t_0,h_n-h_0\rangle_{\mathbf{H}^*,\mathbf{H}}\geq 0$. This and assumption 3.14 imply:

$$\liminf_{h\in\mathcal{H}_{k(n)}:||h-h_0||_c\geq\varepsilon}\{E[m(X,h)'m(X,h)]+\lambda_n(P(h)-P(h_0))\}$$

$$\geq\quad\liminf_{h\in\mathcal{H}_{k(n)}:||h-h_0||_c\geq\varepsilon}\{E[m(X,h)'m(X,h)]+\lambda_n\langle t_0,h-h_0\rangle_{\mathbf{H}^*,\mathbf{H}}+\lambda_n g(||h-h_0||_c)\}$$

$$\geq\quad\liminf_{h\in\mathcal{H}_{k(n)}:||h-h_0||_c\geq\varepsilon}\{E[m(X,h)'m(X,h)]+\lambda_n g(||h-h_0||_c)\}.$$

Thus, for all $\varepsilon>0$, for result (a), since $E[m(X,h)'m(X,h)]\geq 0$,

$$\Pr\left(||\widehat{h}_n-h_0||_c>\varepsilon\right)$$

$$\leq\quad\Pr\left(\begin{array}{c}\inf_{h\in\mathcal{H}_n:||h-h_0||_c\geq\varepsilon}\{E[m(X,h)'m(X,h)]+\lambda_n(P(h)-P(h_0))\}\\\leq K\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)]+O_P(\delta_{m,n}^2)+o_P(\lambda_n)\}\end{array}\right)$$

$$\leq\quad\Pr\left(\lambda_n g(\varepsilon)\leq K\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)]+O_P(\delta_{m,n}^2)+o_P(\lambda_n)\}\right)$$

$$\rightarrow\quad 0\quad\text{if }\max\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)],\delta_{m,n}^2\}=o(\lambda_n);$$

For result (b), under the additional assumption 3.12,

$$\Pr\left(||\widehat{h}_n-h_0||_c>\varepsilon\right)$$

$$\leq\quad\Pr\left(B(k(n))g_m(\varepsilon)+\lambda_n g(\varepsilon)\leq K\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)]+O_P(\delta_{m,n}^2)+o_P(\lambda_n)\}\right)$$

$$\rightarrow\quad 0\quad\text{if }\max\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)],\delta_{m,n}^2\}=o(\max\{B(k(n)),\lambda_n\}).$$

Thus $||\widehat{h}_n-h_0||_c=o_P(1)$. This and assumption 3.14 imply $P(\widehat{h}_n)-P(h_0)\geq o_P(1)$. But Lemma 3.1 (2) also implies $P(\widehat{h}_n)-P(h_0)\leq o_P(1)$. Thus $P(\widehat{h}_n)-P(h_0)=o_P(1)$. Q.E.D.

PROOF OF COROLLARY 3.2: It suffices to show that assumptions 3.16, 3.17 and 3.18 imply that assumptions 3.15 and 3.3 of Theorem 3.4 hold. First, assumptions 3.2(i) and 3.16(i)(iii) imply that every sequence $\{h_k\in\mathcal{H}_k\}$ has a weakly convergent sub-sequence in $\mathcal{H}$, denoting its limit as h_∞, then $h_\infty\in\mathcal{H}$ by assumption 3.16(ii) and Zeidler (1985, corollary 38.8). By assumption 3.17 and $\liminf_{k\to\infty}E[m(X,h_k)'m(X,h_k)]=0$, we have $E[m(X,h_\infty)'m(X,h_\infty)]=0$. This and assumption 3.1(iii) imply $h_\infty=h_0$; hence assumption 3.15 holds with $c=0$. Next, under assumptions 3.1(i), 3.2(i) and 3.8, by theorem 38.A and corollary 38.8 of Zeidler (1985), we have that assumptions 3.16, 3.17 and 3.18 imply assumption 3.3. Finally, we need to establish the claim in Remark 3.2. Under assumptions 3.16 and 3.17', any weakly convergent sequence $\{h_k:k\}$ to h_∞ in $\mathcal{H}$ has an associated convergent sub-sequence $\{m(\cdot,h_k):k\}$ to $m(\cdot,h_\infty)$ in $L^2(f_X)$, since the functional $E[m(X,h)'m(X,h)]:m\in L^2(f_X)\rightarrow[0,+\infty]$ is convex and continuous in $m\in L^2(f_X)$, it follows that $E[m(X,h_k)'m(X,h_k)]\rightarrow E[m(X,h_\infty)'m(X,h_\infty)]$ as $k\to\infty$; hence assumption 3.17 holds. By Remark B.1(2)(c), assumptions 3.16 and 3.17" imply that assumption 3.17 holds. Q.E.D.

PROOF OF THEOREM 3.5: By Lemma B.4(2), we have: for all $\varepsilon>0$,

$$\Pr\left(||\widehat{h}_n-h_0||_c>\varepsilon\right)$$

$$\leq\quad\Pr\left(\begin{array}{c}\inf_{h\in\mathcal{H}_{k(n)}:||h-h_0||_c\geq\varepsilon}\{E[m(X,h)'m(X,h)]+\lambda_n P(h)\}\\\leq K\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)]+O_P(\delta_{m,n}^2)+\lambda_n P(h_0)+o_P(\lambda_n)\}\end{array}\right).$$

Under assumptions 3.2(i), 3.11, 3.16(i)(ii)(iii)', 3.17" and 3.18(b), the infimum,

$$\inf_{h\in\mathcal{H}_n:||h-h_0||_c\geq\varepsilon}\left\{E\left[m(X,h)'m(X,h)\right]+\lambda_n P(h)\right\}$$

exists. If the set $\{E\left[m(X,h)'m(X,h)\right]+\lambda_n P(h)\leq K\left\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)]+\lambda_n P(h_0)+\delta_{m,n}^2+o_P(\lambda_n)\right\}\}$ is empty for all n big enough the desired result will trivially follow. Thus we assume that this set is not empty infinitely often, i.e., it exists a subsequence $(h_n)_n$ that belongs to the aforementioned set and to $\{h\in\mathcal{H}_n:||h-h_0||_c\geq\varepsilon\}$. By assumptions 3.1(ii), 3.16, 3.18(b) and Remark 3.2, it follows that there exists a weak convergence subsequence, denoted as $\{h_{n_k}\}_k$ with weak limit $h_\infty\in\mathcal{H}$. By assumptions 3.16 and 3.17" and Remark 3.2, we have: $E[||m(X,h_\infty)||_E^2]\leq\liminf E[||m(X,h_{n_k})||_E^2]$, and

$$\begin{aligned}0\leq E[||m(X,h_\infty)||_E^2]+\lambda_0 P(h_\infty)&\leq\underline{\lim}\left(E[||m(X,h_{n_k})||_E^2]+\lambda_{n_k}P(h_{n_k})\right)\\&\leq\underline{\lim}K\left(E[||m(X,\Pi_{n_k}h_0)||_E^2]+\lambda_{n_k}P(\Pi_{n_k}h_0)+\delta_{m,n_k}^2\right)\\&\leq K\left(E[||m(X,h_0)||_E^2]+\lambda_0 P(h_0)\right),\end{aligned}$$

where the first inequality follows from assumptions 3.17" and 3.19(ii) and the last follows from the fact that $\lambda_n=\lambda_0+o(1)$ and assumptions 3.2(ii) and 3.11. Assumption 3.19(i) then implies $||h_\infty-h_0||_c=0$. Moreover, all of such weak convergence subsequences having their limits satisfying $||h_\infty-h_0||_c=0$. Thus we have $\liminf_{h_n\in\mathcal{H}_n:||h-h_0||_c\geq\varepsilon}\{\langle t_0,h_n-h_0\rangle_{\mathbf{H}^*,\mathbf{H}}\}=\langle t_0,h_\infty-h_0\rangle_{\mathbf{H}^*,\mathbf{H}}=0$ for t_0 defined in assumption 3.14. Hence, for all $\varepsilon>0$,

$$\begin{aligned}&\Pr\left(||\widehat{h}_n-h_0||_c>\varepsilon\right)\\\leq\quad&\Pr\left(\begin{array}{c}\liminf_{h_n\in\mathcal{H}_n:||h-h_0||_c\geq\varepsilon}\lambda_n\left\{\langle t_0,h_n-h_0\rangle_{\mathbf{H}^*,\mathbf{H}}+g(||h_n-h_0||_c)\right\}\\\leq K\left\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)]+O_P\left(\delta_{m,n}^2\right)+o_P(\lambda_n)\right\}\end{array}\right)\\\leq\quad&\Pr\left(\lambda_n g(\varepsilon)\leq K\left\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)]+O_P\left(\delta_{m,n}^2\right)+o_P(\lambda_n)\right\}\right)\\\rightarrow\quad&0\quad\text{if }\max\left\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)],\delta_{m,n}^2\right\}=o(\lambda_n),\end{aligned}$$

where the last inequality is due to assumption 3.14, and the result now follows. For case (b) the proof is completely analogous to the one in theorem 3.4.

Finally, $||\widehat{h}_n-h_0||_c=o_P(1)$ and assumption 3.14 (or $P(h)$ is lower semicontinuous at h_0) imply $P(\widehat{h}_n)-P(h_0)\geq o_P(1)$ But Lemma 3.1 (2) also implies $P(\widehat{h}_n)-P(h_0)\leq o_P(1)$. Thus $P(\widehat{h}_n)-P(h_0)=o_P(1)$. Q.E.D.

C Convergence Rate

PROOF OF THEOREM 4.1: (1) Let $r_n^2=\max\{\delta_{m,n}^2,||\Pi_n h_0-h_0||^2,\lambda_n|P(\Pi_n h_0)-P(\widehat{h}_n)|\}=o_P(1)$. Since $\widehat{h}_n\in\mathcal{H}_{osn}$ with probability approaching one, we have: for all $M>1$,

$$\begin{aligned}&\Pr\left(\frac{||\widehat{h}_n-h_0||}{r_n}\geq M\right)\\\leq\quad&\Pr\left(\inf_{\{h\in\mathcal{H}_{osn}:||h-h_0||\geq Mr_n\}}\{||\widehat{m}(\cdot,h)||_{n,X}^2+\lambda_n P(h)\}\leq||\widehat{m}(\cdot,\Pi_n h_0)||_{n,X}^2+\lambda_n P(\Pi_n h_0)\right).\end{aligned}$$

By assumption 3.8, we have:

$$(1 - o_P(1))||m(\cdot, \widehat{h}_n)||_X^2 + \lambda_n P(\widehat{h}_n) \leq O_P(\delta_{m,n}^2) + (1 + o_P(1))||m(\cdot, \Pi_n h_0)||_X^2 + \lambda_n P(\Pi_n h_0). \quad \text{(C.1)}$$

which implies

$$(1 - o_P(1))||m(\cdot, \widehat{h}_n)||_X^2 \leq O_P(\delta_{m,n}^2) + (1 + o_P(1))||m(\cdot, \Pi_n h_0)||_X^2 + \lambda_n |P(\Pi_n h_0) - P(\widehat{h}_n)|.$$

This, $||\widehat{h}_n - h_0||_s = o_P(1)$ and assumption 4.1 imply that

$$\Pr\left(\frac{||\widehat{h}_n - h_0||}{r_n} \geq M\right) \leq \Pr\left(M^2 r_n^2 \leq O_P\left\{\delta_{m,n}^2, ||\Pi_n h_0 - h_0||^2, \lambda_n |P(\Pi_n h_0) - P(\widehat{h}_n)|\right\}\right),$$

which, given our choice of r_n, goes to zero as $M \to \infty$; hence $||\widehat{h}_n - h_0|| = O_P(r_n)$ and Theorem 4.1(1) follows.

(2) Using the same argument as that for result (1), we still have inequality (C.1) holds. By assumption 4.2, $\lambda_n\left(P(\widehat{h}_n) - P(\Pi_n h_0)\right) \geq \lambda_n \langle t_0, \widehat{h}_n - \Pi_n h_0 \rangle_{\mathbf{H}^*, \mathbf{H}}$ and thus

$$(1 - o_P(1))||m(\cdot, \widehat{h}_n)||_X^2 + \lambda_n \langle t_0, \widehat{h}_n - \Pi_n h_0 \rangle_{\mathbf{H}^*, \mathbf{H}} \leq O_P(\delta_{m,n}^2) + (1 + o_P(1))||m(\cdot, \Pi_n h_0)||_X^2,$$

thus

$$(1 - o_P(1))||m(\cdot, \widehat{h}_n)||_X^2 \leq O_P(\delta_{m,n}^2) + (1 + o_P(1))||m(\cdot, \Pi_n h_0)||_X^2 + const.\lambda_n ||\widehat{h}_n - \Pi_n h_0||_s$$

By assumption 4.1, theorem 4.1(2) follows by choosing $r_n^2 = \max\{\delta_{m,n}^2, ||\Pi_n h_0 - h_0||^2, \lambda_n||\widehat{h}_n - \Pi_n h_0||_s\} = o_P(1)$. Q.E.D.

PROOF OF THEOREM 4.2: It directly follows from theorem 4.1(1), assumption 4.3 and the definition of $\omega_n(\delta, \mathcal{H}_{osn})$. Q.E.D.

PROOF OF THEOREM 4.3: By setting $\lambda_n = O\left(\frac{\delta_{m,n}^2}{||\widehat{h}_n - \Pi_n h_0||_s}\right)$ the result directly follows from theorem 4.1(2), assumption 4.3 and the definitions of $\omega_n(\delta, \mathcal{H}_{osn})$ and $\omega(\delta, \mathcal{H}_{os})$. Q.E.D.

PROOF OF COROLLARY 4.1: Under the stated condition, we can replace $\widehat{\lambda}_n \widehat{P}_n(h)$ by $\lambda_n P(h)(1 + o_P(1))$ uniformly over $h \in \mathcal{H}_{osn}$. It is then easy to check that all the theorems still hold under their respective assumptions. Q.E.D.

PROOF OF LEMMA 5.1: Result (1) follows directly from the definition of $\omega_n(\delta, \mathcal{H}_{osn})$, as well as the fact that for any $h \in \mathcal{H}_{osn}$, under assumption 5.1,

$$C^{-1}||h||_s^2 = \sum_{j \leq k(n)} |\langle h, q_j \rangle_s|^2 \leq (\max_{j \leq k(n)} b_j^{-1}) \sum_{j \leq k(n)} b_j |\langle h, q_j \rangle_s|^2 \leq \frac{1}{cb_{k(n)}}||h||^2,$$

where the last inequality is due to assumption 5.2(i) and $\{b_j\}$ non-increasing. Similarly, assumption 5.2(ii) implies Result (2) since

$$
\begin{aligned}
||h_0 - \Pi_n h_0||_s^2 &= c \sum_{j > k(n)} |\langle h_0 - \Pi_n h_0, q_j \rangle_s|^2 \\
&\geq c(\min_{j > k(n)} b_j^{-1}) \sum_{j > k(n)} b_j |\langle h_0 - \Pi_n h_0, q_j \rangle_s|^2 \geq \frac{c'}{b_{k(n)}}||h_0 - \Pi_n h_0||^2.
\end{aligned}
$$

Result (3) directly follows from results (1) and (2). *Q.E.D.*

PROOF OF LEMMA 5.2: For any $h \in \mathcal{H}_{os}$ with $||h||^2 \leq O(\delta^2)$, and for any $k \geq 1$, assumptions 5.1, 5.4(i) and 5.3 imply that

$$
\begin{aligned}
C^{-1}||h||_s^2 &= \sum_{j \leq k} \langle h, q_j \rangle_s^2 + \sum_{j > k} \langle h, q_j \rangle_s^2 \\
&\leq (\max_{j \leq k} b_j^{-1}) \sum_j b_j \langle h, q_j \rangle_s^2 + M^2 (\nu_{k+1})^{-2\gamma_h} \leq \frac{1}{c} b_k^{-1} \delta^2 + M^2 (\nu_{k+1})^{-2\gamma_h}.
\end{aligned}
$$

Given that $M > 0$ is a fixed finite number and δ goes to zero as n increases, we can assume $M^2 (\nu_2)^{-2\gamma_h} > \frac{1}{c} \delta^2 / b_1$, which will be satisfied for big enough n. Since $\{b_j\}$ is non-increasing and $\{\nu_j\}_{j=1}^{\infty}$ is non-decreasing in $j \geq 1$, we have: there is a $k^* > 1$ such that

$$
\frac{\delta^2}{b_{k^*-1}} < cM^2 (\nu_{k^*})^{-2\gamma_h} \quad \text{and} \quad \frac{\delta^2}{b_{k^*}} \geq cM^2 (\nu_{k^*})^{-2\gamma_h} \geq cM^2 (\nu_{k^*+1})^{-2\gamma_h},
$$

and

$$
\omega(\delta, \mathcal{H}_{os}) = \sup_{h \in \mathcal{H}_{os}: ||h|| \leq \delta} ||h||_s \leq const. \frac{\delta}{\sqrt{b_{k^*}}}.
$$

The result follows. *Q.E.D*

D   Applications

PROOF OF PROPOSITION 6.1: For the nonparametric additive quantile IV model (6.1), we apply Corollary 3.1 by verifying all its assumptions are satisfied. First, for both Results (1) and (2), Assumptions 3.1(i)(iii) are directly assumed. Assumptions 3.4 and 3.6(ii) are trivially satisfied with $\lambda_n = 0$, or with $\lambda_n > 0$, $\lambda_n = o(1)$, $\widehat{P}_n(h) = P(h) = ||\varpi h_2||_{\mathcal{T}_{p_2, q_2}^{s_2}}$, and $\mathcal{H}^2 = \{h_2 \in L^2(\mathcal{R}^d, f_{Y_2}) : ||\varpi h_2||_{\mathcal{T}_{p,q}^{\gamma_2}} \leq M_0\}$ for $\gamma_2 > 0$ and a known constant $M_0 < \infty$. This is because the embedding of $\mathcal{H}^2$ into the set $\{h_2 \in L^2(\mathcal{R}^d, f_{Y_2}) : ||\varpi h_2||_{\mathcal{T}_{p_2, q_2}^{s_2}} < \infty\}$ is continuous as long as $s_2 \in [0, \gamma_2 - d(p^{-1} - p_2^{-1})]$ and $q \leq q_2$ (see Edmunds and Triebel, 1996, chapter 4). Thus, $P(h) = ||\varpi h_2||_{\mathcal{T}_{p_2, q_2}^{s_2}} \leq const.$ uniformly in $h_2 \in \mathcal{H}^2$ and $\mathcal{H}_n^2 \subseteq \mathcal{H}^2$.

For Result (1): Assumption 3.1(ii) is automatically satisfied with the choice of the spaces $\mathcal{H} = \Lambda_1^{\gamma_1}([0,1]^d) \times \mathcal{H}^2$ and $\mathbf{H} = L^\infty([0,1]^d) \times \{h_2 : \sup_{y_2} |h_2(y_2) w(y_2)| < \infty\}$ with the norm $||h||_c = \sup_{y_1} |h(y_1)| + \sup_{y_2} |h_2(y_2) w(y_2)|$. Moreover, the embedding of $\mathcal{H}$ into $\mathbf{H}$ is compact under the norm $|| \cdot ||_c$ with $\gamma_1 > 0$, $\gamma_2 > d/p$, $\theta > \vartheta \geq 0$ (which implies $\frac{w}{\varpi} \asymp \frac{(1+|y_2|)^{-\theta}}{(1+|y_2|)^{-\vartheta}} \to 0$ as $|y_2| \to \infty$; see Edmunds and Triebel, 1996, chapter 4). Given the choice of the sieve space $\mathcal{H}_n$ and the definition of $|| \cdot ||_c$, we have $\sup_{h \in \mathcal{H}} ||h - \Pi_n h||_c = o(1)$, which implies assumption 3.2(i). For assumption 3.2(ii), notice that

$$
\begin{aligned}
&m(X, h) - m(X, h_0) \\
&= E[F_{Y_3|Y_1,Y_2,X}(h_1(Y_1) + h_2(Y_2)) - F_{Y_3|Y_1,Y_2,X}(h_{01}(Y_1) + h_{02}(Y_2))|X] \\
&= E\{f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2))[h_1(Y_1) - h_{01}(Y_1) + h_2(Y_2) - h_{02}(Y_2)]|X\},
\end{aligned}
$$

thus

$$|m(X,h) - m(X,h_0)|$$

$$\leq \quad E[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2))|X] \times \sup_{y_1} |h_1(y_1) - h_{01}(y_1)|$$

$$+ E[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2)) \frac{1}{w(Y_2)}|X] \times \sup_{y_2} |[h_2(y_2) - h_{02}(y_2)]w(y_2)|.$$

Since $m(X,h_0) = 0$ and $|m(X,h)| = |E[F_{Y_3|Y_1,Y_2,X}(h_1(Y_1) + h_2(Y_2))|X]| \leq 1$ for all h for almost all X, we have

$$E[|m(X,h)|^2] \leq E[|m(X,h) - m(X,h_0)|] \leq E[\sup_{y_3} f_{Y_3|Y_1,Y_2,X}(y_3)\{1 + \frac{1}{w(Y_2)}\}] \times \|h - h_0\|_c.$$

Thus condition 6.1(ii)(iii) and $\|\Pi_n h_0 - h_0\|_c = o(1)$ imply

$$E[|m(X,\Pi_n h_0)|^2] \leq E[\sup_{y_3} f_{Y_3|Y_1,Y_2,X}(y_3)\{1 + \frac{1}{w(Y_2)}\}] \times \|\Pi_n h_0 - h_0\|_c = o(1)$$

hence assumption 3.2(ii) holds. Assumptions 3.6(i) and 3.7(i) follow directly from our choices of $\mathcal{H}$, $\mathcal{H}_n$ and $\|\cdot\|_c$. For assumptions 3.7(ii) and 3.5(i)(ii), notice that for all $h, h' \in \mathcal{H}$,

$$|m(X,h) - m(X,h')|$$

$$\leq \quad E[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2))|X] \times \sup_{y_1} |h_1(y_1) - h'_1(y_1)|$$

$$+ E[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2)) \frac{1}{w(Y_2)}|X] \times \sup_{y_2} |[h_2(y_2) - h'_2(y_2)]w(y_2)|,$$

condition 6.1(ii)(iii) imply assumption 3.5(i) holds with $b(X) = E[\sup_{y_3} f_{Y_3|Y_1,Y_2,X}(y_3)\{1 + \frac{1}{w(Y_2)}\}|X]$ and $\kappa = 1$. Since $|m(X,h)| \leq 1$ for all h for almost all X, assumption 3.5(ii.a) and 3.7(ii) are satisfied. To establish assumptions 3.5(iii), it suffices to show that

$$(i) \quad \sup_{h \in \mathcal{H}_n} n^{-1} \sum_{i=1}^{n} (\widehat{m}(X_i,h) - \widetilde{m}(X_i,h))^2 = O_P\left(\frac{J_n}{n}\right) = o_P(1),$$

and

$$(ii) \quad \sup_{h \in \mathcal{H}_n} n^{-1} \sum_{i=1}^{n} (\widetilde{m}(X_i,h) - m(X_i,h))^2 = o_P(1).$$

For claim (i), assumption 3.9 and the fact that $|\rho(Z,h)| \leq 1$ imply that all the conditions of Lemma B.3(1) are satisfied; hence claim (i) follows from Lemma B.3(1). Regarding claim (ii), for each $h \in \mathcal{H}_n$, $n^{-1} \sum_{i=1}^{n} (\widetilde{m}(X_i,h) - m(X_i,h))^2 = o_P(1)$ follows directly from conditions 6.1(i)(ii)(iv) and 6.3(i), and the LS projection approximation property of the sieve space $p^{J_n}(X)$ as $J_n \to \infty$. From the verification of assumption 3.5(i)(ii.a), compactness of $\mathcal{H}_n$ and $\mathcal{H}$, we have that $n^{-1} \sum_{i=1}^{n} (\widetilde{m}(X_i,h) - m(X_i,h))^2$ is stochastic equicontinous in $\mathcal{H}_n$; hence we obtain claim (ii). Thus all the assumptions of Corollary 3.1 are satisfied and Result (1) follows.

For Result (2): The verifications for Result (2) are essentially the same as those for Result (1). Here we only highlight the parts that are slightly different due to the different choice of $\mathbf{H}$ and $\|h\|_c$. Assumption 3.1(ii) is satisfied with the choice of the spaces $\mathcal{H} = \Lambda_1^{\gamma_1}([0,1]^d) \times \mathcal{H}^2$, and $\mathbf{H} = L^\infty([0,1]^d) \times \{h_2 : \|h_2 w\|_{L^2(\mathcal{R}^d,leb)} < \infty\}$ with the norm $\|h\|_c = \sup_{y_1} |h(y_1)| + \|h_2 w\|_{L^2(\mathcal{R}^d,leb)}$.

The embedding of $\mathcal{H}$ into $\mathbf{H}$ is compact under the norm $||\cdot||_c$ with $\gamma_1 > 0$, $\gamma_2 + d/2 > d/p$, $p^{-1} + (\theta - \vartheta)/d > 1/2$ (see Edmunds and Triebel, 1996, chapter 4). Thus we have $\sup_{h \in \mathcal{H}} ||h - \Pi_n h||_c = o(1)$ which implies assumption 3.2(i). For assumption 3.2(ii), notice that

$$
\begin{aligned}
&|m(X, h) - m(X, h_0)| \\
=\quad & |E[F_{Y_3|Y_1,Y_2,X}(h_1(Y_1) + h_2(Y_2)) - F_{Y_3|Y_1,Y_2,X}(h_{01}(Y_1) + h_{02}(Y_2))|X]| \\
\leq\quad & E[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2)) \times |h_1(Y_1) - h_{01}(Y_1)||X] \\
& + E\left\{ \frac{f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2))}{w(Y_2)} \times \{|h_2(Y_2) - h_{02}(Y_2)|w(Y_2)\}|X \right\}
\end{aligned}
$$

and that $|m(X, h)| \leq 1$, we have, under condition 6.1(ii) and $E[\frac{1}{w(Y_2)}]^2 < \infty$,

$$
\begin{aligned}
&E\{|m(X, h)|^2\} \leq E\{|m(X, h) - m(X, h_0)|\} \\
\leq\quad & \sup_{y_3} f_{Y_3|Y_1,Y_2,X}(y_3) \times ||h_1 - h_{01}||_{L^\infty([0,1]^d, leb)} \\
& + \sup_{y_3} f_{Y_3|Y_1,Y_2,X}(y_3) \times \sqrt{E[\frac{1}{w(Y_2)}]^2} \times ||w[h_2 - h_{02}]||_{L^2(\mathcal{R}^d, leb)},
\end{aligned}
$$

thus assumption 3.2(ii) is satisfied. For assumptions 3.7(ii) and 3.5(i)(ii), notice that for all $h, h' \in \mathcal{H}$,

$$
\begin{aligned}
&|m(X, h) - m(X, h')| \\
\leq\quad & E[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2))|X] \times \sup_{y_1} |h_1(y_1) - h'_1(y_1)| \\
& + E\left\{ \frac{f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2))}{w(Y_2)} \times \{|h_2(Y_2) - h'_2(Y_2)|w(Y_2)\}|X \right\},
\end{aligned}
$$

condition 6.1(ii) and $E[\frac{1}{w(Y_2)}]^2 < \infty$ imply assumption 3.5(i) holds with $b(X) = E[\sup_{y_3} f_{Y_3|Y_1,Y_2,X}(y_3)\{1 + [w(Y_2)]^{-2}\}|X]$ and $\kappa = 1$. Since $|m(X, h)| \leq 1$ for all h for almost all X, assumption 3.5(ii.a) and 3.7(ii) are satisfied. The rest of the verifications are the same as those for Result (1). *Q.E.D.*

PROOF OF PROPOSITION 6.2: For **(1)** We obtain the results by verifying that all the assumptions of Theorem 3.3 (lower semicompact penalty) are satisfied. First, assumption 3.9 is directly imposed. Assumption 3.10(ii) holds by the choice of the sieve basis for $p^{J_n}(X)$ and by condition 6.4 with $b_{m,J_n}^2 = J_n^{-2r_m}$. Next, following the proofs for Results (1) and (2) of Proposition 6.1, we have that for any $M < \infty$, the embedding of the set $\{h \in \mathcal{H} : P(h) = ||\varpi h_2||_{\mathcal{T}_{p,q}^{\gamma_2}} \leq M\}$ into $\mathbf{H}$ is compact under the norm $||\cdot||_c$; hence assumption 3.13 is satisfied. Given the choice of the sieve space $\mathcal{H}_n$ and the definition of $||\cdot||_c$, we have for $h_0 \in \mathcal{H}$,

$$
||h_0 - \Pi_n h_0||_c \leq c\{k_1(n)\}^{-\gamma_1/d} + c'_n\{k_2(n)\}^{-\gamma_2/d} = o(1),
$$

thus assumption 3.2(i) holds. Assumptions 3.2(ii), 3.7(ii) and 3.4(b) are already verified in the proof of Proposition 6.1. Now the results follow from Theorem 3.3 provided that $\max\{\delta_{m,n}^2, E[m(X, \Pi_n h_0)^2]\} =$

$O(\lambda_n)$. We already have $\delta^2_{m,n} = \frac{J_n}{n} + J_n^{-2r_m} = O(\lambda_n)$. By conditions 6.1(ii)(iii)(iv), we also have

$$
\begin{aligned}
&E\{\|m(X, \Pi_n h_0)\|^2_E\} \\
=\quad &E\{\left|E\left[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2))\{\Pi_n h_{01}(Y_1) - h_{01}(Y_1) + \Pi_n h_{02}(Y_2) - h_{02}(Y_2)\}|X\right]\right|^2\} \\
\leq\quad &E\left\{E\left(\left[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1) + \overline{h}_2(Y_2))\{\Pi_n h_{01}(Y_1) - h_{01}(Y_1) + \Pi_n h_{02}(Y_2) - h_{02}(Y_2)\}\right]^2 |X\right)\right\} \\
\leq\quad &CE\left\{[\Pi_n h_{01}(Y_1) - h_{01}(Y_1) + \Pi_n h_{02}(Y_2) - h_{02}(Y_2)]^2\right\} \\
\leq\quad &2CE\{[\Pi_n h_{01}(Y_1) - h_{01}(Y_1)]^2\} + 2CE\{[\Pi_n h_{02}(Y_2) - h_{02}(Y_2)]^2\} \\
\leq\quad &const. \|h_0 - \Pi_n h_0\|^2_c = O\left(\max\left[\{k_1(n)\}^{-2\gamma_1/d}, \{k_2(n)\}^{-2\gamma_2/d}\right]\right) = O(\lambda_n),
\end{aligned}
$$

the result now follows. For **(2)**, it directly follows from Lemma 3.1 (1). *Q.E.D.*

PROOF OF PROPOSITION 6.3: We obtain the results by verifying that all the assumptions of Corollary 3.2 (convex penalty) are satisfied. Again assumptions 3.9 and 3.10 hold with $b^2_{m,J_n} = J_n^{-2r_m}$. Assumptions 3.1(i)(iii) are already assumed, and assumption 3.1(ii) holds trivially given the choice of the norm $\|h\|_c = \sup_{y_1} |h(y_1)| + \|h_2 w\|_{L^2(\mathcal{R}^d, leb)}$ for the spaces $\mathcal{H} = \Lambda_1^{\gamma_1}([0,1]^d) \times \mathcal{H}^2 \subset \mathbf{H} = L^\infty([0,1]^d) \times \{h_2 : \|h_2 w\|_{L^2(\mathcal{R}^d, leb)} < \infty\}$. By the choice of the spaces $\mathcal{H}_n$ and $\mathcal{H}$, we have:

$$\|\Pi_n h_{01} - h_{01}\|_{L^2(f_{Y_1})} \leq \sup_{h_1 \in \mathcal{H}^1} \|\Pi_n h_1 - h_1\|_{L^2(f_{Y_1})} \leq \sup_{h_1 \in \mathcal{H}^1} \sup_{y_1} |\Pi_n h_1(y_1) - h_1(y_1)| \leq c\{k_1(n)\}^{-\gamma_1/d},$$

$$\|\Pi_n h_{02} - h_{02}\|_{L^2(f_{Y_2})} \leq \sqrt{\sup_{y_2} \frac{f_{Y_2}(y_2)}{w^2(y_2)}} \times \|w(\Pi_n h_{02} - h_{02})\|_{L^2(\mathcal{R}^d, leb)} \leq c'\{k_2(n)\}^{-\gamma_2/d},$$

thus assumption 3.2(i) holds. Assumption 3.2(ii) is already verified in the proof of Result (2) of Proposition 6.1. Assumption 3.11 follows from the fact that $\widehat{P}(h) = P(h) = \|(w h_2)\|^2_{L^2(\mathcal{R}^d, leb)}$ and

$$P(\Pi_n h_0) - P(h_0) = \|w(\Pi_n h_{02} - h_{02})\|^2_{L^2(\mathcal{R}^d, leb)} + 2\langle w h_{02}, w(\Pi_n h_{02} - h_{02})\rangle_{L^2(\mathcal{R}^d, leb)} = o(1).$$

Assumption 3.14 follows from

$$P(h) - P(h_0) = \|w(h - h_{02})\|^2_{L^2(\mathcal{R}^d, leb)} + 2\langle w h_{02}, w(h - h_{02})\rangle_{L^2(\mathcal{R}^d, leb)}$$

with $g(\varepsilon) = \varepsilon^2$ and $t_0 = 2 w h_{02}$. Assumption 3.16 follows by our choice of norm and space. Assumption 3.17' is implied by condition 6.5. Finally assumption 3.18(b) follows from the fact that $P(h) = \|w h_2\|^2_{L^2(\mathcal{R}^d, leb)}$ is convex and continuous. Finally, by conditions 6.1(ii)(iii)(iv), we have

$$E\{\|m(X, \Pi_n h_0)\|^2_E\} \leq const. \|h_0 - \Pi_n h_0\|^2_c = O\left(\max\left[\{k_1(n)\}^{-2\gamma_1/d}, \{k_2(n)\}^{-2\gamma_2/d}\right]\right).$$

The result now follows from Corollary 3.2. *Q.E.D.*

PROOF OF PROPOSITION 6.4: We obtain the results by verifying that all the assumptions of Corollary 5.1 are satisfied. As assumptions 3.1, 3.2, 3.9 and 3.10 are already verified in the proofs of Propositions 6.1, 6.2 and 6.3, assumption 5.1 is automatically satisfied. Condition 6.8 implies assumption 5.4 (hence 5.2). It remains to verify assumptions 4.1. For assumption 4.1(i), by condition 6.1(ii) we have

$$\frac{dm(X, h_0)}{dh}[h - h_0] = E\{f_{Y_3|Y_1,Y_2,X}(h_{01}(Y_1) + h_{02}(Y_2))[h_1(Y_1) - h_{01}(Y_1) + h_2(Y_2) - h_{02}(Y_2)]|X\},$$

$$\|h - h_0\|^2 = E\left(\frac{dm(X, h_0)}{dh}[h - h_0]\right)^2 \le const. \|h - h_0\|_s^2,$$

$$\text{where } \|h - h_0\|_s^2 = E\left\{(h_1(Y_1) - h_{01}(Y_1) + h_2(Y_2) - h_{02}(Y_2))^2\right\},$$

hence assumption 4.1(i) holds. For any $h \in \mathcal{H}_{os}$ we recall the linear integral operator $T_h[g_1 + g_2] \equiv E\{f_{Y_3|Y_1,Y_2,X}(h_1(Y_1) + h_2(Y_2))[g_1(Y_1) + g_2(Y_2)]|X\}$ that maps from $Dom(T_h) \to L^2([0,1]^{d_x}, f_X)$. By condition 6.6(i)(ii) and proposition 7.33 of Zeidler (1985), T_h is compact for any $h \in \mathcal{H}_{os}$. Moreover, by conditions 6.6, for all $h \in \mathcal{H}_{os}$, T_h shares the same domain, range, and $a_j(T_h) \asymp a_j(T_{h_0})$; hence $\mu_j(T_h) \asymp \mu_j(T_{h_0})$ for all j (the same speed of singular value decay), and $\|T_h[g]\|_{L^2(f_X)} \asymp \|T_{h_0}[g]\|_{L^2(f_X)}$ for all $g \in Dom(T_h)$ (see Edmunds and Triebel (1996)). By the mean value theorem, for all $h \in \mathcal{H}_{os}$, $E\left[(m(X, h) - m(X, h_0))^2\right] = \|T_{\overline{h}}[h_1 - h_{01} + h_2 - h_{02}]\|_{L^2(f_X)}^2$, where $\overline{h}$ is a convex combination of h and h_0 in $\mathcal{H}_{os}$. While $\|h - h_0\|^2 = \|T_{h_0}[h_1 - h_{01} + h_2 - h_{02}]\|_{L^2(f_X)}^2$ by definition. Thus for all $h \in \mathcal{H}_{os}$, $c^2 \|h - h_0\|^2 \le E\left[(m(X, h) - m(X, h_0))^2\right] \le C^2 \|h - h_0\|^2$, and assumption 4.1(ii) holds. The conclusions now follow directly from Corollary 5.1. Q.E.D.

PROOF OF PROPOSITION 6.5:   We obtain the results by verifying that all the assumptions of Corollary 5.3 are satisfied. First, if the conditions of Proposition 6.2 holds, then we are in the "lower semi-compact case", and most of the verifications follow directly from those for Proposition 6.4. Given our choices of space, sieve and condition 6.8(ii), we only need to verify assumption 5.3, which directly follows from our choice of $\mathcal{H}_{os}^2$.

If the conditions of Proposition 6.3 holds, then it suffices to verify that assumption 4.2 is satisfied with the penalty $P(h_2) = \|(wh_2)\|_{L^2(\mathcal{R}^d, leb)}^2$. We have for all $h \in \mathcal{H}_{osn}$,

$$P(h_2) - P(\Pi_n h_{02}) - \langle 2w\Pi_n h_{02}, w(h_2 - \Pi_n h_{02})\rangle_{L^2(\mathcal{R}^d, leb)} = \|w(h_2 - \Pi_n h_{02})\|_{L^2(\mathcal{R}^d, leb)}^2 \ge 0.$$

Let $t_0 = 2w\Pi_n h_{02}$ then

$$
\begin{aligned}
|\langle t_0, h - \Pi_n h_0\rangle_{\mathbf{H}^*, \mathbf{H}}| &= 2\left|\langle w\Pi_n h_{02}, w(h_2 - \Pi_n h_{02})\rangle_{L^2(\mathcal{R}^d, leb)}\right| \\
&\le 2\|(w\Pi_n h_{02})\|_{L^2(\mathcal{R}^d, leb)} \times \|w(h_2 - \Pi_n h_{02})\|_{L^2(\mathcal{R}^d, leb)}
\end{aligned}
$$

thus assumption 4.2 is satisfied. Q.E.D.

PROOF OF PROPOSITION 6.6:   Result (1)(i) directly follows from Theorem 3.3 (lower semicompact penalty). Result (1)(ii) follows from Result (1)(i) and the Sobolev interpolation inequalities:

$$\|\nabla^k(w[\widehat{h} - h_0])\|_{L^2(\mathcal{R}, leb)} \le C \times (\|w[\widehat{h} - h_0]\|_{L^2(\mathcal{R}, leb)})^{1-\varsigma} \times (\|\nabla^{\overline{k}}(w[\widehat{h} - h_0])\|_{L^2(\mathcal{R}, leb)})^{\varsigma},$$

$$\|\nabla^k[\widehat{h} - h_0]\|_{L^2(f_{Y_2})} \le C \times (\|\widehat{h} - h_0\|_{L^2(f_{Y_2})})^{1-\varsigma} \times (\|\nabla^{\overline{k}}[\widehat{h} - h_0]\|_{L^2(f_{Y_2})})^{\varsigma}$$

for some $\varsigma \in (0, 1)$ depends on $\overline{k} > k$. Result (2) can be easily obtained by applying Theorem 1 of Chen, Linton and van Keilegom (2003). Using their notation, we define $M(\beta, h) \equiv E[\beta - a(Y_2)\nabla^k h(Y_2)]$ and $M_n(\beta, h) \equiv n^{-1} \sum_{i=1}^n (\beta - a(Y_{2i})\nabla^k h(Y_{2i}))$. Then their conditions (1.1) - (1.4) are trivially satisfied with the pseudo-metric $\|h\|_{\mathcal{H}} = \|\nabla^k h\|_{L^2(f_{Y_2})}$ (since $\|\nabla^k[\widehat{h}_n - h_0]\|_{L^2(f_{Y_2})} = o_P(1)$ by result (1)(ii)). Their condition (1.5) is satisfied provided that the class $\{\beta - a(Y_2)\nabla^k h(Y_2) :$

$\beta \in B, h \in \mathcal{H}_{osn}\}$ satisfies Glinvenko-Cantelli. Since the data is i.i.d., the class is Glinvenko-Cantelli provided that its $L^1(f_{Y_2})$−covering number with bracketing is finite, which is true as for any $(\beta, h), (\beta', h') \in B \times \mathcal{H}_{osn}$,

$$\left|(\beta - \beta') - a(Y_{2i})\nabla^k[h(Y_{2i}) - h'(Y_{2i})]\right| \leq |\beta - \beta'| + \sup_{y_2}|a(y_2)| \times \left|\nabla^k[h(Y_{2i}) - h'(Y_{2i})]\right|$$

and $\mathcal{H}_{osn} \subset \{h \in \mathcal{H} : ||h - h_0||_{L^2(f_{Y_2})} = o(1), ||\nabla^k[h - h_0]||_{L^2(f_{Y_2})} = o(1), P(h) \leq const.\}$, condition 6.10 and result (1) imply that $B \times \mathcal{H}_{osn}$ has a finite cover. *Q.E.D.*

PROOF OF PROPOSITION 6.7: We apply Theorem 4.1 of Chen (2007), which is a slight refinement of Theorem 2 of Chen, Linton and van Keilegom (2003). Given Proposition 6.6, it suffices to restrict our attention to $\mathcal{A}_{os} = \{\alpha = (\beta, h) \in B \times \mathcal{H}_{os} : |\beta - \beta_0| = o(1)\}$ and $\mathcal{A}_{osn} = \{\alpha = (\beta, h) \in B \times \mathcal{H}_{osn} : |\beta - \beta_0| = o(1)\}$. Following their notations, $M(\beta, h) \equiv E[\beta - a(Y_2)\nabla^k h(Y_2)] = E[\beta - (-1)^k l^{(k)}(Y_2)h(Y_2)]$ (the second equation is due to condition 6.11(i)), and $M_n(\beta, h) \equiv n^{-1}\sum_{i=1}^n(\beta - a(Y_{2i})\nabla^k h(Y_{2i}))$, conditions (4.1.1) - (4.1.4) of Chen (2007) are trivially satisfied with $\Gamma_1 = \Gamma_1(\beta, h_0) = 1$, $W = 1$ and

$$\begin{aligned}\Gamma_2(\beta_0, h_0)[h - h_0] &= \Gamma_2(\beta, h_0)[h - h_0] = -E\{a(Y_2)\nabla^k[h(Y_2) - h_0(Y_2)]\} \\ &= (-1)^{k+1}E\{l^{(k)}(Y_2)[h(Y_2) - h_0(Y_2)]\}.\end{aligned}$$

Chen's condition (4.1.5) is satisfied given i.i.d. data and the class $\{(\beta - a(Y_{2i})\nabla^k h(Y_{2i})) : (\beta, h) \in \mathcal{A}_{osn}\}$ is a Donsker class. For any $(\beta, h), (\beta', h') \in \mathcal{A}_{osn}$, under condition 6.10, we have

$$\left[E\left(\sup_{|\beta - \beta'|\leq\delta, ||\nabla^k[h-h']||_{L^2(f_{Y_2})}\leq\delta}\left|(\beta - \beta') - a(Y_{2i})\nabla^k[h(Y_{2i}) - h'(Y_{2i})]\right|^2\right)\right]^{1/2}$$

$$\leq 2\delta + 2\sup_{y_2}|a(y_2)|\left[E\left(\sup_{||\nabla^k[h-h']||_{L^2(f_{Y_2})}\leq\delta}\left|\nabla^k[h(Y_{2i}) - h'(Y_{2i})]\right|^2\right)\right]^{1/2} \leq const.\delta.$$

Let $N(\varepsilon, \mathcal{H}_{osn}, ||\nabla^k[.]||_{L^2(f_{Y_2})})$ denote the covering number of the class $\mathcal{H}_{osn} \subset \{h \in \mathcal{H} : ||h - h_0||_{L^2(f_{Y_2})} = o(1), ||\nabla^k[h - h_0]||_{L^2(f_{Y_2})} = o(1), ||\nabla^{\overline{k}}(\varpi h)||_{L^2(\mathcal{R}, leb)} \leq const.\}$. Then

$$\begin{aligned}&N(\varepsilon, \mathcal{H}_{osn}, ||\nabla^k[.]||_{L^2(f_{Y_2})}) \\ &\leq N(\varepsilon, \{h \in \mathcal{H} : ||\nabla^k(wh)||_{L^2(\mathcal{R}, leb)} + ||\nabla^{\overline{k}}(\varpi h)||_{L^2(\mathcal{R}, leb)} \leq const.\}, ||\nabla^k[.]||_{L^2(f_{Y_2})}) \\ &\leq const.\left(\frac{1}{\varepsilon}\right)^{1/(\overline{k}-k)}.\end{aligned}$$

Thus $\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{H}_{osn}, ||\nabla^k[.]||_{L^2(f_{Y_2})})}d\varepsilon < \infty$ provided $\overline{k} - k > 0.5$; hence the class $\{(\beta - a(Y_{2i})\nabla^k h(Y_{2i})) : (\beta, h) \in \mathcal{A}_{osn}\}$ is a Donsker by Theorem 3 of Chen, Linton and van Keilegom (2003).

To verify Chen's condition (4.1.6), we need to establish that $\sqrt{n}\Gamma_2(\beta_0, h_0)[\hat{h} - h_0] = O_P(1)$ and it has an asymptotic linear expansion:

$$\sqrt{n}\Gamma_2(\beta_0, h_0)[\hat{h} - h_0] = (-1)^k\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{dm(X_i, h_0)}{dh}[v^*]\right)'\rho(Z_i, h_0) + o_P(1), \quad\quad (D.1)$$

and thus

$$\sqrt{n}\{M_n(\beta_0, h_0) + \Gamma_2(\beta_0, h_0)[\widehat{h} - h_0]\}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ (\beta_0 - a(Y_{2i})\nabla^k h_0(Y_{2i})) + (-1)^k \left(\frac{dm(X_i, h_0)}{dh}[v^*] \right) \rho(Z_i, h_0) \right\} + o_P(1)$$

and by theorem 4.1 of Chen or theorem 2 Chen, Linton and van Keilegom (2003), we obtain: $\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \Rightarrow N(0, V^{-1})$ with

$$V^{-1} = Var\left\{ (\beta_0 - a(Y_{2i})\nabla^k h_0(Y_{2i})) + (-1)^k (T_{h_0}[v^*]) \rho(Z_i, h_0) \right\}.$$

To finish the proof, it remains to establish (D.1). Denote $\theta_0 \equiv E\{l^{(k)}(Y_2)h_0(Y_2)\}$ and $\widehat{\theta} = E\{l^{(k)}(Y_2)\widehat{h}(Y_2)\}$. It suffices to show that $\sqrt{n}(\widehat{\theta} - \theta_0) = O_P(1)$ and is asymptotically linearly distributed, where $\widehat{h}(Y_2)$ is the penalized SMD estimator of h_0:

$$\widehat{h} = \arg\min_{h \in \mathcal{H}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} ||\widehat{m}(X_i, h)||_E^2 + \lambda_n ||\nabla^{\overline{k}}(\varpi h)||_{L^2(leb)}^2 \right\}.$$

Following AC (2003) or Chen and Pouzo (2008), we first compute the Riesz representor v^* for $\theta - \theta_0$:

$$||v^*||^2 \equiv \sup_{h:0<||h-h_0||<\infty} \frac{(\theta - \theta_0)^2}{||h - h_0||^2} = \sup_{h:0<||h-h_0||<\infty} \frac{(E\{l^{(k)}(Y_2)[h(Y_2) - h_0(Y_2)]\})^2}{E\left[(T_{h_0}[h(Y_2) - h_0(Y_2)])^2 \right]}$$

$$= \sup_{h:0<||h-h_0||<\infty} \frac{\left(\langle l^{(k)}, h - h_0 \rangle_{L^2(f_{Y_2})} \right)^2}{\langle T_{h_0}^* T_{h_0}[h - h_0], h - h_0 \rangle_{L^2(f_{Y_2})}}$$

$$= \left\langle l^{(k)}, (T_{h_0}^* T_{h_0})^{-1} l^{(k)} \right\rangle_{L^2(f_{Y_2})} = \left\| (T_{h_0}^* T_{h_0})^{-\frac{1}{2}} l^{(k)} \right\|_{L^2(f_{Y_2})}^2 < \infty$$

and

$$E\{l^{(k)}(Y_2)[h(Y_2) - h_0(Y_2)]\} = \theta - \theta_0 = \langle v^*, h - h_0 \rangle = E\{(T_{h_0}[v^*]) (T_{h_0}[h - h_0])\}$$

by condition 6.12(i). Moreover, condition 6.12(ii) implies that we can solve v^* in a closed form: $v^* = (T_{h_0}^* T_{h_0})^{-1} l^{(k)} \in L^2(f_{Y_2})$.

Denote $||| \cdot |||^2 \equiv n^{-1} \sum_{i=1}^{n} || \cdot ||_E^2$ and $\widetilde{m}$ as the LS projection of $m(X, h)$ onto the linear sieve basis $p^{J_n}(X)$. By lemma B.3 (for which both assumptions 3.9 and 3.10(i) hold), i.i.d. data and condition 6.15 we obtain:

$$\sup_{h \in \mathcal{N}_{on}} |||\widehat{m}(\cdot, h) - \widetilde{m}(\cdot, h) - \widehat{m}(\cdot, h_0)|||^2$$

$$= \sup_{h \in \mathcal{N}_{on}} n^{-1} \sum_{i=1}^{n} (\widehat{m}(X_i, h) - \widetilde{m}(X_i, h) - \widehat{m}(X_i, h_0))^2$$

$$\leq \sup_{h \in \mathcal{N}_{on}} E \left[p^{J_n}(X_i)(P'P)^{-1}P'(\Delta\epsilon(h))(\Delta\epsilon(h))'P(P'P)^{-1}p^{J_n}(X_i)' \right]$$

$$\leq \sup_{h \in \mathcal{N}_{on}} E \left[p^{J_n}(X_i)(P'P)^{-1}P'E\left[(\Delta\epsilon(h))(\Delta\epsilon(h))'|X_1, \ldots, X_n\right]P(P'P)^{-1}p^{J_n}(X_i)' \right]$$

$$\leq \sup_{h \in \mathcal{N}_{on}} E \left[\Lambda_n \times Tr\left\{ n^{-1}p^{J_n}(X_i)'p^{J_n}(X_i)(P'P/n)^{-1} \right\} \right]$$

$$\leq K \sup_{h \in \mathcal{N}_{on}} E \left[E\left[(\rho(Z_i, h) - \rho(Z_i, h_0))^2 |X\right] \right] \times \frac{J_n}{n}$$

$$\leq K \sup_{h \in \mathcal{N}_{on}} \frac{J_n}{n} ||h - h_0||_s^{2\kappa} \leq O_P\left(\frac{J_n}{n} (\delta_{s,n}^*)^{2\kappa} \right) = o_P(n^{-1}),$$

where $\Lambda_n \equiv E[(\rho(Z,h) - \rho(Z,h_0))^2 | X]$, $\epsilon(Z,h) \equiv \rho(Z,h) - m(X,h)$ (i.e., a populational projection error), $\Delta\epsilon(h) \equiv \epsilon(Z,h) - \epsilon(Z,h_0)$ and Tr is the trace operator. With this, we can now follow the arguments in Chen and Pouzo (2008) and obtain

$$\sup_{h \in \mathcal{N}_{on}} |||\widehat{m}(\cdot,h)|||^2 = C \sup_{h \in \mathcal{N}_{on}} |||\widetilde{m}(\cdot,h) + \widehat{m}(\cdot,h_0)|||^2 + o_P(n^{-1})$$

for a constant $C > 0$. By definition of $\widehat{h}_n$, we have: $|||\widehat{m}(\cdot,\widehat{h}_n)|||^2 + \lambda_n P(\widehat{h}_n) \leq |||\widehat{m}(\cdot,h)|||^2 + \lambda_n P(h)$ for all $h \in \mathcal{N}_{0n}$, we have: for all $h \in \mathcal{N}_{0n}$,

$$C|||\widetilde{m}(\cdot,\widehat{h}_n) + \widehat{m}(\cdot,h_0)|||^2 + \lambda_n P(\widehat{h}_n) \leq C|||\widetilde{m}(\cdot,h) + \widehat{m}(\cdot,h_0)|||^2 + \lambda_n P(h) + o_P(n^{-1}).$$

Denote $\ell(\cdot,h) \equiv \widetilde{m}(\cdot,h) + \widehat{m}(\cdot,h_0)$. Then, by condition 6.17(i), $|||\ell(\cdot,h)|||^2 + C^{-1}\lambda_n P(h)$ is a smooth criterion function with $\widehat{h}_n$ as its approximate minimizer. Let $u_n^* = \pm v_n^*$, then, with $0 < \varepsilon_n = o(1/\sqrt{n})$, we have:

$$
\begin{aligned}
|||\ell(\cdot,\widehat{h}_n)|||^2 &\leq |||\ell(\cdot,\widehat{h}_n + \varepsilon_n u_n^*)|||^2 + C^{-1}\lambda_n\{P(\widehat{h}_n + \varepsilon_n u_n^*) - P(\widehat{h}_n)\} + o_P(n^{-1}). \\
&= |||\widetilde{m}(\cdot,\widehat{h}_n + \varepsilon_n u_n^*) + \widehat{m}(\cdot,h_0)|||^2 + o_P(n^{-1}),
\end{aligned}
$$

where the $o_P(n^{-1})$ in the above equation is due to condition 6.14(iii). After the second order Taylor expansion to the term $|||\ell(\cdot,\widehat{h}_n)|||^2 - |||\ell(\cdot,\widehat{h}_n + \varepsilon_n u_n^*)|||^2$, we have:

$$0 \leq \frac{\varepsilon_n}{n} \sum_{i=1}^{n} \left(\frac{d\widetilde{m}(X_i,\widehat{h}_n)}{dh}[u_n^*] \right)' \left(\widetilde{m}(X_i,\widehat{h}_n) + \widehat{m}(X_i,h_0) \right) + I_n(h(s)) + II_n(h(s)) + o_P(n^{-1}),$$

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

$$I_n(h(s)) \equiv 2\frac{\varepsilon_n^2}{n} \sum_{i=1}^{n} \left(\frac{d^2\widetilde{m}(X_i,h(s))}{dhdh}[u_n^*,u_n^*] \right)' \left(\widetilde{m}(X_i,h(s)) + \widehat{m}(X_i,h_0) \right),$$

$$II_n(h(s)) \equiv 2\frac{\varepsilon_n^2}{n} \sum_{i=1}^{n} \left(\frac{d\widetilde{m}(X_i,h(s))}{dh}[u_n^*] \right)' \left(\frac{d\widetilde{m}(X_i,h(s))}{dh}[u_n^*] \right).$$

Applying Cauchy-Schwarz and condition 6.16(iii), we have:

$$\sup_{h \in \mathcal{N}_{0n}} |I_n(h)| \leq const.\varepsilon_n^2 \sqrt{\sup_{h \in \mathcal{N}_{0n}} \frac{1}{n} \sum_{i=1}^{n} \|\widetilde{m}(X_i,h) + \widehat{m}(X_i,h_0)\|_E^2} = \varepsilon_n^2 \times O_P(\delta_n^* + \sqrt{\frac{J_n}{n}}),$$

where the second equality is due to assumption 4.1, conditions 6.14(i) and 6.15 (ii), and the definition of $\widetilde{m}(X_i,h)$. (Lemma A.1(C) of AC (2003) and condition 6.15 (ii) imply $\frac{1}{n}\sum_{i=1}^{n} \|\widehat{m}(X_i,h_0)\|_E^2 = O_P(\frac{J_n}{n})$. Lemma B.2 and the definition of $\widetilde{m}(X,h)$ imply

$$
\begin{aligned}
\sup_{h \in \mathcal{N}_{0n}} n^{-1} \sum_{i=1}^{n} \|\widetilde{m}(X_i,h)\|_E^2 &\asymp \sup_{h \in \mathcal{N}_{0n}} E\left[\|\widetilde{m}(X,h)\|_E^2\right] \leq \sup_{h \in \mathcal{N}_{0n}} E\left[\|m(X,h)\|_E^2\right] \\
&\asymp \sup_{h \in \mathcal{N}_{0n}} \|h - h_0\|^2 = O(\delta_n^{*2})
\end{aligned}
$$

by assumption 3.8(i) and condition 6.14(i).)

Next, by condition 6.17(i)(ii) and i.i.d. data, we have:

$$\sup_{h\in\mathcal{N}_{0n}}|II_n(h)| \leq const.\varepsilon_n^2\left\{\sup_{h\in\mathcal{N}_{0n}}\frac{1}{n}\sum_{i=1}^n\left\|\frac{d\widetilde{m}(X_i,h)}{dh}[u_n^*]-\frac{d\widetilde{m}(X_i,h_0)}{dh}[u_n^*]\right\|_E^2+\frac{1}{n}\sum_{i=1}^n\left\|\frac{d\widetilde{m}(X_i,h_0)}{dh}[u_n^*]\right\|_E^2\right\}$$

$$= \varepsilon_n^2\times o_P(n^{-1/2})+O_P(\varepsilon_n^2).$$

Therefore, we have

$$0\leq\frac{\varepsilon_n}{n}\sum_{i=1}^n\left(\frac{d\widetilde{m}(X_i,\widehat{h}_n)}{dh}[u_n^*]\right)'\left(\widetilde{m}(X_i,\widehat{h}_n)+\widehat{m}(X_i,h_0)\right)+O_p(\varepsilon_n^2)+o_P(n^{-1})$$

holds for $u_n^*=\pm v_n^*$; and hence

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{d\widetilde{m}(X_i,\widehat{h}_n)}{dh}[v_n^*]\right)'\left(\widehat{m}(X_i,h_0)+\widetilde{m}(X_i,\widehat{h}_n)\right)=o_P(1).$$

Since both $\widehat{m}$ and $\widetilde{m}$ are the LS projections onto the linear sieve basis $p^{J_n}(X)$, we have:

$$\frac{1}{n}\sum_{i=1}^n\left(\frac{dm(X_i,\widehat{h}_n)}{dh}[v_n^*]\right)'\left(\widehat{m}(X_i,h_0)+\widetilde{m}(X_i,\widehat{h}_n)\right)$$

$$= \frac{1}{n}\sum_{i=1}^n\left(\frac{d\widetilde{m}(X_i,\widehat{h}_n)}{dh}[v_n^*]\right)'\left(\widehat{m}(X_i,h_0)+\widetilde{m}(X_i,\widehat{h}_n)\right)=o_P(\frac{1}{\sqrt{n}}).$$

Note that,

$$\left|\frac{1}{n}\sum_{i=1}^n\left(\frac{dm(X_i,\widehat{h}_n)}{dh}[v_n^*]-\frac{dm(X_i,h_0)}{dh}[v_n^*]\right)'\left(\widehat{m}(X_i,h_0)+\widetilde{m}(X_i,\widehat{h}_n)\right)\right|$$

$$\leq const.\sqrt{\frac{1}{n}\sum_{i=1}^n\left\|\frac{dm(X_i,\widehat{h}_n)}{dh}[v_n^*]-\frac{dm(X_i,h_0)}{dh}[v_n^*]\right\|_E^2}\times\sqrt{\frac{1}{n}\sum_{i=1}^n\left\|\widehat{m}(X_i,h_0)+\widetilde{m}(X_i,\widehat{h}_n)\right\|_E^2}$$

$$= o_P(n^{-1/4})\times O_P(\sqrt{\frac{J_n}{n}}+\delta_n^*)=o_P(n^{-1/2}),$$

where the first term is of order $o_P(n^{-1/4})$ by condition 6.17(ii) and Markov inequality. Thus, we obtain:

$$\frac{1}{n}\sum_{i=1}^n\left(\frac{dm(X_i,h_0)}{dh}[v_n^*]\right)'\left(\widehat{m}(X_i,h_0)+\widetilde{m}(X_i,\widehat{h}_n)\right)=o_P(\frac{1}{\sqrt{n}}).$$

Notice that

$$\left|\frac{1}{n}\sum_{i=1}^n\left(\frac{dm(X_i,h_0)}{dh}[v_n^*-v^*]\right)'\left(\widehat{m}(X_i,h_0)+\widetilde{m}(X_i,\widehat{h}_n)\right)\right|$$

$$\leq \sqrt{n^{-1}\sum_{i=1}^n\left\|\frac{dm(X_i,h_0)}{dh}[v_n^*-v^*]\right\|_E^2}\times\sqrt{n^{-1}\sum_{i=1}^n\left\|\widehat{m}(X_i,h_0)+\widetilde{m}(X_i,\widehat{h}_n)\right\|_E^2}$$

$$\leq O_P\left(\|v_n^*-v^*\|\right)\times O_P(\sqrt{\frac{J_n}{n}}+\delta_n^*)=o_P(n^{-1/2}),$$

where the second inequality follows from the Markov inequality, i.i.d. data and the definition of $\|v_n^*-v^*\|$, and the least equality is due to conditions 6.13 and 6.14(i). Therefore,

$$\frac{1}{n}\sum_{i=1}^n\left(\frac{dm(X_i,h_0)}{dh}[v^*]\right)'\left(\widehat{m}(X_i,h_0)+\widetilde{m}(X_i,\widehat{h}_n)\right)=o_P(\frac{1}{\sqrt{n}}).$$

Denote $g^*(X) \equiv \frac{dm(X,h_0)}{dh}[v^*]$, then $g^*() \in L^2(f_X)$. Let $\tilde{g}^*$ denote the LS projection of g^* onto the linear sieve basis $p^{J_n}(X)$. Then by exchanging summation to the left hand side of the above equation, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{g}^*(X_i)'\left(\rho(Z_i,h_0)+m(X_i,\hat{h}_n)\right)=\frac{1}{n}\sum_{i=1}^{n}g^*(X_i)'\left(\hat{m}(X_i,h_0)+\tilde{m}(X_i,\hat{h}_n)\right)=o_P(\frac{1}{\sqrt{n}}).$$

By Markov inequality, i.i.d. data, $E[\rho(Z_i,h_0)|X_1,...,X_n]=0$, and condition 6.15 (ii), we have for all $\varepsilon > 0$,

$$\Pr\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\tilde{g}^*(X_i)-g^*(X_i))'\rho(Z_i,h_0)\right|\geq\varepsilon\right)$$
$$\leq\quad \varepsilon^{-2}E\left(|(\tilde{g}^*(X_i)-g^*(X_i))'\rho(Z_i,h_0)|^2\right)\leq const.\varepsilon^{-2}E\left(|\tilde{g}^*(X_i)-g^*(X_i)|^2\right)=o(1).$$

Also, by Markov inequality, $m(X_i,h_0)=0$, assumption 4.1, conditions 6.14(i) and 6.16(iii), we have for all $\varepsilon > 0$,

$$\Pr\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\tilde{g}^*(X_i)-g^*(X_i))'m(X_i,\hat{h}_n)\right|\geq\varepsilon\right)$$
$$\leq\quad \varepsilon^{-2}\sqrt{n}E\left(|(\tilde{g}^*(X_i)-g^*(X_i))'\{m(X_i,\hat{h}_n)-m(X_i,h_0)\}|\right)$$
$$\leq\quad \varepsilon^{-1}\sqrt{n}\times\sqrt{E\left(|\tilde{g}^*(X_i)-g^*(X_i)|^2\right)}\times O_P(\delta_n^*)=o_P(1).$$

Thus

$$\frac{1}{n}\sum_{i=1}^{n}g^*(X_i)'\rho(Z_i,h_0)+\frac{1}{n}\sum_{i=1}^{n}g^*(X_i)'m(X_i,\hat{h}_n)=o_P(\frac{1}{\sqrt{n}}).$$

By condition 6.16(i)(ii), $\{g^*(X_i)'m(X_i,h):h\in\mathcal{N}_{on}\}$ is a Donsker class, and since $E[(g^*(X_i)'[m(X_i,h)-m(X_i,h_0)])^2]=o(1)$ for all $h\in\mathcal{N}_{on}$, applying Lemma 1 of Chen et al (2003), we obtain that $\{g^*(X_i)'m(X_i,h):h\in\mathcal{N}_{on}\}$ satisfies stochastic equicontinuity; hence, uniformly in $h\in\mathcal{N}_{on}$,

$$\frac{1}{n}\sum_{i=1}^{n}(g^*(X_i))'m(X_i,h)$$
$$=\quad E\left[\left(\frac{dm(X_i,h_0)}{dh}[v^*]\right)'(m(X,h)-m(X,h_0))\right]+o_P(n^{-1/2})$$
$$=\quad E\left[\left(\frac{dm(X_i,h_0)}{dh}[v^*]\right)'\left(\frac{dm(X_i,\overline{h})}{dh}[h-h_0]\right)\right]+o_P(n^{-1/2})$$
$$=\quad \langle v^*,h-h_0\rangle+o_P(n^{-1/2}),$$

for some $\overline{h}\in\mathcal{N}_o$ in the second equality (applying the mean value theorem to $m(X,h)-m(X,h_0)$), and the last equality is due to $\langle v^*,h-h_0\rangle=E\left[\left(\frac{dm(X_i,h_0)}{dh}[v^*]\right)'\left(\frac{dm(X_i,h_0)}{dh}[h-h_0]\right)\right]$ and condition 6.17(iii). Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}g^*(X_i)'\rho(Z_i,h_0)+\langle v^*,\hat{h}_n-h_0\rangle=o_P(\frac{1}{\sqrt{n}}),$$

that is

$$\sqrt{n}(\hat{\theta}-\theta_0)=\sqrt{n}\langle v^*,\hat{h}_n-h_0\rangle=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{dm(X_i,h_0)}{dh}[v^*]\right)'\rho(Z_i,h_0)+o_P(1).$$

56

$$\sqrt{n}\Gamma_2(\beta_0, h_0)[\widehat{h} - h_0] = \sqrt{n}(-1)^{k+1}(\widehat{\theta} - \theta_0)$$
$$= (-1)^k \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{dm(X_i, h_0)}{dh}[v^*] \right)' \rho(Z_i, h_0) + o_P(1).$$

Hence we obtain result (D.1). *Q.E.D.*

REFERENCES

Ai, C. and X. Chen (2003). Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica* **71** 1795-1844.

Ai, C. and X. Chen (2005). Efficient Estimation of Sequential Moment Restrictions Containing Unknown Functions. Mimeo, University of Florida and New York University.

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

Andrews, D. (1995). Nonparametric Kernel Estimation for Semiparametric Models. *Econometric Theory* **11** 560-596.

Blundell, R. X. Chen and D. Kristensen (2007). Semi-nonparametric IV Estimation of Shape-Invariant Engel Curves. *Econometrica* **75** 1613-1670.

Bissantz, N., T. Hohage, A. Munk and F. Ruymgaart (2007). Convergence Rates of General Regularization Methods for Statistical Inverse Problems and Applications. *SIAM J. Numer. Anal.* **45**, 2610-2636.

Carrasco, M., J.-P. Florens and E. Renault (2007). Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization. *The Handbook of Econometrics*, J.J. Heckman and E.E. Leamer (eds.), **6B**. North-Holland, Amsterdam.

Cavalier, L., G. Golubev, D. Picard and A. Tsybakov (2002). Oracle Inequalities in Inverse Problems. *Annals of Statistics* **30** 843-874.

Chen, X. (2007). Large Sample Sieve Estimation of Semi-nonparametric Models. *The Handbook of Econometrics*, J.J. Heckman and E.E. Leamer (eds.), **6B**. North-Holland, Amsterdam.

Chen, X. and S. Ludvigson (2004). Land of Addicts? An Empirical Investigation of Habit-based Asset Pricing Models. NBER Working Paper No. 10503.

Chen, X. and D. Pouzo (2007). Efficient Estimation of Nonparametric Quantile IV Weighted Average Derivative. Mimeo, Yale University and New York University.

Chen, X. and D. Pouzo (2008). Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals. Yale University, Cowles Foundation Discussion Paper No. 1640R.

Chen, X. and M. Reiss (2007). On Rate Optimality for Nonparametric Ill-posed Inverse Problems in Econometrics. Yale University, Cowles Foundation Discussion Paper No. 1626.

CHERNOZHUKOV, V. AND C. HANSEN (2005). An IV Model of Quantile Treatment Effects. *Econometrica* **73**, 245-61.

CHERNOZHUKOV, V., P. GAGLIARDINI, AND O. SCAILLET (2008). Nonparametric Instrumental Variable Estimation of Quantile Structural Effects. Mimeo, MIT, University of Lugano and Swiss Finance Institute.

CHERNOZHUKOV, V., G. IMBENS, AND W. NEWEY (2007). Instrumental Variable Estimation of Nonseparable Models. *Journal of Econometrics* **139**, 4-14.

CHESHER, A. (2003). Identification in Nonseparable Models. *Econometrica* **71**, 1405-1441.

DAROLLES, S., J.-P. FLORENS AND E. RENAULT (2006). Nonparametric Instrumental Regression. mimeo, Toulouse School of Economics.

DAUBECHIES, I., M. DEFRISE AND C. DE MOL (2004). An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint. *Comm. on Pure and Applied Math.* **LVII**, 1413-1457.

EDMUNDS, D. AND H. TRIEBEL (1996). *Function Spaces, Entropy Numbers, Differential Operators*, Cambridge University Press: Cambridge.

EFROMOVICH, S. AND V. KOLTCHINSKII (2001). On Inverse Problems with Unknown Operators. *IEEE Trans. on Information Theory* **47**, 2876-2893.

EGGERMONT, P.P.B. AND V.N. LARICCIA (2001). *Maximum Penalized Likelihood Estimation*, Springer Series in Statistics.

ENGL, H., M. HANKE AND A. NEUBAUER (1996). *Regularization of Inverse Problems*, Kluwer Academic Publishers: London.

FLORENS, JP, J. JOHANNES AND S. VAN BELLEGEM (2007). Identification and Estimation by Penalization in Nonparametric Instrumental Regression. Mimeo, Toulouse School of Economics.

GAGLIARDINI, P. AND O. SCAILLET (2007). Tikhonov Regularization for Nonparametric Instrumental Variable Estimators. Mimeo, University of Lugano and Swiss Finance Institute.

GRENANDER, U. (1981). *Abstract Inference*, Wiley Series: New York.

HALL, P. AND J. HOROWITZ (2005). Nonparametric Methods for Inference in the Presence of Instrumental Variables. *Annals of Statistics* **33**, 2904-2929.

HOFFMANN, M. AND M. REISS (2008). Nonlinear Estimation for Linear Inverse Problems with Error in the Operator. *Annals of Statistics* **36**, 310-336.

HOROWITZ, J. AND S. LEE (2007). Nonparametric Instrumental Variables Estimation of a Quantile Regression Model. *Econometrica* **75**, 1191-1208.

HOROWITZ, J. AND E. MAMMEN (2007). Rate-Optimal Estimation for a General Class of Nonparametric Regression Models with Unknown Link Functions. *Annals of Statistics*, forthcoming.

HUANG, J. (1998). Projection Estimation in Multiple Regression with Application to Functional ANOVA Models. *Annals of Statistics* **26**, 242-272.

Huang, J. (2003). Local Asymptotics for Polynomial Spline Regression. *Annals of Statistics* **31**, 1600-1635.

Matzkin, R. (2007). Nonparametric Identification. *The Handbook of Econometrics*, J.J. Heckman and E.E. Leamer (eds.), **6B**. North-Holland, Amsterdam.

Meyer, Y. (1992). *Wavelets and Operators*. Cambridge University Press.

Newey, W.K. (1991). Uniform Convergence in Probability and Stochastic Equicontinuity. *Econometrica* **59**, 1161-1167.

————— (1997). Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics* **79**, 147-168.

Newey, W.K. and J. Powell (2003). Instrumental Variables Estimation for Nonparametric Models. *Econometrica* **71**, 1565-1578.

Severini, T. and G. Tripathi (2006). Some Identification Issues in Nonparametric Linear Models with Endogenous Regressors. *Econometric Theory,* **22**, 258-278.

Shen, X. and W. Wong (1994). Convergence Rate of Sieve Estimates. *The Annals of Statistics,* **22**, 580-615.

Van de Geer, S. (2000). *Empirical Processes in M-estimation*, Cambridge University Press.

Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics,* **27**, 1564-1599.

Zeidler, E. (1985). *Nonlinear Functional Analysis and its Applications III: Variational methods and optimization*, Springer.