ESTIMATION OF NONPARAMETRIC CONDITIONAL MOMENT MODELS WITH POSSIBLY NONSMOOTH GENERALIZED RESIDUALS

BY

Xiaohong Chen and Demian Pouzo

COWLES FOUNDATION PAPER NO. 1345

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

2012

http://cowles.econ.yale.edu/
ESTIMATION OF NONPARAMETRIC CONDITIONAL MOMENT MODELS WITH POSSIBLY NONSmooth GENERALIZED RESIDUALS

BY XIAOHONG CHEN AND DEMIAN POUZO

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 noncompact 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.

1. INTRODUCTION

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

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

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. $\rho(\cdot)$ is a vector of generalized

---

This paper is a substantially revised version of Cowles Foundation Discussion Paper 1650. We are grateful to two co-editors, three anonymous referees, R. Blundell, V. Chernozhukov, T. Christensen, K. Evdokimov, H. Hong, J. Horowitz, J. Huang, S. Lee, Z. Liao, O. Linton, D. Nekipelov, J. Powell, A. Santos, E. Tamer, and A. Torgovitsky for their constructive comments that led to a much improved revision. Earlier versions were presented in August 2006 at the European Meeting of the Econometric Society, in March 2007 at the Oberwolfach Workshop on Semi/nonparametrics, in June 2007 at the Cemmap Conference on Measurement Matters, in 2008 at the Cowles Summer Conference, and at econometric workshops at many universities. We thank participants of these conferences and workshops for helpful suggestions. Chen acknowledges financial support from the National Science Foundation under Grant SES0631613 and SES0838161. The usual disclaimer applies.

© 2012 The Econometric Society

DOI: 10.3982/ECTA7888
residuals with functional forms known up to a finite-dimensional parameter \( \theta_0 \) and functions of interest \( h_0(\cdot) \equiv (h_{01}(\cdot), \ldots, h_{0q}(\cdot)) \), where each function \( h_{0\ell}(\cdot) \), \( \ell = 1, \ldots, q \), may depend on different components of \( X \) and \( Y \), and some could depend on \( \theta_0 \) and \( h_{0\ell'}(\cdot) \) for \( \ell' \neq \ell \). In this paper, \( \rho(\cdot) \) may depend on the unknown \((\theta_0, h_0)\) 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 nonsmooth with respect to the unknown parameters of interest \((\theta, h)\). As these papers have already illustrated, many semi/nonparametric structural models in economics are special cases of (1). For instance, our model 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 (2011) (DFFR), Chen and Reiss (2011) (CR), and Hall and Horowitz (2005) (HH):

\[
E[Y_1 - h_0(Y_2)|X] = 0.
\]

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):

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

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 \((\theta, h)\). For instance, Chen and Ludvigson (2009) studied 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, for example, Gallant and Tauchen (1989) and 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 \((\theta, h)\). See, for example, Arellano (2008) for an economic general equilibrium model and Chen and Pouzo (2009b) for an econometric study of pricing default risk.

As demonstrated in NP, AC, CIN, and Chen, Chernozhukov, Lee, and Newey (2011) (CCLN), the key difficulty of analyzing the semi-nonparametric model (1) is not the presence of the unknown finite-dimensional parameter \(\theta_0\), but the fact that some of the unknown functions \(h_{0\ell}(\cdot), \ell = 1, \ldots, q\), depend on the endogenous variable \(Y\).\(^2\) Therefore, in this paper, we focus on the nonparametric estimation of \(h_0(\cdot)\), which is identified by the conditional moment restrictions

\[
E\left[ \rho(Y, X; h_{01}(\cdot), \ldots, h_{0q}(\cdot)) | X \right] = 0, \tag{4}
\]

where \(h_0(\cdot) \equiv (h_{01}(\cdot), \ldots, h_{0q}(\cdot))\) depends on \(Y\) and may enter \(\rho(\cdot)\) non-linearly and possibly nonsmoothly.\(^3\) 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,

\[
E\{E[\rho(Z; h(\cdot)) | X]'W(X)E[\rho(Z; h(\cdot)) | X]\}
\]

over sieve spaces \((\mathcal{H}_n)\) that are dense in the infinite-dimensional function space \(\mathcal{H}\).\(^4\) 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., \(\dim(\mathcal{H}_n)/n \to \text{const.} > 0\)), with lower semicompact\(^5\) 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 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).

\(^2\)In some applications, the presence of the parametric part \(\theta_0\) in the semi-nonparametric model (1) aids the identification of the unknown function \(h_0\); see, for example, Chen and Ludvigson (2009) and CCLN.

\(^3\)See Chen and Pouzo (2009a) for semiparametric efficient estimation of the parametric part \(\theta_0\) 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.

\(^4\)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.

\(^5\)See Section 2 for its definition.
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, for example, 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., \( \int \{h(y)\}^2 \, dy \) or \( \int \{\nabla^r h(y)\}^2 \, dy \) with \( \nabla^r h \) being the \( r \)th derivative of \( h \)); see, for example, 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 \( \mathcal{H} \). For example, in the simulation study of the NPQIV model (3), HL approximated 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 used a small number of Chebyshev polynomial series terms to approximate \( h_0 \) so as 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)). 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 many published papers on the asymptotic properties of the SMD and the TR-MD procedures for the linear NPIV model (2). For example, see NP for consistency of the SMD estimator in a (weighted) sup-norm; see BCK and CR for the convergence rate in a root mean squared norm of the SMD estimator; see 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; and see HH and CR for the minimax optimal rate in a root mean squared norm for the NPIV model.

---

\(^6\)This is because numerical optimization algorithms cannot handle infinite-dimensional objects in \( \mathcal{H} \).

\(^7\)This is because a constrained optimization problem is typically more difficult to compute than the corresponding unconstrained optimization problem.

\(^8\)See NP, DFFR, BCK, CFR, Severini and Tripathi (2006), D'Haultfoeuille (2011), Florens, Johannes, and van Bellegem (2011) and others for identification of the NPIV model.
There are currently only a few published papers on the asymptotic properties of any nonparametric 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 pseudometric weaker than $\| \cdot \|_s$. For the NPQIV example (3), 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 presented 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 ($\dim(\mathcal{H}_n) \rightarrow \infty$, $\dim(\mathcal{H}_n)/n \rightarrow 0$) 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 ($\dim(\mathcal{H}_n)/n \rightarrow \text{const.} > 0$) 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 $h_0(\cdot)$ 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 $E[1\{Y_3 \leq h_{01}(Y_1) + h_{02}(Y_2)|X\} |X} = \gamma \in (0, 1)$, 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

---

9See 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.
well in finite samples. Section 8 briefly concludes. Some regularity conditions and general lemmas are stated in the Appendix. The Supplemental Material (Chen and Pouzo (2012)) contains all the proofs, as well as a brief review of some functional spaces and sieve bases.

**Notation**

In this paper, we denote \( f_{AB}(a; b) \) \((F_{AB}(a; b))\) as the conditional probability density (c.d.f.) of random variable \( A \) given \( B \) evaluated at \( a \) and \( b \) and denote \( f_{AB}(a, b) \) \((F_{AB}(a, b))\) as the joint density (c.d.f.) of the random variables \( A \) and \( B \). We denote \( L^p(\Omega, d\mu) \) as the space of measurable functions with \( \|f\|_{L^p(\Omega, d\mu)} \equiv \left\{ \int_{\Omega} |f(t)|^p \, d\mu(t) \right\}^{1/p} < \infty \), where \( \Omega \) 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 \( \epsilon > 0 \), \( \lim_{n \to \infty} \Pr(a_n/b_n > \epsilon) = 0 \). We use w.p.a.1 to denote with probability approaching 1. 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 \( |A| = \|A\|_\ell \equiv \sqrt{A'AA} \) 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 and 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 \( \mathcal{Z} \equiv (Y', X')' \in \mathcal{Z} \equiv \mathcal{Y} \times \mathcal{X} \) and \( \mathcal{X} \subseteq \mathcal{X} \). Suppose that the unknown distribution of \( (Y', X') \) satisfies the conditional moment restriction (4), where \( p: \mathcal{Z} \times \mathcal{H} \to \mathcal{R}^d_Y \) 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 \( \mathcal{H}_j, j = 1, \ldots, q \), being a space of real-valued measurable functions whose arguments vary across indices. We assume that the parameter space \( \mathcal{H} \) is a nonempty, closed, possibly noncompact infinite-dimensional subset of \( \mathcal{H} \equiv \mathcal{H}_1 \times \cdots \times \mathcal{H}_q \), a separable Banach space with norm \( \|h\|_s \equiv \sum_{t=1}^q \|h_t\|_{s,t} \).

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 \( j = 1, \ldots, d_p \). 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 \( j = 1, \ldots, d_p \). (Note that the functional form of \( m_j(X, h) \) is unknown since the conditional distribution \( F_{Y|X} \) is not specified.) Let \( m(X, h) \equiv (m_1(X, h), \ldots, m_{d_p}(X, h))' \) and let \( W(X) \) be a positive-definite finite weight-
ing matrix for almost all \(X\). Under the assumption that model (4) identifies \(h_0 \in \mathcal{H}\), we have

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

One could construct an estimator of \(h_0 \in \mathcal{H}\) by minimizing a sample analog of \(E[\|m(X, h)\|_W^2]\) over the function space \(\mathcal{H}\). Unfortunately, when \(h_0(\cdot)\) depends on the endogenous variables \(Y\), the “\(\| \cdot \|_s\) identifiability 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 \(\liminf_{k \to \infty} E[\|m(X, h_k)\|_W^2] = 0\); that is, the metric \(\| h - h_0 \|_s\) is not continuous with respect to the population criterion function \(E[\|m(X, h)\|_W^2]\), and the problem is ill-posed.\(^{10}\)

When \(E[\|m(X, h)\|_W^2]\) is lower semicontinuous on \((\mathcal{H}, \| \cdot \|_s)\) and \(h_0 \in \mathcal{H}\) is its unique minimizer, one way to ensure the “\(\| \cdot \|_s\) identifiability uniqueness” is to assume that the parameter space \(\mathcal{H}\) is a compact subset of \((\mathcal{H}, \| \cdot \|_s)\); see, for example, 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 \(\Lambda^\alpha\)) or a Sobolev space (denoted as \(W_\alpha^p\)) with \(\alpha > 0\),\(^{11}\) but it could be a noncompact 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 \(\mathcal{H} = \{ h \in W_2^\alpha(f_{Y_2}) : \sup_y |h(y)| \leq 1, \| \nabla^\alpha h \|_{L^2(f_{Y_2})} < \infty \}\) for some \(\alpha \geq 1\), which is a smooth function space, but is neither \(\| \cdot \|_{L^2(f_{Y_2})}\) compact nor \(\| \cdot \|_{L^{\infty}(\text{leb})}\) compact. To allow for wider applicability, in this paper, we assume that the parameter space \(\mathcal{H}\) is an infinite-dimensional, possibly noncompact subset of a separable Banach space \((\mathcal{H}, \| \cdot \|_s)\).\(^{12}\)

To design a consistent estimator for \(h_0 \in \mathcal{H}\) with possibly noncompact parameter space \(\mathcal{H}\), we need to tackle two issues. First, we need to replace the unknown population minimum distance criterion, \(E[\|m(X, h)\|_W^2]\), by a con-

\(^{10}\)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 \(j = 1, \ldots, d_p\).

\(^{11}\)See Chen (2007) and the Supplemental Material for definitions of Hölder space, Sobolev space, Besov space, and other widely used function spaces in economics.

\(^{12}\)A subset of \((\mathcal{H}, \| \cdot \|_s)\) 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 \((\mathcal{H}, \| \cdot \|_s)\) is \(\| \cdot \|_s\) compact if and only if it is finite dimensional.
sistent 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, \( \hat{h}_n \), defined as

\[
\hat{Q}_n(\hat{h}_n) \leq \inf_{h \in \mathcal{H}_n} \hat{Q}_n(h) + \hat{\eta}_n \quad \text{with} \quad \hat{\eta}_n \geq 0, \quad \hat{\eta}_n = O_p(\eta_n)
\]

and

\[
\hat{Q}_n(h) \equiv \frac{1}{n} \sum_{i=1}^{n} \hat{m}(X_i, h) \hat{W}(X_i) \hat{m}(X_i, h) + \lambda_n \hat{P}_n(h),
\]

where \( \{ \eta_n \}_{n=1}^\infty \) is a sequence of positive real values such that \( \eta_n = o(1) \), \( \hat{m}(X, h) \) is any nonparametric consistent estimator of \( m(X, h) \), \( \mathcal{H}_n \equiv \mathcal{H}_1 \times \cdots \times \mathcal{H}_q \) is a sieve parameter space whose complexity (denoted as \( k(n) \equiv \dim(\mathcal{H}_n) \)) grows with sample size \( n \) and becomes dense in the original function space \( \mathcal{H} \), \( \lambda_n \geq 0 \) is a penalization parameter such that \( \lambda_n \to 0 \) as \( n \to \infty \), the penalty \( \hat{P}_n(\cdot) \geq 0 \) is an empirical analog of a nonrandom penalty function \( P : \mathcal{H} \to [0, +\infty) \), and \( \hat{W}(X) \) is a consistent estimator of \( W(X) \) that is introduced to address potential heteroskedasticity. In this paper, we assume that \( \hat{m}(\cdot, h), \hat{W}(\cdot), \) and \( \hat{P}_n(h) \) are 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 \( \hat{h}_n \) exists.\(^{13}\)

The sieve space \( \mathcal{H}_n \) in the definition of the PSMD estimator (6) could be finite dimensional, infinite dimensional, compact, or noncompact (in \( \| \cdot \|_s \)). Commonly used finite-dimensional linear sieves (also called series) take the form:

\[
\mathcal{H}_n = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot) \right\},
\]

\[
k(n) < \infty, \quad k(n) \to \infty \text{ slowly as } n \to \infty,
\]

where \( \{ q_k \}_{k=1}^\infty \) is a sequence of known basis functions of a Banach space \( (\mathcal{H}, \| \cdot \|_s) \) such as wavelets, splines, Fourier series, and Hermite polynomial series.\(^{14}\)

\(^{13}\)In this paper, we implicitly assume that \( \hat{h}_n \) 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 Appendix A for sufficient conditions to ensure measurability.

\(^{14}\)See Chen and Shen (1998), Chen (2007), and the references therein for additional examples of linear sieves (or series) and nonlinear sieves.
Commonly used linear sieves with constraints can be expressed as

\[
H_n = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot), R_n(h) \leq B_n \right\},
\]

where the constraint \( R_n(h) \leq B_n \) reflects prior information about \( h_0 \in \mathcal{H} \), such as smoothness properties. The sieve space \( H_n \) in (9) is finite dimensional and compact (in \( \| \cdot \|_s \)) if and only if \( k(n) < \infty \) and \( H_n \) is closed and bounded; it is infinite dimensional and compact (in \( \| \cdot \|_s \)) if and only if \( k(n) = \infty \) and \( H_n \) is closed and totally bounded. For example, \( H_n = \{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot), \| h \|_s \leq \log(n) \} \) is compact if \( k(n) < \infty \), but it is not compact if \( k(n) = \infty \).

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

\[
\frac{1}{n} \sum_{i=1}^{n} \| \hat{m}(X_i, \hat{h}_n) \|_{\hat{W}}^2 \leq \inf_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \| \hat{m}(X_i, h) \|_{\hat{W}}^2,
\]

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

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

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
\[ f(\nabla^2 h(y))^2 dy \], 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: (i) PSMD using slowly growing finite-dimensional sieves \( (k(n)/n \to 0) \), with small flexible penalty \( (\lambda_nP(\cdot) \searrow 0 \text{ fast}) \) or zero penalty \( (\lambda_nP(\cdot) = 0) \); and (ii) PSMD using large dimensional sieves \( (k(n)/n \to \text{const.} > 0) \), with positive penalty \( (\lambda_nP(\cdot) > 0) \) 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.

On the one hand, 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 \( H_n \) given in (9) be a finite-dimensional compact sieve. However, it also allows for \( \lambda_n \searrow 0 \) fast with \( H_n \) 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 TR-MD procedure as a special case by setting \( H_n = \mathcal{H} \) (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 \( k(n)/n \to \text{const.} > 0 \) and \( \lambda_n \searrow 0 \) slowly, which is computationally much easier than the TR-MD procedure.

When \( n^{-1} \sum_{i=1}^{n} \| \hat{m}(X_i, h) \|_{\hat{P}}^2 \) 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 \( \lambda_n \hat{P}_n(h) \) 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

\[
\hat{Q}_n(h_n) + v_nR_n(h_n) \leq \inf_{h \in \text{clsp}(\mathcal{H}_n)} \{ \hat{Q}_n(h) + v_nR_n(h) \} + O_p(\eta_n),
\]

where \( R_n(h_n) \leq B_n \) and \( v_n \geq 0 \) is such that \( v_n(R_n(h_n) - B_n) = 0 \); 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 \( v_nR_n(h) \). For most applications, it suffices to have either \( \lambda_n \hat{P}_n(h) > 0 \) or \( v_nR_n(h) > 0 \).
Even when \( n^{-1} \sum_{i=1}^{n} \|\hat{m}(X_i, h)\|_{\hat{W}}^2 \) 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 applying either the class of PSMD estimators using slowly growing finite-dimensional sieves with/without small flexible penalty \( (k(n) \to \infty \text{ slowly}, k(n)/n \to 0; \lambda_n \downarrow 0 \) fast or \( \lambda_n = 0 \)) or the class of PSMD estimators using faster growing finite-dimensional sieves with large lower semicompact penalty \( (k(n) \to \infty \text{ faster}, k(n)/n \to 0; \lambda_n = O(k(n)/n)) \). Our subsequent theoretical results and simulation studies indicate that these two classes of estimators perform well in finite samples and they can achieve the optimal rate of convergence under weaker assumptions than can the class of PSMD estimators using large dimensional sieves with large lower semicompact penalty \( (k(n)/n \to \text{const.} > 0; \lambda_n \downarrow 0 \text{ slowly}) \). Between 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 \( \hat{h}_n \) 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 lets \( \hat{W}(\cdot) = W(\cdot) = I \) (identity) and \( \hat{m}(\cdot, h) \) 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 \( \hat{h}_n \), allowing for any nonparametric estimators of \( m(\cdot, h) \) and \( W(\cdot) \) that satisfy a mild regularity assumption, 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

\[
\hat{m}(X, h) = p'^n(X)'(P'P)^{-1} \sum_{i=1}^{n} p'^n(X_i)\rho(Z_i, h),
\]

where \( \{p_j(\cdot)\}_{j=1}^{\infty} \) 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 \( n \to \infty \), \( p^h_n(X) = (p_1(X), \ldots, p_{I_n}(X))' \), \( P = (p^h_n(X_1), \ldots, p^h_n(X_n))' \), and \( (P')^{-1} \) is the generalized inverse of the matrix \( PP \). See NP, AC, BCK, CIN, and CR for more details and applications of this estimator.

3. CONSISTENCY

In Appendix A, 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 the consistency of various PSMD estimators (6).

We first impose three basic conditions on identification: sieve 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) \( E[p(Z, h_0)]X = 0 \) and \( \|h_0 - h\|_2 = 0 \) for any \( h \in (\mathcal{H}, \| \cdot \|_2) \) with \( E[p(Z, h)]X = 0 \); (iii) \( \{h_k: k \geq 1\} \) is a sequence of nonempty closed subsets satisfying \( \mathcal{H}_k \subseteq \mathcal{H}_{k+1} \subseteq \mathcal{H} \), and there is \( \Pi_n \mathcal{H}_0 \in \mathcal{H}_{k(n)} \) such that \( \|\Pi_n \mathcal{H}_0 - h_0\|_2 = o(1) \); (iv) \( E[\|m(X, \Pi_n \mathcal{H}_0)\|_{\mathcal{H}}^2] = o(1) \).

ASSUMPTION 3.2—Penalty: One of the following holds: (a) \( \lambda_n = 0 \), (b) \( \lambda_n > 0 \), \( \lambda_n = o(1) \), \( \sup_{h \in \mathcal{H}_{k(n)}} |\hat{P}_n(h) - P(h)| = O_p(1) \), and \( |P(\Pi_n \mathcal{H}_0) - P(h_0)| = O(1) \) with \( P: \mathcal{H} \to [0, \infty) \), \( P(h_0) < \infty \), or (c) \( \lambda_n > 0 \), \( \lambda_n = o(1) \), \( \sup_{h \in \mathcal{H}_{k(n)}} |\hat{P}_n(h) - P(h)| = o_p(1) \), and \( |P(\Pi_n \mathcal{H}_0) - P(h_0)| = O(1) \) with \( P: \mathcal{H} \to [0, \infty) \), \( P(h_0) < \infty \).

Let \( \{\eta_{0,n}\}_{n=1}^{\infty} \) and \( \{\delta_{m,n}\}_{n=1}^{\infty} \) be sequences of positive real values that decrease to zero as \( n \to \infty \). Let \( \mathcal{H}_{M_0} = \{h \in \mathcal{H}_{k(n)}: \lambda_n P(h) \leq \lambda_n M_0\} \) for a large but finite \( M_0 \) such that \( \Pi_n \mathcal{H}_0 \in \mathcal{H}_{M_0} \) and that \( \hat{h}_n \in \mathcal{H}_{M_0} \) with probability arbitrarily close to 1 for all large \( n \). Given Assumptions 3.2 and 3.3(i), such an \( M_0 \) always exists (see Lemma A.3 in Appendix A).

ASSUMPTION 3.3—Sample Criterion: (i) \( \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, \Pi_n \mathcal{H}_0)\|_{\mathcal{H}}^2 \leq c_0 \times E[\|m(X, \Pi_n \mathcal{H}_0)\|_{\mathcal{H}}^2] + O_p(\eta_{0,n}) \) for some \( \eta_{0,n} = o(1) \) and a finite constant \( c_0 > 0 \); (ii) \( \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, h)\|_{\mathcal{H}}^2 \geq cE[\|m(X, h)\|_{\mathcal{H}}^2] - O_p(\delta_{m,n}^2) \) uniformly over \( \mathcal{H}_{k(n)}^M \) for some \( \delta_{m,n}^2 = o(1) \) and a finite constant \( c > 0 \).

Under Assumption 3.1(ii) (global identification) and (iii) (definition of sieves), Assumption 3.1(iv) is satisfied if \( E[\|m(X, h)\|_{\mathcal{H}}^2] \) is continuous at \( h_0 \) (under \( \| \cdot \|_2 \)). Assumption 3.2(b) and (c) are trivially satisfied when \( \mathcal{H}_{k(n)} = \mathcal{H} \) and \( \hat{P}_n = P \). 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 \( |P(\Pi_n \mathcal{H}_0) - P(h_0)| = o(1) \) 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 Appendix C shows that the series LS estimator \( \widehat{m}(X, h) \) defined in (11) satisfies Assumption 3.3.

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

\[
\inf_{h \in \mathcal{H}_n} \left\{ E\left[ \| m(X, h) \|_W^2 \right] + \lambda_n P(h) \right\} + O_p(\eta_n)
\]

for some positive sequence \( \eta_n = o(1) \), which has a solution, provided that the set \( \{ h \in \mathcal{H}_n : E[\| m(X, h) \|_W^2] + \lambda_n P(h) \leq M \} \) is compact in some topology \( T \) (that may be weaker than the norm \( \| \cdot \|_s \) topology on \( \mathcal{H} \)). Further, when \( E[\| m(X, h) \|_W^2] \) has a unique minimizer \( (h_0) \) on \( (\mathcal{H}, \| \cdot \|_s) \), we establish \( \| \cdot \|_s \) consistency of \( \hat{h}_n \) under some assumptions over the penalty and the smoothing parameters. This explains why \( \| \cdot \|_s \) consistency of \( \hat{h}_n \) can be obtained 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 discussion, 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 \( (k(n)/n \to \text{const.} > 0) \) or infinite-dimensional sieves in separate subsections.

### 3.1. PSMD Using Slowly Growing Finite-Dimensional Sieves

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

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

\[
\max\{ \eta_{0,n}, E[\| m(X, \Pi_n h_0) \|_W^2], \delta_{m,n}^2, \lambda_n \} = o(g(k(n), \varepsilon))
\]

for all \( \varepsilon > 0 \),

then \( \| \hat{h}_n - h_0 \|_s = o_p(1) \), and \( P(\hat{h}_n) = O_p(1) \) 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 \( \lambda_n = 0 \), \( \lim_{k(n) \to \infty} g(k(n), \varepsilon) = \)
\[
\inf_{h \in \mathcal{H}; \|h - h_0\|_s \geq \varepsilon} E[\|m(X, h)\|_W^2].
\]
Thus, given Assumption 3.1(ii) (identification), for all \(\varepsilon > 0\), \(\liminf_{k(n) \to \infty} g(k(n), \varepsilon) > 0\) if \((\mathcal{H}, \| \cdot \|_s)\) is compact; otherwise, \(\liminf_{k(n) \to \infty} g(k(n), \varepsilon)\) could be zero. When \((\mathcal{H}, \| \cdot \|_s)\) is compact, restriction (12) becomes \(\max\{\eta_{0,n}, E[\|m(X, \Pi_n h_0)\|_W]\}\). 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) and \(\mathcal{H} = \{h \in L^2(f_{Y_2}); \|\nabla^r h\|_{L^2(f_{Y_2})} < \infty, r > 0\}\), which is not compact in \(\| \cdot \|_s\). 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 \(\{\mu_k; \phi_{1,k}, \phi_{0,k}\}_k^{\infty}\) where \(\{\mu_k\}_k^{\infty}\) are the singular numbers arranged in nonincreasing order \(\mu_k \geq \mu_{k+1} \wedge 0\), and \(\{\phi_{1,k}(\cdot)\}_k^{\infty}\) and \(\{\phi_{0,k}(\cdot)\}_k^{\infty}\) are eigenfunctions in \(L^2(f_{Y_2})\) and \(L^2(f_X)\), respectively. Let \(\mathcal{H}_n = \{h \in \mathcal{H}; h(Y_2) = \sum_{k=1}^{k(n)} a_k \phi_{1,k}(Y_2), \|\nabla^r h\|_{L^2(f_{Y_2})} \leq \log(n)\}\) and \(\lambda_n P(h) = \lambda_n \|\nabla^r h\|_{L^2(f_{Y_2})}^2\) for \(\lambda_n \geq 0\).

Note that \(E[\|m(X, h)\|_W^2]\) is continuous on \((\mathcal{H}_n, \| \cdot \|_s)\) and

\[
E\left[\|m(X, \Pi_n h_0)\|_W^2\right] = E\left[\left(E[\Pi_n h_0(Y_2) - h_0(Y_2)|X]\right)^2\right] = \sum_{j=k(n)+1}^{\infty} \mu_j^2 \|\phi_{0,j}\|_{L^2(f_{Y_2})}^2 \leq \mu_{k(n)+1}^2 \sum_{j=k(n)+1}^{\infty} \|\phi_{1,j}\|_{L^2(f_{Y_2})}^2 = \mu_{k(n)+1}^2 \|\Pi_n h_0 - h_0\|_s^2.
\]
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

\[
h_n^* = \arg\min_{h \in \mathcal{H}_n; \|h - h_0\|_s \geq \varepsilon} E[\left(E[h(Y_2) - h_0(Y_2)|X]\right)^2].
\]
Then

\[
g(k(n), \varepsilon) \geq E\left[\left(E[h_n^*(Y_2) - h_0(Y_2)|X]\right)^2\right] = \sum_{j=1}^{\infty} \mu_j^2 \|h_j^* - h_0, \phi_{1,j}\|_{L^2(f_{Y_2})}^2.
\]
\[ \geq \mu^2_k \sum_{j=1}^{k(n)} (h_n^* - h_0, \phi_{1,j})_{L^2(f_{Y_2})}^2 \]
\[ = \mu^2_k \| h_n^* - \Pi_nh_0 \|_s^2. \]

Note that \[ \| h_n^* - \Pi_nh_0 \|_s \] is bounded below by a constant \( c(\varepsilon) > 0 \) for all \( k(n) \) large enough; otherwise, there is a large \( k(n) \) such that \( \| h_n^* - \Pi_nh_0 \|_s < (\varepsilon / 3 / 2 \times \| \Pi_nh_0 - h_0 \|_s = o(1) \). By letting \( \max\{ \eta_{0,n}, \delta^2_{m,n}, \lambda_n \} / g(k(n), \varepsilon) = o(1) \), Theorem 3.1 is applicable; hence, \( \| \hat{h}_n - h_0 \|_s = o_P(1) \).

### 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 \( \hat{\eta}_n \) be the PSMD estimator with \( \lambda_n > 0 \) and \( \eta_n = O(\eta_{0,n}) \), and let Assumptions 3.1, 3.2(b), and 3.3 hold. Suppose that \( P(\cdot) \) is lower semicompact and that \( E[\| m(X, h) \|_{W}] \) is lower semicontinuous on \( (H, \| \cdot \|_s) \). If

\[ \max\{ \eta_{0,n}, E[\| m(X, \Pi_nh_0) \|_{W}] \} = O(\lambda_n), \]

then \( \| \hat{h}_n - h_0 \|_s = o_P(1) \) and \( P(\hat{h}_n) = O_p(1) \).

The lower semicompact penalty implies that the effective parameter space \( \{ h \in H : P(h) \leq M_n \} \) with \( M_n \rightarrow \infty \) slowly is compact in the \( \| \cdot \|_s \) topology and, hence, converts an ill-posed problem into a well-posed one.\(^{15}\) Theorem 3.2 applies to the class of PSMD estimators with any positive lower semicompact penalty functions, allowing for \( k(n) = \infty \) or \( k(n)/n \rightarrow \text{const.} \geq 0 \). 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 \( W = I \), \( E[\| m(X, h) \|_W] \) is obviously lower semicontinuous on \( (H, \| \cdot \|_s) \) with a norm \( \| h \|_s = \| h \|_{L^2(\mathbb{R}^d, f_{Y_2})} = \sup_{y \in \mathbb{R}^d} (1 + |y|^2)^{-\theta/2} h(y) \) for some \( \theta \geq 0 \). For a penalty function \( P(h) \) to be lower semicompact, it suffices that the embedding of the

\(^{15}\) We thank Victor Chernozhukov for pointing out this nice property of lower semicompact penalties.
set \( \{ h \in \mathcal{H} : P(h) \leq M \} \) into \( \mathcal{H} \) is compact for all \( M \in [0, \infty) \). For example, if \( \| \cdot \|_s = \| \cdot \|_{L^2(f_{Y_2})} \), then \( P(h) = \|(1 + | \cdot |^2)^{-\theta/2} h(\cdot)\|_{W^{p}_{\alpha}(\mathcal{R}^d)}^p \) with \( 0 < p \leq 2 \), \( \alpha > \frac{d}{2} - \frac{d}{p} \), \( \theta \geq 0 \), and \( f_{Y_2}(y_2) |y_2|^\theta \to 0 \) 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 \( P(h) = \|(1 + | \cdot |^2)^{-\theta/2} h(\cdot)\|_{\Lambda^p(\mathcal{R}^d)}^p \) with \( \alpha > 0 \), \( \theta > \theta \) and \( P(h) = \|(1 + | \cdot |^2)^{-\theta/2} h(\cdot)\|_{W^{p,\alpha}(\mathcal{R}^d)}^p \) with \( 0 < p < \infty \) and \( \alpha > \frac{d}{p} \), \( \theta > \theta \) are lower semicompact; see Edmunds and Triebel (1996). Theorem 3.2 immediately implies \( \| \hat{h}_n - h_0 \|_{L^2(f_{Y_2})} = o_p(1) \) or \( \sup_{y \in \mathcal{R}^d} |(1 + |y|^2)^{-\theta/2}[\hat{h}_n(y) - h_0(y)]| = o_p(1) \). 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 original SMD estimator

\[
\hat{h}_n = \arg \min_{h \in \mathcal{H}_{k(n)} : P(h) \leq M_n} \frac{1}{n} \sum_{i=1}^{n} \| \hat{m}(X_i, h) \|_{W^{\alpha}(\mathcal{R}^d)}^2,
\]

using finite-dimensional compact sieves \( \{ h \in \mathcal{H}_{k(n)} : P(h) \leq M_n \} \) (with \( M_n \to \infty \) slowly), is equivalent to our PSMD estimator (6) using finite-dimensional linear sieves \( \mathcal{H}_{k(n)} \) (e.g., (8)). Therefore, Theorem 3.2 also establishes the consistency of the original SMD estimator using finite-dimensional compact sieves of the type \( \{ h \in \mathcal{H}_{k(n)} : P(h) \leq M_n \} \) 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 \( \mathcal{H} \) be 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 denote a bilinear form \( \langle \cdot, \cdot \rangle_{\mathbf{H}^* \times \mathbf{H}} : \mathbf{H}^* \times \mathbf{H} \to \mathcal{R} \) as the inner product that links the space \( \mathbf{H} \) with its dual \( \mathbf{H}^* \). A Banach space \( \mathbf{H} \) is **reflexive** if and only if (iff) \( (\mathbf{H}^*)^* = \mathbf{H} \). For example, the spaces \( L^p \) for \( 1 < p < \infty \) and the Sobolev spaces \( W^{\alpha}_{p} \) for \( 1 < p < \infty \) are reflexive and separable Banach spaces.

**Assumption 3.4:** (i) There is a \( t_0 \in \mathbf{H}^* \) with \( \langle t_0, \cdot \rangle_{\mathbf{H}^* \times \mathbf{H}} \) a bounded linear functional with respect to \( \| \cdot \|_s \), and a nondecreasing lower semicontinuous function \( g(\cdot) \) with \( g(0) = 0, g(\varepsilon) > 0 \) for \( \varepsilon > 0 \), such that \( P(h) - P(h_0) - \langle t_0, h - h_0 \rangle_{\mathbf{H}^* \times \mathbf{H}} \geq g(\| h - h_0 \|_s) \) for all \( h \in \mathcal{H}_k \) and all \( k \geq 1 \). (ii) \( (\mathbf{H}, \| \cdot \|_s) \) is a reflexive Banach space, and \( \mathcal{H} \) 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 $\| \cdot \|$; that is, there exists a $c > 0$ such that $P(h) - P(h_0) - \langle DP(h_0), h - h_0 \rangle_{W, H} \geq c \| h - h_0 \|_s^2$ for all $h \in \mathcal{H}$, where $DP(h_0) \in \mathbf{H}^*$ is the Gateaux derivative of $P(\cdot)$ at $h_0$. We note that strong convexity is satisfied by commonly used penalization functions; see, for example, Eggermont and LaRiccia (2001). In Assumption 3.4(ii), we note that strong convexity is satisfied by commonly used penalization functions; see, for example, Eggermont and LaRiccia (2001). In Assumption 3.4(ii), the condition that $\mathcal{H}$ is bounded in $\| \cdot \|_s$ (i.e., $\sup_{h \in \mathcal{H}} \| h \|_s \leq K < \infty$) can be replaced by the so-called coercive condition $E[\| m(X, h) \|_{W}^2] + \lambda P(h) \rightarrow +\infty$ as $\| h \|_s \rightarrow +\infty$ for $\lambda \in (0, 1]$.

A functional $G : \mathcal{H} \rightarrow (-\infty, +\infty)$ is weak sequentially lower semicontinuous at $h \in \mathcal{H}$ iff $G(h) \leq \liminf_{j \rightarrow \infty} G(h_j)$ 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 $\lambda_n > 0$ and $\eta_n = o(\eta_{0,n})$, and let Assumptions 3.1, 3.2(c), 3.3, and 3.4 hold. Let $E[\| m(X, h) \|_{W}^2]$ be weak sequentially lower semicontinuous on $\mathcal{H}$. If

\begin{equation}
\max \{ \eta_{0,n}, E[\| m(X, \Pi_n h_0 \|_{W}^2] \} = o(\lambda_n),
\end{equation}

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

**Remark 3.2:** Under Assumption 3.4(ii), $E[\| m(X, h) \|_{W}^2]$ is weak sequentially lower semicontinuous on $\mathcal{H}$ if any one of the following holds:

(i) $E[\| m(X, \cdot \|_{W}^2]$ is convex and lower semicontinuous on $(\mathcal{H}, \| \cdot \|_s)$,

(ii) $E[\| m(X, \cdot \|_{W}^2] \cdot \mathcal{H} \rightarrow [0, \infty)$ has compact Gateaux derivative on $\mathcal{H}$, or

(iii) $\sqrt{W(\cdot)m(\cdot, h) : \mathcal{H} \rightarrow L^2(f_X)}$ is compact and Frechet differentiable.

**NPV 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 $\mathbf{H} = L^2(f_Y)$, $\| \cdot \|_s = \| \cdot \|_{L^2(f_Y)}$ and $\mathcal{H} = \{ h \in L^2(f_Y) : \| h \|_{L^2(f_Y)} \leq M < \infty \}$. $E[\| m(X, h) \|_{W}^2] = E[(E[Y_l - h(Y_l)|X])^2]$ is convex and lower semicontinuous on $\mathcal{H}$ and, hence, is weak sequentially lower semicontinuous on $\mathcal{H}$. Let $P(h) = \| h \|_{L^2(f_Y)}^2$ be the penalty function. Then Assumption 3.4(i) is satisfied with $t_0 = 2h_0$. Theorem 3.3 immediately leads to $\| \hat{h}_n - h_0 \|_{L^2(f_Y)}^2 = o_p(1)$.

### 3.2.3. Choice of Penalty Functions

Comparing Theorem 3.3 to Theorem 3.2, both consistency results allow for noncompact (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.

**Identification via a 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 identification 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 $M_0 \equiv \{h \in \mathcal{H} : E[\|m(X, h)\|_W^2] = 0\}$ uniquely. See Theorem A.1 in Appendix A.

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

\begin{align*}
\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
\end{align*}

for some positive finite constants $M_1$ and $M_0$, and a sufficiently small positive $\epsilon$ 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 pseudometric 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

\begin{align*}
\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 \text{a.s. } \mathcal{X}.
\end{align*}

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

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

**ASSUMPTION 4.1—Local Curvature:** (i) $\mathcal{H}_{os}$ and $\mathcal{H}_{osn}$ are convex, and $m(X, h)$ 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 pseudometric \( \| h - h_0 \| \) 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 NIV model (2) and \( \| h - h_0 \| = \sqrt{E[(E[f(Y)|Y_2](h_0(Y_2))|h_0(Y_2)]|X])^2} \) 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 pseudometric \( \| h - h_0 \| \) 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 NIV model). Assumption 4.1 enables us to obtain a fast convergence rate of \( \| \hat{h} - h_0 \| \) 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 \( \theta_0 \) for the semi/nonparametric conditional moment restrictions \( E[\rho(Y, X_2; \theta_0, h_0(\cdot))|X] = 0 \). Here we use the same trick to derive the nonparametric convergence rate of \( \| \hat{h} - h_0 \|_s \).

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

\[
\omega_n(\delta, \mathcal{H}_{osn}) = \sup_{h \in \mathcal{H}_{osn}: \| h - \Pi_n h_0 \| \leq \delta} \| h - \Pi_n h_0 \|_s,
\]

\[
\omega(\delta, \mathcal{H}_{os}) = \sup_{h \in \mathcal{H}_{os}: \| h - h_0 \| \leq \delta} \| h - h_0 \|_s.
\]

The definition of the modulus of continuity, \( \omega(\delta, \mathcal{H}_{os}) \) does not depend on the choice of any estimation method. Therefore, when \( \omega(\delta, \mathcal{H}_{os}) \) goes to infinity as \( \delta \) 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 \( \tau_n \), defined as

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

\(^{16}\)CCLN imposed a stronger version of Assumption 4.1 in local identification of \( h_0 \) and provided various sufficient conditions.

\(^{17}\)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.
We note that $\tau_n$ is a direct extension of BCK’s sieve measure of ill-posedness,

$$\tau_n = \sup_{h \in \mathcal{H} : \|h - \hat{h}_n\| \neq 0} \frac{\sqrt{\mathbb{E}[\{h(Y_2) - \Pi_n h_0(Y_2)\]^2]}}{\sqrt{\mathbb{E}[\{h(Y_2) - \Pi_n h_0(Y_2)|X\]^2]}}$$

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 $\tau_n$ depend on the choice of the sieve space. Nevertheless, for any sieve space $\mathcal{H}_{osn}$ and for any $\delta > 0$, we have the following properties:

(i) $\omega_n(\delta, \mathcal{H}_{osn}) \leq \tau_n \times \delta$ and $\omega_n(\delta, \mathcal{H}_{osn}) \leq \omega(\delta, \mathcal{H}_{os})$.

(ii) $\omega_n(\delta, \mathcal{H}_{osn})$ and $\tau_n$ 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 $\limsup_{n \to \infty} \tau_n = \sup_{h \in \mathcal{H}_{os} : \|h - h_0\| \neq 0} \frac{\|h - h_0\|}{\|h - h_0\|}$.

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 ($\tau_n$) 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 $\sup_{h \in \mathcal{H}_{osn}} |\hat{P}(h) - P(h)| = O_P(\delta_{P,n})$. Let $\{\delta_{m,n}\}_{n=1}^{\infty}$ be a sequence of positive real values such that $\delta_{m,n} = o(1)$ and $\frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, h)\|_P^2 \geq \text{const}. \mathbb{E}[\|m(X, h)\|_P^2] - O_P(\delta_{m,n}^2)$ uniformly over $\mathcal{H}_{osn}$. By definition, $\delta_{m,n}^2 \leq \delta_{m,n}^2$. In fact, we have $\delta_{m,n}^2 = \eta_{0,n}$ for most commonly used nonparametric estimators $\hat{m}(\cdot)$. For example, Lemma C.2 in Appendix C shows that the series LS estimator $\hat{m}(X, h)$ defined in (11) satisfies $\delta_{m,n}^2 = \eta_{0,n} = \max\{\frac{T_s}{n}, b_{m,1,1}\}$, where $\frac{T_s}{n}$ is the order of the variance and $b_{m,1,1}$ 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 $\|\hat{h}_n - h_0\|_s = O(1)$. Let $\hat{h}_n \in \mathcal{H}_{os}$ and $\Pi_n h_0 \in \mathcal{H}_{osn}$, and let Assumptions 3.1, 3.2, 3.3 with $\eta_{0,n} = O(\delta_{m,n}^2)$, and 4.1 hold. Suppose that one of the following conditions holds:

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

(ii) $\max\{\delta_{m,n}^2, \lambda_n\} = \delta_{m,n}^2 = O(\lambda_n)$ and $P(\cdot)$ is lower semicompact.

(iii) $\max\{\delta_{m,n}^2, \lambda_n \delta_{P,n} \lambda_n \|\hat{h}_n - \Pi_n h_0\|_s \} = O_P(\delta_{m,n}^2)$ and there is a $t_0 \in \mathcal{H}^*$ with $\langle t_0, \cdot \rangle_{\mathcal{H}, H}$ a bounded linear functional with respect to $\| \cdot \|_s$ such that $\lambda_n \|P(h) - P(\Pi_n h_0) - (t_0, h - \Pi_n h_0)\|_{H^*, H} \geq 0$ for all $h \in \mathcal{H}_{osn}$. 


Then \( \| \hat{h}_n - h_0 \|_s = O_p(\| h_0 - \Pi_n h_0 \|_s + \omega_n(\max(\delta_{m,n}, \| \Pi_n h_0 - h_0 \|), \mathcal{H}_{os})) \).

Theorem 4.1 under condition (i) allows for slowly growing finite-dimensional sieves without a penalty (\( \lambda_n = 0 \)) or with any flexible penalty satisfying \( \lambda_n = o(\| \Pi_n h_0 - h_0 \|^2) \); such cases are loosely called the sieve dominating case. We note that condition (iii) 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 (ii) or (iii) allows for an infinite-dimensional sieve (\( k(n) = \infty \)) or large dimensional sieves (\( k(n)/n \to \text{const.} > 0 \)) 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 (i), (ii), or (iii) also allows for finite (but maybe large) dimensional sieves (\( k(n)/n \to \text{const.} \geq 0 \)) satisfying \( \| \Pi_n h_0 - h_0 \|^2 = O(\lambda_n) \); such cases are loosely called the sieve penalization balance case.

**REMARK 4.1:** (i) For PSMD estimators using finite-dimensional sieves (\( k(n) < \infty \)), the conclusion of Theorem 4.1 can be stated as

\[
\| \hat{h}_n - h_0 \|_s = O_p(\| h_0 - \Pi_n h_0 \|_s + \tau_n \times \max(\delta_{m,n}, \| \Pi_n h_0 - h_0 \|)).
\]

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 \( \| \cdot \|_s \) (in addition to the rate in the root mean squared metric).

(ii) 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, 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 \( \omega_n(\delta, \mathcal{H}_{os}) \) (the sieve measure of local ill-posedness \( \tau_n \)), 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 \( H \) 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 \( (H, \| \cdot \|_s) \); that is, any \( h \in H \) can be expressed as \( h = \sum_j \langle h, q_j \rangle q_j \), and there are two finite constants \( c_1, c_2 > 0 \) such that \( c_1\|h\|_s^2 \leq \sum_j |\langle h, q_j \rangle|^2 \leq c_2\|h\|_s^2 \) for all \( h \in H \). See the 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: \( \|h_0 - \sum_{j=1}^{k(n)} \langle h_0, q_j \rangle q_j \|_s = O(\nu_{k(n)}^{-\alpha}) \) for a finite \( \alpha > 0 \) and a positive sequence \( \{\nu_j\}_{j=1}^\infty \) that strictly increases to \( \infty \) 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 \( h \in \mathcal{H}_{osn} \) and (ii) \( \|\Pi_n h_0 - h_0\|^2 \leq C \sum_j \varphi(\nu_j^2) \times |\langle \Pi_n h_0 - h_0, q_j \rangle|_s^2 \).

Assumption 5.1 is a very mild condition about the smoothness of \( h_0 \in \mathcal{H}_{os} \); it suggests that \( \mathcal{H}_n = \text{clsp}\{q_1, \ldots, q_{k(n)}\} \) is a natural sieve to approximate \( h_0 \). For example, if \( (\mathbf{H}, \|\cdot\|_s) = (L^2([0, 1]^d, \text{leb}), \|\cdot\|_{L^2(\text{leb})}) \) and \( h_0 \in W_\alpha^s([0, 1]^d, \text{leb}) \), 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 pseudometric \( \|\cdot\| \) to the strong norm in a shrinking sieve 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 \( \text{const.} \times \delta/\sqrt{\varphi(\nu_{k(n)}^2)} \) and that the sieve measure of (local) ill-posedness \( \tau_n \leq \text{const.} \times \sqrt{\varphi(\nu_{k(n)}^2)} \) (see Lemma B.2). Assumption 5.2(ii) is the so-called stability condition that is required to hold only 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 presented conditions that imply Assumption 5.2(i) and (ii). See Section 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 \( \lambda_n \geq 0 \) and \( \lambda_n = o(1) \), and let all the assumptions of Theorem 4.1(i) hold. Let Assumptions 5.1 and 5.2 hold with \( \mathcal{H}_n = \text{clsp}\{q_1, \ldots, q_{k(n)}\} \) and \( k(n) < \infty \). Let \( \max\{\delta_{m,n}^2, \lambda_n\} = \delta_{m,n}^2 = \text{const.} \times k(n)/n = o(1) \). Then

\[
\|\hat{h}_n - h_0\|_s = O_p\left(\nu_{k(n)}^{-\alpha} + \frac{k(n)}{\sqrt{n \times \varphi(\nu_{k(n)}^2)}} \right) = O_p\left(\nu_{k_{o(n)}}^{-\alpha} \right),
\]

where \( k_{o(n)} \) is such that \( \nu_{k_{o(n)}}^{-2\alpha} \geq \frac{k_{o(n)}}{n} \{\varphi(\nu_{k_{o(n)}}^{-2})\}^{-1} \).

(i) Mildly ill-posed case. If \( \varphi(\tau) = \tau^s \) for some \( s \geq 0 \) and \( \nu_k \asymp k^{1/d} \), then \( \|\hat{h}_n - h_0\|_s = O_p(n^{-s/(2(a+s)+d)}) \) provided that \( k_{o(n)} \asymp n^{d/(2(a+s)+d)} \).

(ii) Severely ill-posed case. If \( \varphi(\tau) = \exp(-\tau^{-s/2}) \) for some \( s > 0 \) and \( \nu_k \asymp k^{1/d} \), then \( \|\hat{h}_n - h_0\|_s = O_p([\ln(n)]^{-a/s}) \) provided that \( k_{o(n)} = c[\ln(n)]^{d/s} \) 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 \((\nu_{k(n)}^{-a})\) and the model complexity (or roughly the standard deviation)
\[
\left( \frac{k(n)}{n} \right) \left( \varphi (\nu_{k(n)}^2) \right)^{-1},
\]
and we let \( \max(\delta_{m,n}^2, \lambda_n) = \delta_{m,n}^* = \text{const.} \times \frac{k(n)}{n} \). For example, if the PSMD estimator \( \hat{h}_n \) is computed using the series LS estimator \( \hat{m}(X, h) \) defined in (11), we can let \( \delta_{m,n}^2 = \eta_0,n = \max(\frac{b_n}{n}, b_{m,n}^1) = \frac{b_n}{n} = \text{const.} \times \frac{k(n)}{n} = o(1) \) (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 \( M > 0 \) and \( \alpha > 0 \), and a strictly increasing positive sequence \( \{\nu_j\}_{j=1}^\infty \) such that \( h - \sum_{j=1}^{k}(h, q_j)q_j, s \| \leq M(\nu_{k+1})^{-\alpha} \) 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 \( \varphi : \mathcal{R}_+ \rightarrow \mathcal{R}_+ \) such that: (i) \( \|h\|^2 \geq c \sum_{j=1}^{\infty} \varphi(\nu_j^{-2})(h, q_j)q_j \| \) for all \( h \in \mathcal{H}_{os} \) and (ii) \( \|h - h_0\|^2 \leq C \sum_{j=1}^{\infty} \varphi(\nu_j^{-2})(h - h_0, q_j)q_j \| \) for all \( h \in \mathcal{H}_{os} \).

Assumption 5.3 obviously implies Assumption 5.1. Assumption 5.3 is automatically satisfied if either \( \mathcal{H}_{os} \subseteq \mathcal{H}_{\text{ellipsoid}} \equiv \{h = \sum_{j=1}^{\infty}(h, q_j)q_j : \sum_{j=1}^{\infty} \nu_j^{2\alpha} \leq M^2\} \) or \( \mathcal{H}_{os} \subseteq \mathcal{H}_{\text{hyperrec}} \equiv \{h = \sum_{j=1}^{\infty}(h, q_j)q_j : \|h, q_j\| \leq M' \nu_j^{-(\alpha+1)/2}, \inf_j \nu_j/j > 0\} \). Both \( \mathcal{H}_{\text{ellipsoid}} \) and \( \mathcal{H}_{\text{hyperrec}} \) 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 \( P(h) \geq \sum_{j=1}^{\infty} \nu_j^{2\alpha} \|h, q_j\|\) for all \( h \in \mathcal{H}_{os} \). Assumption 5.4(i) and (ii) obviously implies Assumption 5.2(i) and (ii), respectively. Assumptions 5.3 and 5.4(i) together provide an upper bound on the modulus of continuity \( \omega_\alpha(\delta, \mathcal{H}_{os}) \) (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 Section 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 \( \delta_{m,n}^2 \) denote the optimal convergence rate of \( \hat{m}(\cdot, h) - m(\cdot, h) \) in the root mean squared metric uniformly over \( \mathcal{H}_{osn} \). By definition, \( \delta_{m,n}^2 \leq \delta_{m,n}^4 \).

**COROLLARY 5.2:** Let \( \hat{h}_n \) be the PSMD estimator with \( \lambda_n > 0 \) and \( \lambda_n = o(1) \), and let all the assumptions of Theorem 4.1(i) hold. Let Assumptions 5.2(ii),
5.3, and 5.4(i) hold with \( \mathcal{H}_n = \text{clsp}\{q_1, \ldots, q_{k(n)}\} \) for \( k(n)/n \to \text{const.} > 0 \) and 
\( \infty \geq k(n) \geq k^* \), where \( k^* = k^* (\delta_{m,n}) \) is such that \( \{\nu_k\}^{-2a} \asymp \delta_{m,n}^2 (\nu_{k^*}^{-2})^{-1} \). Let 
either condition (ii) of Theorem 4.1 hold with \( \lambda_n = O(\delta_{m,n}^2) \) or condition (iii) of 
Theorem 4.1 hold with \( \lambda_n = O(\delta_{m,n}^2 \sqrt{\nu_{k^*}^{-2}}) \). (i) Then

\[ \| \hat{h}_n - h_0 \|_s = O_p(\{\nu_k\}^{-a}) = O_p\left(\delta_{m,n}^s (\nu_{k^*}^{-2})^{-1/2}\right); \]

Thus \( \| \hat{h}_n - h_0 \|_s = O_p(\{\delta_{m,n}^s (\nu_{k^*}^{-2})^{-a}\}) \) if \( \varphi(\tau) = \tau^s \) for some \( s \geq 0 \) and \( \| \hat{h}_n - h_0 \|_s = O_p(\{\ln(\delta_{m,n})\}^{-a/s}) \) if \( \varphi(\tau) = \exp\{-\tau^{-s/2}\} \) for some \( s > 0 \).

(ii) If \( \mathcal{H}_n = \mathcal{H} \) (or \( k(n) = \infty \)), then Assumption 5.2(ii) holds and result (i) 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, \( \delta_{m,n}^2 = \frac{J_n^2}{n} \asymp b_{m,J_n^2}^2 \), where \( J_n^* \) is such that the variance part \( \{\frac{J_n^2}{n}\} \) and the squared bias part \( \{b_{m,J_n^2}^2\} \) 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 let \( \hat{m}(X,h) \) be the series LS estimator satisfying Assumptions C.1 and C.2. Let Assumption 5.4 and all the assumptions of Theorem 4.1(ii) hold with 
\( c_2 \mathbb{E}[\tilde{m}(X,h) - h_0\] \( \|_{\gamma_j}^2 \) for all \( h \in \mathcal{H}_{os} \). Let either \( P(h) \geq \sum_{j=1}^{\infty} \nu_j^2 |\langle h, q_j \rangle|^2 \) for all \( h \in \mathcal{H}_{os} \) or \( \mathcal{H}_{os} \subset \mathcal{H}_{ellipsoid} \). Let \( \lambda_n = O\left(\frac{J_n^2}{n}\right) \), where \( J_n^* \leq k(n) \leq \infty \) and is such that \( \frac{J_n^2}{n} \asymp b_{m,J_n^2}^2 \leq \text{const.}\nu_j^{-2} \varphi(\nu_j^{-2}) \). Then

\[ \| \hat{h}_n - h_0 \|_s = O_p\left(\nu_j^{-a}\right) = O_p\left(\frac{J_n^*}{n \times \varphi(\nu_j^{-2})}\right). \]

Thus, \( \| \hat{h}_n - h_0 \|_s = O_p\left(n^{-a/2((a+s)/d)}\right) \) if \( \varphi(\tau) = \tau^s \) for some \( s \geq 0 \) and \( \nu_k \asymp k^{1/d} \), and \( \| \hat{h}_n - h_0 \|_s = O_p\left(\nu_k^{-2}\right) \) if \( \varphi(\tau) = \exp\{-\tau^{-s/2}\} \) for some \( s > 0 \) and \( \nu_k \asymp k^{1/d} \), \( J_n^* = c[\ln(n)]^{d/s} \) 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, Assumption 5.4, is important for the convergence rate of the PSMD estimator. Denote \( T_{h_0} = \sqrt{W(\cdot)} \frac{d \mathcal{H}(h_0)}{dh} : \mathcal{H}_{os} \subset \mathbb{H} \to L^2(f_X) \), and denote \( T_{h_0}^* \) 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)} = \
\[ \| (T_{h_0}^* T_{h_0})^{1/2} h \|_2^2, \] Hence, Assumption 5.4 can be restated in terms of the operator \( T_{h_0}^* T_{h_0} \): There is a positive increasing function \( \varphi \) such that \[ \| (T_{h_0}^* T_{h_0})^{1/2} h \|_2^2 \asymp \sum_{j=1}^{\infty} \varphi(v_j^{-2}) \| (h, q_j) \|_s^2 \] for all \( h \in H_{o,s} \). This assumption relates the smoothness of the operator \( (T_{h_0}^* T_{h_0})^{1/2} \) to the smoothness of the unknown function \( h_0 \in H_{o,s} \). Assumption 5.4(i) and (ii) are, respectively, the reverse link condition and the link condition imposed in CR’s study of the NPIV model (2). Nair, Pereverzev, and Tautenhahn (2005) also assumed this in their study of a linear ill-posed inverse problem with deterministic noise and a known operator. In the following discussion, we mention some sufficient conditions for Assumption 5.4.

A (nonlinear) operator \( A : H \to L^2(f_X) \) is compact if it is continuous and maps bounded sets in \( H \) 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 \( \sqrt{W(\cdot)} m(\cdot, h) : H \to L^2(f_X) \) is compact and is Frechet differentiable at \( h_0 \in H_{o,s} \); see Zeidler (1985, Proposition 7.33)). Then \( T_{h_0} \) has a singular value decomposition \( \{ \mu_k; \phi_k, \phi_{0k} \}_{k=1}^{\infty} \), where \( \{ \mu_k \}_{k=1}^{\infty} \) are the singular numbers arranged in nonincreasing order (\( \mu_k \geq \mu_{k+1} \rightarrow 0 \)), and \( \{ \phi_{0k}(\cdot) \}_{k=1}^{\infty} \) 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) and (ii)) and \( c_2E[\| m(X, h) \|_H^2] \leq \| h - h_0 \|_2^2 \) for all \( h \in H_{o,s} \). Then the eigenfunction sequence \( \{ \phi_{1k}(\cdot) \}_{k=1}^{\infty} \) is an orthonormal basis (hence a Riesz basis) for \( H_{o,s} \), and \( \| (T_{h_0}^* T_{h_0})^{1/2} h \|_2^2 = \sum_{k=1}^{\infty} \mu_k^2 \| (h, \phi_{1k}) \|_s^2 \) for all \( h \in H_{o,s} \). Thus, Assumption 5.4 is automatically satisfied with \( q_j = \phi_{1j} \) and \( \varphi(v_j^{-2}) = \mu_j^2 \) for all \( j \). Following the proof of Lemma 1 in BCK, we can show that the sieve measure of local ill-posedness \( \tau_n = [\mu_{k(n)}]^{-1} = [\varphi(v_{k(n)}^{-2})]^{-1/2} \) and that the sieve modulus of continuity \( \omega_n(\delta, H_{o,s}) = \delta \tau_n = \delta [\mu_{k(n)}]^{-1} \).

In the numerical analysis literature on ill-posed inverse problems with known operators, it is common to measure the smoothness of \( h_0 \in H_{o,s} \) 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 \( \psi \) with \( \psi(0) = 0 \) such that \( h_0 \in H_{source} \equiv \{ h = \psi(T_{h_0}^* T_{h_0}) \nu : \nu \in H, \| \nu \|_2^2 \leq M^2 \} \) for a finite constant \( M \), and the original source condition corresponds to the choice \( \psi(\eta) = \eta^{1/2} \) (see Engl, Hanke, and Neubauer 1996). 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

\[
(16) \quad h_0 \in H_{source} = \left\{ h = \sum_{j=1}^{\infty} \langle h, \phi_{1j} \rangle_s \phi_{1j} : \sum_{j=1}^{\infty} \langle h, \phi_{1j} \rangle_s \psi^2(\mu_j^{-1}) \leq M^2 \right\},
\]

\[18\text{See Bissantz, Hohage, Munk, and Ruymgaart (2007) for convergence rates of statistical linear ill-posed inverse problems via the Hilbert scale (or general source condition) approach for possibly noncompact but known operators.}
which is a particular Sobolev ellipsoid class of functions $\mathcal{H}_{\ell}$-ellipsoid. Therefore, the general source condition implies our Assumptions 5.3 and 5.4 by setting $q_j = \phi_{ij}$, $\varphi(v_j^{-2}) = \mu_j^2$, and $\psi(\mu_j^2) = v_j^{-a}$ for all $j \geq 1$. Then $\varphi(\tau) = \tau^s$ (mildly ill-posed case) is equivalent to $\psi(\eta) = \eta^{a/(2s)}$; $\varphi(\tau) = \exp\{-\tau^{-s/2}\}$ (severely ill-posed case) is equivalent to $\psi(\eta) = [-\log(\eta)]^{-a/s}$.

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 $\lambda_n \geq 0$ and $\lambda_n = o(1)$, and let all the assumptions of Theorem 4.1(i) hold with $c_2E[\|m(X, h)\|_W^2] \leq \|h - h_0\|^2$ for all $h \in \mathcal{H}_{os}$. Let $T_{h_0} \equiv \sqrt{\mathcal{W}(\cdot)}(\mathcal{m}_{h_0}) : \mathcal{H}_{os} \subset \mathbf{H} \to L^2(f_X)$ be a compact operator with a singular value decomposition $\{\mu_j; \phi_{ij}, \phi_{0j}\}_{j=1}^\infty$. Let $\mathcal{H}_n = \text{clsp}\{\phi_{ij} : j = 1, \ldots, k(n)\}$ for $k(n) \leq \infty$.

(i) **Sieve Dominating Case.** Let $h_0 \in \mathcal{H}_{source}$. If $\max\{\delta_{m,n}^2, \lambda_n\} = \delta_{m,n}^2 = \text{const.} \times \frac{k(n)}{n} = o(1)$, then

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

(ii) **Penalty Dominating Case.** Let $\hat{m}(X, h)$ be the series LS estimator satisfying Assumptions C.1 and C.2. Let either $P(h) \geq \sum_{j=1}^\infty \|\langle \phi_{ij}, \phi_{0j}\rangle\|^2$ for all $h \in \mathcal{H}_{os}$ or $\mathcal{H}_{os} \subset \mathcal{H}_{source}$. Let $0 < \lambda_n = O(\frac{J_n^2}{n}) = o(1)$, where $J_n \leq k(n) \leq \infty$ and is such that $\frac{J_n}{n} \times b_{m,n}^2 \leq \text{const.} \{\psi(\mu_{J_n}^2)\}^{-2} \mu_{J_n}^2$. Then

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

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

**Remark 5.2:** (i) Suppose that $q_j = \phi_{ij}$ (Assumption 5.4 holds), $\varphi(v_j^{-2}) = \mu_j^2 \geq \text{const.} j^{-s}$, $s > 1$ (mildly ill-posed case), and $\mathcal{H}_{os} = \{h = \sum_{j=1}^\infty \langle h, \phi_{ij}\rangle, \phi_{ij} : \|\langle h, \phi_{ij}\rangle\| \leq M_j^{-1/2}\} \backslash \mathcal{H}_{os}$, $\alpha > 0$ (Assumption 5.3 holds). HH established 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 \( 2\alpha + 1 > s \geq \alpha \)). HL extended their result to the NPQIV model (3) (for \( 2\alpha > s \geq \alpha > \frac{1}{2} \)).

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

\[
\inf_{\hat{h}} \sup_{h \in \mathcal{H}_{\alpha s}} E^h [\| \hat{h} - h \|_{S^2}^2] \geq \text{const. } n^{-1} \sum_{j=1}^{k_o} [\varphi(\nu_j^{-2})]^{-1} \{\nu_{k_o}\}^{-2\alpha},
\]

where \( k_o = k_o(n) \) is the largest integer such that \( \frac{1}{n} \sum_{j=1}^{k_o} (\nu_j)^{2\alpha}[\varphi(\nu_j^{-2})]^{-1} \approx 1 \). In addition, suppose that Assumption 5.4(i) holds; CR showed that the BCK estimator \( \hat{h}_n \), 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

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

where \( h_{01} \) and \( h_{02} \) are the unknown functions of interest, and 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 \( Y = [0, 1]^d \times [0, 1]^d \times \mathcal{R} \) and the support of \( X \) be \( X = [0, 1]^{d_1} \) with \( d_1 \geq d \geq 1 \). We estimate \( h_0 = (h_{01}, h_{02}) \in \mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2 \) using the PSMD estimator \( \hat{h}_n \) given in (6), with \( \hat{W} = W = I \) (identity), \( \mathcal{H}_n = \mathcal{H}_1 \times \mathcal{H}_2 \) being either a finite-dimensional \( (\dim(\mathcal{H}_n) = k(n) = k_1(n) + k_2(n) < \infty) \) or an infinite-dimensional \( (k(n) = \infty) \) linear sieve, and \( \hat{P}_n(h) = P(h) \geq 0 \). The conditional mean function \( m(X, h) \) is estimated by the series LS estimator \( \hat{m}(X, h) \) defined in (11). To simplify presentation, we let \( p^{h}(X) \) be a tensor-product linear sieve basis, which is the product of univariate linear sieves. For example, let \( \{\phi_i : i = 1, \ldots, J_{j,n}\} \) denote a P-spline (polynomial spline), B-spline, wavelet, or Fourier series basis for \( L^2(\mathcal{X}_j, \text{leb.}) \), with \( \mathcal{X}_j \) a compact interval in \( \mathcal{R}, 1 \leq j \leq d_1 \). Then the tensor product \( \prod_{j=1}^{d_1} \phi_i(X_j) : i = 1, \ldots, J_{j,n}, j = 1, \ldots, d_1 \) is a P-spline, B-spline, wavelet, Fourier series, or power series basis for \( L^2(\mathcal{X}, \text{leb.}) \).
with $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_{d_x}$. Clearly the number of terms in the tensor-product sieve $p^n(X)$ 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 the following condition.

**CONDITION 6.1:** (i) $\{(Y_l', X_l')\}_{l=1}^m$ 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 $E[p^n(X)p^n(X')]$ is bounded away from zero uniformly in $J_n$, where $p^n(X)$ is a tensor product P-spline, B-spline, wavelet, or cosine sieve with $J_n^2 = o(n)$. (iii) $E[f_{Y_l|Y_1,Y_2,X}(h_1(Y_l) + h_2(Y_l))|X = \cdot ] \in A^a_m([0,1]^{d_x})$ with $\alpha_m > 0$ for all $h \in H^{d_0}_{k(n)}$. (iv) $f_{Y_l|Y_1,Y_2,X}(y_1,y_2) = f_{Y_l|Y_1,Y_2}(y_1,y_2)$ is continuous in $(Y_l, y_1, Y_2, x)$ and $\sup_{\cdot} f_{Y_l|Y_1,Y_2}(y_1,y_2) \leq \text{const.} < \infty$ for almost all $Y_l$, $Y_2$, $X$.

Condition 6.1(i)–(iii) implies that the series LS estimator $\hat{m}(\cdot, h)$ satisfies Assumption 3.3 with $\eta_{\theta,n} = \delta_{m,n}^2 = \max\{J_n^{-1}, J_n^{-2\alpha_m/d_x}\}$ and $\delta_{m,n}^2 = o(1)$ (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 text, we denote $h_0(y_1,y_2) = h_0(y_1) + h_0(y_2)$, $\Delta h (y_1, y_2) = h(y_1, y_2) - h_0(y_1, y_2) = \Delta h_1 (y_1) + \Delta h_2 (y_2)$, and, for $l = 1, 2$,

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

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

Condition 6.2(i) and (iii), respectively, specify the function space and the sieve space for $h = (h_1, h_2)$. 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 $\lambda_n > 0$ and $\eta_n = O(\lambda_n) = o(1)$, and let $\hat{m}(X, h)$ be the series LS estimator. Let
Conditions 6.1 and 6.2 hold, \( P(h) = \|h_1\|_{L^{d_1}} + \|h_2\|_{L^{d_2}} \) and \( \max\{k_1(n)\}^{-2r_1}, [k_2(n)]^{-2r_2}, \frac{J_n}{n} + J^{-2r_m} \} = O(\lambda_n). \) Then
\[
\sup_{y_1 \in [0,1]^d} |\hat{h}_{1,n}(y_1) - h_{01}(y_1)| + \sup_{y_2 \in [0,1]^d} |\hat{h}_{2,n}(y_2) - h_{02}(y_2)| = o_p(1);
\]
hence, \( \|\hat{h}_{1,n} - h_{01}\|_{L^2(f_{Y_1})} + \|\hat{h}_{2,n} - h_{02}\|_{L^2(f_{Y_2})} = o_p(1) \) and \( P(\hat{h}_{1,n}) + P(\hat{h}_{2,n}) = O_p(1). \)

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 \( \|h\|_s^2 \leq \|h\|_{\text{sup}}^2 \) for all \( h \in H \). The above consistency results immediately imply that \( \|\hat{h}_n - h_0\|_s = o_P(1) \). Let \( H_{os} = \{h = (h_1, h_2) \in H : \|h-h_0\|_{\text{sup}} = o(1), P(h) \leq c\}. \) For \( h = (h_1, h_2) \in H_{os} \) and \( l = 1, 2 \), denote
\[
T_{l,0}[h_l - h_{0l}](X) = E[f_{Y_l}Y_l,h_0(Y_l)]h_l(X)
\]
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) \( \|T_{h_0}[h - h_0]\|_{L^2(f_X)} \leq \|K_{1,h}[h_1 - h_0] + K_{2,h}[h_2 - h_{02}]\|_{L^2(f_X)} \) for all \( h = (h_1, h_2) \in H_{os} \cap H_n, \) Range\(T_{1,0}) \cap \text{Range}(T_{2,0}) = \{0\}, \) and \( T_{1,0}[\Delta h_l](X) = 0 \) a.s. \( X \) implies \( \Delta h_l = 0 \) a.s. \( Y_l \) for \( l = 1, 2 \). (ii) There is a continuous increasing function \( \varphi \geq 0 \) such that
\[
\|T_{h}(h - h_0)\|_{L^2(f_X)}^2 \leq \sum_{j=1}^{\infty} \varphi(j^{-2/d})(j_1^2h_{1,j}^2 + j_2^2h_{2,j}^2)
\]
for all \( h = (h_1, h_2) \in H_{os} \cap H_n. \)

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 \( \max\{J_n, J^{-2r_m} \} = \frac{J_n}{n} = \text{const.} \times \frac{k(n)}{n} = o(1), k(n) = k_1(n) + k_2(n), \) and \( k_1(n) \asymp k_2(n) \to \infty, \) then
\[
\|\hat{h}_n - h_0\|_s = o_P\left((k(n))^{-a/d} + n^{-1}n \times \varphi(k(n))^{-2/d}\right).
\]
Thus, \( \| \hat{h}_n - h_0 \|_s = O_p(n^{-\alpha/(2(\alpha + \varsigma) + d)}) \) if \( \varphi(\tau) = \tau^\varsigma \) for some \( \varsigma \geq 0 \) and \( k(n) \asymp n^{d/(2(\alpha + \varsigma) + d)} \), and \( \| \hat{h}_n - h_0 \|_s = O_p((\ln(n))^{-\alpha/\varsigma}) \) 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 \( \varphi(\{k(n)\}^{-2/d}) = \text{const.} \) in Proposition 6.2. The resulting convergence rate \( \| \hat{h}_n - h_0 \|_s \) coincides with the known optimal rate for the additive quantile regression model: \( Y_3 = h_0(Y_2) + U, \Pr(U \leq 0|X) = \gamma; \) see, for example, 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.

**Remark 6.1:** Chen and Pouzo (2009c) established the asymptotic normality of a plug-in PSMD estimator \( \phi(\hat{h}_n) \) of a functional \( \phi(h_0) \) regardless of whether it could be estimated at a root-\( n \) rate. As an illustrative application to the NPQIV model \( Y_3 = h_0(Y_2) + U, \Pr(U \leq 0|X) = \gamma \), Chen and Pouzo (2009c) presented the pointwise asymptotic normality of \( \hat{h}(\cdot) \) which is a PSMD estimator using a slowly growing finite-dimensional linear sieve \( \mathcal{H}_{k(n)} = \{ h(Y_2) = q^{k(n)}(y_2)b = \sum_{j=1}^{k(n)} b_j q_j(y_2) \} \). For a fixed \( y_2 \) in the support of \( Y_2 \), by applying the rate result of this paper to control the order of the bias, Chen and Pouzo (2009c) showed that

\[
\sqrt{n} \frac{\hat{h}(y_2) - h_0(y_2)}{\sqrt{\gamma(1-\gamma)\sigma_n^2}} \xrightarrow{d} \mathcal{N}(0, 1),
\]

where

\[
\sigma_n^2 = q^{k(n)}(y_2)' \left[ E\left( \frac{E[f_{U|Y_2,X}(0)]q^{k(n)}(Y_2)|X]}{E[f_{U|Y_2,X}(0)]q^{k(n)}(Y_2)|X]} \right]^{-1} q^{k(n)}(y_2).
\]

We could obtain the pointwise asymptotic normality for the nonparametric additive quantile IV regression in a similar way.

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 \((Y_2, \tilde{X})\) according to a bivariate Gaussian density whose mean and covariance are set to those estimated from the U.K. Family Expenditure Survey Engel curve data set (see BCK for details). Let \(X = \Phi(\frac{X - \mu_2}{\sigma_2})\) and \(h_0(y) = \Phi(\frac{y - \mu_2}{\sigma_2})\), where \(\Phi\) denotes the standard normal c.d.f., and the means \(\mu_1\), \(\mu_2\) and variances \(\sigma_1\), \(\sigma_2\) are the estimated ones. Second, we generate \(Y_1\) from \(Y_1 = h_0(Y_2) + U\), where \(U = \sqrt{0.075[V - \Phi^{-1}(\gamma + 0.01[E[h_0(Y_2)]\tilde{X}] - h_0(Y_2))]\) with \(V \sim N(0, 1)\). The number of observations is set to \(n = 500\). We have also tried to draw \((Y_2, \tilde{X})\) 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 those reported here.

In this MC study and for the sake of concreteness, we estimate \(h_0(\cdot)\) using the PSMD estimator \(\hat{h}_n\) given in (6), with \(\hat{m}(X, \hat{h})\) being the series LS estimator (11) of \(m(X, h)\), \(\hat{W} = W = I\) (identity), and \(\mathcal{H}_n\) being a finite-dimensional \((\dim(\mathcal{H}_n) = k(n) < \infty)\) 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 \((\lambda_n)\) for the penalization, and one \((J_n)\) for the nonparametric LS estimator of \(\hat{m}(X, h)\). In the previous theoretical sections, we showed that we could obtain the optimal rate in either the “sieve dominating case” (the case of choosing \(k(n) \gg J_n\), \(k(n) \ll J_n\) properly and letting \(\lambda_n = 0\) or \(\lambda_n \propto n^{\frac{2}{q + r}}\) fast) or the sieve penalization balance case (the case of choosing \(k(n) \ll J_n, k(n) \leq J_n\) and \(\lambda_n \propto \frac{J_n}{n}\) properly). In this MC study, we compare the finite sample performance of these two cases. 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 \(k(n) \gg J_n\) and \(k(n)/n \to \text{const.} > 0\)). It was too time-consuming to compute the MC results for this case, and the results were not very stable.

\[J_{16} = \|\nabla h\|^2_{L^2(\text{leb})}. \]
FIGURE 1.—Summary of Monte Carlo results for \( \gamma = \{0.25, 0.5, 0.75\} \). \( h_0 \) (solid thick), \( \hat{h}_n \) (solid thin), MC confidence bands (dashed), a sample of \( Y_1 \) (dots), \( \hat{P}(h) = \|\nabla h\|_{L_2}^2 \), top row: \( k(n) = 8, \lambda_n = 0.003, J_n = 16 \); bottom row: \( k(n) = 16, \lambda_n = 0.006, J_n = 16 \).

similar; thus, 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 percent 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 an 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 and (approximately) 240 minutes for the sieve penalization balance case.

Table I 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 percent to 97.5 percent.\(^12\) Here, for simplicity, we have reported only the estimated quantile with \( \gamma = 0.5 \), but with 750 MC replications. Figure 2 shows the corresponding estimated curves and MC 95 percent confidence bands. In Table I, 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 \((k(n) = 6, 8)\) performs well in terms of I-BIAS\(^2\) and I-VAR (hence

\(^{12}\)The simulations in the Table I were computed using a Intel\textsuperscript{®} processor with 2.8 GHz and 12.0 GB of RAM. We refer the reader to the working paper version for more MC simulations; those simulations yield results very similar to those in this paper.
I-MSE), and is much more economical in terms of computational time. Within the sieve dominating case \( k(n) = 8 \), given the same \( \lambda_n \), the ones with derivative penalty perform better than the one with function-level penalty.
7.2. Empirical Illustration

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

\[ E \left[ 1 \{ Y_{1il} \leq h_{0l}(Y_{2i}) \} | X_i \right] = \gamma \in (0, 1), \quad \ell = 1, \ldots, 7, \]

where \( Y_{1il} \) is the budget share of household \( i \) on good \( \ell \) (in this application, \( 1 = \text{food—out}, 2 = \text{food—in}, 3 = \text{alcohol}, 4 = \text{fares}, 5 = \text{fuel}, 6 = \text{leisure goods}, \) and \( 7 = \text{travel} \)). \( Y_{2i} \) 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. BCK studied the same data set 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 \( H_n \) for \( h \), with different types of penalty functions and \( \hat{p}_n(h) = \| \nabla^k h \|_{L^2(d\mu)} \) for \( k = 1, 2 \) and \( j = 1, 2, \) and Hermite polynomial sieves, cosine sieves, and polynomial splines sieves for the series LS estimator \( \hat{m} \). All combinations yielded very similar results; thus, we present figures only for one sieve dominating case, using P-spline \((2, 5)\) as \( H_n \) 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 the three different quantiles \( \gamma = \{0.25, 0.50, 0.75 \} \) and four goods that were considered in BCK. Figure 3 presents the estimated Engel curves using \( \hat{P}_n(h) = \| \nabla^2 h \|_{L^2(d\mu)}^2 \) with \( \lambda_n = 0.001 \) and \( \hat{P}_n(h) = \| \nabla^2 h \|_{L^2(d\mu)} \) with \( \lambda_n = 0.001 \) in the first and second rows, \( \hat{P}_n(h) = \| \nabla h \|_{L^2(d\mu)}^2 \) with \( \lambda_n = 0.001 \) (third row), \( \lambda_n = 0.003 \) (fourth row), and \( \hat{P}_n(h) = \| \nabla h \|_{L^2(d\mu)}^2 \) 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 \( \lambda_n \) and \( \hat{P}_n(h) \), which is again consistent with the theoretical results for the PSMD estimator in the sieve dominating case.

8. Conclusion

In this paper, we propose the PSMD estimation of conditional moment restrictions containing unknown functions of endogenous variables: \( E[\rho(Y, X; h_{0}(\cdot)) | X] = 0 \). 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 noncompact infinite-dimensional function parameter space; (ii) possibly

\(^{21}\)The results on all seven goods are available on request from the authors.
noncompact 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 coincides 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 U.K. Family Expenditure Survey. These results indicate that PSMD estimators using slowly growing finite-dimensional sieves with a small penalization parameter are easy to compute and perform well in finite samples.
The consistency and the rate results obtained in this paper are crucial for inference on the semi-nonparametric conditional moment restrictions (1): $E[p(Z; \theta_0, h_0(\cdot))|X] = 0$. For the general model (1) when $p(Z, \theta, h(\cdot))$ may not be pointwise smooth in $(\theta, h)$, Chen and Pouzo (2009a) showed that the PSMD estimator using slowly growing finite-dimensional sieves can simultaneously achieve the root-$n$ asymptotic normality of $\hat{\theta}_n - \theta_0$ and the nonparametric optimal rate of convergence for $\hat{h}_n - h_0$. Unfortunately, due to the nonparametric endogeneity, for the general model (1) it is very difficult to verify whether a real-valued functional $\phi(\theta_0, h_0)$ is root-$n$ estimable or not. Using the rate result of this paper, Chen and Pouzo (2009c) established the asymptotic normality of the plug-in PSMD estimator $\phi(\hat{\theta}_n, \hat{h}_n)$ and provided simple PSMD criterion-based inferences on $\phi(\theta_0, h_0)$ regardless of whether it could be estimated at a root-$n$ rate.

APPENDIX A: ADDITIONAL RESULTS FOR CONSISTENCY

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

In the following text, we let $(A, T)$ be a Hausdorff topological space and let $B_T(a)$ be a nonempty open neighborhood (under $T$) around $a \in A \subseteq A$. Let $Pr^*$ denote the outer measure associated with $Pr$. Let $o_{Pr^*}$ and $O_{Pr^*}$, respectively, denote convergence in probability under $Pr^*$ and bounded in probability under $Pr^*$.

**LEMMA A.1:** Let $\hat{\alpha}_n$ be such that $\hat{Q}_n(\hat{\alpha}_n) \leq \inf_{\alpha \in A_{k(n)}} \hat{Q}_n(\alpha) + O_{Pr^*}(\eta_n)$, where $(\eta_n)_{n=1}^\infty$ is a positive real-valued sequence such that $\eta_n = o(1)$. Let $\overline{Q}_n(\cdot): A \to [0, \infty)$ be a sequence of nonrandom measurable functions and let the following conditions hold:

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

$$\inf_{\alpha \in A_k: \alpha \notin B_T(a_0)} \overline{Q}_n(\alpha) \geq g_0(n, k, B) > 0 \text{ for each } n \geq 1, k \geq 1$$

and $\lim \inf_{n \to \infty} g_0(n, k(n), B) \geq 0$ for all $B_T(\alpha_0)$.

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

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

d. (i) $\hat{Q}_n(\Pi_n(\alpha_0)) \leq K_0 \overline{Q}_n(\Pi_n(\alpha_0)) + O_{Pr^*}(c_{0,n})$ for some $c_{0,n} = o(1)$ and a finite constant $K_0 > 0$; (ii) $\hat{Q}_n(\alpha) \geq K \overline{Q}_n(\alpha) - O_{Pr^*}(c_n)$ uniformly over $\alpha \in A_{k(n)}$ for some $c_n = o(1)$ and a finite constant $K > 0$; (iii) $\max(c_{0,n}, c_n, \overline{Q}_n(\Pi_n(\alpha_0), \eta_n) = o(g_0(n, k(n), B))$ for all $B_T(\alpha_0)$.

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

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

**Lemma A.2:** Let \( \hat{\alpha}_n \) be such that \( \hat{Q}_n(\alpha) \leq \inf_{\alpha \in A_k(n)} Q_n(\alpha) + O_p(\eta_n) \) with \( \eta_n = o(1) \). Let \( \tilde{Q}_n(\cdot) : A \to (-\infty, \infty) \) be a sequence of nonrandom measurable functions and let the following conditions hold:

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

\[
\inf_{\alpha \in A_k : |\alpha - \alpha_0| > \varepsilon} \tilde{Q}_n(\alpha) - \tilde{Q}_n(\alpha_0) \geq g_0(n, k, \varepsilon) > 0
\]

for each \( n \geq 1, k \geq 1, \varepsilon > 0 \)

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

b. (i) \( A \subseteq A \) and \( (A, \| \cdot \|) \) is a metric space; (ii) \( A_k \subseteq A_{k+1} \subseteq A \) for all \( k \geq 1 \) and there exists a sequence \( \Pi_n \alpha_0 \in A_k(n) \) such that \( \tilde{Q}_n(\Pi_n \alpha_0) - \tilde{Q}_n(\alpha_0) = o(1) \).

c. (i) \( \hat{Q}_n(\alpha) \) is a measurable function of the data \( \{(Y_i, X_i)\}_{i=1}^n \) for all \( \alpha \in A_k(n) \); (ii) \( \hat{\alpha}_n \) is well defined and measurable.

d. Let \( \hat{\alpha}_n = \sup_{\alpha \in A_k(n)} |\hat{Q}_n(\alpha) - \tilde{Q}_n(\alpha)| = o_p(1) \) and

\[
\max\{\hat{\alpha}_n, \eta_n / g_0(n, k(n), \varepsilon)\} |\hat{Q}_n(\Pi_n \alpha_0) - \tilde{Q}_n(\alpha_0)| = o_p(1)
\]

for all \( \varepsilon > 0 \).

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

**Remark A.1:** (i) Let \( (A, T) \) be a Hausdorff topological space and let \( A_k \) be nonempty for each \( k \). Lemma A.1c is satisfied and \( \hat{\alpha}_n \) is measurable if one of the following two conditions holds: (a) for each \( k \geq 1, A_k \) is a compact subset of \( (A, T) \), and for any data \( \{Z_i\}_{i=1}^n \), \( \hat{Q}_n(\alpha) \) is lower semicontinuous (in the topology \( T \)) on \( A_k \); (b) for any data \( \{Z_i\}_{i=1}^n \), the level set \( \{\alpha \in A_k : \hat{Q}_n(\alpha) \leq r\} \) is compact in \( (A, T) \) for all \( r \in (-\infty, +\infty) \). See Zeidler (1985, Theorem 38.B).

(ii) Let \( (A, \| \cdot \|) \) be a Banach space and let \( A_k \) be nonempty for each \( k \). Lemma A.1c is satisfied and \( \hat{\alpha}_n \) is measurable if one of the following three
conditions holds: (a) \( \mathcal{A}_k \) is a weak sequentially compact subset of \( (A, \| \cdot \|_s) \), and for any data \( \{Z_i\}_{i=1}^\infty \), \( \hat{Q}_n(\alpha) \) is weak sequentially lower semicontinuous on \( \mathcal{A}_k(n) \); and (b) \( \mathcal{A}_k \) is a bounded, and weak sequentially closed subset of a reflexive Banach space \( (A, \| \cdot \|_s) \), and for any data \( \{Z_i\}_{i=1}^\infty \), \( \hat{Q}_n(\alpha) \) 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 \( (A, \| \cdot \|_s) \), and for any data \( \{Z_i\}_{i=1}^\infty \), \( \hat{Q}_n(\alpha) \) 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).

Given Remark A.1, in the rest of the paper we assume that \( \tilde{\alpha}_n \) and our approximate PSMD estimator \( \hat{\alpha}_n \) defined in (6) are measurable.

**Lemma A.3:** Let \( \hat{\alpha}_n \) be the (approximate) PSMD estimator (6). Then \( \hat{\alpha}_n \in \mathcal{H}_n \) w.p.a.1.

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

(i) If Assumption 3.2(b) and \( \max \{ \eta_{0,n}, E[\|m(X, \Pi_n h_0)\|_W^2] \} = O(\lambda_n) \) hold, then \( P(\hat{\alpha}_n) = O_p(1) \).

(ii) If Assumption 3.2(c) and \( \max \{ \eta_{0,n}, E[\|m(X, \Pi_n h_0)\|_W^2] \} = o(\lambda_n) \) hold, then \( P(\hat{\alpha}_n) \leq P(h_0) + o_p(1) = O_p(1) \).

Recall that \( \mathcal{H}^{M_0}_{k(n)} = \{ h \in \mathcal{H}_{k(n)} : \lambda_n P(h) \leq \lambda_n M_0 \} \) for a large but finite \( M_0 \equiv M_0(\epsilon) \in (0, \infty) \), such that \( \Pi_n h_0 \in \mathcal{H}^{M_0}_{k(n)} \) and that for all \( \epsilon > 0 \), \( \Pr(\hat{\alpha}_n \notin \mathcal{H}^{M_0}_{k(n)}) < \epsilon \) for all sufficiently large \( n \), where the bound \( M_0 \equiv M_0(\epsilon) \) in \( \mathcal{H}^{M_0}_{k(n)} \) can depend on \( \epsilon > 0 \) but not on \( n \). Given Assumptions 3.2 and 3.3(i) and Lemma A.3, such an \( M_0 \) always exists. In the following discussion, we denote \( B_T(h_0) \) as any open neighborhood in a topological space \( (\mathcal{H}, T) \) around \( h_0 \).

**Lemma A.4:** Let \( \hat{\alpha}_n \) be the (approximate) PSMD estimator with \( \lambda_n \geq 0 \) and \( \eta_n = O(\eta_{0,n}) \) and let Assumption 3.3 hold. Let Assumption 3.1(iii) hold and allow that the \( T \) topology could be the norm \( \| \cdot \|_s \) topology or weaker ones. Then, for all \( B_T(h_0) \) and all \( \epsilon > 0 \), we have two results:

(i) Under Assumption 3.2(b) and \( \max \{ \eta_{0,n}, E[\|m(X, \Pi_n h_0)\|_W^2] \} = O(\lambda_n) \),

\[
\Pr(\hat{\alpha}_n \notin B_T(h_0)) \leq \Pr\left( \inf_{h \in \mathcal{H}^{M_0}_{k(n)}, h \notin B_T(h_0)} \{ cE[\|m(X, h)\|_W^2] + \lambda_n P(h) \} \right) \\
\leq O_p(\delta_{m,n}) + \lambda_n P(h_0) + O_p(\lambda_n) + \epsilon
\]

for all \( n \) sufficiently large, where the bound \( M_0 \equiv M_0(\epsilon) \) in \( \mathcal{H}^{M_0}_{k(n)} \) can depend on \( \epsilon > 0 \).
Under Assumption 3.2(c) and \(\max\{\eta_{0,n}, E[\|m(X, \Pi_n h_0)\|_W^2]\} = o(\lambda_n)\),
\[
\Pr(\hat{h}_n \notin B_T(h_0)) \leq \Pr\left(\inf_{h \in M_0 \cap \{h \notin B_T(h_0)\}} \left\{ cE\left[\|m(X, h)\|_W^2\right] + \lambda_n P(h)\right\}\right) \leq O_p(\delta_{m,n}^2) + \lambda_n P(h_0) + o_p(\lambda_n) + \epsilon
\]
for all \(n\) sufficiently large, where the bound \(M_0 \equiv M_0(\epsilon)\) in \(H_{k(n)}^M\) can depend on \(\epsilon > 0\).

**Identification via a Strictly Convex Penalty**

When \(E[\|m(X, h)\|_W^2]\) is convex in \(h \in H\), we can relax the identification Assumption 3.1(ii) by using a strictly convex penalty. Let \(M_0^p \equiv \{h \in H : E[\|m(X, h)\|_W^2] = 0\}\). Let \(M_0^p \equiv \{h \in H : h = \arg\inf_{h \in M_0} P(h')\}\) be the set of minimum penalization solutions.

**THEOREM A.1:** Suppose that \(M_0^p\) is nonempty, \(P\) is strictly convex and lower semicontinuous on \((M_0^p, \|\cdot\|_s)\), and \(E[\|m(X, h)\|_W^2]\) is convex and lower semicontinuous on \((H, \|\cdot\|_s)\).

(i) If Assumption 3.4(ii) holds, then \(M_0^p = \{h_0\} \subseteq M_0^p\).

(ii) Let \(\hat{h}_n\) be the PSMD estimator with \(\lambda_n > 0\) and \(\eta_n = O(\eta_{0,n})\) and let Assumptions 3.1(i), (iii), and (iv), 3.2(c), 3.3 and 3.4 hold. Suppose that for any \(k \geq 1, H_k\) is convex, and \(P(\cdot)\) is convex and lower semicontinuous on \((H_k, \|\cdot\|_s)\). If \(P(h_0|\eta_{0,n}, E[\|m(X, \Pi_n h_0)\|_W^2], \delta_{m,n}^2) = o(\lambda_n)\), then \(\|\hat{h}_n - h_0\|_s = o_p(1)\) and \(P(\hat{h}_n) = P(h_0) + o_p(1)\).

**APPENDIX B: LEMMAS FOR CONVERGENCE RATE**

**LEMMA B.1:** Suppose that all the conditions of Theorem 4.1(i) hold. Then we have:

(i) \(\|\hat{h}_n - \Pi_n h_0\| = O_p(\max\{\delta_{m,n}, \sqrt{\lambda_n \delta_{P,n}}, \sqrt{\lambda_n |P(\hat{h}_n) - P(\Pi_n h_0)|}, \|\Pi_n h_0 - h_0\|\})\).

(ii) Under Assumption 3.2(c), \(\|\hat{h}_n - \Pi_n h_0\| = O_p(\max\{\delta_{m,n}, o(\sqrt{\lambda_n}), \|\Pi_n h_0 - h_0\|\})\).

(iii) Under the condition (iii) of Theorem 4.1, \(\|\hat{h}_n - \Pi_n h_0\| = O_p(\max\{\delta_{m,n}, \sqrt{\lambda_n \delta_{P,n}}, \sqrt{\lambda_n \|\hat{h}_n - \Pi_n h_0\|_s}, \|\Pi_n h_0 - h_0\|\})\).

**LEMMA B.2:** Let \(H_n = \text{clsp}\{q_1, \ldots, q_{k(n)}\}\) and \(\{q_j\}_{j=1}^\infty\) be a Riesz basis for \((H, \|\cdot\|_s)\).
(i) If Assumption 5.2(i) holds, then \( \omega_n(\delta, \mathcal{H}_{osn}) \leq \text{const.} \times \delta / \sqrt{\varphi(v_{k(n)}^{-2})} \) and 
\( \tau_n \leq \text{const.} / \sqrt{\varphi(v_{k(n)}^{-2})} \).

(ii) If Assumption 5.2(ii) holds, then 
\[ \|h_0 - \Pi_n h_0\| \leq \text{const.} \sqrt{\varphi(v_{k(n)}^{-2})} \|h_0 - \Pi_n h_0\|_v. \]

(iii) If Assumption 5.2(i) and (ii) holds, then 
\[ \omega_n(\|\Pi_n h_0 - h_0\|, \mathcal{H}_{osn}) \leq c\|\Pi_n h_0 - h_0\|_v. \]

**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 \( \delta^2 / \varphi(v_{k^+}^{-2}) < M^2(v_{k^+}^{-2a}) \) and \( \delta^2 / \varphi(v_{k^+}^{-2}) \geq M^2(v_{k^+}^{-2a}); \) hence we have:

(i) \( \omega(\delta, \mathcal{H}_{os}) \leq \text{const.} \times \delta / \sqrt{\varphi(v_{k^+}^{-2})}. \)

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

**APPENDIX C: LEMMAS FOR SERIES LS ESTIMATOR \( \hat{m}(\cdot) \) OF \( m(\cdot) \)**

Under the following two mild assumptions, we show that the series LS estimator \( \hat{m}(X, h) \) defined in (11) satisfies Assumption 3.3 with \( \eta_{0,n} = \delta_{m,n}^2 = \max\{\frac{\varepsilon}{n}, b_{m, J_n}\} \), where \( \frac{\varepsilon}{n} \) 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) \( X \) is a compact connected subset of \( \mathbb{R}^d \), with Lipschitz continuous boundary, and \( f_X \) is bounded and bounded away from zero over \( X \). (iii) \( \max_{1 \leq j \leq J_n} E[|p_j(X)|^2] \leq \text{const.} \); and the smallest eigenvalue of \( E[p_{J_n}(X)p_{J_n}(X)'] \) is bounded away from zero for all \( J_n \). (iv) Either \( \xi_{n} J_n = o(n) \) with \( \xi \equiv \sup_{x \in \mathcal{X}} \|p_{J_n}(X)\|_t \), 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 \( K I \leq W(x) \leq K' I \) for all \( x \in \mathcal{X}, W(x) \) is positive definite for almost all \( X \in \mathcal{X} \), and \( \sup_{x \in \mathcal{X}} W(x) = W(x)_{ir} = o_p(1) \).

Let \( N_1(\varepsilon, \mathcal{F}_n, \| \cdot \|_{L^2(f_Z)}) \) be the \( L^2(f_Z) \) covering number with bracketing of a class of functions \( \mathcal{F}_n \). For \( j = 1, \ldots, J_n \), denote \( \mathcal{O}_{J_n} \equiv \{p_j(\cdot)\rho(\cdot, h) : h \in \mathcal{H}_{k(n)}^{M_0}\} \) and \( \mathcal{O}_{ojn} \equiv \{p_j(\cdot)\rho(\cdot, h) : h \in \mathcal{H}_{osn}\} \). Denote

\[
C_n(j) \equiv \int_0^1 \sqrt{1 + \log N_1(\varepsilon, \mathcal{O}_{J_n}, \| \cdot \|_{L^2(f_Z)})} \, dw,
\]

\[
C_{on}(j) \equiv \int_0^1 \sqrt{1 + \log N_1(\varepsilon, \mathcal{O}_{ojn}, \| \cdot \|_{L^2(f_Z)})} \, dw.
\]
ASSUMPTION C.2: (i) There is a sequence of measurable functions \( \{\tilde{\rho}_n(Z)\}_{n=1}^{\infty} \) and a finite constant \( K > 0 \), such that \( \sup_{h \in \mathcal{H}^{M_0}_{k(n)}} |\rho(Z, h)| \leq \tilde{\rho}_n(Z) \) and \( E[\tilde{\rho}_n(Z)^2 | X] \leq K \). (ii) There is \( p_{J_n}(X) \) such that \( E[\{\rho(Z/h) - p_{J_n}(X)\}^2] = O(n^{-m_{J_n}}) \) uniformly over \( h \in \mathcal{H}^{M_0}_{k(n)} \). (iii) \( \max_{1 \leq j \leq J_n} C_n(j) \leq \sqrt{C_n} < \infty \) and \( \frac{1}{J_n} C_n = o(1) \). (iv) \( \max_{1 \leq j \leq J_n} C_{on}(j) \leq \sqrt{C} < \infty \).

In Assumption C.1, if \( p_{J_n}(X) \) is a spline, cosine/sine, or wavelet sieve, then \( \xi_n \approx J_n^{1/2} \); see, for example, 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 \( \Lambda^\alpha_{m,c}(X) \)), then Assumption C.2(ii) holds for tensor-product polynomial splines, wavelets, or Fourier series sieves with \( b_{m,J_n} = J_n^{-m_r} \), 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 \( D_T \) denote the distance generated by the topology \( T \) (on \( \mathcal{H} \)) such that \( \mathcal{H}^{M_0}_{k(n)} \) is totally bounded under \( D_T \).

REMARK C.1: Suppose that there are finite constants \( \kappa \in (0, 1] \), \( K > 0 \), such that

\[
\max_{1 \leq j \leq J_n} E \left[ \left\{ p_j(X) \right\}^2 \sup_{h' \in \mathcal{H}^{M_0}_{k(n)}; D_T(h', h) \leq \delta} |\rho(Z, h') - \rho(Z, h)|^2 \right] \leq K^2 \delta^{2\kappa}
\]

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

(i) \( \max_{1 \leq j \leq J_n} N(\|\epsilon\|, O_{m,J_n}; \|\cdot\|_{L^2(f_Z)}) \leq N(1/\kappa, \mathcal{H}^{M_0}_{k(n)}, D_T) \).

(ii) Assumption C.2(iii) is satisfied with

\[
\int_0^1 \sqrt{1 + \log N \left( w^{1/\kappa}, \mathcal{H}^{M_0}_{k(n)}, D_T \right)} \, dw \leq \sqrt{C_n}
\]

and \( \frac{1}{J_n} C_n = o(1) \).

(iii) Assumption C.2(iv) is satisfied with

\[
\int_0^1 \sqrt{1 + \log N \left( w^{1/\kappa}, \mathcal{H}_{m,J_n}, D_T \right)} \, dw \leq \sqrt{C} < \infty.
\]

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

LEMMA C.1: Let \( \tilde{m}(\cdot, h) \) be the series LS estimator defined in (11) and let Assumption C.1 hold.
(i) If \( \text{Var}[\rho(Z, \Pi_n h_0)|X] \leq K \), then
\[
\frac{1}{n} \sum_{i=1}^{n} \left\| \hat{m}(X_i, \Pi_n h_0) - \tilde{m}(X_i, \Pi_n h_0) \right\|_{W}^2 = O_p\left( \frac{J_n}{n} \right).
\]

(ii) If Assumption C.2(i) and (iii) holds, then
\[
\sup_{h \in \mathcal{H}_{k(n)}} \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{m}(X_i, h) - \tilde{m}(X_i, h) \right\|_{W}^2 = O_p\left( \frac{J_n}{n} C_n \right) = o_p(1).
\]

(iii) If Assumption C.2(i) and (iv) holds, then
\[
\sup_{h \in \mathcal{H}_{osn}} \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{m}(X_i, h) - \tilde{m}(X_i, h) \right\|_{W}^2 = O_p\left( \frac{J_n}{n} \right) = o_p(1).
\]

**Lemma C.2:** Let \( \hat{m}(\cdot, h) \) be the series LS estimator defined in (11) and let Assumption C.1 hold.

(i) If Assumption C.2(i) and (ii) holds at \( h = \Pi_n h_0 \), then, with \( \eta_{0,n} = \max\left\{ \frac{J_n}{n}, b_{m,J_n}^2 \right\} \),
\[
cE\left[ \left\| m(X, \Pi_n h_0) \right\|_{W}^2 \right] - O_p(\eta_{0,n}) \leq \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{m}(X_i, \Pi_n h_0) \right\|_{W}^2 
\leq cE\left[ \left\| m(X, \Pi_n h_0) \right\|_{W}^2 \right] + O_p(\eta_{0,n}).
\]

(ii) If Assumption C.2(i)–(iii) holds, then there are finite constants \( K, K' > 0 \) such that, with \( \tilde{\delta}_{m,n}^2 = \frac{J_n}{n} C_n + b_{m,J_n}^2 = o(1) \) and uniformly over \( h \in \mathcal{H}_{k(n)} \),
\[
KE\left[ \left\| m(X, h) \right\|_{W}^2 \right] - O_p(\tilde{\delta}_{m,n}) \leq \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{m}(X_i, h) \right\|_{W}^2 
\leq K'E\left[ \left\| m(X, h) \right\|_{W}^2 \right] + O_p(\tilde{\delta}_{m,n}).
\]

(iii) If Assumption C.2(i), (ii), and (iv) holds, then there are finite constants \( K, K' > 0 \) such that, with \( \delta_{m,n}^2 = \eta_{0,n} = \max\left\{ \frac{J_n}{n}, b_{m,J_n}^2 \right\} \) and uniformly over \( h \in \mathcal{H}_{osn} \),
\[
KE\left[ \left\| m(X, h) \right\|_{W}^2 \right] - O_p(\delta_{m,n}) \leq \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{m}(X_i, h) \right\|_{W}^2 
\leq K'E\left[ \left\| m(X, h) \right\|_{W}^2 \right] + O_p(\delta_{m,n}).
\]
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 (2009c) for a more general version and its applications to derive the convergence rates and limiting distributions of plug-in PSMD estimators of a large class of functionals of \( h_0 \) satisfying \( E[\rho(Y, X; h_0(\cdot))|X] = 0. \)

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

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

where \( M_0 \) is a finite but large number such that \( \hat{h}_n \in N_{os} \) for large \( n \), with probability greater than \( 1 - \epsilon \) for a small \( \epsilon > 0 \). Let \( N_{osn} \equiv \{ h \in \mathcal{H}_{osn} : \|h - h_0\|_s \leq M_0 \delta_{s,n} \} \).

**Lemma C.3:** Let \( \hat{m}(\cdot, h) \) be the series LS estimator defined in (11) and let Assumption C.1 hold. Suppose that the following conditions hold:

(i) There are finite constants \( \kappa \in (0, 1], K > 0 \) such that

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

for all \( h' \in N_{osn} \cup \{h_0\} \) and all positive value \( \delta = o(1) \).

(ii) \( \int_0^1 \sqrt{1 + \log N(w^{1/\kappa}, N_{osn}, \| \cdot \|_s)} \, dw \leq \sqrt{C} < \infty \).

Then

\[
\sup_{N_{osn}} \frac{1}{n} \sum_{i=1}^n \| \hat{m}(X_i, h) - \hat{m}(X_i, h_0) - \tilde{m}(X_i, h) \|_F^2 = O_p \left( \frac{J_n}{n} (\delta_{s,n})^{2\kappa} \right).
\]

**REFERENCES**


SUPPLEMENT TO “ESTIMATION OF NONPARAMETRIC CONDITIONAL MOMENT MODELS WITH POSSIBLY NONSmooth GENERALIZED RESIDUALS”

(Econometrica, Vol. 80, No. 1, January 2012, 277–321)

BY XIAOHONG CHEN AND DEMIAN POUZO

In this supplemental 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 $S(\mathbb{R}^d)$ be the Schwartz space of all complex-valued, rapidly decreasing, infinitely differentiable functions on $\mathbb{R}^d$. Let $S^*(\mathbb{R}^d)$ be the space of all tempered distributions on $\mathbb{R}^d$, which is the topological dual of $S(\mathbb{R}^d)$. For $h \in S(\mathbb{R}^d)$, we let $\hat{h}$ denote the Fourier transform of $h$ (i.e., $\hat{h}(\xi) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \exp\{iy'\xi\}h(y)dy$) and let $(g)^\vee$ denote the inverse Fourier transform of $g$ (i.e., $(g)^\vee(y) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \exp\{iy'\xi\}g(\xi)d\xi$). Let $\varphi_0 \in S(\mathbb{R}^d)$ 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 $x \in \mathbb{R}^d$). Let $\nu \in \mathbb{R}$ and $p, q \in (0, \infty]$. The Besov space $B_{\nu, p/q}(\mathbb{R}^d)$ is the collection of all functions $h \in S^*(\mathbb{R}^d)$ such that $\|h\|_{B_{\nu, p/q}}$ is finite:

$$
\|h\|_{B_{\nu, p/q}} \equiv \left( \sum_{j=0}^{\infty} \left\{ 2^j \nu \left\| (\varphi_j \hat{h})^\vee \right\|_{L_p(\text{leb})} \right\}^q \right)^{1/q} < \infty
$$

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

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

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

© 2012 The Econometric Society DOI: 10.3982/ECTA7888
Let $T^\nu_p (\mathcal{R}^d)$ denote either $B^\nu_p (\mathcal{R}^d)$ or $\mathcal{F}^\nu_p (\mathcal{R}^d)$. Then $T^\nu_p (\mathcal{R}^d)$ gets larger with increasing $q$ (i.e., $T^\nu_{p,q_1} (\mathcal{R}^d) \subseteq T^\nu_{p,q_2} (\mathcal{R}^d)$ for $q_1 \leq q_2$), gets larger with decreasing $p$ (i.e., $T^\nu_{p_1,q} (\mathcal{R}^d) \subseteq T^\nu_{p_2,q} (\mathcal{R}^d)$ for $p_1 \geq p_2$), and gets larger with decreasing $\nu$ (i.e., $T^\nu_{p,q_1} (\mathcal{R}^d) \subseteq T^\nu_{p,q_2} (\mathcal{R}^d)$ for $\nu_1 \geq \nu_2$). Also, $T^\nu_{p,q} (\mathcal{R}^d)$ becomes a Banach space when $p, q \geq 1$. The spaces $T^\nu_{p,q} (\mathcal{R}^d)$ include many well known function spaces as special cases. For example, for $p \in (1, \infty)$, the Hölder space $\Lambda^\nu (\mathcal{R}^d) = B^\nu_{\infty,\infty} (\mathcal{R}^d)$ for any real-valued $r > 0$, the Hilbert–Sobolev space $W^k_2 (\mathcal{R}^d) = B^k_{2,2} (\mathcal{R}^d)$ for integer $k > 0$, and the (fractional) Sobolev space $W^\nu_p (\mathcal{R}^d) = \mathcal{F}^\nu_p (\mathcal{R}^d)$ for any $\nu \in \mathcal{R}$ and $p \in (1, \infty)$, which has the equivalent norm $\|h\|_{W^\nu_p} \equiv \|(1 + |\cdot|^2)^{\nu/2} \hat{h} (\cdot)\|_{L^p (\mathcal{R}^d)} < \infty$ (note that for $\nu > 0$, the norm $\|h\|_{W^\nu_p}$ is a shrinkage in the Fourier domain).

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

We define “weighted” versions of the space $T^\nu_{p,q} (\mathcal{R}^d)$ as follows. Let $w(\cdot) = (1 + |\cdot|^2)^{\xi/2}$, $\xi \in \mathcal{R}$, be a weight function and define $\|h\|_{T^\nu_{p,q}(\mathcal{R}^d,w)} = \|wh\|_{T^\nu_{p,q}(\mathcal{R}^d)}$, that is, $T^\nu_{p,q} (\mathcal{R}^d, w) = \{h : \|wh\|_{T^\nu_{p,q}(\mathcal{R}^d)} < \infty\}$.

Then the embedding of $T^\nu_{p_1, q_1} (\mathcal{R}^d, w_1)$ into $T^\nu_{p_2, q_2} (\mathcal{R}^d, w_2)$ is compact if and only if $\nu_1 - \nu_2 > d(p_1^{-1} - p_2^{-1})$, $w_2 (x)/w_1 (x) \rightarrow 0$ as $|x| \rightarrow \infty$, and $-\infty < \nu_2 < \nu_1 < \infty$, $0 < q_1, q_2 \leq \infty$, and $0 < p_1, p_2 \leq \infty$ ($0 < p_1, p_2 < \infty$ for $\mathcal{F}^\nu_{p,q} (\Omega)$).

If $\mathcal{H} \subseteq \mathcal{H}$ is a Besov space, then a wavelet basis $\{\psi_j\}$ is a natural choice of $\{q_j\}$ to satisfy Assumption 5.1 in Section 5. A real-valued function $\psi$ is called a mother wavelet of degree $\gamma$ if it satisfies (a) $\int_{\mathcal{R}} y^k \psi (y) \, dy = 0$ for $0 \leq k \leq \gamma$, (b) $\psi$ and all its derivatives up to order $\gamma$ decrease rapidly as $|y| \rightarrow \infty$ and (c) $\{2^{k/2} \psi (2^k y - j) : k, j \in \mathcal{Z}\}$ forms a Riesz basis of $L^2 (\mathcal{R} \cap \mathcal{H})$, that is, the linear span of $\{2^{k/2} \psi (2^k y - j) : k, j \in \mathcal{Z}\}$ is dense in $L^2 (\mathcal{R} \cap \mathcal{H})$ and

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

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

Some examples of sieves follow:

**Orthogonal Wavelets:** Given an integer $\gamma > 0$, there exist a father wavelet $\varphi$ of degree $\gamma$ and a mother wavelet $\psi$ of degree $\gamma$, both compactly supported,
such that for any integer $k_0 \geq 0$, any function $h$ in $L^2(\text{leb})$ has the wavelet $\gamma$-regular multiresolution expansion

$$ h(y) = \sum_{j=-\infty}^{\infty} a_{k_0j} \varphi_{k_0j}(y) + \sum_{k=k_0}^{\infty} \sum_{j=-\infty}^{\infty} b_{kj} \psi_{kj}(y), \quad y \in \mathcal{R}, $$

where $\{\varphi_{k_0j}, j \in \mathbb{Z}; \psi_{kj}, k \geq k_0, j \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\text{leb})$; 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 $\gamma$:

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

**Cardinal B-Spline Wavelets of Order $\gamma$:**

(SM.1) \hspace{1cm} $h_n(y) = \psi^{kn}(y)' \Pi = \sum_{k=0}^{K_n} \sum_{j \in K_n} \pi_{kj} 2^{k/2} B_{\gamma}(2^k y - j), \quad k(n) = 2^{K_n} + 1,$

where $B_{\gamma}(\cdot)$ is the cardinal B-spline of order $\gamma$:

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

**Polynomial Splines of Order $q_n$:**

(SM.2) \hspace{1cm} $h_n(y) = \psi^{kn}(y)' \Pi$

$$ = \sum_{j=0}^{q_n} \pi_j(y)^j + \sum_{k=1}^{r_n} \pi_{qn+k}(y - \nu_k)^{q_n}, \quad k(n) = q_n + r_n + 1, $$

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

**Hermite Polynomials of Order $k(n) - 1$:**

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

where $\nu_1$ and $\nu_2$ can be chosen as the sample mean and variance of the data.
B. CONSISTENCY: PROOFS OF THEOREMS

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

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

exists and is strictly positive (under Assumption 3.1(i) and (ii)). Moreover, for fixed $k$, $g(k, \varepsilon)$ increases as $\varepsilon$ increases. For any fixed $\varepsilon > 0$, $g(k, \varepsilon)$ decreases as $k$ increases, and could go to zero as $k$ goes to infinity. Following the proof of Lemma A.4(i) with $T = \|\cdot\|_s$ topology, $\mathcal{H}_k^{M_0} \subseteq \mathcal{H}_k^{M_0(n)}$, $\lambda^n P(h) \geq 0$, and $\eta_n = O(\eta_0, n)$, we have, for all $\varepsilon > 0$ and $n$ sufficiently large,

$$\Pr(\|\hat{h}_n - h_0\|_s \geq \varepsilon) \leq \Pr(\|\hat{h}_n - h_0\|_s \geq \varepsilon, \hat{h}_n \in \mathcal{H}_k^{M_0(n)}) + \varepsilon$$

$$\leq \Pr\left(\inf_{h \in \mathcal{H}_k^{M_0(n)} : \|h - h_0\|_s \geq \varepsilon} \{cE[\|m(X, h)\|_W^2] + \lambda^n P(h)\}\right) \leq c' E[\|m(X, \Pi_n h_0)\|_W^2]$$

$$+ O_p(\eta_0, n) + O_p(\bar{\delta}_{m,n}^2) + \lambda^n P(h_0) + O_p(\lambda_n) + \varepsilon$$

$$\leq \Pr\left(\inf_{h \in \mathcal{H}_k^{M_0} : \|h - h_0\|_s \geq \varepsilon} \{cE[\|m(X, h)\|_W^2]\} \leq c'E[\|m(X, \Pi_n h_0)\|_W^2]$$

$$+ O_p(\eta_0, n) + O_p(\bar{\delta}_{m,n}^2) + \lambda^n P(h_0) + O_p(\lambda_n) + \varepsilon$$

$$\leq \Pr\left(g(k(n), \varepsilon) \leq O_p(\max\{\bar{\delta}_{m,n}^2, \eta_0, n, E(\|m(X, \Pi_n h_0)\|_W^2), \lambda_n\})\right) + \varepsilon,$$

which goes to zero under $\max\{\bar{\delta}_{m,n}^2, \eta_0, n, E(\|m(X, \Pi_n h_0)\|_W^2), \lambda_n\} = o(g(k(n), \varepsilon)).$ Thus $\|\hat{h}_n - h_0\|_s = o_p(1)$.

Q.E.D.

PROOF OF THEOREM 3.2: Under the assumptions that $E[m(X, h)W(X) \times m(X, h)]$ is lower semicontinuous and $P(h)$ is lower semicompact on $(\mathcal{H}, \|\cdot\|_s)$, we have that for all $\varepsilon > 0$,

$$g(\varepsilon) \equiv \min_{h \in \mathcal{H} : \|h - h_0\|_s \geq \varepsilon} E[m(X, h)W(X)m(X, h)]$$
exists (by Theorem 38.B in Zeidler (1985)) and is strictly positive (under Assumption 3.1(i) and (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.4(i) with \( T = \| \cdot \|_s \) topology, \( \mathcal{H}^M \subseteq \mathcal{H}^M_0 \), \( \lambda_n > 0 \), \( P(h) \geq 0 \), \( \eta_n = O(\eta_{0,n}) \), and \( \max\{\eta_{0,n}, E[\|m(X, \Pi_n h_0)\|_W^2]\} = O(\lambda_n) \), we have, for all \( \varepsilon > 0 \) and \( n \) sufficiently large,

\[
\Pr(\|\hat{h}_n - h_0\|_s \geq \varepsilon) \\
\leq \Pr(\|\hat{h}_n - h_0\|_s \geq \varepsilon, \hat{h}_n \in \mathcal{H}_M^M) + \varepsilon \\
\leq \Pr \left( \inf_{h \in \mathcal{H}_M^M; \|h - h_0\|_s \geq \varepsilon} \left\{ cE\left[\|m(X, h)\|_W^2\right] + \lambda_n P(h) \right\} \right) \\
\leq O_p(\delta^2_{m,n} + \lambda_n P(h_0) + O_p(\lambda_n)) + \varepsilon \\
\leq \Pr \left( \inf_{h \in \mathcal{H}_M^M; \|h - h_0\|_s \geq \varepsilon} \left[ cE\left[\|m(X, h)\|_W^2\right] \right] \right) \\
\leq O_p(\delta^2_{m,n} + \lambda_n P(h_0) + O_p(\lambda_n)) + \varepsilon \\
\leq \Pr \left( g(\varepsilon) \leq O_p(\max\{\delta^2_{m,n}, \lambda_n\}) \right) + \varepsilon,
\]

which goes to zero under \( \max\{\delta^2_{m,n}, \lambda_n\} = o(1) \). Thus \( \|\hat{h}_n - h_0\|_s = o_p(1) \).

**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, \( \hat{h}_n \in \mathcal{H}_k(n) \) with probability approaching 1. By Lemma A.3(ii) with \( \max\{\eta_{0,n}, E[\|m(X, \Pi_n h_0)\|_W^2]\} = o(\lambda_n) \) and \( \eta_n = O(\eta_{0,n}) \), we have \( P(\hat{h}_n) - P(h_0) \leq o_p(1) \); thus we can focus on the set \( \{ h \in \mathcal{H}_k(n) : P(h) \leq M_0 \} = \mathcal{H}_M^M_0 \) for all \( n \) large enough. Let \( B_w(h_0) \) denote any open neighborhood (in the weak topology) around \( h_0 \), and let \( B^c_w(h_0) \) denote its complement (under the weak topology) in \( \mathcal{H} \). By Lemma A.4(ii) with \( B_w(h_0) = B_w(h_0), \lambda_n P(h) \geq 0, \mathcal{H}_M^M_0 \subseteq \mathcal{H}, \eta_n = O(\eta_{0,n}), \) and \( \max\{\eta_{0,n}, E[\|m(X, \Pi_n h_0)\|_W^2]\} = o(\lambda_n) \), we have, for all nonempty open balls \( B_w(h_0), \) all \( \varepsilon > 0, \) and \( n \) sufficiently large,

\[
\Pr(\hat{h}_n \notin B_w(h_0)) \\
\leq \Pr(\hat{h}_n \notin B_w(h_0), \hat{h} \in \mathcal{H}_M^M_0) + \varepsilon \\
\leq \Pr \left( \inf_{h \in \mathcal{H}_M^M_0; \hat{h} \notin B_w(h_0)} \left\{ cE\left[\|m(X, h)\|_W^2\right] + \lambda_n P(h) \right\} \right).
\]
\[
\begin{align*}
\leq O_p(\delta_{m,n}^2) + &\lambda_nP(h_0) + o(\lambda_n) + \varepsilon \\
\leq &\Pr\left( \inf_{x \in B(h_0)} E[\|m(X, h)\|_W^2] \leq O_p(\max\{\delta_{m,n}, \lambda_n\}) \right) + \varepsilon.
\end{align*}
\]

Let \( E[\|m(X, h)\|_W^2] \) be weak sequentially lower semicontinuous on \( \mathcal{H} \). Since \( \mathcal{H} \cap B_{\delta}(h_0) \) is weakly compact (weakly closed and bounded), by Assumption 3.4(ii) and Theorem 38.1 in Zeidler (1985), there exists \( h^*(B) \in \mathcal{H} \cap B_{\delta}(h_0) \) such that \( \inf_{x \in B(h_0)} E[\|m(X, h)\|_W^2] = E[\|m(X, h^*(B))\|_W^2] \). It must hold that \( g(B) \equiv E[\|m(X, h^*(B))\|_W^2] > 0 \); otherwise, by Assumption 3.1(i) and (ii), \( \|h^*(B) - h_0\|_s = 0 \). But if this is the case, then for any \( t \in \mathcal{H}^* \) we have \( \langle t, h^*(B) - h_0 \rangle_{\mathcal{H}^*, \mathcal{H}} \leq \text{const.} \times \|h^*(B) - h_0\|_s = 0 \), a contradiction to the fact that \( h^*(B) \notin B_{\delta}(h_0) \). Thus
\[
\Pr(h_n \notin B_{\delta}(h_0), \hat{h}_n \in \mathcal{H}^{BM}) \leq \Pr(E[\|m(X, h^*(B))\|_W^2] \leq O_p(\max\{\delta_{m,n}, \lambda_n\})),
\]
which goes to zero since \( \max\{\delta_{m,n}, \lambda_n\} = o(1) \). Hence \( \Pr(h_n \notin B_{\delta}(h_0)) \rightarrow 0 \).

**Step 2.** Consistency under the weak topology implies that \( \langle t_0, \hat{h}_n - h_0 \rangle_{\mathcal{H}^*, \mathcal{H}} = o_p(1) \). By Assumption 3.4(i), \( P(\hat{h}_n) - P(h_0) \geq \langle t_0, \hat{h}_n - h_0 \rangle_{\mathcal{H}^*, \mathcal{H}} + g(\|\hat{h}_n - h_0\|_s) \). Lemma A.3(ii) implies that \( P(\hat{h}_n) - P(h_0) \leq o_p(1) \) under \( \max\{\eta_{0,n}, E[\|m(X, \Pi_0h_0)\|_W^2]\} = o(\lambda_n), \eta_n = O(\eta_{0,n}) \). Thus \( g(\|\hat{h}_n - h_0\|_s) = o_p(1) \) and \( \|\hat{h}_n - h_0\|_s = o_p(1) \) by our assumption over \( g(\cdot) \). This, \( \langle t_0, \hat{h}_n - h_0 \rangle_{\mathcal{H}^*, \mathcal{H}} = o_p(1) \), and Assumption 3.4(ii) imply that \( P(\hat{h}_n) - P(h_0) \geq o_p(1) \). But \( P(\hat{h}_n) \leq P(h_0) + o_p(1) \) by Lemma A.3(ii). Thus \( P(\hat{h}_n) = P(h_0) = o_p(1) \).

**Verification of Remark 3.2:** Claim (i) follows from Proposition 38.7 of Zeidler (1985). Claim (ii) follows from Corollary 41.9 of Zeidler (1985). For claim (iii), the fact that \( \sqrt{W(\cdot)}m(\cdot, h) : \mathcal{H} \rightarrow L^2(f_X) \) is compact and Frechet differentiable implies that its Frechet derivative is also a compact operator; see Zeidler (1985, Proposition 7.33). This and the chain rule imply that the functional \( E[\|m(X, \cdot)\|_W^2] : \mathcal{H} \rightarrow [0, \infty) \) is Frechet differentiable and its Frechet derivative is compact on \( \mathcal{H} \). Hence \( E[\|m(X, h)\|_W^2] \) has a compact Gateaux derivative on \( \mathcal{H} \) and, by claim (ii), is weak sequentially lower semicontinuous on \( \mathcal{H} \).

**Proof of Theorem A.1:** For result (i), we first show that the set of minimum penalization solutions, \( \mathcal{M}_{0}^p \), is not empty. Since \( E[\|m(X, h)\|_W^2] \) 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 nonempty). Since \( P(\cdot) \) is convex and lower semicontinuous on \( \mathcal{M}_0 \), by applying Proposition 38.15 of Zeidler (1985), we have that the set \( \mathcal{M}_0^p \) is nonempty, convex, closed, and a...
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 since $\mathcal{M}_0$ is convex, $h' = \lambda h_1 + (1 - \lambda) h_0 \in \mathcal{M}_0$. Since $P(\cdot)$ 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 have established result (i).

For result (ii), first, as already shown earlier, $\hat{h}_n \in \mathcal{H}_{k(n)}$ with probability approaching 1. 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 $B_w(h_0)$ denote any open neighborhood (in the weak topology) around $h_0$, and let $B_cw(h_0)$ denote its complement (under the weak topology) in $\mathcal{H}$. By Lemma A.3(ii), $P(\hat{h}_n) = O_p(1)$. By Lemma A.4(ii) with $B_T(h_0) = B_w(h_0)$, $\mathcal{H}_{k(n)} \subseteq \mathcal{H}_{k(n)}^{M_0}$, $\eta_n = O(\eta_0, n)$, and $\max\{\hat{\delta}^2, \eta_0, n, E[\|m(X, \Pi_h h_0)\|_W^2]\} = o(\lambda_n)$, we have, for all nonempty open balls $B_w(h_0)$,

$$
\Pr(\hat{h}_n /\notin B_w(h_0), \hat{h}_n \in \mathcal{H}_{k(n)}^{M_0}) \\
\leq \Pr\left(\inf_{\mathcal{H}_{k(n)} \cap B_w(h_0)} \{cE[\|m(X, h)\|_W^2] + \lambda_n P(h)\} \right) \\
\leq \lambda_n P(h_0) + o_p(\lambda_n) \\
\leq \Pr\left(\inf_{\mathcal{H}_{k(n)} \cap B_w(h_0)} \{cE[\|m(X, h)\|_W^2] + \lambda_n P(h)\} \right) \\
\leq \lambda_n P(h_0) + o_p(\lambda_n).
$$

By Assumptions 3.1(iii) and 3.4(ii), $\mathcal{H}_{k(n)}$ is weakly sequentially compact. Since $B_w(h_0)$ is closed under the weak topology, the set $\mathcal{H}_{k(n)} \cap B_w(h_0)$ is weakly sequentially compact. By Assumption 3.4(ii) and the assumption that $E[\|m(X, h)\|_W^2] + \lambda_n P(h)$ is weakly sequentially lower semicontinuous on $\mathcal{H}_{k(n)}$, we have its minimizer as $h_n(\varepsilon) \in \mathcal{H}_{k(n)} \cap B_w(h_0)$. Hence, with $\max\{\hat{\delta}^2, \eta_0, n, E[\|m(X, \Pi_h h_0)\|_W^2]\} = o(\lambda_n)$ and $\lambda_n > 0$, we have

$$
\Pr(\hat{h}_n /\notin B_w(h_0), \hat{h}_n \in \mathcal{H}_{k(n)}^{M_0}) \\
\leq \Pr\left(\inf_{\mathcal{H}_{k(n)} \cap B_w(h_0)} \{cE[\|m(X, h_n(\varepsilon))\|_W^2] + \lambda_n P(h_n(\varepsilon))\} \right) \\
\leq \Pr\left(\frac{g(k(n), \varepsilon, \lambda_n) - \lambda_n P(h_0)}{\lambda_n} \leq o_p(1)\right).
$$
If \( \liminf_{n} E[\|m(X, h_{n}(\varepsilon))\|_{W}^{2}] = \text{const.} > 0 \), then \( \text{Pr}(\hat{h}_{n} \notin B_{w}(h_{0}), \hat{h}_{n} \in H_{M_{0}}^{0}) \to 0 \) trivially. So we assume \( \liminf_{n} E[\|m(X, h_{n}(\varepsilon))\|_{W}^{2}] = \text{const.} = 0 \). Since \( \mathcal{H} \cap B_{w}^{c}(h_{0}) \) is weakly compact, there exists a subsequence \( \{h_{n_k}(\varepsilon)\}_{k} \) that converges (weakly) to \( h_{\infty}(\varepsilon) \in \mathcal{H} \cap B_{w}^{c}(h_{0}) \). By weakly lower semicontinuity of \( E[\|m(X, h)\|_{W}^{2}] \) on \( \mathcal{H} \), \( h_{\infty}(\varepsilon) \in M_{0} \). By definition of \( h_{0} \) and the assumption that \( P(h) \) is strictly convex in \( h \in M_{0} \), it must be that

\[
P(h_{\infty}(\varepsilon)) - P(h_{0}) \geq \text{const.} > 0 \text{ by result (i). Note that this is true for any convergent subsequence.}
\]

Therefore, we have established that

\[
\liminf_{n} \frac{g(k(n), \varepsilon, \lambda_{n}) - \lambda_{n} P(h_{0})}{\lambda_{n}} \geq \text{const.} > 0;
\]

thus \( \text{Pr}(\hat{h}_{n} \notin B_{w}(h_{0}), \hat{h}_{n} \in H_{M_{0}}^{0}) \to 0 \). Hence, by similar calculations to those in Lemma A.4(ii), for any \( \varepsilon > 0 \) and sufficiently large \( n \), \( \text{Pr}(\hat{h}_{n} \notin B_{w}(h_{0})) \leq \text{Pr}(\hat{h}_{n} \notin B_{w}(h_{0}), \hat{h}_{n} \in \mathcal{H}_{M_{0}}^{0} + \varepsilon \leq 2\varepsilon.
\]

Given the consistency under the weak topology, Assumption 3.4(i) and Lemma A.3(ii), we obtain \( \|\hat{h}_{n} - h_{0}\| = o_{p}(1) \) and \( P(\hat{h}_{n}) - P(h_{0}) = o_{p}(1) \) by following Step 2 in the proof of Theorem 3.3.

Q.E.D.

C. CONSISTENCY: PROOFS OF LEMMAS

PROOF OF LEMMA A.1: By definition of the infimum, \( \hat{\alpha}_{n} \) always exists, and \( \hat{\alpha}_{n} \in A_{k(n)} \) with outer probability approaching 1 (\( \hat{\alpha}_{n} \) may not be measurable). It follows that for all \( B_{T}(\alpha_{0}) \),

\[
\text{Pr}^{*}(\hat{\alpha}_{n} \in A_{k(n)}, \hat{\alpha}_{n} \notin B_{T}(\alpha_{0})) \leq \text{Pr}^{*}\left( \inf_{\alpha \in A_{k(n)}: \alpha \notin B_{T}(\alpha_{0})} \hat{Q}_{n}(\alpha) \leq \hat{Q}_{n}(\Pi_{n} \alpha_{0}) + O_{p_{s}}(\eta_{n}) \right)
\]

\[
\leq \text{Pr}^{*}\left( \inf_{\alpha \in A_{k(n)}: \alpha \notin B_{T}(\alpha_{0})} \{K\hat{Q}_{n}(\alpha) - O_{p_{s}}(c_{n})\} \right)
\]

\[
\leq K_{0}\hat{Q}_{n}(\Pi_{n} \alpha_{0}) + O_{p_{s}}(c_{0,n}) + O_{p_{s}}(\eta_{n})
\]

\[
\leq \text{Pr}^{*}\left( \inf_{\alpha \in A_{k(n)}: \alpha \notin B_{T}(\alpha_{0})} \hat{Q}_{n}(\alpha) \leq O_{p_{s}}(\max\{c_{n}, c_{0,n}, \hat{Q}_{n}(\Pi_{n} \alpha_{0}), \eta_{n}\}) \right)
\]

\[
\leq \text{Pr}^{*}(g_{0}(n, k(n), B) \leq O_{p_{s}}(\max\{c_{n}, c_{0,n}, \hat{Q}_{n}(\Pi_{n} \alpha_{0}), \eta_{n}\}))
\]

by condition a(ii) in Lemma A.1,

which goes to 0 by condition d(iii) in Lemma A.1.

Q.E.D.
PROOF OF LEMMA A.2: Under condition c(ii) of Lemma A.2, \( \hat{\alpha}_n \) is well defined and measurable. It follows that for any \( \epsilon > 0 \),

\[
\Pr(\|\hat{\alpha}_n - \alpha_0\|_s > \epsilon) \\
\leq \Pr\left( \inf_{\alpha \in A_k(n)} \hat{Q}_n(\alpha) \leq \hat{Q}_n(\Pi_n \alpha_0) + O_p(\eta_n) \right) \\
\leq \Pr\left( \inf_{\alpha \in A_k(n)} |\hat{Q}_n(\alpha) - \overline{Q}_n(\alpha)| \right) \\
\leq \overline{Q}_n(\Pi_n \alpha_0) + |\hat{Q}_n(\Pi_n \alpha_0) - \overline{Q}_n(\Pi_n \alpha_0)| + O_p(\eta_n) \\
\leq \Pr\left( \inf_{\alpha \in A_k(n)} \overline{Q}_n(\alpha) \leq 2\varepsilon_n + \overline{Q}_n(\Pi_n \alpha_0) + O_p(\eta_n) \right) \\
\leq \Pr\left( \inf_{\alpha \in A_k(n)} \overline{Q}_n(\alpha) - \overline{Q}_n(\alpha_0) \leq 2\varepsilon_n + \overline{Q}_n(\Pi_n \alpha_0) - \overline{Q}_n(\alpha_0) + O_p(\eta_n) \right) \\
\leq \Pr\left( g_0(n, k(n), \epsilon) \leq 2\varepsilon_n + |\overline{Q}_n(\Pi_n \alpha_0) - \overline{Q}_n(\alpha_0)| + O_p(\eta_n) \right)
\]

which goes to 0 by condition d of Lemma A.2. Q.E.D.

PROOF OF LEMMA A.3: We first show that \( \hat{h}_n \in \mathcal{H}_n \) w.p.a.1. The infimum \( \inf_{\hat{h}_n} \hat{Q}_n(h) \) exists w.p.a.1 and hence, for any \( \epsilon > 0 \), there is a sequence, \( (h_{j,n}(\epsilon)) \subseteq \mathcal{H}_n \) such that \( \hat{Q}_n(h_{j,n}(\epsilon)) \leq \inf_{\hat{h}_n} \hat{Q}_n(h) + \epsilon \) w.p.a.1. Let \( \hat{h}_n \equiv h_{n,n}(\eta_n) \). Then such a choice satisfies \( \hat{h}_n \in \mathcal{H}_n \) w.p.a.1.

Next, by definition of \( \hat{h}_n \), we have for any \( \lambda_n > 0 \),

\[
\lambda_n \hat{P}_n(\hat{h}_n) \leq \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, \hat{h}_n)\|_W^2 + \lambda_n \hat{P}_n(\hat{h}_n) \\
\leq \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, \Pi_n h_0)\|_W^2 + \lambda_n \hat{P}_n(\Pi_n h_0) + O_p(\eta_n)
\]

and

\[
\lambda_n \{P(\hat{h}_n) - P(h_0)\} + \lambda_n \{\hat{P}_n(\hat{h}_n) - P(\hat{h}_n)\} \\
\leq \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, \Pi_n h_0)\|_W^2 + \lambda_n \{\hat{P}_n(\Pi_n h_0) - P(\Pi_n h_0)\} \\
+ \lambda_n \{P(\Pi_n h_0) - P(h_0)\} + O_p(\eta_n).
\]
Thus
\[
\lambda_n \{ P(\hat{h}_n) - P(h_0) \} \\
\leq \frac{1}{n} \sum_{i=1}^{n} \| \hat{m}(X_i, \Pi_n h_0) \|_{W}^{2} + 2\lambda_n \sup_{h \in H_n} |\hat{P}_n(h) - P(h)| \\
+ \lambda_n |P(\Pi_n h_0) - P(h_0)| + O_p(\eta_n) \\
\leq O_p(\eta_{0, n} + E[\|m(X, \Pi_n h_0)\|_{W}^{2}]) + 2\lambda_n \sup_{h \in H_n} |\hat{P}_n(h) - P(h)| \\
+ \lambda_n |P(\Pi_n h_0) - P(h_0)|,
\]
where the last inequality is due to Assumption 3.3(i) and \( \eta_n = O(\eta_{0, n}). \) Therefore, for all \( M > 0, \)
\[
\Pr(P(\hat{h}_n) - P(h_0) > M) \\
= \Pr(\lambda_n \{ P(\hat{h}_n) - P(h_0) \} > \lambda_n M) \\
\leq \Pr\left( O_p(\eta_{0, n} + E[\|m(X, \Pi_n h_0)\|_{W}^{2}]) + 2\lambda_n \sup_{h \in H_n} |\hat{P}_n(h) - P(h)| \\
+ \lambda_n |P(\Pi_n h_0) - P(h_0)| > \lambda_n M \right).
\]

(i) Under Assumption 3.2(b), \( \lambda_n \sup_{h \in H_n} |\hat{P}_n(h) - P(h)| + \lambda_n |P(\Pi_n h_0) - P(h_0)| = O_p(\lambda_n), \) we have
\[
\Pr(P(\hat{h}_n) - P(h_0) > M) \\
\leq \Pr\left( O_p\left( \max\{ \eta_{0, n} + E[\|m(X, \Pi_n h_0)\|_{W}^{2}], \lambda_n \} \right) > \lambda_n M \right) \\
\leq \Pr\left( O_p\left( \frac{\eta_{0, n} + E[\|m(X, \Pi_n h_0)\|_{W}^{2}]}{\lambda_n} \right) + O_p(1) > M \right),
\]
which, under \( \max\{ \eta_{0, n}, E[\|m(X, \Pi_n h_0)\|_{W}^{2}] \} = O(\lambda_n), \) goes to zero as \( M \to \infty. \)

Thus \( P(\hat{h}_n) - P(h_0) = O_p(1). \) Since \( 0 \leq P(h_0) < \infty, \) we have \( P(\hat{h}_n) = O_p(1). \)

(ii) Under Assumption 3.2(c), \( \lambda_n \sup_{h \in H_n} |\hat{P}_n(h) - P(h)| + \lambda_n |P(\Pi_n h_0) - P(h_0)| = o_p(\lambda_n), \) we have
\[
\Pr(P(\hat{h}_n) - P(h_0) > M) \\
\leq \Pr\left( O_p\left( \frac{\eta_{0, n} + E[\|m(X, \Pi_n h_0)\|_{W}^{2}]}{\lambda_n} \right) + o_p(1) > M \right),
\]
which, under $\max\{\eta_0, E[\|m(X, \Pi_n h_0)\|_W^2]\} = o(\lambda_n)$, goes to zero for all $M > 0$. Thus $P(\hat{h}_n) - P(h_0) \leq o_p(1)$. Q.E.D.

**PROOF OF LEMMA A.4:** It suffices to consider $\lambda_n P(\cdot) > 0$ only. By the fact that $\Pr(A) \leq \Pr(A \cap B) + \Pr(B^c)$ for any measurable sets $A$ and $B$, we have

$$\Pr(\hat{h}_n \notin B_T(h_0)) \leq \Pr(\hat{h}_n \notin B_T(h_0), P(\hat{h}_n) \leq M_0) + \Pr(P(\hat{h}_n) > M_0).$$

For any $\varepsilon > 0$, choose $M_0 \equiv M_0(\varepsilon)$ such that $\Pr(P(\hat{h}_n) > M_0) < \varepsilon$ for sufficiently large $n$. Note that such a $M_0$ always exists by Lemma A.3. 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

$$\Pr(\hat{h}_n \notin B_T(h_0), P(\hat{h}_n) \leq M_0).$$

By definition of $\hat{h}_n$ and $\Pi_n h_0$, Assumptions 3.3 and 3.1(iii), and $\eta_n = O(\eta_0)$, we have, for all $B_T(h_0)$,

$$\Pr(\hat{h}_n \notin B_T(h_0), \hat{h}_n \in \mathcal{H}_{k(n)}^{M_0})$$

$$\leq \Pr\left(\inf_{h \in \mathcal{H}_{k(n)}, h \notin B_T(h_0)} \left\{ \frac{1}{n} \sum_{i=1}^n \|\hat{m}(X_i, h)\|_W^2 + \lambda_n \hat{P}(h) \right\} \right)$$

$$\leq \frac{1}{n} \sum_{i=1}^n \|\hat{m}(X_i, \Pi_n h_0)\|_W^2 + \lambda_n \hat{P}(\Pi_n h_0) + O_p(\eta_n)$$

$$\leq \Pr\left(\inf_{h \in \mathcal{H}_{k(n)}, h \notin B_T(h_0)} \left\{ cE[\|m(X, h)\|_W^2] + \lambda_n \hat{P}(h) \right\} \right)$$

$$\leq \lambda_n \sup_{h \in \mathcal{H}_{k(n)}} |\hat{P}(h) - P(h)| = O_p(\lambda_n)$$

By Assumption 3.2(b), we have $\lambda_n \sup_{h \in \mathcal{H}_{k(n)}} |\hat{P}(h) - P(h)| = O_p(\lambda_n)$ and $\lambda_n |P(\Pi_n h_0) - P(h_0)| = O(\lambda_n)$. Thus, with $\max\{\eta_0, E[\|m(X, \Pi_n h_0)\|_W^2]\} = O(\lambda_n)$, for all $B_T(h_0)$,

$$\Pr(\hat{h}_n \notin B_T(h_0), \hat{h}_n \in \mathcal{H}_{k(n)}^{M_0})$$

$$\leq \Pr\left(\inf_{h \in \mathcal{H}_{k(n)}, h \notin B_T(h_0)} \left\{ cE[\|m(X, h)\|_W^2] + \lambda_n P(h) \right\} \right)$$

$$\leq \lambda_n \sup_{h \in \mathcal{H}_{k(n)}} |\hat{P}(h) - P(h)| = O_p(\lambda_n)$$

By Assumption 3.2(c), we have $\lambda_n \sup_{h \in \mathcal{H}_{k(n)}} |\hat{P}(h) - P(h)| = O(\lambda_n)$ and $\lambda_n |P(\Pi_n h_0) - P(h_0)| = o(\lambda_n)$ for $\lambda_n > 0$. Thus, with $\max\{\eta_0, E[\|m(X, \Pi_n h_0)\|_W^2]\} = O(\lambda_n)$, for all $B_T(h_0)$,
\[ \Pi_n(h_0) = o(\lambda_n), \text{ for all } B_T(h_0), \]

\[ \Pr(\hat{h}_n \notin B_T(h_0), \hat{h}_n \in \mathcal{H}_{k(n)}^{M_0}) \leq \Pr \left( \inf_{h \in \mathcal{H}_{k(n)}^{M_0}, h \notin B_T(h_0)} \left\{ cE \left[ \|m(X, h)\|_W^2 \right] + \lambda_n P(h) \right\} \right) \]

\[ \leq O_p(\delta_{m,n}^3) + \lambda_n P(h_0) + o_p(\lambda_n). \]

Hence we obtain results (i) and (ii). \[ \text{Q.E.D.} \]

**D. CONVERGENCE RATE: PROOFS OF THEOREMS**

The proof of Theorem 4.1 directly follows from Lemma B.1 and the definition of \( \omega_n(\delta, \mathcal{H}_{os}) \). The proof of Corollary 5.1 directly follows from Theorem 4.1 and Lemma B.2. The proof of Corollary 5.2 directly follows from Theorem 4.1 and Lemmas B.2 and B.3.

**PROOF OF COROLLARY 5.3:** By Theorem 4.1, Lemmas B.2 and B.3(ii), results of Corollary 5.2 are obviously true. We now specialize Corollary 5.2 to the PSMD estimator using a series LS estimator \( \hat{m}(X, h) \). For this case, we have \( \delta_{m,n} = \frac{\mu_n}{n} \times b_{m,J_n}^2 \).

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 \( h \in \mathcal{H}_{os} \),

\[ c_2 E[m(X, h) W(X) m(X, h)] \leq \|h - h_0\|^2 \]

\[ \leq \text{const.} \sum_{j=1}^{\infty} \langle \varphi(v_j^{-2}) \rangle \|h - h_0, q_j\|^2. \]

On the other hand, the choice of penalty and the definition of \( \mathcal{H}_{os} \) imply that \( \sum_j \nu_j^{2\alpha} \langle h - h_0, q_j \rangle_s^2 \leq \text{const.} \) for all \( h \in \mathcal{H}_{os} \). Denote \( \eta_j = \langle \varphi(v_j^{-2}) \rangle \|h - h_0, q_j\|^2 \). Then \( \sum_j \nu_j^{2\alpha} \langle \varphi(v_j^{-2}) \rangle^{-1} \eta_j \leq M \). Therefore, the class \( \{ g \in L^2(X, \| \cdot \|_{L^2_s(X)}), g(h) = \sqrt{W(h)} m(\cdot, h), h \in \mathcal{H}_{os} \} \) is embedded in the ellipsoid \( \{ g \in L^2(X, \| \cdot \|_{L^2_s(X)}), \|g\|_{L^2_s(X)}^2 = \sum_j \eta_j, and \sum_j \nu_j^{2\alpha} \langle \varphi(v_j^{-2}) \rangle^{-1} \eta_j \leq M' \} \) for some finite constant \( M' \). By invoking the results of Yang and Barron (1999), it follows that the \( J_n \)th approximation error rate of this ellipsoid satisfies \( b_{m,J_n}^2 \leq \text{const.} \nu_j^{-2\alpha} \langle \varphi(v_j^{-2}) \rangle \). Hence \( \delta_{m,n} = \frac{\mu_n}{n} \times b_{m,J_n}^2 \leq \text{const.} \nu_j^{-2\alpha} \langle \varphi(v_j^{-2}) \rangle \) and

\[ \|\hat{h} - h_0\|_s = O_p(\nu_j^{-\alpha}) = O_p(\frac{\mu_n}{n} \langle \varphi(v_j^{-2}) \rangle^{-1}). \]

\[ \text{Q.E.D.} \]
E. CONVERGENCE RATE: PROOFS OF LEMMAS

PROOF OF LEMMA B.1: Let $r_n^2 = \max\{\delta_{m,n}^2, \lambda_n \delta_{P,n}, \|\Pi_nh_0 - h_0\|^2, \lambda_n|P(\Pi_nh_0) - P(\hat{h}_n)|\} = o_p(1)$. Since $\hat{h}_n \in \mathcal{H}_{osn}$ with probability approaching 1, we have, for all $M > 1$,

$$\Pr\left(\frac{\|\hat{h}_n - h_0\|}{r_n} \geq M\right)
\leq \Pr\left(\inf_{h \in \mathcal{H}_{osn} : \|h - h_0\| \geq Mr_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, h)\|^2_{\Pi} + \lambda_n \hat{P}_n(h) \right\} \right)
\leq \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, \Pi_nh_0)\|^2_{\Pi} + \lambda_n \hat{P}_n(\Pi_nh_0) + O_p(\eta_n)
\leq \Pr\left(\inf_{h \in \mathcal{H}_{osn} : \|h - h_0\| \geq Mr_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, h)\|^2_{\Pi} + \lambda_n P(h) \right\} \right)
\leq \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, \Pi_nh_0)\|^2_{\Pi} + \lambda_n \hat{P}_n(\Pi_nh_0) + 2\lambda_n \delta_{P,n} + O_p(\eta_n)
,$$

where the last inequality is due to $\sup_{h \in \mathcal{H}_{osn}} |\hat{P}_n(h) - P(h)| = O_p(\delta_{P,n}) = o_p(1)$.

By Assumption 3.3 with $\eta_{0,n} = O(\delta_{m,n}^2)$ and $\eta_n = O(\eta_{0,n})$, and definitions of $\mathcal{H}_{osn}$ and $\delta_{m,n}^2$, there are two finite constants $c, c_0 > 0$ such that

\begin{equation}
SM.4) \quad cE(\|m(X, \hat{h}_n)\|_W^2) + \lambda_n P(\hat{h}_n)
\leq O_p(\delta_{m,n}^2 + \lambda_n \delta_{P,n}) + c_0 E(\|m(X, \Pi_nh_0)\|_W^2) + \lambda_n \hat{P}_n(\Pi_nh_0),
\end{equation}

which implies

$$cE(\|m(X, \hat{h}_n)\|_W^2) \leq O_p(\delta_{m,n}^2 + \lambda_n \delta_{P,n}) + c_0 E(\|m(X, \Pi_nh_0)\|_W^2) + \lambda_n |P(\Pi_nh_0) - P(\hat{h}_n)|.$$

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

$$\Pr\left(\frac{\|\hat{h}_n - h_0\|}{r_n} \geq M\right)
\leq \Pr\left(M^2r_n^2 \right)
\leq O_p\left(\max\{\delta_{m,n}^2, \lambda_n \delta_{P,n}, \|\Pi_nh_0 - h_0\|^2, \lambda_n |P(\Pi_nh_0) - P(\hat{h}_n)|\}\right)$$.  


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

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

Under Assumption 3.2(c), $\lambda_n^2|P(\Pi_n h_0) - P(\hat{h}_n)| = o_p(\lambda_n)$ and $\delta_{P,n} = o_p(1)$; hence result (ii) follows.

For result (iii), using the same argument as that for results (i) and (ii), inequality (SM.4) still holds. By condition (iii) of Theorem 4.1, $\lambda_n(P(\hat{h}_n) - P(\Pi_n h_0)) \geq \lambda_n\langle t_0, \hat{h}_n - \Pi_n h_0 \rangle_{\mathcal{H}}$. Thus

$$cE(\|m(X, \hat{h}_n)\|_W^2) + \lambda_n\langle t_0, \hat{h}_n - \Pi_n h_0 \rangle_{\mathcal{H}}^2 \leq O_p(\delta_{m,n}^2 + \lambda_n\delta_{P,n} + c_0E(\|m(X, \Pi_n h_0)\|_W^2));$$

hence

$$cE(\|m(X, \hat{h}_n)\|_W^2) \leq O_p(\delta_{m,n}^2 + \lambda_n\delta_{P,n} + c_0E(\|m(X, \Pi_n h_0)\|_W^2)) + \text{const.} \lambda_n\|\hat{h}_n - \Pi_n h_0\|_s.$$

By Assumption 4.1, Lemma B.1(iii) follows by choosing $r_n^2 = \max\{\delta_{m,n}^2, \lambda_n\delta_{P,n}, \|\Pi_n h_0 - h_0\|_s, \lambda_n\|\hat{h}_n - \Pi_n h_0\|_s\} = o_p(1)$.

**Q.E.D.**

**Proof of Lemma B.2:** To simplify notation, we denote $b_j = \varphi(\nu_j^{-2})$. Result (i) 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

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

where the last inequality is due to Assumption 5.2(i) and $\{b_j\}$ nonincreasing. Similarly, Assumption 5.2(ii) implies result (ii) since

$$c_2\|h_0 - \Pi_n h_0\|_s^2 \geq \sum_{j > k(n)} |\langle h_0 - \Pi_n h_0, q_j \rangle_s|^2 \geq c\left( \min_{j > k(n)} b_j^{-1} \right) \sum_{j > k(n)} b_j|\langle h_0 - \Pi_n h_0, q_j \rangle_s|^2 \geq \frac{c'}{b_{k(n)}}\|h_0 - \Pi_n h_0\|^2.$$
for some finite positive constants $c_2$, $c$, and $c'$. Result (iii) directly follows from results (i) and (ii).

PROOF OF LEMMA B.3: Denote $b_j = \varphi(\nu_j^{-2})$. For any $h \in \mathcal{H}_{os}$ with $\|h\|_s^2 \leq O(\delta^2)$ and for any $k \geq 1$, Assumptions 5.3 and 5.4(i) imply that there are finite positive constants $c_1$ and $c_2$ such that

$$c_1\|h\|_s^2 \leq \sum_{j\leq k}(h, q_j)_s^2 + \sum_{j> k}(h, q_j)_s^2 \leq \left(\max_{j \leq k} b_j^{-1}\right) \sum_{j} b_j(h, q_j)_s^2 + M^2(\nu_{k+1})^{-2a} \leq \frac{1}{c} b_k^{-1}\delta^2 + M^2(\nu_{k+1})^{-2a}.$$ 

Given that $M > 0$ is a fixed finite number and $\delta$ is small, we can assume $M^2(\nu_{k+1})^{-2a} > \frac{1}{c}\delta^2/b_1$. Since $\{b_j\}$ is nonincreasing and $\{\nu_j\}_{j=1}^\infty$ is strictly increasing in $j \geq 1$, we have that there is a $k^* \equiv k^*(\delta) \in (1, \infty)$ such that

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

and

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

thus result (i) holds. Result (ii) follows from Lemma B.2 and result (i).

Q.E.D.

F. PROOFS OF LEMMAS FOR SERIES LS ESTIMATION OF $m(\cdot)$

Denote $\tilde{m}(X, h) \equiv \rho^T(X)(P'P)^{-1}P'm(h)$ and $m(h) = (m(X_1, h), \ldots, m(X_n, h))^T$.

LEMMA SM.1: Let Assumptions C.1 and C.2(i) hold. Then there are finite constants $c, c' > 0$ such that, w.p.a.1,

$$cEX\left[\|\tilde{m}(X, h)\|_W^2\right] \leq \frac{1}{n} \sum_{i=1}^n \|\tilde{m}(X_i, h)\|_{W_i}^2 \leq c'EX\left[\|\tilde{m}(X, h)\|_W^2\right]$$

uniformly in $h \in \mathcal{H}_{k(\alpha)}^{M_0}$.
**Proof:** Denote \((g, \overline{g})_{n,X} \equiv \frac{1}{n} \sum_{i=1}^{n} g(X_i)\overline{g}(X_i)\) and \(\langle g, \overline{g} \rangle \equiv E_X[g(X) \times \overline{g}(X)]\), where \(g(X)\) and \(\overline{g}(X)\) are square integrable functions of \(X\). We want to show that for all \(t > 0\),

\[
(\text{SM}.5) \quad \lim_{n \to \infty} \text{Pr} \left( \sup_{h \in \mathcal{H}_{k(n)}^p} \left| \frac{\langle \tilde{m}(\cdot, h), \tilde{m}(\cdot, h) \rangle_{n,X} - \langle \tilde{m}(\cdot, h), \tilde{m}(\cdot, h) \rangle_X}{\langle \tilde{m}(\cdot, h), \tilde{m}(\cdot, h) \rangle_X} \right| > t \right) = 0.
\]

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

\[
\sup_{h \in \mathcal{H}_{k(n)}^p} \left| \frac{\langle \tilde{m}(\cdot, h), \tilde{m}(\cdot, h) \rangle_{n,X} - \langle \tilde{m}(\cdot, h), \tilde{m}(\cdot, h) \rangle_X}{\langle \tilde{m}(\cdot, h), \tilde{m}(\cdot, h) \rangle_X} \right| \leq \sup_{g \in G_n} \left| \langle g, g \rangle_{n,X} - \langle g, g \rangle_X \right|.
\]

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)–(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 \(h \in \mathcal{H}_{k(n)}^M\),

\[
0.5E_X[\|\tilde{m}(X, h)\|^2] \leq \frac{1}{n} \sum_{i=1}^{n} \|\tilde{m}(X_i, h)\|^2 \leq 2E_X[\|\tilde{m}(X, h)\|^2]
\]

except for an event w.p.a.0. By Assumption C.1(v), there are finite constants \(K, K' > 0\) such that \(K'I \leq W(X) \leq K'I\) for almost all \(X\). Thus, \(K\|\tilde{m}(X, h)\|^2 \leq \|\tilde{m}(X, h)\|^2_{\overline{W}} \leq K'\|\tilde{m}(X, h)\|^2_{I}\) for almost all \(X\). Also by Assumption C.1(v), uniformly over \(h \in \mathcal{H}_{k(n)}^M\),

\[
\|\tilde{m}(X, h)\|^2_{\overline{W}} \leq \sup_{x \in \mathcal{X}} |\hat{W}(x) - W(x)| \times \|\tilde{m}(X, h)\|^2_{I} + \|\tilde{m}(X, h)\|^2_{\overline{W}} \leq (K' + o_p(1))\|\tilde{m}(X, h)\|^2_{I}.
\]

Similarly,

\[
\|\tilde{m}(X, h)\|^2_{\overline{W}} \geq (K - o_p(1))\|\tilde{m}(X, h)\|^2_{I}.
\]
Note that for \( n \) large, \( \min\{K', K\} \pm o_p(1) > 0 \). Therefore, uniformly over \( h \in \mathcal{H}_{k(n)}^M \),

\[
\text{const.} \times E_X\left[\|\hat{m}(X, h)\|_W^2\right] \leq \frac{1}{n} \sum_{i=1}^{n} \|\tilde{m}(X_i, h)\|_W^2
\]

\[
\leq \text{const.'} \times E_X\left[\|\hat{m}(X, h)\|_W^2\right]
\]

except for a set w.p.a.0. \( \square \)

\textbf{PROOF OF LEMMA C.1:} By Assumption C.1(i) and (v) it suffices to establish the results for \( W = I \). Result (i) directly follows from our Assumption C.1 and Lemma A.1 Part (C) of Ai and Chen (2003).

Result (iii) can be established in the same way as that of result (ii). For result (ii), let \( \varepsilon(Z, h) \equiv \rho(Z, h) - m(X, h) \) and \( \varepsilon(h) \equiv (\varepsilon(Z_1, h), \ldots, \varepsilon(Z_n, h))' \).

For any symmetric and positive matrix \( \Omega(d \times d) \), we have the spectral decomposition \( \Omega = U\Lambda U' \), where \( \Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_d\} \) with \( \lambda_i > 0 \) and \( \Lambda = I_d \).

Denote \( \lambda_{\min}(\Omega) \) as the smallest eigenvalue of the matrix \( \Omega \). By definition, we have

\[
\sup_{h \in \mathcal{H}_{k(n)}^M} \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(X_i, h) - \tilde{m}(X_i, h)\|_I^2
\]

\[
= \sup_{h \in \mathcal{H}_{k(n)}^M} \frac{1}{n} \sum_{i=1}^{n} \text{Tr}\{p^{hn}(X_i)'(P'P)^{-1}P'\varepsilon(h)\varepsilon(h)'P(P'P)^{-1}p^{hn}(X_i)\}
\]

\[
= \sup_{h \in \mathcal{H}_{k(n)}^M} \frac{1}{n} \sum_{i=1}^{n} \text{Tr}\{\varepsilon(h)'P(P'P)^{-1}p^{hn}(X_i)p^{hn}(X_i)'(P'P)^{-1}P'\varepsilon(h)\}
\]

\[
= \sup_{h \in \mathcal{H}_{k(n)}^M} \frac{1}{n} \text{Tr}\left\{\varepsilon(h)'P(P'P)^{-1}\sum_{i=1}^{n} p^{hn}(X_i)p^{hn}(X_i)'(P'P)^{-1}P'\varepsilon(h)\right\}
\]

\[
= \sup_{h \in \mathcal{H}_{k(n)}^M} \frac{1}{n} \text{Tr}\{\varepsilon(h)'P(P'P)^{-1}P'\varepsilon(h)\}
\]

\[
= \sup_{h \in \mathcal{H}_{k(n)}^M} \frac{1}{n} \text{Tr}\{\varepsilon(h)'P(P'P/n)^{-1}P'\varepsilon(h)\}
\]

\[
\leq (\lambda_{\min}(P'P/n))^{-1} \times \sup_{h \in \mathcal{H}_{k(n)}^M} \frac{1}{n^2} \text{Tr}\{\varepsilon(h)'PP'\varepsilon(h)\}.
\]
Note that
\[ \epsilon(h)' P P' \epsilon(h) = \sum_{j=1}^{J_n} \left( \sum_{i=1}^{n} p_j(X_i) \epsilon(Z_i, h) \right)^2. \]

Let \( r_n = \frac{j_n}{n} C_n \). We have, for all \( M \geq 1 \),
\[
\Pr \left( \sup_{h \in \mathcal{H}_{k(n)}^{M_0}} \frac{1}{n} \sum_{i=1}^{n} \| \hat{m}(X_i, h) - \tilde{m}(X_i, h) \|_2^2 > M r_n \right) \\
\leq \Pr \left( (\lambda_{\min}(PP/n))^{-1} X \sup_{h \in \mathcal{H}_{k(n)}^{M_0}} \sum_{j=1}^{J_n} \left( \frac{1}{n} \sum_{i=1}^{n} p_j(X_i) \epsilon(Z_i, h) \right)^2 > M r_n \right) \\
\leq \Pr \left( (\lambda_{\min}(PP/n))^{-1} X J_n \max_{1 \leq j \leq J_n} \frac{1}{n} \sum_{i=1}^{n} p_j(X_i) \epsilon(Z_i, h) \right)^2 > M r_n \right). \\
\]
Following Newey (1997, p. 162) and under Assumption C.1(i)–(iv), we have: 
\( (\lambda_{\min}(PP/n))^{-1} = O_p(1) \). Thus, to bound \( \Pr(\sup_{h \in \mathcal{H}_{k(n)}^{M_0}} n^{-1} \sum_{i=1}^{n} \| \hat{m}(X_i, h) - \tilde{m}(X_i, h) \|_2^2 > M r_n) \), it suffices to bound the probability 
\[
\Pr \left( \sum_{j=1}^{J_n} \left( \sup_{h \in \mathcal{H}_{k(n)}^{M_0}} \frac{1}{n} \sum_{i=1}^{n} p_j(X_i) \epsilon(Z_i, h) \right)^2 > M r_n \right) \\
\leq \frac{1}{M r_n} E_{Z_n} \left[ \sum_{j=1}^{J_n} \left( \sup_{h \in \mathcal{H}_{k(n)}^{M_0}} \frac{1}{n} \sum_{i=1}^{n} p_j(X_i) \epsilon(Z_i, h) \right)^2 \right] \\
\leq \frac{J_n}{n r_n M} \max_{1 \leq j \leq J_n} E_{Z_n} \left[ \left( \sup_{h \in \mathcal{H}_{k(n)}^{M_0}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} p_j(X_i) \epsilon(Z_i, h) \right)^2 \right],
\]
where the first inequality is by Markov inequality and \( E_{Z_n}(\cdot) \) denotes the expectation with respect to \( Z_n \equiv (Z_1, \ldots, Z_n) \). By Theorem 2.14.5 in Van der Vaart
and Wellner (1996) (VdV-W; also see Pollard (1990)), we have

\[
\max_{1 \leq j \leq J_n} E_{Z_n} \left[ \left( \sup_{h \in \mathcal{H}_{K(n)}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} p_j(X_i) \varepsilon(Z_i, h) \right)^2 \right]
\leq \max_{1 \leq j \leq J_n} \left( E_{Z_n} \left[ \sup_{h \in \mathcal{H}_{K(n)}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} p_j(X_i) \varepsilon(Z_i, h) \right] \right.

\qquad + \left. \sqrt{E[|p_j(X)\tilde{\rho}_n(Z)|^2]} \right)^2.
\]

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

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

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

\[
\max_{1 \leq j \leq J_n} E_{Z_n} \left[ \sup_{h \in \mathcal{H}_{K(n)}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} p_j(X_i) \varepsilon(Z_i, h) \right]
\leq \max_{1 \leq j \leq J_n} \left\{ \sqrt{E[|p_j(X)\tilde{\rho}_n(Z)|^2]} \right. 

\qquad \times \int_{0}^{1} \sqrt{1 + \log N_1(wK, \mathcal{E}_{jn}, \| \cdot \|_{L^2(f_Z)})} \, dw \right\}
\leq K \max_{1 \leq j \leq J_n} \int_{0}^{1} \sqrt{1 + \log N_1(wK, \mathcal{E}_{jn}, \| \cdot \|_{L^2(f_Z)})} \, dw,
\]

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)||\varepsilon(Z, h) - \varepsilon(Z, h')| 

\qquad + |E[\varepsilon(Z, h)|X] - E[\varepsilon(Z, h'|X)]| \}
\]

and

\[
|p_j(X)||E[\varepsilon(Z, h)|X] - E[\varepsilon(Z, h'|X)]| \leq |p_j(X)|E[|\varepsilon(Z, h) - \varepsilon(Z, h')||X].
\]
Recall that \( \mathcal{O}_{jn} \equiv \{ p_j(\cdot) \varphi(\cdot, h) : h \in \mathcal{H}_{k(n)}^{M_0} \} \) and that
\[
\max_{1 \leq j \leq J_n} \int_0^1 \sqrt{1 + \log N_{\| \cdot \|_{L^2(fZ)}}(wK, \mathcal{O}_{jn})} \, dw \leq \sqrt{C_n} < \infty
\]
by Assumption C.2(iii). We have:
\[
\max_{1 \leq j \leq J_n} \int_0^1 \sqrt{1 + \log N_{\| \cdot \|_{L^2(fZ)}}(wK, \mathcal{E}_{jn})} \, dw \leq \text{const.} \times \sqrt{C_n} < \infty
\]
and hence
\[
\max_{1 \leq j \leq J_n} \mathbb{E}_{Z^n} \left[ \left( \sup_{h \in \mathcal{H}_{k(n)}^{M_0}} \frac{1}{\sqrt{n}} \sum_{i=1}^n p_j(X_i) \varphi(Z_i, h) \right)^2 \right] \leq \text{const.} \times \sqrt{C_n}.
\]
It then follows that
\[
\frac{J_n}{nr_nM} \max_{1 \leq j \leq J_n} \mathbb{E}_{Z^n} \left[ \left( \sup_{h \in \mathcal{H}_{k(n)}^{M_0}} \frac{1}{\sqrt{n}} \sum_{i=1}^n p_j(X_i) \varphi(Z_i, h) \right)^2 \right] \leq \text{const.} \times \frac{J_n C_n}{nr_nM},
\]
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 (i) and (iii) are the same as that of result (ii). For result (ii), by the fact \((a - b)^2 + b^2 \geq \frac{1}{2} a^2\), we have that uniformly over \( h \in \mathcal{H}_{k(n)}^{M_0} \),
\[
\frac{1}{n} \sum_{i=1}^n \| \tilde{m}(X_i, h) \|_{\hat{W}}^2 \geq \frac{1}{2} \frac{1}{n} \sum_{i=1}^n \| \tilde{m}(X_i, h) \|_{\hat{W}}^2
\]
\[
- \frac{1}{n} \sum_{i=1}^n \| \tilde{m}(X_i, h) - \tilde{m}(X_i, h) \|_{\hat{W}}^2.
\]
By Lemma SM.1, there is a finite constant \( c > 0 \) such that, w.p.a.1 and uniformly over \( h \in \mathcal{H}_{k(n)}^{M_0} \),
\[
\frac{1}{n} \sum_{i=1}^n \| \tilde{m}(X_i, h) \|_{\hat{W}}^2 \geq \frac{c}{2} \mathbb{E}_X \left[ \| \tilde{m}(X, h) \|_{\hat{W}}^2 \right] - \frac{1}{n} \sum_{i=1}^n \| \tilde{m}(X_i, h) - \tilde{m}(X_i, h) \|_{\hat{W}}^2.
\]
\[
\frac{c}{4} E_X \left[ \| m(X, h) \|_W^2 \right] - \left( \frac{c}{2} E_X \left[ \| m(X, h) - \tilde{m}(X, h) \|_W^2 \right] \right) \\
+ \frac{1}{n} \sum_{i=1}^{n} \| \tilde{m}(X_i, h) - \tilde{m}(X_i, h) \|_W^2 \\
\geq K E_X \left[ \| m(X, h) \|_W^2 \right] - O_p \left( \frac{b_{m, J_n}^2 + J_n C_n}{n} \right),
\]

where the second inequality is due to the fact that \((a - b)^2 + b^2 \geq \frac{1}{2} a^2\) and the last inequality is due to Lemma C.1, Assumption C.2(ii), and \(c_4 \equiv k_0 > 0\).

Similarly, by the fact \((a + b)^2 \leq 2a^2 + 2b^2\), we have that uniformly over \(h \in \mathcal{H}_{k(n)}\),
\[
\frac{1}{n} \sum_{i=1}^{n} \| \tilde{m}(X_i, h) \|_W^2 \leq 2 \frac{1}{n} \sum_{i=1}^{n} \| \tilde{m}(X_i, h) \|_W^2 \\
+ 2 \frac{1}{n} \sum_{i=1}^{n} \| \tilde{m}(X_i, h) - \tilde{m}(X_i, h) \|_W^2.
\]

By Lemma SM.1, there is a finite constant \(c' > 0\) such that, w.p.a.1 and uniformly over \(h \in \mathcal{H}_{k(n)}\),
\[
\frac{1}{n} \sum_{i=1}^{n} \| \tilde{m}(X_i, h) \|_W^2 \\
\leq 2c' E_X \left[ \| \tilde{m}(X, h) \|_W^2 \right] + 2 \frac{1}{n} \sum_{i=1}^{n} \| \tilde{m}(X_i, h) - \tilde{m}(X_i, h) \|_W^2 \\
\leq 4c' E_X \left[ \| m(X, h) \|_W^2 \right] + \left( 4c' E_X \left[ \| \tilde{m}(X, h) - m(X, h) \|_W^2 \right] \right) \\
+ \frac{2}{n} \sum_{i=1}^{n} \| \tilde{m}(X_i, h) - \tilde{m}(X_i, h) \|_W^2 \\
\leq K' E_X \left[ \| m(X, h) \|_W^2 \right] + O_p \left( \frac{b_{m, J_n}^2 + J_n C_n}{n} \right),
\]

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 \(4c' \equiv K' < \infty\).

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

\[
\sup_{h \in \mathcal{N}_{osn}} \frac{1}{n} \sum_{i=1}^{n} \| \hat{m}(X_i, h) - \hat{m}(X_i, h_0) - \tilde{m}(X_i, h) \|_2^2
\]

\[
= \sup_{h \in \mathcal{N}_{osn}} \frac{1}{n^2} \text{Tr}\left\{[\varepsilon(h) - \varepsilon(h_0)] P (P' P/n)^{-1} P' [\varepsilon(h) - \varepsilon(h_0)]\right\}
\]

\[
\leq (\lambda_{\min}(P' P/n))^{-1}
\]

\[
\times \sup_{h \in \mathcal{N}_{osn}} \frac{1}{n^2} \text{Tr}\left\{[\varepsilon(h) - \varepsilon(h_0)] PP' [\varepsilon(h) - \varepsilon(h_0)]\right\}
\]

\[
= (\lambda_{\min}(P' P/n))^{-1}
\]

\[
\times \sup_{h \in \mathcal{N}_{osn}} \frac{1}{n^2} \sum_{j=1}^{J_n} \left( \left| \sum_{i=1}^{n} p_j(X_i) [\varepsilon(Z_i, h) - \varepsilon(Z_i, h_0)] \right| \right)^2.
\]

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

\[
\Pr\left( \sup_{h \in \mathcal{N}_{osn}} \frac{1}{n} \sum_{i=1}^{n} \| \hat{m}(X_i, h) - \hat{m}(X_i, h_0) - \tilde{m}(X_i, h) \|_2^2 > Mr_n \right),
\]

it suffices to bound the probability

\[
\Pr\left( \sum_{j=1}^{J_n} \left( \sup_{h \in \mathcal{N}_{osn}} \left| \frac{1}{n} \sum_{i=1}^{n} p_j(X_i) [\varepsilon(Z_i, h) - \varepsilon(Z_i, h_0)] \right| \right)^2 > Mr_n \right)
\]

\[
\leq \frac{J_n}{nr_n M}
\]

\[
\times \max_{1 \leq j \leq J_n} E_{Z_n} \left[ \left( \sup_{h \in \mathcal{N}_{osn}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} p_j(X_i) \Delta \varepsilon(Z_i, h) \right| \right)^2 \right].
\]

Let \( \Delta \varepsilon(Z_i, h) \equiv \varepsilon(Z_i, h) - \varepsilon(Z_i, h_0) \). By Theorem 2.14.5 in VdV-W, we have

\[
\max_{1 \leq j \leq J_n} E_{Z_n} \left[ \left( \sup_{h \in \mathcal{N}_{osn}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} p_j(X_i) \Delta \varepsilon(Z_i, h) \right| \right)^2 \right]
\]

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

By Jensen’s inequality,

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

Hence

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

by condition (i) in Lemma C.3.

By Theorem 2.14.2 in VdV-W, Remark C.1, and conditions (i) and (ii) of Lemma C.3, we have (up to some constant)

\[ \max_{1 \leq j \leq n} E_{Z^n} \left[ \sup_{h \in \mathcal{N}_{non}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} p_j(X_i) \Delta \epsilon(Z_i, h) \right] \leq (\delta_{s,n})^\kappa \int_{0}^{1} \left( 1 + \log N_{\|w\|} \right) (\delta_{s,n})^\kappa, \]

\[ \{p_j(\cdot)\Delta \epsilon(\cdot, h) : h \in \mathcal{N}_{non}, \| \cdot \|_{L^2(f_Z)} \}^{1/2} dw \]

\[ \leq (\delta_{s,n})^\kappa \int_{0}^{1} \sqrt{1 + \log N(w^{1/\kappa}, \mathcal{N}_{non}, \| \cdot \|_s)} \, dw \leq \text{const.} \times (\delta_{s,n})^\kappa. \]

Hence

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

The desired result follows. Q.E.D.
G. 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 $\hat{W} = W = I$.

First, Assumption 3.1(i) is trivially satisfied with $W = I$. For any $h \in \mathcal{H}$, we denote $h(y_1, y_2) = h_1(y_1) + h_2(y_2)$, $\Delta h(y_1, y_2) = h(y_1, y_2) - h_0(y_1, y_2) = \Delta h_1(y_1) + \Delta h_2(y_2)$, and $\Delta h_l(y_l) = h_l(y_l) - h_{0l}(y_l)$ 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

\[(\text{SM.6}) \quad m(X, h) - m(X, h_0) = E[F_{Y|Y_1,Y_2}(h_1(Y_1) + h_2(Y_2))] - F_{Y|Y_1,Y_2}(h_0(Y_1) + h_{02}(Y_2))|X] \]
\[= E\left[\left\{\int_0^1 f_{Y|Y_1,Y_2}(h_0(Y_1, Y_2) + t\Delta h(Y_1, Y_2)) \, dt\right\} \right. \]
\[\left. \times [\Delta h_1(Y_1) + \Delta h_2(Y_2)]|X \right]\]
\[= K_{1,h}[\Delta h_1](X) + K_{2,h}[\Delta h_2](X). \]

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 $\mathcal{H}$ and $\mathcal{H}_n$ (Condition 6.2(i) and (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 $h_0 \in \mathcal{H}$, that there is $\Pi_n h_0 \in \mathcal{H}_n$ such that

$$\|h_0 - \Pi_n h_0\|_s = \|h_0 - \Pi_n h_0\|_{sup} \leq c_k(n)^{-r_1} + c^\prime_k(n)^{-r_2} = o(1), \quad \text{with} \quad r_i = \alpha_i/d,$$

thus Assumption 3.1(iii) holds. For any $h \in \mathcal{H}$ with $\Delta h(y_1, y_2) = \Delta h_1(y_1) + \Delta h_2(y_2)$ and $\Delta h_l(y_l) = h_l(y_l) - h_{0l}(y_l)$, $l = 1, 2$, equation (SM.6) implies that

$$|m(X, h) - m(X, h_0)| \leq E\left[\sup_{t \in [0, 1]} f_{Y|Y_1,Y_2}(h_0(Y_1, Y_2) + t\Delta h(Y_1, Y_2))|X\right] \times \left[\sup_{y_1} |\Delta h_1(y_1)| + \sup_{y_2} |\Delta h_2(y_2)|\right].$$
Since \( m(X, h_0) = 0 \) and by Condition 6.1(iv), we have
\[
E[|m(X, h)|^2] = E[|m(X, h) - m(X, h_0)|^2] \\
\leq E\left[\left(\sup_{t \in [0, 1]} f_{Y_1,Y_2,X}(h_0(Y_1, Y_2) + t\Delta h(Y_1, Y_2))|X\right)^2\right] \\
\times (\|h - h_0\|_s)^2 \\
\leq \text{const.} \times (\|h - h_0\|_s)^2.
\]
This and \( \|\Pi_nh_0 - h_0\|_s = o(1) \) imply
\[
E[|m(X, \Pi_nh_0)|^2] \leq \text{const.} \|\Pi_nh_0 - h_0\|_s^2 \\
\leq c[k_1(n)]^{-2r_1} + c'[k_2(n)]^{-2r_2} = o(1);
\]
hence Assumption 3.1(iv) holds. Assumption 3.2(b) directly follows from our choice of \( \hat{P}(\cdot) = P(\cdot) \).

Next, Condition 6.1(i) and (ii) and \( \hat{W} = W = I \) imply that Assumption C.1 holds. Assumption C.2(ii) follows trivially with \( \hat{\rho}_n(Z) \equiv 1 \) since \( \sup_{h, r} |\rho(Z, h)| \leq 1 \). Condition 6.1(ii) and (iii) implies that Assumption C.2(ii) holds with \( \beta_{m, J_n} = J_n^{-2m} \). Thus Lemma C.2 result (i) is applicable and Assumption 3.3(i) is satisfied with \( \eta_{0,n} = \frac{\beta_n}{n} + J_n^{-2m} \). This, \( E[(m(X, \Pi_nh_0))^2] = O(\max([k_1(n)]^{-2r_1}, [k_2(n)]^{-2r_2}), [k_1(n)]^{-2r_1}, [k_2(n)]^{-2r_2}, \frac{\beta_n}{n} + J_n^{-2m}] = O(\lambda_n) \) together imply that Lemma A.3(i) holds. Moreover, it follows by our choice of penalty that \( P(\Pi_nh_0) = O(1) \) and \( P(\hat{h}_n) = O_P(1) \). By our choice of sieves space it follows that \( \log N(u^{1/2}, \mathcal{H}^{\beta_{m, J_n}}_K, \|\cdot\|_{L^\infty}) \leq \min\{\frac{1}{2}k(n) \log(1/w), \text{const.}(1/w)^{d/2\alpha}\} \), where \( \alpha \equiv \min\{\alpha_1, \alpha_2\} > 0 \) (and const. can depend on \( M_0 \), but not \( n \)); see, for example, 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 that Assumption C.2(iii) is satisfied with either \( C_n \leq \text{const.} \times k(n) \) if \( \alpha \leq d \) or \( C_n = \text{const.} < \infty \) if \( \alpha > d \). By Lemma C.2 result (ii) and the fact that \( \frac{\lambda_{k(n)}}{n} = o(1) \), it follows that \( \hat{\delta}_{m,n}^2 = o(1) \) and hence Assumption 3.3(ii) holds.

By the mean value theorem and Condition 6.1(iv), we have, for all \( h, h' \in \mathcal{H}, \|h - h\|_s = \|h - h\|_{\sup} = \sup_{y_1} |h_1(y_1) - h'_1(y_1)| + \sup_{y_2} |h_2(y_2) - h'_2(y_2)| \). Then
\[
|m(X, h) - m(X, h')| \\
\leq E\left[\sup_{t \in [0, 1]} f_{Y_1,Y_2,X}(h'(Y_1, Y_2) + t(h - h')(Y_1, Y_2))|X\right]\|h - h'\|_s.
\]
This, Condition 6.1(iv), and \( \sup_{x \in X, h \in \mathcal{H}} |m(x, h)| \leq 1 \) imply that
\[
E[|m(X, h)|^2] - E[|m(X, h')|^2] \leq 2E[|m(X, h) - m(X, h')|] \\
\leq \text{const.} \times \|h - h'\|_s.
\]

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 \( \{ h \in \mathcal{H} : P(h) = \|h_1\|_{A_1} + \|h_2\|_{A_2} \leq M \} \) into \( \mathbf{H} \) is compact under the norm \( \| \cdot \|_s \); hence \( P(\cdot) \) is lower semicompact.

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

**REFERENCES**


Cowles Foundation for Research in Economics, Yale University, 30 Hillhouse, Box 208281, New Haven, CT 06520, U.S.A.; xiaohong.chen@yale.edu

and

Department of Economics, UC at Berkeley, 508-1 Evans Hall 3880, Berkeley, CA 94704-3880, U.S.A.; dpouzo@econ.berkeley.edu.

Manuscript received April, 2008; final revision received December, 2010.