ESTIMATION OF NONPARAMETRIC CONDITIONAL MOMENT MODELS WITH POSSIBLY NONSMOOTH GENERALIZED RESIDUALS

By

Xiaohong Chen and Demian Pouzo

April 2008
Revised July 2009
Revised January 2011

COWLES FOUNDATION DISCUSSION PAPER NO. 1650RR

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 [*]

Xiaohong Chen [†]and Demian Pouzo [‡]

Abstract

This paper studies nonparametric estimation of conditional moment restrictions 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; others 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.

[*] First version: August 2005, Final revised version: December 2010. This is a substantially revised version of Cowles Foundation Discussion Paper 1650. We are grateful to a co-editor, three anonymous referees, R. Blundell, V. Chernozhukov, T. Christensen, K. Evdokimov, H. Hong, J. Horowitz, J. Huang, S. Lee, Z. Liao, D. Nekipelov, J. Powell, A. Santos 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, O. Linton, 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. The usual disclaimer applies.

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

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

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:

$$E[\rho(Y, X_z; \theta_0, h_{01}(\cdot), ..., h_{0q}(\cdot))|X] = 0, \tag{1}$$

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) extends the semi/nonparametric conditional moment framework previously considered in Chamberlain (1992), Newey and Powell (2003) (henceforth NP) and Ai and Chen (2003) (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). For instance, it includes the model of a shape-invariant system of Engel curves with endogenous total expenditure of Blundell, Chen and Kristensen (2007) (BCK), which itself is an extension of the nonparametric mean instrumental variables regression (NPIV) analyzed in NP, Darolles, Fan, Florens and Renault (2010) (DFFR) and Hall and Horowitz (2005) (HH):

$$E[Y_1 - h_0(Y_2)|X] = 0. \tag{2}$$

Model (1) also nests the quantile instrumental variables (IV) treatment effect model of Chernozhukov and Hansen (2005) (CH), and the nonparametric quantile instrumental variables regression (NPQIV) of Chernozhukov, Imbens and Newey (2007) (CIN) and Horowitz and Lee (2007) (HL):

$$E[1\{Y_1 \leq h_0(Y_2)\}|X] = \gamma \in (0, 1), \tag{3}$$

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), 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 (2009) study a consumption-based asset pricing model with an unknown habit formation. Their model is an example of (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); 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 and Chen and Pouzo (2009b, 2010) for an econometric study of pricing default risk.

As demonstrated in NP, AC, CIN and Chen, Chernozhukov, Lee and Newey (2010) (CCLN), the key difficulty of analyzing the semi/nonparametric model (1) is not the presence of the unknown finite dimensional parameter θ_0, but the fact that some of the unknown functions $h_{0\ell}(\cdot), \ell=1,...,q$ depend on the endogenous variable Y.[1] Therefore, in this paper we shall focus on the nonparametric estimation of $h_0()$, which is identified by the following conditional moment restrictions:

$$E[\rho(Y,X_z;h_{01}(\cdot),...,h_{0q}(\cdot))|X]=0, \tag{4}$$

where $h_0(\cdot)\equiv(h_{01}(\cdot),...,h_{0q}(\cdot))$ depends on Y and may enter $\rho()$ nonlinearly and possibly nonsmoothly.[2] 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 (4). All of the PSMD procedures minimize a possibly penalized consistent estimate of the minimum distance criterion, $E\left\{E[\rho(Z;h(\cdot))|X]'W(X)E[\rho(Z;h(\cdot))|X]\right\}$, over sieve spaces $(\mathcal{H}_n)$ that are dense in the infinite dimensional function space $\mathcal{H}$.[3] 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; others use *large dimensional sieves* (i.e., $\dim(\mathcal{H}_n)/n\to$ *const.* >0), with lower semicompact[4] and/or convex penalties. Under relatively low level sufficient conditions and without assuming $||\cdot||_s-$compactness of the function parameter space $\mathcal{H}$, we establish consistency and the convergence rates under norm $||\cdot||_s$ for these PSMD estimators. Our

[1] In some applications the presence of the parametric part θ_0 in the semi/nonparametric model (1) aids the identification of the unknown function h_0; see, e.g., Chen and Ludvigson (2009) and CCLN.

[2] See Chen and Pouzo (2009a) for semiparametric efficient estimation of the parametric part θ_0 for the general semi/nonparametric model (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 this paper.

[3] In this paper, W denotes a weighting matrix, n is the sample size, and $\dim(\mathcal{H}_n)$ is the dimension of the sieve space.

[4] See Section 2 for its definition.

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 (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 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 whole 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., DFFR, HH, HL, Carrasco, Florens and Renault (2007) (CFR), Chernozhukov, Gagliardini and Scaillet (2010) (CGS) and the references therein. When h_0 enters the residual function $\rho(Z;h_0)$ linearly such as in the NPIV model (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 (3), the numerical implementations of TR-MD estimators typically involve some finite dimensional sieve approximations to functions in $\mathcal{H}$.[5] For example, in the simulation study of the NPQIV model (3), HL approximate the unknown function $h_0(\cdot)$ by a Fourier series with a large number of terms; hence they could ignore the Fourier series approximation error and view their implemented procedure as a 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 Chebyshev 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 (equation (9)), simulation studies in BCK and Chen and Pouzo (2009a) indicate that it is easier to compute a penalized SMD estimator using finite dimensional linear sieves (equation (8)).[6] In summary, some versions of our proposed PSMD procedures have already been numerically implemented in the existing literature, but their asymptotic properties have not been established for the general model (4).

There are some published papers on the asymptotic properties of the SMD and the TR-MD procedures for the linear NPIV model (2).[7] For example, 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, DFFR, and Gagliardini and Scaillet (2010) (GS) for the convergence rate in a root mean squared norm of their kernel based TR-MD estimators; HH and Chen and Reiss (2010) (CR) for the minimax optimal rate in a root mean squared norm for the NPIV model.

There are currently only a few published papers on the asymptotic properties of any nonpara-

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

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

[7] See NP, DFFR, BCK, CFR, Severini and Tripathi (2006), D'Haultfoeuille (2010), Florens, Johannes and van Bellegem (2010) and others for identification of the NPIV model.

metric estimators of h_0 when it could enter the conditional moment restrictions (4) nonlinearly. Assuming that the function space $\mathcal{H}$ is compact (in $||\cdot||_s$) and that the residual function $\rho(Z, h(\cdot))$ is pointwise smooth in h, NP established the $||\cdot||_s-$consistency of the SMD estimator, and AC derived some convergence rate of the SMD estimator in a pseudo metric weaker than $||\cdot||_s$. For the NPQIV example (3),[8] CIN obtained the consistency (in a sup-norm) of the SMD estimator when the function space $\mathcal{H}$ is sup-norm compact, and HL established the convergence rate (in a root mean squared norm) of a kernel based TR-MD estimator. In a recent working paper on the same NPQIV model, CGS present the consistency (in a root mean squared norm) and pointwise asymptotic normality of their kernel based TR-MD estimator. To the best of our knowledge, there is no published work that establishes the convergence rate (in $||\cdot||_s$) of any estimator of h_0 for the general model (4).

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 (4), without assuming the $||\cdot||_s-$compactness of the function space $\mathcal{H}$. Our PSMD procedures using large dimensional linear sieves ($\dim(\mathcal{H}_n)/n \to const. > 0$) and lower semicompact and/or convex penalties are computable extensions of the current TR-MD procedures for the NPIV and the NPQIV models to all conditional moment models (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 (4), and then present the PSMD procedures. Section 3 provides sufficient conditions for consistency in a Banach space norm $||\cdot||_s$ and Section 4 derives the convergence rate. Under relatively low level sufficient conditions, Section 5 presents the rate of convergence in a Hilbert norm $||\cdot||_s$ and shows that the rate for the general model (4) coincides with the optimal minimax rate for the NPIV model (2). Throughout these sections, we use the NPIV example (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 (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. Some regularity conditions and general lemmas are stated in the appendix. The online supplemental material

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

contains all the proofs, as well as a brief review of some functional spaces and sieve bases.

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 positive sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, $a_n \asymp b_n$ 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 $\lim_{c\to\infty} \limsup_n \Pr(a_n/b_n > c) = 0$; and $a_n = o_p(b_n)$ means that for all $\varepsilon > 0$, $\lim_{n\to\infty} \Pr(a_n/b_n > \varepsilon) = 0$. We use "wpa1" to denote "with probability approaching one". For any vector-valued A, we let A' denote its transpose and $||A||_W \equiv \sqrt{A'WA}$ for its weighted norm, although sometimes we also use $|A| = ||A||_I \equiv \sqrt{A'A}$ without too much confusion. We use $\mathcal{H}_n \equiv \mathcal{H}_{k(n)}$ to denote sieve spaces.

2 Penalized Sieve Minimum Distance Estimation

Suppose that observations $\{(Y_i', X_i')\}_{i=1}^n$ are strictly stationary ergodic, that for each i, the distribution of (Y_i', X_i') is the same as that 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 (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 the parameter space $\mathcal{H}$ is a non-empty, closed, possibly non-compact 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 $m(X,h) \equiv \left(m_1(X,h), ..., m_{d_\rho}(X,h)\right)'$ and $W(X)$ be a positive-definite finite weighting matrix for almost all X. Under the assumption that model (4) identifies $h_0 \in \mathcal{H}$, we have

$$E\left[||m(X,h)||_W^2\right] \geq 0 \text{ for all } h \in \mathcal{H}; \text{ and } = 0 \text{ if and only if } h = h_0. \quad (5)$$

One could construct an estimator of $h_0 \in \mathcal{H}$ by minimizing a sample analog of $E\left[||m(X,h)||_W^2\right]$ over the function space $\mathcal{H}$. Unfortunately, when $h_0(\cdot)$ depends on the endogenous variables Y, the "$||\cdot||_s$−identifiable uniqueness" condition for $||\cdot||_s$−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

$\liminf_{k\to\infty} E\left[||m(X,h_k)||_W^2\right] = 0$; that is, the metric $||h-h_0||_s$ is not continuous with respect to the population criterion function $E\left[||m(X,h)||_W^2\right]$, and the problem is ill-posed.[9]

When $E\left[||m(X,h)||_W^2\right]$ is lower semicontinuous on $(\mathcal{H},||\cdot||_s)$ and $h_0\in\mathcal{H}$ is its unique minimizer, one way to ensure the "$||\cdot||_s-$identifiable uniqueness" is to assume that the parameter space $\mathcal{H}$ is a compact subset of $(\mathbf{H},||\cdot||_s)$; see, e.g., NP, CIN, AC and BCK for imposing such a compactness condition to establish $||\cdot||_s-$consistency of their SMD estimators.

In many economic applications, although the functional forms of structural functions h_0 (such as Engel curves or cost functions) are unknown, they are believed to be Hölder continuous or to have continuous derivatives. Thus, it is reasonable to assume that the parameter space $\mathcal{H}$ is a subset of a Hölder space (denoted as Λ^α) or a Sobolev space (denoted as W_p^α) with $\alpha>0$,[10] but, it could be a non-compact subset of a space of smooth functions. For example, when applying the NPIV (2) or the NPQIV (3) model to estimate an Engel curve h_0, it is sensible to assume that h_0 belongs to $\mathcal{H}=\{h\in W_2^\alpha(f_{Y_2}):\sup_y|h(y)|\le 1,||\nabla^\alpha h||_{L^2(f_{Y_2})}<\infty\}$ for some $\alpha\ge 1$, which is a smooth function space, but is not $\|\cdot\|_{L^2(f_{Y_2})}-$compact nor $\|\cdot\|_{L^\infty(leb)}-$compact. To allow for wider applicability, in this paper we assume that the parameter space $\mathcal{H}$ is an infinite dimensional, possibly non-compact subset of a separable Banach space $(\mathbf{H},||\cdot||_s)$.[11]

In order to design a consistent estimator for $h_0\in\mathcal{H}$ with possibly non-compact parameter space $\mathcal{H}$ we need to tackle two issues. First, we need to replace the unknown population minimum distance criterion, $E\left[||m(X,h)||_W^2\right]$, by a consistent empirical estimate. Second, we need to regularize the problem to make the metric $||h-h_0||_s$ continuous with respect to the criterion function.

2.1 PSMD estimators

In this paper we consider a class of (approximate) penalized sieve minimum distance (PSMD) estimators, $\widehat{h}_n$, defined as:

$$\widehat{Q}_n(\widehat{h}_n) \le \inf_{h\in\mathcal{H}_n}\widehat{Q}_n(h)+\widehat{\eta}_n, \quad \text{with } \widehat{\eta}_n\ge 0,\ \widehat{\eta}_n=O_p(\eta_n), \tag{6}$$

$$\text{and} \quad \widehat{Q}_n(h) \equiv \frac{1}{n}\sum_{i=1}^n \widehat{m}(X_i,h)'\widehat{W}(X_i)\widehat{m}(X_i,h)+\lambda_n\widehat{P}_n(h), \tag{7}$$

where $\{\eta_n\}_{n=1}^\infty$ is a sequence of positive real values such that $\eta_n=o(1)$, $\widehat{m}(X,h)$ 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 $\mathcal{H}$; $\lambda_n\ge 0$ is a penalization parameter such that $\lambda_n\to 0$ as $n\to\infty$; the penalty $\widehat{P}_n()\ge 0$

[9] 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 $j=1,...,d_\rho$.

[10] See Chen (2007) and the online supplemental material for definitions of Hölder space, Sobolev space, Besov space and other widely used function spaces in economics.

[11] A subset of $(\mathbf{H},||\cdot||_s)$ is $\|\cdot\|_s-$compact if and only if it is closed and *totally* bounded (in $\|\cdot\|_s$). It is well known that a closed and bounded subset of $(\mathbf{H},||\cdot||_s)$ is $\|\cdot\|_s-$compact if and only if it is finite dimensional.

is an empirical analog of a non-random penalty function $P : \mathcal{H} \to [0, +\infty)$; $\widehat{W}(X)$ is a consistent estimator of $W(X)$ that is introduced to address potential heteroskedasticity. In this paper we assume that each of $\widehat{m}(\cdot, h)$, $\widehat{W}(\cdot)$ and $\widehat{P}_n(h)$ is jointly measurable in the data $\{(Y_i', X_i')\}_{i=1}^n$ and the parameter $h \in \mathcal{H}$, and hence the approximate PSMD estimator $\widehat{h}_n$ exists.[12]

The sieve space $\mathcal{H}_n$ in the definition of the PSMD estimator (6) 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:

$$\mathcal{H}_n = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot) \right\}, \quad k(n) < \infty, \;\; k(n) \to \infty \text{ slowly as } n \to \infty, \tag{8}$$

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:

$$\mathcal{H}_n = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot), \;\; R_n(h) \le B_n \right\}, \quad B_n \to \infty \text{ slowly as } n \to \infty, \tag{9}$$

where the constraint $R_n(h) \le B_n$ reflects prior information about $h_0 \in \mathcal{H}$ such as smoothness properties. The sieve space $\mathcal{H}_n$ in (9) 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, $\mathcal{H}_n = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot), \;\; \|h\|_s \le \log(n) \right\}$ 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 (6) is typically convex and/or *lower semicompact* (i.e., the set $\{h \in \mathcal{H} : P(h) \le M\}$ is compact in $(\mathbf{H}, \|\cdot\|_s)$ for all $M \in [0, \infty)$), and reflects prior information about $h_0 \in \mathcal{H}$. For instance, when $\mathcal{H} \subseteq L^p(d\mu)$, $1 \le p < \infty$, 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 $\{h : ||h||_{L^2(d\mu)}^2 + ||\nabla^r h||_{L^p(leb)}^p < \infty\}$, $1 \le p < \infty$, $r \ge 1$, 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 $\widehat{P}_n(h) = ||\nabla^k h||_{L^p(leb)}^p$ for some $k \in [1, r]$.

Our definition of PSMD estimators includes many existing estimators as special cases. For example, when $\widehat{\eta}_n = 0$, $\lambda_n = 0$ and $\mathcal{H}_n$ given in (9) is a finite-dimensional (i.e., $k(n) < \infty$) compact sieve space of $\mathcal{H}$, the (approximate) PSMD estimator (6) becomes solution to:

$$\frac{1}{n} \sum_{i=1}^{n} ||\widehat{m}(X_i, \widehat{h}_n)||_{\widehat{W}}^2 \le \inf_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} ||\widehat{m}(X_i, h)||_{\widehat{W}}^2,$$

[12] In this paper we implictly assume that $\widehat{h}_n$ is measurable with respect to the underlying probability. If not, its asymptotic properties remain valid after being stated under the outer measure. See Remark A.1 in the appendix for sufficient conditions to ensure measurability.

[13] See Chen and Shen (1998), Chen (2007) and the references therein for additional examples of linear sieves (or series), and nonlinear sieves.

which is the original SMD estimator proposed in NP, AC and CIN. When $\widehat{\eta}_n = 0$, $\lambda_n \widehat{P}_n() > 0$, $\widehat{P}_n() = P()$ and $\mathcal{H}_n = \mathcal{H}$ (i.e., $k(n) = \infty$), the (approximate) PSMD estimator (6) becomes solution to:

$$\frac{1}{n}\sum_{i=1}^{n} ||\widehat{m}(X_i, \widehat{h}_n)||^2_{\widehat{W}} + \lambda_n P(h) \le \inf_{h \in \mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} ||\widehat{m}(X_i, h)||^2_{\widehat{W}} + \lambda_n P(h) \right\},$$

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 DFFR, 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 (6) 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 letting $\lambda_n = 0$ and $\mathcal{H}_n$ given in (9) be a finite dimensional compact sieve. However, it also allows for $\lambda_n \searrow 0$ fast with $\mathcal{H}_n$ given in (8) being 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 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 $n^{-1}\sum_{i=1}^{n} ||\widehat{m}(X_i, h)||^2_{\widehat{W}}$ 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., R_n is a positive convex function in the definition of the sieve space (9)). To see why, let $clsp(\mathcal{H}_n)$ denote the closed linear span of $\mathcal{H}_n$ (in $||\cdot||_s$). Then the PSMD procedure (6) is equivalent to

$$\widehat{Q}_n(\widehat{h}_n) + \nu_n R_n(\widehat{h}_n) \le \inf_{h \in clsp(\mathcal{H}_n)} \left\{ \widehat{Q}_n(h) + \nu_n R_n(h) \right\} + O_p(\eta_n), \tag{10}$$

where $R_n(\widehat{h}_n) \leq B_n$ and $\nu_n \geq 0$ is such that $\nu_n(R_n(\widehat{h}_n) - B_n) = 0$; 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 $\nu_n R_n(h)$. For most applications, it suffices to have either $\lambda_n \widehat{P}_n(h) > 0$ or $\nu_n R_n(h) > 0$.

Even when $n^{-1}\sum_{i=1}^{n} ||\widehat{m}(X_i, h)||^2_{\widehat{W}}$ 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 (8)) with small penalization $\lambda_n > 0$.

Which class of PSMD estimators to use? In most economics applications, the unknown structural function h_0 is Hölder continuous or has continuous derivatives or satisfies some shape restrictions (such as monotonicity or concavity). To estimate such smooth functions for the model (4), we recommend to apply either the class of PSMD estimators using slowly growing finite dimensional sieves with/without small flexible penalty ($k(n) \to \infty$ slowly, $k(n)/n \to 0$; $\lambda_n \searrow 0$ fast or $\lambda_n = 0$), or the class of PSMD estimators using faster growing finite dimensional sieves with big lower semicompact penalty ($k(n) \to \infty$ faster, $k(n)/n \to 0$; $\lambda_n = O\left(k(n)/n\right)$). Our subsequent theoretical results and simulation studies indicate that these two classes of estimators perform well in finite samples, and can achieve the optimal rate of convergence under weaker assumptions than the class of PSMD estimators using large dimensional sieves with big lower semicompact penalty ($k(n)/n \to const. > 0$; $\lambda_n \searrow 0$ slowly) can. Among these two, the subclass of PSMD estimators using slowly growing finite dimensional linear sieves (i.e., series (8)) with small flexible penalty is our favorite since it is easier to compute and performs very well in finite samples.

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

To compute the PSMD estimator $\widehat{h}_n$ defined in (6), nonparametric estimators of the conditional mean function $m(\cdot, h) \equiv E[\rho(Z, h)|X = \cdot]$ and of the weighting matrix $W(\cdot)$ are needed. Without an analysis of asymptotic efficiency in nonparametric estimation of h_0, one typically let $\widehat{W}(\cdot) = W(\cdot) = I$ (identity), and $\widehat{m}(\cdot, h)$ be any nonparametric least squares (LS) estimator of $m(\cdot, h)$ such as the ones based on kernel, local linear, sieve (or series) methods.

In this paper, we establish the asymptotic properties of the PSMD estimator $\widehat{h}_n$, allowing for any nonparametric estimators of $m(\cdot, h)$ and $W(\cdot)$ that satisfy a mild regularity assumption 3.3. All the commonly used nonparametric consistent estimators such as the kernel estimators and the series LS estimators can be shown to satisfy assumption 3.3. For the sake of concreteness, in the empirical application and Monte Carlo simulations we use a series LS estimator

$$\widehat{m}(X, h) = p^{J_n}(X)'(P'P)^{-}\sum_{i=1}^{n} p^{J_n}(X_i)\rho(Z_i, h), \qquad (11)$$

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 $P'P$. See NP, AC, BCK, CIN, CR and others for more details and applications of this estimator.

3 Consistency

In the appendix, we provide a general consistency result (Lemma A.1) for any approximate penalized sieve extremum estimator, allowing for both well-posed and ill-posed problems, as well as time series observations. Here, in the main text we present consistency of various PSMD estimators (6).

We first impose three basic conditions on identification, sieves spaces, penalty functions and sample criterion function.

Assumption 3.1. *(**identification, sieves**) (i) $W(X)$ is a positive-definite finite weighting matrix for almost all X; (ii) $E[\rho(Z, h_0)|X] = 0$, and $\|h_0 - h\|_s = 0$ for any $h \in (\mathcal{H}, \|\cdot\|_s)$ with $E[\rho(Z, h)|X] = 0$; (iii) $\{\mathcal{H}_k : k \geq 1\}$ is a sequence of non-empty closed subsets satisfying $\mathcal{H}_k \subseteq \mathcal{H}_{k+1} \subseteq \mathcal{H}$, and for any $h \in \mathcal{H}$, there is $\Pi_n h \in \mathcal{H}_{k(n)}$ such that $||\Pi_n h - h||_s = o(1)$; (iv) $E[||m(X, \Pi_n h_0)||^2_W] = o(1)$.*

Assumption 3.2. *(**penalty**) either (a) or (b) or (c) holds: (a) $\lambda_n = 0$; (b) $\lambda_n > 0$, $\lambda_n = o(1)$, $\sup_{h\in\mathcal{H}_{k(n)}} |\widehat{P}_n(h) - P(h)| = O_p(1)$ and $|P(\Pi_n h_0) - P(h_0)| = O(1)$ with $P : \mathcal{H} \to [0, \infty)$, $P(h_0) < \infty$; (c) $\lambda_n > 0$, $\lambda_n = o(1)$, $\sup_{h\in\mathcal{H}_{k(n)}} |\widehat{P}_n(h) - P(h)| = o_p(1)$ and $|P(\Pi_n h_0) - P(h_0)| = o(1)$ with $P : \mathcal{H} \to [0, \infty)$, $P(h_0) < \infty$.*

Let $\{\eta_{0,n}\}_{n=1}^{\infty}$ and $\{\bar{\delta}^2_{m,n}\}_{n=1}^{\infty}$ be sequences of positive real values that decrease to zero as $n \to \infty$. Let $\mathcal{H}^{M_0}_{k(n)} \equiv \{h \in \mathcal{H}_{k(n)} : \lambda_n P(h) \leq \lambda_n M_0\}$ for a large but finite M_0 such that $\Pi_n h_0 \in \mathcal{H}^{M_0}_{k(n)}$ and that $\widehat{h}_n \in \mathcal{H}^{M_0}_{k(n)}$ with probability arbitrarily close to one for all large n. Given assumptions 3.2 and 3.3(i), such a M_0 always exists (see Lemma A.2 in the appendix).

Assumption 3.3. *(**sample criterion**) (i) $\frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i, \Pi_n h_0)||^2_{\widehat{W}} \leq c_0 E\left[||m(X, \Pi_n h_0)||^2_W\right] + O_p(\eta_{0,n})$ for some $\eta_{0,n} = o(1)$ and a finite constant $c_0 > 0$; (ii) $\frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i, h)||^2_{\widehat{W}} \geq cE\left[||m(X, h)||^2_W\right] - O_p(\bar{\delta}^2_{m,n})$ uniformly over $\mathcal{H}^{M_0}_{k(n)}$ for some $\bar{\delta}^2_{m,n} = o(1)$ and a finite constant $c > 0$.*

Under assumption 3.1(ii) (global identification) and assumption (iii) (definition of sieves), assumption 3.1(iv) is satisfied if $E[||m(X, h)||^2_W]$ is continuous at h_0 (under $\|\cdot\|_s$). Assumptions 3.2(b) and (c) are trivially satisfied when $\mathcal{H}_{k(n)} = \mathcal{H}$ and $\widehat{P}_n = P$. Assumption 3.2(c) is a stronger version of assumption 3.2(b). Under assumption 3.1(iii) and $P(h_0) < \infty$, a sufficient condition for $|P(\Pi_n h_0) - P(h_0)| = o(1)$ is that $P(\cdot)$ is continuous at h_0.

Assumption 3.3 is satisfied by most nonparametric estimators of $m(\cdot, h)$ and $W(\cdot)$. Note that assumption 3.3(i) only needs to hold at $\Pi_n h_0$. Lemma C.2 in the appendix shows that the series LS estimator $\widehat{m}(X, h)$ defined in (11) satisfies assumption 3.3.

Under the above regularity conditions, one can show that the PSMD estimator $\widehat{h}_n$ defined in (6) approximately solves the optimization problem:

$$\inf_{h\in\mathcal{H}_n}\left\{E\left[||m(X,h)||_W^2\right]+\lambda_n P(h)\right\}+O_p(\eta_n)\quad\text{for some sequence } \eta_n=o(1),$$

which has a solution, provided that the set $\left\{h\in\mathcal{H}_n : E\left[||m(X,h)||_W^2\right]+\lambda_n P(h)\leq M\right\}$ is compact in some topology $\mathcal{T}$ (that may be weaker than the norm $\|\cdot\|_s$ −topology on $\mathbf{H}$) for all $M\in[0,\infty)$. Further, when $E\left[||m(X,h)||_W^2\right]$ has a unique minimizer (h_0) on $(\mathcal{H},||\cdot||_s)$, we establish $\mathcal{T}$−consistency of $\widehat{h}_n$ under choices of smoothing parameters $k(n)\equiv\dim(\mathcal{H}_n)$ and λ_n, which in turn leads to $\|\cdot\|_s$ −consistency of $\widehat{h}_n$ under some assumptions over the penalty and the smoothing parameters. This explains why one could obtain $\|\cdot\|_s$ −consistency of $\widehat{h}_n$ by regularizing either the sieve space $\mathcal{H}_n$ or the penalty $\lambda_n P(\cdot)>0$ or both, without the need to assume the $\|\cdot\|_s$ −compactness of the whole parameter space $\mathcal{H}$.

In the following, for easy reference, we present consistency results for PSMD estimators using slowly growing finite dimensional sieves ($k(n)/n\to 0$) and PSMD estimators using large ($k(n)/n\to$ *const.* >0) or infinite dimensional sieves in separate subsections.

3.1 PSMD using slowly growing finite dimensional sieves

Denote $g\left(k(n),\varepsilon\right)\equiv\inf_{h\in\mathcal{H}_{k(n)}^{M_0}:||h-h_0||_s\geq\varepsilon}E\left[\|m(X,h)\|_W^2\right]$ for any $\varepsilon>0$.

Theorem 3.1. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n\geq 0$, $\eta_n=O(\eta_{0,n})$, and assumptions 3.1, 3.2(a)(b) and 3.3 hold. Suppose that for each integer $k<\infty$, $\dim(\mathcal{H}_k)<\infty$, $\mathcal{H}_k$ is bounded and $E\left[\|m(X,h)\|_W^2\right]$ is lower semicontinuous on $(\mathcal{H}_k,||\cdot||_s)$. Let $k(n)<\infty$ and $k(n)\to\infty$ as $n\to\infty$. If*

$$\max\left\{\eta_{0,n},E\left[||m(X,\Pi_n h_0)||_W^2\right],\bar{\delta}_{m,n}^2,\lambda_n\right\}=o\left(g\left(k(n),\varepsilon\right)\right)\quad\textit{for all } \varepsilon>0,\tag{12}$$

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 compact sieves, allowing for no penalty ($\lambda_n=0$), or any flexible penalty $P(h)$ with $\lambda_n>0$. It is clear that, when $\lambda_n=0$, $\liminf_{k(n)\to\infty}g\left(k(n),\varepsilon\right)=\inf_{h\in\mathcal{H}:||h-h_0||_s\geq\varepsilon}E\left[\|m(X,h)\|_W^2\right]$. Thus, given assumption 3.1(ii) (identification), for all $\varepsilon>0$, $\liminf_{k(n)\to\infty}g\left(k(n),\varepsilon\right)>0$ if $(\mathcal{H},||\cdot||_s)$ is compact; otherwise $\liminf_{k(n)\to\infty}g\left(k(n),\varepsilon\right)$ could be zero. When $(\mathcal{H},||\cdot||_s)$ is compact, restriction (12) becomes $\max\left\{\eta_{0,n},E\left[||m(X,\Pi_n h_0)||_W^2\right],\bar{\delta}_{m,n}^2,\lambda_n\right\}=o\left(1\right)$ and is trivially satisfied. Theorem 3.1 (with $\lambda_n=0$) not only recovers the consistency results of NP, AC and CIN when $(\mathcal{H},||\cdot||_s)$ is compact, but also implies consistency of the original SMD estimator when $\mathcal{H}$ is a class of smooth functions that is not compact in $||\cdot||_s$.

NPIV example (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 $W=I$ (identity weighting), $\mathcal{H}=\{h\in L^2(f_{Y_2}):$

$||\nabla^r h||^2_{L^2(leb)} < \infty\}$ (for some $r > 0$), which is not compact in $||\cdot||_s = ||\cdot||_{L^2(f_{Y_2})}$. 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 $\mathcal{H}_n = \{h \in \mathcal{H} : h(y_2) = \sum_{k=1}^{k(n)} a_k \phi_{1,k}(y_2), ||\nabla^r h||_{L^2(leb)} \leq \log(n)\}$ and $\lambda_n P(h) = \lambda_n ||\nabla^r h||^2_{L^2(leb)}$ for $\lambda_n \geq 0$. Note that $E[\|m(X,h)\|^2_W]$ is continuous on $(\mathcal{H}_n, \|\cdot\|_s)$ and

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

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

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

Note that $||h_n^* - \Pi_n h_0||^2_s$ 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 $||h_n^* - \Pi_n h_0||^2_s < (\varepsilon/3)^2$ 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 $k(n)$. Thus $E\left[||m(X,\Pi_n h_0)||^2_W\right] / g\left(k(n), \varepsilon\right) \leq const. \times ||\Pi_n h_0 - h_0||^2_s = o(1)$. By letting $\max\left\{\eta_{0,n}, \bar{\delta}^2_{m,n}, \lambda_n\right\} / g\left(k(n), \varepsilon\right) = o(1)$, Theorem 3.1 is applicable hence $||\widehat{h}_n - h_0||_{L^2(f_{Y_2})} = o_p(1)$.

3.2 PSMD using large or infinite dimensional sieves

In this subsection we present consistency results for PSMD estimators using large or infinite dimensional sieves, depending on the properties of the penalty function.

3.2.1 Lower semicompact penalty

Theorem 3.2. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n > 0$, $\eta_n = O(\eta_{0,n})$, and assumptions 3.1, 3.2(b) and 3.3 hold. Suppose that $P()$ is lower semicompact and $E[\|m(X,h)\|^2_W]$ is lower semicontinuous on $(\mathcal{H}, \|\cdot\|_s)$. If*

$$
\max\left\{\eta_{0,n}, E\left[||m(X,\Pi_n h_0)||^2_W\right]\right\} = O\left(\lambda_n\right), \tag{13}
$$

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

The lower semicompact penalty implies that the effective parameter space, $\{h \in \mathcal{H} : P(h) \leq M_n\}$ with $M_n \nearrow \infty$ slowly, is compact in the $\|\cdot\|_s$ −topology, and hence converts an ill-posed problem to a well-posed one.[14] Theorem 3.2 applies to the class of PSMD estimators with any positive lower semicompact penalty functions, allowing for $k(n) = \infty$ or $k(n)/n \to const. \geq 0$. To apply this theorem, it suffices to choose the penalization parameter $\lambda_n > 0$ to ensure restriction (13).

NPIV example (2): For this model with identity weighting $W = I$, $E[\|m(X,h)\|_W^2]$ is obviously lower semicontinuous on $(\mathcal{H}, \|\cdot\|_s)$ with a 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 $\theta \geq 0$. For a penalty function $P(h)$ to be lower semicompact, it suffices that the embedding of the set $\{h \in \mathcal{H} : P(h) \leq M\}$ into $(\mathcal{H}, || \cdot ||_s)$ is compact for all $M \in [0, \infty)$. For example, if $|| \cdot ||_s = || \cdot ||_{L^2(f_{Y_2})}$ then $P(h) = ||(1 + | \cdot |^2)^{-\vartheta/2} h(\cdot)||^p_{W^\alpha_p(\mathcal{R}^d)}$ with $0 < p \leq 2$, $\alpha > \frac{d}{p} - \frac{d}{2}$, $\vartheta \geq 0$, $f_{Y_2}(y_2)|y_2|^\vartheta \to 0$ as $|y_2| \to \infty$ will yield the desired result. 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 $P(h) = ||(1 + | \cdot |^2)^{-\vartheta/2} h(\cdot)||^p_{W^\alpha_p(\mathcal{R}^d)}$ with $0 < p < \infty$, $\alpha > \frac{d}{p}$, $\theta > \vartheta$ are lower semicompact; see Edmunds and Triebel (1996). 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 lower semicompact and convex, under assumption 3.1(ii) (identification), the PSMD estimator $\hat{h}_n$ using finite dimensional linear sieves (8) $\mathcal{H}_{k(n)}$ is equivalent to the original SMD estimator using finite dimensional compact sieves $\{h \in \mathcal{H}_{k(n)} : \widehat{P}_n(h) \leq M_n\}$:*

$$\hat{h}_n = \arg \inf_{h \in \mathcal{H}_{k(n)} : \widehat{P}_n(h) \leq M_n} \frac{1}{n} \sum_{i=1}^n \|\widehat{m}(X_i, h)\|^2_{\widehat{W}} \quad \textit{with} \quad M_n \to \infty \textit{ slowly.}$$

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 (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

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 $\mathbf{H}^*$. 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^α_p for $1 < p < \infty$ are reflexive and separable Banach spaces.

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

Assumption 3.4. *(i) 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 $k \geq 1$; (ii) $(\mathbf{H}, ||\cdot||_s)$ is a reflexive Banach space; $\mathcal{H}$ is a closed, bounded and convex subset in $(\mathbf{H}, ||\cdot||_s)$.*

Assumption 3.4(i) 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; see, e.g., Eggermont and LaRiccia (2001). In assumption 3.4(ii) the condition that $\mathcal{H}$ is bounded in $||\cdot||_s$ (i.e., $\sup_{h\in\mathcal{H}} ||h||_s \leq K < \infty$) is implied by the so-called *coercive* condition: $E\left[\|m(X,h)\|_W^2\right] + \lambda P(h) \to +\infty$ *as* $||h||_s \to +\infty$ *for* $\lambda \in (0,1]$.

A functional $G : \mathcal{H} \to (-\infty, +\infty)$ is *weak sequentially lower semicontinuous* at $h \in \mathcal{H}$ iff $G(h) \leq \liminf_{j\to\infty} G(h_j)$ for each sequence $\{h_j\}$ in $\mathcal{H}$ that converges weakly to h.

Theorem 3.3. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n > 0$, $\eta_n = O(\eta_{0,n})$, and assumptions 3.1, 3.2(c), 3.3 and 3.4 hold. Let $E\left[\|m(X,h)\|_W^2\right]$ be weak sequentially lower semicontinuous on $\mathcal{H}$. If*

$$\max\left\{\eta_{0,n}, E\left[||m(X, \Pi_n h_0)||_W^2\right]\right\} = o\left(\lambda_n\right), \tag{14}$$

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

Remark 3.2. *Under assumption 3.4(ii), $E\left[||m(X,h)||_W^2\right]$ is weak sequentially lower semicontinuous on $\mathcal{H}$ if either (1) or (2) or (3) holds: (1) $E\left[||m(X,\cdot)||_W^2\right]$ is convex and lower semicontinuous on $(\mathcal{H}, \|\cdot\|_s)$; or (2) $E\left[||m(X,\cdot)||_W^2\right] : \mathcal{H} \to [0,\infty)$ has compact Gateaux derivative on $\mathcal{H}$; or (3) $\sqrt{W(\cdot)}m(\cdot,h) : \mathcal{H} \to L^2(f_X)$ is compact and Frechet differentiable.*

NPIV example (2): For this model with $W = I$, the assumption that $\mathbf{H}$ is reflexive rules out the (weighted) sup-norm case; but assumption 3.4(ii) is readily satisfied by $\mathbf{H} = L^2(f_{Y_2})$, $||\cdot||_s = ||\cdot||_{L^2(f_{Y_2})}$ and $\mathcal{H} = \{h \in L^2(f_{Y_2}) : ||h||_{L^2(f_{Y_2})} \leq M < \infty\}$. $E\left[\|m(X,h)\|_W^2\right] = E\left[(E[Y_1 - h(Y_2)|X])^2\right]$ is convex and lower semicontinuous on $(\mathcal{H}, \|\cdot\|_s)$ and hence is weak sequentially lower semicontinuous on $\mathcal{H}$. Let $P(h) = ||h||_{L^2(f_{Y_2})}^2$ be the penalty function, then assumption 3.4(i) is satisfied with $t_0 = 2h_0$. Theorem 3.3 immediately leads to $||\widehat{h}_n - h_0||_{L^2(f_{Y_2})} = o_p(1)$.

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, Theorem 3.2 for $\hat{h}_n$ using a lower semicompact penalty allows for consistency under sup-norm and mild restriction (13) on smoothing parameters, while Theorem 3.3 for $\hat{h}_n$ using a convex penalty does not. Therefore, if one has some prior information about smoothness of h_0 in the sense that $P(h_0) < \infty$ and the set $\{h \in \mathcal{H} : P(h) \leq M\}$ is compact in $(\mathcal{H}, \|\cdot\|_s)$ for all $M \in [0,\infty)$, then one should apply either a PSMD procedure using large dimensional sieves with a lower semicompact penalty, or a PSMD procedure using slowly growing finite dimensional sieves $\{h \in \mathcal{H}_{k(n)} : P(h) \leq M_n\}$ with $k(n)$, $M_n \nearrow \infty$ slowly.

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

$$\mathcal{H}_{os} \equiv \{h \in \mathcal{H} : ||h - h_0||_s \leq \epsilon, ||h||_s \leq M_1, \lambda_n P(h) \leq \lambda_n M_0\} \quad \text{and } \mathcal{H}_{osn} \equiv \mathcal{H}_{os} \cap \mathcal{H}_n \tag{15}$$

for some positive finite constants M_1, M_0, and a sufficiently small positive ϵ such that $\Pr(\hat{h}_n \notin \mathcal{H}_{os}) < \epsilon$. 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 could be weaker than $||\cdot||_s$. Define the first pathwise derivative in the direction $[h - h_0]$ evaluated at h_0 as

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

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

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

Assumption 4.1. ***(local curvature)*** *(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 $||h - h_0||^2 \leq c_1 E[\|m(X, h)\|_W^2]$ holds for all $h \in \mathcal{H}_{osn}$; and $c_2 E[\|m(X, \Pi_n h_0)\|_W^2] \leq ||\Pi_n h_0 - h_0||^2$.*

Assumption 4.1(i) implies that the pseudo-metric $||h - h_0||$ is well-defined in $\mathcal{H}_{os}$ and is weaker than $||h - h_0||_s$. For example, let $W(X) = I$, then $||h - h_0|| = \sqrt{E[(E[h(Y_2) - h_0(Y_2)|X])^2]}$ for the NPIV model (2) and $||h - h_0|| = \sqrt{E[(E[f_{Y_1|Y_2,X}(h_0(Y_2))\{h(Y_2) - h_0(Y_2)\}|X])^2]}$ for the NPQIV model (3). Both are weaker than the root mean squared metric $||h - h_0||_s = \sqrt{E[\{h(Y_2) - h_0(Y_2)\}^2]}$ and the sup-norm metric $||h - h_0||_s = \sup_y |h(y) - h_0(y)|$. Assumption 4.1(ii) implies that the weaker pseudo-metric $||h - h_0||$ is Lipschitz continuous with respect to the population criterion function $E[\|m(X, h)\|_W^2]$ for all $h \in \mathcal{H}_{osn}$. It restricts local curvature of the criterion function, and is automatically satisfied by linear problems (such as the NPIV model). Assumption 4.1 enables us to obtain fast convergence rate of $||\widehat{h} - h_0||$ even when the convergence rate in the strong metric $||\cdot||_s$ could be very slow. Previously, AC used this insight to establish root-n asymptotic normality and efficiency of their SMD estimator of finite dimensional parameter θ_0 for the semi/nonparametric conditional moment restrictions $E[\rho(Y,X_z;\theta_0,h_0(\cdot))|X]=0$. Here we shall use the same trick to drive the nonparametric convergence rate of $||\widehat{h}-h_0||_s$.[15]

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

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

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:

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

We note that τ_n is a direct extension of BCK's sieve measure of ill-posedness,

$$\tau_n=\sup_{h\in\mathcal{H}_n:||h-\Pi_n h_0||\neq 0}\frac{\sqrt{E\left[\{h(Y_2)-\Pi_n h_0(Y_2)\}^2\right]}}{\sqrt{E[\{E[h(Y_2)-\Pi_n h_0(Y_2)|X]\}^2]}}\quad\text{for the NPIV model (2)},$$

to the general nonlinear nonparametric conditional moment model (4). 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) $\omega_n(\delta,\mathcal{H}_{osn})\le\tau_n\times\delta$ and $\omega_n(\delta,\mathcal{H}_{osn})\le\omega(\delta,\mathcal{H}_{os})$;

(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 $||\cdot||_s$. Let $\{\delta_{P,n}\}_{n=1}^{\infty}$ be a sequence of positive real values such that $\delta_{P,n}=O(1)$ and $\sup_{h\in\mathcal{H}_{osn}}|\widehat{P}_n(h)-P(h)|=O_p(\delta_{P,n})$. Let $\{\delta^2_{m,n}\}_{n=1}^{\infty}$ be a sequence of positive real values such that $\delta^2_{m,n}=o(1)$ and $\frac{1}{n}\sum_{i=1}^n||\widehat{m}(X_i,h)||^2_{\widehat{W}}\ge const.E\left[||m(X,h)||^2_W\right]-O_p(\delta^2_{m,n})$ uniformly over $\mathcal{H}_{osn}$. By definition $\delta^2_{m,n}\le\bar{\delta}^2_{m,n}$. In fact we have $\delta^2_{m,n}=\eta_{0,n}$ for most commonly used nonparametric estimators $\widehat{m}()$.

[15] Recently, CCLN (2010) impose a stronger version of assumption 4.1 in local identification of h_0, and provide various sufficient conditions.

[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.

For example, Lemma C.2 in the appendix shows that the series LS estimator $\widehat{m}(X,h)$ defined in (11) satisfies $\delta^2_{m,n} = \eta_{0,n} = \max\{\frac{J_n}{n}, b^2_{m,J_n}\}$, where $\frac{J_n}{n}$ is the order of the variance and b_{m,J_n} is the order of the bias of the series LS estimator of $m(\cdot, h)$.

Theorem 4.1. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n \geq 0$, $\eta_n = O(\eta_{0,n})$ and $||\hat{h}_n - h_0||_s = o_p(1)$. Let $h_0 \in \mathcal{H}_{os}$, $\Pi_n h_0 \in \mathcal{H}_{osn}$, assumptions 3.1, 3.2, 3.3 with $\eta_{0,n} = O(\delta^2_{m,n})$, and 4.1 hold. Suppose either condition (1) or (2) or (3) holds:*

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

(2) $\max\{\delta^2_{m,n}, \lambda_n\} = \delta^2_{m,n} = O(\lambda_n)$ *and $P()$ is lower semicompact.*

(3) $\max\left\{\delta^2_{m,n}, \lambda_n \delta_{P,n}, \lambda_n ||\widehat{h}_n - \Pi_n h_0||_s\right\} = O_p(\delta^2_{m,n})$ *and 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 $h \in \mathcal{H}_{osn}$.*

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

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”. We note that condition (3) controls the linear approximation of the penalty function around $\Pi_n h_0$, which is similar to assumption 3.4(i). It is satisfied when the penalty $P(h)$ is convex in $\Pi_n h_0$. 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 using 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 (2) to the general model (4). It allows for any sieve approximation error rates and other nonparametric estimators of $m(X,h)$ (beyond the series LS estimator (11)). It leads to convergence rates in any Banach norm $||\cdot||_s$ (besides the rate in the root mean squared metric).

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

5 Convergence Rates in a Hilbert Norm

To apply the general convergence rate theorem 4.1, one needs to compute upper bounds on the sieve approximation error $||h_0 - \Pi_n h_0||_s$, the sieve modulus of continuity $\omega_n(\delta, \mathcal{H}_{osn})$ (the sieve measure of local ill-posedness τ_n), or the modulus of continuity $\omega(\delta, \mathcal{H}_{os})$. In this section we provide sufficient conditions to bound these terms, which then lead to more concrete convergence rate results.

Throughout this section, we assume that $\mathcal{H}_{os}$ (given in (15)) is an infinite dimensional subset of a real-valued 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 the online supplemental material 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 using slowly growing finite dimensional sieves

Assumption 5.1. ***(sieve approximation error)*** *$||h_0 - \sum_{j=1}^{k(n)} \langle h_0, q_j\rangle_s q_j||_s = O(\{\nu_{k(n)}\}^{-\alpha})$ for a finite $\alpha > 0$ and a positive sequence $\{\nu_j\}_{j=1}^{\infty}$ that strictly increases to ∞ as $j \to \infty$.*

Assumption 5.2. ***(sieve link condition)*** *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) $||\Pi_n h_0 - h_0||^2 \leq C\sum_j \varphi(\nu_j^{-2})|\langle \Pi_n h_0 - h_0, q_j\rangle_s|^2$.*

Assumption 5.1 is a very mild condition about the smoothness of $h_0 \in \mathcal{H}_{os}$, it suggests that $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ is a natural sieve to approximate h_0. For example, if $(\mathbf{H}, ||\cdot||_s) = (L^2([0,1]^d, leb), ||\cdot||_{L^2(leb)})$ and $h_0 \in W_2^{\alpha}([0,1]^d, leb)$, then assumption 5.1 is satisfied with spline, wavelet, power series and Fourier series bases, and $\nu_{k(n)} = \{k(n)\}^{1/d}$. Assumption 5.2(i) relates the weak pseudo-metric $||h||$ to the strong norm in a sieve shrinking neighborhood $\mathcal{H}_{osn}$ (of h_0). It implies that the sieve modulus of continuity $\omega_n(\delta, \mathcal{H}_{osn})$ is bounded above by $const. \times \delta/\sqrt{\varphi(\nu_{k(n)}^{-2})}$ and that the sieve measure of (local) ill-posedness $\tau_n \leq const./\sqrt{\varphi(\nu_{k(n)}^{-2})}$ (see Lemma B.2). 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 (2), BCK and CR present conditions that imply assumptions 5.2(i) and (ii). See subsection 5.3 below for further discussion.

Theorem 4.1 and Lemma B.2 together 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 $\max\{\delta^2_{m,n}, \lambda_n\} = \delta^2_{m,n} = const. \times \frac{k(n)}{n} = o(1)$*. Then:*

$$||\hat{h}_n - h_0||_s = O_p\left(\left\{\nu_{k(n)}\right\}^{-\alpha} + \sqrt{\frac{k(n)}{n \times \varphi(\nu^{-2}_{k(n)})}}\right) = O_p\left(\left\{\nu_{k_o(n)}\right\}^{-\alpha}\right)$$

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

(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* $k_o(n) = c[\ln(n)]^{d/\varsigma}$ *for some* $c \in (0, 1)$.

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) ($\sqrt{\frac{k(n)}{n}\{\varphi(\nu^{-2}_{k(n)})\}^{-1}}$), and let $\max\{\delta^2_{m,n}, \lambda_n\} = \delta^2_{m,n} = const. \times \frac{k(n)}{n}$. For example, if the PSMD estimator $\hat{h}_n$ is computed using the series LS estimator $\widehat{m}(X, h)$ defined in (11), one can let $\delta^2_{m,n} = \eta_{0,n} = \max\{\frac{J_n}{n}, b^2_{m,J_n}\} = \frac{J_n}{n} = const. \times \frac{k(n)}{n} = o(1)$ (by Lemma C.2). This corollary extends the rate results of BCK for the NPIV model (2) to the general model (4), allowing for more general parameter space $\mathcal{H}$ and other nonparametric estimators of $m(X, h)$.

5.2 PSMD using large or infinite dimensional sieves

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

Assumption 5.4. *(**link condition over** $\mathcal{H}_{os}$) There are finite constants* $c, C > 0$ *and a continuous increasing function* $\varphi : \mathcal{R}_+ \to \mathcal{R}_+$ *such that: (i)* $||h||^2 \geq c\sum^{\infty}_{j=1}\varphi(\nu^{-2}_j)|\langle h, q_j\rangle_s|^2$ *for all* $h \in \mathcal{H}_{os}$*; (ii)* $||h - h_0||^2 \leq C\sum^{\infty}_{j=1}\varphi(\nu^{-2}_j)|\langle h - h_0, q_j\rangle_s|^2$ *for all* $h \in \mathcal{H}_{os}$.

Assumption 5.3 obviously implies assumption 5.1. Assumption 5.3 is automatically satisfied if either $\mathcal{H}_{os} \subseteq \mathcal{H}_{ellipsoid} \equiv \left\{h = \sum^{\infty}_{j=1}\langle h, q_j\rangle_s q_j : \sum^{\infty}_{j=1}\nu^{2\alpha}_j|\langle h, q_j\rangle_s|^2 \leq M^2\right\}$ or $\mathcal{H}_{os} \subseteq \mathcal{H}_{hyperrec} \equiv \left\{h = \sum^{\infty}_{j=1}\langle h, q_j\rangle_s q_j : |\langle h, q_j\rangle_s| \leq M'\nu_j^{-(\alpha+\frac{1}{2})},\ \inf_j \nu_j/j > 0\right\}$. Both $\mathcal{H}_{ellipsoid}$ and $\mathcal{H}_{hyperrec}$ are smooth function classes that are widely used in nonparametric estimation. Given our definition of $\mathcal{H}_{os}$ in (15), assumption 5.3 is also satisfied if the penalty function is such that $P(h) \geq \sum^{\infty}_{j=1}\nu^{2\alpha}_j|\langle h, q_j\rangle_s|^2$ for all $h \in \mathcal{H}_{os}$. Assumptions 5.4(i) and (ii) obviously imply assumptions 5.2(i) and (ii) respectively. Assumptions 5.3 and 5.4(i) together provide a upper bound on the modulus of continuity $\omega_n(\delta, \mathcal{H}_{os})$ (see Lemma B.3). Various versions of assumptions 5.3 and 5.4 have been imposed in the literature on minimax optimal rates for linear ill-posed inverse problems. See subsection 5.3 below for further discussion.

Theorem 4.1, Lemmas B.2 and B.3 together 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 uniformly over $\mathcal{H}_{osn}$. By definition $\delta^{*2}_{m,n} \le \delta^2_{m,n}$.

Corollary 5.2. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n > 0$, $\lambda_n = o(1)$, and all the assumptions of Theorem 4.1(1) hold. Let assumptions 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 $\infty \ge k(n) \ge k^*$, where $k^* = k^*(\delta^*_{m,n})$ is such that $\{\nu_{k^*}\}^{-2\alpha} \asymp \delta^{*2}_{m,n}\{\varphi(\nu^{-2}_{k^*})\}^{-1}$. Let either condition (2) of Theorem 4.1 hold with $\lambda_n = O(\delta^{*2}_{m,n})$, or condition (3) of Theorem 4.1 hold with $\lambda_n = O\left(\delta^*_{m,n}\sqrt{\varphi(\nu^{-2}_{k^*})}\right)$. Then:*

$$(1)\quad ||\hat{h}_n - h_0||_s = O_p\left(\{\nu_{k^*}\}^{-\alpha}\right) = O_p\left(\delta^*_{m,n}\{\varphi(\nu^{-2}_{k^*})\}^{-\frac{1}{2}}\right);$$

thus $||\hat{h}_n - h_0||_s = O_p\left((\delta^*_{m,n})^{\frac{\alpha}{\alpha+\varsigma}}\right)$ if $\varphi(\tau) = \tau^\varsigma$ for some $\varsigma \ge 0$; 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.

5.2.1 PSMD with large dimensional sieves and a series LS estimator of $m(X,h)$

The next rate result is applicable to the PSMD estimator using a series LS estimator of $m(X,h)$, and hence $\delta^{*2}_{m,n} = \frac{J^*_n}{n} \asymp b^2_{m,J^*_n}$ where J^*_n is such that the variance part $(\frac{J^*_n}{n})$ and the squared bias part (b^2_{m,J^*_n}) are of the same order.

Corollary 5.3. *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 satisfying assumptions C.1 and C.2. Let assumption 5.4 and all the assumptions of Theorem 4.1(2) hold with $c_2 E[||m(X,h)||^2_W] \le ||h-h_0||^2$ for all $h \in \mathcal{H}_{os}$. Let either $P(h) \ge \sum_{j=1}^{\infty}\nu_j^{2\alpha}|\langle h, q_j\rangle_s|^2$ for all $h \in \mathcal{H}_{os}$ or $\mathcal{H}_{os} \subseteq \mathcal{H}_{ellipsoid}$. Let $\lambda_n = O(\frac{J^*_n}{n})$, where $J^*_n \le k(n) \le \infty$ and is such that $\frac{J^*_n}{n} \asymp b^2_{m,J^*_n} \le const.\{\nu_{J^*_n}\}^{-2\alpha}\varphi(\nu^{-2}_{J^*_n})$. Then:*

$$||\hat{h}_n - h_0||_s = O_p\left(\{\nu_{J^*_n}\}^{-\alpha}\right) = O_p\left(\sqrt{\frac{J^*_n}{n \times \varphi(\nu^{-2}_{J^*_n})}}\right).$$

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 $\varsigma \ge 0$ and $\nu_k \asymp k^{1/d}$; and $||\widehat{h}_n - h_0||_s = O_p\left([\ln(n)]^{-\alpha/\varsigma}\right)$ if $\varphi(\tau) = \exp\{-\tau^{-\varsigma/2}\}$ for some $\varsigma > 0$ and $\nu_k \asymp k^{1/d}$, $J^*_n = c[\ln(n)]^{d/\varsigma}$ for some $c \in (0,1)$.

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 $T_{h_0} \equiv \sqrt{W(\cdot)}\frac{dm(\cdot,h_0)}{dh}$:

$\mathcal{H}_{os} \subset \mathbf{H} \to L^2(f_X)$, 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 $||h||^2 \equiv ||T_{h_0} h||^2_{L^2(f_X)} = ||(T^*_{h_0} T_{h_0})^{1/2} h||^2_s$. 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* $||(T^*_{h_0} T_{h_0})^{1/2} h||^2_s \asymp \sum_{j=1}^{\infty} \varphi(\nu_j^{-2}) |\langle h, q_j \rangle_s|^2$ *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 (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. In the following we mention some sufficient conditions for assumption 5.4.

A (nonlinear) operator $A : \mathcal{H} \to L^2(f_X)$ is *compact* iff it is continuous and maps bounded sets in $\mathcal{H}$ into relatively compact sets in $L^2(f_X)$. Suppose that T_{h_0} is a compact operator, which is a mild condition (for example, T_{h_0} is compact if $\sqrt{W(\cdot)} m(\cdot, h) : \mathcal{H} \subseteq \mathbf{H} \to L^2(f_X)$ 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 (e.g., $(T^*_{h_0} T_{h_0})^{1/2} \phi_{1k} = \mu_k \phi_{1k}$ for all k). Suppose that $T^*_{h_0} T_{h_0}$ is non-singular (i.e., T_{h_0} is injective), which is satisfied under the global identification condition (assumption 3.1(i)(ii)) and $c_2 E[\|m(X,h)\|^2_W] \leq ||h - h_0||^2$ for all $h \in \mathcal{H}_{os}$. Then the eigenfunction sequence $\{\phi_{1k}()\}_{k=1}^{\infty}$ is an orthonormal basis (hence a Riesz basis) for $\mathcal{H}_{os}$, and $||(T^*_{h_0} T_{h_0})^{1/2} h||^2_s = \sum_{k=1}^{\infty} \mu_k^2 |\langle h, \phi_{1k} \rangle_s|^2$ for all $h \in \mathcal{H}_{os}$. Thus, assumption 5.4 is automatically satisfied with $q_j = \phi_{1j}$ and $\varphi(\nu_j^{-2}) = \mu_j^2$ for all j. Following the proof of Lemma 1 in BCK, we can show that the sieve measure of local ill-posedness $\tau_n = [\mu_{k(n)}]^{-1} = [\varphi(\nu_{k(n)}^{-2})]^{-1/2}$ and that the sieve modulus of continuity $\omega_n(\delta, \mathcal{H}_{osn}) = \delta \tau_n = \delta [\mu_{k(n)}]^{-1}$.

In the numerical analysis literature on ill-posed inverse problems with *known* operators, it is common to measure the smoothness of $h_0 \in \mathcal{H}_{os}$ 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 increasing function ψ with $\psi(0) = 0$ such that $h_0 \in \mathcal{H}_{source} \equiv \left\{ h = \psi(T^*_{h_0} T_{h_0}) v : v \in \mathbf{H}, ||v||^2_s \leq M^2 \right\}$ 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)). When T_{h_0} is compact with a singular value system $\{\mu_j; \phi_{1j}, \phi_{0j}\}_{j=1}^{\infty}$, this general source condition becomes:

$$h_0 \in \mathcal{H}_{source} = \left\{ h = \sum_{j=1}^{\infty} \langle h, \phi_{1j} \rangle_s \phi_{1j} : \quad \sum_{j=1}^{\infty} \frac{\langle h, \phi_{1j} \rangle_s^2}{\psi^2(\mu_j^2)} \leq M^2 \right\}, \tag{16}$$

which is a particular Sobolev ellipsoid class of functions $\mathcal{H}_{ellipsoid}$. Therefore, the general source condition implies our assumptions 5.4 and 5.3 by setting $q_j = \phi_{1j}$, $\varphi(\nu_j^{-2}) = \mu_j^2$ and $\psi(\mu_j^2) = \nu_j^{-\alpha}$

[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.

for all $j \geq 1$. Then $\varphi(\tau) = \tau^{\varsigma}$ (mildly ill-posed case) is equivalent to $\psi(\eta) = \eta^{\alpha/(2\varsigma)}$; $\varphi(\tau) = \exp\{-\tau^{-\varsigma/2}\}$ (severely ill-posed case) is equivalent to $\psi(\eta) = [-\log(\eta)]^{-\alpha/\varsigma}$.

The above discussion and Corollaries 5.1 and 5.3 immediately imply the following rate results.

Remark 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 with $c_2 E[\|m(X,h)\|_W^2] \leq ||h - h_0||^2$ for all $h \in \mathcal{H}_{os}$. Let $T_{h_0} \equiv \sqrt{W(\cdot)}\frac{dm(\cdot,h_0)}{dh} : \mathcal{H}_{os} \subset \mathbf{H} \to L^2(f_X)$ be a compact operator with a singular value decomposition $\{\mu_j; \phi_{1j}, \phi_{0j}\}_{j=1}^{\infty}$. Let $\mathcal{H}_n = clsp\{\phi_{1j} : j = 1, ..., k(n)\}$ for $k(n) \leq \infty$.*

(1) (sieve dominating case) Let $h_0 \in \mathcal{H}_{source}$. If $\max\{\delta_{m,n}^2, \lambda_n\} = \delta_{m,n}^2 = const. \times \frac{k(n)}{n} = o(1)$, then:

$$||\hat{h}_n - h_0||_s = O_p\left(\psi(\mu_{k(n)+1}^2) + \sqrt{\frac{k(n)}{n \times \mu_{k(n)}^2}}\right).$$

(2) (penalty dominating case) Let $\widehat{m}(X,h)$ be the series LS estimator satisfying assumptions C.1 and C.2. Let either $P(h) \geq \sum_{j=1}^{\infty}\{\psi(\mu_j^2)\}^{-2}|\langle h, \phi_{1j}\rangle_s|^2$ for all $h \in \mathcal{H}_{os}$ or $\mathcal{H}_{os} \subseteq \mathcal{H}_{source}$. Let $0 < \lambda_n = O(\frac{J_n^*}{n}) = o(1)$, where $J_n^* \leq k(n) \leq \infty$ and is such that $\frac{J_n^*}{n} \asymp b_{m,J_n^*}^2 \leq const.\{\psi(\mu_{J_n^*}^2)\}^{-2}\mu_{J_n^*}^2$. Then:

$$||\hat{h}_n - h_0||_s = O_p\left(\psi(\mu_{J_n^*}^2)\right) = O_p\left(\sqrt{\frac{J_n^*}{n \times \mu_{J_n^*}^2}}\right).$$

Note that applications of Corollaries 5.1 and 5.3 do not require knowledge of the singular value decomposition $\{\mu_j; \phi_{1j}, \phi_{0j}\}_{j=1}^{\infty}$ of the injective, compact derivative operator T_{h_0}, but applications of the rate results stated in Remark 5.1 do. In particular, Result (1) of Remark 5.1 is applicable only when the eigenfunction sequence $\{\phi_{1j} : j = 1, ..., k(n)\}$ is used as the sieve basis to construct the PSMD estimator; Result (2) is applicable if the choice of penalty satisfies $P(h_0) < \infty$ and $P(h) \geq \sum_{j=1}^{\infty}\{\psi(\mu_j^2)\}^{-2}|\langle h, \phi_{1j}\rangle_s|^2$ for all $h \in \mathcal{H}_{os}$.

Remark 5.2. ***(1)*** *Suppose that $q_j = \phi_{1j}$ (assumption 5.4 holds), $\varphi(\nu_j^{-2}) = \mu_j^2 \geq const.j^{-\varsigma}$, $\varsigma > 1$ (mildly ill-posed case), and $\mathcal{H}_{os} = \left\{h = \sum_{j=1}^{\infty}\langle h, \phi_{1j}\rangle_s \phi_{1j} : |\langle h, \phi_{1j}\rangle_s| \leq M' j^{-(\alpha + \frac{1}{2})}\right\}$, $\alpha > 0$ (assumption 5.3 holds), HH establish that their kernel based function space TR-MD estimator of the NPIV model (2) achieves the minimax lower bound in the metric $||\cdot||_s = ||\cdot||_{L^2(f_{Y_2})}$ (for $2\alpha + 1 > \varsigma \geq \alpha$); and HL extend their result to the NPQIV model (3) (for $2\alpha > \varsigma \geq \alpha > \frac{1}{2}$).*

(2) *For the NPIV model (2), under assumptions 5.3 and 5.4(ii), CR establish the minimax lower bound in the metric $||\cdot||_s = ||\cdot||_{L^2(f_{Y_2})}$:*

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

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. The rates stated in Corollaries 5.1 and 5.3 for the PSMD estimators of the general model (4) achieve the minimax lower bound of CR. Note that our rate results allow for both mildly ill-posed and severely ill-posed cases.*

6 Application to Nonparametric Additive Quantile IV Regression

In this section we present an application of the PSMD estimation of the nonparametric additive quantile IV regression model:

$$Y_3 = h_{01}(Y_1) + h_{02}(Y_2) + U, \quad \Pr(U \leq 0|X) = \gamma, \tag{17}$$

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 $\gamma \in (0,1)$. To map into the general model (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 concreteness and illustration, we let the support of $Y = (Y_1, Y_2, Y_3)'$ be $\mathcal{Y} = [0,1]^d \times [0,1]^d \times \mathcal{R}$, and the support of X be $\mathcal{X} = [0,1]^{d_x}$ with $d_x \geq d \geq 1$. We estimate $h_0 = (h_{01}, h_{02}) \in \mathcal{H} = \mathcal{H}^1 \times \mathcal{H}^2$ using the PSMD estimator $\widehat{h}_n$ given in (6), with $\widehat{W} = W = I$ (identity), $\mathcal{H}_n = \mathcal{H}_n^1 \times \mathcal{H}_n^2$ being either a finite dimensional ($\dim(\mathcal{H}_n) \equiv k(n) = k_1(n) + k_2(n) < \infty$) or an infinite dimensional ($k(n) = \infty$) *linear* sieve, and $\widehat{P}_n(h) = P(h) \geq 0$. The conditional mean function $m(X, h)$ is estimated by the series LS estimator $\widehat{m}(X, h)$ defined in (11). 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 P-spline (polynomial spline), B-spline, wavelet, or Fourier 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 P-spline, B-spline, wavelet, Fourier series, or power series basis for $L^2(\mathcal{X}, leb.)$, with $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_{d_x}$. Clearly the number of terms in the tensor-product sieve $p^{J_n}(X)$ is given by $J_n = \prod_{j=1}^{d_x} J_{j,n}$. See Newey (1997), Huang (1998) and Chen (2007) for details about tensor-product linear sieves. We assume:

Condition 6.1. *(i) $\{(Y_i', X_i')\}_{i=1}^n$ is a random sample from a probability density $f_{Y,X}$ on $\mathcal{Y} \times \mathcal{X}$, and $0 < \inf_{x\in\mathcal{X}} f_X(x) < \sup_{x\in\mathcal{X}} f_X(x) < \infty$; (ii) The smallest eigenvalue of $E\left[p^{J_n}(X)p^{J_n}(X)'\right]$ is bounded away from zero uniformly in J_n; where $p^{J_n}(X)$ is a tensor product P-spline, B-spline, wavelet, cosine sieve with $J_n^2 = o(n)$; (iii) $E[F_{Y_3|Y_1,Y_2,X}(h_1(Y_1) + h_2(Y_2))|X = \cdot] \in \Lambda_c^{\alpha_m}([0,1]^{d_x})$ with $\alpha_m > 0$ for all $h \in \mathcal{H}_{k(n)}^{M_0}$; (iv) $f_{Y_3|(Y_1,Y_2,X)=(y_1,y_2,x)}(y_3)$ 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.*

Condition 6.1(i)(ii)(iii) implies that the series LS estimator $\widehat{m}(\cdot, h)$ satisfies assumption 3.3 with $\eta_{0,n} = \delta_{m,n}^2 = \max\{\frac{J_n}{n}, J_n^{-2\alpha_m/d_x}\}$ and $\bar{\delta}_{m,n}^2 = o(1)$ (by Lemma C.2). Condition 6.1(iv) implies that

$E\{[m(X,h)]^2\}$ is continuous on $(\mathcal{H}, \|\cdot\|_{\sup})$, $\|h\|_{\sup} = \sup_{y_1}|h_1(y_1)| + \sup_{y_2}|h_2(y_2)|$, and provides sufficient condition to bound $E\{[m(X,\Pi_n h_0)]^2\}$ (assumption 3.1(iv)).

In the following we denote $h_0(y_1,y_2) = h_{01}(y_1) + h_{02}(y_2)$, $\Delta h(y_1,y_2) = h(y_1,y_2) - h_0(y_1,y_2) = \Delta h_1(y_1) + \Delta h_2(y_2)$, and for $l = 1,2$,

$$K_{l,h}[\Delta h_l](X) \equiv E\left(\left\{\int_0^1 f_{Y_3|Y_1Y_2X}(h_0(Y_1,Y_2) + t\Delta h(Y_1,Y_2))dt\right\}\Delta h_l(Y_l)|X\right).$$

Condition 6.2. *(i) $\mathcal{H} = \mathcal{H}^1 \times \mathcal{H}^2$ with $\mathcal{H}^l = \Lambda^{\alpha_l}([0,1]^d)$ for $\alpha_l > 0$; (ii) for any $h \in \mathcal{H}$, Range$(K_{1,h}) \cap$ Range$(K_{2,h}) = \{0\}$; and $K_{l,h}[\Delta h_l](X) = 0$ a.s.-$\mathcal{X}$ implies $\Delta h_l = 0$ a.s.-$\mathcal{Y}_l$ for $l = 1,2$; (iii) $\mathcal{H}_n = \mathcal{H}_n^1 \times \mathcal{H}_n^2$, where $\mathcal{H}_n^l$ is a tensor product P-spline, B-spline, wavelet or cosine series closed linear subspace of $\mathcal{H}^l$ for $l = 1,2$.*

Conditions 6.2(i) and (iii) specify the function space and the sieve space for $h = (h_1,h_2)$ respectively. Condition 6.2(ii) is a global identification condition (assumption 3.1(ii)), which extends the identification condition for the NPQIV model (3) of CH to the nonparametric additive quantile IV model (17). See CH, CIN and CCLN for sufficient conditions for identification.

Denote $r_m \equiv \alpha_m/d_x$ and $r_l \equiv \alpha_l/d$ for $l = 1,2$. The following consistency result is a simple application of Theorem 3.2.

Proposition 6.1. *For the model (17), let $\hat{h}_n$ be the PSMD estimator with $\lambda_n > 0$, $\eta_n = O(\lambda_n) = o(1)$, and $\widehat{m}(X,h)$ be the series LS estimator. Let conditions 6.1 and 6.2 hold, $P(h) = ||h_1||_{\Lambda^{\alpha_1}} + ||h_2||_{\Lambda^{\alpha_2}}$ and $\max\left\{[k_1(n)]^{-2r_1}, [k_2(n)]^{-2r_2}, \frac{J_n}{n} + J_n^{-2r_m}\right\} = O(\lambda_n)$. Then:*

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

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 $P(\widehat{h}_{1,n}) + P(\widehat{h}_{2,n}) = O_p(1)$.

We now turn to the calculation of the convergence rate of our PSMD estimator. For the model (17), let $||h||_s^2 = E\{[h_1(Y_1)]^2\} + E\{[h_2(Y_2)]^2\}$, then $||h||_s^2 \le \|h\|_{\sup}^2$ for all $h \in \mathcal{H}$. The above consistency results immediately imply that $||\widehat{h}_n - h_0||_s = o_P(1)$. Let $\mathcal{H}_{os} = \{h = (h_1,h_2) \in \mathcal{H} : \|h - h_0\|_{\sup} = o(1), P(h) \le c\}$. For $h = (h_1,h_2) \in \mathcal{H}_{os}$, and $l = 1,2$, denote

$$T_{l,0}[h_l - h_{0l}](X) \equiv E\left(f_{Y_3|Y_1,Y_2,X}(h_{01}(Y_1) + h_{02}(Y_2))[h_l(Y_l) - h_{0l}(Y_l)]|X\right),$$

and

$$T_{h_0}[h - h_0](X) = T_{1,0}[h_1 - h_{01}](X) + T_{2,0}[h_2 - h_{02}](X).$$

Condition 6.3. *(i) $||T_{h_0}[h - h_0]||_{L^2(f_X)} \asymp ||K_{1,h}[h_1 - h_{01}] + K_{2,h}[h_2 - h_{02}]||_{L^2(f_X)}$ for all $h = (h_1,h_2) \in \mathcal{H}_{os} \cap \mathcal{H}_n$; Range$(T_{1,0}) \cap$ Range$(T_{2,0}) = \{0\}$ and $T_{l,0}[\Delta h_l](X) = 0$ a.s.-$\mathcal{X}$ implies $\Delta h_l = 0$ a.s.-$\mathcal{Y}_l$ for $l = 1,2$; (ii) there is a continuous increasing function $\varphi \ge 0$ such that*

$$||T_{h_0}[h - h_0]||_{L^2(f_X)}^2 \asymp \sum_{j=1}^{\infty}\varphi(j^{-2/d})\left(\langle h_1 - h_{01}, q_{1,j}\rangle_{L^2(f_{Y_1})}^2 + \langle h_2 - h_{02}, q_{2,j}\rangle_{L^2(f_{Y_2})}^2\right)$$

for all $h = (h_1,h_2) \in \mathcal{H}_{os} \cap \mathcal{H}_n$.

Condition 6.3(i) implies assumption 4.1 (local curvature), and Condition 6.3(ii) implies assumption 5.2. Applying Corollary 5.1, we obtain the following convergence rate for the PSMD estimator using slowly growing finite dimensional sieves. Denote $\alpha = \min\{\alpha_1, \alpha_2\}$.

Proposition 6.2. *For the model (17), let all the conditions of Proposition 6.1 and condition 6.3 hold. Let $\alpha > d$. If $\max\{\frac{J_n}{n}, J_n^{-2r_m}, \lambda_n\} = \frac{J_n}{n} = const. \times \frac{k(n)}{n} = o(1)$, $k(n) = k_1(n) + k_2(n)$ and $k_1(n) \asymp k_2(n) \to \infty$. Then:*

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

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 $k(n) \asymp n^{\frac{d}{2(\alpha+\varsigma)+d}}$; and $||\widehat{h}_n - h_0||_s = O_p\left([\ln(n)]^{-\alpha/\varsigma}\right)$ if $\varphi(\tau) = \exp\{-\tau^{-\varsigma/2}\}$ for some $\varsigma > 0$ and $k(n) = c[\ln(n)]^{d/\varsigma}$ for some $c \in (0,1)$.

When Y_1 and Y_2 are measurable functions of X, we have $\varphi([k(n)]^{-2/d}) = const.$ in Proposition 6.2. 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 Lee (2005) and Horowitz and Mammen (2007). See the working paper version (Chen and Pouzo, 2008) for additional consistency and convergence rate results, in which the support of Y_2 could be unbounded, h_{02} could belong to a function space $\mathcal{H}^2$ different from the Holder space $\Lambda^{\alpha_2}([0,1]^d)$, and $P(h)$ could take other functional forms as well.

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 (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 (6), with $\widehat{m}(X,h)$ being the series LS estimator (11) of $m(X,h)$, $\widehat{W} = W = I$ (identity), 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 (6): 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 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 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 of the quantiles. 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 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 2 shows the corresponding estimated curves and

[18] In the working paper version (Chen and Pouzo, 2008) 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 $k(n)/n \to const. > 0$). It was too time consuming to compute the MC results for this case and the results were not very stable either.

MC 95% confidence bands. In Table 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.

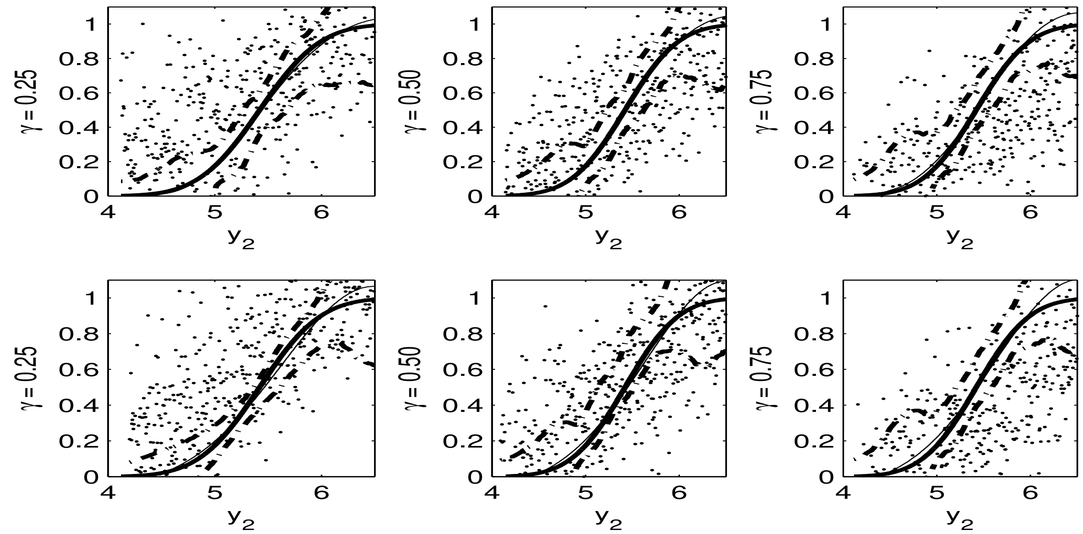

Figure 1: h_0 (solid thick), $\hat{h}_n$ (solid thin), MC confidence bands (dashed), a sample of Y_1 (dots), $\hat{P}(h) = ||\nabla h||^2_{L^2}$, 1st row: $k(n) = 8, \lambda_n = 0.003, J_n = 16$; 2nd row: $k(n) = 16, \lambda_n = 0.006, J_n = 16$.

Table 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

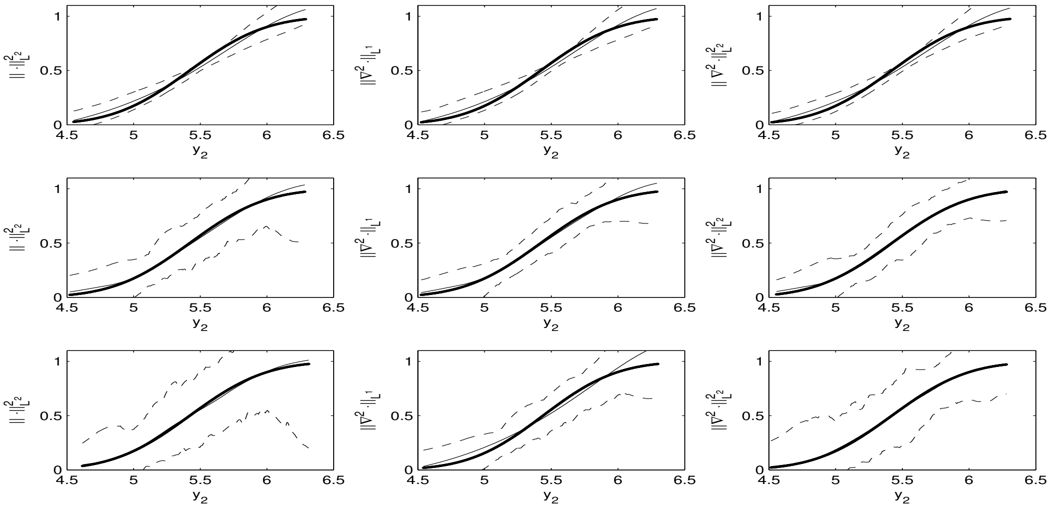

Figure 2: Table 1 experiments. 1st row: $k(n) = 6, \lambda_n = 0.00001, J_n = 16$. 2nd row: $k(n) = 8, \lambda_n = 0.0001, J_n = 16$. 3nd row: $k(n) = 16, \lambda_n = 0.0005, J_n = 16$.

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 (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 and $\widehat{W} = W = I$ (identity). 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", using P-Spline(2,5) as $\mathcal{H}_n$ and P-Spline(5,10) for $\hat{m}$ (hence $k(n) = 8$, $J_n = 16$). Due to the lack of space, in Figure 3 we report the estimated quantile IV Engel curves only for three different quantiles $\gamma = \{0.25, 0.50, 0.75\}$ and for four goods that has been considered in BCK.[19] Figure 3 presents the estimated Engel curves using $\widehat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\widehat{\mu})}$

[19] The results on all seven goods are available upon request from the authors.

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".

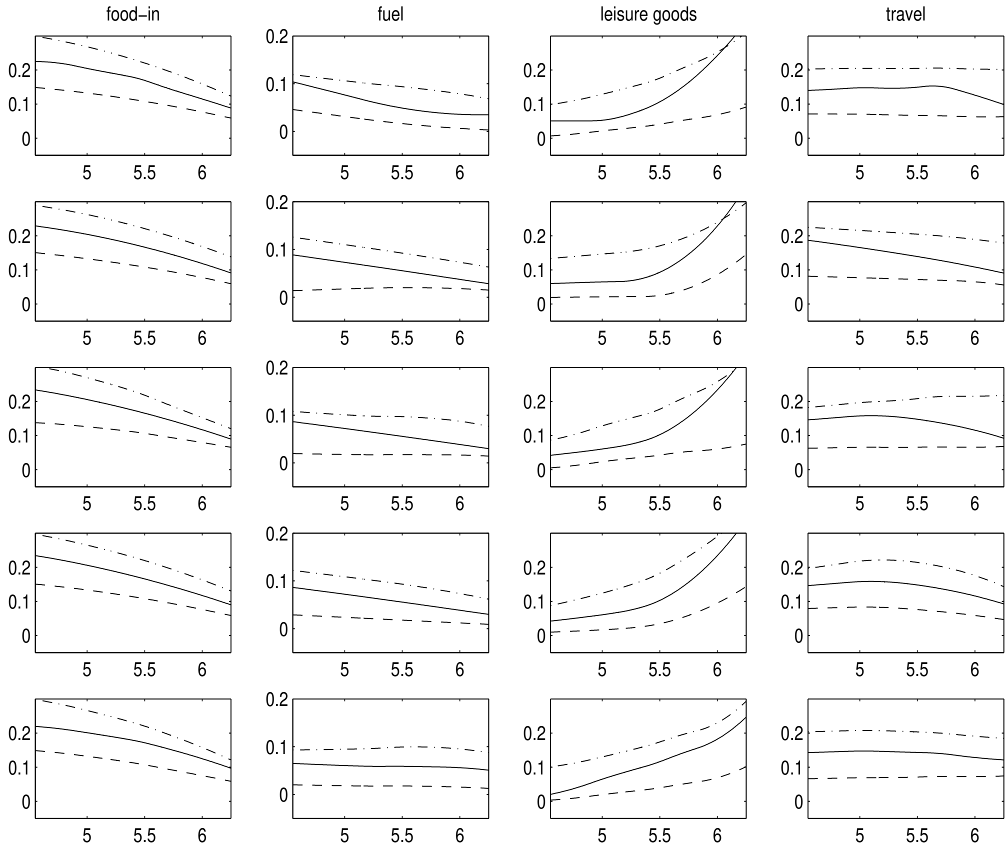

Figure 3: Engel curves for quantiles $\gamma = 0.25$ (dash), 0.50 (solid), 0.75 (dot-dash). $k(n) = 8, J_n = 16$ (all rows). $\hat{P}_n(h) = ||\nabla^2 h||^2_{L^2(d\hat{\mu})}$ with $\lambda_n = 0.001$ (1st row); $\hat{P}_n(h) = ||\nabla^2 h||_{L^1(d\hat{\mu})}$ with $\lambda_n = 0.001$ (2nd row); $\hat{P}_n(h) = ||\nabla h||^2_{L^2(d\hat{\mu})}$ 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).

8 Conclusion

In this paper, we propose the PSMD estimation of conditional moment restrictions 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 establish 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 show that the convergence rate under a Hilbert space norm coincide with the known minimax optimal rate for the NPIV model (2). We illustrate the general theory with a nonparametric additive quantile IV regression. We also present a simulation study and estimate 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 (2009a), we consider 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), and 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 optimal rate of convergence for $\widehat{h}_n - h_0$. In Chen and Pouzo (2010), we provide inference and limiting distributions of plug-in PSMD estimators of possibly irregular functionals of (θ_0, h_0).

A Additional Results for 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.

In the following we let $(\mathbf{A}, \mathcal{T})$ be a Hausdorff topological space and $\mathcal{B}_{\mathcal{T}}(a)$ be a non-empty open neighborhood (under $\mathcal{T}$) around $a \in \mathcal{A} \subseteq \mathbf{A}$. Let $\Pr^*$ denote the outer measure associated with Pr. Let o_{p^*} and O_{p^*} respectively denote convergence in probability under $\Pr^*$ and bounded in probability under $\Pr^*$.

Lemma A.1. *Let $\widehat{\alpha}_n$ be such that $\widehat{Q}_n(\widehat{\alpha}_n) \leq \inf_{\alpha \in \mathcal{A}_{k(n)}} \widehat{Q}_n(\alpha) + O_{p^*}(\eta_n)$, where $\{\eta_n\}_{n=1}^{\infty}$ is a positive real-valued sequence such that $\eta_n = o(1)$. Let $\overline{Q}_n() : \mathcal{A} \to [0, \infty)$ be a sequence of non-random measurable functions and the following conditions (A.1.1) - (A.1.4) hold:*

(A.1.1) (i) $0 \leq \overline{Q}_n(\alpha_0) = o(1)$; (ii) there is a positive function $g_0(n, k, \mathcal{B})$ such that:

$$\inf_{\alpha \in \mathcal{A}_k : \alpha \notin \mathcal{B}_{\mathcal{T}}(\alpha_0)} \overline{Q}_n(\alpha) \geq g_0(n, k, \mathcal{B}) > 0 \quad \textit{for each } n \geq 1, k \geq 1,$$

and $\liminf_{n\to\infty} g_0(n, k(n), \mathcal{B}) \geq 0$ for all $\mathcal{B}_{\mathcal{T}}(\alpha_0)$.

(A.1.2) (i) $\mathcal{A} \subseteq \mathbf{A}$ and $(\mathbf{A}, \mathcal{T})$ is a Hausdorff topological space; (ii) $\mathcal{A}_k \subseteq \mathcal{A}_{k+1} \subseteq \mathcal{A}$ for all $k \geq 1$, and there is a sequence $\{\Pi_n \alpha_0 \in \mathcal{A}_{k(n)}\}$ such that $\overline{Q}_n(\Pi_n \alpha_0) = o(1)$.

(A.1.3) $\widehat{Q}_n(\alpha)$ is jointly measurable in the data $\{(Y_i', X_i')\}_{i=1}^n$ and the parameter $\alpha \in \mathcal{A}_{k(n)}$.

(A.1.4) (i) $\widehat{Q}_n(\Pi_n\alpha_0) \leq K_0\overline{Q}_n(\Pi_n\alpha_0) + O_{p^*}(c_{0,n})$ for some $c_{0,n} = o(1)$ and a finite constant $K_0 > 0$; (ii) $\widehat{Q}_n(\alpha) \geq K\overline{Q}_n(\alpha) - O_{p^*}(c_n)$ uniformly over $\alpha \in \mathcal{A}_{k(n)}$ for some $c_n = o(1)$ and a finite constant $K > 0$; (iii) $\max\left\{c_{0,n}, c_n, \overline{Q}_n(\Pi_n\alpha_0), \eta_n\right\} = o(g_0(n, k(n), \mathcal{B}))$ for all $\mathcal{B}_{\mathcal{T}}(\alpha_0)$.

Then: for all $\mathcal{B}_{\mathcal{T}}(\alpha_0)$, $\Pr^*(\hat{\alpha}_n \notin \mathcal{B}_{\mathcal{T}}(\alpha_0)) \to 0$ as $n \to \infty$.

In the online supplemental material we present another consistency lemma for penalized sieve extremum estimators, which is a special case of Lemma A.1, but is still general enough for most applications.

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) $\lim_{j\to\infty}\langle v, \alpha_j\rangle_{\mathbf{A}^*,\mathbf{A}} = \langle v, \alpha\rangle_{\mathbf{A}^*,\mathbf{A}}$ 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; is *lower semicontinuous on* $(\mathcal{A}, \|\cdot\|_s)$ iff the set $\{\alpha \in \mathcal{A} : F(\alpha) \leq M\}$ is closed under $||\cdot||_s$ for all $M \in [0, \infty)$.

Remark A.1. ***(1)*** *Let $(\mathbf{A}, \mathcal{T})$ be a Hausdorff topological space and $\mathcal{A}_k$ be non-empty for each k. Condition (A.1.3) is satisfied and $\widehat{\alpha}_n$ is measurable 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 (A.1.3) is satisfied and $\widehat{\alpha}_n$ is measurable 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).*

Give Remark A.1, in the rest of the paper we will assume that $\widehat{\alpha}_n$ and our approximate PSMD estimator $\widehat{h}_n$ defined in (6) are measurable.

Lemma A.2. *Let $\hat{h}_n$ be the (approximate) PSMD estimator (6). Then: $\widehat{h}_n \in \mathcal{H}_n$ wpa1.*

Further, let assumption 3.3(i) hold with $\eta_n = O(\eta_{0,n})$.

(1) If assumption 3.2(b) and $\max\left\{\eta_{0,n}, E\left[||m(X, \Pi_n h_0)||_W^2\right]\right\} = O(\lambda_n)$ hold, then $P(\widehat{h}_n) = O_p(1)$.

(2) If assumption 3.2(c) and $\max\left\{\eta_{0,n}, E\left[||m(X,\Pi_n h_0)||_W^2\right]\right\} = o(\lambda_n)$ *hold, then* $P(\widehat{h}_n) \leq P(h_0) + o_p(1) = O_p(1)$.

Recall that $\mathcal{H}_{k(n)}^{M_0} \equiv \{h \in \mathcal{H}_{k(n)} : \lambda_n P(h) \leq \lambda_n M_0\}$ for a large but finite $M_0 \equiv M_0(\varepsilon) \in (0, \infty)$ such that $\Pi_n h_0 \in \mathcal{H}_{k(n)}^{M_0}$ and that for all $\varepsilon > 0$, $\Pr\left(\hat{h}_n \notin \mathcal{H}_{k(n)}^{M_0}\right) < \varepsilon$ for all sufficiently large n, where the bound $M_0 \equiv M_0(\varepsilon)$ in $\mathcal{H}_{k(n)}^{M_0}$ can depend on $\varepsilon > 0$ but *not* on n. Given assumptions 3.2 and 3.3(i) and Lemma A.2, such a M_0 always exists. In the following we denote $\mathcal{B}_{\mathcal{T}}(h_0)$ as any open neighborhood in a topological space $(\mathcal{H}, \mathcal{T})$ around h_0.

Lemma A.3. *Let* $\hat{h}_n$ *be the (approximate) PSMD estimator with* $\lambda_n \geq 0$, $\eta_n = O(\eta_{0,n})$ *and assumption 3.3 hold. Let assumption 3.1(iii) hold and the* $\mathcal{T}-$*topology could be the norm* $||\cdot||_s-$*topology or weaker ones. Then, for all* $\mathcal{B}_{\mathcal{T}}(h_0)$ *and all* $\varepsilon > 0$,

(1) under assumption 3.2(b) and $\max\{\eta_{0,n}, E[||m(X,\Pi_n h_0)||_W^2]\} = O(\lambda_n)$,

$$\begin{aligned} & \Pr\left(\widehat{h}_n \notin \mathcal{B}_{\mathcal{T}}(h_0)\right) \\ \leq \; & \Pr\left(\inf_{h \in \mathcal{H}_{k(n)}^{M_0} : h \notin \mathcal{B}_{\mathcal{T}}(h_0)} \left\{cE\left[||m(X,h)||_W^2\right] + \lambda_n P(h)\right\} \leq O_p\left(\bar{\delta}_{m,n}^2\right) + \lambda_n P(h_0) + O_p(\lambda_n)\right) + \varepsilon \end{aligned}$$

for all n *sufficiently large, where the bound* $M_0 \equiv M_0(\varepsilon)$ *in* $\mathcal{H}_{k(n)}^{M_0}$ *can depend on* $\varepsilon > 0$.

(2) under assumption 3.2(c) and $\max\{\eta_{0,n}, E[||m(X,\Pi_n h_0)||_W^2]\} = o(\lambda_n)$,

$$\begin{aligned} & \Pr\left(\widehat{h}_n \notin \mathcal{B}_{\mathcal{T}}(h_0)\right) \\ \leq \; & \Pr\left(\inf_{h \in \mathcal{H}_{k(n)}^{M_0} : h \notin \mathcal{B}_{\mathcal{T}}(h_0)} \left\{cE\left[||m(X,h)||_W^2\right] + \lambda_n P(h)\right\} \leq O_p\left(\bar{\delta}_{m,n}^2\right) + \lambda_n P(h_0) + o_p(\lambda_n)\right) + \varepsilon \end{aligned}$$

for all n *sufficiently large, where the bound* $M_0 \equiv M_0(\varepsilon)$ *in* $\mathcal{H}_{k(n)}^{M_0}$ *can depend on* $\varepsilon > 0$.

Identification via strictly convex penalty. When $E[\|m(X,h)\|_W^2]$ is convex in $h \in \mathcal{H}$ (e.g. the NPIV model), we can relax the global identification condition (assumption 3.1(ii)) by using a strictly convex penalty function, that is, we can use a strictly convex penalty to select one h_0 out of the solution set $\mathcal{M}_0 \equiv \{h \in \mathcal{H} : E[\|m(X,h)\|_W^2] = 0\}$ uniquely.

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

Theorem A.1. *Suppose that* $\mathcal{M}_0$ *is non-empty,* P *is strictly convex and lower semicontinuous on* $(\mathcal{M}_0, \|\cdot\|_s)$, *and* $E\left[\|m(X,h)\|_W^2\right]$ *is convex and lower semicontinuous on* $(\mathcal{H}, \|\cdot\|_s)$.

(1) If assumptions 3.4(ii) holds, then: $\mathcal{M}_0^P = \{h_0\} \subseteq \mathcal{M}_0$.

(2) Let $\hat{h}_n$ *be the PSMD estimator with* $\lambda_n > 0$, $\eta_n = O(\eta_{0,n})$ *and assumptions 3.1(i)(iii)(iv), 3.2(c), 3.3 and 3.4 hold. Suppose that for any* $k \geq 1$, $\mathcal{H}_k$ *is convex,* $P(\cdot)$ *is convex and lower semicontinuous on* $(\mathcal{H}_k, \|\cdot\|_s)$. *If* $\max\{\eta_{0,n}, E\left[||m(X,\Pi_n h_0)||_W^2\right], \bar{\delta}_{m,n}^2\} = o(\lambda_n)$, *then:* $||\widehat{h}_n - h_0||_s = o_p(1)$, *and* $P(\widehat{h}_n) = P(h_0) + o_p(1)$.

B Lemmas for Convergence Rate

Lemma B.1. *Suppose that all the conditions of Theorem 4.1(1) hold. Then:*
(1) $||\hat{h}_n - \Pi_n h_0|| = O_p\left(\max\{\delta_{m,n}, \sqrt{\lambda_n \delta_{P,n}}, \sqrt{\lambda_n |P(\hat{h}_n) - P(\Pi_n h_0)|}, ||\Pi_n h_0 - h_0||\}\right)$.
(2) Under assumption 3.2(c), $||\hat{h}_n - \Pi_n h_0|| = O_p\left(\max\{\delta_{m,n}, o(\sqrt{\lambda_n}), ||\Pi_n h_0 - h_0||\}\right)$.
(3) Under condition (3), $||\hat{h}_n - \Pi_n h_0|| = O_p\left(\max\{\delta_{m,n}, \sqrt{\lambda_n \delta_{P,n}}, \sqrt{\lambda_n ||\hat{h}_n - \Pi_n h_0||_s}, ||\Pi_n h_0 - h_0||\}\right)$.

Lemma B.2. *Let* $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$ *and* $\{q_j\}_{j=1}^{\infty}$ *be a Riesz basis for* $(\mathbf{H}, ||\cdot||_s)$.

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

Lemma B.3. *Let assumptions 5.3 and 5.4(i) hold. Then: for small* $\delta > 0$, *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* $\mathcal{H}_n = clsp\{q_1, ..., q_{k(n)}\}$.

C Lemmas for Series LS estimator $\widehat{m}()$ of $m()$

Under the following two mild assumptions, we show that the series LS estimator $\widehat{m}(X, h)$ defined in (11) satisfies assumption 3.3 with $\eta_{0,n} = \delta_{m,n}^2 = \max\{\frac{J_n}{n}, b_{m,J_n}^2\}$, where $\frac{J_n}{n}$ is the order of the variance and b_{m,J_n} is the order of the bias of the series LS estimator of $m(\cdot, h)$.

Assumption C.1. *(i)* $\{(Y_i', X_i')\}_{i=1}^n$ *is a random sample from the distribution of* (Y', X')*; (ii)* $\mathcal{X}$ *is a compact connected subset of* $\mathcal{R}^{d_x}$ *with Lipschitz continuous boundary, and* f_X *is bounded and bounded away from zero over* $\mathcal{X}$*; (iii)* $\max_{1\leq j\leq J_n} E[|p_j(X)|^2] \leq const.$*; the smallest eigenvalue of* $E\left[p^{J_n}(X)p^{J_n}(X)'\right]$ *is bounded away from zero for all* J_n*; (iv) either* $\xi_n^2 J_n = o(n)$ *with* $\xi_n \equiv \sup_{X\in\mathcal{X}} \left\|p^{J_n}(X)\right\|_I$*, or* $J_n \log(J_n) = o(n)$ *for* $p^{J_n}(X)$ *a polynomial spline sieve; (v) there are finite constants* $K, K' > 0$ *such that* $KI \leq W(x) \leq K'I$ *for all* $x \in \mathcal{X}$*;* $\widehat{W}(X)$ *is positive definite for almost all* $X \in \mathcal{X}$*;* $\sup_{x\in\mathcal{X}} \left\|\widehat{W}(x) - W(x)\right\|_{tr} = o_p(1)$.

Let $N_{[]}(\epsilon, \mathcal{F}_n, ||.||_{L^2(f_Z)})$ be the $L^2(f_Z)-$covering number with bracketing of a class of functions $\mathcal{F}_n$. For $j = 1, ..., J_n$, denote $\mathcal{O}_{jn} \equiv \{p_j(\cdot)\rho(\cdot, h) : h \in \mathcal{H}_{k(n)}^{M_0}\}$ and $\mathcal{O}_{ojn} \equiv \{p_j(\cdot)\rho(\cdot, h) : h \in \mathcal{H}_{osn}\}$. Denote

$$\mathcal{C}_n(j) \equiv \int_0^1 \sqrt{1 + \log N_{[]}(w, \mathcal{O}_{jn}, ||.||_{L^2(f_Z)})}dw,\ \mathcal{C}_{on}(j) \equiv \int_0^1 \sqrt{1 + \log N_{[]}(w, \mathcal{O}_{ojn}, ||.||_{L^2(f_Z)})}dw.$$

Assumption C.2. *(i) There are a sequence of measurable functions* $\{\bar{\rho}_n(Z)\}_{n=1}^{\infty}$ *and a finite constant* $K > 0$*, such that* $\sup_{h\in\mathcal{H}_{k(n)}^{M_0}} |\rho(Z, h)| \leq \bar{\rho}_n(Z)$ *and* $E[\bar{\rho}_n(Z)^2|X] \leq K$*; (ii) there is* $p^{J_n}(X)'\pi$

such that $E\{[m(X,h)-p^{J_n}(X)'\pi]^2\} = O(b^2_{m,J_n})$ *uniformly over* $h \in \mathcal{H}^{M_0}_{k(n)}$*; (iii)* $\max_{1\le j\le J_n} \mathcal{C}_n(j) \le \sqrt{C_n} < \infty$ *and* $\frac{J_n}{n}C_n = o(1)$*; (iv)* $\max_{1\le j\le J_n} \mathcal{C}_{on}(j) \le \sqrt{C} < \infty$.

In assumption C.1, 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 C.2(ii) is satisfied by typical smooth function classes of $\{m(\cdot,h) : h \in \mathcal{H}^{M_0}_{k(n)}\}$ and typical linear sieves $p^{J_n}(X)$. For example, if $\{m(\cdot,h) : h \in \mathcal{H}^{M_0}_{k(n)}\}$ is a subset of a Hölder ball (denoted as $\Lambda^{\alpha_m}_c(\mathcal{X})$), then assumption C.2(ii) holds for tensor product polynomial splines, wavelets or Fourier series sieves with $b_{m,J_n} = J_n^{-r_m}$ where $r_m = \alpha_m/d_x$.

The following remark is a special case of lemma 4.2(i) of Chen (2007), which is derived in the proof of theorem 3 in Chen, Linton and van Keilegom (2003). Let $D_{\mathcal{T}}$ denote the distance generated by the topology $\mathcal{T}$ (on $\mathcal{H}$) such that $\mathcal{H}^{M_0}_{k(n)}$ is totally bounded under $D_{\mathcal{T}}$.

Remark C.1. *Suppose that there are finite constants* $\kappa \in (0,1]$, $K > 0$ *such that*

$$\max_{1\le j\le J_n} E\left[[p_j(X)]^2 \sup_{h'\in\mathcal{H}^{M_0}_{k(n)}:D_{\mathcal{T}}(h',h)\le\delta} \left|\rho(Z,h')-\rho(Z,h)\right|^2\right] \le K^2\delta^{2\kappa} \tag{18}$$

for all $h \in \mathcal{H}^{M_0}_{k(n)}$ *and all positive value* $\delta = o(1)$*. Then:*

(1) $\max_{1\le j\le J_n} N_{[]}(\epsilon, \mathcal{O}_{jn}, ||.||_{L^2(f_Z)}) \le N([\frac{\epsilon}{2K}]^{1/\kappa}, \mathcal{H}^{M_0}_{k(n)}, D_{\mathcal{T}})$*;*

(2) Assumption C.2(iii) is satisfied with $\int_0^1 \sqrt{1+\log N(w^{1/\kappa}, \mathcal{H}^{M_0}_{k(n)}, D_{\mathcal{T}})dw} \le \sqrt{C_n}$ *and* $\frac{J_n}{n}C_n = o(1)$*;*

(3) Assumption C.2(iv) is satisfied with $\int_0^1 \sqrt{1+\log N(w^{1/\kappa}, \mathcal{H}_{osn}, D_{\mathcal{T}})dw} \le \sqrt{C} < \infty$*.*

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

Lemma C.1. *Let* $\widehat{m}(.,h)$ *be the series LS estimator defined in (11) and assumption C.1 hold. (1) If* $Var[\rho(Z,\Pi_n h_0)|X] \le K$ *then*

$$\frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i,\Pi_n h_0) - \widetilde{m}(X_i,\Pi_n h_0)||^2_{\widehat{W}} = O_p\left(\frac{J_n}{n}\right).$$

(2) If assumption C.2(i)(iii) holds, then:

$$\sup_{h\in\mathcal{H}^{M_0}_{k(n)}} \frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i,h) - \widetilde{m}(X_i,h)||^2_{\widehat{W}} = O_p\left(\frac{J_n}{n}C_n\right) = o_p(1).$$

(3) If assumption C.2(i)(iv) holds, then:

$$\sup_{h\in\mathcal{H}_{osn}} \frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i,h) - \widetilde{m}(X_i,h)||^2_{\widehat{W}} = O_p\left(\frac{J_n}{n}\right) = o_p(1).$$

Lemma C.2. *Let $\widehat{m}(.,h)$ be the series LS estimator defined in (11) and assumption C.1 hold. (1) If assumption C.2(i)(ii) holds at $h = \Pi_n h_0$, then, with $\eta_{0,n} = \max\{\frac{J_n}{n}, b^2_{m,J_n}\}$,*

$$cE\left[||m(X,\Pi_n h_0)||^2_W\right] - O_p\left(\eta_{0,n}\right) \leq \frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,\Pi_n h_0)||^2_{\widehat{W}} \leq c'E\left[||m(X,\Pi_n h_0)||^2_W\right] + O_p\left(\eta_{0,n}\right).$$

(2) If assumption C.2(i)(ii)(iii) holds, then there are finite constants $K, K' > 0$ such that, with $\bar{\delta}^2_{m,n} = \frac{J_n}{n}C_n + b^2_{m,J_n} = o(1)$ and uniformly over $h \in \mathcal{H}^{M_0}_{k(n)}$,

$$KE\left[||m(X,h)||^2_W\right] - O_p\left(\bar{\delta}^2_{m,n}\right) \leq \frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)||^2_{\widehat{W}} \leq K'E\left[||m(X,h)||^2_W\right] + O_p\left(\bar{\delta}^2_{m,n}\right).$$

(3) If assumption C.2(i)(ii)(iv) holds, then there are finite constants $K, K' > 0$ such that, with $\delta^2_{m,n} = \eta_{0,n} = \max\{\frac{J_n}{n}, b^2_{m,J_n}\}$ and uniformly over $h \in \mathcal{H}_{osn}$,

$$KE\left[||m(X,h)||^2_W\right] - O_p\left(\delta^2_{m,n}\right) \leq \frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)||^2_{\widehat{W}} \leq K'E\left[||m(X,h)||^2_W\right] + O_p\left(\delta^2_{m,n}\right).$$

The next lemma is of independent interest. It is a version of Lemma A.1(1) in Chen and Pouzo (2009a), and our proof here corrects a typo in their proof. See Chen and Pouzo (2010) for a more general version and its applications to derive the convergence rates and limiting distributions of plug-in PSMD estimators of any functionals of h_0 satisfying $E[\rho(Y,X_z;h_0(\cdot))|X] = 0$.

Let $\{\delta_{s,n}\}_{n=1}^{\infty}$ be a sequence of positive real values such that $\delta_{s,n} = o(1)$, and

$$\mathcal{N}_{os} \equiv \{h \in \mathcal{H}_{os} : ||h - h_0||_s \leq M_0\delta_{s,n}\},$$

where M_0 is a finite but large number such that $\hat{h}_n \in \mathcal{N}_{os}$ for large n, with probability greater than $1-\epsilon$, for a small $\epsilon > 0$.

Lemma C.3. *Let $\widehat{m}(.,h)$ be the series LS estimator defined in (11) and assumption C.1 hold. Suppose the following condition hold:*

(C.3.1) (i) there are finite constants $\kappa \in (0,1]$, $K > 0$ such that

$$\max_{1\leq j\leq J_n} E\left[[p_j(X)]^2 \sup_{h'\in\mathcal{N}_{os}:||h'-h||_s\leq\delta}\left|\rho(Z,h') - \rho(Z,h)\right|^2\right] \leq K^2\delta^{2\kappa}$$

for all $h \in \mathcal{N}_{os}$ and all positive value $\delta = o(1)$; (ii) $\int_0^1 \sqrt{1+\log N(w^{1/\kappa}, \mathcal{N}_{os}, ||\cdot||_s)}dw \leq \sqrt{C} < \infty$.

$$\textit{Then: } \sup_{\mathcal{N}_{os}}\frac{1}{n}\sum_{i=1}^{n}\|\widehat{m}(X_i,h) - \widehat{m}(X_i,h_0) - \widetilde{m}(X_i,h)\|^2_{\widehat{W}} = O_P\left(\frac{J_n}{n}\left(\delta_{s,n}\right)^{2\kappa}\right).$$

References

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

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

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

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

[5] 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.

[6] 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.

[7] Chamberlain, G. (1992). Efficiency Bounds for Semiparametric Regression. *Econometrica*, **60**, 567-596.

[8] 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.

[9] Chen, X. and S. Ludvigson (2009). Land of Addicts? An Empirical Investigation of Habit-based Asset Pricing Models. *Journal of Applied Econometrics* **24**, 1057-1093.

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

[11] Chen, X. and D. Pouzo (2009a). Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals. *Journal of Econometrics* **152**, 46-60.

[12] Chen, X. and D. Pouzo (2009b). On Nonlinear Ill-posed Inverse Problems with Applications to Pricing of Defaultable Bonds and Option Pricing. *Science in China, Series A: Mathematics* **52**, 1157-1168.

[13] Chen, X. and D. Pouzo (2010). On Inference of PSMD Estimators of Nonparametric Conditional Moment Models. Mimeo, Yale University.

[14] Chen, X. and M. Reiss (2010). On Rate Optimality for Nonparametric Ill-posed Inverse Problems in Econometrics. Yale University, Cowles Foundation Discussion Paper No. 1626, forthcoming in *Econometric Theory*.

[15] Chen, X. and X. Shen (1998). Sieve extremum estimates for weakly dependent data. *Econometica* **66**, 289-314.

[16] Chen, X., V. Chernozhukov, S. Lee and W. Newey (2010). Identification in Semiparametric and Nonparametric Conditional Moment Models. Mimeo, Yale University.

[17] Chen, X., O. Linton and I. van Keilegom (2003). Estimation of Semiparametric Models when the Criterion Function is not Smooth. *Econometrica* **71**, 1591-1608.

[18] Chernozhukov, V. and C. Hansen (2005). An IV Model of Quantile Treatment Effects. *Econometrica* **73**, 245-61.

[19] Chernozhukov, V., P. Gagliardini, and O. Scaillet (2010). Nonparametric Instrumental Variable Estimation of Quantile Structural Effects. Mimeo, MIT, University of Lugano and Swiss Finance Institute.

[20] Chernozhukov, V., G. Imbens, and W. Newey (2007). Instrumental Variable Estimation of Nonseparable Models. *Journal of Econometrics* **139**, 4-14.

[21] Chesher, A. (2003). Identification in Nonseparable Models. *Econometrica* **71**, 1405-1441.

[22] Darolles, S., Y. Fan, J.-P. Florens and E. Renault (2010). *Nonparametric Instrumental Regression.* Mimeo, Toulouse School of Economics.

[23] D'Haultfoeuille, X. (2010). On the Completeness Condition in Nonparametric Instrumental Problems. Forthcoming in *Econometric Theory.*

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

[25] Eggermont, P.P.B. and V.N. LaRiccia (2001). *Maximum Penalized Likelihood Estimation*, Springer Series in Statistics.

[26] Engl, H., M. Hanke and A. Neubauer (1996). *Regularization of Inverse Problems*, Kluwer Academic Publishers: London.

[27] Florens, JP, J. Johannes and S. van Bellegem (2010). Identification and Estimation by Penalization in Nonparametric Instrumental Regression. Forthcoming in *Econometric Theory.*

[28] Gagliardini, P. and O. Scaillet (2010). Tikhonov Regularization for Nonparametric Instrumental Variable Estimators. Mimeo, University of Lugano and Swiss Finance Institute.

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

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

[31] Horowitz, J. and S. Lee (2005). Nonparametric Estimation of an Additive Quantile Regression Model. *Journal of the American Statistical Association* **100**, 1238-1249.

[32] Horowitz, J. and S. Lee (2007). Nonparametric Instrumental Variables Estimation of a Quantile Regression Model. *Econometrica* **75**, 1191-1208.

[33] Horowitz, J. and E. Mammen (2007). Rate-Optimal Estimation for a General Class of Nonparametric Regression Models with Unknown Link Functions. *Annals of Statistics* **35**, 2589-2619.

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

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

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

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

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

[39] Newey, W.K. (1997). Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics* **79**, 147-168.

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

[41] Pollard, D. (1990). *Empirical Processes: Theory and Applications.* NSF-CBMS Regional Conference Series in Probability and Statistics Vol. 2. Institute of Mathematical Statistics.

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

[43] Van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*, New York: Springer-Verlag.

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

[45] Zeidler, E. (1985). *Nonlinear Functional Analysis and its Applications III: Variational methods and optimization*, New York: Springer-Verlag.

Supplementary Material of "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals" by Xiaohong Chen and Demian Pouzo

In this document we first provide a brief summary of commonly used function spaces and sieve spaces. We then provide mathematical proofs of all the theorems, corollaries, propositions and lemmas that appear in the main text and the appendix.

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* $\mathcal{B}^{\nu}_{p,q}\left(\mathcal{R}^d\right)$ 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* $\mathcal{F}^{\nu}_{p,q}\left(\mathcal{R}^d\right)$ 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 $\mathcal{F}^{-\nu}_{p',q'}\left(\mathcal{R}^d\right)$ $(\mathcal{B}^{-\nu}_{p',q'}\left(\mathcal{R}^d\right))$ is the dual space of $\mathcal{F}^{\nu}_{p,q}\left(\mathcal{R}^d\right)$ $(\mathcal{B}^{\nu}_{p,q}\left(\mathcal{R}^d\right))$ with $1/p' + 1/p = 1$ and $1/q' + 1/q = 1$.

Let $\mathcal{T}^{\nu}_{p,q}\left(\mathcal{R}^d\right)$ denote either $\mathcal{B}^{\nu}_{p,q}\left(\mathcal{R}^d\right)$ or $\mathcal{F}^{\nu}_{p,q}\left(\mathcal{R}^d\right)$. Then $\mathcal{T}^{\nu}_{p,q}\left(\mathcal{R}^d\right)$ gets larger with increasing q (i.e., $\mathcal{T}^{\nu}_{p,q_1}\left(\mathcal{R}^d\right) \subseteq \mathcal{T}^{\nu}_{p,q_2}\left(\mathcal{R}^d\right)$ for $q_1 \leq q_2$), gets larger with decreasing p (i.e., $\mathcal{T}^{\nu}_{p_1,q}\left(\mathcal{R}^d\right) \subseteq \mathcal{T}^{\nu}_{p_2,q}\left(\mathcal{R}^d\right)$ for $p_1 \geq p_2$), and gets larger with decreasing ν (i.e., $\mathcal{T}^{\nu_1}_{p,q}\left(\mathcal{R}^d\right) \subseteq \mathcal{T}^{\nu_2}_{p,q}\left(\mathcal{R}^d\right)$ for $\nu_1 \geq \nu_2$). Also, $\mathcal{T}^{\nu}_{p,q}\left(\mathcal{R}^d\right)$ becomes a *Banach* space when $p, q \geq 1$. The spaces $\mathcal{T}^{\nu}_{p,q}\left(\mathcal{R}^d\right)$ include many well-known function spaces as special cases. For example, $L^p(\mathcal{R}^d, leb) = \mathcal{F}^0_{p,2}\left(\mathcal{R}^d\right)$ for $p \in (1, \infty)$; the *Hölder* space $\Lambda^r(\mathcal{R}^d) = \mathcal{B}^r_{\infty,\infty}\left(\mathcal{R}^d\right)$ 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 $||h||_{W_p^\nu} \equiv \left\| \left((1 + |\cdot|^2)^{\nu/2} \widehat{h}(\cdot) \right)^\vee \right\|_{L^p(leb)} < \infty$ (note that for $\nu > 0$, the norm $||h||_{W_p^{-\nu}}$ is a shrinkage in the Fourier domain).

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

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

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 $0 \leq k \leq \gamma$; (b) ψ and all its derivatives up to order γ decrease rapidly as $|y| \to \infty$; (c) $\{2^{k/2}\psi(2^k y - j) : k, j \in \mathbb{Z}\}$ forms a Riesz basis of $L^2(leb)$, that is, the linear span of $\{2^{k/2}\psi(2^k y - j) : k, j \in \mathbb{Z}\}$ is dense in $L^2(leb)$ and

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

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

Some examples 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 $k_0 \geq 0$, any function h in $L^2(leb)$ has the following wavelet $\gamma-$ regular multiresolution expansion:

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

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

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

Cardinal B-spline wavelets of order γ:

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

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

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

Polynomial splines of order q_n:

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

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 $k(n)-1$:

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

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

Consistency: Proof of Theorems

Proof of Theorem 3.1: Under the assumption that $E\left[m(X,h)'W(X)m(X,h)\right]$ is lower semicontinuous on finite dimensional closed and bounded sieve spaces $\mathcal{H}_k$, we have that for all $\varepsilon > 0$ and each fixed $k \geq 1$,

$$g\left(k,\varepsilon\right) \equiv \inf_{h\in\mathcal{H}_k^{M_0}:||h-h_0||_s\geq\varepsilon} E\left[\|m(X,h)\|_W^2\right] \geq \min_{h\in\mathcal{H}_k:||h-h_0||_s\geq\varepsilon} E\left[\|m(X,h)\|_W^2\right]$$

exists, and is strictly positive (under assumption 3.1(i)(ii)). Moreover, for fixed k, $g\left(k,\varepsilon\right)$ increases as ε increases. For any fixed $\varepsilon > 0$, $g\left(k,\varepsilon\right)$ decreases as k increases, and $g\left(k,\varepsilon\right)$ could go to zero as k goes to infinity. Following the proof of Lemma A.3(1) with $\mathcal{T} = ||\cdot||_s$ topology, $\mathcal{H}_{k(n)}^{M_0} \subseteq \mathcal{H}_{k(n)}$, $\lambda_n P(h) \geq 0$ and $\eta_n = O(\eta_{0,n})$, we have: for all $\varepsilon > 0$ and n sufficiently large,

$$\begin{array}{rl}
& \Pr\left(||\widehat{h}_n - h_0||_s \geq \varepsilon\right) \\
\leq & \Pr\left(||\widehat{h}_n - h_0||_s \geq \varepsilon, \widehat{h}_n \in \mathcal{H}_{k(n)}^{M_0}\right) + \varepsilon \\
\leq & \Pr\left(\begin{array}{c}\inf_{h\in\mathcal{H}_{k(n)}^{M_0}:||h-h_0||_s\geq\varepsilon}\left\{cE\left[||m(X,h)||_W^2\right] + \lambda_n P(h)\right\} \\ \leq c'E[||m(X,\Pi_n h_0)||_W^2] + O_p\left(\eta_{0,n}\right) + O_p\left(\bar{\delta}_{m,n}^2\right) + \lambda_n P(h_0) + O_p(\lambda_n)\end{array}\right) + \varepsilon \\
\leq & \Pr\left(\begin{array}{c}\inf_{h\in\mathcal{H}_{k(n)}^{M_0}:||h-h_0||_s\geq\varepsilon}\left\{cE\left[||m(X,h)||_W^2\right]\right\} \\ \leq c'E[||m(X,\Pi_n h_0)||_W^2] + O_p\left(\eta_{0,n}\right) + O_p\left(\bar{\delta}_{m,n}^2\right) + \lambda_n P(h_0) + O_p(\lambda_n)\end{array}\right) + \varepsilon \\
\leq & \Pr\left(g(k(n),\varepsilon) \leq O_p\left(\max\{\bar{\delta}_{m,n}^2, \eta_{0,n}, E(\|m(X,\Pi_n h_0)\|_W^2), \lambda_n\}\right)\right) + \varepsilon
\end{array}$$

which goes to zero under $\max\{\bar{\delta}^2_{m,n}, \eta_{0,n}, E(\|m(X,\Pi_n h_0)\|^2_W), \lambda_n\} = o\left(g(k(n),\varepsilon)\right)$. Thus $||\widehat{h}_n - h_0||_s = o_p(1)$. *Q.E.D.*

Proof of Theorem 3.2: Under the assumptions that $E\left[m(X,h)'W(X)m(X,h)\right]$ is lower semicontinuous and $P(h)$ is lower semicompact on $(\mathcal{H}, ||\cdot||_s)$, we have that for all $\varepsilon > 0$,

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

exists (by theorem 38.B in Zeidler (1985)) and is strictly positive (under assumption 3.1(i)(ii)) for $\mathcal{H}^M = \{h \in \mathcal{H} : P(h) \leq M\}$ with some large but finite $M \geq M_0$. By Lemma A.3(1) with $\mathcal{T} = ||\cdot||_s$ topology, $\mathcal{H}^{M_0}_{k(n)} \subseteq \mathcal{H}^M$, $\lambda_n > 0$, $P(h) \geq 0$, $\eta_n = O(\eta_{0,n})$ and $\max\{\eta_{0,n}, E[||m(X,\Pi_n h_0)||^2_W]\} = O(\lambda_n)$, we have: for all $\varepsilon > 0$ and n sufficiently large,

$$\begin{aligned}
& \Pr\left(||\widehat{h}_n - h_0||_s \geq \varepsilon\right) \\
\leq\; & \Pr\left(||\widehat{h}_n - h_0||_s \geq \varepsilon, \widehat{h}_n \in \mathcal{H}^{M_0}_{k(n)}\right) + \varepsilon \\
\leq\; & \Pr\left(\inf_{h\in\mathcal{H}^{M_0}_{k(n)}:||h-h_0||_s\geq\varepsilon} \left\{cE\left[||m(X,h)||^2_W\right] + \lambda_n P(h)\right\} \leq O_p\left(\bar{\delta}^2_{m,n}\right) + \lambda_n P(h_0) + O_p(\lambda_n)\right) + \varepsilon \\
\leq\; & \Pr\left(\inf_{h\in\mathcal{H}^M:||h-h_0||_s\geq\varepsilon} E\left[||m(X,h)||^2_W\right] \leq O_p\left(\bar{\delta}^2_{m,n}\right) + \lambda_n P(h_0) + O_p(\lambda_n)\right) + \varepsilon \\
\leq\; & \Pr\left(g(\varepsilon) \leq O_p\left(\max\{\bar{\delta}^2_{m,n}, \lambda_n\}\right)\right) + \varepsilon
\end{aligned}$$

which goes to zero under $\max\{\bar{\delta}^2_{m,n}, \lambda_n\} = o\left(1\right)$. Thus $||\widehat{h}_n - h_0||_s = o_p(1)$. *Q.E.D.*

Proof of Theorem 3.3: We divide the proof in two steps; first we show consistency under the weak topology; second we establish consistency under the strong norm.

Step 1 We can establish consistency in the weak topology by applying Lemma A.1, either verifying its conditions or following its proof directly. Under stated conditions, $\widehat{h}_n \in \mathcal{H}_{k(n)}$ with probability approaching one. By Lemma A.2(2) with $\max\{\eta_{0,n}, E[||m(X,\Pi_n h_0)||^2_W]\} = o(\lambda_n)$ and $\eta_n = O(\eta_{0,n})$, we have $P(\widehat{h}_n) - P(h_0) \leq o_p(1)$, thus we can focus on the set $\{h \in \mathcal{H}_{k(n)} : P(h) \leq M_0\} = \mathcal{H}^{M_0}_{k(n)}$ for all n large enough. Let $\mathcal{B}_w(h_0)$ denote any open neighborhood (in the weak topology) around h_0, and $\mathcal{B}^c_w(h_0)$ its complement (under the weak topology) in $\mathcal{H}$. By Lemma A.3(2) with $\mathcal{B}_{\mathcal{T}}(h_0) = \mathcal{B}_w(h_0)$, $\lambda_n P(h) \geq 0$, $\mathcal{H}^{M_0}_{k(n)} \subseteq \mathcal{H}$, $\eta_n = O(\eta_{0,n})$, and $\max\{\eta_{0,n}, E[||m(X,\Pi_n h_0)||^2_W]\} = o(\lambda_n)$, we have: for all non-empty open ball $\mathcal{B}_w(h_0)$, all $\varepsilon > 0$ and n sufficiently large,

$$\begin{aligned}
& \Pr\left(\widehat{h}_n \notin \mathcal{B}_w(h_0)\right) \\
\leq\; & \Pr\left(\widehat{h}_n \notin \mathcal{B}_w(h_0), \widehat{h} \in \mathcal{H}^{M_0}_{k(n)}\right) + \varepsilon \\
\leq\; & \Pr\left(\inf_{\mathcal{H}^{M_0}_{k(n)}:h\notin\mathcal{B}_w(h_0)} \left\{cE[||m(X,h)||^2_W] + \lambda_n P(h)\right\} \leq O_p\left(\bar{\delta}^2_{m,n}\right) + \lambda_n P(h_0) + o(\lambda_n)\right) + \varepsilon \\
\leq\; & \Pr\left(\inf_{\mathcal{H}:h\notin\mathcal{B}_w(h_0)} E[||m(X,h)||^2_W] \leq O_p\left(\max\{\bar{\delta}^2_{m,n}, \lambda_n\}\right)\right) + \varepsilon.
\end{aligned}$$

Let $E\left[\|m(X,h)\|_W^2\right]$ be weak sequentially lower semicontinuous on $\mathcal{H}$. Since $\mathcal{H}\cap\mathcal{B}_w^c(h_0)$ is weakly compact (weakly closed and bounded), by assumption 3.4(ii) and theorem 38.A in Zeidler (1985), there exists $h^*(\mathcal{B})\in\mathcal{H}\cap\mathcal{B}_w^c(h_0)$ such that $\inf_{\mathcal{H}:h\notin\mathcal{B}_w(h_0)}E[||m(X,h)||_W^2]=E[||m(X,h^*(\mathcal{B}))||_W^2]$. It must hold that $g(\mathcal{B})\equiv E[||m(X,h^*(\mathcal{B}))||_W^2]>0$; otherwise, by assumption 3.1(i)(ii) $||h^*(\mathcal{B})-h_0||_s=0$. But, if this is the case, then for any $t\in\mathbf{H}^*$ we have $|\langle t,h^*(\mathcal{B})-h_0\rangle_{\mathbf{H}^*,\mathbf{H}}|\leq const.\times||h^*(\mathcal{B})-h_0||_s=0$, a contradiction to the fact that $h^*(\mathcal{B})\notin\mathcal{B}_w(h_0)$. Thus

$$\Pr\left(\widehat{h}_n\notin\mathcal{B}_w(h_0),\widehat{h}_n\in\mathcal{H}_{k(n)}^{M_0}\right)\leq\Pr\left(E[||m(X,h^*(\mathcal{B}))||_W^2]\leq O_p\left(\max\{\bar{\delta}_{m,n}^2,\lambda_n\}\right)\right),$$

which goes to zero since $\max\{\bar{\delta}_{m,n}^2,\lambda_n\}=o(1)$. Hence $\Pr\left(\widehat{h}_n\notin\mathcal{B}_w(h_0)\right)\to 0$.

Step 2 Consistency under the weak topology implies that $\langle t_0,\widehat{h}_n-h_0\rangle_{\mathbf{H}^*,\mathbf{H}}=o_p(1)$. By assumption 3.4(i), $P(\widehat{h}_n)-P(h_0)\geq\langle t_0,\widehat{h}_n-h_0\rangle_{\mathbf{H}^*,\mathbf{H}}+g(||\widehat{h}_n-h_0||_s)$. Lemma A.2(2) implies that $P(\widehat{h}_n)-P(h_0)\leq o_p(1)$ under $\max\{\eta_{0,n},E[||m(X,\Pi_nh_0)||_W^2]\}=o(\lambda_n)$, $\eta_n=O(\eta_{0,n})$. Thus $g(||\widehat{h}_n-h_0||_s)=o_p(1)$, and $||\widehat{h}_n-h_0||_s=o_p(1)$ by our assumption over $g(.)$. This, $\langle t_0,\widehat{h}_n-h_0\rangle_{\mathbf{H}^*,\mathbf{H}}=o_p(1)$ and assumption 3.4(i) imply that $P(\widehat{h}_n)-P(h_0)\geq o_p(1)$. But $P(\widehat{h}_n)\leq P(h_0)+o_p(1)$ by Lemma A.2(2). Thus $P(\widehat{h}_n)-P(h_0)=o_p(1)$.

Verification of Remark 3.2 Claim (1) follows from proposition 38.7 of Zeidler (1985). Claim (2) follows from corollary 41.9 of Zeidler (1985). For Claim (3), the fact that $\sqrt{W(\cdot)}m(\cdot,h):\mathcal{H}\to L^2(f_X)$ is compact and Frechet differentiable imply that its Frechet derivative is also a compact operator; see Zeidler (1985, proposition 7.33). This and the chain rule imply that the functional $E\left[||m(X,\cdot)||_W^2\right]:\mathcal{H}\to[0,\infty)$ is Frechet differentiable and its Frechet derivative is compact on $\mathcal{H}$. Hence $E\left[||m(X,h)||_W^2\right]$ has a compact Gateaux derivative on $\mathcal{H}$, and by Claim (2), is weak sequentially lower semicontinuous on $\mathcal{H}$. *Q.E.D.*

Proof of Theorem A.1: For result (1), we first show that the set of minimum penalization solution, $\mathcal{M}_0^P$, is not empty. Since $E\left[||m(X,h)||_W^2\right]$ is convex and lower semicontinuous in $h\in\mathcal{H}$ and $\mathcal{H}$ is a convex, closed and bounded subset of a reflexive Banach space (assumption 3.4(ii)), by proposition 38.15 of Zeidler (1985), $\mathcal{M}_0$ is convex, closed and bounded (and non-empty). Since $P(.)$ is convex and lower semi-continuous on $\mathcal{M}_0$, 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, $h'=\lambda h_1+(1-\lambda)h_0\in\mathcal{M}_0$. Since $P(.)$ 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, $\widehat{h}_n\in\mathcal{H}_{k(n)}$ with probability approaching one. We now show its consistency under the weak topology. To establish this, we adapt Step 1 in the proof of Theorem 3.3 to the case where assumption 3.1(ii) (identification) may not hold, but h_0 is the minimum penalization solution. Let $\mathcal{B}_w(h_0)$ denote any open neighborhood (in the weak topology) around h_0, and $\mathcal{B}_w^c(h_0)$ denote its complement (under the weak topology) in $\mathcal{H}$. By Lemma A.2(2), $P(\widehat{h}_n) = O_p(1)$. By Lemma A.3(2) with $\mathcal{B}_{\mathcal{T}}(h_0) = \mathcal{B}_w(h_0)$, $\mathcal{H}_{k(n)}^{M_0} \subseteq \mathcal{H}_{k(n)}$, $\eta_n = O(\eta_{0,n})$ and $\max\{\bar{\delta}_{m,n}^2, \eta_{0,n}, E\left[||m(X,\Pi_n h_0)||_W^2\right]\} = o(\lambda_n)$, we have: for all non-empty open ball $\mathcal{B}_w(h_0)$,

$$
\begin{aligned}
& \Pr\left(\widehat{h}_n \notin \mathcal{B}_w(h_0), \widehat{h}_n \in \mathcal{H}_{k(n)}^{M_0}\right) \\
\leq\ & \Pr\left(\inf_{\mathcal{H}_{k(n)}^{M_0}: h \notin \mathcal{B}_w(h_0)} \left\{cE[||m(X,h)||_W^2] + \lambda_n P(h)\right\} \leq \lambda_n P(h_0) + o_p(\lambda_n)\right) \\
\leq\ & \Pr\left(\inf_{\mathcal{H}_{k(n)}: h \notin \mathcal{B}_w(h_0)} \left\{cE[||m(X,h)||_W^2] + \lambda_n P(h)\right\} \leq \lambda_n P(h_0) + o_p(\lambda_n)\right).
\end{aligned}
$$

By assumptions 3.4(ii) and 3.1(iii), $\mathcal{H}_{k(n)}$ 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 assumption 3.4(ii) and the assumption that $E\left[||m(X,h)||_W^2\right]$ is convex and lower semicontinuous on $\mathcal{H}$, $cE[||m(X,h)||_W^2] + \lambda_n P(h)$ is weakly sequentially lower semicontinuous on $\mathcal{H}_{k(n)}$. Thus $g(k(n),\varepsilon,\lambda_n) \equiv \inf_{\mathcal{H}_{k(n)} \cap \mathcal{B}_w^c(h_0)} \left\{cE[||m(X,h)||_W^2] + \lambda_n P(h)\right\} \geq 0$ exists, and we denote its minimizer as $h_n(\varepsilon) \in \mathcal{H}_{k(n)} \cap \mathcal{B}_w^c(h_0)$. Hence, with $\max\{\bar{\delta}_{m,n}^2, \eta_{0,n}, E\left[||m(X,\Pi_n h_0)||_W^2\right]\} = o(\lambda_n)$ and $\lambda_n > 0$, we have:

$$
\begin{aligned}
& \Pr\left(\widehat{h}_n \notin \mathcal{B}_w(h_0), \widehat{h}_n \in \mathcal{H}_{k(n)}^{M_0}\right) \\
\leq\ & \Pr\left(cE[||m(X,h_n(\varepsilon))||_W^2] + \lambda_n P(h_n(\varepsilon)) \leq \lambda_n P(h_0) + o_p(\lambda_n)\right) \\
=\ & \Pr\left(\frac{g(k(n),\varepsilon,\lambda_n) - \lambda_n P(h_0)}{\lambda_n} \leq o_p(1)\right).
\end{aligned}
$$

If $\liminf_n E[||m(X,h_n(\varepsilon))||_W^2] = const. > 0$ then $\Pr\left(\widehat{h}_n \notin \mathcal{B}_w(h_0), \widehat{h}_n \in \mathcal{H}_{k(n)}^{M_0}\right) \to 0$ trivially. So we assume $\liminf_n E[||m(X,h_n(\varepsilon))||_W^2] = const. = 0$. Since $\mathcal{H} \cap \mathcal{B}_w^c(h_0)$ is weakly compact, 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 of $E\left[||m(X,h)||_W^2\right]$ on $\mathcal{H}$, $h_\infty(\varepsilon) \in \mathcal{M}_0$. By definition of h_0 and the assumption that $P(h)$ is strictly convex in $h \in \mathcal{M}_0$, it must be that $P(h_\infty(\varepsilon)) - P(h_0) \geq const. > 0$ by result (1). Note that this is true for any convergent subsequence. Therefore, we have established that

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

thus $\Pr\left(\widehat{h}_n \notin \mathcal{B}_w(h_0), \widehat{h}_n \in \mathcal{H}_{k(n)}^{M_0}\right) \to 0$. Hence, by similar calculations to those in Lemma A.3(2), for any $\varepsilon > 0$ and sufficiently large n, $\Pr\left(\widehat{h}_n \notin \mathcal{B}_w(h_0)\right) \leq \Pr\left(\widehat{h}_n \notin \mathcal{B}_w(h_0), \widehat{h}_n \in \mathcal{H}_{k(n)}^{M_0}\right) + \varepsilon \leq 2\varepsilon$.

Given the consistency under the weak topology, assumption 3.4(i) and Lemma A.2(2), we obtain $||\widehat{h}_n - h_0||_s = o_p(1)$ and $P(\widehat{h}_n) - P(h_0) = o_p(1)$ by following Step 2 in the proof of Theorem 3.3. *Q.E.D.*

Consistency: Proofs of Lemmas

Proof of Lemma A.1: By definition of the infimum, $\widehat{\alpha}_n$ always exists, and $\hat{\alpha}_n \in \mathcal{A}_{k(n)}$ with outer probability approaching one ($\widehat{\alpha}_n$ may not be measurable). Note that conditions (A.1.1)(ii), (A.1.2)(ii) and (A.1.4)(iii) imply that there is a sequence $\{\Pi_n\alpha_0 \in \mathcal{A}_{k(n)} \cap \mathcal{B}_{\mathcal{T}}(\alpha_0)\}$ for all $\mathcal{B}_{\mathcal{T}}(\alpha_0)$. It follows that for all $\mathcal{B}_{\mathcal{T}}(\alpha_0)$,

$$\begin{aligned}
&Pr^*\left(\hat{\alpha}_n \in \mathcal{A}_{k(n)}, \hat{\alpha}_n \notin \mathcal{B}_{\mathcal{T}}(\alpha_0)\right) \\
&\leq Pr^*\left(\inf_{\alpha\in\mathcal{A}_{k(n)}:\alpha\notin\mathcal{B}_{\mathcal{T}}(\alpha_0)} \widehat{Q}_n(\alpha) \leq \widehat{Q}_n(\Pi_n\alpha_0) + O_{p*}(\eta_n)\right) \\
&\leq Pr^*\left(\inf_{\alpha\in\mathcal{A}_{k(n)}:\alpha\notin\mathcal{B}_{\mathcal{T}}(\alpha_0)} \left\{K\overline{Q}_n(\alpha) - O_{p*}(c_n)\right\} \leq K_0\overline{Q}_n(\Pi_n\alpha_0) + O_{p*}(c_{0,n}) + O_{p*}(\eta_n)\right) \\
&\leq Pr^*\left(\inf_{\alpha\in\mathcal{A}_{k(n)}:\alpha\notin\mathcal{B}_{\mathcal{T}}(\alpha_0)} \overline{Q}_n(\alpha) \leq O_{p*}\left(\max\left\{c_n, c_{0,n}, \overline{Q}_n(\Pi_n\alpha_0), \eta_n\right\}\right)\right) \\
&\leq Pr^*\left(g_0(n, k(n), \mathcal{B}) \leq O_{p*}\left(\max\left\{c_n, c_{0,n}, \overline{Q}_n(\Pi_n\alpha_0), \eta_n\right\}\right)\right) \text{ by condition (A.1.1)(ii),}
\end{aligned}$$

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

Next we present another consistency lemma for penalized sieve extremum estimators, which is a special case of Lemma A.1, but is general enough and easily applicable in most applications.

Lemma SM.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 $\eta_n = o(1)$. Let $\overline{Q}_n() : \mathcal{A} \to (-\infty, \infty)$ be a sequence of non-random measurable functions and following conditions (SM.1.1) - (SM.1.4) hold:*

(SM.1.1) (i) $-\infty < \overline{Q}_n(\alpha_0) < \infty$; (ii) there is a positive function $g_0(n, k, \varepsilon)$ such that:

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

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

(SM.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 $\overline{Q}_n(\Pi_n\alpha_0) - \overline{Q}_n(\alpha_0) = o(1)$.

(SM.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.

(SM.1.4) Let $\hat{c}_n \equiv \sup_{\alpha\in\mathcal{A}_{k(n)}} \left|\widehat{Q}_n(\alpha) - \overline{Q}_n(\alpha)\right| = o_p(1)$.

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

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

Proof of Lemma SM.1: Under condition (SM.1.3)(ii) $\widehat{\alpha}_n$ is well-defined and measurable. Note that conditions (SM.1.1)(ii), (SM.1.2)(ii) and (SM.1.4) imply that there exists a sequence

$\Pi_n \alpha_0 \in \mathcal{A}_{k(n)}$ such that $||\Pi_n \alpha_0 - \alpha_0||_s = o(1)$. It follows that for any $\varepsilon > 0$,

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

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

Proof of Lemma A.2: We first show that $\widehat{h}_n \in \mathcal{H}_n$ wpa1. The infimum $\inf_{\mathcal{H}_n} \widehat{Q}_n(h)$ exists wpa1, and hence for any $\epsilon > 0$, there is a sequence, $(h_{j,n}(\epsilon))_j \subseteq \mathcal{H}_n$ such that $\widehat{Q}_n(h_{j,n}(\epsilon)) \leq \inf_{\mathcal{H}_n} \widehat{Q}_n(h) + \epsilon$ wpa1. Let $\widehat{h}_n \equiv h_{n,n}(\eta_n)$ and then such a choice satisfies $\widehat{h}_n \in \mathcal{H}_n$ wpa1.

Next, by definition of $\widehat{h}_n$, we have for any $\lambda_n > 0$,

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

and

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

Thus

$$\begin{aligned}
& \lambda_n\{P(\widehat{h}_n) - P(h_0)\} \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n} ||\widehat{m}(X_i, \Pi_n h_0)||^2_{\widehat{W}} + 2\lambda_n \sup_{h \in \mathcal{H}_n} \left|\widehat{P}_n(h) - P(h)\right| + \lambda_n \left|P(\Pi_n h_0) - P(h_0)\right| + O_p(\eta_n) \\
\leq\ & O_p\left(\eta_{0,n} + E[||m(X, \Pi_n h_0)||^2_W]\right) + 2\lambda_n \sup_{h \in \mathcal{H}_n} \left|\widehat{P}_n(h) - P(h)\right| + \lambda_n \left|P(\Pi_n h_0) - P(h_0)\right|
\end{aligned}$$

where the last inequality is due to assumption 3.3(i) and $\eta_n = O(\eta_{0,n})$. Therefore, for all $M > 0$,

$$\begin{aligned}
& \Pr\left(P(\widehat{h}_n) - P(h_0) > M\right) = \Pr\left(\lambda_n\{P(\widehat{h}_n) - P(h_0)\} > \lambda_n M\right) \\
\leq\ & \Pr\left(O_p\left(\eta_{0,n} + E[||m(X, \Pi_n h_0)||^2_W]\right) + 2\lambda_n \sup_{h \in \mathcal{H}_n} \left|\widehat{P}_n(h) - P(h)\right| + \lambda_n \left|P(\Pi_n h_0) - P(h_0)\right| > \lambda_n M\right).
\end{aligned}$$

(1) Under assumption 3.2(b), $\lambda_n \sup_{h\in\mathcal{H}_n} \left|\widehat{P}_n(h) - P(h)\right| + \lambda_n \left|P(\Pi_n h_0) - P(h_0)\right| = O_p(\lambda_n)$, we have:

$$\begin{aligned}
\Pr\left(P(\widehat{h}_n) - P(h_0) > M\right) &\leq \Pr\left(O_p\left(\max\left\{\eta_{0,n} + E[||m(X,\Pi_n h_0)||_W^2], \lambda_n\right\}\right) > \lambda_n M\right) \\
&\leq \Pr\left(O_p(\frac{\eta_{0,n} + E[||m(X,\Pi_n h_0)||_W^2]}{\lambda_n}) + O_p(1) > M\right)
\end{aligned}$$

which, under $\max\{\eta_{0,n}, E[||m(X,\Pi_n h_0)||_W^2]\} = O(\lambda_n)$, 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 3.2(c), $\lambda_n \sup_{h\in\mathcal{H}_n} \left|\widehat{P}_n(h) - P(h)\right| + \lambda_n \left|P(\Pi_n h_0) - P(h_0)\right| = o_p(\lambda_n)$, we have:

$$\Pr\left(P(\widehat{h}_n) - P(h_0) > M\right) \leq \Pr\left(O_p(\frac{\eta_{0,n} + E[||m(X,\Pi_n h_0)||_W^2]}{\lambda_n}) + o_p(1) > M\right)$$

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

Proof of Lemma A.3: It suffices to consider $\lambda_n P() > 0$ only. By the fact that $\Pr(A) \leq \Pr(A \cap B) + \Pr(B^c)$ for any measurable sets A and B, we have:

$$\Pr\left(\widehat{h}_n \notin \mathcal{B}_{\mathcal{T}}(h_0)\right) \leq \Pr\left(\widehat{h}_n \notin \mathcal{B}_{\mathcal{T}}(h_0), P(\hat{h}_n) \leq M_0\right) + \Pr\left(P(\hat{h}_n) > M_0\right).$$

For any $\varepsilon > 0$, choose $M_0 \equiv M_0(\varepsilon)$ such that $\Pr\left(P(\hat{h}_n) > M_0\right) < \varepsilon$ for sufficiently large n. Note that such a M_0 always exists by Lemma A.2. Thus, we can focus on the set $\mathcal{H}_{k(n)}^{M_0} \equiv \{h \in \mathcal{H}_{k(n)} : \lambda_n P(h) \leq \lambda_n M_0\}$ and bound $\Pr\left(\widehat{h}_n \notin \mathcal{B}_{\mathcal{T}}(h_0), P(\hat{h}_n) \leq M_0\right)$.

By definition of $\widehat{h}_n$ and $\Pi_n h_0$, assumptions 3.3 and 3.1(iii) and $\eta_n = O(\eta_{0,n})$, we have: for all $\mathcal{B}_{\mathcal{T}}(h_0)$,

$$\begin{aligned}
&\Pr\left(\widehat{h}_n \notin \mathcal{B}_{\mathcal{T}}(h_0), \widehat{h}_n \in \mathcal{H}_{k(n)}^{M_0}\right) \\
\leq\ &\Pr\left(\begin{array}{c} \inf_{h\in\mathcal{H}_{k(n)}^{M_0}: h\notin\mathcal{B}_{\mathcal{T}}(h_0)} \{\frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i,h)||_{\widehat{W}}^2 + \lambda_n \widehat{P}(h)\} \\ \leq \frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i,\Pi_n h_0)||_{\widehat{W}}^2 + \lambda_n \widehat{P}(\Pi_n h_0) + O_p(\eta_n) \end{array}\right) \\
\leq\ &\Pr\left(\begin{array}{c} \inf_{h\in\mathcal{H}_{k(n)}^{M_0}: h\notin\mathcal{B}_{\mathcal{T}}(h_0)} \{cE\left[||m(X,h)||_W^2\right] + \lambda_n \widehat{P}(h)\} \\ \leq c' E\left[||m(X,\Pi_n h_0)||_W^2\right] + O_p(\bar{\delta}_{m,n}^2) + \lambda_n \widehat{P}(\Pi_n h_0) + O_p(\eta_{0,n}) \end{array}\right).
\end{aligned}$$

By assumption 3.2(b), 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, with $\max\{\eta_{0,n}, E[||m(X,\Pi_n h_0)||_W^2]\} = O(\lambda_n)$, for all $\mathcal{B}_{\mathcal{T}}(h_0)$,

$$\begin{aligned}
&\Pr\left(\widehat{h}_n \notin \mathcal{B}_{\mathcal{T}}(h_0), \widehat{h}_n \in \mathcal{H}_{k(n)}^{M_0}\right) \\
\leq\ &\Pr\left(\inf_{h\in\mathcal{H}_{k(n)}^{M_0}: h\notin\mathcal{B}_{\mathcal{T}}(h_0)} \{cE\left[||m(X,h)||_W^2\right] + \lambda_n P(h)\} \leq O_p\left(\bar{\delta}_{m,n}^2\right) + \lambda_n P(h_0) + O_p(\lambda_n)\right).
\end{aligned}$$

By assumption 3.2(c), we have: $\lambda_n \sup_{h\in\mathcal{H}_n} |\widehat{P}(h) - P(h)| = o_p(\lambda_n)$ and $\lambda_n |P(\Pi_n h_0) - P(h_0)| = o(\lambda_n)$ for $\lambda_n > 0$. Thus, with $\max\{\eta_{0,n}, E[||m(X, \Pi_n h_0)||^2_W]\} = o(\lambda_n)$, for all $\mathcal{B}_{\mathcal{T}}(h_0)$,

$$\begin{aligned} & \Pr\left(\widehat{h}_n \notin \mathcal{B}_{\mathcal{T}}(h_0), \widehat{h}_n \in \mathcal{H}^{M_0}_{k(n)}\right) \\ \leq \quad & \Pr\left(\inf_{h\in\mathcal{H}^{M_0}_{k(n)}: h\notin\mathcal{B}_{\mathcal{T}}(h_0)} \left\{cE\left[||m(X,h)||^2_W\right] + \lambda_n P(h)\right\} \leq O_p\left(\bar{\delta}^2_{m,n}\right) + \lambda_n P(h_0) + o_p(\lambda_n)\right). \end{aligned}$$

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

Convergence Rate: Proofs of Theorems

Proof of Theorem 4.1: Directly follows from Lemma B.1 and the definition of $\omega_n(\delta, \mathcal{H}_{osn})$. *Q.E.D.*

Proof of Corollary 5.1: Directly follows from Theorem 4.1 and Lemma B.2. *Q.E.D.*

Proof of Corollary 5.2: Directly follows from Theorem 4.1 and Lemmas B.2 and B.3. *Q.E.D.*

Proof of Corollary 5.3: By Theorem 4.1, Lemma B.2 and Lemma B.3(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 $\delta^{*2}_{m,n} = \frac{J^*_n}{n} \asymp b^2_{m,J^*_n}$.

By assumption 5.4(ii) and the condition that either $P(h) \geq \sum_{j=1}^{\infty} \nu_j^{2\alpha} |\langle h, q_j\rangle_s|^2$ for all $h \in \mathcal{H}_{os}$ or $\mathcal{H}_{os} \subseteq \mathcal{H}_{ellipsoid}$, we have: for all $h \in \mathcal{H}_{os}$,

$$c_2 E[m(X,h)' W(X) m(X,h)] \leq ||h - h_0||^2 \leq const. \sum_{j=1}^{\infty} \{\varphi(\nu_j^{-2})\} \langle h - h_0, q_j\rangle^2_s.$$

On the other hand, the choice of penalty and the definition of $\mathcal{H}_{os}$ imply that $\sum_j \nu_j^{2\alpha} \langle h - h_0, q_j\rangle^2_s \leq const.$ for all $h \in \mathcal{H}_{os}$. Denote $\eta_j = \{\varphi(\nu_j^{-2})\}\langle h - h_0, q_j\rangle^2_s$. Then $\sum_j \nu_j^{2\alpha} \{\varphi(\nu_j^{-2})\}^{-1} \eta_j \leq M$. Therefore, the class $\{g \in L^2(\mathcal{X}, ||\cdot||_{L^2(f_X)}) : g(\cdot) = \sqrt{W(\cdot)} m(\cdot, h),\ h \in \mathcal{H}_{os}\}$ is embedded in the ellipsoid $\{g \in L^2(\mathcal{X}, ||\cdot||_{L^2(f_X)}) : ||g||^2_{L^2(f_X)} = \sum_j \eta_j,\ and\ \sum_j \nu_j^{2\alpha}\{\varphi(\nu_j^{-2})\}^{-1}\eta_j \leq M\}$. By invoking the results of Yang and Barron (1999), it follows that the J_n-th approximation error rate of this ellipsoid satisfies $b^2_{m,J_n} \leq const. \nu_{J_n}^{-2\alpha}\{\varphi(\nu_{J_n}^{-2})\}$. Hence $\delta^{*2}_{m,n} = \frac{J^*_n}{n} \asymp b^2_{m,J^*_n} \leq const. \nu_{J^*_n}^{-2\alpha}\{\varphi(\nu_{J^*_n}^{-2})\}$, 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*

Convergence Rate: Proofs of Lemmas

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

$$\begin{aligned}
& \Pr\left(\frac{||\widehat{h}_n - h_0||}{r_n} \geq M\right) \\
\leq \;& \Pr\left(\begin{array}{c} \inf_{\{h\in\mathcal{H}_{osn}:||h-h_0||\geq Mr_n\}}\{\frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i,h)||^2_{\widehat{W}} + \lambda_n\widehat{P}_n(h)\} \\ \leq \frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i,\Pi_n h_0)||^2_{\widehat{W}} + \lambda_n\widehat{P}_n(\Pi_n h_0) + O_p(\eta_n) \end{array}\right) \\
\leq \;& \Pr\left(\begin{array}{c} \inf_{\{h\in\mathcal{H}_{osn}:||h-h_0||\geq Mr_n\}}\{\frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i,h)||^2_{\widehat{W}} + \lambda_n P(h)\} \\ \leq \frac{1}{n}\sum_{i=1}^n ||\widehat{m}(X_i,\Pi_n h_0)||^2_{\widehat{W}} + \lambda_n P(\Pi_n h_0) + 2\lambda_n\delta_{P,n} + O_p(\eta_n) \end{array}\right),
\end{aligned}$$

where the last inequality is due to $\sup_{h\in\mathcal{H}_{osn}}|\widehat{P}_n(h) - P(h)| = O_p(\delta_{P,n}) = O_p(1)$. By assumption 3.3 with $\eta_{0,n} = O(\delta_{m,n}^2)$, $\eta_n = O(\eta_{0,n})$ and definitions of $\mathcal{H}_{osn}$ and $\delta_{m,n}^2$, there are two finite constants $c, c_0 > 0$ such that:

$$cE\left(||m(X,\widehat{h}_n)||^2_W\right) + \lambda_n P(\widehat{h}_n) \leq O_p(\delta_{m,n}^2 + \lambda_n\delta_{P,n}) + c_0 E\left(||m(X,\Pi_n h_0)||^2_W\right) + \lambda_n P(\Pi_n h_0) \quad \text{(SM.4)}$$

which implies

$$cE\left(||m(X,\widehat{h}_n)||^2_W\right) \leq O_p(\delta_{m,n}^2 + \lambda_n\delta_{P,n}) + c_0 E\left(||m(X,\Pi_n h_0)||^2_W\right) + \lambda_n|P(\Pi_n h_0) - P(\widehat{h}_n)|.$$

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

$$\begin{aligned}
& \Pr\left(\frac{||\widehat{h}_n - h_0||}{r_n} \geq M\right) \\
\leq \;& \Pr\left(M^2 r_n^2 \leq O_p\left(\max\left\{\delta_{m,n}^2, \lambda_n\delta_{P,n}, ||\Pi_n h_0 - h_0||^2, \lambda_n|P(\Pi_n h_0) - P(\widehat{h}_n)|\right\}\right)\right),
\end{aligned}$$

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

By definition of $\mathcal{H}_{osn}$ (or under assumption 3.2(a)(b)), $\lambda_n|P(\Pi_n h_0) - P(\widehat{h}_n)| = O_p(\lambda_n)$ and $\delta_{P,n} = O_p(1)$; hence Result (1) follows.

Under assumption 3.2(c), $\lambda_n|P(\Pi_n h_0) - P(\widehat{h}_n)| = o_p(\lambda_n)$ and $\delta_{P,n} = o_p(1)$; hence Result (2) follows.

For Result (3), using the same argument as that for Results (1)(2), inequality (SM.4) still holds. By condition (3) of Theorem 4.1, $\lambda_n\left(P(\widehat{h}_n) - P(\Pi_n h_0)\right) \geq \lambda_n\langle t_0, \widehat{h}_n - \Pi_n h_0\rangle_{\mathbf{H}^*,\mathbf{H}}$. Thus

$$cE\left(||m(X,\widehat{h}_n)||^2_W\right) + \lambda_n\langle t_0, \widehat{h}_n - \Pi_n h_0\rangle_{\mathbf{H}^*,\mathbf{H}} \leq O_p(\delta_{m,n}^2 + \lambda_n\delta_{P,n}) + c_0 E\left(||m(X,\Pi_n h_0)||^2_W\right),$$

hence

$$cE\left(||m(X,\widehat{h}_n)||^2_W\right) \leq O_p(\delta_{m,n}^2 + \lambda_n\delta_{P,n}) + c_0 E\left(||m(X,\Pi_n h_0)||^2_W\right) + const.\lambda_n||\widehat{h}_n - \Pi_n h_0||_s.$$

By assumption 4.1, Lemma B.1(3) follows by choosing $r_n^2 = \max\{\delta_{m,n}^2, \lambda_n \delta_{P,n}, ||\Pi_n h_0 - h_0||^2, \lambda_n ||\widehat{h}_n - \Pi_n h_0||_s\} = o_p(1)$. *Q.E.D.*

Proof of Lemma B.2: 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 $\{q_j\}_{j=1}^{\infty}$ is a Riesz basis, and hence for any $h \in \mathcal{H}_{osn}$, 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, assumption 5.2(ii) implies 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 B.3: 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 5.3 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 is small, we can assume $M^2(\nu_2)^{-2\alpha} > \frac{1}{c}\delta^2 / b_1$. 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

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

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 B.2 and Result (1). *Q.E.D*

Proofs of Lemmas for Series LS estimation of $m()$

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

Lemma SM.2. *Let assumptions C.1 and C.2(i) hold. Then: there are finite constants* $c, c' > 0$ *such that, wpa1,*

$$cE_X\left[||\widetilde{m}\left(X,h\right)||^2_W\right] \leq \frac{1}{n}\sum_{i=1}^{n}||\widetilde{m}\left(X_i,h\right)||^2_{\widehat{W}} \leq c'E_X\left[||\widetilde{m}\left(X,h\right)||^2_W\right] \text{ uniformly in } h \in \mathcal{H}^{M_0}_{k(n)}.$$

Proof of Lemma SM.2: Denote $\langle g, \overline{g}\rangle_{n,X} \equiv \frac{1}{n}\sum_{i=1}^{n} g(X_i)\overline{g}(X_i)$ and $\langle g, \overline{g}\rangle_X \equiv E_X[g(X)\overline{g}(X)]$, where $g(X)$ and $\overline{g}(X)$ are square integrable functions of X. We want to show that for all $t > 0$,

$$\lim_{n\to\infty} \Pr\left(\sup_{h\in\mathcal{H}^P_{k(n)}}\left|\frac{\langle \widetilde{m}(\cdot,h),\widetilde{m}(\cdot,h)\rangle_{n,X} - \langle \widetilde{m}(\cdot,h),\widetilde{m}(\cdot,h)\rangle_X}{\langle \widetilde{m}(\cdot,h),\widetilde{m}(\cdot,h)\rangle_X}\right| > t\right) = 0 \qquad \text{(SM.5)}$$

Let $G_n \equiv \{g : g(x) = \sum_{k=1}^{J_n}\pi_k p_k(x);\ \pi_k \in \mathcal{R},\ \sup_x|g(x)| < \infty\}$. By construction $\widetilde{m}(X,h) = \arg\min_{g\in G_n} n^{-1}\sum_{i=1}^{n}||m(X_i,h) - g(X_i)||^2_I$; so $\widetilde{m}(X,h) \in G_n$ and

$$\sup_{h\in\mathcal{H}^P_{k(n)}}\left|\frac{\langle \widetilde{m}(\cdot,h),\widetilde{m}(\cdot,h)\rangle_{n,X} - \langle \widetilde{m}(\cdot,h),\widetilde{m}(\cdot,h)\rangle_X}{\langle \widetilde{m}(\cdot,h),\widetilde{m}(\cdot,h)\rangle_X}\right| \leq \sup_{g\in G_n:||g||_X=1}\left|\langle g,g\rangle_{n,X} - \langle g,g\rangle_X\right|.$$

Define $A_n \equiv \sup_{g\in G_n}\frac{\sup_x|g(x)|}{\sqrt{E[(g(X))^2]}}$. Then, under assumption C.1(i)(ii)(iii) and the definition of G_n, we have $A_n \asymp \xi_n$. Thus, by assumption C.1(iv), 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, equation (SM.5) holds. So with $t = 0.5$, we obtain that uniformly over $h \in \mathcal{H}^{M_0}_{k(n)}$,

$$0.5E_X\left[||\widetilde{m}(X,h)||^2_I\right] \leq \frac{1}{n}\sum_{i=1}^{n}||\widetilde{m}(X_i,h)||^2_I \leq 2E_X\left[||\widetilde{m}(X,h)||^2_I\right].$$

expect for an event wpa0. By assumption C.1(v), there are finite constants $K, K' > 0$ such that $KI \leq W(X) \leq K'I$ for almost all X. Thus, $K||\widetilde{m}(X,h)||^2_I \leq ||\widetilde{m}(X,h)||^2_W \leq K'||\widetilde{m}(X,h)||^2_I$ for almost all X. Also by assumption C.1(v), uniformly over $h \in \mathcal{H}_{k(n)}$,

$$\begin{aligned}||\widetilde{m}(X,h)||^2_{\widehat{W}} &= \widetilde{m}(X,h)'\{\widehat{W}(X) - W(X) + W(X)\}\widetilde{m}(X,h)\\ &\leq \sup_{x\in\mathcal{X}}|\widehat{W}(x) - W(x)| \times ||\widetilde{m}(X,h)||^2_I + ||\widetilde{m}(X,h)||^2_W \leq (K' + o_p(1))||\widetilde{m}(X,h)||^2_I.\end{aligned}$$

Similarly,

$$||\widetilde{m}(X,h)||^2_{\widehat{W}} \geq (K - o_p(1))||\widetilde{m}(X,h)||^2_I.$$

Note that for n large, $\min\{K', K\} \pm o_p(1) > 0$. Therefore, uniformly over $h \in \mathcal{H}^{M_0}_{k(n)}$,

$$const \times E_X\left[||\widetilde{m}(X,h)||^2_W\right] \leq \frac{1}{n}\sum_{i=1}^{n}||\widetilde{m}(X_i,h)||^2_{\widehat{W}} \leq const' \times E_X\left[||\widetilde{m}(X,h)||^2_W\right]$$

except for a set wpa0. *Q.E.D.*

Proof of Lemma C.1: By assumption C.1(i)(v) it suffices to establish the results for $W = I$. Result (1) directly follows from our assumption C.1 and Lemma A.1 Part (C) of Ai and Chen (2003).

Result (3) can be established in the same way as that of Result (2). For Result (2), let $\varepsilon(Z,h) \equiv \rho(Z,h) - m(X,h)$ and $\varepsilon(h) \equiv (\varepsilon(Z_1,h),...,\varepsilon(Z_n,h))'$. For any symmetric and positive matrix Ω $(d \times d)$, we have the spectral decomposition $\Omega = U\Lambda U'$ where $\Lambda = diag\{\lambda_1,...,\lambda_d\}$ with $\lambda_i > 0$ and $UU' = I_d$. Denote $\lambda_{\min}(\Omega)$ as the smallest eigenvalue of the matrix Ω. By definition we have:

$$\begin{aligned}
&\sup_{h\in\mathcal{H}^{M_0}_{k(n)}} \frac{1}{n}\sum_{i=1}^{n} ||\widehat{m}(X_i,h) - \widetilde{m}\,(X_i,h)\,||^2_I \\
&= \sup_{h\in\mathcal{H}^{M_0}_{k(n)}} \frac{1}{n}\sum_{i=1}^{n} Tr\{p^{J_n}(X_i)'(P'P)^{-1}P'\varepsilon(h)\varepsilon(h)'P(P'P)^{-1}p^{J_n}(X_i)\} \\
&= \sup_{h\in\mathcal{H}^{M_0}_{k(n)}} \frac{1}{n}\sum_{i=1}^{n} Tr\{\varepsilon(h)'P(P'P)^{-1}p^{J_n}(X_i)p^{J_n}(X_i)'(P'P)^{-1}P'\varepsilon(h)\} \\
&= \sup_{h\in\mathcal{H}^{M_0}_{k(n)}} \frac{1}{n} Tr\{\varepsilon(h)'P(P'P)^{-1}\sum_{i=1}^{n}\{p^{J_n}(X_i)p^{J_n}(X_i)'\}(P'P)^{-1}P'\varepsilon(h)\} \\
&= \sup_{h\in\mathcal{H}^{M_0}_{k(n)}} \frac{1}{n} Tr\{\varepsilon(h)'P(P'P)^{-1}P'\varepsilon(h)\} = \sup_{h\in\mathcal{H}^{M_0}_{k(n)}} \frac{1}{n^2} Tr\{\varepsilon(h)'P(P'P/n)^{-1}P'\varepsilon(h)\} \\
&\le (\lambda_{\min}(P'P/n))^{-1} \times \sup_{\mathcal{H}^{M_0}_{k(n)}} \frac{1}{n^2} Tr\{\varepsilon(h)'PP'\varepsilon(h)\}.
\end{aligned}$$

Note that

$$\varepsilon(h)'PP'\varepsilon(h) = \sum_{j=1}^{J_n}\left(\left|\sum_{i=1}^{n} p_j(X_i)\varepsilon(Z_i,h)\right|\right)^2.$$

Let $r_n = \frac{J_n}{n}C_n$. We have for all $M \ge 1$,

$$\begin{aligned}
&\Pr\left(\sup_{h\in\mathcal{H}^{M_0}_{k(n)}} \frac{1}{n}\sum_{i=1}^{n} ||\widehat{m}(X_i,h) - \widetilde{m}\,(X_i,h)\,||^2_I > Mr_n\right) \\
&\le \Pr\left((\lambda_{\min}(P'P/n))^{-1} \sup_{h\in\mathcal{H}^{M_0}_{k(n)}} \sum_{j=1}^{J_n}\left(\left|\frac{1}{n}\sum_{i=1}^{n} p_j(X_i)\varepsilon(Z_i,h)\right|\right)^2 > Mr_n\right) \\
&\le \Pr\left((\lambda_{\min}(P'P/n))^{-1} \sum_{j=1}^{J_n}\left(\sup_{h\in\mathcal{H}^{M_0}_{k(n)}}\left|\frac{1}{n}\sum_{i=1}^{n} p_j(X_i)\varepsilon(Z_i,h)\right|\right)^2 > Mr_n\right).
\end{aligned}$$

Following Newey (1997, p. 162) and under assumption C.1(i)(ii)(iii)(iv), we have: $(\lambda_{\min}(P'P/n))^{-1} = O_p(1)$. Thus, to bound $\Pr\left(\sup_{\mathcal{H}^{M_0}_{k(n)}} n^{-1}\sum_{i=1}^{n} ||\widehat{m}(X_i,h) - \widetilde{m}(X_i,h)||^2_I > Mr_n\right)$, it suffices to bound the probability

$$\begin{aligned}
&\Pr\left(\sum_{j=1}^{J_n}\left(\sup_{h\in\mathcal{H}_{k(n)}^{M_0}}\left|\frac{1}{n}\sum_{i=1}^{n}p_j(X_i)\varepsilon(Z_i,h)\right|\right)^2 > Mr_n\right)\\
\leq\ &\frac{1}{Mr_n}E_{Z^n}\left[\sum_{j=1}^{J_n}\left(\sup_{h\in\mathcal{H}_{k(n)}^{M_0}}\left|\frac{1}{n}\sum_{i=1}^{n}p_j(X_i)\varepsilon(Z_i,h)\right|\right)^2\right]\\
\leq\ &\frac{J_n}{nr_nM}\max_{1\leq j\leq J_n}E_{Z^n}\left[\left(\sup_{h\in\mathcal{H}_{k(n)}^{M_0}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}p_j(X_i)\varepsilon(Z_i,h)\right|\right)^2\right],
\end{aligned}$$

where the first inequality is by Markov inequality, and $E_{Z^n}(\cdot)$ denotes the expectation with respect to $Z^n \equiv (Z_1, ..., Z_n)$. By Theorem 2.14.5 in Van der Vaart and Wellner (VdV-W, 1996) (also see Pollard, 1990), we have:

$$\begin{aligned}
&\max_{1\leq j\leq J_n}E_{Z^n}\left[\left(\sup_{h\in\mathcal{H}_{k(n)}^{M_0}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}p_j(X_i)\varepsilon(Z_i,h)\right|\right)^2\right]\\
\leq\ &\max_{1\leq j\leq J_n}\left(E_{Z^n}\left[\sup_{h\in\mathcal{H}_{k(n)}^{M_0}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}p_j(X_i)\varepsilon(Z_i,h)\right|\right]+\sqrt{E[|p_j(X)\bar{\rho}_n(Z)|^2]}\right)^2.
\end{aligned}$$

By assumption C.2(i) and $\max_{1\leq j\leq J_n}E[|p_j(X)|^2]\leq const.$, we have

$$\max_{1\leq j\leq J_n}E[|p_j(X)\bar{\rho}_n(Z)|^2]\leq const. < \infty.$$

By Theorem 2.14.2 in VdV-W (1996), we have (up to some constant)

$$\begin{aligned}
&\max_{1\leq j\leq J_n}E_{Z^n}\left[\sup_{h\in\mathcal{H}_{k(n)}^{M_0}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}p_j(X_i)\varepsilon(Z_i,h)\right|\right]\\
\leq &\max_{1\leq j\leq J_n}\left\{\sqrt{E[|p_j(X)\bar{\rho}_n(Z)|^2]}\int_0^1\sqrt{1+\log N_{[]}(wK,\mathcal{E}_{jn},||.||_{L^2(f_Z)})}dw\right\}\\
\leq &K\max_{1\leq j\leq J_n}\int_0^1\sqrt{1+\log N_{[]}(wK,\mathcal{E}_{jn},||.||_{L^2(f_Z)})}dw,
\end{aligned}$$

where $\mathcal{E}_{jn}\equiv\{p_j(\cdot)\varepsilon(\cdot,h): h\in\mathcal{H}_{k(n)}^{M_0}\}$. Note that for any $h, h'\in\mathcal{H}_{k(n)}^{M_0}$, we have:

$$|p_j(X)(\varepsilon(Z,h)-\varepsilon(Z,h'))|\leq|p_j(X)|\{|\rho(Z,h)-\rho(Z,h')|+|E[\rho(Z,h)|X]-E[\rho(Z,h')|X]|\},$$

and

$$|p_j(X)||E[\rho(Z,h)|X]-E[\rho(Z,h')|X]|\leq|p_j(X)|E[|\rho(Z,h)-\rho(Z,h')||X].$$

Recall that $\mathcal{O}_{jn}\equiv\{p_j(\cdot)\rho(\cdot,h): h\in\mathcal{H}_{k(n)}^{M_0}\}$ and that $\max_{1\leq j\leq J_n}\int_0^1\sqrt{1+\log N_{[]}(wK,\mathcal{O}_{jn},||.||_{L^2(f_Z)})}dw\leq\sqrt{C_n}<\infty$ by assumption C.2(iii). We have: $\max_{1\leq j\leq J_n}\int_0^1\sqrt{1+\log N_{[]}(wK,\mathcal{E}_{jn},||.||_{L^2(f_Z)})}dw\leq$

$const.\sqrt{C_n} < \infty$ and hence

$$\max_{1\le j\le J_n} E_{Z^n}\left[\sup_{h\in\mathcal{H}_{k(n)}^{M_0}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} p_j(X_i)\varepsilon(Z_i,h)\right|\right] \le const. \times \sqrt{C_n}$$

It then follows

$$\frac{J_n}{nr_nM}\max_{1\le j\le J_n} E_{Z^n}\left[\left(\sup_{h\in\mathcal{H}_{k(n)}^{M_0}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} p_j(X_i)\varepsilon(Z_i,h)\right|\right)^2\right] \le const. \times \frac{J_nC_n}{nr_nM},$$

so $r_n = \frac{J_n}{n}C_n$ and letting $M \to \infty$, the desired result follows. *Q.E.D.*

Proof of Lemma C.2: The proofs of Results (1) and (3) are the same as that of Result (2). For Result (2), by the fact $(a-b)^2 + b^2 \ge \frac{1}{2}a^2$, we have that uniformly over $h \in \mathcal{H}_{k(n)}$,

$$\frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)||_{\widehat{W}}^2 \ge \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}||\widetilde{m}(X_i,h)||_{\widehat{W}}^2 - \frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)-\widetilde{m}(X_i,h)||_{\widehat{W}}^2.$$

By Lemma SM.2, there is a finite constant $c > 0$ such that, wpa1 and uniformly over $h \in \mathcal{H}_{k(n)}^{M_0}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)||_{\widehat{W}}^2 \ge \frac{c}{2}E_X[||\widetilde{m}(X,h)||_W^2] - \frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)-\widetilde{m}(X_i,h)||_{\widehat{W}}^2\\
&\ge \frac{c}{4}E_X[||m(X,h)||_W^2] - \left(\frac{c}{2}E_X[||m(X,h)-\widetilde{m}(X,h)||_W^2] + \frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)-\widetilde{m}(X_i,h)||_{\widehat{W}}^2\right)\\
&\ge KE_X[||m(X,h)||_W^2] - O_p\left(b_{m,J_n}^2 + \frac{J_n}{n}C_n\right),
\end{aligned}$$

where the second inequality is due to the fact $(a-b)^2 + b^2 \ge \frac{1}{2}a^2$, and the last inequality is due to lemma C.1, assumption C.2(ii) and $\frac{c}{4} \equiv K > 0$.

Similarly, by the fact $(a+b)^2 \le 2a^2 + 2b^2$, we have that uniformly over $h \in \mathcal{H}_{k(n)}$,

$$\frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)||_{\widehat{W}}^2 \le 2\frac{1}{n}\sum_{i=1}^{n}||\widetilde{m}(X_i,h)||_{\widehat{W}}^2 + 2\frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)-\widetilde{m}(X_i,h)||_{\widehat{W}}^2.$$

By Lemma SM.2, there is a finite constant $c' > 0$ such that, wpa1 and uniformly over $h \in \mathcal{H}_{k(n)}^{M_0}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)||_{\widehat{W}}^2 \le 2c'E_X[||\widetilde{m}(X,h)||_W^2] + 2\frac{1}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)-\widetilde{m}(X_i,h)||_{\widehat{W}}^2\\
&\le 4c'E_X[||m(X,h)||_W^2] + \left(4c'E_X[||\widetilde{m}(X,h)-m(X,h)||_W^2] + \frac{2}{n}\sum_{i=1}^{n}||\widehat{m}(X_i,h)-\widetilde{m}(X_i,h)||_{\widehat{W}}^2\right)\\
&\le K'E_X[||m(X,h)||_W^2] + O_p\left(b_{m,J_n}^2 + \frac{J_n}{n}C_n\right),
\end{aligned}$$

where the second inequality is again due to the fact $(a+b)^2 \leq 2a^2 + 2b^2$, and the last inequality is due to lemma C.1, assumption C.2(ii) and $4c' \equiv K' < \infty$. *Q.E.D.*

Proof of Lemma C.3: By assumption C.1(i)(v) it suffices to establish the results for $W = I$. Using the same notations and following the steps as in the proof of Lemma C.1, we obtain:

$$\begin{aligned}
&\sup_{h\in\mathcal{N}_{os}} \frac{1}{n}\sum_{i=1}^{n} \|\widehat{m}(X_i,h) - \widehat{m}(X_i,h_0) - \widetilde{m}(X_i,h)\|_I^2 \\
&= \sup_{h\in\mathcal{N}_{os}} \frac{1}{n^2} Tr\{[\varepsilon(h)-\varepsilon(h_0)]'P(P'P/n)^{-1}P'[\varepsilon(h)-\varepsilon(h_0)]\} \\
&\leq (\lambda_{\min}(P'P/n))^{-1} \times \sup_{h\in\mathcal{N}_{os}} \frac{1}{n^2} Tr\{[\varepsilon(h)-\varepsilon(h_0)]'PP'[\varepsilon(h)-\varepsilon(h_0)]\} \\
&= (\lambda_{\min}(P'P/n))^{-1} \times \sup_{h\in\mathcal{N}_{os}} \frac{1}{n^2}\sum_{j=1}^{J_n}\left(\left|\sum_{i=1}^{n} p_j(X_i)[\varepsilon(Z_i,h)-\varepsilon(Z_i,h_0)]\right|\right)^2.
\end{aligned}$$

Let $r_n = \frac{J_n}{n}(\delta_{s,n})^{2\kappa}$. For all $M \geq 1$, to bound

$$\Pr\left(\sup_{h\in\mathcal{N}_{os}} \frac{1}{n}\sum_{i=1}^{n} \|\widehat{m}(X_i,h) - \widehat{m}(X_i,h_0) - \widetilde{m}(X_i,h)\|_I^2 > Mr_n\right),$$

it suffices to bound the probability

$$\begin{aligned}
&\Pr\left(\sum_{j=1}^{J_n}\left(\sup_{h\in\mathcal{N}_{os}}\left|\frac{1}{n}\sum_{i=1}^{n} p_j(X_i)[\varepsilon(Z_i,h)-\varepsilon(Z_i,h_0)]\right|\right)^2 > Mr_n\right) \\
&\leq \frac{J_n}{nr_nM}\max_{1\leq j\leq J_n} E_{Z^n}\left[\left(\sup_{h\in\mathcal{N}_{os}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} p_j(X_i)[\varepsilon(Z_i,h)-\varepsilon(Z_i,h_0)]\right|\right)^2\right].
\end{aligned}$$

Let $\Delta\varepsilon(Z_i,h) \equiv \varepsilon(Z_i,h) - \varepsilon(Z_i,h_0)$. By Theorem 2.14.5 in Van der Vaart and Wellner (VdV-W, 1996), we have

$$\begin{aligned}
&\max_{1\leq j\leq J_n} E_{Z^n}\left[\left(\sup_{h\in\mathcal{N}_{os}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} p_j(X_i)\Delta\varepsilon(Z_i,h)\right|\right)^2\right] \\
&\leq \max_{1\leq j\leq J_n}\left(E_{Z^n}\left[\sup_{h\in\mathcal{N}_{os}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} p_j(X_i)\Delta\varepsilon(Z_i,h)\right|\right] + \sqrt{E[\sup_{h\in\mathcal{N}_{os}} |p_j(X)\Delta\varepsilon(Z,h)|^2]}\right)^2.
\end{aligned}$$

By Jensen's inequality,

$$E\left[\sup_{h\in\mathcal{N}_{os}} |p_j(X)\{m(X,h)-m(X,h_0)\}|^2\right] \leq E\left[\sup_{h\in\mathcal{N}_{os}} |p_j(X)\{\rho(Z,h)-\rho(Z,h_0)\}|^2\right].$$

Hence

$$\max_{1\leq j\leq J_n}\sqrt{E\left[\sup_{h\in\mathcal{N}_{os}} |p_j(X)\Delta\varepsilon(Z,h)|^2\right]} \leq \max_{1\leq j\leq J_n}\sqrt{2E\left[\sup_{h\in\mathcal{N}_{os}} |p_j(X)\{\rho(Z,h)-\rho(Z,h_0)\}|^2\right]} \leq const.\times(\delta_{s,n})^{\kappa}$$

by condition (C.3.1)(i).

By Theorem 2.14.2 in VdV-W (1996), Remark C.1 and condition (C.3.1)(i)(ii), we have (up to some constant),

$$
\begin{aligned}
&\max_{1\le j\le J_n} E_{Z^n}\left[\sup_{h\in\mathcal{N}_{os}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} p_j(X_i)\Delta\varepsilon(Z_i,h)\right|\right] \\
\le &\max_{1\le j\le J_n}\left\{(\delta_{s,n})^{\kappa}\int_0^1\sqrt{1+\log N_{[]}(w(\delta_{s,n})^{\kappa},\{p_j(\cdot)\Delta\varepsilon(\cdot,h):h\in\mathcal{N}_{os}\},||.||_{L^2(f_Z)})}dw\right\} \\
\le &(\delta_{s,n})^{\kappa}\int_0^1\sqrt{1+\log N(w^{1/\kappa},\mathcal{N}_{os},||.||_s)}dw \le const.\times(\delta_{s,n})^{\kappa}.
\end{aligned}
$$

Hence

$$
\max_{1\le j\le J_n} E_{Z^n}\left[\left(\sup_{h\in\mathcal{N}_{os}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} p_j(X_i)\Delta\varepsilon(Z_i,h)\right|\right)^2\right] = O((\delta_{s,n})^{2\kappa}).
$$

The desired result follows. *Q.E.D.*

Application: Proofs of Propositions

Proof of Proposition 6.1: We obtain the result by verifying that all the assumptions of Theorem 3.2 (lower semicompact penalty) are satisfied with $\widehat{W}=W=I$.

First, assumption 3.1(i) is trivially satisfied with $W=I$. For any $h\in\mathcal{H}$ we denote $h(y_1,y_2)=h_1(y_1)+h_2(y_2)$, $\Delta h(y_1,y_2)=h(y_1,y_2)-h_0(y_1,y_2)=\Delta h_1(y_1)+\Delta h_2(y_2)$, $\Delta h_l(y_l)=h_l(y_l)-h_{0l}(y_l)$ for $l=1,2$. By the mean value theorem, condition 6.1(iv) and the definitions of $K_{l,h}[\Delta h_l](X)$, we have:

$$
\begin{aligned}
& m(X,h)-m(X,h_0) \\
= \;& E[F_{Y_3|Y_1,Y_2,X}(h_1(Y_1)+h_2(Y_2))-F_{Y_3|Y_1,Y_2,X}(h_{01}(Y_1)+h_{02}(Y_2))|X] \\
= \;& E\left(\left\{\int_0^1 f_{Y_3|Y_1,Y_2,X}(h_0(Y_1,Y_2)+t\Delta h(Y_1,Y_2))dt\right\}[\Delta h_1(Y_1)+\Delta h_2(Y_2)]|X\right) \qquad \text{(SM.6)} \\
= \;& K_{1,h}[\Delta h_1](X)+K_{2,h}[\Delta h_2](X).
\end{aligned}
$$

Therefore, for any $h\in\mathcal{H}$ such that $m(X,h)-m(X,h_0)=0$ almost surely X, under condition 6.2(ii), we have: $K_{1,h}[\Delta h_1](X)=0$, $K_{2,h}[\Delta h_2](X)=0$ almost surely X, which implies $\Delta h_l=0$ almost surely Y_l for $l=1,2$ (by condition 6.2(ii)). Thus, the identification assumption 3.1(ii) holds. Given our choices of $\mathcal{H}$, $\mathcal{H}_n$ (condition 6.2(i)(iii)), and $\|h\|_s=\|h\|_{\sup}=\sup_{y_1}|h_1(y_1)|+\sup_{y_2}|h_2(y_2)|$, the sieve space $\mathcal{H}_n$ is closed, and we have for $h_0\in\mathcal{H}$, there is $\Pi_n h_0\in\mathcal{H}_n$ such that

$$
\|h_0-\Pi_n h_0\|_s=\|h_0-\Pi_n h_0\|_{\sup}\le c\{k_1(n)\}^{-r_1}+c'\{k_2(n)\}^{-r_2}=o(1), \quad \text{with } r_l=\alpha_l/d,
$$

thus assumption 3.1(iii) holds. For any $h\in\mathcal{H}$ with $\Delta h(y_1,y_2)=\Delta h_1(y_1)+\Delta h_2(y_2)$, $\Delta h_l(y_l)=$

$h_l(y_l) - h_{0l}(y_l)$, $l = 1, 2$, equation (SM.6) implies that

$$\begin{aligned} & |m(X,h) - m(X,h_0)| \\ \leq \; & E\left(\sup_{t\in[0,1]} f_{Y_3|Y_1,Y_2,X}(h_0(Y_1,Y_2) + t\Delta h(Y_1,Y_2))|X\right)\left[\sup_{y_1}|\Delta h_1(y_1)| + \sup_{y_2}|\Delta h_2(y_2)|\right]. \end{aligned}$$

Since $m(X,h_0) = 0$ and by condition 6.1(iv), we have

$$\begin{aligned} E[|m(X,h)|^2] &= E[|m(X,h) - m(X,h_0)|^2] \\ &\leq E\left[\left(\sup_{t\in[0,1]} f_{Y_3|Y_1,Y_2,X}(h_0(Y_1,Y_2) + t\Delta h(Y_1,Y_2))|X\right)\right]^2 (\|h-h_0\|_s)^2 \\ &\leq const. \times [\|h-h_0\|_s]^2 . \end{aligned}$$

This and $\|\Pi_n h_0 - h_0\|_s = o(1)$ imply

$$E[|m(X,\Pi_n h_0)|^2] \leq const. \|\Pi_n h_0 - h_0\|_s^2 \leq c\{k_1(n)\}^{-2r_1} + c'\{k_2(n)\}^{-2r_2} = o(1),$$

hence assumption 3.1(iv) holds. Assumption 3.2(b) directly follows from our choice of $\widehat{P}() = P()$.

Next, condition 6.1(i)(ii) and $\widehat{W} = W = I$ imply that assumption C.1 holds. Assumption C.2(i) follows trivially with $\bar{\rho}_n(Z) \equiv 1$ since $\sup_{h\in\mathcal{H}}|\rho(Z,h)| \leq 1$. Condition 6.1(ii)(iii) implies that assumption C.2(ii) holds with $b^2_{m,J_n} = J_n^{-2r_m}$. Thus Lemma C.2 result (1) is applicable and assumption 3.3(i) is satisfied with $\eta_{0,n} = \frac{J_n}{n} + J_n^{-2r_m}$. This, $E([m(X,\Pi_n h_0)]^2) = O\left(\max\left[\{k_1(n)\}^{-2r_1}, \{k_2(n)\}^{-2r_2}\right]\right)$, and $\max\left[\{k_1(n)\}^{-2r_1}, \{k_2(n)\}^{-2r_2}, \frac{J_n}{n} + J_n^{-2r_m}\right] = O(\lambda_n)$ together imply that Lemma A.2(1) holds. Moreover, it follows by our choice of penalty that $P(\Pi_n h_0) = O(1)$ and $P(\widehat{h}_n) = O_p(1)$. By our choice of sieves space, it follows that $\log N(w^{1/2}, \mathcal{H}^{M_0}_{k(n)}, ||.||_{L^\infty}) \leq \min\left\{\frac{1}{2}k(n)\log(1/w), const.(1/w)^{d/2\alpha}\right\}$ where $\alpha \equiv \min\{\alpha_1, \alpha_1\} > 0$ (and *const.* can depend on M_0, but not n); see, e.g., Chen (2007) and Chen, Linton and van Keilegom (2003). Following the verifications of examples 1 and 2 in Chen, Linton and van Keilegom (2003), we have that condition (18) in Remark C.1 holds with $\kappa = 1/2$. Hence, by Remark C.1 (with $\kappa = 1/2$), we have assumption C.2(iii) is satisfied with either $C_n \leq const. \times k(n)$ if $\alpha \leq d$ or $C_n = const. < \infty$ if $\alpha > d$. By Lemma C.2 result (2) and the fact that $\frac{J_n k(n)}{n} = o(1)$, it follows $\bar{\delta}^2_{m,n} = o(1)$ and hence assumption 3.3(ii) holds.

By the mean value theorem and condition 6.1(iv), we have: for all $h, h' \in \mathcal{H}$,

$$\begin{aligned} & |m(X,h) - m(X,h')| \\ \leq \; & E\left(\sup_{t\in[0,1]} f_{Y_3|Y_1,Y_2,X}(h'(Y_1,Y_2) + t\{h-h'\}(Y_1,Y_2))|X\right)\left[\sup_{y_1}\left|h_1(y_1) - h_1'(y_1)\right| + \sup_{y_2}\left|h_2(y_2) - h_2'(y_2)\right|\right]. \end{aligned}$$

This, condition 6.1(iv), and $\sup_{x\in\mathcal{X},h\in\mathcal{H}}|m(x,h)| \leq 1$ imply that

$$E[|m(X,h)|^2] - E[|m(X,h')|^2] \leq 2E\left(|m(X,h) - m(X,h')|\right) \leq const. \times \left\|h - h'\right\|_s .$$

Thus $E[|m(X,h)|^2]$ is continuous on $(\mathcal{H}, \|\cdot\|_s)$. We have that for any $M < \infty$, the embedding of the set $\{h \in \mathcal{H} : P(h) = ||h_1||_{\Lambda^{\alpha_1}} + ||h_2||_{\Lambda^{\alpha_2}} \leq M\}$ into $\mathbf{H}$ is compact under the norm $\|\cdot\|_s = || \cdot ||_{\sup}$; hence $P(\cdot)$ is lower semicompact.

The condition $\max\left[\{k_1(n)\}^{-2r_1}, \{k_2(n)\}^{-2r_2}, \frac{J_n}{n} + J_n^{-2r_m}\right] = O(\lambda_n)$ and Theorem 3.2 now imply the desired consistency results. *Q.E.D.*

Proof of Proposition 6.2: We obtain the results by verifying that all the assumptions of Corollary 5.1 are satisfied.

We first show that $\delta_{m,n} = \frac{J_n}{n} + b^2_{m,J_n}$. Similar to the proof of Lemma 6.1, $\log N(w^{1/2}, \mathcal{H}_{osn}, ||.||_{L^\infty}) \leq \min\left\{\frac{1}{2}k(n)\log(1/w), const.(1/w)^{d/2\alpha}\right\}$ where $\alpha \equiv \min\{\alpha_1, \alpha_1\} > d$. By Remark C.1 (with $\kappa = 1/2$), we have that assumption C.2(iv) is satisfied with $C < \infty$. Thus Lemma C.2 result (3) is applicable and yields $\delta^2_{m,n} = \frac{J_n}{n} + J_n^{-2r_m} = o(1)$.

Assumptions 3.1, 3.2 and 3.3 (with $\eta_{0,n} = \delta^2_{m,n} = \frac{J_n}{n} + J_n^{-2r_m}$) are already verified in the proof of Proposition 6.1. Given the choice of the norm $||h||_s$, assumption 5.1 is satisfied with $\|h_0 - \Pi_n h_0\|_s = O\left(\{k(n)\}^{-r}\right)$ with $r = \alpha/d$. Condition 6.3(ii) implies assumption 5.2. It remains to verify assumption 4.1. By condition 6.1(iv) we have

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

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

hence assumption 4.1(i) holds. Since

$$m(X,h) - m(X,h_0) = K_{1,h}[h_1 - h_{01}](X) + K_{2,h}[h_2 - h_{02}](X),$$

condition 6.3(i) implies assumption 4.1(ii). The results now follow from Corollary 5.1. *Q.E.D.*