ESTIMATION OF NONPARAMETRIC CONDITIONAL MOMENT MODELS WITH POSSIBLY NONSMOOTH GENERALIZED RESIDUALS

By

Xiaohong Chen and Demian Pouzo

April 2008
Revision July 2009

COWLES FOUNDATION DISCUSSION PAPER NO. 1650R

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 Generalized Residuals[1]

Xiaohong Chen[2] and Demian Pouzo[3]

First version: August 2005, Revised version: July 2009

Abstract

This paper studies nonparametric estimation of conditional moment models in which the generalized residual functions can be nonsmooth in the unknown functions of endogenous variables. This is a nonparametric nonlinear instrumental variables (IV) problem. We propose a class of penalized sieve minimum distance (PSMD) estimators which are minimizers of a penalized empirical minimum distance criterion over a collection of sieve spaces that are dense in the infinite dimensional function parameter space. Some of the PSMD procedures use slowly growing finite dimensional sieves with flexible penalties or without any penalty; some use large dimensional sieves with lower semicompact and/or convex penalties. We establish their consistency and the convergence rates in Banach space norms (such as a sup-norm or a root mean squared norm), allowing for possibly non-compact infinite dimensional parameter spaces. For both mildly and severely ill-posed nonlinear inverse problems, our convergence rates in Hilbert space norms (such as a root mean squared norm) achieve the known minimax optimal rate for the nonparametric mean IV regression. We illustrate the theory with a nonparametric additive quantile IV regression. We present a simulation study and an empirical application of estimating nonparametric quantile IV Engel curves.

KEYWORDS: Nonlinear ill-posed inverse, penalized sieve minimum distance, modulus of continuity, convergence rate, nonparametric additive quantile IV, quantile IV Engel curves.

JEL Classification: C13, C14, D12.

1 Introduction

This paper is about estimation of the unknown functions $h_0(\cdot) \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))$ satisfying the following conditional moment restrictions:

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

[1]This is a substantially revised version of Cowles Foundation Discussion Paper 1650. We are grateful to the coeditor W. Newey, three anonymous referees, R. Blundell, V. Chernozhukov, J. Horowitz, S. Lee, Z. Liao and A. Torgovitsky for their constructive comments that lead to a much improved revision. Earlier versions were presented in August 2006 European Meeting of the Econometric Society, March 2007 Oberwolfach Workshop on Semi/nonparametrics, June 2007 Cemmap Conference on Measurement Matters, 2008 Cowles Summer Conference, and econometric workshops at many universities. We thank J. Florens, I. Komunjer, O. Linton, E. Mammen, J. Powell, A. Santos, E. Tamer, E. Vytlacil, and other participants of these conferences and workshops for helpful suggestions. Chen acknowledges financial support from the National Science Foundation under Award Number SES0631613 and SES0838161. Any errors are the responsibility of the authors.

[2]Cowles Foundation for Research in Economics, Yale University, 30 Hillhouse, Box 208281, New Haven, CT 06520, USA. E-mail: xiaohong.chen@yale.edu

[3]Department of Economics, UC at Berkeley, 508-1 Evans Hall 3880, Berkeley, CA 94704-3880. E-mail: dpouzo@econ.berkeley.edu

where $Z \equiv (Y', X_z')'$, Y is a vector of endogenous (or dependent) variables, X_z is a subset of the conditioning (or instrumental) variables X and the conditional distribution of Y given X is not specified. $\rho()$ is a vector of generalized residuals with functional forms known up to a finite dimensional parameter θ_0 and functions of interest $h_0(\cdot) \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))$, where each function $h_{0\ell}(\cdot), \ell = 1, ..., q$ may depend on different components of X and Y, and some could depend on θ_0 and $h_{0\ell'}(\cdot)$ for $\ell' \neq \ell$. In this paper $\rho()$ may depend on the unknown (θ_0, h_0) nonlinearly and pointwise nonsmoothly.

Model (1.1) extends the semi/nonparametric conditional moment framework previously considered in Chamberlain (1992), Newey and Powell (2003) (henceforth NP) and Ai and Chen (2003) (henceforth AC) to allow for the generalized residual function $\rho(Z; \theta, h)$ to be pointwise non-smooth with respect to the unknown parameters of interest (θ, h). As already illustrated by these papers, many semi/nonparametric structural models in economics are special cases of (1.1). For instance, it includes the model of a shape-invariant system of Engel curves with endogenous total expenditure of Blundell, Chen and Kristensen (2007) (henceforth BCK), which itself is an extension of the nonparametric mean instrumental variables regression (NPIV) model analyzed in NP, Darolles, Florens and Renault (2006) (henceforth DFR) and Hall and Horowitz (2005) (henceforth HH):

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

Model (1.1) also nests the quantile instrumental variables (IV) treatment effect model of Chernozhukov and Hansen (2005) (henceforth CH), and the nonparametric quantile instrumental variables regression (NPQIV) model of Chernozhukov, Imbens and Newey (2007) (henceforth CIN) and Horowitz and Lee (2007) (henceforth HL):

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

where $1\{\cdot\}$ denotes the indicator function. Additional examples include a partially linear quantile IV regression $E[1\{Y_1 \leq h_0(Y_2) + Y_3'\theta_0\}|X] = \gamma$, a single index quantile IV regression $E[1\{Y_1 \leq h_0(Y_2'\theta_0)\}|X] = \gamma$, an additive quantile IV regression $E[1\{Y_3 \leq h_{01}(Y_1) + h_{02}(Y_2)\}|X] = \gamma$ and many more.

Most asset pricing models also imply the conditional moment restriction (1.1), in which the generalized residual function $\rho(Z; \theta, h)$ takes the form of some asset returns multiplied by a pricing kernel (or stochastic discount factor). Different asset pricing models correspond to different functional form specifications of the pricing kernel up to some unknown parameters (θ, h). For instance, Chen and Ludvigson (2006) study a consumption-based asset pricing model with an unknown habit formation. Their model is an example of (1.1), in which the generalized residual function $\rho(Z; \theta, h)$ is highly nonlinear, but smooth, in the unknown habit function h. Many durable goods and investment based asset pricing models with flexible pricing kernels also belong to the framework (1.1); see,

e.g., Gallant and Tauchen (1989), Bansal and Viswanathan (1993). In some asset pricing models involving cash-in-advance constraints, or in which the underlying asset is a defaultable bond, the pricing kernels (hence the generalized residual functions) are not pointwise smooth in (θ, h). See, e.g., Arellano (2008) for an economic general equilibrium model of pricing default risk, and Chen and Pouzo (2009) for an econometric study.

As demonstrated in NP, AC and CIN, the key difficulty of analyzing the semi/nonparametric model (1.1) is not due to the presence of the unknown finite dimensional parameter θ_0, but due to the fact that some of the unknown functions $h_{0\ell}(\cdot), \ell = 1, ..., q$ depend on the endogenous variable Y.[4] Therefore, in this paper we shall focus on the nonparametric estimation of $h_0()$, which is identified by the following conditional moment restrictions:

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

where $h_0(\cdot) \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))$ depends on Y and may enter $\rho()$ nonlinearly and possibly nonsmoothly.[5] Suppose that $h_0(\cdot)$ belongs to a function space $\mathcal{H}$, which is an infinite dimensional subset of a Banach space with norm $||\cdot||_s$, such as the space of bounded continuous functions with the sup-norm $||h||_s = \sup_y |h(y)|$, or the space of square integrable functions with the root mean squared norm $||h||_s = \sqrt{E[h(Y)^2]}$. We are interested in consistently estimating $h_0(\cdot)$ and determining the rate of convergence of the estimator under $||\cdot||_s$.

In this paper, we first propose a broad class of penalized sieve minimum distance (PSMD) estimation procedures for the general model (1.4). All of the PSMD procedures minimize a possibly penalized consistent estimate of the minimum distance criterion, $E\left(E[\rho(Z; h(\cdot))|X]' E[\rho(Z; h(\cdot))|X]\right)$, over sieve spaces, $\mathcal{H}_n$,[6] that are dense in the infinite dimensional function space $\mathcal{H}$. Some of the PSMD procedures use *slowly growing finite dimensional sieves* (i.e., $\dim(\mathcal{H}_n) \to \infty$, $\dim(\mathcal{H}_n)/n \to 0$), with flexible penalties or without any penalty. Some use *large dimensional sieves* (i.e., $\dim(\mathcal{H}_n)/n \to const. > 0$), with lower semicompact[7] and/or convex penalties. Under relatively low-level sufficient conditions and without assuming compactness of the function parameter space $\mathcal{H}$, we establish consistency and the convergence rates under norm $||\cdot||_s$ for these PSMD estimators. Our convergence rates in the case when $\mathcal{H}$ is an infinite dimensional subset of a Hilbert space coincide with the known minimax optimal rate for the NPIV example (1.2).

The existing literature on estimation of nonparametric IV models consists of two separate approaches: the sieve minimum distance (SMD) method and the function space Tikhonov regularized

[4] In some applications the presence of the parametric part θ_0 in the semi/nonparametric model (1.1) aids the identification of the unknown function h_0; see, e.g., Chen and Ludvigson (2006).

[5] See Chen and Pouzo (2008b) for semiparametric efficient estimation of the parametric part θ_0 for the general semi/nonparametric model (1.1) with possibly nonsmooth residuals. Their results depend crucially on the consistency and convergence rates of the nonparametric estimation of h_0, which are established in our this paper.

[6] In this paper we let n denote the sample size, and $\dim(\mathcal{H}_n)$ the dimension of the sieve space.

[7] See Section 2 for its definition.

minimum distance (TR-MD) method. The SMD procedure minimizes a consistent estimate of the minimum distance criterion over some finite dimensional compact sieve space; see, e.g., NP, AC, CIN and BCK. The TR-MD procedure minimizes a consistent penalized estimate of the minimum distance criterion over the original infinite dimensional function space $\mathcal{H}$, in which the penalty function is of the classical Tikhonov type (e.g., $\int\{h(y)\}^2dy$ or $\int\{\nabla^r h(y)\}^2dy$ with $\nabla^r h$ being the r-th derivative of h); see, e.g., DFR, HH, HL, Carrasco, Florens and Renault (2007) (henceforth CFR), Chernozhukov, Gagliardini and Scaillet (2008) (henceforth CGS) and the references therein. When h_0 enters the residual function $\rho(Z;h_0)$ linearly such as in the NPIV model (1.2), both SMD and TR-MD estimators can be computed analytically. But, when h_0 enters the residual function $\rho(Z;h_0)$ nonlinearly, such as in the NPQIV model (1.3), the numerical implementations of TR-MD estimators typically involve some finite dimensional sieve approximations to functions in $\mathcal{H}$.[8] For example, in the simulation study of the NPQIV model (1.3), HL approximate the unknown function $h_0(\cdot)$ by a Fourier series with a large number of terms; hence they can ignore the Fourier series approximation error and view their implemented procedure as the solution to the infinite dimensional TR-MD problem. In another simulation study and empirical illustration of the NPQIV model, CGS use a small number of spline and polynomial series terms to approximate h_0 in order to compute their function space TR-MD estimator. Although one could numerically compute the SMD estimator using finite dimensional compact sieves, simulation studies indicate that it is easier to compute a penalized SMD estimator using finite dimensional linear sieves (see, e.g., BCK).[9] In summary, some versions of the PSMD family of procedures have already been implemented in the existing literature, but their asymptotic properties have not been established for the general model (1.4).

There is a rapidly growing literature on the consistent estimation of $h_0()$ for two popular special cases of the general model (1.4): the NPIV and the NPQIV. For the NPIV model (1.2), see NP for consistency of the SMD estimator in a (weighted) sup-norm; BCK for the convergence rate in a root mean squared norm of the SMD estimator; HH, DFR, and Gagliardini and Scaillet (2008) (henceforth GS) for the convergence rate in a root mean squared norm of their respective kernel based TR-MD estimators; HH and Chen and Reiss (2007) (henceforth CR) for the minimax optimal rate in a root mean squared norm.[10] For the NPQIV model (1.3), see CIN for consistency of the SMD estimator in a sup-norm; HL and CGS for the convergence rates in a root mean squared norm of their respective kernel based TR-MD estimators.[11]

[8] This is because numerical optimization algorithms cannot handle infinite dimensional objects in $\mathcal{H}$.

[9] This is because a constraint optimization problem is typically more difficult to compute than the corresponding unconstraint optimization problem.

[10] See NP, DFR, BCK, CFR, Severini and Tripathi (2006), D'Haultfoeuille (2008), Florens, Johannes and van Bellegem (2008) and the references therein for identification of the NPIV model.

[11] See CH and CIN for identification of the NPQIV model; also see Chesher (2003), Matzkin (2007) and the references therein for identification of nonseparable models.

To the best of our knowledge, there are currently only two published papers that consider consistent estimation of h_0 for the general model (1.4) when $h_0(\cdot)$ depends on Y. Under the assumption that the infinite dimensional function space $\mathcal{H}$ is compact (in $||\cdot||_s$), NP established the consistency of the SMD estimator of h when it enters the residual function $\rho(Z, h(\cdot))$ pointwise continuously, and CIN derived the consistency of the SMD estimator when h may enter $\rho(Z, h(\cdot))$ pointwise nonsmoothly. However, except for the NPIV (1.2) and the NPQIV (1.3) examples, there is no published work that establishes the convergence rate (in $||\cdot||_s$) of any estimator of h_0 for the general model (1.4). Even for the NPQIV model (1.3), there are no published results on the convergence rate of the SMD estimator of h_0.

The original SMD procedures of NP, AC and CIN can be viewed as PSMD procedures using slowly growing finite dimensional linear sieves ($\dim(\mathcal{H}_n) \to \infty$, $\dim(\mathcal{H}_n)/n \to 0$) with lower semicompact penalty functions; hence our theoretical results immediately imply the consistency and the rates of convergence (in $||\cdot||_s$) of the original SMD estimators for the general model (1.4), without assuming the $||\cdot||_s-$compactness of the whole function parameter space $\mathcal{H}$. More interestingly, our theoretical results also allow for the series minimum distance procedure with slowly growing finite dimensional linear sieves without any penalty. The PSMD procedure using large dimensional linear sieves ($\dim(\mathcal{H}_n)/n \to const. > 0$) and lower semicompact and/or convex penalties can be viewed as computable extensions of the current TR-MD procedures for the NPIV and the NPQIV models to all conditional moment models (1.4), and allow for much more flexible penalty functions.

In Section 2, we first explain the technical hurdle associated with nonparametric estimation of $h_0()$ for the general model (1.4), and then present the PSMD procedure. Section 3 provides relatively low level sufficient conditions for consistency when the parameter space is a Banach space with norm $||\cdot||_s$ and Section 4 derives the convergence rate. Section 5 derives the rate of convergence under relatively low level sufficient conditions for the case when the parameter space is a Hilbert space with norm $||\cdot||_s$ and shows that the rate for the general model (1.4) coincides with the optimal minimax rate for the NPIV model (1.2). Throughout these sections, we use the NPIV example (1.2) to illustrate key sufficient conditions and various theoretical results. Section 6 specializes the general theoretical results to a nonparametric additive quantile IV model: $E[1\{Y_3 \leq h_{01}(Y_1) + h_{02}(Y_2)\}|X] = \gamma \in (0,1)$ where $h_0 = (h_{01}, h_{02})$. In Section 7, we first present a simulation study of the NPQIV model (1.3) to assess the finite sample performance of the PSMD estimators. We then provide an empirical application of nonparametric quantile IV Engel curves using data from the British Family Expenditure Survey (FES). Based on our simulation and empirical studies, the PSMD estimators using slowly growing finite dimensional linear sieves with flexible penalties are not only easy to compute but also perform well in finite samples. Section 8 briefly concludes. Appendix A presents a brief review of some functional spaces and sieve bases, and the other appendices contain mathematical proofs.

Notation. 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 exist 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 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' to denote its transpose and $||x||_E$ to denote its Euclidean norm (i.e., $||x||_E \equiv \sqrt{x'x}$), although sometimes we will also use $|x| = ||x||_E$ without too much confusion.

2 PSMD Estimators

Suppose that observations $\{(Y_i', X_i')\}_{i=1}^n$ 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 (1.4), where $\rho : \mathcal{Z} \times \mathcal{H} \to \mathcal{R}^{d_\rho}$ is a measurable mapping known up to a vector of unknown functions, $h_0 \in \mathcal{H} \equiv \mathcal{H}^1 \times \cdots \times \mathcal{H}^q$, with each $\mathcal{H}^j, j = 1, ..., q$, being a space of real-valued measurable functions whose arguments vary across indices. We assume that $\mathcal{H}$ is an infinite dimensional subset of $\mathbf{H} \equiv \mathbf{H}^1 \times \cdots \times \mathbf{H}^q$, a separable Banach space with norm $||h||_s \equiv \sum_{\ell=1}^q ||h_\ell||_{s,\ell}$.

Denote by $m_j(X,h) \equiv \int \rho_j(y, X_z, h(\cdot)) dF_{Y|X}(y)$ the conditional mean function of $\rho_j(Y, X_z, h(\cdot))$ given X for $j = 1, ..., d_\rho$. Then m_j is a (nonlinear) mapping (or operator) from $\mathcal{H}$ into $L^2(f_X)$ such that $m_j(\cdot, h_0)$ is a zero function in $L^2(f_X)$ for all $j = 1, ..., d_\rho$. (Note that the functional form of $m_j(X,h)$ is unknown since the conditional distribution $F_{Y|X}$ is not specified.) Let $m(X,h) \equiv \big(m_1(X,h), ..., m_{d_\rho}(X,h)\big)'$. Under the assumption that model (1.4) identifies $h_0 \in \mathcal{H}$, we have

$$E\left[m(X,h)'m(X,h)\right] \geq 0 \text{ for all } h \in \mathcal{H}; \text{ and } = 0 \text{ if and only if } h = h_0. \tag{2.1}$$

One could construct an estimator of $h_0 \in \mathcal{H}$ by minimizing a sample analog of $E[m(X,h)'m(X,h)]$ over the function space $\mathcal{H}$. Unfortunately, when $h_0(\cdot)$ depends on the endogenous variables Y, the "identifiable uniqueness" condition for consistency might fail in the sense that for any $\varepsilon > 0$ there are sequences $\{h_k\}_{k=1}^\infty$ in $\mathcal{H}$ with $\liminf_{k\to\infty} ||h_k - h_0||_s \geq \varepsilon > 0$ but $\liminf_{k\to\infty} E[m(X,h_k)'m(X,h_k)] = 0$; that is, the metric $||h - h_0||_s$ is not continuous with respect to the population criterion function $E[m(X,h)'m(X,h)]$, and the problem is ill-posed.[12]

[12] An alternative way to explain the ill-posed problem is that the inverse of the unknown (nonlinear) mapping $m_j : (\mathcal{H}, \|\cdot\|_s) \to (L^2(f_X), \|\cdot\|_{L^2(f_X)})$ is not continuous for at least one $j = 1, ..., d_\rho$.

Therefore, in order to design a consistent estimator for $h_0(\cdot)$ we need to tackle two issues. First, since the functional form of $m(X,h)$ is unknown, we need to replace the population minimum distance criterion, $E\left[||m(X,h)||_E^2\right]$, by a consistent empirical estimate. Second, we need to regularize the problem in order to make the metric $||h-h_0||_s$ continuous with respect to the criterion function.

2.1 PSMD estimators

In this paper we consider the class of penalized sieve minimum distance (PSMD) estimators:

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

where $\widehat{m}(X,h)$ is any nonparametric consistent estimator of $m(X,h)$; $\mathcal{H}_n \equiv \mathcal{H}_n^1 \times \cdots \times \mathcal{H}_n^q$ is a sieve parameter space whose complexity (denoted as $k(n) \equiv \dim(\mathcal{H}_n)$) grows with sample size n and becomes dense in the original function space $\mathcal{H}$; $\lambda_n \geq 0$ is a penalization parameter such that $\lambda_n \to 0$ as $n \to \infty$; the penalty $\widehat{P}_n() \geq 0$ is an empirical analog of a non-random penalty function $P : \mathcal{H} \to [0, +\infty)$.

The sieve space $\mathcal{H}_n$ in the definition of the PSMD estimator (2.2) could be finite-dimensional, infinite-dimensional, compact or non-compact (in $\|\cdot\|_s$). Commonly used finite-dimensional linear sieves (also called series) take the form:

$$\mathcal{H}_n = \left\{h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot)\right\}, \quad k(n) < \infty,\ k(n) \to \infty \text{ slowly as } n \to \infty, \tag{2.3}$$

where $\{q_k\}_{k=1}^{\infty}$ is a sequence of known basis functions of a Banach space $(\mathbf{H}, \|\cdot\|_s)$ such as wavelets, splines, Fourier series, Hermite polynomial series, etc.[13] Commonly used linear sieves with constraints can be expressed as:

$$\mathcal{H}_n = \left\{h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot),\ Q_n(h) \leq B_n\right\}, \quad B_n \to \infty \text{ slowly as } n \to \infty, \tag{2.4}$$

where the constraint $Q_n(h) \leq B_n$ reflects prior information about $h_0 \in \mathcal{H}$ such as smoothness properties. The sieve space $\mathcal{H}_n$ in (2.4) is finite dimensional and compact (in $\|\cdot\|_s$) if and only if $k(n) < \infty$ and $\mathcal{H}_n$ is closed and bounded; it is infinite dimensional and compact (in $\|\cdot\|_s$) if and only if $k(n) = \infty$ and $\mathcal{H}_n$ is closed and totally bounded. For example, $\mathcal{H}_n = \left\{h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot),\ \|h\|_s \leq \log(n)\right\}$ is compact if $k(n) < \infty$, but it is not compact if $k(n) = \infty$.

The penalty function $P()$ in the definition of the PSMD estimator (2.2) is typically convex and/or *lower semicompact* (i.e., the set $\{h \in \mathcal{H} : P(h) \leq M\}$ is compact in $(\mathbf{H}, \|\cdot\|_s)$ for all $M \in$

[13] See Newey (1997), Chen (2007) and the references therein for additional examples of linear sieves (or series), and nonlinear sieves.

$[0,\infty))$, and reflects prior information about $h_0 \in \mathcal{H}$. For instance, when $\mathcal{H} \subseteq L^p(d\mu)$, $1 \leq p < \infty$, a commonly used penalty function is $\widehat{P}_n(h) = ||h||_s^p = ||h||_{L^p(d\mu)}^p$ for a known measure $d\mu$, or $\widehat{P}_n(h) = ||h||_{L^p(d\widehat{\mu})}^p$ for an empirical measure $d\widehat{\mu}$ when $d\mu$ is unknown. When $\mathcal{H}$ is a mixed weighted Sobolev space $\{h : ||h||_{L^2(d\mu)}^2 + ||\nabla^r h||_{L^p(leb)}^p < \infty\}$, $1 \leq p < \infty$, where $\nabla^r h$ is the r-th derivative of h for some integer $r \geq 1$, we can let $||\cdot||_s$ be the $L^2(d\mu)-$norm, and $\widehat{P}_n(h) = ||h||_{L^2(d\widehat{\mu})}^2 + ||\nabla^k h||_{L^p(leb)}^p$ or $\widehat{P}_n(h) = ||\nabla^k h||_{L^p(leb)}^p$.

Our definition of PSMD estimators includes many existing estimators as special cases. For example, when $\lambda_n = 0$ and $\mathcal{H}_n$ is a finite-dimensional (i.e., $k(n) < \infty$), compact sieve space of $\mathcal{H}$, the PSMD estimator (2.2) becomes:

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

which is the original SMD estimator proposed in NP, AC and CIN. When $\lambda_n \widehat{P}_n() > 0$, $\widehat{P}_n() = P()$ and $\mathcal{H}_n = \mathcal{H}$ (i.e., $k(n) = \infty$), the PSMD estimator (2.2) becomes:

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

which is a function space penalized minimum distance estimator. When the penalty $P(h)$ is of the classical Tikhonov type (e.g., $\int \{h(y)\}^2 dy$ or $\int \{\nabla^r h(y)\}^2 dy$), such an estimator is also called the TR-MD estimator. See DFR, HH, CFR, GS, HL and CGS for their TR-MD estimators for the NPIV and NPQIV models.

To solve the ill-posed inverse problem, the PSMD procedure (2.2) effectively combines two types of regularization methods: the regularization by sieves and the regularization by penalization. The family of PSMD procedures consists of two broad subclasses: (1) PSMD using slowly growing finite dimensional sieves ($k(n)/n \to 0$), with small flexible penalty ($\lambda_n P() \searrow 0$ fast) or zero penalty ($\lambda_n P() = 0$); (2) PSMD using large dimensional sieves ($k(n)/n \to const. > 0$), with positive penalty ($\lambda_n P() > 0$) that is convex and/or lower semicompact. The first subclass of PSMD procedures mainly follows the regularization by sieves approach, while the second subclass adopts the regularization by penalizing criterion function approach.

The class of PSMD procedures using slowly growing finite dimensional sieves ($k(n)/n \to 0$) solves the ill-posed inverse problem by restricting the complexity of the sieve spaces (and the sieve tuning parameter $k(n)$), while imposing very mild restrictions on the penalty. It includes the original SMD procedure as a special case by setting $\lambda_n = 0$ and taking $\mathcal{H}_n$ to be a finite dimensional compact sieve. However, it also allows for $\lambda_n \searrow 0$ fast with $\mathcal{H}_n$ a finite dimensional linear sieve (i.e., series), which is computationally easier than the original SMD procedure.

On the other hand, the class of PSMD procedures using large dimensional sieves solves the ill-posed inverse problem by imposing strong restrictions on the penalty (and the penalization tuning

parameter $\lambda_n > 0$), but very mild restrictions on the sieve spaces. It includes the function space TR-MD procedure as a special case by setting $\mathcal{H}_n = \mathcal{H}$ (i.e., $k(n) = \infty$) and $\lambda_n \searrow 0$ slowly. Moreover, it also allows for large but finite dimensional ($k(n) < \infty$) linear sieves with $k(n)/n \to const. > 0$ and $\lambda_n \searrow 0$ slowly, which is computationally much easier than the function space TR-MD procedure.

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 the space $\mathcal{H}$ is closed, convex (but not compact in $||\cdot||_s$), it is computationally attractive to use a convex penalization function $\lambda_n \widehat{P}_n(h)$ in h, and a closed convex sieve space $\mathcal{H}_n$ (e.g., Q_n is a positive convex function in the definition of the sieve space (2.4)). To see why, let $clsp(\mathcal{H}_n)$ denote the closed linear span of $\mathcal{H}_n$ (in $||\cdot||_s$). Then the PSMD procedure (2.2) is equivalent to

$$\inf_{h \in clsp(\mathcal{H}_n)} \left\{ \frac{1}{n}\sum_{i=1}^n \widehat{m}(X_i,h)'\widehat{m}(X_i,h) + \lambda_n \widehat{P}_n(h) + \nu_n Q_n(h) \right\}, \tag{2.5}$$

where $Q_n(\widehat{h}_n) \leq B_n$ and $\nu_n \geq 0$ is such that $\nu_n(Q_n(\widehat{h}_n) - B_n) = 0$; see Eggermont and LaRiccia (2001). Therefore, in this case we can recast the constrained optimization problem that represents our PSMD estimator as an unconstrained problem with penalization $\nu_n Q_n(h)$. For most applications, it suffices to have either $\lambda_n \widehat{P}_n(h) > 0$ or $\nu_n Q_n(h) > 0$.

Even when $n^{-1}\sum_{i=1}^n \widehat{m}(X_i,h)'\widehat{m}(X_i,h)$ is not convex in h, our Monte Carlo simulations indicate that it is still much easier to compute PSMD estimators using finite dimensional linear sieves (i.e., series) with small penalization $\lambda_n > 0$.

2.2 Nonparametric estimation of $m(\cdot, h)$

In order to compute the PSMD estimator $\widehat{h}_n$ defined in (2.2), a nonparametric estimator of the conditional mean function $m(\cdot,h) \equiv E[\rho(X,h)|X = \cdot]$ is needed. In the subsequent theoretical sections we establish the asymptotic properties of the PSMD estimator (2.2) allowing for any nonparametric estimator $\widehat{m}(\cdot,h)$ of $m(\cdot,h)$, provided that it satisfies the following assumption regarding the rate of convergence. Let $\{\delta_{m,n}\}_{n=1}^\infty$ be a sequence of positive real values that decreases to zero as $n \to \infty$.

Assumption 2.1. *(i)* $\sup_{h\in\mathcal{H}_n} E\left[\|\widehat{m}(X,h) - m(X,h)\|_E^2\right] = O_p(\delta_{m,n}^2)$*; (ii) there are finite constants* $c, c' > 0$ *such that, except on an event whose probability goes to zero as* $n \to \infty$, $cE\left[||\widehat{m}(X,h)||_E^2\right] \leq n^{-1}\sum_{i=1}^n ||\widehat{m}(X_i,h)||_E^2 \leq c'E\left[||\widehat{m}(X,h)||_E^2\right]$ *uniformly over* $h \in \mathcal{H}_n$.

Many commonly used nonparametric estimators of the conditional mean function $m(X,h)$ can be shown to satisfy assumption 2.1. In the empirical application and Monte Carlo simulations we use a series least square (LS) estimator

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

where $\{p_j()\}_{j=1}^\infty$ is a sequence of known basis functions that can approximate any square integrable function of X well, J_n is the number of approximating terms such that $J_n \to \infty$ 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.

Under the following two mild assumptions 2.2 and 2.3, one can show that the series LS estimator $\widehat{m}(X, h)$ defined in (2.6) satisfies assumption 2.1 with $\delta_{m,n}^2 = \max\{\frac{J_n}{n}, b_{m,J_n}^2\}$ (see Lemmas B.2 and B.3 in the Appendix), where $\{b_{m,J}\}_{J=1}^{\infty}$ is a sequence of positive real values that decreases to zero as $J \to \infty$, denoting the bias of the series approximation error, and $\frac{J_n}{n}$ is the order of the variance of the series LS estimator.

Assumption 2.2. *(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 $E\left[p^{J_n}(X)p^{J_n}(X)'\right]$ are bounded and bounded away from zero for all J_n; (iii) Denote $\xi_n \equiv \sup_{X\in\mathcal{X}} \left\|p^{J_n}(X)\right\|_E$. Either $\xi_n^2 J_n = o(n)$ or $J_n \log(J_n) = o(n)$ for $p^{J_n}(X)$ a polynomial spline sieve.*

Assumption 2.3. *(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 $p^{J_n}(X)$ a sieve with $\xi_n = O(J_n^{1/2})$.*

In assumption 2.2, if $p^{J_n}(X)$ is a spline, cosine/sine or wavelet sieve, then $\xi_n \asymp J_n^{1/2}$; see e.g. Newey (1997) or Huang (1998). Assumption 2.3(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 a Hölder ball (denoted as $\Lambda_c^{\alpha_m}(\mathcal{X})$) or a Sobolev ball (denoted as $W_{2,c}^{\alpha_m}(\mathcal{X}, leb)$)[14] with $\alpha_m > 0$, then assumption 2.3(ii) (a) and (b) hold for tensor product polynomial splines, wavelets or Fourier series sieves with $b_{m,J_n} = J_n^{-r_m}$ where $r_m = \alpha_m/d_x$.

3 Consistency

In Appendix B we provide high-level regularity conditions for general consistency results for an approximate penalized sieve extremum estimator that applies to both well-posed and ill-posed problems. Here in the main text we provide low-level sufficient conditions for consistency of the PSMD estimator (2.2).

[14] See Appendix A for definitions of Hölder ball, Sobolev ball, Hölder space, Sobolev space and other widely used function spaces in economics.

We first impose the following three basic regularity assumptions on identification, the parameter space, the sieve space and the penalty function.

Assumption 3.1. *(i)* $\{(Y_i', X_i')\}_{i=1}^n$ *is a random sample from the joint distribution of* (Y', X')*; (ii)* $\mathcal{H}$ *is a non-empty subset of* $\mathbf{H}$*, and* $\mathbf{H} \equiv \mathbf{H}^1 \times \cdots \times \mathbf{H}^q$ *is a separable Banach space under a metric* $\|h\|_s \equiv \sum_{\ell=1}^q \|h_\ell\|_{s,\ell}$*; (iii)* $E[\rho(Z, h_0)|X] = 0$*, and* $\|h_0 - h\|_s = 0$ *for any* $h \in \mathcal{H}$ *with* $E[\rho(Z, h)|X] = 0$*.*

Assumption 3.2. *(i)* $\{\mathcal{H}_k : k \geq 1\}$ *are non-empty 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* $||\Pi_n h_0 - h_0||_s = o(1)$*; (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* $\|\cdot\|_s$*.*

Assumption 3.3. *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* $\lambda_n |P(\Pi_n h_0) - P(h_0)| = O(\lambda_n) = o(1)$*.*

In the following, for the sake of easy reference, we present consistency results for PSMD estimators using slowly growing finite dimensional sieves and PSMD estimators using large or infinite dimensional sieves in separate subsections. None of the consistency theorems require the $\|\cdot\|_s$ −compactness of the function parameter space $\mathcal{H}$, but they differ in terms of the choice of the key tuning parameters (sieve number of terms $k(n)$ vs penalization parameter λ_n).

3.1 PSMD using slowly growing finite dimensional sieves

Assumption 3.4. *For each integer* $k < \infty$*, (i)* $\dim(\mathcal{H}_k) < \infty$*; the sieve spaces* $\mathcal{H}_k$ *is closed and bounded under* $|| \cdot ||_s$*; (ii)* $E[m(X, h)' m(X, h)]$ *is lower semicontinuous on* $(\mathcal{H}_k, \|\cdot\|_s)$*, i.e., the set* $\{h \in \mathcal{H}_k : E[m(X, h)' m(X, h)] \leq M\}$ *is closed under* $|| \cdot ||_s$ *for all* $M \in [0, \infty)$*.*

Assumptions 2.1, 3.3 and 3.4 ensure that the PSMD $\hat{h}_n$ estimator is well-defined. Under assumption 3.4, for all $\varepsilon > 0$ and each fixed $k \geq 1$,

$$g(k, \varepsilon) \equiv \min_{h \in \mathcal{H}_k : ||h - h_0||_s \geq \varepsilon} E\left[m(X, h)' m(X, h)\right]$$

exists, and is strictly positive (under assumption 3.1(iii)). Moreover, for fixed k, $g(k, \varepsilon)$ increases as ε increases. For any fixed $\varepsilon > 0$, $g(k, \varepsilon)$ decreases as k increases, and $g(k, \varepsilon)$ could go to zero as k goes to infinity.

Theorem 3.1. *Let* $\hat{h}_n$ *be the PSMD estimator with* $\lambda_n \geq 0$*,* $\lambda_n = o(1)$*, and* $\widehat{m}(X, h)$ *any nonparametric estimator of* $m(X, h)$ *satisfying assumption 2.1. Suppose that assumptions 3.1, 3.2, 3.3, and 3.4 hold. Let* $k(n) < \infty$ *and* $k(n) \to \infty$ *as* $n \to \infty$*. If*

$$\max\left\{\delta_{m,n}^2, E(\|m(X, \Pi_n h_0)\|_E^2), \lambda_n\right\} = o\left(g\left(k(n), \varepsilon\right)\right) \quad \text{for all } \varepsilon > 0, \tag{3.1}$$

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

Theorem 3.1 applies to a PSMD estimator using slowly growing finite dimensional ($k(n) < \infty$, $k(n)/n \to 0$) compact sieves, allowing for no penalty ($\lambda_n = 0$), or any flexible penalty $P(h)$ with $\lambda_n > 0$. It is clear that $\liminf_{k\to\infty} g(k,\varepsilon) = \inf_{h\in\mathcal{H}:||h-h_0||_s\geq\varepsilon} E[m(X,h)'m(X,h)]$. Thus, given assumption 3.1(iii), for all $\varepsilon > 0$, $\liminf_{k\to\infty} g(k,\varepsilon) > 0$ if $\mathcal{H}$ is compact in $||\cdot||_s$; otherwise $\liminf_{k\to\infty} g(k,\varepsilon)$ could be zero. Restriction (3.1) allows for $\liminf_{k(n)\to\infty} g(k(n),\varepsilon) = 0$. For a PSMD estimator with positive penalty, restriction (3.1) is trivially satisfied by setting $\lambda_n = O(\max\{\delta_{m,n}^2, E(\|m(X,\Pi_n h_0)\|_E^2)\}) = o(g(k(n),\varepsilon))$. For a PSMD estimator using slowly growing finite dimensional sieves without a penalty ($\lambda_n = 0$), one has to choose the sieve space $\mathcal{H}_n$ and the sieve number of terms $k(n)$ to ensure restriction (3.1). Theorem 3.1 (with $\lambda_n = 0$) implies consistency for the original SMD estimator of NP, AC and CIN without assuming $||\cdot||_s-$compactness of $\mathcal{H}$.

NPIV example (1.2): For this model, $m(X,h_0) = E[Y_1 - h_0(Y_2)|X] = 0$ and $m(X,h) = E[Y_1 - h(Y_2)|X] = E[h_0(Y_2) - h(Y_2)|X]$. Let $\mathcal{H} = \{h \in L^2(f_{Y_2}) : ||h||_{L^2(f_{Y_2})} \leq M < \infty\}$ and $||\cdot||_s = ||\cdot||_{L^2(f_{Y_2})}$. Under very mild regularity conditions on the conditional density of Y_2 given X, $E[\cdot|X]$ is a compact operator mapping from $\mathcal{H} \subseteq L^2(f_{Y_2})$ to $L^2(f_X)$ (see, e.g., BCK), which 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 $\{\phi_{0k}()\}_{k=1}^\infty$ are eigenfunctions in $L^2(f_{Y_2})$ and $L^2(f_X)$ respectively. Let $\mathcal{H}_n = \{h \in \mathcal{H} : h(y_2) = \sum_{k=1}^{k(n)} a_k\phi_{1,k}(y_2)\}$. Then

$$\begin{aligned} E(\|m(X,\Pi_n h_0)\|_E^2) &= E[(E[\Pi_n h_0(Y_2) - h_0(Y_2)|X])^2] = \sum_{j=k(n)+1}^{\infty} \mu_j^2 |\langle h_0, \phi_{1,j}\rangle_{L^2(f_{Y_2})}|^2 \\ &\leq \mu_{k(n)+1}^2 \sum_{j=k(n)+1}^{\infty} |\langle h_0, \phi_{1,j}\rangle_{L^2(f_{Y_2})}|^2 = \mu_{k(n)+1}^2 ||\Pi_n h_0 - h_0||_s^2. \end{aligned}$$

Since $\mathcal{H}_n$ is finite dimensional, bounded and closed, it is compact; thus there is an element $h_n^* \in \mathcal{H}_n$ and $||h_n^* - h_0||_s \geq \varepsilon$ such that $h_n^* = \arg\min_{h\in\mathcal{H}_n:||h-h_0||_s\geq\varepsilon} E[(E[h(Y_2) - h_0(Y_2)|X])^2]$. Then

$$\begin{aligned} g(k(n),\varepsilon) &= E[(E[h_n^*(Y_2) - h_0(Y_2)|X])^2] = \sum_{j=1}^{\infty} \mu_j^2 |\langle h_n^* - h_0, \phi_{1,j}\rangle_{L^2(f_{Y_2})}|^2 \\ &\geq \mu_{k(n)}^2 \sum_{j=1}^{k(n)} |\langle h_n^* - h_0, \phi_{1,j}\rangle_{L^2(f_{Y_2})}|^2 = \mu_{k(n)}^2 ||h_n^* - \Pi_n h_0||_s^2 \end{aligned}$$

Note that the term $||h_n^* - \Pi_n h_0||_s^2$ is bounded below by a constant $c(\varepsilon) > 0$ for all $k(n)$ large enough; for otherwise there is a large $k(n)$ such that $||h_n^* - \Pi_n h_0||_s^2 < (\varepsilon/3)^2$ and thus $||h_n^* - h_0||_s \leq \varepsilon/3 + ||\Pi_n h_0 - h_0||_s < 2\varepsilon/3 < \varepsilon$. This, however, contradicts the fact that $||h_n^* - h_0||_s \geq \varepsilon$ for all $k(n)$. Hence, $E(\|m(X,\Pi_n h_0)\|_E^2)/g(k(n),\varepsilon) \leq const. \times ||\Pi_n h_0 - h_0||_s^2$. By assuming $||\Pi_n h_0 - h_0||_s^2 = o(1)$ and $\max\left\{\delta_{m,n}^2, \lambda_n\right\} = o\left(\mu_{k(n)}^2\right)$, restriction (3.1) is satisfied hence conclusion of Theorem 3.1 holds.

Furthermore, if we let $\lambda_n = 0$ and $\widehat{m}(X,h)$ be the series LS estimator of $m(X,h)$, then, under the conditions $||\Pi_n h_0 - h_0||_s^2 = o(1)$ and $\max\left\{\frac{J_n}{n}, b_{m,J_n}^2\right\} = \frac{J_n}{n} = const.\frac{k(n)}{n} = o\left(\mu_{k(n)}^2\right)$, we obtain the consistency of the original SMD estimator for the NPIV model when $\mathcal{H}$ is not compact in $||\cdot||_s = ||\cdot||_{L^2(f_{Y_2})}$. We note that these conditions are the same as those in theorem 2 of BCK.

3.2 PSMD using large or infinite dimensional sieves

In this subsection we present two consistency results for the PSMD estimator using large or infinite dimensional sieves ($k(n)/n \to const. > 0$), depending on the properties of the penalty function.

3.2.1 Lower semicompact penalty

Assumption 3.5. *$P()$ is lower semicompact, i.e., the set $\{h \in \mathcal{H} : P(h) \leq M\}$ is compact under $||\cdot||_s$ for all $M \in [0,\infty)$.*

Assumption 3.6. *(i) the sieve spaces $\mathcal{H}_{k(n)}$ are closed under $||\cdot||_s$; (ii) $E[m(X,h)'m(X,h)]$ is lower semicontinuous on $\mathcal{H}$ under $\|\cdot\|_s$.*

Assumptions 2.1, 3.3(b), 3.5 and 3.6 ensure that the PSMD $\hat{h}_n$ estimator is well-defined. The next consistency result indicates that the lower semicompact penalty converts an ill-posed problem to a well-posed one.[15]

Theorem 3.2. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n > 0$, $\lambda_n = o(1)$, and $\widehat{m}(X,h)$ any nonparametric estimator of $m(X,h)$ satisfying assumption 2.1. Suppose that assumptions 3.1, 3.2, 3.3(b), 3.5 and 3.6 hold. If*

$$\max\left\{\delta_{m,n}^2, E\{\|m(X,\Pi_n h_0)\|_E^2\}\right\} = O\left(\lambda_n\right), \tag{3.2}$$

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

Theorem 3.2 applies to the PSMD estimator with positive lower semicompact penalty functions, allowing for $k(n) = \infty$ or $k(n)/n \to const. > 0$. To apply this theorem with lower semicompact penalty, it suffices to choose the penalization parameter $\lambda_n > 0$ to ensure restriction (3.2).

NPIV example (1.2): For this model, assumption 3.6(ii) is trivially satisfied with the norm $||h||_s = ||h||_{L^2(\mathcal{R}^d, f_{Y_2})}$ or $= \sup_{y\in\mathcal{R}^d}|(1+|y|^2)^{-\theta/2}h(y)|$ for some $\theta \geq 0$, and assumption 3.6(i) is satisfied by a wide range of linear sieves. To verify assumption 3.5, it suffices to choose a penalty function such that the embedding of the set $\{h \in \mathcal{H} : P(h) \leq M\}$ into $(\mathcal{H}, ||\cdot||_s)$ is compact. For example, if $||\cdot||_s = ||\cdot||_{L^2(f_{Y_2})}$ then $P(h) = ||(1+|\cdot|^2)^{-\vartheta/2}h(\cdot)||_{W_p^\alpha(\mathcal{R}^d)}^p$ with $0 < p \leq 2$, $\alpha > \frac{d}{p} - \frac{d}{2}$ and $\vartheta \geq 0$, $f_{Y_2}(y_2)|y_2|^\vartheta \to 0$ as $|y_2| \to \infty$ will yield the desired result (see Appendix A). If $||h||_s = \sup_{y\in\mathcal{R}^d}|(1+|y|^2)^{-\theta/2}h(y)|$ then both $P(h) = ||(1+|\cdot|^2)^{-\vartheta/2}h(\cdot)||_{\Lambda^\alpha(\mathcal{R}^d)}$ with $\alpha > 0$, $\theta > \vartheta$ and $P(h) = ||(1+|\cdot|^2)^{-\vartheta/2}h(\cdot)||_{W_p^\alpha(\mathcal{R}^d)}^p$ with $0 < p < \infty$, $\alpha > \frac{d}{p}$, $\theta > \vartheta$ are lower

[15] We are grateful to Victor Chernozhukov for pointing out this nice property of lower semicompact penalties.

semicompact (see Appendix A). Theorem 3.2 immediately implies $||\widehat{h}_n - h_0||_{L^2(f_{Y_2})} = o_P(1)$ or $\sup_{y\in\mathcal{R}^d} |(1+|y|^2)^{-\theta/2}[\widehat{h}_n(y) - h_0(y)]| = o_P(1)$. Moreover, these examples of lower semicompact penalties $P(h)$ are also convex when $p \geq 1$, but are not convex when $0 < p < 1$, which illustrates that one can have penalties that are lower semicompact but not convex.

Remark 3.1. *When $P(h)$ is both lower semicompact and convex, under assumption 3.1(iii), the PSMD 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 $\{h \in \mathcal{H}_{k(n)} : \widehat{P}_n(h) \leq M_n\}$:*

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

Therefore, Theorem 3.2 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 $||\cdot||_s-$compactness of the function parameter space $\mathcal{H}$. In particular, this immediately implies the consistency of the SMD estimators of the NPIV model (1.2) studied in NP and BCK without requiring that $\mathcal{H}$ is a compact subset of the space $L^2(f_{Y_2})$.

3.2.2 Convex penalty

In this subsection we present consistency results for PSMD estimators with general penalty functions that may not be lower semicompact, but satisfy the following assumption.

Assumption 3.7. *$\lambda_n > 0$, $\lambda_n \sup_{h\in\mathcal{H}_n} |\widehat{P}_n(h) - P(h)| = o_P(\lambda_n)$, with $P(\cdot)$ a non-negative real-valued measurable function of $h \in \mathcal{H}$, $P(h_0) < \infty$ and $\lambda_n|P(\Pi_n h_0) - P(h_0)| = o(\lambda_n)$.*

Assumption 3.7 is a stronger version of assumption 3.3(b). 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 $\|\cdot\|_s$*. Note that assumptions 3.3(b) and 3.7 are trivially satisfied when $\mathcal{H}_n = \mathcal{H}$ and $\widehat{P}_n = P$.

For a Banach space $\mathbf{H}$ we denote $\mathbf{H}^*$ as the dual of $\mathbf{H}$ (i.e., the space of all bounded linear functionals on $\mathbf{H}$), and a bilinear form $\langle\cdot,\cdot\rangle_{\mathbf{H}^*,\mathbf{H}} : \mathbf{H}^* \times \mathbf{H} \to \mathcal{R}$ as the inner product that links the space $\mathbf{H}$ with its dual $\mathbf{H}^*$.

Assumption 3.8. *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 $||\cdot||_s$, and a non-decreasing lower semicontinuous function $g()$ with $g(0) = 0, g(\varepsilon) > 0$ for $\varepsilon > 0$, such that $P(h) - P(h_0) - \langle t_0, h - h_0\rangle_{\mathbf{H}^*,\mathbf{H}} \geq g(||h - h_0||_s)$ for all $h \in \mathcal{H}_k$ and all $k \geq 1$.*

When $\mathcal{H}$ is convex, Assumption 3.8 is satisfied if $P(h)$ is *strongly convex* at h_0 under $||\cdot||_s$, that is, there exists a $c > 0$ such that $P(h) - P(h_0) - \langle DP(h_0), h - h_0\rangle_{\mathbf{H}^*,\mathbf{H}} \geq c||h - h_0||_s^2$ 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 strong convexity

is satisfied by commonly used penalization functions, and it obviously implies that $P(h)$ is *strictly convex* at h_0 (i.e., $P(\lambda h + (1-\lambda)h_0) < \lambda P(h) + (1-\lambda)P(h_0)$ for all $\lambda \in (0,1)$ and all $h \in \mathcal{H}$ with $||h - h_0||_s > 0$). See, e.g., Eggermont and LaRiccia (2001).

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^α for $1 < p < \infty$ are reflexive and separable Banach spaces.

Assumption 3.9. *(i)* $(\mathbf{H}, ||\cdot||_s)$ *is a reflexive Banach space; (ii)* $\mathcal{H}$ *is a non-empty, closed and convex subset in* $(\mathbf{H}, ||\cdot||_s)$*; (iii)* $\mathcal{H}$ *is bounded in* $||\cdot||_s$ *(i.e.,* $\sup_{h\in\mathcal{H}} ||h||_s \leq K < \infty$*).*

Assumption 3.9(iii) is implied by the so-called *coercive* condition, denoted as
Assumption 3.9(iii)'*:* $E[m(X,h)'m(X,h)] + \lambda P(h) \to +\infty$ *as* $||h||_s \to +\infty$ *for* $h \in \mathcal{H}$ *and* $\lambda \in (0,1]$.

Assumption 3.10. *Either (a) or (b) holds: (a)* $m(\cdot,h) : \mathcal{H} \subseteq \mathbf{H} \to L^2(f_X)$ *is compact (i.e., continuous and maps bounded sets in* $\mathcal{H}$ *into relatively compact sets in* $L^2(f_X)$*); (b)* $E[m(X,h)'m(X,h)]$ *is convex and lower semicontinuous on* $\mathcal{H}$ *(in* $\|\cdot\|_s$*).*

Assumption 3.11. *Either (a) or (b) holds: (a)* $\mathcal{H}_k$ *are compact under* $||\cdot||_s$*, and* $P(h)$ *is lower semicontinuous on* $\mathcal{H}_k$ *(in* $\|\cdot\|_s$*); (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* $\|\cdot\|_s$*).*

Assumptions 3.9, 3.10 and 3.11 are simple sufficient conditions for the general consistency Lemma B.6 stated in Appendix B.

Theorem 3.3. *Let* $\hat{h}_n$ *be the PSMD estimator with* $\lambda_n > 0$*,* $\lambda_n = o(1)$*, and* $\widehat{m}(X,h)$ *any nonparametric estimator of* $m(X,h)$ *satisfying assumption 2.1. Let assumptions 3.1, 3.2, 3.7 -3.11 hold. If*

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

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

Comparing Theorem 3.3 to Theorem 3.2, both consistency results allow for non-compact (in $||\cdot||_s$) parameter space $\mathcal{H}$ and infinite dimensional sieve spaces. Nevertheless, under the global identification assumption 2.1(iii), the condition $\max\{\delta_{m,n}^2, E(\|m(X,\Pi_n h_0)\|_E^2)\} = o(\lambda_n)$ imposed in Theorem 3.3 for a general penalty can be improved to the condition $\max\{\delta_{m,n}^2, E(\|m(X,\Pi_n h_0)\|_E^2)\} = O(\lambda_n)$ in Theorem 3.2 for a lower semicompact penalty. In addition, using a lower semicompact penalty, Theorem 3.2 leads to consistency without imposing assumptions 3.8 and 3.9. This means that by applying Theorem 3.2, one can obtain sup-norm consistency of the PSMD estimator using a lower semicompact penalty.

Identification via strictly convex penalty. When $E[m(X,h)'m(X,h)]$ is convex in $h \in \mathcal{H}$ (e.g. the NPIV model), we can relax the global identification condition (assumption 3.1(iii)) by

using a strictly convex penalty function, that is, we can use a strictly convex penalty to uniquely identify h_0 out of the solution set $\mathcal{M}_0 \equiv \{h \in \mathcal{H} : E[m(X,h)'m(X,h)] = 0\}$.

Assumption 3.12. *(i) $\mathcal{M}_0$ is non-empty; (ii) P is lower semicontinuous and strictly convex on $\mathcal{M}_0$ (in $\|\cdot\|_s$).*

Let $\mathcal{M}_0^P$ be the set of minimum penalization solutions, i.e., $\mathcal{M}_0^P \equiv \{h \in \mathcal{H} : h = \arg\inf_{h' \in \mathcal{M}_0} P(h')\}$.

Theorem 3.4. *(1) Let assumptions 3.9, 3.10(b) and 3.12 hold. Then: $\mathcal{M}_0^P = \{h_0\} \subseteq \mathcal{M}_0$.*

(2) Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n > 0$, $\lambda_n = o(1)$, and $\widehat{m}(X,h)$ any nonparametric estimator of $m(X,h)$ satisfying assumption 2.1. Let assumptions 3.1(i)(ii), 3.2, 3.7, 3.8, 3.9, 3.10(b), 3.11(b) and 3.12 hold. If $\max\{\delta_{m,n}^2, E(\|m(X,\Pi_n h_0)\|_E^2)\} = o(\lambda_n)$, then: $||\widehat{h}_n - h_0||_s = o_P(1)$, and $P(\widehat{h}_n) = P(h_0) + o_P(1)$.

NPIV example (1.2): For this model, Assumption 3.9(i) rules out the (weighted) sup-norm case; but assumption 3.9 is readily satisfied by $\mathbf{H} = L^2(f_{Y_2})$, $||\cdot||_s = ||\cdot||_{L^2(f_{Y_2})}$ and $\mathcal{H} = \{h \in L^2(f_{Y_2}) : ||h||_{L^2(f_{Y_2})} \leq M < \infty\}$. Assumption 3.10(b) is trivially satisfied. Let $P(h) = ||h||^2_{L^2(f_{Y_2})}$ be the penalty function, then Assumption 3.8 is satisfied with $t_0 = 2h_0$. If assumption 3.1(iii) holds then Theorem 3.3 immediately leads to $||\widehat{h}_n - h_0||_{L^2(f_{Y_2})} = o_P(1)$ where $\mathcal{M}_0 = \{h_0\}$. If assumption 3.1(iii) fails to hold in the sense that $\{h_0\} \subset \mathcal{M}_0$, then, since $P(h) = ||h||^2_{L^2(f_{Y_2})}$ is strictly convex, Theorem 3.4 is applicable and we obtain $||\widehat{h}_n - h_0||_{L^2(f_{Y_2})} = o_P(1)$ where $\mathcal{M}_0^P = \{h_0\} \subset \mathcal{M}_0$.

4 Convergence Rates in a Banach Norm

Given the consistency results stated in Section 3, we can now restrict our attention to a shrinking $||\cdot||_s-$neighborhood around h_0. Let $\mathcal{H}_{os} \equiv \{h \in \mathcal{H} : ||h - h_0||_s \leq \epsilon, ||h||_s \leq M_1, P(h) \leq M_0\}$ and $\mathcal{H}_{osn} \equiv \{h \in \mathcal{H}_n : ||h - \Pi_n h_0||_s \leq \epsilon, ||h||_s \leq M_1, P(h) \leq M_0\}$ for a sufficiently small positive ϵ and some known positive finite constants M_1, 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.

We first introduce a pseudo-metric on $\mathcal{H}_{os}$ that is weaker than $||\cdot||_s$. Define the first pathwise derivative in the direction $[h - h_0]$ evaluated at h_0 as

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

Following AC, 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 $||h - h_0||^2 \leq c_1 E[m(X,h)'m(X,h)]$ holds 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 weaker pseudo-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 $h \in \mathcal{H}_{osn}$.*

Assumption 4.2 controls the linear approximation of the penalty function around $\Pi_n h_0$. This assumption is similar to assumption 3.8. It is satisfied when the penalty $P(h)$ is convex in $\Pi_n h_0$.

Before we establish the convergence rate under $||\cdot||_s$, 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

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

The definition of the modulus of continuity,[16] $\omega(\delta, \mathcal{H}_{os})$, 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 *locally ill-posed in rate*.

The definition of the sieve modulus of continuity, $\omega_n(\delta, \mathcal{H}_{osn})$, is closely related to the notion of the *sieve measure of (local) ill-posedness*, τ_n, defined as:

$$\tau_n \equiv \sup_{h \in \mathcal{H}_{osn}: ||h - \Pi_n h_0|| \neq 0} \frac{||h - \Pi_n h_0||_s}{||h - \Pi_n h_0||},$$

which is a direct extension of the one introduced in BCK for the NPIV model (1.2). By definition, the values of $\omega_n(\delta, \mathcal{H}_{osn})$ and τ_n depend on the choice of the sieve space. Nevertheless, for any sieve space $\mathcal{H}_{osn}$ and for any $\delta > 0$, we have: (i) $\omega_n(\delta, \mathcal{H}_{osn}) \leq \tau_n \times \delta$ and $\omega_n(\delta, \mathcal{H}_{osn}) \leq \omega(\delta, \mathcal{H}_{os})$; (ii) $\omega_n(\delta, \mathcal{H}_{osn})$ and τ_n increase as $k(n) = \dim(\mathcal{H}_{osn})$ increases; (iii) $\limsup_{n \to \infty} \omega_n(\delta, \mathcal{H}_{osn}) = \omega(\delta, \mathcal{H}_{os})$ and $\limsup_{n \to \infty} \tau_n = \sup_{h \in \mathcal{H}_{os}: ||h - h_0|| \neq 0} \frac{||h - h_0||_s}{||h - h_0||} = \frac{\omega(\delta, \mathcal{H}_{os})}{\delta}$. In particular, the problem of estimating h_0 under $||\cdot||_s$ is *locally ill-posed in rate* if and only if $\limsup_{n \to \infty} \tau_n = \infty$. These properties of the sieve modulus of continuity $(\omega_n(\delta, \mathcal{H}_{osn}))$ and the sieve measure of (local) ill-posedness (τ_n) justify their use in convergence rate analysis.

We now present a general theorem on the convergence rates under a Banach norm $||\cdot||_s$. Notice that once after we establish $||\hat{h}_n - h_0||_s = o_P(1)$ (consistency), the convergence rate results can be derived without the global identification assumption 3.1(iii).

[16] Our definition of modulus of continuity is inspired by that of Nair, Pereverzev and Tautenhahn (2005) in their study of a linear ill-posed inverse problem with deterministic noise and a known operator.

Theorem 4.1. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n \geq 0$, $\lambda_n = o(1)$, $\widehat{P}_n(h) = P(h)$, and $\widehat{m}(X,h)$ any nonparametric estimator of $m(X,h)$ satisfying assumption 2.1. Let $h_0 \in \mathcal{H}_{os}$ and $\hat{h}_n \in \mathcal{H}_{osn}$ with probability approaching one. If assumptions 3.1(i)(ii), 3.2, 3.3 and 4.1 hold. Then:*

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

under either one of the following three conditions:

(1) $\max\{\delta_{m,n}^2, \lambda_n\} = \delta_{m,n}^2$.

(2) $\max\{\delta_{m,n}^2, \lambda_n\} = \delta_{m,n}^2 = O(\lambda_n)$ *and assumption 3.5 holds.*

(3) $\max\left\{\delta_{m,n}^2, \lambda_n||\widehat{h}_n - \Pi_n h_0||_s\right\} = \delta_{m,n}^2$ *and assumption 4.2 holds.*

Theorem 4.1 under condition (1) allows for slowly growing finite dimensional sieves without a penalty ($\lambda_n = 0$) or with any flexible penalty satisfying $\lambda_n = o\left(||\Pi_n h_0 - h_0||^2\right)$; such cases are loosely called the "sieve dominating case". Theorem 4.1 under conditions (2) or (3) allows for an infinite dimensional sieve ($k(n) = \infty$) or large dimensional sieves ($k(n)/n \to const. > 0$) satisfying $||\Pi_n h_0 - h_0||^2 = o(\lambda_n)$; such cases are loosely called the "penalization dominating case". Theorem 4.1 under conditions (1) or (2) or (3) also allows for finite (but maybe large) dimensional sieves ($k(n)/n \to const. \geq 0$) satisfying $||\Pi_n h_0 - h_0||^2 = O(\lambda_n)$; such cases are loosely called the "sieve penalization balance case".

Remark 4.1. *(1) For PSMD estimators with finite dimensional sieves ($k(n) < \infty$), the conclusion of Theorem 4.1 can be stated as: $||\hat{h}_n - h_0||_s = O_P\left(||h_0 - \Pi_n h_0||_s + \tau_n \times \max\{\delta_{m,n}, ||\Pi_n h_0 - h_0||\}\right)$. This result extends theorem 2 of BCK for the NPIV model to the general model (1.4), allowing for more general sieve approximation error rate and a nonparametric estimator $\widehat{m}(X,h)$ different from the series LS estimator of $m(X,h)$.*

(2) For PSMD estimators with infinite dimensional sieves ($k(n) = \infty$), the conclusion of Theorem 4.1 can be stated as: $||\hat{h}_n - h_0||_s = O_P\left(\omega(\delta_{m,n}, \mathcal{H}_{os})\right)$.

The following corollary establishes the convergence rate for the PSMD 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 PSMD estimator with $\lambda_n = o(1)$ and $\widehat{m}(X,h)$ any nonparametric estimator of $m(X,h)$ satisfying assumption 2.1. 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 Theorem 4.1 remains true.*

To apply Theorem 4.1 and Corollary 4.1, one needs to compute upper bounds on the sieve modulus of continuity $\omega_n(\delta, \mathcal{H}_{osn})$ or on the modulus of continuity $\omega(\delta, \mathcal{H}_{os})$. See section 5 for sufficient conditions to bound these terms.

5 Convergence Rates in a Hilbert Norm

In this section we shall present some sufficient conditions to bound the sieve modulus of continuity and the modulus of continuity. Throughout this section, we restrict $||\cdot||_s$ to be a Hilbert space norm for simplicity. We assume that $\mathcal{H}_{os}$ is an infinite dimensional subset of a separable Hilbert space $\mathbf{H}$ with an inner product $\langle\cdot,\cdot\rangle_s$ and the inner product induced norm $||\cdot||_s$.

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 PSMD with slowly growing finite dimensional sieves

We first provide some sufficient conditions to bound the sieve approximation error rate and the sieve modulus of continuity $\omega_n(\delta, \mathcal{H}_{osn})$.

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) $||h_0 - \sum_{j=1}^{k(n)} \langle h_0, q_j\rangle_s q_j||_s = O(\{\nu_{k(n)}\}^{-\alpha})$ for a finite $\alpha > 0$ and a positive sequence $\{\nu_j\}_{j=1}^{\infty}$ that strictly increases to ∞ as $j \to \infty$.*

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 $h_0 \in W_2^{\alpha}([0,1]^d, leb))$, then assumption 5.1(i) is satisfied with spline, wavelet, 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 continuous increasing function $\varphi : \mathcal{R}_+ \to \mathcal{R}_+$ such that: (i) $||h||^2 \geq c\sum_{j=1}^{\infty} \varphi(\nu_j^{-2})|\langle h, q_j\rangle_s|^2$ for all $h \in \mathcal{H}_{osn}$; (ii) $||h_0 - \Pi_n h_0||^2 \leq C\sum_j \varphi(\nu_j^{-2})|\langle h_0 - \Pi_n h_0, q_j\rangle_s|^2$.*

Assumption 5.2(i) is a low-level sufficient condition that relates the weak norm $||h||$ to its strong norm in a sieve shrinking neighborhood $\mathcal{H}_{osn}$ (of h_0). Assumption 5.2(ii) is the so-called "stability condition" that is only required to hold in terms of the sieve approximation error $h_0 - \Pi_n h_0$. In their convergence rate study of the NPIV model (1.2), BCK and CR actually impose conditions that imply assumption 5.2(i) and (ii). See subsection 5.3 below for further discussion.

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: $\omega_n(\delta, \mathcal{H}_{osn}) \leq const.\times\delta/\sqrt{\varphi(\nu_{k(n)}^{-2})}$ and $\tau_n \leq const./\sqrt{\varphi(\nu_{k(n)}^{-2})}$.

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

(3) If assumption 5.2(i)(ii) holds, then: $\omega_n(||\Pi_n h_0 - h_0||, \mathcal{H}_{osn}) \leq c||\Pi_n h_0 - h_0||_s$.

Theorem 4.1 and Lemma 5.1 together immediately imply the following corollary for the convergence rate of the PSMD estimator using a slowly growing finite dimensional sieve (i.e., $k(n)/n \to 0$):

Corollary 5.1. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n \geq 0$, $\lambda_n = o(1)$, and all the assumptions of Theorem 4.1(1) hold. Let assumptions 5.1 and 5.2 hold with $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ and $k(n) < \infty$. Let $\max\{\delta_{m,n}^2, \lambda_n\} = \delta_{m,n}^2 = const. \times \frac{k(n)}{n} = o(1)$. Then:*

$$||\hat{h}_n - h_0||_s = O_P\left(\max\left(\{\nu_{k(n)}\}^{-\alpha}, \sqrt{\frac{k(n)}{n \times \varphi(\nu_{k(n)}^{-2})}}\right)\right) = O_P\left(\{\nu_{k_o(n)}\}^{-\alpha}\right)$$

where $k_o(n)$ is such that $\{\nu_{k_o(n)}\}^{-2\alpha} \asymp \frac{k_o(n)}{n}\{\varphi(\nu_{k_o(n)}^{-2})\}^{-1}$.

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

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

Corollary 5.1 allows for both the sieve dominating case and the sieve penalization balance case. To apply this corollary to obtain a convergence rate for $||\hat{h}_n - h_0||_s$, we choose $k(n)$ to balance the sieve approximation error rate ($\{\nu_{k(n)}\}^{-\alpha}$) and the model complexity (or roughly the standard deviation) ($\sqrt{\frac{k(n)}{n}\{\varphi(\nu_{k(n)}^{-2})\}^{-1}}$), and let $\max\{\delta_{m,n}^2, \lambda_n\} = \delta_{m,n}^2 = const. \times \frac{k(n)}{n}$. For example, if the PSMD estimator $\hat{h}_n$ is computed using the series LS estimator $\widehat{m}(X,h)$ defined in (2.6), one can let $\delta_{m,n}^2 = \max\{\frac{J_n}{n}, b_{m,J_n}^2\} = \frac{J_n}{n} = const. \times \frac{k(n)}{n} = o(1)$. This corollary extends the rate results of BCK for the NPIV model to the general model (1.4), allowing for more general parameter space $\mathcal{H}$ and other nonparametric estimators of $m(X,h)$.

5.2 PSMD with large or infinite dimensional sieves

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

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

Assumption 5.3 obviously implies assumption 5.1(ii). Under assumption 5.1(i), assumption 5.3 is implied by the so-called *ellipsoid* class:

Assumption 5.3': *There are finite constants $M > 0$, $\alpha > 0$ and a strictly increasing positive sequence $\{\nu_j\}_{j=1}^{\infty}$ such that $\sum_{j=1}^{\infty}\nu_j^{2\alpha}|\langle h, q_j\rangle_s|^2 \leq M^2$ for all $h \in \mathcal{H}_{os}$.*

Assumption 5.3 is the approximation condition imposed in CR in their study of the minimax rate for the NPIV model (1.2). See CR for other sufficient conditions.

Assumption 5.4. *There are finite constants c, $C > 0$ and a continuous increasing function $\varphi : \mathcal{R}_+ \to \mathcal{R}_+$ such that: (i) $||h||^2 \geq c\sum_{j=1}^{\infty}\varphi(\nu_j^{-2})|\langle h, q_j\rangle_s|^2$ for all $h \in \mathcal{H}_{os}$; (ii) $||h||^2 \leq C\sum_{j=1}^{\infty}\varphi(\nu_j^{-2})|\langle h, q_j\rangle_s|^2$ for all $h \in \mathcal{H}_{os}$.*

It is obvious that assumptions 5.4(i) and (ii) imply assumptions 5.2(i) and (ii) respectively. This stronger condition is commonly imposed in the literature on minimax optimal rates for ill-posed inverse problems, but in terms of operator formulations. See subsection 5.3 below for further discussion.

Lemma 5.2. *Let assumptions 5.1(i), 5.3 and 5.4(i) hold. Then: for any $\delta > 0$, there is an integer $k^* \equiv k^*(\delta) \in (1, \infty)$ such that $\delta^2/\varphi(\nu_{k^*-1}^{-2}) < M^2(\nu_{k^*})^{-2\alpha}$ and $\delta^2/\varphi(\nu_{k^*}^{-2}) \geq M^2(\nu_{k^*})^{-2\alpha}$; hence*

(1) $\omega(\delta, \mathcal{H}_{os}) \leq const. \times \delta/\sqrt{\varphi(\nu_{k^*}^{-2})}$.

(2) $\omega_n(\delta, \mathcal{H}_{osn}) \leq const.\times\delta/\sqrt{\varphi(\nu_{\overline{k}}^{-2})}$, with $\overline{k} \equiv \min\{k(n), k^*\} \in (1, \infty)$ and $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$.

Theorem 4.1, Lemmas 5.1 and 5.2 immediately imply the following corollary for the convergence rate of a PSMD estimator using large or infinite dimensional sieves with lower semicompact and/or convex penalties. Let $\delta_{m,n}^*$ denote the optimal convergence rate of $\widehat{m}(\cdot, h) - m(\cdot, h)$ in the root mean squared metric. By definition $\delta_{m,n}^{*2} \leq \delta_{m,n}^2$.

Corollary 5.2. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n > 0$, $\lambda_n = o_P(1)$, and all the assumptions of Theorem 4.1(1) hold. Let assumptions 5.1(i), 5.2(ii), 5.3 and 5.4(i) hold with $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ for $k(n)/n \to const. > 0$ and $\infty \geq k(n) \geq k^*$, where $k^* = k^*(\delta_{m,n}^*)$ is such that $\{\nu_{k^*}\}^{-2\alpha} \asymp \delta_{m,n}^{*2}\{\varphi(\nu_{k^*}^{-2})\}^{-1}$. Let either assumption 3.5 hold with $\lambda_n = O(\delta_{m,n}^{*2})$, or assumption 4.2 hold with $\lambda_n = O\left(\delta_{m,n}^*\sqrt{\varphi(\nu_{k^*}^{-2})}\right)$. Then:*

$$(1) \quad ||\hat{h}_n - h_0||_s = O_P\left(\{\nu_{k^*}\}^{-\alpha}\right) = O_P\left(\delta_{m,n}^*\{\varphi(\nu_{k^*}^{-2})\}^{-\frac{1}{2}}\right).$$

Thus $||\hat{h}_n - h_0||_s = O_P\left(\left(\delta_{m,n}^*\right)^{\frac{\alpha}{\alpha+\varsigma}}\right)$ if $\varphi(\tau) = \tau^{\varsigma}$ for some $\varsigma \geq 0$; and $||\widehat{h}_n - h_0||_s = O_P\left([-\ln(\delta_{m,n}^*)]^{-\alpha/\varsigma}\right)$ if $\varphi(\tau) = \exp\{-\tau^{-\varsigma/2}\}$ for some $\varsigma > 0$.

(2) If $\mathcal{H}_n = \mathcal{H}$ (or $k(n) = \infty$), then assumption 5.2(ii) holds, and result (1) remains true.

The next rate result specializes the above corollary to the PSMD estimator using a series LS estimator of $m(X, h)$, and hence $\delta_{m,n}^{*2} = \frac{J_n^*}{n} \asymp b_{m,J_n^*}^2$ where J_n^* is such that the variance part $(\frac{J_n^*}{n})$ and the squared bias part $(b_{m,J_n^*}^2)$ are of the same order.

Corollary 5.3. *Suppose that all the conditions of Corollary 5.2 hold, and that $m(X, h)$ is estimated using the series LS estimator satisfying assumptions 2.2 and 2.3.*

(1) If $b_{m,J_n} = O(J_n^{-r_m})$ for some $r_m > 0$, then: the conclusions of Corollary 5.2 hold with $\delta_{m,n}^* = O\left(n^{-r_m/(2r_m+1)}\right)$.

(2) If assumptions 5.2(ii) and 5.3 are replaced by assumptions 5.4(ii) and 5.3', then:

$$||\hat{h}_n - h_0||_s = O_P\left(\{\nu_{J_n^*}\}^{-\alpha}\right) = O_P\left(\sqrt{\frac{J_n^*}{n}\{\varphi(\nu_{J_n^*}^{-2})\}^{-1}}\right)$$

where J_n^* is the largest integer such that $\frac{J_n^*}{n} \asymp b_{m,J_n^*}^2 \leq const.\{\nu_{J_n^*}\}^{-2\alpha}\varphi(\nu_{J_n^*}^{-2})$. Thus $||\hat{h}_n - h_0||_s = O_P\left(n^{-\frac{\alpha}{2(\alpha+\varsigma)+d}}\right)$ if $\varphi(\tau) = \tau^\varsigma$ for some $\varsigma \geq 0$ and $\nu_k \asymp k^{1/d}$.

5.3 Further discussion

Given the results of the previous two subsections, it is clear that assumption 5.2 or its stronger version 5.4 is important for the convergence rate of the PSMD estimator. 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 $||h||^2 \equiv ||T_{h_0}h||^2_{L^2(f_X)} = ||(T_{h_0}^*T_{h_0})^{1/2}h||_s^2$. Hence assumption 5.4 can be restated in terms of the operator $T_{h_0}^*T_{h_0}$: *there is a positive increasing function φ such that: $||(T_{h_0}^*T_{h_0})^{1/2}h||_s^2 \asymp \sum_{j=1}^\infty \varphi(\nu_j^{-2})|\langle h, q_j\rangle_s|^2$ for all $h \in \mathcal{H}_{os}$.* This assumption relates the smoothness of the operator $(T_{h_0}^*T_{h_0})^{1/2}$ to the smoothness of the unknown function $h_0 \in \mathcal{H}_{os}$. Assumptions 5.4(i) and (ii) are respectively the *reverse link condition* and the *link condition* imposed in CR in their study of the NPIV model (1.2). It is also assumed in Nair, Pereverzev and Tautenhahn (2005) in their study of a linear ill-posed inverse problem with deterministic noise and a known operator.

Remark 5.1. ***(1)*** *Under assumptions 5.1(i), 5.3 and 5.4(ii), CR establish the minimax lower bound under the metric $||\cdot||_s = ||\cdot||_{L^2(f_{Y_2})}$ for estimation of the NPIV model (1.2) $E[Y_1 - h_0(Y_2)|X] = 0$:*

$$\inf_{\widetilde{h}} \sup_{h\in\mathcal{H}_{os}} E_h[||\widetilde{h} - h||^2_{L^2(f_{Y_2})}] \geq const.n^{-1}\sum_{j=1}^{k_o}[\varphi(\nu_j^{-2})]^{-1} \asymp \{\nu_{k_o}\}^{-2\alpha}$$

where $k_o = k_o(n)$ is the largest integer such that: $\frac{1}{n}\sum_{j=1}^{k_o}\{\nu_j\}^{2\alpha}[\varphi(\nu_j^{-2})]^{-1} \asymp 1$. In addition, suppose that assumption 5.4(i) holds, CR show that the BCK estimator $\widehat{h}_n$, which is a PSMD estimator using a slowly growing finite dimensional sieve and a series LS estimator of $m(X,h)$, achieves this minimax lower bound in probability. HH establish that their kernel based function space TR-MD estimator of the NPIV model achieves the minimax lower bound for the mildly ill-posed case.

(2) *For the NPQIV model (1.3), HL show that their kernel based function space TR-MD estimator achieves the minimax lower bound when the problem is mildly ill-posed. The rates obtained in Corollary 5.1 and Corollary 5.3 (2) for our PSMD estimators of the general model (1.4) achieve the minimax lower bound of CR.*

We conclude this section by mentioning two obvious sufficient conditions for assumption 5.4.

Suppose 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)).[17] 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 $||(T_{h_0}^* T_{h_0})^{1/2} h||_s^2 = \sum_{k=1}^{\infty} \mu_k^2 |\langle h, \phi_{1k} \rangle_s|^2$ for all $h \in \mathbf{H}$. Thus, assumptions 5.1(i) and 5.4 are automatically satisfied with $q_j = \phi_{1j}$ and $b_j \asymp \varphi(\nu_j^{-2}) = \mu_j^2$ for all j.

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 $\psi(0) = 0$ and $\eta^{-1/2}\psi(\eta)$ non-decreasing (in $\eta > 0$) such that

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

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

for a finite constant M, and the original "source condition" corresponds to the choice $\psi(\eta) = \eta^{1/2}$ (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 $\psi(\mu_j^2) = \nu_j^{-\alpha}$ for all $j \geq 1$. Then $\varphi(\tau) = \tau^{\varsigma}$ is equivalent to $\psi(\eta) = \eta^{\alpha/(2\varsigma)}$ and $\eta^{-1/2}\psi(\eta)$ non-decreasing iff $\alpha \geq \varsigma$; $\varphi(\tau) = \exp\{-\tau^{-\varsigma/2}\}$ is equivalent to $\psi(\eta) = [-\log(\eta)]^{-\alpha/\varsigma}$.

6 Application to Nonparametric Additive Quantile IV Regression

In this section we present a detailed application to illustrate the general results obtained in the previous sections. The nonparametric additive quantile IV regression 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 $h_0 \equiv (h_{01}, h_{02}) \in \mathcal{H} \equiv \mathcal{H}^1 \times \mathcal{H}^2$ using the PSMD estimator $\widehat{h}_n$ given in (2.2), with $\mathcal{H}_n = \mathcal{H}_n^1 \times \mathcal{H}_n^2$ 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

[17] 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.

$\widehat{P}_n(h) = P(h_2) \geq 0$. The conditional mean function $m(X,h)$ is estimated by the series LS estimator $\widehat{m}(X,h)$ defined in (2.6).

We present two propositions on consistency. The first one considers a lower semicompact penalty and the second one uses a convex (but not lower semicompact) penalty. For both 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)* $E[(|Y_2|)^{2\theta}] < \infty$ *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 6.1(ii)(iii)(iv) provide sufficient condition to bound $E\{m(X,\Pi_n h_0)^2\}$ (assumption 3.2(ii)).

Condition 6.2. *(i)* $\mathcal{H}^1 = \{h_1 \in \Lambda_1^{\alpha_1}([0,1]^d) : h_1(y_1^*) = 0\}$ *for* $\alpha_1 > 0$ *and some* y_1^**, and* $\mathcal{H}^2 \subset L^2(\mathcal{R}^d, f_{Y_2})$*; (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.*

Condition 6.2 is a global identification condition. Condition 6.2(ii) is similar to the global identification condition for the NPQIV model (1.3) imposed in CH and HL. See CH and CIN for further discussion and sufficient conditions for identification of the NPQIV model (1.3).

Condition 6.3. *(i) assumption 2.2 holds with* $p^{J_n}(X)$ *being a tensor product P-spline, B-spline, wavelet or cosine linear sieve; (ii)* $\mathcal{H}_n = \mathcal{H}_n^1 \times \mathcal{H}_n^2$*, where* $\mathcal{H}_n^1$ *is a tensor product P-spline, B-spline, wavelet, 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* $\mathcal{H}^2$*.*

Condition 6.3(ii) specifies the sieve basis for $h = (h_1,h_2)$. Condition 6.3(i) specifies the basis for the series LS estimator $\widehat{m}(\cdot,h)$. In addition, conditions 6.3(i) and 6.4 together imply a series LS approximation bias rate for $m(\cdot,h)$.

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

We let $||\cdot||_{\mathcal{T}_{p,q}^\alpha}$ denote the norm of a Banach space $\mathcal{T}_{p,q}^\alpha(\mathcal{R}^d, leb)$, which is either a Besov space $\mathcal{B}_{p,q}^\alpha(\mathcal{R}^d, leb)$ for $p,q \in [1,\infty]$ or an F-space $\mathcal{F}_{p,q}^\alpha(\mathcal{R}^d, leb)$ for $p \in [1,\infty), q \in [1,\infty]$; see Appendix A for their definitions and properties. We also denote $r_m \equiv \alpha_m/d_x$, $r_1 \equiv \alpha_1/d$ and $r_2 \equiv \alpha_2/d$.

We now present the first consistency result in which the parameter space $\mathcal{H}^2$ is not compact but the penalty is lower semicompact. It is an application of Theorem 3.2.

Proposition 6.1. *For the model (6.1), let* $\hat{h}_n$ *be the PSMD estimator with* $\lambda_n > 0$*,* $\lambda_n = o(1)$ *and let* $\widehat{m}(X,h)$ *be the series LS estimator. Let conditions 6.1, 6.2, 6.3 and 6.4 hold. Let* $\mathcal{H}^2 = \{h_2 \in L^2(\mathcal{R}^d, f_{Y_2}) : ||\left(1 + |\cdot|^2\right)^{-\vartheta/2} h_2||_{\mathcal{T}_{p,q}^{\alpha_2}} < \infty\}$ *for* $\alpha_2 > 0, p, q \in [1,\infty]$ *(and* $p < \infty$ *for*

$\mathcal{T}_{p,q}^{\alpha_2} = \mathcal{F}_{p,q}^{\alpha_2}$). *Let* $P(h_2) = ||\left(1+|\cdot|^2\right)^{-\vartheta/2} h_2||_{\mathcal{T}_{p,q}^{\alpha_2}}$. *Let* $\max\left\{[k_1(n)]^{-2r_1}, [k_2(n)]^{-2r_2}\right\} = O(\lambda_n)$ *and* $\frac{J_n}{n} + J_n^{-2r_m} = O(\lambda_n)$. *Then:*

(1) If $r_2 > 1/p$ *and* $\theta > \vartheta \geq 0$, *then:*

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

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

(2) If $r_2 + 1/2 > 1/p$, $p^{-1} + (\theta - \vartheta)/d > 1/2$, *then:*

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

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

We next present a second consistency result in which the parameter space $\mathcal{H}^2$ is not compact but the penalty is convex. It is an application of Theorem 3.3. We assume:

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

Condition 6.5 implies that the mapping $m(\cdot, h) : \mathcal{H} \to L^2(f_X)$ is compact (assumption 3.10(a)).

Proposition 6.2. *For the model (6.1), let* $\hat{h}_n$ *be the PSMD 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* $\mathcal{H}^2 = \{\left(1+|\cdot|^2\right)^{-\theta/2} h_2 \in W_2^{\alpha_2}(\mathcal{R}^d, leb) : ||\left(1+|\cdot|^2\right)^{-\theta/2} h_2||_{L^2(leb)} \leq M\}$ *for* $\alpha_2 > 0$ *and* $P(h) = ||\left(1+|\cdot|^2\right)^{-\theta/2} h_2||^2_{L^2(leb)}$. *Let* $\max\left\{[k_1(n)]^{-2r_1}, [k_2(n)]^{-2r_2}\right\} = o(\lambda_n)$ *and* $\frac{J_n}{n} + J_n^{-2r_m} = o(\lambda_n)$. *Then:*

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

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

Without unknown h_1, Proposition 6.2 is very similar to that of HL except that we allow for sieve approximation and for the support of Y_2 to be unbounded.

Given these consistency results, we now turn to the calculation of the convergence rate of our PSMD estimator. For the model (6.1), let $||h||_s^2 = E\{[h_1(Y_1) + h_2(Y_2)]^2\}$, then $||h||_s^2 \leq 2[||h_1||^2_{L^2(f_{Y_1})} + ||h_2||^2_{L^2(f_{Y_2})}]$ for all $h \in \mathcal{H}$. The above consistency results immediately imply that $||\widehat{h}_n - h_0||_s = o_P(1)$. Recall that $\mathcal{H}_{os} \equiv \{h = (h_1, h_2) \in \mathcal{H} : ||h - h_0||_s = o(1), ||h||_s \leq c, P(h) \leq c\}$. 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 $\mathcal{B}\left(Dom(T_h), L^2(f_X)\right)$ be 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)):

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

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 $h \in \mathcal{H}_{os}$.*

This condition implies that assumption 4.1 on the local curvature of $E[m(X,h)'m(X,h)]$ in terms of the weak pseudo metric $\|h - h_0\|^2$ holds for all $h \in \mathcal{H}_{os}$.

Condition 6.7. *(i) Y_1 and Y_2 are independent; (ii) there is a continuous increasing function $\varphi \geq 0$ such that $||T_{h_0}[g_1 + g_2]||^2_{L^2(f_X)} \asymp \sum_{j=1}^{\infty} \varphi(j^{-2/d})|\langle g_1 + g_2, q_{1,j} + q_{2,j}\rangle_s|^2$ for all $g_1 + g_2 \in Dom(T_{h_0}) \cap \mathcal{H}_{os}$.*

Condition 6.7 implies that assumption 5.4 (hence 5.2) holds. Applying Corollary 5.1, we obtain the following convergence rate for the PSMD estimator using slowly growing finite dimensional sieves.

Proposition 6.3. *For the model (6.1), suppose that conditions 6.6 and 6.7 hold. Let either the conditions of Proposition 6.1 hold with $\max\{\frac{J_n}{n}, J_n^{-2r_m}, \lambda_n\} = \frac{J_n}{n}$, or the conditions of Proposition 6.2 hold with $\max\{\frac{J_n}{n}, J_n^{-2r_m}, o(\lambda_n)\} = \frac{J_n}{n}$. Let $\frac{J_n}{n} = const. \times \frac{k(n)}{n} = o(1)$, $k(n) = k_1(n) + k_2(n)$ and $k_1(n) \asymp k_2(n) \to \infty$. Denote $\alpha = \min\{\alpha_1, \alpha_2\}$. Then:*

$$||\widehat{h}_n - h_0||_s = O_P\left(\max\left\{\{k(n)\}^{-\alpha/d}, \sqrt{\frac{k(n)}{n \times \varphi([k(n)]^{-2/d})}}\right\}\right) = O_P\left(\{k_o(n)\}^{-\alpha/d}\right),$$

where $k_o(n)$ is such that $\{k_o(n)\}^{-2\alpha/d} \asymp \frac{k_o(n)}{n}\{\varphi([k_o(n)]^{-2/d})\}^{-1}$; and the rate results (1) and (2) of Corollary 5.1 hold.

When Y_1 and Y_2 are measurable with respect to X, we have $\varphi([k(n)]^{-2/d}) = const.$ in Proposition 6.3. The resulting convergence rate $||\widehat{h}_n - h_0||_s = O_P\left(n^{-\frac{\alpha}{2\alpha+d}}\right)$ coincides with the known optimal rate 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).

Note that, by applying Corollary 5.3 with the series LS estimator $\widehat{m}(X,h)$, for the PSMD estimator $\widehat{h}_n$ using large or infinite dimensional sieves ($k(n)/n \to const. > 0$) with convex and/or compact penalties, we can obtain the same final convergence rates as in Proposition 6.3. See the working paper version (Chen and Pouzo, 2008a) for the precise statement.

7 Simulation and Empirical Illustration

7.1 Monte Carlo Simulation

We report a small Monte Carlo (MC) study of PSMD 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. 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 for details). Let $X = \Phi\left(\frac{\widetilde{X}-\mu_x}{\sigma_x}\right)$ 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 variances σ_x, σ_2 are the estimated ones. Second, we generate Y_1 from $Y_1 = h_0(Y_2) + U$, where $U = \sqrt{0.075}[V - \Phi^{-1}\left(\gamma + 0.01\{E[h_0(Y_2)|\widetilde{X}] - h_0(Y_2)\}\right)]$, with $V \sim N(0,1)$. The number of observations 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 a 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 PSMD estimator $\widehat{h}_n$ given in (2.2), 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 of the polynomial and r being the number of knots, so $k(n) = q(n) + r(n) + 1$.

There are three kinds of smoothing parameters in the PSMD procedure (2.2): one ($k(n)$) for the sieve approximation $\mathcal{H}_n$, one (λ_n) for the penalization, and one (J_n) for the nonparametric LS estimator of $\widehat{m}(X,h)$. In the previous theoretical sections, we showed that we could obtain the 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). In this MC study, we compare the finite sample performance of these two cases.[18]

Figure 7.1 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 computed PSMD estimators using Hermite polynomial sieves for $\mathcal{H}_n$, Fourier basis, B-spline basis, Hermite basis for $\hat{m}$, and $\widehat{P}_n(h) = ||\nabla^j h||_{L^1(leb)}$ or $||\nabla^j h||_{L^1(d\widehat{\mu})}$ for $j = 1$ or 2. As long as the choices 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 7.1, each panel shows the true function (solid thick line), the corresponding estimator (solid thin line, which is the pointwise average over the 500 MC simulation), the Monte Carlo 95% confidence bands (dashed), and a sample realization of Y_1 (that is arbitrarily picked from the last MC iteration). Both estimators perform very well for all

[18] In the working paper version (Chen and Pouzo, 2008a) we analyzed a third case: 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 $k(n)/n \to const. > 0$). The MC implementations of this case are too time consuming to report.

of the quantiles, with the "sieve dominating case" ($k(n) = 8$) 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 (with 500 repetitions) written in FORTRAN took (approximately) 50 minutes to finish for the "sieve dominating case", whereas it took (approximately) 240 minutes to finish for the "sieve penalization balance case".

Table 7.1 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%. Here for simplicity we have only reported the estimated quantile with $\gamma = 0.5$ and 250 MC replications. Figure 7.2 shows the corresponding estimated curves and MC 95% confidence bands. In Table 7.1, the rows with $k(n) = 6$, 8 belong to the "sieve dominating case"; the rows with $k(n) = 16$ belong to the "sieve penalization balance case". For this MC study, the "sieve dominating case" ($k(n) = 6$, 8) perform well in terms of $I - BIAS^2$ and $I - VAR$ (hence $I - MSE$), and are much more economical in terms of computational time. Within the "sieve penalization balance case"($k(n) = 16$), given the same λ_n the ones with derivative penalty perform slightly better than the one with function level penalty.

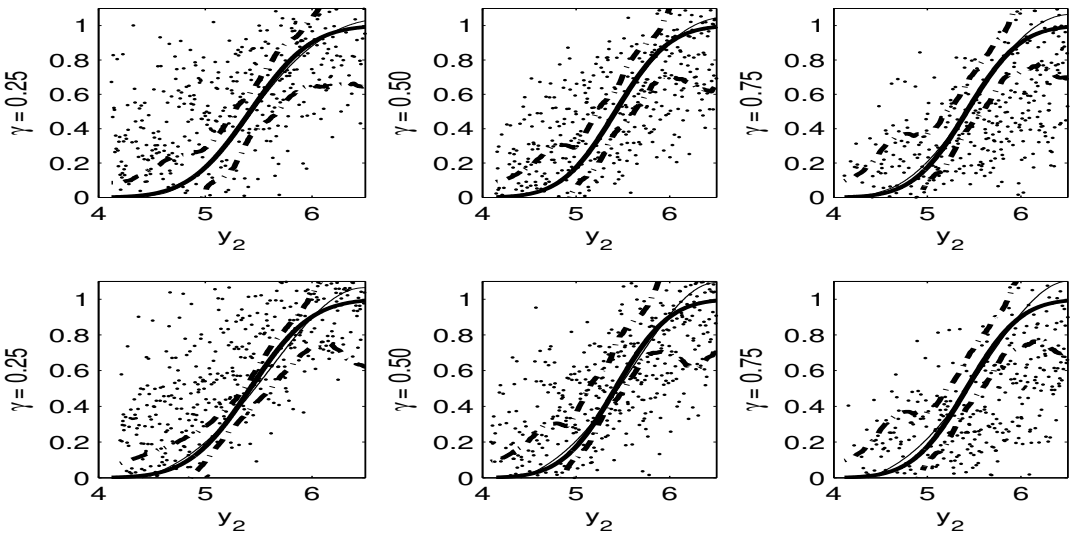

Figure 7.1: h_0 (solid thick), $\hat{h}_n$ (solid thin), MC confidence bands (dashed), a sample of Y_1 (dots), $\hat{P}(h) = ||\nabla h||^2_{L^2}$, 1st row: $k(n) = 8, \lambda_n = 0.003, J_n = 16$; 2nd row: $k(n) = 16, \lambda_n = 0.006, J_n = 16$.

Table 7.1: Simulation Results for $\gamma = 0.5$ quantile IV curve, 250 MC runs

$(k(n), J_n)$	$I-BIAS^2$	$I-VAR$	$I-MSE$	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
$(16,16)$	0.00378	0.02141	0.02520	$\|\|\nabla^2\cdot\|\|_{L^1}$	0.00050	84
$(16,16)$	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

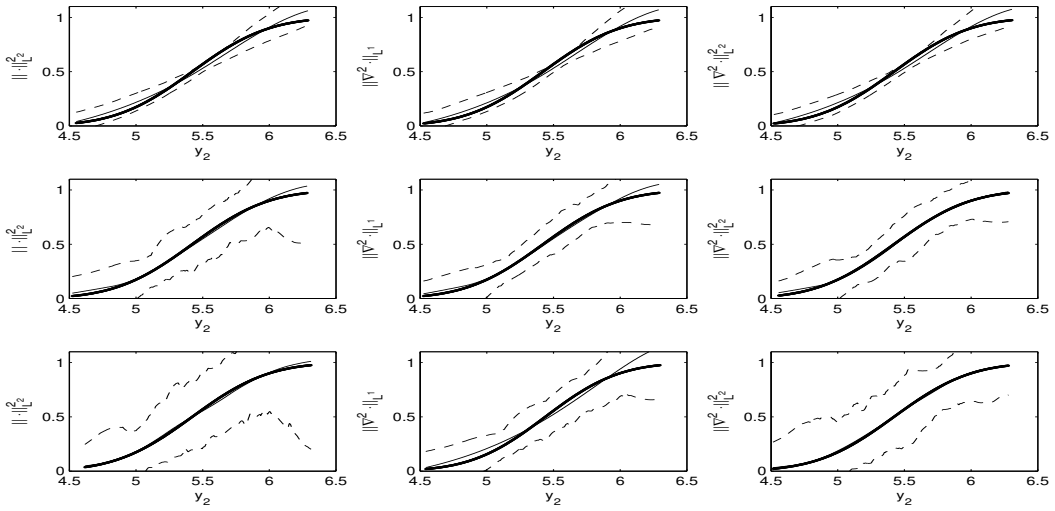

Figure 7.2: Table 7.1 experiments. 1st row: $k(n) = 6, \lambda_n = 0.00001, J_n = 16$. 2nd row: $k(n) = 8, \lambda_n = 0.0001, J_n = 16$. 3nd row: $k(n) = 16, \lambda_n = 0.0005, J_n = 16$.

7.2 Empirical Illustration

We apply the PSMD procedure to nonparametric quantile IV estimation of Engel curves using the UK Family Expenditure Survey data. The model is

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

where $Y_{1i\ell}$ 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, which 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 for the NPIV model (1.2).

As an illustration, we apply the PSMD procedure 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 also computed PSMD estimators with $||\nabla^k h||^j_{L^j(d\widehat{\mu})} \equiv n^{-1}\sum_{i=1}^n |\nabla^k h(Y_{2i})|^j$ for $k = 1, 2$ and $j = 1, 2$, and Hermite polynomial sieves, cosine sieves, 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 7.3 we report the estimated Engel curves only for three different quantiles $\gamma = \{0.25, 0.50, 0.75\}$ and for four selected goods, using P-Spline(2,5) as $\mathcal{H}_n$ and P-Spline(5,10) for $\hat{m}$ (hence $k(n) = 8$, $J_n = 16$). Figure 7.3 presents the estimated Engel curves using $\widehat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\widehat{\mu})}$ with $\lambda_n = 0.001$ and $\widehat{P}_n(h) = ||\nabla^2 h||_{L^1(d\widehat{\mu})}$ with $\lambda_n = 0.001$ in the first and second rows; $\widehat{P}_n(h) = ||\nabla h||^2_{L^2(d\widehat{\mu})}$ 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 PSMD estimator in the "sieve dominating case".

8 Conclusion

In this paper, we proposed the PSMD estimation of conditional moment models containing unknown functions of endogenous variables: $E[\rho(Y, X_z; h_0(\cdot))|X] = 0$. The estimation problem is a difficult nonlinear ill-posed inverse problem with an unknown operator. We established the consistency and the convergence rate of the PSMD estimator of $h_0(\cdot)$, allowing for (i) a possibly non-compact infinite dimensional function 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 and/or convex penalty, or the SMD estimator with slowly growing finite dimensional linear sieves without a penalty; and (v) mildly or severely ill-posed inverse problems. Under relatively low-level sufficient conditions, we showed that the convergence rate under a Hilbert space norm coincide with the known minimax optimal rate for the NPIV model (1.2). We illustrated

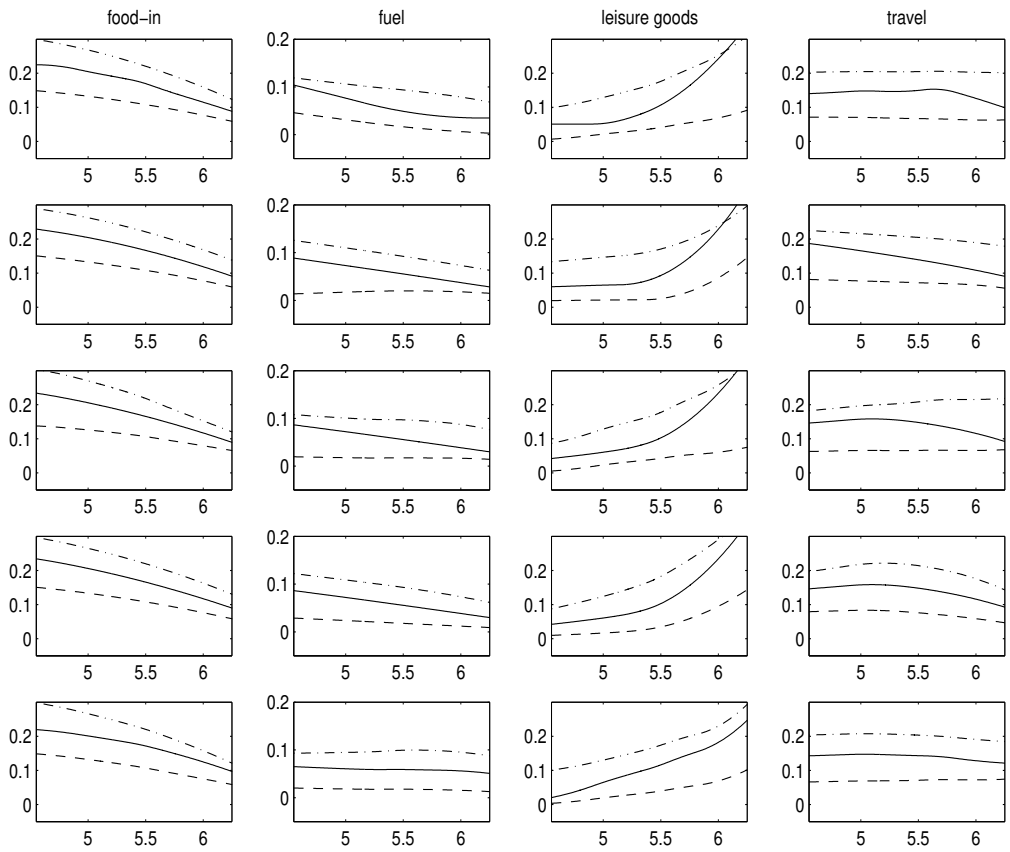

Figure 7.3: Engel curves for quantiles $\gamma = 0.25$ (dash), 0.50 (solid), 0.75 (dot-dash). $k(n) = 8, J_n = 16$ (all rows). $\hat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\hat{\mu})}$ with $\lambda_n = 0.001$ (1st row); $\hat{P}_n(h) = ||\nabla^2 h||_{L^1(d\hat{\mu})}$ with $\lambda_n = 0.001$ (2nd row); $\hat{P}_n(h) = ||\nabla h||^2_{L^2(d\hat{\mu})}$ 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).

the general theory with a nonparametric additive quantile IV regression. We also presented a simulation study and estimated a system of nonparametric quantile IV Engel curves using the UK Family Expenditure Survey. These results indicate that PSMD estimators using slowly growing finite dimensional sieves with small penalization parameter are easy to compute and perform well in finite samples.

In Chen and Pouzo (2008b), which considers the general semi/nonparametric conditional moment restrictions $E[\rho(Y, X_z; \theta_0, h_0(\cdot))|X] = 0$ when $\rho(Y, X_z, \theta, h(\cdot))$ may not be pointwise smooth in (θ, h), we show that the PSMD estimator using slowly growing finite dimensional sieves can simultaneously achieve the root-n asymptotic normality of $\widehat{\theta}_n - \theta_0$ and the nonparametric minimax optimal rate of convergence for $\widehat{h}_n - h_0$.

A A Brief Summary of Function Spaces and Sieves

Here we briefly summarize some definitions and properties of function spaces that are used in the main text; see Edmunds and Triebel (1996) for details. 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* $\mathcal{B}^\nu_{p,q}(\mathcal{R}^d)$ is the collection of all functions $h \in \mathcal{S}^*(\mathcal{R}^d)$ such that $\|h\|_{\mathcal{B}^\nu_{p,q}}$ is finite:

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

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

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

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

Let $\mathcal{T}^\nu_{p,q}(\mathcal{R}^d)$ denote either $\mathcal{B}^\nu_{p,q}(\mathcal{R}^d)$ or $\mathcal{F}^\nu_{p,q}(\mathcal{R}^d)$. Then $\mathcal{T}^\nu_{p,q}(\mathcal{R}^d)$ gets larger with increasing q (i.e., $\mathcal{T}^\nu_{p,q_1}(\mathcal{R}^d) \subseteq \mathcal{T}^\nu_{p,q_2}(\mathcal{R}^d)$ for $q_1 \leq q_2$), gets larger with decreasing p (i.e., $\mathcal{T}^\nu_{p_1,q}(\mathcal{R}^d) \subseteq \mathcal{T}^\nu_{p_2,q}(\mathcal{R}^d)$ for $p_1 \geq p_2$), and gets larger with decreasing ν (i.e., $\mathcal{T}^{\nu_1}_{p,q}(\mathcal{R}^d) \subseteq \mathcal{T}^{\nu_2}_{p,q}(\mathcal{R}^d)$ for $\nu_1 \geq \nu_2$). Also, $\mathcal{T}^\nu_{p,q}(\mathcal{R}^d)$ becomes a *Banach* space when $p, q \geq 1$. The spaces $\mathcal{T}^\nu_{p,q}(\mathcal{R}^d)$ include many well-known function spaces as special cases. For example, $L^p(\mathcal{R}^d, leb) = \mathcal{F}^0_{p,2}(\mathcal{R}^d)$ for $p \in (1, \infty)$; the *Hölder* space $\Lambda^r(\mathcal{R}^d) = \mathcal{B}^r_{\infty,\infty}(\mathcal{R}^d)$ for any real-valued $r > 0$; the *Hilbert-Sobolev* space $W^k_2(\mathcal{R}^d) =$

$\mathcal{B}^k_{2,2}\left(\mathcal{R}^d\right)$ for integer $k > 0$; and the *(fractional) Sobolev* space $W^\nu_p(\mathcal{R}^d) = \mathcal{F}^\nu_{p,2}\left(\mathcal{R}^d\right)$ for any $\nu \in \mathcal{R}$ and $p \in (1,\infty)$, which has the equivalent norm $||h||_{W^\nu_p} \equiv \left\| \left((1+|\cdot|^2)^{\nu/2} \widehat{h}(\cdot) \right)^\vee \right\|_{L^p(leb)} < \infty$ (note that for $\nu > 0$, the norm $||h||_{W_p^{-\nu}}$ is a shrinkage in the Fourier domain).

Let $\mathcal{T}^\nu_{p,q}(\Omega)$ be the corresponding space on an (arbitrary) bounded domain Ω in $\mathcal{R}^d$. Then the embedding of $\mathcal{T}^{\nu_1}_{p_1,q_1}(\Omega)$ into $\mathcal{T}^{\nu_2}_{p_2,q_2}(\Omega)$ is compact if $\nu_1 - \nu_2 > d\max\left\{p_1^{-1} - p_2^{-1}, 0\right\}$, and $-\infty < \nu_2 < \nu_1 < \infty$, $0 < q_1, q_2 \le \infty$, $0 < p_1, p_2 \le \infty$ ($0 < p_1, p_2 < \infty$ for $\mathcal{F}^\nu_{p,q}(\Omega)$).

We define "weighted" versions of the space $\mathcal{T}^\nu_{p,q}(\mathcal{R}^d)$ as follows. Let $w(\cdot) = (1+|\cdot|^2)^{\zeta/2}$, $\zeta \in \mathcal{R}$ be a weight function and define $||h||_{\mathcal{T}^\nu_{p,q}(\mathcal{R}^d,w)} = ||wh||_{\mathcal{T}^\nu_{p,q}(\mathcal{R}^d)}$, that is, $\mathcal{T}^\nu_{p,q}(\mathcal{R}^d, w) = \{h : ||wh||_{\mathcal{T}^\nu_{p,q}(\mathcal{R}^d)} < \infty\}$. Then the embedding of $\mathcal{T}^{\nu_1}_{p_1,q_1}\left(\mathcal{R}^d, w_1\right)$ into $\mathcal{T}^{\nu_2}_{p_2,q_2}\left(\mathcal{R}^d, w_2\right)$ is compact if and only if $\nu_1 - \nu_2 > d(p_1^{-1} - p_2^{-1})$, $w_2(x)/w_1(x) \to 0$ as $|x| \to \infty$, and $-\infty < \nu_2 < \nu_1 < \infty$, $0 < q_1, q_2 \le \infty$, $0 < p_1 \le p_2 \le \infty$ ($0 < p_1 \le p_2 < \infty$ for $\mathcal{F}^\nu_{p,q}(\Omega)$).

If $\mathcal{H} \subseteq \mathbf{H}$ 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 5. 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 $0 \le k \le \gamma$; (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

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

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 of sieves:

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 $k_0 \ge 0$, 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 $\{\varphi_{k_0 j}, j \in \mathbb{Z}; \psi_{kj}, k \ge k_0, j \in \mathbb{Z}\}$ 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:

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

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,

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

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 $\{\nu_k\}_{k=1,\ldots,r_n}$ are the knots. In the empirical application, for any given number of knots value r_n, the knots $\{\nu_k\}_{k=1,\ldots,r_n}$ 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 all approximate penalized sieve extremum estimation problems, be they 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:

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

and $\liminf_{n\to\infty} g_0(n,k(n),\varepsilon) \geq 0$ for all $\varepsilon > 0$.

(B.1.2) (i) $\mathcal{A} \subseteq \mathbf{A}$ and $(\mathbf{A}, ||\cdot||_s)$ 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 $||\Pi_n\alpha_0 - \alpha_0||_s \to 0$ 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 $\hat{c}^Q(k(n)) \equiv \sup_{\alpha\in\mathcal{A}_{k(n)}}\left|\widehat{Q}_n(\alpha) - \overline{Q}_n(\alpha)\right| = o_P(1)$.

$$\frac{\max\left\{\hat{c}^Q(k(n)), \eta_n, \left|\overline{Q}_n(\Pi_n\alpha_0) - \overline{Q}_n(\alpha_0)\right|\right\}}{g_0(n,k(n),\varepsilon)} = o(1) \quad \text{for all } \varepsilon > 0.$$

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

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

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

which goes to 0 by condition (B.1.4). *Q.E.D.*

We recall some standard definitions. A sequence $\{\alpha_j\}_{j=1}^{\infty}$ in a Banach space $(\mathbf{A}, ||\cdot||_s)$ converges *weakly* to α if and only if (iff) $\lim_{j\to\infty}\langle v, \alpha_j - \alpha_0\rangle_{\mathbf{A}^*,\mathbf{A}} = \langle v, \alpha - \alpha_0\rangle_{\mathbf{A}^*,\mathbf{A}}$ for all $v \in \mathbf{A}^*$. A set $\mathcal{A} \subseteq \mathbf{A}$ is *weak sequentially compact* iff each sequence in $\mathcal{A}$ possesses a weakly convergent subsequence with limit value in $\mathcal{A}$. A set $\mathcal{A} \subseteq \mathbf{A}$ is *weak sequentially closed* iff each weakly convergent sequence in $\mathcal{A}$ has its limit value in $\mathcal{A}$. A functional $F : \mathcal{A} \subseteq \mathbf{A} \to [-\infty, +\infty]$ is said to be *weak sequentially lower semicontinuous* at $\alpha \in \mathcal{A}$ iff $F(\alpha) \leq \liminf_{j\to\infty} F(\alpha_j)$ for each sequence $\{\alpha_j\}$ in $\mathcal{A}$ that converges weakly to α.

Remark B.1. ***(1)*** *Let $(\mathbf{A}, \mathcal{T})$ be a topological space and $\mathcal{A}_k$ be non-empty for each k. 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 level set $\{\alpha \in \mathcal{A}_k : \widehat{Q}_n(\alpha) \leq r\}$ is compact in $(\mathbf{A}, \mathcal{T})$ for all $r \in (-\infty, +\infty)$. See Zeidler (1985, theorem 38.B).*

(2) *Let $(\mathbf{A}, ||\cdot||_s)$ be a Banach space and $\mathcal{A}_k$ be non-empty for each k. Condition (B.1.3) is satisfied if one of the following three conditions holds: (a) $\mathcal{A}_k$ is a weak sequentially compact subset of $(\mathbf{A}, ||\cdot||_s)$, and for any data $\{Z_i\}_{i=1}^n$, $\widehat{Q}_n(\alpha)$ is weak sequentially lower semicontinuous on $\mathcal{A}_{k(n)}$. (b) $\mathcal{A}_k$ is a bounded, and weak sequentially closed subset of a reflexive Banach space $(\mathbf{A}, ||\cdot||_s)$, 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 $(\mathbf{A}, ||\cdot||_s)$, 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 38.12, theorem 38.A, corollary 38.8).*

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 2.2 and 2.3(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 $n \to \infty$,

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

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 2.2 and the definition of G_n, we have $A_n \asymp \xi_n$. Thus, by assumption 2.2(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 2.2 and 2.3(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 2.3(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:

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

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

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 2.3(i), and the last inequality follows from assumption 2.2(ii).

Given Result (1), assumption 2.2(ii) and 2.3(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.*

The next condition assumes the existence of the PSMD estimator; see Remark B.1 for general sufficient conditions. In the main text we presented low level sufficient conditions for assumption B.1.

Assumption B.1. *$\widehat{h}_n \in \mathcal{H}_{k(n)}$ is well-defined with probability approaching one.*

Lemma B.4. *Let $\hat{h}_n$ be the PSMD 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 2.1. Let assumptions 3.1(i)(ii), 3.2(i) and B.1 hold. Then: (1) under assumption 3.3, for all $\varepsilon > 0$,*

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

(2) under assumption 3.7, for all $\varepsilon > 0$,

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

Proof of Lemma B.4: By definition of $\widehat{h}_n$ and $\Pi_n h_0$ and assumptions 3.1(i)(ii), 3.2(i) and B.1, we have: for any $\varepsilon > 0$,

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

By the i.i.d. sample, and assumption 2.1(ii) for any consistent estimator $\widehat{m}$ (or assumptions 2.2 - 2.3(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:

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

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 2.1(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 2.1(i) for any consistent estimator $\widehat{m}$, 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 3.3, 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 $\varepsilon>0$,

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

By assumption 3.7, 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 $\varepsilon>0$,

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

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

Lemma B.5. *Let $\hat{h}_n$ be the PSMD estimator satisfying assumption B.1 with $\lambda_n>0$, $\lambda_n=o_P(1)$, and let $\widehat{m}(X,h)$ be any consistent estimator of $m(X,h)$ satisfying assumption 2.1 at $h=\Pi_n h_0$.*

(1) Under assumption 3.3(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 3.7 and $\max\{\delta_{m,n}^2, E[||m(X,\Pi_n h_0)||_E^2]\}=o(\lambda_n)$, $P(\widehat{h}_n)\leq P(h_0)+o_P(1)$.

Proof of Lemma B.5: 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

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

Thus

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

where the last inequality is due to assumption 2.1 for $h = \Pi_n h_0$. Therefore, for all $M > 0$,

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

(1) Under assumption 3.3, $\lambda_n \sup_{h\in\mathcal{H}_n}\left|\widehat{P}_n(h) - P(h)\right| + \lambda_n\left|P(\Pi_n h_0) - P(h_0)\right| = O_P(\lambda_n)$, 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 3.7, $\lambda_n \sup_{h\in\mathcal{H}_n}\left|\widehat{P}_n(h) - P(h)\right| + \lambda_n\left|P(\Pi_n h_0) - P(h_0)\right| = o_P(\lambda_n)$, 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 3.1: It is clear that assumptions 2.1, 3.3 and 3.4 imply assumption B.1. Under assumptions 3.1(iii) and 3.4, $g(k,\varepsilon) \equiv \min_{h\in\mathcal{H}_k:||h-h_0||_s\geq\varepsilon} E\left[m(X,h)'m(X,h)\right]$ exists and is positive for each $k \geq 1, \varepsilon > 0$. By Lemma B.4(1) and $\lambda_n P(h) \geq 0$, we have: for all $\varepsilon > 0$,

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

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

Proof of Theorem 3.2: Denote $\Delta_n \equiv K\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)] + \delta_{m,n}^2 + \lambda_n P(h_0) + O(\lambda_n)\}$. By Lemma B.4(1), for all $\varepsilon > 0$,

$$\Pr\left(||\widehat{h}_n - h_0||_s > \varepsilon\right) \leq \Pr\left(\inf_{h\in\mathcal{H}_n:||h-h_0||_s\geq\varepsilon}\left\{E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h)\right\} \leq \Delta_n\right).$$

We divide $\mathcal{H}_{k(n)}(\varepsilon) \equiv \{\mathcal{H}_n : ||h - h_0||_s \geq \varepsilon\}$ into two disjoint sets: $\mathcal{H}_{k(n)}^+(\varepsilon) \equiv \{h \in \mathcal{H}_{k(n)}(\varepsilon) : P(h) \leq \lambda_n^{-1}\Delta_n + M\}$ for any $M > 0$, and $\mathcal{H}_{k(n)}^-(\varepsilon) \equiv \mathcal{H}_{k(n)}(\varepsilon) \setminus \mathcal{H}_{k(n)}^+(\varepsilon)$. Note that

$$\inf_{h\in\mathcal{H}_{k(n)}^-(\varepsilon)}\left\{E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h)\right\} \geq \Delta_n + \lambda_n M > \Delta_n$$

Thus $\Pr\left(\inf_{h\in\mathcal{H}_{k(n)}^-(\varepsilon)}\{E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h)\} \leq \Delta_n\right) = 0$; hence

$$\begin{aligned} \Pr\left(||\widehat{h}_n - h_0||_s > \varepsilon\right) \; & \leq \; \Pr\left(\inf_{h\in\mathcal{H}_{k(n)}^+(\varepsilon)}\left\{E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h)\right\} \leq \Delta_n\right) \\ & \leq \; \Pr\left(\inf_{h\in\mathcal{H}^+(\varepsilon)}\left\{E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h)\right\} \leq \Delta_n\right), \end{aligned}$$

where $\mathcal{H}^+(\varepsilon) \equiv \left\{h \in \mathcal{H} : ||h - h_0||_s \geq \varepsilon, P(h) \leq \lambda_n^{-1}\Delta_n + M\right\}$.

Given that assumption 3.5(i), the fact that $\{h \in \mathcal{H} : ||h - h_0||_s \geq \varepsilon\}$ is closed, and that $\max\{\delta_{m,n}^2, E(\|m(X, \Pi_n h_0)\|_E^2)\} = O(\lambda_n)$, we have that the set $\mathcal{H}^+(\varepsilon)$ is compact under $||\cdot||_s$. Moreover, $E[m(X,h)'m(X,h)]$ is lower semicontinuous on $\mathcal{H}$ under $||\cdot||_s$ (assumption 3.6(ii)). Theorem 38.B of Zeidler (1985) now implies that the minimization problem,

$$\min_{\mathcal{H}_{k(n)}^+(\varepsilon)} \left\{E\left[m(X,h)'m(X,h)\right] + \lambda_n P(h)\right\}$$

has a solution, h_n, which belongs to the set $\mathcal{H}^+(\varepsilon)$. Therefore, the sequence $\{h_n\}$ must have a further subsequence, denoted as $\{h_{n_k}\}$, that converges to a limit h_∞ in $||\cdot||_s$ and $h_\infty \in \{h \in \mathcal{H} : ||h - h_0||_s \geq \varepsilon, P(h) \leq \overline{M}\}$ for some $\overline{M} \in [0, +\infty)$. By assumption 3.6(ii) and $P(h) \geq 0$, we have:

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

This and assumption 3.1(iii) together imply that $||h_\infty - h_0||_s = 0$, which contradicts $h_\infty \in \{h \in \mathcal{H} : ||h - h_0||_s \geq \varepsilon, P(h) \leq \overline{M}\}$. Thus $||\widehat{h}_n - h_0||_s = o_P(1)$. Lemma B.5 (1) implies $P(\widehat{h}_n) = O_P(1)$. *Q.E.D.*

Denote $A_n(\varepsilon) \equiv \inf_{h \in \mathcal{H}_{k(n)}: ||h-h_0||_s \geq \varepsilon} \{E[m(X,h)'m(X,h)] + \lambda_n \langle t_0, h - h_0\rangle_{\mathbf{H}^*,\mathbf{H}}\}$.

Assumption B.2. $\liminf_{n\to\infty} \frac{A_n(\varepsilon)}{\lambda_n} \geq 0$ *for all* $\varepsilon > 0$.

Lemma B.6. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n > 0$, $\lambda_n = o(1)$, and $\widehat{m}(X,h)$ any consistent estimator of $m(X,h)$ satisfying assumption 2.1. Let assumptions 3.1(i)(ii), 3.2, B.1, 3.7, 3.8, and B.2 hold. If $\max\{\delta_{m,n}^2, E(\|m(X,\Pi_n h_0)\|_E^2)\} = o(\lambda_n)$, Then: $||\widehat{h}_n - h_0||_s = o_P(1)$, and $P(\widehat{h}_n) = P(h_0) + o_P(1)$.*

Proof of Lemma B.6: By Lemma B.4(2), we have: for all $\varepsilon > 0$,

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

By assumption 3.8

$$\begin{aligned} &\inf_{h \in \mathcal{H}_{k(n)}: ||h-h_0||_s \geq \varepsilon} \left\{E\left[m(X,h)'m(X,h)\right] + \lambda_n(P(h) - P(h_0))\right\} \\ \geq\ &\inf_{h \in \mathcal{H}_{k(n)}: ||h-h_0||_s \geq \varepsilon} \left\{E\left[m(X,h)'m(X,h)\right] + \lambda_n\langle t_0, h - h_0\rangle_{\mathbf{H}^*,\mathbf{H}} + \lambda_n g(||h - h_0||_s)\right\} \\ \geq\ &\inf_{h \in \mathcal{H}_{k(n)}: ||h-h_0||_s \geq \varepsilon} \left\{E\left[m(X,h)'m(X,h)\right] + \lambda_n\langle t_0, h - h_0\rangle_{\mathbf{H}^*,\mathbf{H}}\right\} + \lambda_n g(\varepsilon) \equiv A_n(\varepsilon) + \lambda_n g(\varepsilon), \end{aligned}$$

thus

$$\begin{aligned} &\Pr\left(||\widehat{h}_n - h_0||_s > \varepsilon\right) \\ \leq\ &\Pr\left(A_n(\varepsilon) + \lambda_n g(\varepsilon) \leq K\left\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)] + O_P\left(\delta_{m,n}^2\right) + o_P(\lambda_n)\right\}\right) \\ \leq\ &\Pr\left(\frac{A_n(\varepsilon)}{\lambda_n} + g(\varepsilon) \leq \frac{K\left\{E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)] + O_P\left(\delta_{m,n}^2\right) + o_P(\lambda_n)\right\}}{\lambda_n}\right) \\ =\ &\Pr\left(\frac{A_n(\varepsilon)}{\lambda_n} + g(\varepsilon) \leq o_P(1)\right) \end{aligned}$$

since $\lambda_n > 0$ and $\lambda_n^{-1} \max\left\{\delta_{m,n}^2, E(\|m(X, \Pi_n h_0)\|_E^2), o(\lambda_n)\right\} = o(1)$. Thus, to obtain the desired result it suffices to show that for any $\varepsilon > 0$, $\frac{A_n(\varepsilon)}{\lambda_n} + g(\varepsilon) \geq c(\varepsilon) > 0$, for all n large enough. By assumption B.2, for any $\delta = \delta(\varepsilon) > 0$, it follows $\frac{A_n(\varepsilon)}{\lambda_n} \geq -\delta$ for $n \geq n(\delta)$. In particular, let $\delta = -0.5g(\varepsilon)$ we obtain: $\frac{A_n(\varepsilon)}{\lambda_n} + g(\varepsilon) \geq 0.5g(\varepsilon) > 0$ for all n large enough. Thus $||\widehat{h}_n - h_0||_s = o_P(1)$. This and assumption 3.8 imply $P(\widehat{h}_n) - P(h_0) \geq o_P(1)$. But Lemma B.5 (2) also implies $P(\widehat{h}_n) - P(h_0) \leq o_P(1)$. Thus $P(\widehat{h}_n) - P(h_0) = o_P(1)$ under $\max\left\{\delta_{m,n}^2, E(\|m(X, \Pi_n h_0)\|_E^2)\right\} = o(\lambda_n)$. *Q.E.D.*

Assumption B.3. *$E\left[m(X,h)'m(X,h)\right]$ is weak sequentially lower semicontinuous on $\mathcal{H}$.*

Theorem B.1. *Assumptions 3.1(iii), 3.2(i), 3.9, B.3 imply assumption B.2.*

PROOF OF THEOREM B.1: First note that

$$A_n(\varepsilon) \geq \inf_{h \in \mathcal{H}} \left\{E[m(X,h)'m(X,h)] + \lambda_n \langle t_0, h - h_0 \rangle_{\mathbf{H}^*,\mathbf{H}}\right\}.$$

Under assumption 3.9, $\mathcal{H}$ is convex, closed and bounded and thus it is weak sequentially compact. By assumption B.3, $E[m(X,h)'m(X,h)]$ is weak sequentially lower semi-continuous. Therefore, for any $\lambda_n \geq 0$, the minimization problem, $\inf_{h \in \mathcal{H}} \{E[m(X,h)'m(X,h)] + \lambda_n \langle t_0, h - h_0 \rangle_{\mathbf{H}^*,\mathbf{H}}\}$, has a solution, $h_n \equiv h^*(\lambda_n)$, which belongs to $\mathcal{H}$ (see Zeidler, 1985, corollary 38.8). Thus

$$A_n(\varepsilon) \geq E[m(X,h_n)'m(X,h_n)] + \lambda_n \langle t_0, h_n - h_0 \rangle_{\mathbf{H}^*,\mathbf{H}} \quad \text{for all } \varepsilon > 0.$$

Since $\mathcal{H}$ is weakly compact, the sequence $\{h_n\}$ has a sub-sequence that converges (weakly) to $h_\infty \in \mathcal{H}$. (To simplify notation we still use $\{h_n\}$ to denote the weakly convergent subsequence.) By assumption 3.9(iii), $\langle t_0, h - h_0 \rangle_{\mathbf{H}^*,\mathbf{H}} = O(1)$ uniformly in $h \in \mathcal{H}$. If $\liminf_{n\to\infty} E\left[m(X,h_n)'m(X,h_n)\right] \equiv c > 0$, we have $\liminf_{n\to\infty} \frac{A_n(\varepsilon)}{\lambda_n} = \infty > 0$ and assumption B.2 holds. So we focus on the case where $c = 0$. By assumption B.3 and $\liminf_{n\to\infty} E\left[m(X,h_n)'m(X,h_n)\right] \equiv c = 0$, we have $E\left[m(X,h_\infty)'m(X,h_\infty)\right] = 0$. This and assumption 3.1(iii) imply $h_\infty = h_0$. Therefore,

$$\liminf_n \langle t_0, h_n - h_0 \rangle_{\mathbf{H}^*,\mathbf{H}} = \lim_n \langle t_0, h_n - h_0 \rangle_{\mathbf{H}^*,\mathbf{H}} = \lim_n \langle t_0, h_n - h_\infty \rangle_{\mathbf{H}^*,\mathbf{H}} = 0;$$

hence assumption B.2 holds with $\liminf_n \frac{A_n(\varepsilon)}{\lambda_n} = 0$. *Q.E.D.*

PROOF OF THEOREM 3.3: First, we need to show that assumptions 3.9 and 3.10 imply assumption B.3. Under assumptions 3.9 and 3.10(a), 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 $E\left[m(X,h)'m(X,h)\right] : m \in L^2(f_X) \to [0, +\infty]$ is convex and continuous in $m \in L^2(f_X)$, it follows that $E\left[m(X,h_k)'m(X,h_k)\right] \to E\left[m(X,h_\infty)'m(X,h_\infty)\right]$ as $k \to \infty$; hence assumption B.3 holds. By Remark B.1(2)(c), assumptions 3.9 and 3.10(b) imply that assumption B.3 holds.

Next, under assumptions 3.1(i), 3.2(i) and 2.1, by theorem 38.A and corollary 38.8 of Zeidler (1985), we have that assumptions 3.9, B.3 and 3.11 imply assumption B.1.

Therefore, the desired result follows directly from lemma B.6 and theorem B.1. *Q.E.D.*

PROOF OF THEOREM 3.4: For result (1), we first show that the set of minimum penalization solution, $\mathcal{M}_0^P$, is not empty. Since $m(X,h)$ is convex and lower semicontinuous (assumption 3.10(b)) and $\mathcal{H}$ is a convex, closed and bounded subset of a reflexive Banach space (assumption

3.9), by proposition 38.15 of Zeidler (1985), $\mathcal{M}_0$ is convex, closed and bounded (and non-empty by assumption 3.12(i)). Since $P(.)$ is convex and lower semi-continuous on $\mathcal{M}_0$ (assumption 3.12(ii)), applying proposition 38.15 of Zeidler (1985), we have that the set $\mathcal{M}_0^P$ is non-empty, convex, closed and bounded subset of $\mathcal{M}_0$. Next, we show uniqueness of the minimum penalization solution. Suppose that there exist $h_1, h_0 \in \mathcal{M}_0^P$ such that $||h_1 - h_0||_s > 0$. Since $\mathcal{M}_0^P$ is a subset of $\mathcal{M}_0$, and $\mathcal{M}_0$ is convex, $h' = \lambda h_1 + (1-\lambda)h_0 \in \mathcal{M}_0$. By assumption 3.12(ii), $P(.)$ is strictly convex on $\mathcal{M}_0$ (in $||\cdot||_s$), thus $P(h') < P(h_0)$, but this is a contradiction since h_0 is a minimum penalization solution. Thus we established result (1).

For result (2), first, as already shown earlier, the PSMD estimator $\widehat{h}_n$ is well-defined. We now show its consistency under the weak topology. Let $\mathcal{B}_w(h_0)$ denote any open ball (under the weak topology) centered at h_0, and $\mathcal{B}_w^c(h_0)$ denote its complement (under the weak topology) in $\mathcal{H}$. Following similar calculations to the ones in the proof of lemma B.4(2) (for which all assumptions hold), we have:

$$\Pr\left(\widehat{h}_n \in \mathcal{B}_w^c(h_0)\right) \leq \Pr\left(\inf_{\mathcal{H}_{k(n)} \cap \mathcal{B}_w^c(h_0)} \left\{E[m(X,h)'m(X,h)] + \lambda_n P(h)\right\} \leq O(\Delta_n) + \lambda_n P(h_0)\right)$$

where $\Delta_n \equiv E[m(X,\Pi_n h_0)'m(X,\Pi_n h_0)] + \delta_{m,n}^2 + o(\lambda_n) = o(\lambda_n)$. By assumptions 3.9(ii)(iii) and 3.11(b), $\mathcal{H}_{k(n)}$ is weakly sequentially compact. Since $\mathcal{B}_w^c(h_0)$ is closed under the weak topology, the set $\mathcal{H}_{k(n)} \cap \mathcal{B}_w^c(h_0)$ is weakly sequentially compact. By assumptions 3.10(b) and 3.11(b), $E[m(X,h)'m(X,h)] + \lambda_n P(h)$ is weakly sequentially lower semicontinuous. Thus $g(k(n), \varepsilon, \lambda_n) \equiv \inf_{\mathcal{H}_{k(n)} \cap \mathcal{B}_w^c(h_0)} \{E[m(X,h)'m(X,h)] + \lambda_n P(h)\} \geq 0$ exists, and we denote its minimizer as $h_n(\varepsilon) \in \mathcal{H}_{k(n)} \cap \mathcal{B}_w^c(h_0)$. If $\liminf_n E[m(X,h_n(\varepsilon))'m(X,h_n(\varepsilon))] = const. > 0$ then $\Pr\left(\widehat{h}_n \in \mathcal{B}_w^c(h_0)\right) \to 0$ trivially (since $\lambda_n = o(1)$). So we assume $\liminf_n E[m(X,h_n(\varepsilon))'m(X,h_n(\varepsilon))] = const. = 0$. Since we have weakly compactness of $\mathcal{H} \cap \mathcal{B}_w^c(h_0)$ there exists a subsequence $\{h_{n_k}(\varepsilon)\}_k$ that converges (weakly) to $h_\infty(\varepsilon) \in \mathcal{H} \cap \mathcal{B}_w^c(h_0)$. By weakly lower semicontinuity, $h_\infty(\varepsilon) \in \mathcal{M}_0$. By definition of h_0 it must be that $P(h_\infty(\varepsilon)) \geq P(h_0)$, moreover if this holds with equality, then $||h_\infty(\varepsilon) - h_0||_s = 0$ by result (1). Since for $t \in \mathbf{H}^*$, $|\langle t, h_\infty(\varepsilon) - h_0\rangle_{\mathbf{H}^*,\mathbf{H}}| \leq const. \times ||h_\infty(\varepsilon) - h_0||_s = 0$, then it cannot be that $h_\infty(\varepsilon) \in \mathcal{B}_w^c(h_0)$. Therefore, $P(h_\infty(\varepsilon)) - P(h_0) \geq const. > 0$. Note that this is true for any convergent subsequence. Therefore, we showed

$$\liminf_n \frac{g(k(n), \varepsilon, \lambda_n) - \lambda_n P(h_0)}{\lambda_n} \geq const. > 0$$

thus $\Pr\left(\widehat{h}_n \in \mathcal{B}_w^c(h_0)\right) \to 0$ (since $\Delta_n = o(\lambda_n)$).

We now show that, under assumption 3.8, consistency under the weak topology implies consistency under the strong norm. By assumption 3.8, $P(\widehat{h}_n)) - P(h_0) + \langle t_0, \widehat{h}_n - h_0\rangle_{\mathbf{H}^*,\mathbf{H}} \geq g(||\widehat{h}_n - h_0||_s)$. Lemma B.5(2) implies that $P(\widehat{h}_n) - P(h_0) \leq o_P(1)$; hence $g(||\widehat{h}_n - h_0||_s) = o_P(1)$, and $||\widehat{h}_n - h_0||_s = o_P(1)$ by our assumption over $g(.)$. *Q.E.D.*

C Convergence Rate

Lemma C.1. *Suppose that all the conditions of Theorem 4.1(1) hold. Then:*

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

(2) Under assumption 3.7, $||\hat{h}_n - \Pi_n h_0|| = O_P\left(\max\{\delta_{m,n}, o(\sqrt{\lambda_n}), ||\Pi_n h_0 - h_0||\}\right)$.

(3) Under assumption 4.2, $||\hat{h}_n - \Pi_n h_0|| = O_P\left(\max\{\delta_{m,n}, \sqrt{\lambda_n ||\hat{h}_n - \Pi_n h_0||_s}, ||\Pi_n h_0 - h_0||\}\right)$.

PROOF OF LEMMA C.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 $M > 1$,

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

By assumption 2.1, there are two finite constants $K, K' > 0$ such that:

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

which implies

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

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 $||\widehat{h}_n - h_0|| = O_P(r_n)$.

Under assumption 3.3, $\lambda_n |P(\Pi_n h_0) - P(\widehat{h}_n)| = O_P(\lambda_n)$; hence Result (1) follows.

Under assumption 3.7, $\lambda_n |P(\Pi_n h_0) - P(\widehat{h}_n)| = o_P(\lambda_n)$; hence Result (2) follows.

For Result (3), using the same argument as that for Results (1)(2), inequality (C.1) still holds. By assumption 4.2, $\lambda_n\left(P(\widehat{h}_n) - P(\Pi_n h_0)\right) \geq \lambda_n \langle t_0, \widehat{h}_n - \Pi_n h_0\rangle_{\mathbf{H}^*, \mathbf{H}}$. Thus

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

hence

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

By assumption 4.1, Lemma C.1(3) 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 4.1: Directly follows from Lemma C.1 and the definition of $\omega_n(\delta, \mathcal{H}_{osn})$. *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 the conclusion of Theorem 4.1 remains true under the stated assumptions. *Q.E.D.*

PROOF OF LEMMA 5.1: To simplify notation we denote $b_j = \varphi(\nu_j^{-2})$. 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(i) (the Riesz basis), there is a finite constant $c_1 > 0$ such that

$$c_1 ||h||_s^2 \leq \sum_{j \leq k(n)} |\langle h, q_j\rangle_s|^2 \leq (\max_{j \leq k(n)} b_j^{-1}) \sum_{j \leq k(n)} b_j |\langle h, q_j\rangle_s|^2 \leq \frac{1}{c b_{k(n)}} ||h||^2,$$

where the last inequality is due to assumption 5.2(i) and $\{b_j\}$ non-increasing. Similarly, assumptions 5.1(i) and 5.2(ii) imply result (2) since

$$\begin{aligned} c_2||h_0 - \Pi_n h_0||_s^2 &\geq \sum_{j>k(n)} |\langle h_0 - \Pi_n h_0, q_j\rangle_s|^2 \\ &\geq c(\min_{j>k(n)} b_j^{-1}) \sum_{j>k(n)} b_j |\langle h_0 - \Pi_n h_0, q_j\rangle_s|^2 \geq \frac{c'}{b_{k(n)}} ||h_0 - \Pi_n h_0||^2 \end{aligned}$$

for some finite positive constants c_2, c and c'. Result (3) directly follows from results (1) and (2). *Q.E.D.*

PROOF OF LEMMA 5.2: Denote $b_j = \varphi(\nu_j^{-2})$. For any $h \in \mathcal{H}_{os}$ with $||h||^2 \leq O(\delta^2)$, and for any $k \geq 1$, assumptions 5.1(i), 5.3 and 5.4(i) imply that there are finite positive constants c_1 and c such that:

$$\begin{aligned} c_1||h||_s^2 &\leq \sum_{j\leq k} \langle h, q_j\rangle_s^2 + \sum_{j>k} \langle h, q_j\rangle_s^2 \\ &\leq (\max_{j\leq k} b_j^{-1}) \sum_j b_j \langle h, q_j\rangle_s^2 + M^2(\nu_{k+1})^{-2\alpha} \leq \frac{1}{c} b_k^{-1}\delta^2 + M^2(\nu_{k+1})^{-2\alpha}. \end{aligned}$$

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

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

and

$$\omega(\delta, \mathcal{H}_{os}) \equiv \sup_{h\in\mathcal{H}_{os}: ||h-h_0||\leq\delta} ||h - h_0||_s \leq const. \frac{\delta}{\sqrt{b_{k^*}}}$$

thus Result (1) holds. Result (2) follows from Lemma 5.1 and Result (1). *Q.E.D*

PROOF OF COROLLARY 5.3: By Theorem 4.1, Lemma 5.1 and Lemma 5.2(2), Results of Corollary 5.2 are obviously true. We now specialize Corollary 5.2 to the PSMD estimator using a series LS estimator $\widehat{m}(X, h)$. For this case we have $\delta_{m,n}^{*2} = \frac{J_n^*}{n} \asymp b_{m,J_n^*}^2$.

By assumptions 4.1 and 5.4(ii), we have: for all $h \in \mathcal{H}_{os}$,

$$E\{m(X,h)'m(X,h)\} \leq ||h - h_0||^2 \leq const. \sum_{j=1}^{\infty} \{\varphi(\nu_j^{-2})\} \langle h - h_0, q_j\rangle_s^2.$$

On the other hand, assumption 5.3' implies that $\sum_j \nu_j^{2\alpha} \langle h - h_0, q_j\rangle_s^2 \leq const.$ for all $h \in \mathcal{H}_{os}$. Denote $\eta_j = \{\varphi(\nu_j^{-2})\}\langle h - h_0, q_j\rangle_s^2$. Then $\sum_j \nu_j^{2\alpha}\{\varphi(\nu_j^{-2})\}^{-1}\eta_j \leq M$. Therefore, the class $\{g \in L^2(\mathcal{X}, ||\cdot||_{L^2(f_X)}) : g(\cdot) = m(\cdot, h),\ h \in \mathcal{H}_{os}\}$ is embedded in the ellipsoid $\{g \in L^2(\mathcal{X}, ||\cdot||_{L^2(f_X)}) : ||g||_{L^2(f_X)}^2 = \sum_j \eta_j,\ and\ \sum_j \nu_j^{2\alpha}\{\varphi(\nu_j^{-2})\}^{-1}\eta_j \leq M\}$. By invoking the results of Yang and Barron (1999), it follows that the J_n-th approximation error rate of this ellipsoid satisfies $b_{m,J_n}^2 \leq const.\nu_{J_n}^{-2\alpha}\{\varphi(\nu_{J_n}^{-2})\}$. Hence $\delta_{m,n}^{*2} = \frac{J_n^*}{n} \asymp b_{m,J_n^*}^2 \leq const.\nu_{J_n^*}^{-2\alpha}\{\varphi(\nu_{J_n^*}^{-2})\}$, and $||\widehat{h} - h_0||_s = O_P\left(\nu_{J_n^*}^{-\alpha}\right) = O_P\left(\sqrt{\frac{J_n^*}{n}\{\varphi(\nu_{J_n^*}^{-2})\}^{-1}}\right)$. *Q.E.D*

D Application

Let $\varpi(y_2) \equiv \left(1+|y_2|^2\right)^{-\vartheta/2}$ for some $\vartheta \geq 0$ and $w(y_2) \equiv \left(1+|y_2|^2\right)^{-\theta/2}$ for some $\theta > 0$.

Proof of Proposition 6.1: We obtain both results by verifying that all the assumptions of Theorem 3.2 (lower semicompact penalty) and Lemma B.5 (1) are satisfied.

For result (**1**), Assumption 3.1(i) is directly assumed, assumption 3.1(ii) follows from our choice of $\mathcal{H}$ and assumption 3.1(iii) follows from Condition 6.2(ii). Assumption 2.2 is directly imposed. Assumption 2.3(i) follows trivially since $|\rho(Z,h)| \leq 1$. Assumption 2.3(ii) holds by the choice of the sieve basis for $p^{J_n}(X)$ and by condition 6.4 with $b^2_{m,J_n} = J_n^{-2r_m}$.

Given the choice of the sieve space $\mathcal{H}_n$ and the definition of $||\cdot||_s$, we have for $h_0 \in \mathcal{H}$,

$$\|h_0 - \Pi_n h_0\|_s \leq c\{k_1(n)\}^{-\alpha_1/d} + c_n'\{k_2(n)\}^{-\alpha_2/d} = o(1),$$

thus assumption 3.2(i) holds. 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

$$\begin{aligned} & |m(X,h) - m(X,h_0)| \\ \leq \; & E[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1)+\overline{h}_2(Y_2))|X] \times \sup_{y_1}|h_1(y_1)-h_{01}(y_1)| \\ & + E[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1)+\overline{h}_2(Y_2))\frac{1}{w(Y_2)}|X] \times \sup_{y_2}|[h_2(y_2)-h_{02}(y_2)]w(y_2)|. \end{aligned}$$

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 and for almost all X, we have

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

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

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

hence assumption 3.2(ii) holds.

We have that for any $M < \infty$, the embedding of the set $\{h \in \mathcal{H} : P(h) = ||\varpi h_2||_{\mathcal{I}^{\alpha_2}_{p,q}} \leq M\}$ into $\mathbf{H}$ is compact under the norm $||\cdot||_s$; hence assumption 3.5 is satisfied.

Assumption 3.6(i) follows directly from our choices of $\mathcal{H}$, $\mathcal{H}_n$ and $\|\cdot\|_s$. For assumption 3.6(ii) notice that for all $h, h' \in \mathcal{H}$,

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

conditions 6.1(ii)(iii) and the fact that $|m(X,h)| \leq 1$ imply assumption 3.6(ii) is satisfied.

Assumption 3.3(b) directly follows. Now the results follow from Theorem 3.2 provided that $\max\{\delta_{m,n}^2, E[m(X,\Pi_n h_0)^2]\} = O(\lambda_n)$. We already have $\delta_{m,n}^2 = \frac{J_n}{n} + J_n^{-2r_m} = O(\lambda_n)$. By conditions 6.1(ii)(iii)(iv), we also have

$$\begin{aligned}
& E\{\|m(X,\Pi_n h_0)\|_E^2\} \\
= \;& E\{\left|E\left[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1)+\overline{h}_2(Y_2))\{\Pi_n h_{01}(Y_1) - h_{01}(Y_1) + \Pi_n h_{02}(Y_2) - h_{02}(Y_2)\}|X\right]\right|^2\} \\
\leq \;& E\left\{E\left(\left[f_{Y_3|Y_1,Y_2,X}(\overline{h}_1(Y_1)+\overline{h}_2(Y_2))\{\Pi_n h_{01}(Y_1) - h_{01}(Y_1) + \Pi_n h_{02}(Y_2) - h_{02}(Y_2)\}\right]^2 |X\right)\right\} \\
\leq \;& CE\left\{[\Pi_n h_{01}(Y_1) - h_{01}(Y_1) + \Pi_n h_{02}(Y_2) - h_{02}(Y_2)]^2\right\} \\
\leq \;& 2CE\{[\Pi_n h_{01}(Y_1) - h_{01}(Y_1)]^2\} + 2CE\{[\Pi_n h_{02}(Y_2) - h_{02}(Y_2)]^2\} \\
\leq \;& const.\, \|h_0 - \Pi_n h_0\|_s^2 = O\left(\max\left[\{k_1(n)\}^{-2\alpha_1/d}, \{k_2(n)\}^{-2\alpha_2/d}\right]\right) = O(\lambda_n),
\end{aligned}$$

from which the result (1) now follows.

For result **(2)**, the verifications 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 $||h||_s$.

For assumption 3.2(ii), notice that

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

and that $|m(X,h)| \leq 1$, we have, under conditions 6.1(ii)(iii),

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

thus assumption 3.2(ii) is satisfied. For assumption 3.6(ii), notice that for all $h, h' \in \mathcal{H}$,

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

Therefore conditions 6.1(ii)(iii) and the fact that $|m(X,h)| \leq 1$ for all h and for almost all X, imply assumption 3.6(ii). The rest of the verifications are the same as those for Result (1). *Q.E.D.*

Proof of Proposition 6.2: We obtain the results by verifying that all the assumptions of Theorem 3.3 (convex penalty) are satisfied. Again assumptions 2.2 and 2.3 hold with $b_{m,J_n}^2 =$

$J_n^{-2r_m}$. Assumptions 3.1(i)(iii) were already verified in the previous proof, and assumption 3.1(ii) holds trivially given the choice of the norm $||h||_s = \sup_{y_1} |h(y_1)| + ||h_2 w||_{L^2(\mathcal{R}^d, leb)}$ for the spaces $\mathcal{H} = \Lambda_1^{\alpha_1}([0,1]^d) \times \mathcal{H}^2 \subset \mathbf{H} = L^\infty([0,1]^d) \times \{h_2 : ||h_2 w||_{L^2(\mathcal{R}^d, leb)} < \infty\}$. By the choice of the spaces $\mathcal{H}_n$ and $\mathcal{H}$, we have:

$$||\Pi_n h_{01} - h_{01}||_{L^2(f_{Y_1})} \leq \sup_{h_1 \in \mathcal{H}^1} ||\Pi_n h_1 - h_1||_{L^2(f_{Y_1})} \leq \sup_{h_1 \in \mathcal{H}^1} \sup_{y_1} |\Pi_n h_1(y_1) - h_1(y_1)| \leq c\{k_1(n)\}^{-\alpha_1/d},$$

$$||\Pi_n h_{02} - h_{02}||_{L^2(f_{Y_2})} \leq \sqrt{\sup_{y_2} \frac{f_{Y_2}(y_2)}{w^2(y_2)}} \times ||w(\Pi_n h_{02} - h_{02})||_{L^2(\mathcal{R}^d, leb)} \leq c'\{k_2(n)\}^{-\alpha_2/d},$$

thus assumption 3.2(i) holds. Assumption 3.2(ii) is already verified in the proof of Result (1)(ii) of Proposition 6.1. Assumption 3.7 follows from the fact that $\widehat{P}(h) = P(h) = ||(wh_2)||^2_{L^2(\mathcal{R}^d, leb)}$ and

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

Assumption 3.8 follows from

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

with $g(\varepsilon) = \varepsilon^2$ and $t_0 = 2wh_{02}$. Assumption 3.9 follows by our choice of norm and space. Assumption 3.10(a) is implied by condition 6.5. Assumption 3.11(b) follows from the fact that $P(h) = ||wh_2||^2_{L^2(\mathcal{R}^d, leb)}$ is convex and continuous. Finally, by conditions 6.1(ii)(iii)(iv), we have

$$E\{\|m(X, \Pi_n h_0)\|_E^2\} \leq const. \|h_0 - \Pi_n h_0\|_s^2 = O\left(\max\left[\{k_1(n)\}^{-2r_1}, \{k_2(n)\}^{-2r_2}\right]\right).$$

The result now follows from Theorem 3.3. *Q.E.D.*

Proof of Proposition 6.3: We obtain the results by verifying that all the assumptions of Corollary 5.1 are satisfied. As assumptions 3.1, 3.2, 2.2 and 2.3 are already verified in the proofs of Propositions 6.1 and 6.2, assumption 5.1 is automatically satisfied. Condition 6.7 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

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

$$\begin{aligned} \|h - h_0\|^2 &= E\left(\frac{dm(X, h_0)}{dh}[h - h_0]\right)^2 \leq const. \|h - h_0\|_s^2, \\ \text{where } \|h - h_0\|_s^2 &= E\left\{(h_1(Y_1) - h_{01}(Y_1) + h_2(Y_2) - h_{02}(Y_2))^2\right\}, \end{aligned}$$

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 $h \in \mathcal{H}_{os}$, $E\left[(m(X,h)-m(X,h_0))^2\right] = ||T_{\overline{h}}[h_1-h_{01}+h_2-h_{02}]||^2_{L^2(f_X)}$, where $\overline{h}$ is a convex combination of h and h_0 in $\mathcal{H}_{os}$. While $\|h-h_0\|^2 = ||T_{h_0}[h_1-h_{01}+h_2-h_{02}]||^2_{L^2(f_X)}$ by definition. Thus for all $h \in \mathcal{H}_{os}$, $c^2\|h-h_0\|^2 \leq E\left[(m(X,h)-m(X,h_0))^2\right] \leq C^2\|h-h_0\|^2$, and assumption 4.1(ii) holds. The conclusions now follow directly from Corollary 5.1. *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.

Arellano, C. (2008). Default Risk and Income Fluctuations in Emerging Economies. *American Economic Review* **98** 690-713.

Bansal, R. and S. Viswanathan (1993). No Arbitrage and Arbitrage Pricing: A New Approach. *The Journal of Finance* **48**, 1231-1262.

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.

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 (2006). Land of Addicts? An Empirical Investigation of Habit-based Asset Pricing Models. NBER Working Paper No. 10503, forthcoming in *Journal of Applied Econometrics*.

Chen, X. and D. Pouzo (2008a). Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Moments. Yale University, Cowles Foundation Discussion Paper No. 1650.

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

Chen, X. and D. Pouzo (2009). On Plug-in PSMD Estimation of Functionals of Semi/nonparametric Conditional and Unconditional Moment Models. Mimeo, Yale University and New York University.

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.

D'Haultfoeuille, X. (2008). On the Completeness Condition in Nonparametric Instrumental Problems. mimeo, ENSAE, CREST-INSEE.

Edmunds, D. and H. Triebel (1996). *Function Spaces, Entropy Numbers, Differential Operators*, Cambridge University Press: Cambridge.

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 (2008). Identification and Estimation by Penalization in Nonparametric Instrumental Regression. Mimeo, Toulouse School of Economics.

Gagliardini, P. and O. Scaillet (2008). Tikhonov Regularization for Nonparametric Instrumental Variable Estimators. Mimeo, University of Lugano and Swiss Finance Institute.

Gallant, A. and G. Tauchen (1989). Semiparametric Estimation of Conditional Constrained Heterogenous Processes: Asset Pricing Applications. *Econometrica*, 57, 1091-1120.

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

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.

Nair, M., S. Pereverzev and U. Tautenhahn (2005). Regularization in Hilbert scales under general smoothing conditions. *Inverse Problems* **21**, 1851-1869.

Newey, W.K. (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.

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.