Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions

Chunrong Ai; Xiaohong Chen


Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28200311%2971%3A3%3C1795%3AEEOMWC%3E2.0.CO%3B2-H

Econometrica is currently published by The Econometric Society.
EFFICIENT ESTIMATION OF MODELS WITH CONDITIONAL MOMENT RESTRICTIONS CONTAINING UNKNOWN FUNCTIONS

BY CHUNRONG AI AND XIAOHONG CHEN

We propose an estimation method for models of conditional moment restrictions, which contain finite dimensional unknown parameters ($\theta$) and infinite dimensional unknown functions ($h$). Our proposal is to approximate $h$ with a sieve and to estimate $\theta$ and the sieve parameters jointly by applying the method of minimum distance. We show that: (i) the sieve estimator of $h$ is consistent with a rate faster than $n^{-1/2}$ under certain metric; (ii) the estimator of $\theta$ is $\sqrt{n}$ consistent and asymptotically normally distributed; (iii) the estimator for the asymptotic covariance of the $\theta$ estimator is consistent and easy to compute; and (iv) the optimally weighted minimum distance estimator of $\theta$ attains the semiparametric efficiency bound. We illustrate our results with two examples: a partially linear regression with an endogenous nonparametric part, and a partially additive IV regression with a link function.

KEYWORDS: Semi-/nonparametric conditional moment restrictions, sieve minimum distance, continuous updating, endogeneity, semiparametric efficiency.

1. INTRODUCTION

A GENERAL FRAMEWORK for analyzing economic data ($Y$, $X$) is to presume that the data satisfy some conditional moment restrictions,

\begin{equation}
E[\rho(Z, \theta_o, h_o(\cdot))|X] = 0, 
\end{equation}

where $Z' = (Y', X'_z)$, $X_z$ is a subset of $X$, $\rho(\cdot)$ is a vector of known (residual) functions, and $E[\rho(Z, \theta_o, h_o)|X]$ is the conditional expectation of $\rho(Z, \theta_o, h_o)$ given $X$. The true conditional distribution of $Y$ given $X$ is assumed unknown, and the parameters of interest $\alpha_o \equiv (\theta_o, h_o)$ contain a vector of finite dimensional unknown parameters $\theta_o$ and possibly a vector of infinite dimensional unknown functions $h_o(\cdot) = (h_{o1}(\cdot), \ldots, h_{oq}(\cdot))$. Model (1) is semiparametric in the sense that it contains unknown functions $h_o$. Without the unknown functions $h_o$, model (1) is just the classical model of conditional moment restrictions

\begin{equation}
E[\rho(Z, \theta_o)|X] = 0, 
\end{equation}

We thank a co-editor, two anonymous referees, and W. Newey for their valuable suggestions that greatly improved the paper. We also received helpful comments from D. Andrews, R. Blundell, B. Hansen, J. Hahn, L. Hansen, H. Hong, B. Honore, K. Hyndman, R. Klein, O. Linton, J. Powell, P. Robinson, C. Sims, and seminar participants at Texas Austin, Rochester, Chicago, LSE, UCL, Carol III/Madrid, Duke, Rutgers, Princeton, Yale, Stanford, Ohio State, Pittsburgh, Wisconsin, Rice, and the 1999 North American Winter Meetings of the Econometric Society. Ai's research is supported by the Warrington College of Business Administration Summer Grants for 1998–2000 at the University of Florida, while Chen's research is supported by the Department of Economics research fund for 2000–2002 at the London School of Economics and ESRC/UK Grant R00023952.
which has been studied extensively in the econometrics literature. For instance, Hansen (1982) proposes a generalized method of moment (GMM) estimation of $\theta_o$ for model (2) and shows that his estimator is $\sqrt{n}$ consistent and asymptotically normally distributed. Chamberlain (1987) computes the efficiency bound of $\theta_o$ for model (2); Newey (1990a, 1993) proposes an estimator of $\theta_o$ that attains Chamberlain’s bound. In this paper, we shall extend the results obtained for model (2) to the more general semiparametric conditional moment restrictions given by (1). Our primary focus is the consistent estimation of $\alpha_o = (\theta_o, h_o)$ as well as the asymptotic distribution and efficiency of the estimator of $\theta_o$.

The inclusion of unknown functions $h_o(\cdot)$ enables model (1) to encompass many important classes of semiparametric and nonparametric models. For example, it includes the partially linear regression $\rho(Z, \alpha_o) = Y - X'_i\theta_o - h_o(X_2)$ studied by Robinson (1988) and the index regression $\rho(Z, \alpha_o) = Y - h_o(X'_i\theta_o)$ studied by Powell, Stock, and Stoker (1989) and Ichimura (1993). It also includes the transformation model $\rho(Z, \alpha_o) = h_o(Y) - X'_i\theta_o$ with monotone but unknown link $h_o$, and Chamberlain’s (1992) generalized semiparametric regression $E[\rho(Z, \theta_o, h_{o1}(\delta_1(Z)), \ldots, h_{oq}(\delta_q(Z))) | X] = 0$ with known $\delta_j(Z)$, $j = 1, \ldots, q$. All these examples are special cases of the class:

$$E[\rho(Z, \theta_o, h_{o1}(\delta_1(Z), \theta_o)), \ldots, h_{oq}(\delta_q(Z, \theta_o))) | X] = 0,$$

where $\delta_j(Z, \theta_o)$, $j = 1, \ldots, q$, are known measurable functions of $Z$ up to an unknown vector of parameters $\theta_o$. Model (1) encompasses all of the above examples, and many others. It permits dependence of $h_{oj}(\cdot)$ on other unknown functions. Nonparametric censored regressions are examples where an unknown function depends on another unknown function. It also permits dependence of $h_{oj}(\cdot)$ on unobserved variables. Random coefficients models are examples where the unknown density of the random coefficients is a function of unobserved random coefficients. Additional examples that fit into model (1) can be found in Powell (1994) and Horowitz (1998).

Under the assumption that model (1) identifies $\alpha_o = (\theta_o, h_o)$, we propose a sieve minimum distance (hereafter SMD) estimator $\hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n)$. The SMD estimator $\hat{\alpha}_n$, which can be interpreted as either a GMM or a two-stage nonlinear least squares (2SNLS) estimator, is both easy to implement and intuitive. Under a set of sufficient conditions, we show that $\hat{\alpha}_n$ converges to $\alpha_o$ at a rate faster than $n^{-1/4}$ under certain metric to be defined later, and that $\hat{\theta}_n$ is both $\sqrt{n}$ consistent and asymptotically normally distributed. We also provide a consistent and easy to compute estimator for the asymptotic covariance of $\hat{\theta}_n$. Furthermore, we compute the efficiency bound for model (1) and show that, when the conditional covariance matrix $\Sigma_o(X) \equiv \text{Var}[\rho(Z, \alpha_o) | X]$ is used as the weighting matrix, the corresponding estimator $\hat{\theta}_n$ (hereafter the optimally weighted SMD estimator) attains the efficiency bound. Our efficiency bound
result generalizes the work of Chamberlain (1992) to the more general framework of model (1).

The literature on the general theory of semiparametric efficient estimation is growing. The papers most closely related to ours are those by Wong and Severini (1991), Severini and Wong (1992), Shen (1997), Ai (1997), and Newey (1990a, 1993). The first four papers studied efficient estimation in the semiparametric maximum likelihood setting, while Newey investigated efficient estimation of the classical conditional moment restrictions model, which is summarized by (2). Our results extend all these studies to the more general setting given by (1). There are also studies on efficient estimation of particular semiparametric models.3 While our model (1) encompasses these prior studies’ models as special cases, they may be able to exploit the specific structures of their particular models to derive the $\sqrt{n}$ efficiency of their estimators for $\theta_o$ under a set of conditions that might be weaker than ours. We emphasize that our aim is to provide a general methodology for all semiparametric models of conditional moment restrictions given by (1). There is no published work prior to ours on $\sqrt{n}$ consistent and efficient estimation of $\theta_o$ when the unknown functions $h_o$ depend on the endogenous variables $Y$.

After the initial submission, an anonymous referee drew our attention to the work of Newey and Powell (2003) on a nonparametric instrumental variables model of the form $E[p(Z, h_o(\delta(Z)))|X] = 0$. Our approach is similar to theirs in that both use sieves to approximate the unknown functions $h_o$ and estimate the unknown parameters by applying the minimum distance procedure; however, we differ in our focus. While Newey and Powell focus on the identification and consistent estimation of the nonparametric components $h_o(\cdot)$, we focus on the $\sqrt{n}$ asymptotic normality and efficiency of the parametric components $\theta_o$. Thus our work complements theirs.

The rest of the paper is organized as follows. Section 2 formally introduces the SMD estimator $\hat{\alpha}_n$. Section 3 shows the consistency of $\hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n)$ and computes its convergence rate. Section 4 derives the $\sqrt{n}$ asymptotic normality of $\hat{\theta}_n$. Section 5 provides a consistent estimator for the asymptotic covariance of $\hat{\theta}_n$. Section 6 studies the efficiency property of $\hat{\theta}_n$. Throughout Sections 2–6, we illustrate the SMD procedure and the general theory with two examples: a partially linear regression with an endogenous nonparametric part $E[Y_1 - X_1'\theta_o - h_o(Y_2) | X_1, X_2] = 0$; and a partially additive IV regression with a link function $E[Y_1 - F(Y_2; \theta_o + \sum_{j=1}^{q} h_{o_j}(X_j)) | X_o, \ldots, X_q] = 0$. We note that the $\sqrt{n}$ normality and efficiency of $\hat{\theta}_n$ for these examples are new and useful additions to the literature. Section 7 evaluates the finite sample properties of

---

2There are also many papers in econometrics on general theory of $\sqrt{n}$-asymptotic normality of semiparametric estimators; see, e.g., Andrews (1994), Newey (1994), Pakes and Olley (1995), Chen and Shen (1998), and Chen, Linton, and van Keilegom (2003).

3See, for example, Robinson’s (1988) partially linear regression with homoskedastic error, and Klein and Spady’s (1993) binary choice model with independent error.
the SMD estimator via a small scale Monte Carlo study of the partially linear regression with an endogenous nonparametric part. Section 8 concludes the paper with suggestions for future research. All technical proofs are presented in the Appendices.

2. THE SMD ESTIMATOR

Suppose that the observations \((Y_i, X_i) : i = 1, 2, \ldots, n\) are drawn independently from the distribution of \((Y, X)\) with support \(\mathcal{Y} \times \mathcal{X}\), where \(\mathcal{Y}\) is a subset of \(\mathcal{R}^{d_y}\) and \(\mathcal{X}\) is a compact subset of \(\mathcal{R}^{d_x}\). Suppose that the unknown distribution of \((Y, X)\) satisfies the conditional moment restriction given by (1), where \(\rho : \mathcal{Z} \times \mathcal{A} \to \mathcal{R}^{d_y}\) is a known mapping, up to an unknown vector of parameters, \(\alpha_o \equiv (\theta_o, h_o) \in \mathcal{A} \equiv \Theta \times \mathcal{H}\). We assume that \(\Theta \subseteq \mathcal{R}^{d_y}\) is compact with nonempty interior and that \(\mathcal{H} \equiv \mathcal{H}_1 \times \cdots \times \mathcal{H}_q\) is a space of continuous functions. We further assume that \(\mathcal{Z} \equiv (Y', X')' \in \mathcal{Z} \equiv \mathcal{Y} \times \mathcal{X}'_z\) and \(\mathcal{X}_z \subseteq \mathcal{X}'\).

Under the assumption that model (1) identifies \(\alpha_o\), we propose to estimate \(\alpha_o\) by the methods of minimum distance and sieves. Heuristically, if the functional form of the conditional distribution of \(Y\) given \(X\), \(F_{Y|X}\), were known, then the functional form of the conditional mean function \(m(x, \alpha) = \int \rho(y, x, \alpha)dF_{Y|X}(y)\) would be known. The minimum distance estimation of \(\alpha_o\) would be appropriate since model (1) implies that \(\alpha_o\) minimizes:

\[
\inf_{\alpha=(\theta, h) \in \Theta \times \mathcal{H}} E\left[ m(X, \alpha) [\Sigma(X)]^{-1} m(X, \alpha) \right]
\]

where \(\Sigma(X)\) is a positive definite matrix for any given \(X\). The true value \(\alpha_o\) could then be estimated by minimizing the sample analog of (3).

However, this method has two deficiencies. First, the functional form of \(F_{Y|X}\), and therefore \(m(\cdot, \cdot)\) is, in fact, unknown; hence, this method is infeasible. Second, even if it were feasible, this method could yield an inconsistent estimator for \(\alpha_o\) or a consistent estimator that converges arbitrarily slowly when the function space \(\mathcal{H}\) is large. The first deficiency can be rectified by replacing \(m(X, \alpha)\) with a consistent nonparametric estimator \(\hat{m}(X, \alpha)\). To remedy the second deficiency, we follow the sieve literature (Grenander (1981)) by replacing \(\mathcal{H}\) with a sieve space \(\mathcal{H}_n = \mathcal{H}_n^1 \times \cdots \times \mathcal{H}_n^q\), which is a computable and often finite-dimensional compact parameter space that becomes dense in \(\mathcal{H}\) as \(n\) increases. To summarize, the SMD estimator of \(\alpha_o\) minimizes the sample analog of a nonparametric version of (3) with \(h\) restricted to the sieve space \(\mathcal{H}_n^1\):

\[
\hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n) : \min_{(\hat{\theta}, \hat{h}) \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n} \hat{m}(X, \alpha)' [\hat{\Sigma}(X)]^{-1} \hat{m}(X, \alpha),
\]

where \(\hat{\Sigma}(X)\) is a consistent estimator of \(\Sigma(X)\) that is introduced to address potential heteroscedasticity.
To compute the SMD estimator, we need a consistent estimator of \( m(X, \alpha) \). Here we present a linear sieve estimator.\(^{\text{4}}\) Let \( \{p_{j0}(X), j = 1, 2, \ldots\} \) denote a sequence of known basis functions (such as power series, splines, Fourier series, etc.), with the property that its linear combination can approximate any square integrable real-valued function of \( X \) well. Denote \( p^{k_n}(X) = (p_{01}(X), \ldots, p_{0k_n}(X))' \) and \( m(X, \alpha) = (m_1(X, \alpha), \ldots, m_{d_p}(X, \alpha))' \). Then for each arbitrarily fixed \( \alpha \) and \( l = 1, 2, \ldots, d_p \), \( m_l(X, \alpha) \) can be approximated by \( p^{k_n}(X)' \pi_l \) for some vector of coefficients \( \pi_l \) as \( k_n \to +\infty \). The linear sieve estimator for \( m_l(X, \alpha) \) is \( \hat{m}_l(X, \alpha) = p^{k_n}(X)' \hat{\pi}_l \), where \( \hat{\pi}_l \) is the ordinary least squares estimate obtained by regressing \( p_l(Z, \alpha) \) on \( p^{k_n}(X) \). To simplify presentation, we let \( p^{k_n}(X) \) be a tensor-product linear sieve basis in the main text. A tensor-product sieve is simply the product of univariate sieves. For example, let \( \{\phi_{i,j} : i = 1, \ldots, J_{j,n}\} \) denote a B-spline (Fourier series or power series) basis for \( L^2(X_j) \), the space of square Lebesgue integrable functions on \( X_j \), where \( X_j \) is a compact interval in \( \mathcal{R} \), \( 1 \leq j \leq d_s \). Then the tensor product \( \prod_{j=1}^{d_s} \phi_{i,j}(X_j) : i = 1, \ldots, J_{j,n}, j = 1, \ldots, d_s \) is a B-spline (or Fourier series or power series) basis for \( L^2(X) \), the space of square Lebesgue integrable functions on \( X = X_1 \times \cdots \times X_{d_s} \). Clearly the number of terms in the tensor-product sieve \( p^{k_n}(X) \) is given by \( k_n = \prod_{j=1}^{d_s} J_{j,n} \). See Schumaker (1981) and Newey (1997) for more details about tensor-product B-splines and other linear sieves. To summarize, a consistent linear sieve estimator for \( m(X, \alpha) \) is given by \( \hat{m}(X, \alpha) = (\hat{m}_1(X, \alpha), \ldots, \hat{m}_{d_p}(X, \alpha))' \) with

\[
\hat{m}_l(X, \alpha) = \sum_{j=1}^{n} \rho_l(Z_j, \alpha) p^{k_n}(X_j)' (P'P)^{-1} p^{k_n}(X) \quad (l = 1, \ldots, d_p),
\]

where \( P = (p^{k_n}(X_1), \ldots, p^{k_n}(X_n))' \). The integer \( k_n \) is called a smoothing parameter, which is required to grow with \( n \) so that the approximation error decreases to zero.

One attraction of the linear sieve estimator \( \hat{m}(X, \alpha) \) is that the proposed SMD estimator can be interpreted as a GMM (or a system of nonlinear 2SLS) estimator. To see this, note that with \( \hat{\Sigma}(X_i) = I \), the SMD estimator \( \hat{\alpha}_n \) is the solution to

\[
\min_{(\theta, h) \in \Theta \times \mathcal{H}_n} \left( \sum_{j=1}^{n} \rho(Z_j, \alpha) \otimes p^{k_n}(X_j) \right)' \left( I \otimes (P'P)^{-1} \right) \times \left( \sum_{j=1}^{n} \rho(Z_j, \alpha) \otimes p^{k_n}(X_j) \right),
\]

\(^{4}\)One can estimate \( m(X, \alpha) \) by any nonparametric methods such as kernel, nearest neighbor, or local linear regression. See the first version of this paper for the kernel estimation of \( m(X, \alpha) \).
where $\otimes$ denotes the Kronecker product and $I$ the $d_\rho \times d_\rho$-identity matrix. The criterion in (6) is exactly the GMM criterion based on the following (increasing number of) unconditional moment restrictions:

\begin{equation}
E[p_{0j}(X)\rho(Z, \theta_o, h_o)] = 0 \quad (j = 1, 2, \ldots, k_n)
\end{equation}

with $I \otimes (P'P)$ as the weighting matrix. This GMM interpretation suggests a lower bound on $k_n$: $d_\rho k_n \geq d_\theta + k_{1n}$, where $k_{1n} = \text{dim}(\mathcal{H}_n)$ denotes the number of unknown sieve parameters in $h \in \mathcal{H}_n = \mathcal{H}_{n1} \times \cdots \times \mathcal{H}_{n\rho}$, and is another smoothing parameter.

One advantage of the SMD procedure (4) is that it is natural and easy to implement. Once $h \in \mathcal{H}$ is replaced by $h_n \in \mathcal{H}_n$, the estimation problem effectively becomes a parametric one; hence commonly used statistical and econometric packages can be used to compute the proposed estimator. The sieve method is particularly convenient when $h_n$ enters $\rho(Z, \theta_o, h_o)$ nonlinearily or satisfies some shape restrictions such as monotonicity, concavity, additivity, exclusion restrictions, etc. The general theory that we present allows for $\mathcal{H}_n$ to be any of the commonly used sieve spaces such as power series, Fourier series, splines, Hermite polynomials, wavelets, or neural networks. In a specific application, one can decide upon his/her favorite sieve space $\mathcal{H}_n$ based on how well it approximates $\mathcal{H}$ and how easily it can be computed. Since, in most applications, how well a sieve can approximate $\mathcal{H}$ depends on the smoothness of functions in $\mathcal{H}$, we introduce a typical space of smooth functions, the Hölder space $\Lambda^\gamma(\mathcal{X})$ of order $\gamma > 0$, as an illustration. Let $\gamma$ denote the largest integer satisfying $\gamma < \gamma$. The Hölder space is a space of functions $g : \mathcal{X} \to \mathcal{R}$ such that the first $\gamma$ derivatives are bounded, and the $\gamma$-th derivatives are Hölder continuous with the exponent $\gamma - \gamma \in (0, 1]$ (i.e.,

\[
\max_{\sum_{i=1}^{d_x} a_i = \gamma} \left| \nabla^{a} g(x) - \nabla^{a} g(x') \right| \leq \text{const.}(\|x - x'\|_E)^{\gamma - \gamma}
\]

for all $x, x' \in \mathcal{X}$, where $a = (a_1, a_2, \ldots, a_{d_x})$ is a vector of nonnegative integers,

\[
\nabla^{a} g(x) = \frac{\partial^{a_1 + a_2 + \cdots + a_{d_x}} g(x)}{\partial x_{1}^{a_1} \cdots \partial x_{d_x}^{a_{d_x}}}
\]

denotes the $\sum_{i=1}^{d_x} a_i$-th derivative, and $\| \cdot \|_E$ denotes the Euclidean norm). The Hölder space becomes a Banach space when endowed with the Hölder norm:

\[
\|g\|_{\Lambda^\gamma} = \sup_{x} |g(x)| + \max_{a_1 + a_2 + \cdots + a_{d_x} = \gamma} \sup_{x \neq x'} \frac{|\nabla^{a} g(x) - \nabla^{a} g(x')|}{(\|x - x'\|_E)^{\gamma - \gamma}} < \infty.
\]

We call $\Lambda^\gamma(\mathcal{X}) = \{ g \in \Lambda^\gamma(\mathcal{X}) : \|g\|_{\Lambda^\gamma} \leq c < \infty \}$ a Hölder ball (with radius $c$). It is known that power series, Fourier series, splines, and wavelets all can approximate functions in $\Lambda^\gamma(\mathcal{X})$ well.
We are now ready to illustrate the SMD procedure with two useful examples.

**Example 2.1** (Partially linear regression with an endogenous nonparametric part):

\[
\rho(Z_i, \alpha_o) = Y_{i|1} - X'_{i|1}\theta_o - h_o(Y_{2i}), \quad E[\rho(Z_i, \alpha_o)|X_{i1}, X_{2i}] = 0,
\]

where \( \alpha_o = (\theta_o, h_o) \), \( Y_1 \) is a scalar, \( Z = (Y_1, Y'_1, X') \), \( X = (X'_1, X'_2) \). To keep things simple in this example, we assume homoscedasticity \( E[p(Z, \alpha_o)^2|X] = \Sigma_o \); \( \dim(X_2) = \dim(Y_2) = d \) and \( \dim(X_1) = \dim(X) = d + d \). Since the intercept can be absorbed in the unknown function \( h_o(Y_2) \), we assume that the constant is excluded from \( X_1 \) and \( Y_2 \). This model is an instrumental variable alternative to the model studied by Newey, Powell, and Vella (1999, p. 584, equation (7.1)). When \( Y_2 = X_2 \), it becomes Robinson's (1988) partially linear regression, which has received considerable attention in the literature. However, when \( Y_2 \) enters the unknown function \( h_o \) endogenously, there is no published work prior to ours on \( \sqrt{n} \) consistent and efficient estimation of \( \theta_o \).

To allow for the possibility that the conditional distribution of \( Y_2 \) given \( X \) might be a multivariate normal c.d.f., we assume that the support of \( Y_2 \) is \( \mathcal{R}^d \). We also assume that the unknown true function \( h_o \) belongs to a Hölder ball \( \mathcal{H} = \Lambda_{\gamma_1}(\mathcal{R}^d) \) with \( \gamma_1 > d/2 \). Finally, we use a univariate B-spline of order \( r > \gamma_1 \), given by

\[
B_r(u) = \frac{1}{(r-1)!} \sum_{i=0}^{r} (-1)^i \binom{r}{i} \left( \max(0, u - i) \right)^{r-i}
\]

to build a spline-wavelet sieve basis \( \{2^{k/2} B_r(2^k Y_{2i} - i) : i = 0, \pm 1, \pm 2, \ldots, k = 0, \ldots, K_n\} \) to approximate functions in \( \Lambda_{\gamma_1}(\mathcal{R}) \). Recall that the function \( B_r \) is a piecewise polynomial of highest degree \( r - 1 \), is \( r - 1 \) times differentiable over its support \([0, r]\), and satisfies \( \sum_{i=-\infty}^{\infty} B_r(u - i) = 1 \) for all \( u \in \mathcal{R} \). A function \( g \in \Lambda_{\gamma_1}(\mathcal{R}) \) can be approximated as \( g(x) \approx \sum_{k=0}^{K_n} \sum_{i=-\infty}^{\infty} a_i 2^{k/2} B_{2^k}(2^k x - i) \); see, e.g., Chui (1992) and Chen, Hansen, and Scheinman (1997).

The tensor-product of the univariate basis \( \{2^{k/2} B_r(2^k Y_{2i} - i) : i = 0, \pm 1, \pm 2, \ldots, k = 0, \ldots, K_n\} \) then forms a multivariate basis for functions in \( \Lambda_{\gamma_1}(\mathcal{R}^d) \). We denote the resulting tensor-product sieve basis as \( q^{k_1n}(Y_2) \), a \( k_1 \times 1 \)-vector with \( k_1 = (2^{K_n})^d \), and the sieve space as

\[
\mathcal{H}_n = \{ h(Y_2) = q^{k_1n}(Y_2)' \beta \text{ for all } \beta \text{ satisfying } \|h\|_{\Lambda_{\gamma_1}} \leq c_1 \}.
\]

The SMD procedure with identity weighting is just a 2SLS estimation applied to \( Y_{i1} = X'_{i1}\theta_o + q^{k_1n}(Y_{2i})' \beta + u_i \) with \( p^{k_2}(X_i) \) as instruments. We will show consistency of the SMD estimator of \( \alpha_o \) in Proposition 3.1, and the \( \sqrt{n} \) normality and efficiency of the SMD estimator of \( \theta_o \) in Propositions 4.1 and 6.1.
While the above example is linear in \( a \) with a homoscedastic error, the next example is nonlinear in \( a \) with a heteroscedastic error.

**Example 2.2** (Partially additive IV regression with a known link function):

\[
\rho(Z_i, \alpha_o) = Y_{1i} - F\left(Y_{2i} \theta_o + \sum_{j=1}^{q} h_{oj}(X_{ji})\right), \quad E[\rho(Z_i, \alpha_o)|X_i] = 0
\]

where \( \alpha_o = (\theta_o, h_{o1}, \ldots, h_{oq})' \), \( F(\cdot) \) is a known function, \( Y_1 \) is a scalar, \( Z = (Y_1, Y_2', X_2')' \), \( X_z = (X_1', \ldots, X_q')' \), \( X = (X_0', X_2')' \). For simplicity, we assume \( \text{dim}(X_0) = \text{dim}(Y_2) = d_\theta \) and \( \text{dim}(X_j) = 1 \) for \( j = 1, \ldots, q \), hence \( \text{dim}(X) = d_x = d_\theta + q \). In a recent work, Horowitz and Mammen (2002) considered this model with \( Y_2 \) being a constant, and they derived the pointwise distribution of the estimator of \( h_o(x_j) \), \( j = 1, \ldots, q \). Here we allow \( Y_2 \) to be endogenous and focus on the \( \sqrt{n} \)-efficient estimation of \( \theta_o \). Note that, when \( F(\cdot) \) is the identity function, Example 2.2 reduces to the model (3.5) in Chamberlain (1992, p. 579). Although Chamberlain derived the efficiency bound for this simpler case, there is no published work on the efficient estimation of \( \theta_o \) even when \( F(\cdot) \) is the identity function.

To keep things simple, we assume that \( h_{oj} \in \mathcal{H}_1 \equiv \mathcal{H}_1^1([-1, 1]), \gamma_1 > 1/2 \), for \( j = 1, \ldots, q \). For identification we assume that \( Y_2 \) contains a constant with \( \text{dim}(Y_2) > 1 \), and \( h_{oj}(0) = 0 \) for \( j = 1, \ldots, q \). We consider the Fourier series sieves for \( j = 1, \ldots, q \):

\[
\mathcal{H}_n^j = \left\{ \begin{array}{l}
    h_j(X_j) = a_0 + \sum_{l=1}^{J_n} \left[ a_{1l} \cos(\pi l X_j) + a_{2l} \sin(\pi l X_j) \right], \\
    h_j(0) = 0, \quad a_0 + \sum_{l=1}^{J_n} l^{2p} (a_{1l}^2 + a_{2l}^2) \leq c_l^2
  \end{array} \right\},
\]

where \( p \in (1/2, \gamma_1) \) is a constant arbitrarily close to \( \gamma_1 \).

We estimate \( \alpha_o = (\theta_o, h_o) \) by applying the SMD procedure described above with \( \mathcal{H}_n \equiv \mathcal{H}_n^1 \times \cdots \times \mathcal{H}_n^q \), \( \mathcal{H}_n^j \) given in (12) and \( k_n = q(2J_n + 1) \). The large sample properties of the SMD estimator for this example can be found in Propositions 3.2, 4.2, and 6.2.

### 3. Consistency and Convergence Rates

In this section, we first apply the result of Newey and Powell (2003) to obtain consistency of the SMD estimator \( \hat{\alpha}_n \) for \( \alpha_o \) under a metric \( \| \cdot \| \), such as the sup or \( L_2 \) metric. We then establish that \( \hat{\alpha}_n \) converges to \( \alpha_o \) at a rate faster than \( n^{-1/4} \) under a (weaker) metric \( \| \cdot \| \). The rate result will be used in Section 4 to establish the asymptotic normality of \( \hat{\theta}_n \). The following definitions are introduced to simplify the exposition in the main text.
DEFINITION 3.1: A real-valued measurable function \( g(Z, \alpha) \) is \( H\ddot{o}lder \) continuous in \( \alpha \in \mathcal{A} \) if there exist a constant \( \kappa \in (0, 1] \) and a measurable function \( c_2(Z) \) with \( E[c_2(Z)^2 | X] \) bounded, such that \( |g(Z, \alpha_1) - g(Z, \alpha_2)| \leq c_2(Z)\|\alpha_1 - \alpha_2\|^\kappa \) for all \( Z \in \mathcal{Z}, \alpha_1, \alpha_2 \in \mathcal{A} \).

DEFINITION 3.2: A real-valued measurable function \( g(Z, \alpha) \) satisfies an envelope condition over \( \alpha \in \mathcal{A} \) if there exists a measurable function \( c_1(Z) \) with \( E[c_1(Z)^4] < \infty \) such that \( |g(Z, \alpha)| \leq c_1(Z) \) for all \( Z \in \mathcal{Z} \) and \( \alpha \in \mathcal{A} \).

The following conditions are quite similar to those imposed by Newey and Powell (2003):

**Assumption 3.1:**
(i) The data \( \{(Y_t, X_t)\}_{t=1}^n \) are i.i.d.; (ii) \( \mathcal{X} \) is compact with nonempty interior; (iii) the density of \( X \) is bounded and bounded away from zero.

**Assumption 3.2:**
(i) The smallest and the largest eigenvalues of \( E[p_k(X) \times p_k(X)'|X] \) are bounded and bounded away from zero for all \( k_n \); (ii) for any \( g(\cdot) \) with \( E[g(X)^2] < \infty \), there exists \( p_k(X)^\prime \pi \) such that \( E[(g(X) - p_k(X)^\prime \pi)^2] = o(1) \).

**Assumption 3.3:** \( \alpha_o \in \mathcal{A} \) is the only \( \alpha \in \mathcal{A} \) satisfying \( E[\rho(Z, \alpha)|X] = 0 \).

**Assumption 3.4:**
(i) \( \hat{\Sigma}(X) = \Sigma(X) + o_P(1) \) uniformly over \( X \in \mathcal{X} \); (ii) \( \Sigma(X) \) is finite positive definite uniformly over \( X \in \mathcal{X} \).

**Assumption 3.5:**
(i) There is a metric \( \| \cdot \|_s \) such that \( \mathcal{A} \equiv \Theta \times \mathcal{H} \) is compact under \( \| \cdot \|_s \); (ii) for any \( \alpha \in \mathcal{A} \), there exists \( \Pi_{n\alpha} \in \mathcal{A}_n \equiv \Theta \times \mathcal{H}_n \) such that \( \| \Pi_{n\alpha} - \alpha \|_s = o(1) \).

**Assumption 3.6:**
(i) \( E[|\rho(Z, \alpha_o)|^2 |X] \) is bounded; (ii) \( \rho(Z, \alpha) \) is \( H\ddot{o}lder \) continuous in \( \alpha \in \mathcal{A} \).

**Assumption 3.7:**
(i) \( d_n k_n \geq d_\theta + k_{1n}, k_{1n} \rightarrow +\infty, \) and \( k_n/n \rightarrow 0 \).

Since Assumptions 3.1–3.3, 3.5, and 3.6 are essentially the same conditions imposed in Newey and Powell (2003), here we provide only a brief discussion. Assumption 3.1(i) rules out time series observations. This condition can be easily relaxed for the results on consistency, rate of convergence and limiting distribution; however it is needed for the efficiency result in Section 6. Assumptions 3.1(ii) and (iii) and 3.2(i) and (ii) are typical conditions imposed for series (or linear sieve) estimation of conditional mean functions. Assumption 3.3 is an identification condition, which has to be verified in each application. Assumption 3.4 is a condition on the estimated weighting matrix,
which is trivially satisfied if the weighting matrix is known (say the identity matrix). Assumption 3.5(i) restricts the parameter space as well as the choice of the metric $\| \cdot \|_i$. It is a commonly imposed condition in the nonparametric and semiparametric econometrics literature, and is satisfied when the infinite-dimensional parameter space $\mathcal{H}$ consists of bounded and smooth functions. See Gallant and Nychka (1987) for detailed discussions about this assumption, and also see Propositions 3.1 and 3.2 below for examples. Assumption 3.5(ii) is simply the definition of a sieve space. Assumption 3.6 is typically imposed on the residual function even in the literature about parametric nonlinear estimation. Assumption 3.7(i) simply states the relationship between the two smoothing parameters $k_n$ and $k_{1n}$ implied by the GMM interpretation. Applying Theorem 4.1 in Newey and Powell (2003), we obtain the following lemma.

**Lemma 3.1:** Under Assumptions 3.1, 3.2(i) and (ii), 3.3, 3.4(i) and (ii), 3.5(i) and (ii), 3.6(i) and (ii), and 3.7(i), we have $\| \hat{\alpha}_n - \alpha_0 \|_i = o_p(1)$.

Lemma 3.1 provides a consistency result under the pseudo metric $\| \cdot \|_i$, which is not enough to establish the asymptotic normality of $\hat{\alpha}_n$. It is well-known that to derive the asymptotic normality of $\hat{\alpha}_n$, one typically needs that $\hat{\alpha}_n$ converges to $\alpha_0$ at a rate faster than $n^{-1/4}$; see, e.g., Newey (1994) and Pakes and Olley (1995). However, as noted by Newey and Powell (2003), for model (1) where the unknown $h_0$ could depend on endogenous variables $Y$, it is generally difficult to obtain a fast convergence rate under the metric $\| \cdot \|_i$, which is either a weighted sup-norm or $L_2$-norm. In this paper, we consider another (weaker) metric $\| \cdot \|_1$ and compute the convergence rate under this metric. We now introduce the $\| \cdot \|_1$-metric and the $L_2$-metric.

Suppose that the parameter space $\mathcal{A}$ is connected in the sense that for any two points $\alpha_1, \alpha_2 \in \mathcal{A}$, there exists a continuous path $\{\alpha(\tau) : \tau \in [0, 1]\}$ in $\mathcal{A}$ such that $\alpha(0) = \alpha_1$ and $\alpha(1) = \alpha_2$. Also, suppose that $\mathcal{A}$ is convex at the true value $\alpha_0$ in the sense that, for any $\alpha \in \mathcal{A}$, $(1 - \tau)\alpha_0 + \tau\alpha \in \mathcal{A}$ for small $\tau > 0$. Furthermore, suppose that for almost all $Z$, $\rho(Z, (1 - \tau)\alpha_0 + \tau\alpha)$ is continuously differentiable at $\tau = 0$. Denote the first pathwise derivative at the direction $[\alpha - \alpha_0]$ evaluated at $\alpha_0$ by:

$$\frac{d\rho(Z, \alpha_0)}{d\alpha}[\alpha - \alpha_0] = \frac{d\rho(Z, (1 - \tau)\alpha_0 + \tau\alpha)}{d\tau}\bigg|_{\tau=0} \quad \text{a.s.} \ Z$$

and for any $\alpha_1, \alpha_2 \in \mathcal{A}$ denote

$$\frac{d\rho(Z, \alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] = \frac{d\rho(Z, \alpha_0)}{d\alpha}[\alpha_1 - \alpha_0] - \frac{d\rho(Z, \alpha_0)}{d\alpha}[\alpha_2 - \alpha_0],$$

$$\frac{dm(X, \alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] = \mathbb{E}\left\{\frac{d\rho(Z, \alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]|X\right\}.$$
For any $\alpha_1, \alpha_2 \in \mathcal{A}$, the metric $\| \cdot \|$ is defined as
\begin{equation}
\| \alpha_1 - \alpha_2 \|
= \sqrt{E \left\{ \left( \frac{dm(X, \alpha_0)}{d\alpha} [\alpha_1 - \alpha_2] \right)' \Sigma(X)^{-1} \frac{dm(X, \alpha_0)}{d\alpha} [\alpha_1 - \alpha_2] \right\}},
\end{equation}
and the $L_2$-metric is
\begin{equation}
\| \alpha_1 - \alpha_2 \|_2
= \sqrt{E \left\{ \left( \frac{d\rho(Z, \alpha_0)}{d\alpha} [\alpha_1 - \alpha_2] \right)' \Sigma(X)^{-1} \frac{d\rho(Z, \alpha_0)}{d\alpha} [\alpha_1 - \alpha_2] \right\}}.
\end{equation}
It is clear that $\| \alpha - \alpha_0 \| \leq \| \alpha - \alpha_0 \|_2$ for all $\alpha \in \mathcal{A}$, and the equality holds if and only if
\begin{equation}
\frac{d\rho(Z, \alpha_0)}{d\alpha} [\alpha - \alpha_0] = E \left\{ \frac{d\rho(Z, \alpha_0)}{d\alpha} [\alpha - \alpha_0] \mid X \right\} \quad \text{a.s.-}X
\end{equation}
(i.e. if and only if the pathwise derivative
\begin{equation}
\frac{d\rho(Z, \alpha_0)}{d\alpha} [\alpha - \alpha_0]
\end{equation}
does not depend on the endogenous variables $Y$). In particular, for conditional moment models where unknown functions $h$ depend on $Y$, we could have slow convergence rates under $\| \cdot \|_2$ and fast convergence rates under $\| \cdot \|$. Luckily to derive $\sqrt{n}$-normality of the SMD estimator $\hat{\theta}_n$ in the next section, it suffices to have a fast convergence rate under the weaker metric $\| \cdot \|$. We now state additional conditions that are needed to compute the convergence rate.

**Assumption 3.2:** (iii) For any $g(\cdot)$ in $\Lambda_c^\gamma(\mathcal{X})$ with $\gamma > d_x/2$, there exists $p_{\gamma}(\cdot, \pi) \in \Lambda_c^\gamma(\mathcal{X})$ such that $\sup_{X \in \mathcal{X}} |g(X) - p_{\gamma}(X, \pi)| = O(k_{n^{-\gamma/d_x}})$, and $k_{n^{-\gamma/d_x}} = o(n^{-1/4}).$

**Assumption 3.4:** (iii) $\widehat{\Sigma}(X) = \Sigma(X) + o_p(n^{-1/4})$ uniformly over $X \in \mathcal{X}$.

**Assumption 3.5:** (iii) There is a constant $\mu_1 > 0$ such that for any $\alpha \in \mathcal{A}$, there is $\Pi_n \alpha \in \mathcal{A}_n$ satisfying $\| \Pi_n \alpha - \alpha \| = O(k_{1n^{-\mu_1}})$, and $k_{1n^{-\mu_1}} = o(n^{-1/4}).$

**Assumption 3.6:** (iii) Each element of $\rho(Z, \alpha)$ satisfies an envelope condition in $\alpha \in \mathcal{A}_n$; (iv) each element of $m(\cdot, \alpha) \in \Lambda_c^\gamma(\mathcal{X})$ with $\gamma > d_x/2$, for all $\alpha \in \mathcal{A}_n$. 
Let $\xi_{0n} \equiv \sup_{x \in \mathcal{X}} \| p^k(x) \|_E$, which is nondecreasing in $k_n$. Denote $N(\delta, \mathcal{A}_n, \| \cdot \|_s)$ as the minimal number of radius $\delta$ covering balls of $\mathcal{A}_n$ under the $\| \cdot \|_s$ metric.

**Assumption 3.7:** (ii) $k_{1n} \times \ln n \times \xi_{0n}^2 \times n^{-1/2} = o(1)$.

**Assumption 3.8:** $\ln[N(\varepsilon^{1/\kappa}, \mathcal{A}_n, \| \cdot \|_s)] \leq \text{const.} \times k_{1n} \times \ln(k_{1n}/\varepsilon)$.

**Assumption 3.9:** (i) $A$ is convex in $\alpha_o$, and $\rho(Z, \alpha)$ is pathwise differentiable at $\alpha_o$; (ii) for some $c_1, c_2 > 0$,

$$c_1 E \left\{ m(X, \alpha) \Sigma(X)^{-1} m(X, \alpha) \right\} \leq \| \alpha - \alpha_o \|^2$$

$$\leq c_2 E \left\{ m(X, \alpha) \Sigma(X)^{-1} m(X, \alpha) \right\}$$

holds for all $\alpha \in \mathcal{A}_n$ with $\| \alpha - \alpha_o \|_s = o(1)$.

The following result is proved in Appendix B.

**Theorem 3.1:** Under Assumptions 3.1–3.9, we have $\| \hat{\alpha}_n - \alpha_o \| = o_p(n^{-1/4})$.

**Discussion of Assumptions:** Assumption 3.2(iii) quantifies the approximation error of functions in $\mathcal{A}_n^\gamma(\mathcal{X})$ by the linear sieve basis functions $p^k(X)$. It is satisfied by commonly used sieve basis functions such as splines, power series, wavelets, and Fourier series. Assumption 3.6(iii) and (iv) impose a dominance condition on $\rho(\cdot)$ and a smoothness condition on $m(\cdot)$. Assumptions 3.2(iii) and 3.6(iv) imply that the bias part of the estimator $\hat{m}(X, \alpha)$ shrinks to zero at the order $O(k_n^{-\gamma/d_x}) = o(n^{-1/4})$ uniformly over $X \in \mathcal{X}, \alpha \in \mathcal{A}_n$, while Assumptions 3.1, 3.2(i), 3.6(ii) and (iii), and 3.7(ii) together determine the rate $o_p(n^{-1/4})$ at which the residual part $\hat{m}(X, \alpha) - E_X[\hat{m}(X, \alpha)]$ goes to zero uniformly over $X \in \mathcal{X}, \alpha \in \mathcal{A}_n$. All these assumptions can be relaxed with more tedious proofs. See Lemma A.1 in the Appendix for a general uniform convergence result.

Assumption 3.4(iii) strengthens Assumption 3.4(i) by quantifying the estimation error of the weighting matrix. It is trivially satisfied by the identity weighting matrix.

Assumption 3.5(iii) quantifies the approximation error of functions in $\mathcal{H}$ by the sieve $\mathcal{H}_n$. This condition is satisfied by commonly used function spaces $\mathcal{H}$ (e.g., Sobolev, Besov, Holder), whose elements are sufficiently smooth, and by popular sieves (e.g., power series, Fourier series, Hermite series, finite-element, splines, wavelet, neural networks). Assumption 3.8 requires that the size of the sieve space $\mathcal{H}_n$ does not grow too fast in terms of the covering number. It is satisfied by commonly used linear and nonlinear sieves. For instance, for power series, Fourier series, splines, and wavelet linear sieves, we have
\[ \ln[N(\varepsilon, \mathcal{A}_n, \| \cdot \|, \varepsilon)] = c k_{1n} \ln(1/\varepsilon) \] (see, e.g., Chen and Shen (1998)); and for neural network and ridgelet nonlinear sieves, we have \[ \ln[N(\varepsilon, \mathcal{A}_n, \| \cdot \|, \varepsilon)] = c k_{1n} \ln(k_{1n}/\varepsilon) \] (see, e.g., Chen and White (1999)). See Propositions 3.1 and 3.2 below for concrete examples of both assumptions. Additional examples can be found in papers such as Fenton and Gallant (1996) for Hermite polynomials, and Newey (1997) for splines and power series.

Assumption 3.9(i) implies that the metric \( \| \cdot \| \) is well-defined. This condition is assumed in this paper only for simplicity, and is not needed if we use general nonlinear smooth path \( \{ \alpha(\tau) : \tau \in [0, 1] \} \) in \( \mathcal{A} \) going through \( \alpha_0 \). Assumption 3.9(ii) implies that the population criterion function can be approximated locally by \( \| \alpha - \alpha_0 \|^2 \). This condition is very useful for deriving the convergence rate of \( \hat{\alpha}_n - \alpha_0 \) and the \( \sqrt{n} \)-normality of \( \hat{\theta}_n - \theta_0 \). It is trivially satisfied over the entire \( \mathcal{A}_n \) when \( \rho \) is linear in \( \alpha \) (see Example 2.1). When \( \rho \) is nonlinear in \( \alpha \), this condition can still be satisfied in a neighborhood of \( \alpha_0 \) (see Example 2.2).

Recall that the SMD procedure requires two sieve approximations: one \( (p^{k_n}(X)) \) for the conditional mean function \( m(X, \alpha) \) and the other \( (\mathcal{H}_n) \) for the parameter space \( \mathcal{H} \). Assumptions 3.2(iii), 3.5(iii), 3.7, and 3.8 jointly determine the growth orders of the two smoothing parameters \( k_n \) and \( k_{1n} \). In particular, Assumptions 3.2(iii) and 3.5(iii) are restrictions on the approximation errors, while Assumptions 3.7(ii) and 3.8 are restrictions on the sizes of both sieves. Assumption 3.7(ii) is easy to verify since \( \xi_{n0} = \sup_{x \in X} \| p^{k_n}(x) \|_{E} \) can be computed upon the specification of the linear sieve \( p^{k_n}(X) \). For example, \( \xi_{n0} = k_n^{1/2} \) if \( p^{k_n}(X) \) is a tensor-product B-spline basis of order \( [\gamma] + 1 \) or a Fourier series sieve; \( \xi_{n0} = k_n \) if \( p^{k_n}(X) \) is a tensor-product polynomial power series sieve; see Newey (1997) for more examples.

We conclude this section by specializing Lemma 3.1 and Theorem 3.1 to the two examples.

**Proposition 3.1:** Let \( \hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n) \) be the SMD estimator for Example 2.1 with identity weighting, \( \mathcal{H}_n \) given in (10), and

\[
\| \alpha \|_s = \| \theta \|_E + \sup_{y_2} |h(y_2) \times (1 + \| y_2 \|_E^2)^{-a/2}| \quad \text{for some} \quad a > \gamma_1 > 0.
\]

Suppose that Assumptions 3.1, 3.2(i), \( E\{\rho(Z, \alpha_0)^2 \mid X\} = \Sigma_\rho > 0 \), and the following hold:

3.1.1. (i) The conditional distribution of \( Y_2 \) given \( X \) is complete; (ii) \( h_0 \in \mathcal{H} = \Lambda_{\gamma_1}^\gamma (R^d); \) (iii) \( E[h_0(Y_2) \mid X] \notin \text{linear span} (X_1) \), and \( E[X_1X_1^\prime] \) is finite positive-definite.

3.1.2. (i) \( E[|Y_1|^4] < \infty; \) (ii) \( E[(1 + Y_2)^a \mid X] \) bounded for some \( a > \gamma_1 \).

3.1.3. \( E[Y_1 \mid X = x] \), \( E[q^{k_{1n}}(Y_2) \mid X = x] \in \Lambda_\gamma^\gamma (X) \), \( \gamma > d_s/2. \)

3.1.4. Either (i) \( p^{k_n}(X) \) is a tensor-product Fourier series (or B-spline basis of order \( [\gamma] + 1 \) or wavelets), and \( k_n k_{1n} \ln(n)/\sqrt{n} = o(1) \); or (ii) \( p^{k_n}(X) \) is a tensor-product power series, and \( k_n^2 k_{1n} \ln(n)/\sqrt{n} = o(1) \).
3.1.5. (i) \( k_n \geq d_n + k_{1n} \); (ii) \( k_n^{-\gamma/d} = o(n^{-1/4}) \); (iii) \( k_n^{-\gamma_1/d} = o(n^{-1/4}) \).

Then: \( \| \hat{\alpha}_n - \alpha_0 \|_\tau = o_p(1) \); \( \| \hat{\theta}_n - \theta_0 \|_E = o_p(n^{-1/4}) \); \( E(\{ E[\hat{h}_n(Y_2) - h_o(Y_2)|X]\}^2) = o_p(n^{-1/2}) \).

Conditions for Proposition 3.1 are mild. In particular, as long as their higher order moments exist, the endogenous variables \( Y_1, Y_2 \) are allowed to have unbounded supports.

**Proposition 3.2**: Let \( \hat{\alpha}_n = (\hat{\alpha}_n, \hat{h}_n) \) be the SMD estimator for Example 2.2 with \( H_n \) given in (12), and \( \| \alpha \|_\tau = \| \theta \|_E + \max_{j=1,...,q} \sup_{x_j} | h_j(x_j) | \). Suppose that Assumptions 3.1, 3.2(i), 3.4, 3.6(i) and the followings hold:

3.2.1. (i) \( F(\cdot) \) is twice differentiable; (ii) \( h_{oj} \in H^\gamma = \Lambda_2^{\gamma}([-1, 1]), \gamma_1 > 1/2, \) and \( h_{oj}(0) = 0 \) for \( j = 1, \ldots, q \); (iii) \( E(\{ E[\nabla F(Y_2^\theta + \sum_{j=1}^q h_{oj}(X_j))Y_2^\tau |X]\}^2) \) has full rank, \( |E[\nabla F(Y_2^\theta + \sum_{j=1}^q h_{oj}(X_j))|X]| \) is bounded and bounded away from zero for all \( X \);

3.2.2. (i) \( E[|Y_1^2|] < \infty \) and \( E[\{ \sup_{\alpha \in A_n} | F(Y_2^\theta + \sum_{j=1}^q h_j(X_j)) | \}^4] < \infty \);

(ii) \( \sup_{\alpha \in A} | \nabla F(Y_2^\theta + \sum_{j=1}^q h_j(X_j)) | \leq c(Z) \) with \( E[\{ c(Z)^2 \{ \| Y_2^2 \|_E + 1 \} |X] \) bounded;

3.2.3. \( E[Y_1 | X = \cdot] \) and \( E[\nabla F(Y_2^\theta + \sum_{j=1}^q h_j(X_j)) | X = \cdot] \) are in \( \Lambda_2^{\gamma}(X) \) for all \( \alpha \in A_n \);

3.2.4. \( p^{k_n}(X) \) is a tensor-product Fourier series (or B-spline basis of order \( [\gamma] + 1 \) or wavelets) sieve, and \( k_n J_n \ln(n) / \sqrt{n} = o(1) \);

3.2.5. (i) \( k_n \geq d_n + q(2J_n + 1) \); (ii) \( k_n^{-\gamma/d} = o(n^{-1/4}) \); (iii) \( J^{-\gamma_1} = o(n^{-1/4}) \);

3.2.6. \( \sup_{\alpha \in A_n, \| \alpha - \alpha_0 \| = o(1)} | \nabla^2 F(Y_2^\theta + \sum_{j=1}^q h_j(X_j)) | \leq C(Z) \) such that \( E[| C(Z) |X], E[C(Z) Y_2^\tau |X] \) and \( E[| C(Z) Y_2 Y_2^\tau |X] \) are bounded.

Then: \( \| \hat{\alpha}_n - \alpha_0 \|_\tau = o_p(1) \); \( \| \hat{\theta}_n - \theta_0 \|_E = o_p(n^{-1/4}) \); \( E[\{ \sum_{i=1}^q (\hat{h}_n(X_i) - h_{oi}(X_i)) \}^2] = o_p(n^{-1/2}) \).

Conditions for Proposition 3.2 are also mild. In addition to the restriction on the higher order moments of the endogenous variables, the others are the dominance, the smoothness, and the boundedness conditions, which are familiar in the nonlinear econometrics literature.

4. **ASYMPTOTIC NORMALITY**

We now derive the asymptotic distribution of \( \hat{\theta}_n \). As before, we must first introduce some notations. Let \( \overline{V} \) denote the closure of the linear span of \( A - \{ \alpha_0 \} \) under the metric \( \| \cdot \| \). Then \( (\overline{V}, \| \cdot \|) \) is a Hilbert space with the inner product:

\[
\langle v_1, v_2 \rangle = E \left\{ \left( \frac{d m(X, \alpha_0)}{d \alpha} [v_1] \right) \Sigma(X)^{-1} \left( \frac{d m(X, \alpha_0)}{d \alpha} [v_2] \right) \right\}.
\]
MODELS WITH CONDITIONAL MOMENT RESTRICTIONS

1809

For any fixed and nonzero \( \lambda \in \mathcal{R}^d \), \( f(\alpha) \equiv \lambda' \theta \) is clearly a linear functional on \( \bar{V} \). By the results in Van der Vaart (1991) and Shen (1997), \( f(\alpha) = \lambda' \theta \) has to be bounded (i.e. \( \sup_{\theta \neq \theta_0, \alpha \in \bar{V}} (|f(\alpha) - f(\alpha_o)|/\|\alpha - \alpha_o\|) < \infty \)) in order for \( f(\alpha_o) \equiv \lambda' \theta_o \) to be estimated at a \( \sqrt{n} \)-rate.

With \( \bar{V} = \mathcal{R}^d \times \mathcal{W} \) and \( \mathcal{V} = \mathcal{H} - \{h_o\} \), we write

\[
\frac{dm(X, \alpha_o)}{d\alpha} [\alpha - \alpha_o] = \frac{dm(X, \alpha_o)}{d\theta} (\theta - \theta_o) + \frac{dm(X, \alpha_o)}{dh} [h - h_o].
\]

For each component \( \theta_j \) (of \( \theta \)), \( j = 1, \ldots, d_\theta \), let \( w_j^* \in \mathcal{W} \) denote the solution to

\[
\min_{w_j \in \mathcal{W}} E \left\{ \left( \frac{dm(X, \alpha_o)}{d\theta_j} - \frac{dm(X, \alpha_o)}{dh} [w_j] \right)' \Sigma(X)^{-1} \times \left( \frac{dm(X, \alpha_o)}{d\theta_j} - \frac{dm(X, \alpha_o)}{dh} [w_j] \right) \right\}.
\]

Define

\[
w^* = (w_1^*, \ldots, w_{d_\theta}^*),
\]

\[
\frac{dm(X, \alpha_o)}{dh} [w^*] = \left( \frac{dm(X, \alpha_o)}{dh} [w_1^*], \ldots, \frac{dm(X, \alpha_o)}{dh} [w_{d_\theta}^*] \right), \text{ and}
\]

\[
D_{w^*}(X) \equiv \frac{dm(X, \alpha_o)}{d\theta} - \frac{dm(X, \alpha_o)}{dh} [w^*].
\]

It is easy to show that for \( f(\alpha) \equiv \lambda' \theta \) with \( \lambda \in \mathcal{R}^d \), \( \lambda \neq 0 \),

\[
\sup_{0 \neq \alpha - \alpha_o \in \bar{V}} \frac{|f(\alpha) - f(\alpha_o)|^2}{\|\alpha - \alpha_o\|^2} = \lambda' \left( E[D_{w^*}(X)' \Sigma(X)^{-1} D_{w^*}(X)] \right)^{-1} \lambda.
\]

Thus \( f(\alpha) = \lambda' \theta \) is bounded if and only if \( E[D_{w^*}(X)' \Sigma(X)^{-1} D_{w^*}(X)] \) is finite positive-definite, in which case we have

\[
f(\alpha) - f(\alpha_o) \equiv \lambda' (\theta - \theta_o) = (v^*, \alpha - \alpha_o) \quad \text{for all} \ \alpha \in \mathcal{A}
\]

where \( v^* \equiv (v^*_\theta, v^*_h) \in \bar{V} \) with \( v^*_\theta = (E[D_{w^*}(X)' \Sigma(X)^{-1} D_{w^*}(X)])^{-1} \lambda \), \( v^*_h = -w^* \times v^*_\theta \).

The following conditions are sufficient for the \( \sqrt{n} \)-normality of \( \hat{\theta}_n \):

ASSUMPTION 4.1: (i) \( E[D_{w^*}(X)' \Sigma(X)^{-1} D_{w^*}(X)] \) is positive-definite; (ii) \( \theta_o \in \text{int}(\Theta) \); (iii) \( \Sigma_o(X) \equiv \text{var}[\rho(Z, \alpha_o)|X] \) is positive definite for all \( X \in \mathcal{X} \).

ASSUMPTION 4.2: There is a \( v^*_n = (v^*_\theta, -\Pi_n w^* \times v^*_\theta) \in \mathcal{A}_n - \alpha_o \) such that \( \|v^*_n - v^*\| = O(n^{-1/4}) \).
Let $\mathcal{N}_{ou} \equiv \{\alpha \in \mathcal{A}_n : \|\alpha - \alpha_o\|_2 = o(1), \|\alpha - \alpha_o\| = o(n^{-1/4})\}$ and define $\mathcal{N}_o$ the same way with $\mathcal{A}_n$ replaced by $\mathcal{A}$. Also for any $v \in \overline{\mathcal{V}}$, we denote

$$\frac{d\rho(Z, \alpha)}{d\alpha}[v] \equiv \left. \frac{d\rho(Z, \alpha + \tau v)}{d\tau} \right|_{\tau=0} \text{ a.s. } Z,$$

and

$$\frac{dm(X, \alpha)}{d\alpha}[v] \equiv E\left\{ \frac{d\rho(Z, \alpha)}{d\alpha}[v] \mid X \right\} \text{ a.s. } X.$$

**ASSUMPTION 4.3:** For all $\alpha \in \mathcal{N}_o$, the pathwise first derivative $(d\rho(Z, \alpha(t))/d\alpha)[v]$ exists a.s. $Z \in \mathcal{Z}$. Moreover, (i) each element of $(d\rho(Z, \alpha)/d\alpha)[v^*_n]$ satisfies an envelope condition and is Hölder continuous in $\alpha \in \mathcal{N}_{ou}$; (ii) each element of $(dm(X, \alpha)/d\alpha)[v^*_n]$ is in $\mathcal{A}_c(X)$, $\gamma > d_x/2$ for all $\alpha \in \mathcal{N}_o$.

**ASSUMPTION 4.4:** Uniformly over $\alpha \in \mathcal{N}_{ou}$,

$$E\left( \left\| \frac{dm(X, \alpha)}{d\alpha}[v^*_n] - \frac{dm(X, \alpha_o)}{d\alpha}[v^*_n] \right\|_E^2 \right) = o(n^{-1/2}).$$

**ASSUMPTION 4.5:** Uniformly over $\alpha \in \mathcal{N}_o$, $\tilde{\alpha} \in \mathcal{N}_{ou}$,

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

**ASSUMPTION 4.6:** For all $\alpha \in \mathcal{N}_{ou}$, the pathwise second derivative $d^2\rho(Z, \alpha + \tau v^*_n)/d\tau^2|_{\tau=0}$ exists a.s. $Z \in \mathcal{Z}$, and is bounded by a measurable function $c_5(Z)$ with $E[c_5(Z)^2] < \infty$.

Assumption 4.1(i) is a local identification condition for $\theta_o$. This condition cannot be relaxed. Assumption 4.2 is so-called “no asymptotic bias” condition, which is needed due to the presence of unknown $h_o$. Here for simplicity we assume that the same sieve space $\mathcal{H}_n$ approximates the space $\overline{\mathcal{H}} \equiv \overline{\mathcal{H}} - \{h_o\}$ well. Nevertheless, for Theorem 4.1 it is enough to assume that $v^*_h = -w^*v^*_g$ could be approximated well by any sieve spaces, possibly different from $\mathcal{H}_n$. Assumption 4.3 is similar to Assumption 3.6, except that it is imposed on the pathwise derivative $(d\rho(Z, \alpha)/d\alpha)[v^*_n]$. Notice that when $\rho$ is linear in $\alpha$, Assumption 4.3 is trivially satisfied. If $\rho$ is nonlinear in $\alpha$, this assumption can still be satisfied; see the verification of Example 2.2 below.
The final three assumptions are only needed when \( \alpha \) enters \( \rho \) in a highly nonlinear manner. More specifically, Assumptions 4.4 and 4.5 are needed to control the asymptotic bias when \( \alpha \) enters \( \rho \) nonlinearly. They are automatically satisfied when \( \rho \) is linear in \( \alpha \) and are easily satisfied when \( \rho \) is nonlinear in \( \alpha \) and when \( h \) does not depend on \( Y \). When \( \rho \) is nonlinear in \( h \) and \( h \) depends on \( Y \), these conditions may not hold within the \( o(n^{-1/4}) \)-neighborhood \( \mathcal{N}_o \), but they could still hold over a smaller neighborhood \( \{ \alpha \in A_n : \| \alpha - \alpha_o \|_2 = o(1), \| \alpha - \alpha_o \| = o(\delta_n) \} \) for some \( \delta_n = o(n^{-1/4}) \). In this case, Theorem 4.1 still can be established if conditions of Theorem 3.1 are strengthened so that \( \| \hat{\alpha}_n - \alpha_o \| = o_p(\delta_n) \) for \( \delta_n = o(n^{-1/4}) \). Assumption 4.6 is needed to control the higher order terms in the asymptotic expansion.

The following result is proved in Appendix C.

**THEOREM 4.1:** Under Assumptions 3.1–3.9 and 4.1–4.6, \( \sqrt{n}(\hat{\theta}_n - \theta_o) \rightarrow N(0, \mathbf{V}^{-1}) \) where

\[
\mathbf{V}^{-1} \equiv \left( E[D_{w^*}(X)'\Sigma(X)^{-1}D_{w^*}(X)] \right)^{-1} \\
\times \left( E[D_{w^*}(X)'\Sigma(X)^{-1}\Sigma_o(X)\Sigma(X)^{-1}D_{w^*}(X)] \right)^{-1} \\
\times \left( E[D_{w^*}(X)'\Sigma(X)^{-1}D_{w^*}(X)] \right)^{-1}.
\]

Theorem 4.1 shows that the SMD estimator for the parametric component \((\theta_o)\) is \( \sqrt{n} \)-consistent and asymptotically normally distributed. The asymptotic covariance has three terms. If we can interpret \( D_{w^*}(X) \) as regressors in a system of equations with \( \rho(Z, \alpha_o) \) as dependent variables, then \( \mathbf{V}^{-1} \) can be thought of as the covariance matrix of the weighted LS estimator. The three terms in \( \mathbf{V}^{-1} \) correspond to the three terms in the covariance of the (incorrectly) weighted LS estimator in the presence of heteroskedasticity.

We conclude this section by applying Theorem 4.1 to the two examples.

**PROPOSITION 4.1:** For Example 2.1, let \( w^* = (w^*_1, \ldots, w^*_{d_\theta}) \) be the solution to

\[
\min_{w_j \in \mathbb{W}} E \left[ (X_{ij} - E[w_j(Y_2)|X])^2 \right] \quad (j = 1, \ldots, d_\theta).
\]

Suppose all conditions of Proposition 3.1, Assumption 4.1(ii) and the following are satisfied:

4.1.1. For \( j = 1, \ldots, d_\theta \), \( E[w^*_j(Y_2)|X = \cdot] \in \Lambda^*_\gamma(X), \gamma > d_\theta/2 \).

Then: \( \sqrt{n}(\hat{\theta}_n - \theta_o) \rightarrow N(0, \mathbf{V}^{-1}) \), where \( \mathbf{V} = E[D_{w^*}(X)'\Sigma_o^{-1}D_{w^*}(X)] \) and \( D_{w^*}(X) = X^*_1 - E[w^*(Y_2)|X] \).

We note that the \( \sqrt{n} \) asymptotic normality of \( \hat{\theta}_n \) is established using the convergence rate \( \| \hat{\alpha}_n - \alpha_o \| = o_p(n^{-1/4}) \), even though \( \hat{h}_n(Y_2) \) may converge to \( h_o(Y_2) \) arbitrarily slowly under the metric \( \| \cdot \|_2 \).
PROPOSITION 4.2: For Example 2.2, let \( w^* = (w^*_1, \ldots, w^*_{d_\theta}) \) be the solution to
\[
\inf_{w_j \in \mathcal{W}} E \left[ D_{w_j} (X) \Sigma (X)^{-1} D_{w_j} (X) \right] \quad (j = 1, \ldots, d_\theta),
\]
\[
D_{w_j} (X) = E \left[ \nabla F \left( Y_2 \theta_\alpha + \sum_{l=1}^q h_{ol} (X_l) \right) \times \left\{ Y_{2j} - \sum_{l=1}^q w_j^l (X_l) \right\} |X \right].
\]

Suppose all conditions of Proposition 3.2, Assumption 4.1(ii) and (iii) and the followings are satisfied:

4.2.1. For \( j = 1, \ldots, d_\theta, \) \( w^*_j (X_i) \in \Lambda^r (\mathcal{X}), \gamma > d_x / 2. \)
4.2.2. (i) \( E \left[ \sup_{\alpha \in \mathcal{N}_n} \left| \nabla F (Y_2 \theta + \sum_{l=1}^q h_l (X_l)) \right| \times \left( \| Y_2 \|_E + 1 \right) \right] < \infty; \)
(ii) \( E \left[ \sup_{\alpha \in \mathcal{N}_n} \left| \nabla^2 F (Y_2 \theta + \sum_{l=1}^q h_l (X_l)) \right|^2 \left( \| Y_2 \|_E + 1 \right) \right] < \infty. \)

Then: \( \sqrt{n} (\hat{\theta}_n - \theta_\alpha) \to N (0, V^{-1}), \) with \( V^{-1} \) given in (16) and \( D_{w^*} (X) \) given in (19).

This proposition demonstrates that even for a nonlinear IV model with unknown heteroscedastic variance, the SMD estimator \( \hat{\theta}_n \) is \( \sqrt{n} \) consistent under mild conditions.

5. COVARIANCE ESTIMATOR

To conduct statistical inference on the parametric component, a consistent estimate of the covariance matrix is needed. In this section, we provide one such estimator by consistently estimating each component of \( V^{-1} \) given in (16). Note that a consistent estimator of \( \Sigma (X) \) already exists. The true conditional covariance \( \Sigma_o (X) \equiv \text{var} [\rho (Z, \alpha) |X] \) can be estimated by regressing \( \rho (Z_i, \hat{\alpha}_n) \rho (Z_i, \hat{\alpha}_n)' \) on \( p^{k_m} (X_i) \). That is, for any \( X \in \mathcal{X}, \alpha \in \mathcal{A}_n \), define the following:
\[
\Sigma_o (X, \alpha) \equiv E \left[ \rho (Z, \alpha) \rho (Z, \alpha)' |X \right] \equiv \left\{ \sigma_{ojl} (X, \alpha) \right\}_{j,l=1,\ldots,d_p};
\]
\[
\hat{\Sigma}_o (X, \alpha) \equiv \left\{ \hat{\sigma}_{ojl} (X, \alpha) \right\}_{j,l=1,\ldots,d_p};
\]
\[
\hat{\sigma}_{ojl} (X, \alpha) \equiv \sum_{i=1}^n \rho_j (Z_i, \alpha) \rho_l (Z_i, \alpha) p^{k_m} (X_i)' (P' P)^{-1} p^{k_m} (X_i).
\]

Then our estimate of \( \Sigma_o (X) \) is given by \( \hat{\Sigma}_o (X) \equiv \hat{\Sigma}_o (X, \alpha_n). \)

To estimate \( D_{w^*} (X) \), we must first estimate \( w^* = (w^*_1, \ldots, w^*_{d_\theta}) \), defined in (15) above. For each component \( \theta_j, j = 1, \ldots, d_\theta \), we estimate \( w^*_j \) by \( \bar{w}_j^* \),
which is the solution to the minimization problem:

$$\min_{w_j \in \mathcal{W}_n} \sum_{i=1}^{n} \left( \frac{d \hat{m}(X_i, \hat{\alpha}_n)}{d \theta_j} - \frac{d \hat{m}(X_i, \hat{\alpha}_n)}{d h} [w_j] \right) \left( \hat{\Sigma}(X_i) \right)^{-1}$$

$$\times \left( \frac{d \hat{m}(X_i, \hat{\alpha}_n)}{d \theta_j} - \frac{d \hat{m}(X_i, \hat{\alpha}_n)}{d h} [w_j] \right)$$

where

$$\frac{d \hat{m}(X, \alpha)}{d \theta_j} - \frac{d \hat{m}(X, \alpha)}{d h} [w_j]$$

$$= \sum_{i=1}^{n} \left( \frac{d \rho(Z_i, \alpha)}{d \theta_j} - \frac{d \rho(Z_i, \alpha)}{d h} [w_j] \right) p^{kn}(X_i)'(P'P)^{-1}p^{kn}(X).$$

Finally, if we let \( \hat{w}^* = (w_1^*, \ldots, w_{d_\theta}^*) \), then \( D_{\hat{w}^*}(X) \) can be estimated by

$$\hat{D}_{\hat{w}^*}(X) = \frac{d \hat{m}(X, \hat{\alpha}_n)}{d \theta_j} - \frac{d \hat{m}(X, \hat{\alpha}_n)}{d h} [\hat{w}^*].$$

Thus, the estimator of \( V^{-1} \) is

$$\hat{V}^{-1} \equiv \left( \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{\hat{w}^*}(X_i) \hat{\Sigma}(X_i)^{-1} \hat{D}_{\hat{w}^*}(X_i) \right)^{-1}$$

$$\times \left( \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{\hat{w}^*}(X_i) \hat{\Sigma}(X_i)^{-1} \hat{\Sigma}_o(X_i) \hat{\Sigma}(X_i)^{-1} \hat{D}_{\hat{w}^*}(X_i) \right)^{-1}$$

We show in the Appendix D that the following additional conditions are sufficient for \( \hat{V} \) to be a consistent estimator of \( V \).

**Assumption 5.1:** Each element of \( \rho(Z, \alpha) \rho(Z, \alpha)' \) satisfies an envelope condition and is Hölder continuous in \( \alpha \in \mathcal{N}_o \).

**Assumption 5.2:** For \( j = 1, \ldots, d_\theta \),

$$\frac{d \rho(Z, \alpha)}{d \theta_j} - \frac{d \rho(Z, \alpha)}{d h} [w_j]$$

satisfies an envelope condition and is Hölder continuous in \( \alpha \in \mathcal{N}_o \) and \( w_j \in \{ v \in \mathcal{W} : \| v \|_s \leq c < \infty \} \).
THEOREM 5.1: Under Assumptions 3.1–3.9, 4.1–4.2, 5.1–5.2, we have: $\hat{\Sigma}^{-1} = V^{-1} + o_p(1)$.

Theorem 5.1 shows that the covariance estimator is consistent. It is worth noting that, for linear sieves, computing $w_\theta^T(\cdot)$ does not require nonlinear optimization. All that is needed is a pooled and weighted LS regression of the derivative of $\rho$ with respect to $\theta_j$ on the derivatives of $\rho$ with respect to $h$. Thus, for linear sieves, our covariance estimator is easy to compute.

6. EFFICIENCY

Having established the asymptotic distribution of $\hat{\theta}_n$, we now investigate its efficiency. Clearly, the efficiency depends on the choice of the weighting matrix $Z(X)$. It is straightforward to show that the optimal weighting matrix in the sense of minimizing the asymptotic covariance of the estimator $\hat{\theta}_n$ is $\Sigma(X) = \Sigma_o(X) = \text{var}(\rho(Z, \alpha_o)|X)$. Thus, when $\Sigma(X)$ is set to $\Sigma_o(X)$, the corresponding estimator is the best in the class of all SMD estimators, and will be called the optimally weighted SMD estimator. The asymptotic covariance of the optimally weighted SMD estimator is $V_o^{-1}$, where

$$V_o = E[D_{w_o}(X)'[\Sigma_o(X)]^{-1}D_{w_o}(X)]$$

and $w_o = (w_{o1}, \ldots, w_{odo})$ solves (16) with $\Sigma(X)$ replaced by $\Sigma_o(X)$.

The issue now is whether the optimally weighted SMD estimator is also the best in the class of all $\sqrt{n}$ consistent and regular estimators for model (1). To address this issue, we compute the efficiency bound for model (1). We follow the approach described in Newey (1990b) by first characterizing the tangent space and then computing the residuals of the projection of the score with respect to $\theta$ onto the tangent space.

Let $q_o(y, x, \alpha_o)$ denote the true joint density of $(Y, X)$. Since $A$ is convex at $\alpha_o$ by assumption, for each fixed $h \in \mathcal{H}$, $h_o + \xi(h - h_o) \in \mathcal{H}$ for sufficiently small constant $\xi > 0$. Hence, $p(y, x, \theta, \xi) = q_o(y, x, \theta, h_o + \xi(h - h_o))$ is a parametric submodel passing through $q_o(y, x, \alpha_o)$ at the true values $\theta = \theta_o$ and $\xi = 0$. The following theorem shows that $V_o$ is the efficiency bound.

THEOREM 6.1: For every fixed $h \in \mathcal{H}$, suppose that $p(y, x, \theta, \xi)$ is smooth in the sense of Newey (1990b, p. 127). Then $V_o$ is the semiparametric efficiency bound for the parametric component in model (1).

Theorem 6.1 presents two results. First, it shows that the optimally weighted SMD estimator is efficient. Second, it tells us that the $V_o$, as defined in (22), is the semiparametric efficiency bound for model (1). The first result provides an efficient estimator without explicitly computing the efficiency bound, which
is very useful since computing the efficiency bound is often difficult. The second result extends Chamberlain’s (1992) efficiency bound result to the more general setting of model (1).

To obtain a feasible efficient SMD estimator, we propose the following simple two-step procedure:

**Step 1**: Obtain an initial consistent SMD estimator $\hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n)$ by

$$
\min_{\alpha = (\theta, h) \in \Theta \times \mathcal{N}_n} \frac{1}{n} \sum_{i=1}^{n} \hat{m}(X_i, \alpha)' \hat{m}(X_i, \alpha)
$$

where $\hat{m}(X_i, \alpha)$ is given by (5).

**Step 2**: Obtain the optimally weighted SMD estimator $\tilde{\alpha}_n = (\tilde{\theta}_n, \tilde{h}_n)$ by

$$
\min_{\alpha = (\theta, h) \in \mathcal{N}_n} \frac{1}{n} \sum_{i=1}^{n} \hat{m}(X_i, \alpha)' (\hat{\Sigma}_o(X_i, \alpha))^{-1} \hat{m}(X_i, \alpha),
$$

with $\hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n)$ as starting point, where $\hat{\Sigma}_o(X_i, \alpha)$ is given by (21).

The second step is a locally continuously updated procedure. We now state conditions to ensure that the two-step procedure is efficient. For any fixed and nonzero $\lambda \in \mathcal{R}^{d_{\theta}}$, define

$$v_o \equiv (v_{o\theta}, v_{oh}), \quad v_{o\theta} = [E[D_{w_o}(X)\Sigma_o(X)^{-1}D_{w_o}(X)]]^{-1}\lambda,$$

$$v_{oh} = -w_o \times v_{o\theta}.$$

**Assumption 6.1**: (i) Each element of $\Sigma_o(\cdot, \alpha) \equiv E[\rho(Z, \alpha)\rho(Z, \alpha)' | X = x]$ is in $\Lambda_{\gamma}^{d}(\mathcal{X})$ with $\gamma > d_x/2$, for all $\alpha \in \mathcal{N}_n$; (ii) $\Sigma_o(X, \alpha)$ is positive definite uniformly over $X \in \mathcal{X}, \alpha \in \mathcal{N}_n$.

**Assumption 6.2**: Uniformly over $\alpha \in \mathcal{N}_n$,

$$E \left( \left\| \left\{ \frac{d m(X, \alpha)}{d \alpha} [\Pi_n v_o] \right\}' \Sigma_o(X, \alpha)^{-1} \right. \right.$$

$$\left. - \left\{ \frac{d m(X, \alpha)}{d \alpha} [\Pi_n v_o] \right\}' \Sigma_o(X, \alpha_o)^{-1} \right\|_{E}^{2} = o(n^{-1/2}).$$

**Assumption 6.3**: For all $\alpha \in \mathcal{N}_n$, each element of

$$\rho(Z, \alpha) \left\{ \frac{d^2 \rho(Z, \alpha + \tau \Pi_n v_o)}{d \tau^2} \bigg|_{\tau = 0} \right\}'$$

is bounded by measurable function $c_6(Z)$ with $E[c_6(Z)^2] < \infty$. 
Assumptions 6.1 and 5.1 together imply that \( \hat{\Sigma}_o(X, \alpha) \) converges to \( \Sigma_o(X, \alpha) \) faster than \( n^{-1/4} \) uniformly over \( X \in \mathcal{X}, \alpha \in \mathcal{A}_{on} \). Assumption 6.2 replaces Assumption 4.4, while Assumption 6.3 is slightly stronger than Assumption 4.6. These assumptions are easily satisfied when \( \hat{\Sigma}_o(X, \alpha) \) does not depend on \( \alpha \).

**THEOREM 6.2:** Under Assumptions 3.1–3.3, 3.5–3.9, 4.1–4.3, 4.5–4.6 with \( \Sigma(X) = \Sigma_o(X) \) and \( w^* = w_o \), 5.1 and 6.1–6.3, the two-step SMD estimator satisfies
\[
\sqrt{n}(\hat{\theta}_n - \theta_o) \longrightarrow N(0, V_o^{-1}),
\]
with \( V_o \) given in (22).

Hansen, Heaton, and Yaron (1996) and Newey and Smith (2000) have shown that for the parametric unconditional moment restrictions model \( E[p(Z, \theta_o)W(X)] = 0 \), continuously updated GMM procedures have better finite sample properties. One might expect the same result holds for the following continuously updated SMD procedure:

\[
\min_{\alpha=(\theta, h) \in \Theta \times \mathcal{A}_n} \frac{1}{n} \sum_{i=1}^{n} \hat{m}(X_i, \alpha)' \left( \hat{\Sigma}_o(X_i, \alpha) \right)^{-1} \hat{m}(X_i, \alpha),
\]

where \( \hat{m}(X_i, \alpha) \) and \( \hat{\Sigma}_o(X_i, \alpha) \) are given by (5) and (21), respectively. Under the additional condition that Assumptions 5.1 and 6.1 hold uniformly over \( \alpha \in \mathcal{A}_n \), we can easily establish that the procedure (23) produces an asymptotically efficient estimator for \( \theta_o \).

We conclude this section by applying Theorem 6.2 to the two examples discussed throughout the text. For Example 2.1, under homoskedastic variance \( \Sigma_o(X) = \Sigma_o \), the optimally weighted SMD estimator is the same as the identity weighted SMD estimator. Hence we obtain the following proposition.

**PROPOSITION 6.1:** Let \( \hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n) \) be the first step SMD estimator for Example 2.1. Suppose that all conditions of Proposition 4.1 hold. Then: \( \hat{\theta}_n \) is asymptotically efficient for \( \theta_o \) with asymptotic variance \( V_o^{-1} = V^{-1} \), where \( V \) is given in Proposition 4.1.

**PROPOSITION 6.2:** Let \( \hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n) \) be the second step SMD estimator for Example 2.2. Suppose that all conditions of Proposition 4.2 with \( \Sigma(X) = \Sigma_o(X) \) and \( w^* = w_o \) hold, and that the following are satisfied:

6.2.1. For all \( \alpha \in \mathcal{A}_{on} \), (i) \( E[(Y_1 - F(Y_2' \theta + \sum_{j=1}^{q} h_j(X_j))]^2 | X = \cdot) \) is in \( L^q(X) \); (ii) \( E[(Y_1 - F(Y_2' \theta + \sum_{j=1}^{q} h_j(X_j))]^2 | X = \cdot) \) is bounded away from zero;

6.2.2. (i) \( E[\sup_{\alpha \in \mathcal{A}_n} |Y_1 - F(Y_2' \theta + \sum_{j=1}^{q} h_j(X_j))]^2 | Y_2 | < \infty \); (ii) \( E[\sup_{\alpha \in \mathcal{A}_n} \| \nabla^2 F(Y_2' \theta + \sum_{j=1}^{q} h_j(X_j))]^2 | Y_2 \|_F + 1)^8 < \infty \).

Then: \( \sqrt{n}(\hat{\theta}_n - \theta_o) \longrightarrow N(0, V_o^{-1}), \) with
\[
V_o = E[D_{w_o}(X) \Sigma_o(X)^{-1} D_{w_o}(X)],
\]
\[
D_{w_o}(X) = E[\nabla F(Y_2' \theta_o + \sum_{i=1}^{q} h_{oi}(X_i)) \{ Y_2 - \sum_{i=1}^{q} w_{oi}(X_i) \} | X].
\]
Note that when $F()$ is the identity function, Example 2.2 becomes Chamberlain's (1992) model (3.5), and

$$V_o = \inf \mathbb{E} \left\{ \left( \mathbb{E}[Y_2^2|X] - \sum_{i=1}^{q} w'(X_i) \right)^2 \Sigma_o(X)^{-1} \times \left( \mathbb{E}[Y_2^2|X] - \sum_{i=1}^{q} w'(X_i) \right)^2 \right\},$$

which is the efficiency bound derived by Chamberlain for his model (3.5).

7. SIMULATION

To illustrate the implementation of the SMD procedure, and to evaluate the finite sample performance of the SMD estimator $\tilde{\alpha}_n$ (especially the parametric part $\tilde{\theta}_n$), we conduct a small scale simulation study. The simulation is based on a partially linear nonparametric IV regression: $Y_1 = X_1 \theta_o + h_o(Y_2) + U$ with true values $\theta_o = 1$ and $h_o(Y_2) = \exp(Y_2)/(1 + \exp(Y_2))$. We assume that $Y_2$ is given by $Y_2 = X_1 + X_2 + R \times U + \varepsilon$ with $R = 0.9$ or $0.1$. The regressors $X_1$ and $X_2$ are independent and uniformly distributed over $[0, 1]$. $\varepsilon$ is independent of $X = (X_1, X_2)'$ and has a standard normal distribution. Conditional on $X$, $U$ has a normal distribution with mean zero and variance $X_1^2 + X_2^2$. Thus, by design, the regressor $Y_2$ is endogenous and the error term $U$ is heteroscedastic. Conditional on $X$, the covariance between $Y_2$ and $U$ is $R(X_1^2 + X_2^2)$. Thus, $R = 0.9$ indicates a high level of correlation while $R = 0.1$ indicates a low level of correlation.

The SMD procedure requires approximating the nonparametric part $h_o(Y_2)$ and the conditional mean function $\mathbb{E}[Y_1 - X_1 \theta - h(Y_2)|X]$ by sieves. For the unknown $h_o(Y_2)$, since it is bounded between zero and one, we approximate it by a power series multiplied by the cumulative distribution function of a standard normal. However, note that the number of terms ($k_{1n}$) in the approximation is not uniquely determined by the sufficient conditions in Sections 3–6. Those conditions only suggest that $k_{1n}$ could be in the order of $n^{1/7}$. For a sample size of $n = 1000$, $n^{1/7}$ is roughly 4. Hence we approximate $h_o(Y_2)$ by

$$h_o(Y_2) \approx \Phi(Y_2)(\pi_o + \pi_1 Y_2 + \pi_2 Y_2^2 + \pi_3 Y_2^3)$$

where $\Phi(Y_2)$ denotes the standard normal cumulative distribution function. Note that the true value of $\pi_o$ is 1. Substituting for $h_o(Y_2)$ by the polynomial sieve, we have

$$Y_1 \approx X_1 \theta_o + \pi_0 \Phi(Y_2) + \pi_1 \Phi(Y_2) Y_2 + \pi_2 \Phi(Y_2) Y_2^2 + \pi_3 \Phi(Y_2) Y_2^3 + U.$$

Now treating the above equation as if it were the true model, a natural estimation method is the instrumental variable estimation method. This is exactly
what our procedure proposes to do. Since the error term \( U \) is orthogonal to
the regressors \( X \), we may use power series of \( X \) as instruments. Clearly the
number of instruments \( (k_r) \) must be at least 5 (the number of unknown coef-
ficients), but the exact number is undetermined. We use an ad-hoc rule that
the number of instruments is twice the number of unknown coefficients. This
leads to the following set of instruments: \( \{1, X_1, X_2, X_1^2, X_1X_2, X_1^3, X_1^2X_2,
X_1^3, X_2^3\} \), which is in fact a tensor product polynomial sieve used to approx-
imate the conditional mean function. This same set of instruments is also used
to estimate the optimal weighting matrix.

A sample of 1000 data points was generated according to the above de-
design. The proposed procedure was applied to the data twice: first with iden-
tity weighting matrix \( \Sigma(X) = I \) (corresponding to an inefficient estimator),
then with estimated optimal weighting matrix (corresponding to a feasible ef-
ficient estimator). The estimated coefficients were recorded. Then, a new sam-
ple of 1000 data points was drawn and the estimated coefficients were com-
puted again. This procedure was repeated 1000 times. The mean (MEAN),
standard deviation (STD), mean absolute deviation (MAD), mean squared er-
ror (MSE), and 25, 50, and 75 percentile of the \( \theta_o \) estimator are reported in
Table I. To evaluate the performance of the proposed SMD estimator of the
nonparametric component, we also compute and report the integrated mean
squared error and the integrated mean absolute deviation. The integrated
mean squared error (IMSE) is computed according to the following discrete
expression: \( 0.01 \sum_{j=0}^{399} \text{mean}\{[h_o(-2 + 0.01j) - \hat{h}(-2 + 0.01j)]^2\} \), where \( \text{mean}\{\cdot\} \)
denotes the average over all 1000 SMD estimators \( \hat{h} \). The integrated mean ab-
solute deviation (IMAD) is computed according to: \( 0.01 \sum_{j=0}^{399} \text{mean}\{|h_o(-2 +
0.01j) - \hat{h}(-2 + 0.01j)|\} \).

Since a power series might have erratic tail behaviors, the above power se-
ries may not be a good sieve for the unknown function \( h_o(Y_2) \), which has un-
bounded support. We also used a spline sieve for \( h_o() \). The spline we used
is some constant over \((-\infty, -2)\), a third degree polynomial over \([-2, 2]\), and
some other constant over \((2, +\infty)\). The spline was chosen so that it is continu-
ous over the entire real line. The set of instruments is the same as before. The
results are also reported in Table I.

We have the following findings in Table I. First, both the inefficient and fea-
sible efficient SMD estimators of the parametric component \( \theta_o \) are mean and
median biased, and the mean biases are within the standard deviation of the
true value. The biases depend on the correlation parameter \( (R) \) between the
endogenous regressor and the error term. When there is small correlation be-
tween \( Y_2 \) and \( U \) (i.e., when \( R = 0.1 \)), both estimators are essentially unbiased.
Second, the feasible efficient SMD estimator of \( \theta_o \) performs slightly better than
the inefficient SMD estimator. The bias of the efficient SMD estimator is gen-
erally smaller than the bias of the inefficient SMD estimator, and the standard
deviation of the efficient estimator is generally about 30 percent lower than
TABLE I

FINITE SAMPLE PERFORMANCE OF THE SMD ESTIMATOR

<table>
<thead>
<tr>
<th>Estimator</th>
<th>MEAN</th>
<th>STD</th>
<th>MAD</th>
<th>MSE</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>IMSE</th>
<th>IMAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( R = 0.9, ) Power Series</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inefficient SMD Estimator of ( \theta )</td>
<td>0.964</td>
<td>0.140</td>
<td>0.108</td>
<td>0.145</td>
<td>0.879</td>
<td>0.962</td>
<td>1.049</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficient SMD Estimator of ( \theta )</td>
<td>0.979</td>
<td>0.104</td>
<td>0.089</td>
<td>0.116</td>
<td>0.903</td>
<td>0.975</td>
<td>1.051</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inefficient SMD Estimator of ( h ):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficient SMD Estimator of ( h ):</td>
<td>0.478</td>
<td>0.360</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( R = 0.9, ) Splines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inefficient SMD Estimator of ( \theta )</td>
<td>0.965</td>
<td>0.122</td>
<td>0.098</td>
<td>0.127</td>
<td>0.879</td>
<td>0.960</td>
<td>1.046</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficient SMD Estimator of ( \theta )</td>
<td>1.016</td>
<td>0.109</td>
<td>0.084</td>
<td>0.110</td>
<td>0.944</td>
<td>1.014</td>
<td>1.083</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inefficient SMD Estimator of ( h ):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficient SMD Estimator of ( h ):</td>
<td>0.081</td>
<td>0.469</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( R = 0.1, ) Power Series</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inefficient SMD Estimator of ( \theta )</td>
<td>0.993</td>
<td>0.141</td>
<td>0.108</td>
<td>0.141</td>
<td>0.907</td>
<td>0.991</td>
<td>1.083</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficient SMD Estimator of ( \theta )</td>
<td>0.998</td>
<td>0.110</td>
<td>0.086</td>
<td>0.110</td>
<td>0.926</td>
<td>0.997</td>
<td>1.067</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inefficient SMD Estimator of ( h ):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficient SMD Estimator of ( h ):</td>
<td>0.503</td>
<td>1.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( R = 0.1, ) Splines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inefficient SMD Estimator of ( \theta )</td>
<td>0.999</td>
<td>0.127</td>
<td>0.101</td>
<td>0.127</td>
<td>0.918</td>
<td>0.993</td>
<td>1.081</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficient SMD Estimator of ( \theta )</td>
<td>1.014</td>
<td>0.085</td>
<td>0.067</td>
<td>0.087</td>
<td>0.959</td>
<td>1.016</td>
<td>1.070</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inefficient SMD Estimator of ( h ):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficient SMD Estimator of ( h ):</td>
<td>0.157</td>
<td>0.722</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

the standard deviation of the inefficient estimator. Third, the choice of sieves (spline or power series) for \( h_0 \) seems to have little impact on the performance of the SMD estimators of \( \theta_\). Fourth, the SMD estimator of the nonparametric component appears to fit the true function \( h_0 \) well. However, in this case the choice of estimation method and sieves seems to have a big impact on the performance of the SMD estimator of \( h_0 \). For instance, the feasible efficient estimation method reduces the IMSE by about 60 to 90 percent, whereas the spline sieve for \( h_0 \) reduces it by about 80 percent.

8. CONCLUSION

In this paper, we propose a SMD estimator for the general semiparametric conditional moments restrictions (1). We present convergence rates for the SMD estimator of the nonparametric component. We also derive the \( \sqrt{n} \) asymptotic normality of the SMD estimator of the parametric component and provide a consistent estimator of its asymptotic covariance matrix. We show that the optimally weighted SMD estimator of the parametric component can attain the semiparametric efficiency bound for model (1), and provide a two-step procedure to obtain a feasible efficient estimator. In addition, we provide two examples and a small scale Monte Carlo experiment to show that our SMD procedure has both theoretical and practical values.
Our theoretical results can be extended in two directions. First, model (1) assumes all moment restrictions hold conditional on the same set of regressors. This assumption may not hold in some applications (see Newey, Powell, and Vella (1999), for example). Typically, in those applications a different set of moment restrictions holds conditional on a different set of regressors. Our procedure and results need to be extended to models of this sort. Second, although our results on convergence rates and $\sqrt{n}$ asymptotic normality can be easily extended to weakly dependent time series data, the problem of a semiparametric efficiency bound with time series data is nontrivial. We plan to address this issue by extending the work of Hansen (1985) in a separate paper.

Dept. of Economics, University of Florida, Gainesville, FL 32611, U.S.A.; aic@ufl.edu

and

Dept. of Economics, New York University, 269 Mercer Street, 7th Floor, New York, NY 10003, U.S.A.; xiaohong.chen@nyu.edu.

Manuscript received August, 1999; final revision received April, 2003.

APPENDIX A: USEFUL UNIFORM CONVERGENCE RESULTS

Recall that $\overrightarrow{A}$ denotes the linear completion of $A$ under the metric $\| \cdot \|$, $A_n$ denotes a sieve approximation of $A$, and $N(\varepsilon, A_n, \| \cdot \|_i)$ denotes the minimal number of $\varepsilon$-radius covering balls of $A_n$ under the metric $\| \cdot \|$. Also, $\| \cdot \|_E$ denotes the Euclidean norm, $\xi_0 \equiv \sup_{X \in \mathcal{X}} \| p^{k_n}(X)\|_E$, and

$$\xi_{1n} \equiv \sup_{x \in \mathcal{X}} \left\| \frac{\partial p^{k_n}(x)}{\partial \alpha} \right\|_E.$$

In the following lemma, $g : \mathcal{Z} \times \overrightarrow{A} \to \mathcal{R}$ denotes a generic measurable function of the data $Z \in \mathcal{Z}$ and the parameter $\alpha \in \overrightarrow{A}$. Define $\varepsilon(Z, \alpha) = g(Z, \alpha) - \mathbb{E}[g(Z, \alpha)|X]$ and $\varepsilon(\alpha) = (\varepsilon(Z_1, \alpha), \ldots, \varepsilon(Z_n, \alpha))'$.

**LEMMA A.1:** (A) Suppose that Assumptions 3.1, 3.2(i), and the following are satisfied:

(i) there exist a constant $c_1$, and a measurable function $c_1(Z) : \mathcal{Z} \to [0, \infty)$ with $E[c_1(Z)] < \infty$ for some $p \geq 4$ such that $|g(Z, \alpha)| \leq c_1 c_1(Z)$ for all $\alpha \in A_n$ and $Z \in \mathcal{Z}$;

(ii) there exist a constant $K \in (0, 1]$ and a measurable function $c_2(Z) : \mathcal{Z} \to [0, \infty)$ with $E[c_2(Z)^2] < \infty$ such that $|g(Z, \alpha_1) - g(Z, \alpha_2)| \leq c_2(Z) \| \alpha_1 - \alpha_2 \|_E^p$ holds for all $Z \in \mathcal{Z}$ and $\alpha_1, \alpha_2 \in A_n$;

(iii) there exists a positive value $\delta_{1n} = o(1)$ such that

$$\frac{n \delta_{1n}^2}{\ln \left( \frac{\delta_{1n}}{\delta_{0n}} \right)^4} N \left( \left\{ \frac{\delta_{1n}}{\delta_{0n}} \right\}^{1/4}, A_n, \| \cdot \|_i \right) \times \max \left\{ \xi_0, \xi_{1n}^2, \xi_{2n}^{2+2/p}, \xi_{3n}^{2+2/p}, \xi_{4n}^{1+2/p} \right\} \to +\infty.$$

Then $p^{k_n}(X)'(P^t P)^{-1} P^t \varepsilon(\delta_{1n}) = o_p(\delta_{1n})$ uniformly over $(X, \alpha) \in \mathcal{X} \times A_n$.

(B) Suppose that Assumptions 3.1, 3.2(i), and the following are satisfied:

(iv) there exist a positive value $\delta_{2n} = o(1)$ and coefficients $\pi(\alpha)$ such that $E[g(Z, \alpha)|X] = p^{k_n}(X)' \pi(\alpha) + o(\delta_{2n})$ holds for all $X \in \mathcal{X}$ and $\alpha \in A_n$. 


Then \( (1/n) \sum_{i=1}^{n} (P^n_i (X_i)'(PP^{-1}) \sum_{j=1}^{n} P^n_j (X_j) E[\varepsilon(Z, \alpha) | X_j] - E[\varepsilon(Z, \alpha) | X_j] \) = \( O_p(\delta_{2n}^2) \) uniformly over \( \alpha \in A_n \).

\begin{align*}
(\text{C}) \text{ Suppose that Assumptions 3.1, 3.2(i) and the following are satisfied:} \\
\text{(v) For fixed } \overline{\alpha_n} \in \overline{A}, \text{ } E[\varepsilon(Z, \overline{\alpha})^2 | X] \leq \text{const for all } n \geq 1 \text{ and } X \in \mathcal{X}.
\end{align*}

Then \( (1/n) \sum_{i=1}^{n} (P^n_i (X_i)'(PP^{-1}) \sum_{j=1}^{n} P^n_j (X_j) e(Z, \overline{\alpha}) \) = \( O_p(k_n/n) \).

\text{PROOF (A): Let } c \text{ denote a generic constant that may have different values in different expressions. Let } W_n = \mathcal{X} \times A_n. \text{ For any pair } (X^1, \alpha^1) \in W_n \text{ and } (X^2, \alpha^2) \in W_n, \text{ we write}

\begin{align*}
&\sup_{X \in \mathcal{X}, \alpha \in A_n} \|p^n_i (X_i)'(PP^{-1}) \sum_{j=1}^{n} P^n_j (X_j) e(Z_i, \alpha) \|_E \\
&\leq \frac{1}{n \lambda_n} \sum_{i=1}^{n} \|c_1(Z_i) + E[c_1(Z_i) | X_i] \|^2.
\end{align*}

\text{Assumption 3.2(i) implies } \lambda_n = O_p(1) \text{ (see p. 162 of Newey (1997)). Applying the weak law of large numbers and } E[\varepsilon(Z_i | X_i)] = \varepsilon(Z_i), \text{ we obtain}

\begin{align*}
&\left( 1/n \lambda_n \right) \sum_{i=1}^{n} \|c_1(Z_i) + E[c_1(Z_i) | X_i] \|^2 = O_p(1). \\
\text{Thus, there exists a constant } c \text{ such that}
\end{align*}

\( \sup_{X \in \mathcal{X}, \alpha \in A_n} \|p^n_i (X_i)'(PP^{-1}) \sum_{j=1}^{n} P^n_j (X_j) e(Z_i, \alpha) \|_E \leq \varepsilon_{0n} \|c_1(Z_i) + E[c_1(Z_i) | X_i] \|^2 \)

holds for any small \( \eta > 0 \) and sufficiently large \( n \). Condition (ii) implies

\begin{align*}
\left( 1/(n \lambda_n) \right) \sum_{i=1}^{n} \|c_2(Z_i) \|^2 = O_p(1). \text{ It follows that for some constant } c:
\end{align*}

\( \sup_{X \in \mathcal{X}, \alpha \in A_n} \left\{ \left( \|p^n_i (X_i)'(PP^{-1}) \sum_{j=1}^{n} P^n_j (X_j) e(Z_i, \alpha) \|_E \right) / \|c_1(Z_i) + E[c_1(Z_i) | X_i] \|^2 \right\} > c \varepsilon_{0n} \) < \( \eta \)

for sufficiently large \( n \).

For any small \( \varepsilon \), partition \( W_n \) into \( b_n \) mutually exclusive subsets \( W_{nm}, m = 1, 2, \ldots, b_n \), where \( (X^1, \alpha^1) \in W_{nm} \) and \( (X^2, \alpha^2) \in W_{nm} \) imply \( \|X^1 - X^2\|_E \leq \varepsilon_{0m}/(c_{1n} \varepsilon_{0n} c) \) and \( \|\alpha^2 - \alpha^1\|_E \leq \varepsilon_{0m}/(\varepsilon_{0n} c) \). Then with probability approaching one, we have

\begin{align*}
\left| p^n_i (X_i)'(PP^{-1}) \sum_{j=1}^{n} P^n_j (X_j) e(Z_i, \alpha) - p^n_i (X_i)'(PP^{-1}) \sum_{j=1}^{n} P^n_j (X_j) e(Z_i, \alpha) \right| \leq 2 \varepsilon \delta_{1n}.
\end{align*}

Let \( (X^m, \alpha^m) \) denote a fixed point in \( W_{nm} \). For any \( (X, \alpha) \), there exists an \( m \) such that \( \|X - X_m\|_E \leq \varepsilon_{1n}/(c_{1n} \varepsilon_{0n} c) \) and \( \|\alpha - \alpha^m\|_E \leq \varepsilon_{0m}/(\varepsilon_{0n} c) \). Then, with probability approaching one, \( \sup_{(X, \alpha) \in \mathcal{X} \times A_n} \left| p^n_i (X_i)'(PP^{-1}) \sum_{j=1}^{n} P^n_j (X_j) e(Z_i, \alpha) \right| \leq 2 \varepsilon \delta_{1n} + \max_m \left| p^n_i (X^m_i)'(PP^{-1}) \sum_{j=1}^{n} P^n_j (X^m_j) e(Z_i, \alpha^m) \right| \).
Hence
\[
P\left( \sup_{(X, \alpha) \in X \times A_n} |p^k(X)'(P'P)^{-1}P'\varepsilon(\alpha)| > 4\varepsilon \right)\
< 2\eta + P\left( \max_k |p^k(X)^m)'(P'P)^{-1}P'\varepsilon(\alpha^m)| > 2\varepsilon \right).
\]

For some constant \(c\), let
\[
M_n = \left( \frac{c \xi_{0n} c_n}{\delta_{1n}} \right)^{2/p}.
\]

Define \(d_{in} = 1[c_1(Z_i) \leq M_n]\). Define \(g_1(Z_i, \alpha) = d_{in}g(Z_i, \alpha)\) and \(g_2(Z_i, \alpha) = (1 - d_{in})g(Z_i, \alpha)\). Define \(\varepsilon_1(Z_i, \alpha)\) and \(\varepsilon_2(Z_i, \alpha)\) accordingly. It follows that
\[
P\left( \max_m |p^k(X)^m)'(P'P)^{-1}P'\varepsilon(\alpha^m)| > 2\varepsilon \right)
\leq P\left( \max_m |p^k(X)^m)'(P'P)^{-1} \sum_{i=1}^n p^k(X_i)e_1(Z_i, \alpha^m)| > \varepsilon \right)
+ P\left( \max_m |p^k(X)^m)'(P'P)^{-1} \sum_{i=1}^n p^k(X_i)e_2(Z_i, \alpha^m)| > \varepsilon \right)
= P_1 + P_2.
\]

Applying the Markov inequality yields
\[
P_2 \leq \frac{E[\max_m |p^k(X)^m)'(P'P)^{-1} \sum_{i=1}^n p^k(X_i)e_2(Z_i, \alpha^m)|]}{\varepsilon \delta_{1n}}.
\]

Note that
\[
\frac{|\varepsilon_2(Z_i, \alpha^m)|}{c_{in}} \leq (1 - d_{in})c_1(Z_i) + E[(1 - d_{in})c_1(Z_i)|X_i]
\leq (1 - d_{in})c_1(Z_i) + \sqrt{E[(1 - d_{in})^2|X_i]|E[c_1(Z_i)^2|X_i]].
\]

Then
\[
|p^k(X)^m)'(P'P)^{-1} \sum_{i=1}^n p^k(X_i)e_2(Z_i, \alpha^m)|
\leq \frac{c_{in} \xi_{0n}}{\lambda_n} \sqrt{\frac{1}{n} \sum_{i=1}^n |\varepsilon_2(Z_i, \alpha^m)|^2}
\leq 2\frac{c_{in} \xi_{0n}}{\lambda_n} \sqrt{\frac{1}{n} \sum_{i=1}^n (1 - d_{in})c_1(Z_i)^2} + \frac{2c_{in} \xi_{0n}}{\lambda_n} \sqrt{\frac{1}{n} \sum_{i=1}^n E[(1 - d_{in})^2|X_i]E[c_1(Z_i)^2|X_i]].
\]

Applying the Markov inequality yields
\[
\frac{1}{n} \sum_{i=1}^n (1 - d_{in})c_1(Z_i)^2 \leq \frac{1}{n} \sum_{i=1}^n (1 - d_{in}) \sqrt{\frac{1}{n} \sum_{i=1}^n c_1(Z_i)^4} = O_p\left( \frac{1}{M_n^p} \right).
\]
Using $E[E[u|X_i]^2] \leq E[u^2]$ for any random variable $u$, we have

$$\frac{1}{n} \sum_{i=1}^{n} E[(1 - d_{in})|X_i]E[c_1(Z_i)^2|X_i] \leq O_P\left(\frac{1}{M_n^p}\right).$$

Combining the above results proves

$$P_2 \leq \frac{c \xi_{0n}^2 c_1}{M_n^{p/2} \epsilon \delta_{1n}} = \eta.$$

Some calculations yield

$$\sigma^2_n \equiv nE\left[\left(\sum_{i=1}^{n} p^{k_n}(X_i^m)\right)^2 \right] = O(c_1^2 \xi_{0n}^2) \quad \text{and} \quad \left|p^{k_n}(X_i^m)\right|^{(P'P/n)^{-1}} \leq \frac{M_n \xi_{0n}^2 c_1}{\lambda_n}.$$

Note that

$$P\left(\left|\sum_{i=1}^{n} p^{k_n}(X_i^m)\epsilon_1(Z_i, \alpha^m)\right| > \epsilon \delta_{1n}\right) = E \left[P\left(\left|\sum_{i=1}^{n} p^{k_n}(X_i^m)\epsilon_1(Z_i, \alpha^m)\right| > \epsilon \delta_{1n} | X_1, \ldots, X_n\right)\right].$$

Applying the Bernstein inequality for independent processes, we obtain

$$P\left(\left|\sum_{i=1}^{n} p^{k_n}(X_i^m)\epsilon_1(Z_i, \alpha^m)\right| > \epsilon \delta_{1n}\right) \leq 2E\left[\exp\left(-n\epsilon^2 \delta_{1n}^2/(c \sigma^2_n + M_n \xi_{0n}^2 c_1 \epsilon \delta_n)\right)\right].$$

where $E[.]$ is taken with respect to the joint distribution of $(X_1, \ldots, X_n)$. Hence,

$$P_1 < 2b_n E\left[\exp\left(-n\epsilon^2 \delta_{1n}^2/(c \sigma^2_n + M_n \xi_{0n}^2 c_1 \epsilon \delta_n)\right)\right],$$

which is arbitrarily small if

$$\frac{n \delta_{1n}^2}{\max\{\xi_{0n}^2 c_1^2, M_n \xi_{0n}^2 c_1 \delta_{1n}\}} - \ln(b_n) \to +\infty.$$

Since $\mathcal{X}$ is a compact subset in $\mathcal{R}^{d_x}$, we have

$$b_n = O\left(\frac{\delta_{1n}}{c_1 \xi_{1n}}\right)^{-d_x} \times N\left(\frac{\delta_{1n}}{\xi_{0n}}^{1/\kappa}, \mathcal{A}_n, \|\cdot\|_3\right).$$

Substituting for $M_n$ and $b_n$, we obtain

$$\frac{n \delta_{1n}^2}{\ln(b_n) \max\{\xi_{0n}^2 c_1^2, M_n \xi_{0n}^2 c_1 \delta_{1n}\}} = O\left(\frac{n \delta_{1n}^2}{\ln\left(\left(\frac{\delta_{1n}}{\xi_{1n}}\right)^{-d_x} \times N\left(\left(\frac{\delta_{1n}}{\xi_{0n}}\right)^{1/\kappa}, \mathcal{A}_n, \|\cdot\|_3\right)\right) \max\{\xi_{0n}^2 c_1^2, \xi_{0n}^2 \delta_{1n}^{1-2/p} c_1^{1+2/p}\}}\right).$$

Thus, $P_1 < \eta$ for sufficiently large $n$ by condition (iii).
PROOF (B): It follows from condition (iv) and

\[ \frac{1}{n} \sum_{i=1}^{n} \left( p^{k_{x}}(X_{i})'(P')^{-1} \sum_{j=1}^{n} p^{k_{x}}(X_{j})E[g(Z, \alpha)|X_{i}] - E[g(Z, \alpha)|X_{i}] \right)^2 \]

\[ \leq \frac{1}{n} \sum_{i=1}^{n} (E[g(Z, \alpha)|X_{i}] - p^{k_{x}}(X_{i})'(\alpha)) \right)^2 + o_{p}(\delta^2_{2n}). \]

PROOF (C): Note that

\[ E\left[ \frac{1}{n} \sum_{i=1}^{n} \left( p^{k_{x}}(X_{i})'(P')^{-1} \sum_{j=1}^{n} p^{k_{x}}(X_{j})e(Z_j, \overline{\alpha}_n) \right)^2 \right] \]

\[ = E\left[ \frac{1}{n} \sum_{i=1}^{n} p^{k_{x}}(X_{i})'(P')^{-1} p^{k_{x}}(X_{i}) \cdot E[e(Z, \overline{\alpha}_n)^2|X_{i}] \right] \]

\[ \leq E\left[ \frac{c}{n} \sum_{i=1}^{n} p^{k_{x}}(X_{i})'(P')^{-1} p^{k_{x}}(X_{i}) \right] = O\left( \frac{k_{n}}{n} \right). \]

Part (C) now follows from applying the Markov inequality. Q.E.D.

Corollary A.1: (i) Under Assumptions 3.1-3.2, 3.6-3.8, \( (1/n) \sum_{i=1}^{n} \| m(X_i, \alpha) - m(X_i, \alpha_0) \|_E^2 = o_p(n^{-1/2}) \) uniformly over \( \alpha \in A_n \); (ii) under Assumptions 3.1-3.3 and 4.1, \( (1/n) \sum_{i=1}^{n} \| m(X_i, \alpha_0) \|_E^2 = O_p(k_n/n) \).

PROOF: For part (i), we apply parts (A) and (B) of Lemma A.1 to each element of \( \rho(Z, \alpha) \) with \( c_{in} = 1, \ p = 4, \) and \( \delta_{in} = \delta_{in} = n^{-1/4}. \) Note that conditions (i), (ii), and (iv) are satisfied by Assumptions 3.1-3.2 and 3.6, while condition (iii) becomes \( \ln[\{n^{1/4} \xi_{in}\}] + \ln[\{n^{1/4} \xi_{in}\}^{-1/\alpha}, A_n, \| \cdot \|_2] \times \delta_{in}^2 \times n^{-1/2} = o(1) \), which is satisfied by Assumptions 3.7(ii) and 3.8.

Part (ii) follows from applying part (C) of Lemma A.1 to each element of \( \rho(Z, \alpha_0) \). Q.E.D.

Corollary A.2: (i) Under Assumptions 3.1-3.3, 3.4(ii), 3.6(iii), and 3.9, we obtain uniformly over \( \alpha \in A_n : \| \alpha - \alpha_0 \| = o(1) \): \( (1/n) \sum_{i=1}^{n} \| m(X_i, \alpha) \|_E^2 - E[\| m(X, \alpha) \|_E^2] = o_p(n^{-1/2}) \).

(ii) Under Assumptions 3.1-3.4, 3.6-3.9, we obtain uniformly over \( \alpha \in A_n, \| \alpha - \alpha_0 \| \leq o(\eta_n) : (1/n) \sum_{i=1}^{n} \| m(X_i, \alpha) \|_E^2 = O_p(\eta_n^2) \) and \( (1/n) \sum_{i=1}^{n} \| m(X_i, \alpha) \|_E^2 = o_p(\eta_n^2) \), where \( \eta_n = n^{\tau - 1/4} \) with \( \tau \leq 1/4 \).

PROOF: (i) Since \( m(X_i, \alpha_0) = 0 \) for all \( X_i \in X \) and \( \| \alpha - \alpha_0 \| = o(1) \), by Lemma 1 in Chen, Linton, and van Keilegom (2003), the result follows after we establish that (i.1) \( \| m(X, \alpha) \|_E^2 : \alpha \in A_n \) is a Donsker class; and (i.2) \( E[\| m(X, \alpha) \|_E^2 - \| m(X, \alpha_0) \|_E^2] \rightarrow 0 \) as \( \| \alpha - \alpha_0 \| \rightarrow 0 \).

Note that Assumptions 3.4(ii) and 3.9 imply that \( E[\| m(X, \alpha) \|_E^2] \) and \( \| \alpha - \alpha_0 \|_2 \) are equivalent. By Assumptions 3.3, 3.9, and 3.6(iii), we have

\[ E\left[ \| m(X, \alpha) \|_E^2 \right] \leq E\left[ \| m(X, \alpha) \|_E^2 \right] \times \sup_{X, \alpha} \| m(X, \alpha) \|_E^2 \leq \text{const.} \times \| \alpha - \alpha_0 \|_2. \]
Hence (i.2) is satisfied. Assumption 3.6(iii) implies \( \|m(X, \alpha)\|_{L^2}^2 : \alpha \in \mathcal{A}_n \) is a subset of \( A_2(X) \) and \( A_2'(X) \) is a Donsker class by Theorem 2.5.6 of van der Vaart and Wellner (1996). Hence we obtain (i.1).

Part (ii) follows from Corollaries A.1(i) and A.2(i), and \( E[\|m(X, \alpha)\|_{L^2}^2] = o(\eta_n^2) \) by Assumptions 3.4(ii) and 3.9.

Q.E.D.

APPENDIX B: CONVERGENCE RATE FOR SIEVE ESTIMATOR

Let

\[
\hat{L}_n(\alpha) \equiv \frac{-1}{2n} \sum_{i=1}^{n} \hat{m}(X_i, \alpha)[\hat{\Sigma}(X_i)]^{-1}\hat{m}(X_i, \alpha) \quad \text{and}
\]

\[
L_n(\alpha) \equiv \frac{-1}{2n} \sum_{i=1}^{n} m(X_i, \alpha)[\Sigma(X_i)]^{-1}m(X_i, \alpha).
\]

We need the following result to prove Theorem 3.1.

COROLLARY B.1: Assumptions 3.1–3.4, and 3.6–3.9 imply: (i) \( \hat{L}_n(\alpha) - L_n(\alpha) = o_p(n^{-1/4}) \) uniformly over \( \alpha \in \mathcal{A}_n \); and (ii) \( \hat{L}_n(\alpha) - \hat{L}_n(\alpha_o) - (L_n(\alpha) - L_n(\alpha_o)) = o_p(\eta_n n^{-1/4}) \) uniformly over \( \alpha \in \mathcal{A}_n \) with \( \|\alpha - \alpha_o\| \leq o(\eta_n) \), where \( \eta_n = n^{-\tau} \) with \( \tau \leq 1/4 \).

PROOF: Part (i) follows from applying Corollary A.1(i) and Assumption 3.4. Part (ii) follows from applying Corollary A.1 and Corollary A.2(ii).

Q.E.D.

PROOF OF THEOREM 3.1: Let \( \iota = 1/4, \eta_{0n} = o(n^{-\iota}), \) and \( \delta_{0n} = 2\sqrt{\eta_{0n}} = o(n^{-1/2}). \) To simplify notation, denote \( \alpha_{no} = \Pi_\alpha \alpha_o. \) For any \( k \geq 1, \)

\[
P(\|\hat{\alpha}_n - \alpha_o\| \geq k \delta_n) \leq P \left( \sup_{(\|\alpha - \alpha_o\| \geq k \delta_{0n}, \alpha \in \mathcal{A}_n)} \hat{L}_n(\alpha) \geq \hat{L}_n(\alpha_{no}) \right)
\]

\[
\leq P \left( \sup_{(\alpha \in \mathcal{A}_n)} |\hat{L}_n(\alpha) - L_n(\alpha)| > \eta_{0n} \right)
\]

\[
+ P \left( \sup_{(\alpha \in \mathcal{A}_n)} |\hat{L}_n(\alpha) - L_n(\alpha)| \leq \eta_{0n} \right) \cap \left\{ \sup_{(\|\alpha - \alpha_o\| \geq k \delta_{0n}, \alpha \in \mathcal{A}_n)} \hat{L}_n(\alpha) \geq \hat{L}_n(\alpha_{no}) \right\}
\]

\[
\leq P \left( \sup_{(\alpha \in \mathcal{A}_n)} |\hat{L}_n(\alpha) - L_n(\alpha)| > \eta_{0n} \right)
\]

\[
+ P \left( \sup_{(\|\alpha - \alpha_o\| \geq k \delta_{0n}, \alpha \in \mathcal{A}_n)} L_n(\alpha) \geq L_n(\alpha_{no}) - 2\eta_{0n} \right)
\]

\[
\equiv P_1 + P_2.
\]

Corollary B.1(i) implies \( P_1 \to 0 \) as \( n \to \infty. \) To show \( P_2 \to 0, \) we notice that all conditions for Theorem 1 in Shen and Wong (1994) are trivially satisfied by our assumptions. Hence \( P_2 \to 0 \) by their Theorem 1. This proves: \( \|\hat{\alpha}_n - \alpha_o\| = o_p(n^{-1/2}). \).
Next, we refine the convergence rate by exploiting the local curvature of \( \hat{L}_n(\alpha) \) around \( \alpha \). Let 
\[ \eta_{1n} = n^{-1/4}\delta_{0n} = o(n^{-i(i+1/2)}) \text{ and } \delta_{1n} = 2/\eta_{1n} = o(n^{-i(i+1/2)}). \]
For any fixed \( k > 1 \), we have
\[
P(\|\hat{a}_n - \alpha_o\| \geq k\delta_{1n}) \leq P\left( \sup_{\delta_{0n} \geq 1-n-\alpha_o \| \geq k\delta_{1n}} (\hat{L}_n(\alpha) - \hat{L}_n(\alpha_o)) \geq L_n(\alpha) - \hat{L}_n(\alpha_o) \right) \leq P\left( \sup_{\delta_{0n} \geq 1-n-\alpha_o \| \geq k\delta_{1n}} |\hat{L}_n(\alpha) - \hat{L}_n(\alpha_o) - (L_n(\alpha) - L(\alpha_o))| > \eta_{1n} \right)
\]
\[ + P\left( \sup_{\delta_{0n} \geq 1-n-\alpha_o \| \geq k\delta_{1n}} |\hat{L}_n(\alpha) - \hat{L}_n(\alpha_o) - (L_n(\alpha) - L(\alpha_o))| \leq \eta_{1n} \right) \cap \left\{ \sup_{\delta_{0n} \geq 1-n-\alpha_o \| \geq k\delta_{1n}} (\hat{L}_n(\alpha) - \hat{L}_n(\alpha_o)) \geq L_n(\alpha_o) - \hat{L}_n(\alpha_o) \right\} \]
\[ \leq P\left( \sup_{\delta_{0n} \geq 1-n-\alpha_o \| \geq k\delta_{1n}} |\hat{L}_n(\alpha) - \hat{L}_n(\alpha_o) - (L_n(\alpha) - L(\alpha_o))| > \eta_{1n} \right)
\]
\[ + P\left( \sup_{\delta_{0n} \geq 1-n-\alpha_o \| \geq k\delta_{1n}} L_n(\alpha) - L_n(\alpha_o) \geq L_n(\alpha_o) - L(\alpha_o) - 2\eta_{1n} \right) \leq P_3 + P_4. \]

Corollary B.1(ii) implies

\[ \hat{L}_n(\alpha) - \hat{L}_n(\alpha_o) - (L_n(\alpha) - L_n(\alpha_o)) = o_p(n^{-1/4}\delta_{0n}) = o_p(\eta_{1n}), \]

hence \( P_3 \to 0 \) as \( n \to \infty \). Applying Theorem 1 of Shen and Wong (1994) again yields \( P_4 \to 0 \) as \( n \to \infty \). This proves \( \|\hat{a}_n - \alpha_o\| = o_p(n^{-i(i+1/2)}). \)

Repeating the above proof an infinite number of times, we obtain \( \|\hat{a}_n - \alpha_o\| = o_p(n^{-i(i+1/2+2+\cdots)}) = o_p(n^{-1/4}). \) This completes the proof.

Q.E.D.

Let
\[ \|h - h_o\|^2 \equiv E\left\| \frac{dm(X,\alpha_o)}{dh} [h - h_o] \right\|^2_E. \]

Recall the notations \( w^* \) and \( \bar{V} \) defined in Section 4.

**Lemma B.1:** Under Assumption 3.4(ii), if \( E[D_{w^*}(X)'D_{w^*}(X)] \) is positive definite, then \( \|\hat{\theta}_n - \theta_o\| = O_p(\|\hat{a}_n - \alpha_o\|) \); if further
\[
\text{tr}\left[ E\left\{ \left( \frac{dm(X,\alpha_o)}{dh} [w^*] \right)' \left( \frac{dm(X,\alpha_o)}{dh} [w^*] \right) \right\} \right]
\]
is finite, then \( \|\hat{h}_n - h_o\| = O_p(\|\hat{a}_n - \alpha_o\|). \)
MODELS WITH CONDITIONAL MOMENT RESTRICTIONS

PROOF: For any \( \alpha - \alpha_o = (\theta - \theta_o, h - h_o) \in \mathcal{V} \), we write

\[
\|\alpha - \alpha_o\|^2 = (\theta - \theta_o)' E[D_{w^*}(X)' \Sigma(X)^{-1} D_{w^*}(X)] (\theta - \theta_o) \\
+ E \left\{ \left( \frac{dm(X, \alpha_o)}{dh} [h - h_o + w^*(\theta - \theta_o)] \right)' \Sigma(X)^{-1} \times \left( \frac{dm(X, \alpha_o)}{dh} [h - h_o + w^*(\theta - \theta_o)] \right) \right\}.
\]

Clearly under Assumption 3.4, \( \|\theta - \theta_o\|^2 \leq c \|\alpha - \alpha_o\|^2 \) for some constant \( c > 0 \) if \( E[D_{w^*}(X)' D_{w^*}(X)] \) is positive definite. \( \|h - h_o\|^2 \leq c \|\alpha - \alpha_o\|^2 \) for some constant \( c > 0 \) if

\[
\text{tr} \left[ E \left( \left\| \frac{dm(X, \alpha_o)}{dh} [w^*]^2 \right\|_{E^*}^2 \right) \right]
\]

is finite. \( Q.E.D. \)

PROOF OF PROPOSITION 3.1: We complete the proof by verifying Assumptions 3.1–3.9 of Theorem 3.1. Assumptions 3.1 and 3.2(i) are directly assumed. Assumption 3.2(ii) is satisfied with the linear sieves assumed in condition 3.1.4. Assumption 3.2(iii) is implied by conditions 3.1.4 and 3.1.5(ii). From Newey and Powell (2003), Assumption 3.3 is satisfied by condition 3.1.1. Assumption 3.4 is satisfied with \( A = \Theta \times \mathcal{H} \), where \( \mathcal{H} = \Lambda_0^n (R^d) \), is compact under the norm \( \|\alpha\| = \|\theta\|_E + \|h\|_{\infty, \omega, r} \), and \( \|h\|_{\infty, \omega, r} = \sup_{y_2} |h(y_2) \omega(y_2)| \) where \( \omega(y_2) = (1 + \|y_2\|^2)^{-\alpha/2} \), \( \alpha > \gamma_1 > 0 \). From Chen, Hansen, and Scheinkman (1997), for any \( \alpha = (\theta, h) \in \mathcal{A} \), \( \|\alpha - \Pi_n \alpha\| = \|h - \Pi_n h\|_{\infty, \omega, r} = \sup_{y_2} |h(y_2) - \Pi_n h(y_2)| \omega(y_2) \| \leq c(k_{in})^{-\gamma/d} \). Hence Assumption 3.5 is satisfied with \( \mu_1 = \gamma_1/d \) and condition 3.1.5(iii). Assumption 3.6(i) is trivially satisfied given the homoskedastic error. Also, for any \( \alpha_1, \alpha_2 \in \mathcal{A} \),

\[
|\rho(Z, \alpha_1) - \rho(Z, \alpha_2)| \leq \|X'_1(\theta_1 - \theta_2)\| + h_1(Y_2) - h_2(Y_2)
\]

\[
\leq \|X'_1\|_E \|\theta_1 - \theta_2\|_E
\]

\[
+ \|\omega(Y_2)\|^{-1} \|h_1 - h_2\|_{\infty, \omega, \gamma} \leq c(Z) \|\alpha_1 - \alpha_2\|.
\]

Thus Assumption 3.6(ii) is satisfied with \( \kappa = 1 \), \( c_2(Z) = \|X_1\|_E + \|\omega(Y_2)\|^{-1} \) satisfying \( E[|c(Z)|^2 |X|] < \infty \) under Assumption 3.1(ii) and condition 3.1.2(ii). For any \( \alpha \in \mathcal{A} = \Theta \times \mathcal{H} \), with \( \mathcal{H}_n \) given in (10), \( \sup_{\alpha \in \mathcal{A}_n} |\rho(Z, \alpha)| \leq |Y_1| + \sup_{\alpha \in \mathcal{A}_n} |X'_1\| \theta + \sup_{\alpha \in \mathcal{H}_n} |h(Y_2)| \leq c_1(Z) \), with \( c_1(Z) = |Y_1| + \sup_{\alpha \in \mathcal{A}_n} |X'_1\| \theta + \sup_{\alpha \in \mathcal{H}_n} \|h\|_{\mathcal{A}_n} \). Hence Assumption 3.6(iii) is satisfied by Assumption 3.1(ii) and condition 3.1.2(i). Since \( m(X, \alpha) = E[Y_1|X] - X'_1 \theta - E[h(Y_2)|X] \), Assumption 3.6(iv) is satisfied by condition 3.1.3 and the sieve space (10). Assumption 3.7 is satisfied with conditions 3.1.4 and 3.1.5(i). From Chen and Shen (1998), with \( \mathcal{H}_n \), given in (10) and \( \kappa = 1 \), we have \( \ln[N(e^{1/\kappa}, \mathcal{A}_n, \|\cdot\|)] \leq c \times k_n \ln(1/\epsilon) \) and Assumption 3.8 is satisfied. Assumption 3.9 is satisfied because \( \|\alpha - \alpha_o\|^2 = E[m(X, \alpha)m(X, \alpha')] \). Finally condition 3.1.1(iii) implies that \( E[D_{w^*}(X)' D_{w^*}(X)] \) is positive-definite. The rates now follow from Lemmas 3.1, B.1, and Theorem 3.1.

Q.E.D.

PROOF OF PROPOSITION 3.2: Assumptions 3.1, 3.2(i), and 3.4 are directly assumed. Assumptions 3.2(ii) and (iii) is satisfied by conditions 3.2.4 and 3.2.5(ii). Assumption 3.3 is satisfied by condition 3.2.1. For Assumption 3.5, note that \( A = \Theta \times \mathcal{H} \) with \( \mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_d \) and \( \mathcal{H}_d = \Lambda_0^n (\{-1, 1\}) \) is compact under the norm \( \|\alpha\| = \|\theta\|_E + \max_{j=1, \ldots, q} \|h_j\|_{\infty, \gamma} \) where \( \|h_j\|_{\infty, \gamma} \equiv \sup_{x_j} \|h(x_j)\| \). Also for any \( \alpha = (\theta, h) \in \mathcal{A} \), \( \|\alpha - \Pi_n \alpha\| = \max_{j=1, \ldots, q} \|h_j - \Pi_n h_j\|_{\infty, \gamma} \leq c_j \gamma^n \); see Chen and Shen (1998). Hence Assumption 3.5 is satisfied with \( \mu_1 = \gamma_1 \) and condition 3.2.5(iii).
Assumption 3.6(i) is directly assumed. For any $\alpha_1, \alpha_2 \in A$,
\[ |\rho(Z, \alpha_1) - \rho(Z, \alpha_2)| \]
\[ = \left| F\left( Y_2^* \theta + \sum_{j=1}^{q} h_j(X_j) \right) - F\left( Y_2^* \theta_2 + \sum_{j=1}^{q} h_j(X_j) \right) \right| \]
\[ = \left| \nabla F\left( Y_2^* \tilde{\theta} + \sum_{j=1}^{q} \tilde{h}_j(X_j) \right) \times \left\{ Y_2^* [\theta - \theta_2] + \sum_{j=1}^{q} [h_j(X_j) - h_j(X_j)] \right\} \right| \]
\[ \leq \sup_{\tilde{\theta}, \tilde{h}_j} \left| \nabla F\left( Y_2^* \tilde{\theta} + \sum_{j=1}^{q} \tilde{h}_j(X_j) \right) \right| \left( \|Y_2\|_E + c\|\alpha_1 - \alpha_2\|_s \right) = c_2(Z)\|\alpha_1 - \alpha_2\|_s. \]

Assumption 3.6(ii) is satisfied with $\kappa = 1$ since $E[c_2(Z)^2|X] \leq c'$ by condition 3.2.2(ii). Furthermore, $\sup_{\alpha \in \mathcal{A}} |\rho(Z, \alpha)| \leq |Y_1| + \sup_{\alpha \in \mathcal{A}} |F(Y_2^* \theta + \sum_{j=1}^{q} h_j(X_j))| = c_1(Z).$ Assumption 3.6(iii) is satisfied by condition 3.2.2(i). Assumption 3.6(iv) is satisfied by condition 3.2.3. Assumption 3.7 is satisfied by conditions 3.2.4 and 3.2.5(i). From Chen and Shen (1998), with $\kappa = 1$, we have $\ln[N(e^{1/\kappa}, \mathcal{A}_n, \|\cdot\|_s)] \leq c \times q(2J_n + 1) \ln(1/e)$ and Assumption 3.8 is satisfied. Assumption 3.9(i) is satisfied with
\[ \frac{d\rho(Z, \alpha_o)}{d\alpha} [\alpha - \alpha_o] \]
\[ = -\nabla F\left( Y_2^* \tilde{\theta}_o + \sum_{j=1}^{q} h_{ij}(X_j) \right) \times \left\{ Y_2^* [\theta - \theta_o] + \sum_{j=1}^{q} [h_j(X_j) - h_{ij}(X_j)] \right\}. \]

Under Assumption 3.4(ii), we have for all $\alpha \in \mathcal{A}$ with $\|\alpha - \alpha_o\|_s = o(1)$,
\[ E\{m(X, \alpha)\Sigma(X)^{-1}m(X, \alpha)\} \]
\[ = E\left[ \nabla F\left( Y_2^* \tilde{\theta} + \sum_{j=1}^{q} \tilde{h}_j(X_j) \right) \times \left\{ Y_2^* [\theta - \theta_o] + \sum_{j=1}^{q} [h_j(X_j) - h_{ij}(X_j)] \right\} \right]^2 \]
\[ = (\theta - \theta_o)'E[D_w(X, \tilde{\alpha})D_w(X, \tilde{\alpha})](\theta - \theta_o) \]
where $Y_2^* \tilde{\theta} + \sum_{j=1}^{q} \tilde{h}_j(X_j)$ is some convex combination of $Y_2^* \theta_o + \sum_{j=1}^{q} h_{ij}(X_j)$ and $Y_2^* \theta + \sum_{j=1}^{q} h_j(X_j),$ and
\[ D_w(X, \alpha) \equiv E \left[ \nabla F\left( Y_2^* \theta + \sum_{j=1}^{q} h_j(X_j) \right) \left\{ Y_2 - \sum_{i=1}^{q} w_i(X_i) \right\} \right]. \]

Also
\[ \|\alpha - \alpha_o\|^2 \asymp (\theta - \theta_o)'E[D_w(X, \alpha_o)'D_w(X, \alpha_o)](\theta - \theta_o). \]

Hence under condition 3.2.6, $\|\alpha - \alpha_o\|^2 \asymp E[m(X, \alpha)\Sigma(X)^{-1}m(X, \alpha)]$ locally, and Assumption 3.9(ii) is satisfied. By Lemma B.1 and condition 3.2.1, $\|\theta - \theta_o\|_E^2 \leq c\|\alpha - \alpha_o\|^2$ and $E((\sum_{j=1}^{q}[h_j(X_j) - h_{ij}(X_j)])^2) \leq c\|\alpha - \alpha_o\|^2.$ The results now follow from Lemmas 3.1, B.1, and Theorem 3.1.

Q.E.D.
Denote
\[ \frac{d\hat{m}(X, \alpha)}{d\alpha} [v_n^*] = \sum_{i=1}^{n} \left\{ \frac{dp(Z_i, \alpha)}{d\alpha} [v_n^*] \right\} P^{k_n}(X_i)'(P'P)^{-1} P^{k_n}(X). \]

Recall the definition of neighborhoods \( N_{on} \) and \( N_0 \) introduced in Section 4.

**Corollary C.1:** Assumptions 3.1–3.2, 3.7–3.8, and 4.1–4.4 imply:

\[(i) \quad \sup_{\alpha \in N_{on}} \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{dm(X_i, \tilde{\alpha})}{d\alpha} [v_n^*] \right\}_{E}^2 = o_p(n^{-1/2});

\[(ii) \quad \sup_{\alpha \in N_{on}} \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{dm(X_i, \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{dm(X_i, \alpha_0)}{d\alpha} [v_n^*] \right\}_{E}^2 = o_p(n^{-1/2}).\]

**Proof:** Part (i) can be proved by applying Lemma A.1 to each component of \( (dp(Z_i, \tilde{\alpha}) / d\alpha)[v_n^*] \), similar to the proof of Corollary A.1(i).

To prove Part (ii), note that
\[ F = \left\{ \left\{ \frac{dm(X_i, \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{dm(X_i, \alpha_0)}{d\alpha} [v_n^*] \right\}_{E}^2 : \tilde{\alpha} \in N_{on} \right\} \]

is a subset of
\[ \left\{ \left\{ g(X_i) - \frac{dm(X_i, \alpha_0)}{d\alpha} [v_n^*] \right\}_{E}^2 : g \in A_n(X) \right\}, \]

which is a Donsker class. Furthermore, Assumption 4.4 implies uniformly over \( \tilde{\alpha} \in N_{on}, \)

\[ E\left( \left\{ \frac{dm(X, \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] \right\}_{E}^2 \right) \]

\[ \leq 2E\left( \left\{ \frac{dm(X, \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] \right\}_{E}^2 \right) \]

\[ \times \sup_{X, \tilde{\alpha}} \left\{ \left\{ \frac{dm(X, \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] \right\}_{E}^2 \right\} \]

\[ \rightarrow 0 \quad \text{as} \quad \|\tilde{\alpha} - \alpha_0\| \rightarrow 0. \]

Applying Lemma 1 in Chen, Linton, and van Keilegom (2003), we have

\[ \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{dm(X_i, \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{dm(X_i, \alpha_0)}{d\alpha} [v_n^*] \right\}_{E}^2 \]

\[ - E\left( \left\{ \frac{dm(X_i, \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{dm(X_i, \alpha_0)}{d\alpha} [v_n^*] \right\}_{E}^2 \right) \]

\[ = o_p(n^{-1/2}) \quad \text{uniformly over} \quad \tilde{\alpha} \in N_{on}. \]
By Assumption 4.4, uniformly over $\hat{\alpha} \in \mathcal{N}_0$,

$$E\left(\left\| \frac{d^2m(X_i, \hat{\alpha})}{d\alpha^2} [v_n^*, v_n^*] - \frac{dm(X_i, \alpha_0)}{d\alpha} [v_n^*] \right\|_E^2 \right) = o_p(n^{-1/2}).$$

Result (ii) now follows. \textit{Q.E.D.}

Denote

$$\frac{d^2\hat{m}(X, \alpha)}{d\alpha^2} [v_n^*, v_n^*] = \sum_{i=1}^n \left\{ \frac{d^2p(Z_i, \alpha)}{d\alpha^2} [v_n^*, v_n^*] \right\} p^{(X_i)(P'P)^{-1}p^{(X_i)}}.$$

**COROLLARY C.2:** (i) \textit{Under Assumptions 3.1–3.2, 3.4–3.9, and 4.6, we have}

$$\sup_{\hat{\alpha} \in \mathcal{N}_0} \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d^2\hat{m}(X, \hat{\alpha})}{d\alpha^2} [v_n^*, v_n^*] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \hat{\alpha}) = o_p(n^{-1/4}).$$

(ii) \textit{Under Assumptions 3.1–3.2, 3.4, 3.7–3.8, 4.1(ii), 4.2–4.4, we have}

$$\sup_{\hat{\alpha} \in \mathcal{N}_0} \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d^2\hat{m}(X, \hat{\alpha})}{d\alpha^2} [v_n^*, v_n^*] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \left\{ \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] \right\} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] \right\} + o_p(n^{-1/4}).$$

**PROOF:** (i) For some constant $c$, Assumption 3.4 implies

$$\left| \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d^2\hat{m}(X_i, \hat{\alpha})}{d\alpha^2} [v_n^*, v_n^*] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \hat{\alpha}) \right| \leq c \sqrt{\frac{1}{n} \sum_{i=1}^n \left\| \frac{d^2\hat{m}(X_i, \hat{\alpha})}{d\alpha^2} [v_n^*, v_n^*] \right\|_E^2 \left( \frac{1}{n} \sum_{i=1}^n \left\| \hat{m}(X_i, \hat{\alpha}) \right\|_E^2 \right)}.$$

Assumption 4.6 implies uniformly over $\hat{\alpha} \in \mathcal{N}_0$:

$$\frac{1}{n} \sum_{i=1}^n \left\| \frac{d^2\hat{m}(X_i, \hat{\alpha})}{d\alpha^2} [v_n^*, v_n^*] \right\|_E^2 \leq \frac{1}{n} \sum_{i=1}^n c_3(Z)^2 = O_p(1).$$

The result now follows from applying Corollary A.2.

(ii) \text{Uniformly over } \hat{\alpha} \in \mathcal{N}_0,

$$\frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm(X, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \left\{ \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] \right\} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \left\{ \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] \right\}$$

$$+ \frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \left\{ \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] - \frac{dm(X, \alpha_0)}{d\alpha} [v_n^*] \right\}.$$
Part (ii) now follows from Corollary C.1 and Assumption 3.4.

**Corollary C.3**: (i) Under Assumptions 3.1–3.2, 3.4–3.9, 4.1(ii), 4.2–4.4, we have uniformly over $\tilde{\alpha} \in N_{on}$:

\[
\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{dm(X_i, \alpha_o)}{d\tilde{\alpha}} \right] \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \tilde{\alpha})
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{dm(X_i, \alpha_o)}{d\alpha} \right] \left[ \Sigma(X_i) \right]^{-1} \hat{m}(X_i, \tilde{\alpha}) + o_p(n^{-1/2}).
\]

(ii) Under Assumptions 3.1–3.2, 4.1(ii), 4.2–4.5, we have uniformly over $\tilde{\alpha} \in N_{on}$:

\[
\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{dm(X_i, \alpha_o)}{d\tilde{\alpha}} \right] \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \tilde{\alpha}) - \hat{m}(X_i, \alpha_o)
\]

\[
= \langle v^*, \tilde{\alpha} - \alpha_o \rangle + o_p(n^{-1/2}).
\]

(iii) Under Assumptions 3.1–3.2, 4.1(iii), 3.4(ii), 4.2, 3.7–3.8, we have

\[
\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{dm(X_i, \alpha_o)}{d\alpha} \right] \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \alpha_o)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{dm(X_i, \alpha_o)}{d\alpha} \right] \left[ \Sigma(X_i) \right]^{-1} \rho(Z_i, \alpha_o) + o_p(n^{-1/2}).
\]

**Proof**: (i) We write

\[
\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{d\hat{m}(X_i, \tilde{\alpha})}{d\tilde{\alpha}} \right] \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \tilde{\alpha})
\]

\[
- \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{dm(X_i, \alpha_o)}{d\alpha} \right] \left[ \Sigma(X_i) \right]^{-1} \hat{m}(X_i, \tilde{\alpha})
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{d\hat{m}(X_i, \tilde{\alpha})}{d\alpha} \right] \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \tilde{\alpha})
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{dm(X_i, \alpha_o)}{d\alpha} \right] \left[ \left[ \hat{\Sigma}(X_i) \right]^{-1} - \left[ \Sigma(X_i) \right]^{-1} \right] \hat{m}(X_i, \tilde{\alpha})
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{dm(X_i, \alpha_o)}{d\alpha} \right] \left[ v_n^* - v^* \right] \left[ \Sigma(X_i) \right]^{-1} \hat{m}(X_i, \tilde{\alpha})
\]

Then $A1 = o_p(n^{-1/2})$ by Corollaries C.1 and A.2(ii) and Assumption 3.4; $A2 = o_p(n^{-1/2})$ by Corollary A.2(ii) and Assumption 3.4; $A3 = o_p(n^{-1/2})$ by Corollary A.2(ii) and Assumption 4.2.
With
\[ g(X, v^*) = \left( \frac{dm(X, \alpha_0)}{d\alpha} [v^*] \right)\} [\Sigma(X)]^{-1}, \]
define
\[ F = \left\{ g(X, v^*)\tilde{m}(X, \alpha) : \alpha \in \mathcal{N}_{on}, \tilde{m} \in A'_\gamma(X) \right\} \]
with
\[ \sup_{x \in X, \alpha \in \mathcal{N}_{on}} |\tilde{m}(x, \alpha) - m(x, \alpha)| = o(1). \]
Notice that, by Assumptions 3.2(iii) and 3.7, \( F \subset \{g(X, v^*)f(x) : f \in A'_\gamma(X)\} \) with \( \gamma > d_x/2 \).
Since for any \( \tilde{f}, \bar{f} \in A'_\gamma, |g(X, v^*)\tilde{f}(X) - g(X, v^*)\bar{f}(X)| \leq |g(X, v^*)| \times |\tilde{f}(X) - \bar{f}(X)| \) and \( E(|g(X, v^*)|^2) \) is finite by Assumptions 3.4(ii) and 4.1(ii), we have
\[ \log\left[N(\delta, \{g(X, v^*)f(X) : f \in A'_\gamma(X)\}, \| \cdot \|_{L_2(P)}^\gamma)\right] \leq C \times \log\left[N(\delta, \{g(X, v^*)f(X) : f \in A'_\gamma(X)\}, \| \cdot \|_\infty)\right] \leq C \times \left( \frac{1}{\delta} \right)^{\frac{d_x}{\gamma}}. \]
Hence \( \{g(X, v^*)f(X) : f \in A'_\gamma(X)\} \) with \( \gamma > d_x/2 \) is a Donsker class by Theorem 2.5.6 of van der Vaart and Wellner (1996). By Assumptions 3.1, 3.2, 3.6–3.8, \( \tilde{m}(X, \alpha) \in A'_\gamma(X) \) uniformly over \( \alpha \in \mathcal{N}_{on} \) with probability approaching one. Moreover, \( E[\|g(X, v^*)\tilde{m}(X, \alpha) - g(X, v^*)m(X, \alpha)\|^2] = o_p(1) \) uniformly over \( \alpha \in \mathcal{N}_{on} \). Applying Lemma 1 of Chen, Linton, and van Keilegom (2003) we have uniformly over \( \alpha \in \mathcal{N}_{on} \):
\[ \frac{1}{n} \sum_{i=1}^{n} g(X_i, v^*)\{\tilde{m}(X_i, \alpha) - m(X_i, \alpha)\} - E(g(X_i, v^*)\{\tilde{m}(X_i, \alpha) - m(X_i, \alpha)\}) \]
\[ = o_p(n^{-1/2}); \]
\[ \frac{1}{n} \sum_{i=1}^{n} g(X_i, v^*)\{\tilde{m}(X_i, \alpha_0) - m(X_i, \alpha_0)\} - E(g(X_i, v^*)\{\tilde{m}(X_i, \alpha_0) - m(X_i, \alpha_0)\}) \]
\[ = o_p(n^{-1/2}). \]
Hence
\[ \frac{1}{n} \sum_{i=1}^{n} g(X_i, v^*)\{\tilde{m}(X_i, \hat{\alpha}) - m(X_i, \alpha_0)\} - E(g(X_i, v^*)\{\tilde{m}(X_i, \hat{\alpha}) - m(X_i, \alpha_0)\}) \]
\[ = \frac{1}{n} \sum_{i=1}^{n} g(X_i, v^*)\{m(X_i, \hat{\alpha}) - m(X_i, \alpha_0)\} \]
\[ - E(g(X_i, v^*)\{m(X_i, \hat{\alpha}) - m(X_i, \alpha_0)\}) + o_p(n^{-1/2}). \]
Note that
\[ E(g(X_i, v^*)\{\tilde{m}(X_i, \hat{\alpha}) - m(X_i, \alpha_0)\}) = E(\hat{g}(X_i, v^*)\{m(X_i, \hat{\alpha}) - m(X_i, \alpha_0)\}), \]
where \( \hat{g}(X, v^*) = P^x(X)'(P^xP)^{-1} \sum_{i=1}^{n} P^k(X_i)g(X_i, v^*), \) and that
\[ E(\hat{g}(X_i, v^*)\{m(X_i, \hat{\alpha}) - m(X_i, \alpha_0)\}) - E(g(X_i, v^*)\{m(X_i, \hat{\alpha}) - m(X_i, \alpha_0)\}) \]
\[ = E(\{\hat{g}(X_i, v^*) - g(X_i, v^*)\}\{m(X_i, \hat{\alpha}) - m(X_i, \alpha_0)\}) = o_p(n^{-1/2}). \]
It follows that

\[
\frac{1}{n} \sum_{i=1}^{n} g(X_i, v^*) \{ \hat{m}(X_i, \hat{\alpha}) - \hat{m}(X_i, \alpha_0) \} = \frac{1}{n} \sum_{i=1}^{n} g(X_i, v^*) \{ m(X_i, \hat{\alpha}) - m(X_i, \alpha_0) \} + o_p(n^{-1/2}).
\]

Consider the class of functions \( \{ g(X_i, v^*) m(X_i, \alpha) : \alpha \in \mathcal{N}_a \} \). Again \( \{ g(X_i, v^*) m(X_i, \alpha) : \alpha \in \mathcal{N}_a \} \) is a Donsker class. Hence, we have

\[
\sup_{\alpha \in \mathcal{N}_a} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i, v^*) \{ m(X_i, \alpha) - m(X_i, \alpha_0) \} - E \left( g(X_i, v^*) \{ m(X_i, \alpha) - m(X_i, \alpha_0) \} \right) \right| = o_p(n^{-1/2}).
\]

It follows that

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

\[
= \langle v^*, \hat{\alpha} - \alpha_0 \rangle + E \left( g(X_i, v^*) \left\{ \frac{dm(X_i, \hat{\alpha})}{d\alpha}[\hat{\alpha} - \alpha_0] - \frac{dm(X_i, \alpha_0)}{d\alpha}[\hat{\alpha} - \alpha_0] \right\} \right)
\]

\[
+ o_p(n^{-1/2})
\]

\[
= \langle v^*, \hat{\alpha} - \alpha_0 \rangle + o_p(n^{-1/2})
\]

for some \( \hat{\alpha} \in \mathcal{N}_a \), a convex combination of \( \hat{\alpha} \) and \( \alpha_0 \). The last equation follows from Assumptions 4.1(ii) and 4.5.

(iii) By exchanging summation, we obtain

\[
\frac{1}{n} \sum_{i=1}^{n} g(X_i, v^*) \hat{m}(X_i, \alpha_0) - \frac{1}{n} \sum_{i=1}^{n} g(X_i, v^*) \rho(Z_i, \alpha_0)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \{ \hat{g}(X_i, v^*) - g(X_i, v^*) \} \rho(Z_i, \alpha_0).
\]

Notice that \( \hat{g}(X, v^*) - g(X, v^*) = o_p(1) \) uniformly over \( X \) and that \( E[\rho(Z_i, \alpha_0)|X_1, X_2, \ldots, X_n] = 0 \). Part (iii) follows from applying the (second moment) Markov inequality. Q.E.D.

PROOF OF THEOREM 4.1: Recall the notation \( \hat{L}_n(\alpha) \) introduced in Appendix B. Let \( 0 < \varepsilon_n = o(n^{-1/2}) \) and \( u^* = \pm v^* \). To simplify notation, denote \( u_n^* = \Pi_n u^* \). We take a continuous path \( \{ \alpha(t) : t \in [0, 1] \} \) in \( \mathcal{N}_a \) such that \( \alpha(0) = \hat{\alpha} \) and \( \alpha(1) = \hat{\alpha} + \varepsilon_n u_n^* \in \mathcal{N}_a \). For example, \( \alpha(t) = \hat{\alpha} + t \varepsilon_n u_n^* \). By Assumptions 4.3 and 4.6, \( \hat{L}_n(\alpha(t)) \) is twice continuously differentiable with respect to \( t \). Denote

\[
\frac{d}{dt} \hat{m}(X, \alpha(t)) \bigg|_{t=0} = \frac{d}{d\alpha} \hat{m}(X, \alpha(\tau)) [\varepsilon_n u_n^*],
\]

\[
\frac{d^2}{dt^2} \hat{m}(X, \alpha(t)) \bigg|_{t=0} = \frac{d^2}{d\alpha d\alpha} \hat{m}(X, \alpha(\tau)) [\varepsilon_n u_n^*, \varepsilon_n u_n^*].
\]
By definition of \( \hat{\alpha} \), and a Taylor expansion around \( t = 0 \) up to second order, we have

\[
0 \leq \hat{L}_n(\hat{\alpha}) - \hat{L}_n(\hat{\alpha} + \varepsilon_n u^*_n) = \hat{L}_n(\alpha(0)) - \hat{L}_n(\alpha(1)) = \frac{d\hat{L}_n(\alpha(t))}{dt} \bigg|_{t=0} - \frac{1}{2} \frac{d^2\hat{L}_n(\alpha(t))}{dt^2} \bigg|_{t=s} \quad \text{for some } s \in [0, 1].
\]

Then

\[
0 \leq \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \hat{\alpha})}{d\alpha} [\varepsilon_n u^*_n] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \hat{\alpha}) + \frac{1}{2} \sum_{i=1}^{n} \left\{ \frac{d^2\hat{m}(X_i, \alpha(s))}{d\alpha^2} [\varepsilon_n u^*_n, \varepsilon_n u^*_n] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \left\{ \frac{d\hat{m}(X_i, \alpha(s))}{d\alpha} [\varepsilon_n u^*_n] \right\}
\]

where \( \alpha(s) = \hat{\alpha} + s \varepsilon_n u^*_n \equiv \hat{\alpha} \in \mathcal{N}_{on} \). By Corollary C.2 we have uniformly over \( \alpha(s) \in \mathcal{N}_{on} \),

\[
\frac{1}{2n} \sum_{i=1}^{n} \left\{ \frac{d^2\hat{m}(X_i, \alpha(s))}{d\alpha^2} [u^*_n, u^*_n] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \alpha(s)) = o_p(n^{-1/4}),
\]

\[
\frac{1}{2n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \alpha(s))}{d\alpha} [u^*_n] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \left\{ \frac{d\hat{m}(X_i, \alpha(s))}{d\alpha} [u^*_n] \right\} = O_p(1).
\]

Hence uniformly over \( \alpha(s) \in \mathcal{N}_{on} \),

\[
0 \leq \frac{\varepsilon_n}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \hat{\alpha})}{d\alpha} [u^*_n] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \hat{\alpha}) + O_p(\varepsilon_n^2).
\]

Since \( u^* = \pm u^* \) and \( \varepsilon_n = o(n^{-1/2}) > 0 \), we obtain

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \hat{\alpha})}{d\alpha} [u^*_n] \right\} \left[ \hat{\Sigma}(X_i) \right]^{-1} \hat{m}(X_i, \hat{\alpha}) = o_p(n^{-1/2}).
\]

By Corollary C.3,

\[
\sqrt{n}(u^*, \hat{\alpha} - \alpha_o) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{dm(X_i, \alpha_o)}{d\alpha} [u^*] \right\} \left[ \Sigma(X_i) \right]^{-1} \rho(Z_i, \alpha_o) + o_p(1).
\]

Since \( \{u^*, \hat{\theta} - \theta_o\} = \lambda(\hat{\theta} - \theta_o) \) for arbitrarily fixed \( \lambda \in \mathbb{R}^d \) with \( |\lambda| \neq 0 \), we obtain Theorem 4.1 by applying a standard CLT for i.i.d. data. Q.E.D.

**Proof of Proposition 4.1:** We prove this by verifying that Assumptions 4.1–4.6 of Theorem 4.1 are satisfied. Note that

\[
\frac{dp(Z, \alpha)}{d\alpha} [v] = -X'_iv_0 - v_0(Y_2),
\]

\[
\frac{dm(X, \alpha)}{d\alpha} [v] = -X'_iv_0 - E[v_0(Y_2)X], \quad \text{and}
\]

\[
D_{v^*}(X) = X'_i - E[v^*(Y_2)X]
\]
with \( w^*(Y_2) \) given by (17). Assumption 4.1(i) is implied by condition 3.1.1(iii). Assumption 4.1(ii) and (iii) are directly assumed. Assumption 4.2 is satisfied with conditions 4.1.1 and 3.1.3 and the sieve space (10). Assumption 4.3(i) is implied by Assumption 3.1 and condition 4.1.1. Assumption 4.3(ii) is satisfied with conditions 4.1.1 and 3.1.3 and the sieve space (10). Since

\[
\frac{dm(X, \alpha)}{d\alpha}[v] - \frac{dm(X, \alpha_0)}{d\alpha}[v] = 0
\]

for all \( \alpha \) and \( v \), Assumptions 4.4 and 4.5 are trivially satisfied. Since

\[
\frac{d^2 p(Z, \alpha)}{d\alpha d\alpha}[v, v] = 0
\]

for all \( \alpha \) and \( v \), Assumption 4.6 is automatically satisfied. \( Q.E.D. \)

PROOF OF PROPOSITION 4.2: We prove this by verifying that Assumptions 4.1–4.6 of Theorem 4.1 are satisfied. Note that

\[
\frac{dp(Z, \alpha)}{d\alpha}[v] = -\nabla F \left( Y_2^* \theta + \sum_{l=1}^q h_l(X_l) \right) \times \left\{ Y_2^* v_\theta + \sum_{l=1}^q [v_h(X_l)] \right\},
\]

and \( w^* \) and \( D_w(X) \) are given by (18) and (19). Assumption 4.1(i) is implied by condition 3.2.1; 4.1(ii) and (iii) are directly assumed. Assumption 4.2 is satisfied with condition 4.2.1 and the sieve space (12). Since

\[
\left| \frac{dp(Z, \alpha)}{d\alpha}[v_n^*] \right| \leq \sup_{\alpha \in \mathbb{N}^n} \left| \nabla F \left( Y_2^* \theta + \sum_{l=1}^q h_l(X_l) \right) \right| \times \left\{ Y_2^* v_\theta^* + \sum_{l=1}^q \sup_n |v_h^*(X_l)| \right\},
\]

and

\[
\left| \frac{dp(Z, \alpha_1)}{d\alpha}[v_n^*] - \frac{dp(Z, \alpha_2)}{d\alpha}[v_n^*] \right| \leq \nabla F \left( Y_2^* \theta_1 + \sum_{l=1}^q h_l(X_l) \right) - \nabla F \left( Y_2^* \theta_2 + \sum_{l=1}^q h_l(X_l) \right) \times \left\{ Y_2^* v_\theta^* + \sum_{l=1}^q \sup_n |v_h^*(X_l)| \right\} \leq c \left\{ \left\| Y_2 \right\|_E + c \right\} \times \| \alpha_1 - \alpha_2 \|_1 = c \| \alpha_1 - \alpha_2 \|_1,
\]

Assumption 4.3(i) is satisfied under conditions 4.2.1 and 4.2.2. Assumption 4.3(ii) is again satisfied with conditions 4.2.1 and the sieve space (12). For Assumption 4.4,

\[
\left| \frac{dm(X, \alpha)}{d\alpha}[v_n^*] - \frac{dm(X, \alpha_o)}{d\alpha}[v_n^*] \right| = -E \left\{ \nabla F \left( Y_2^* \theta + \sum_{l=1}^q h_l(X_l) \right) - \nabla F \left( Y_2^* \theta_o + \sum_{l=1} h_l(X_l) \right) \right\} \times \left\{ Y_2^* v_\theta + \sum_{l=1}^q [v_h^*(X_l)] \right\} \times \right\{ X \right\}
\]

\[
= -E \left\{ \nabla^2 F \left( Y_2^* \theta + \sum_{l=1}^q h_l(X_l) \right) \right\} \times \left\{ Y_2^* (\theta - \theta_o) + \sum_{l=1} [h_l(X_l) - h_l(X_o)] \right\},
\]
where \( Y_2 \tilde{\theta} + \sum_{i=1}^{q} \tilde{h}_i(X_i) \) is a convex combination of \( Y_2 \tilde{\theta} + \sum_{i=1}^{q} \tilde{h}_i(X_i) \) and \( Y_2 \tilde{\theta}_o + \sum_{i=1}^{q} \tilde{h}_i(X_i) \). Hence Assumption 4.4 is satisfied given Proposition 3.2, conditions 3.2.1 and 3.2.6. Similarly Assumption 4.5 is satisfied given Proposition 3.2, conditions 3.2.1 and 3.2.6 since

\[
\lim_{n \to \infty} \frac{d^2}{d\alpha d\alpha} \left( \frac{d^2 \rho(Z, \alpha)}{d\alpha d\alpha} \right) = -\nabla^2 F \left( Y_2 \tilde{\theta} + \sum_{i=1}^{q} \tilde{h}_i(X_i) \right) \times \left( Y_2^* v_o^* + \sum_{i=1}^{q} v_i^*(X_i) \right)^2,
\]

Assumption 4.6 is satisfied under conditions 4.2.1 and 4.2.2(ii). Q.E.D.

APPENDIX D: ASYMPTOTIC VARIANCE ESTIMATOR

COROLLARY D.1: Suppose that Assumptions 3.1–3.2, 3.7–3.8, and 5.1 hold. Then: uniformly over \( X \in \mathcal{X}, \alpha \in \mathcal{N}_\alpha \), (i) \( \tilde{\Sigma}_o(X, \alpha) = \Sigma_o(X, \alpha) + o_\rho(1) \); (ii) \( \tilde{\Sigma}_o(X, \alpha) = \Sigma_o(X, \alpha_o) + o_\rho(1) \).
PROOF: (i) We apply Parts A and B of Lemma A.1 to each element of \( \rho(Z, \alpha) \rho(Z, \alpha)' \) with \( c_{1n} = 1, \ \delta_{1n} = \delta_{2n} = o(1) \). Noting that conditions (i)–(iv) are satisfied by Assumptions 3.1–3.2, 3.7, and 5.1, we obtain uniformly over \( X \in \mathcal{X} \) and \( \alpha \in \mathcal{N}_{on} \),

\[
\left| \Sigma_\alpha(X, \alpha) - \Sigma_\alpha(X, \alpha_o) \right| \leq E \left[ \left| \rho(Z, \alpha) \rho(Z, \alpha)' - \rho(Z, \alpha_o) \rho(Z, \alpha_o)' \right| \right] \leq E[c_{6}(Z)|X] \times \| \alpha - \alpha_o \|_*^* = o(1)
\]

where the last relation is due to Assumption 5.1(ii) and the definition of \( \mathcal{N}_{on} \). Q.E.D.

Denote

\[
D_{w_j}(X, \alpha) = E \left\{ \frac{dp(Z, \alpha)}{d\theta_j} - \frac{dp(Z, \alpha)}{dh \cdot [w_j]} \big| X \right\}
\]

and

\[
\hat{D}_{w_j}(X, \alpha) = \frac{d\hat{m}(X, \alpha)}{d\theta_j} - \frac{d\hat{m}(X, \alpha)}{dh \cdot [w_j]} \quad \text{for } j = 1, \ldots, d_\theta.
\]

COROLLARY D.2: Suppose that Assumptions 3.1–3.2, 3.7–3.8, and 5.2 hold. Then: uniformly over \( X \in \mathcal{X}, \alpha \in \mathcal{N}_{on}, w_j \in \mathcal{H}_n, j = 1, \ldots, d_\theta \), (i) \( \hat{D}_{w_j}(X, \alpha) = D_{w_j}(X, \alpha) + o_p(1) \); (ii) \( \hat{D}_{w_j}(X, \alpha) = D_{w_j}(X, \alpha_o) + o_p(1) \).

PROOF: The proof is the same as that for Corollary D.1, except using \( (dp(Z, \alpha)/d\theta_j) - (dp(Z, \alpha)/dh) [w_j] \) and \( D_{w_j}(X, \alpha) \) and Assumption 5.2. Q.E.D.

PROOF OF THEOREM 5.1: Given Assumptions 3.1, 3.4(i) and (ii), and Corollary D.1, to prove the consistency of the estimated covariance matrix, it remains to show that for each \( j, \hat{D}_{w_j}(X) \) converges to \( D_{w_j}(X) \) uniformly over \( X \in \mathcal{X} \). We note that \( (1/n) \sum_{i=1}^{n} \hat{D}_{w_j}(X, \alpha_n) [\hat{\Sigma}(X_i)]^{-1} \) converges to \( D_{w_j}(X, \alpha_o) \) is globally convex in \( w_j \). The solution \( \hat{w}_j(\cdot) \) must be bounded by \( \| \hat{w}_j(\cdot) \|_s \leq c \) for some constant \( c \). Thus, we only need to be concerned with the subset \( \{ v \in \mathcal{W} : \| v \|_s \leq c \} \). We still use the sieve space \( \mathcal{H}_n \) to approximate this subset. Note that

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{D}_{w_j}(X, \alpha_n) [\hat{\Sigma}(X_i)]^{-1} \hat{D}_{w_j}(X, \alpha_n) - \frac{1}{n} \sum_{i=1}^{n} D_{w_j}(X, \alpha_o) [\Sigma(X)]^{-1} D_{w_j}(X, \alpha_o)
\]

\[
\quad = \frac{1}{n} \sum_{i=1}^{n} [\hat{D}_{w_j}(X, \alpha_n) - D_{w_j}(X, \alpha_o)] [\hat{\Sigma}(X_i)]^{-1} \hat{D}_{w_j}(X, \alpha_n)
\]

\[
\quad + \frac{1}{n} \sum_{i=1}^{n} D_{w_j}(X, \alpha_o) [\hat{\Sigma}(X_i)]^{-1} [\hat{D}_{w_j}(X, \alpha_n) - D_{w_j}(X, \alpha_o)]
\]

\[
\quad + \frac{1}{n} \sum_{i=1}^{n} D_{w_j}(X, \alpha_o) [\hat{\Sigma}(X_i)]^{-1} - [\Sigma(X_i)]^{-1} D_{w_j}(X, \alpha_o),
\]

and that \( \alpha_n \in \mathcal{N}_{on} \) with probability approaching one by Theorem 3.1. Applying Corollary D.2(ii) and Assumption 3.4(i) and (ii), we obtain uniformly over \( w_j \in \mathcal{H}_n \) with \( \| w_j \|_s \leq c \):

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{D}_{w_j}(X, \alpha_n) [\hat{\Sigma}(X_i)]^{-1} \hat{D}_{w_j}(X, \alpha_n) = \frac{1}{n} \sum_{i=1}^{n} D_{w_j}(X, \alpha_o) [\Sigma(X)]^{-1} D_{w_j}(X, \alpha_o) + o_p(1).
\]
By Assumptions 3.5(i), 3.7(i), 4.1(i), 5.2, and Lemma A.1 in Newey and Powell (2003), we have \( \| \bar{w}_t^*(\cdot) - w^*_t(\cdot) \|_2 = o_p(1) \). This, Assumption 5.2, and Corollary D.2 imply \( D_{\bar{w}_t^*}(X_t) = D_{w^*_t}(X_t) + o_p(1) \) uniformly over \( X_t \in X \). The theorem now follows.

Q.E.D.

**APPENDIX E: SEMIPARAMETRIC EFFICIENCY**

**PROOF OF THEOREM 6.1:** We follow the approach described in Newey (1990b) by first characterizing the tangent space and then computing the projection. Let \( u = \rho(Z, \alpha_o) \). Decompose \( Y \) into \( (Y_1, Y_2) \) so that \( \dim(Y_1) = \dim(u) \). Let \( f_o(u, y_2, x) \) denote the true conditional density of \( (u, Y_2) \) given \( X \) and let \( g_o(x) \) denote the true marginal density of \( X \). The true joint density of \( (Y, X) \) (with respect to a dominating measure \( \mu \)) is \( q_o(y, x, \alpha_o) = J(z, \alpha_0) f_o(\rho(Z, \alpha_o), y_2, x) g_o(x) \), where \( J(z, \alpha_o) \) is the Jacobian of the transformation from \( u = \rho(Z, \alpha_o) \) to \( Y \).

Since \( \mathcal{H} \) is convex at \( h_o, h_o + \xi_1(h - h_o) \in \mathcal{H} \) for any \( h \in \mathcal{H} \) and small \( \xi_1 > 0 \). For small \( \xi_2 > 0 \), define \( \Delta_t(u, y_2, x, \xi_2) = (1 + \xi_2 t_1(u, y_2, x) - t_1(x) - t_2(u)t_2(x)) \), where \( t_1(\cdot) \) and \( t_2(\cdot) \) are any bounded functions with bounded first and second derivatives with respect to \( u \), \( E[ut_2(u)|x] \) is nonsingular for any \( x, t_2(x) = (E[ut_2(u)|x])^{-1}E[ut_1(u, y_2, x)|x] \) and \( t_1(x) = E[t_1(u, y_2, x)|x] - E[t_2(u)t_2(x)] \). By construction \( f_o(u, y_2, x, \xi_2) = f_o(u, y_2, x) \Delta_t(u, y_2, x, \xi_2) \) is a density for small \( \xi_2 \). Similarly, \( g(x, \xi_3) = g_o(x) \Delta_t(x, \xi_3) \) with \( \Delta_t(x, \xi_3) = (1 + \xi_3 t_3(x) - E[t_3(x)]) \) is a density for small \( \xi_3 \) and bounded \( t_3(x) \). With \( \delta = (\theta, \xi_1, \xi_2, \xi_3)' \), define \( \Delta(y, x, \delta) = \Delta_t(\rho(z, \theta, h_o + \xi_1(h - h_o)), y_2, x, \xi_2) \). Then

\[
q(y, x, \delta) = q_o(y, x, \theta, h_o + \xi_1(h - h_o)) \Delta(y, x, \delta)
\]

is a parametric submodel passing through the true model. Next, we show that this submodel is smooth; see Newey (1990b, p. 127) for definition. For any sequence \( \delta_t \rightarrow \delta \), since \( t_1(\cdot) \) and \( t_2(\cdot) \) have bounded first and second derivatives with respect to \( u \), it is straightforward to show that

\[
E \left\{ \left( \Delta(y, x, \delta_t)^{1/2} - \Delta(y, x, \delta) \right)^2 \right\} \rightarrow 0 \quad \text{and} \quad \int \left\{ \frac{\partial \Delta(y, x, \delta_t)}{\partial \delta} - \frac{\partial \Delta(y, x, \delta)}{\partial \delta} \right\}^2 d\mu \rightarrow 0.
\]

This and the assumption that \( q_o(y, x, \theta, h_o + \xi_1(h - h_o)) \) is smooth imply that \( q(y, x, \delta) \) is smooth.

Let \( S_0(y, x), S_{\xi_1}(y, x), S_{\xi_2}(y, x), \) and \( S_{\xi_3}(y, x) \) denote the scores of the log likelihood function of the submodel (of one observation), evaluated at the true values. Then

\[
S_0(y, x) = \frac{q_o(y, x, \alpha_o)}{q_o(y, x, \alpha_o)}, \quad S_{\xi_1}(y, x) = \frac{q_o(y, x, \alpha_o, [h - h_o])}{q_o(y, x, \alpha_o)}, \quad S_{\xi_2}(y, x) = t_1(u, y_2, x) - t_1(x) - t_2(u)t_2(x), \quad S_{\xi_3}(y, x) = t_3(x) - E[t_3(x)].
\]

Since any unbounded functions can be approximated by a sequence of bounded functions, the tangent space is \( A_h + A_f + A_s \), where

\[
A_h = \left\{ \frac{q_h(y, x, \alpha_o, w)}{q_o(y, x, \alpha_o)} : \text{for all } w \in \mathbb{W} \right\}, \quad A_f = \left\{ t_3(x) - E[t_3(x)] : \text{for any } t_3 \text{ with } E[t_3(x)^2] < \infty \right\}, \quad \text{and}
\]

\[
A_s = \left\{ \Delta_t(x, y_2, x, \xi_2) : \text{for any } \xi_2 \right\}.
\]
\[ \Lambda_f = \left\{ t_1(u, y_2, x) - t_1(x) - t_2(u)'t_2(x) : \text{for any } t_1, t_2 \text{ with } E[t_1(u, y_2, x)^2] < \infty, E[t_2(u)^2] < \infty, E[u_t(u)|x] \text{ nonsingular for all } x. \right\}. \]

We now compute the projection. For any component \( \theta_j \), we choose \( w_0(\cdot), \Delta^*_f(u, y_2, x) \), and \( \Delta^*_g(x) \) to solve:

\[
\min_{w(.) \in \mathcal{W}} E\left\{ \left( S_{\theta_j}(y, x) - \frac{q_h[y, x, \alpha_o, w]}{q_o(y, x, \alpha_o)} - \Delta_f(u, y_2, x) - \Delta_g(x) \right)^2 \right\}.
\]

Note that the above minimization problem is equivalent to

\[
\min_{\Delta(\cdot) \in \Delta_f, \Delta_g(\cdot) \in \Delta_g} E\left\{ \left( S_{\theta_j}(y, x) - \frac{q_h[z, \alpha_o, w]}{q_o(z, \alpha_o)} - \Delta_f(u, y_2, x) - \Delta_g(x) \right)^2 \right\}.
\]

For any fixed \( w(\cdot) \in \mathcal{W} \), let \( \Delta^*_{fw}(u, y_2, x), \Delta^*_{gw}(x) \) denote the solution to

\[
\min_{\Delta(\cdot) \in \Delta_f, \Delta_g(\cdot) \in \Delta_g} E\left\{ \left( S_{\theta_j}(y, x) - \frac{q_h[y, x, \alpha_o, w]}{q_o(y, x, \alpha_o)} - \Delta_f(u, y_2, x) - \Delta_g(x) \right)^2 \right\}.
\]

It is very easy to show that \( \Delta^*_{gw}(x) = 0 \). Let

\[
t_f^*(u, y_2, x) = E\left\{ S_{\theta_j}(y, x) - \frac{q_h[y, x, \alpha_o, w]}{q_o(y, x, \alpha_o)} \right\}|u, y_2, x, \Delta^{*}_{fw}(u, y_2, x) = 0,
\]

\[
t^*_f(u) = u, \quad \tilde{t}^*_f(x) = \Sigma_o(x)^{-1} E\left\{ \left( S_{\theta_j}(y, x) - \frac{q_h[y, x, \alpha_o, w]}{q_o(y, x, \alpha_o)} \right) u|x \right\},
\]

\[
\Delta^*_{fw}(u, y_2, x) = t^*_f(u, y_2, x) - u't^*_f(x).
\]

By Assumption 4.1(iii), \( E[t^*_f(u)t^*_f(u)'|x] \) is nonsingular. By the assumptions for Theorem 6.1, \( E[t^*_f(u, y_2, x)^2|x] \) is finite. It is straightforward to show that \( \Delta^*_{fw}(u, y_2, x) \in \Lambda_f \). Now by differentiating both sides of \( \rho(z, \alpha_o)q_o(y, x, \alpha_o)dy \equiv 0 \) with respect to \( \theta_o \), we obtain

\[
E[S_{\theta_j}(y, x)u|x] = E\left\{ \frac{d\rho(z, \alpha_o)}{d\theta_j} |x \right\}.
\]

Similarly,

\[
E\left\{ \frac{q_h[y, x, \alpha_o, w]}{q_o(y, x, \alpha_o)} u|x \right\} = E\left\{ \frac{d\rho(z, \alpha_o)}{dh} [w(\cdot)] |x \right\}.
\]

Also, note that

\[
E\left\{ S_{\theta_j}(y, x) - \frac{q_h[y, x, \alpha_o, w]}{q_o(y, x, \alpha_o)} \right\}|u, y_2, x = S_{\theta_j}(y, x) - \frac{q_h[y, x, \alpha_o, w]}{q_o(y, x, \alpha_o)}.
\]

Thus the projection residual is

\[
u'\Sigma_o(x)^{-1} \left( E\left\{ \frac{d\rho(z, \alpha_o)}{d\theta_j} |x \right\} - E\left\{ \frac{d\rho(z, \alpha_o)}{dh} [w(\cdot)] |x \right\} \right) \Sigma_o(x)^{-1}.
\]

Combining all these results, we obtain \( w_0(\cdot) \) solving

\[
\inf_{w(.) \in \mathcal{W}} \left\{ \left( E\left\{ \frac{d\rho(z, \alpha_o)}{d\theta_j} - \frac{d\rho(z, \alpha_o)}{dh} [w(\cdot)] |x \right\} \right)' \Sigma_o(x)^{-1} \left( E\left\{ \frac{d\rho(z, \alpha_o)}{d\theta_j} - \frac{d\rho(z, \alpha_o)}{dh} [w(\cdot)] |x \right\} \right) \right\}.
\]
This completes the proof. Q.E.D.

**Corollary E.1:** Suppose that Assumptions 3.1-3.2, 3.7-3.8, 5.1, and 6.1 hold. Then: \( \hat{\Sigma}_\alpha(X, \alpha) = \Sigma_\alpha(X, \alpha) + o_p(n^{-1/4}) \) uniformly over \( X \in \mathcal{X}, \alpha \in \mathcal{N}_o. \)

**Proof:** The result follows from applying Parts A and B of Lemma A.1 to each element of \( \rho(Z, \alpha) \rho(Z, \alpha) \) with \( c_{1n} = 1, \delta_{1n} = \delta_{2n} = o(n^{-1/4}). \) Q.E.D.

**Proof of Theorem 6.2:** By Lemma 3.1 and Theorem 3.1, the SMD estimator \( \hat{\alpha}_n \) in Step 1 satisfies \( \| \hat{\alpha}_n - \alpha_0 \| = o_p(1) \) and \( \| \hat{\alpha}_n - \alpha_0 \| = o_p(n^{-1/4}). \) Hence \( \hat{\alpha}_n \in \mathcal{N}_o. \) Using the proof similar to those of Theorem 3.1, we can also show that \( \hat{\alpha}_n \in \mathcal{N}_o. \) For any fixed nonzero \( \lambda \in \mathcal{R}^d, \) we have

\[
\lambda' (\theta - \theta_0) = E \left( \left\{ \frac{d m(X, \alpha_0)}{d \alpha} \right\}_{v_0} \left[ \Sigma_\alpha(X) \right]^{-1} \frac{d m(X, \alpha_0)}{d \alpha} [\alpha - \alpha_0] \right) \quad \text{for all } \alpha \in \mathcal{A}.
\]

Let \( \varepsilon_n = o(n^{-1/2}) > 0 \) and \( u = \pm v_0. \) To simplify notation, denote \( v_{n0} = \Pi_{n0} v_0. \) We take a continuous path \( \alpha(t) = \tilde{\alpha} \pm t \varepsilon_n v_{n0}. \) Then \( \{ \alpha(t) : t \in [0, 1] \} \in \mathcal{N}_o. \) Let

\[
\tilde{L}_n(\alpha(t)) = -\frac{1}{2n} \sum_{i=1}^{n} \tilde{m}(X_i, \alpha(t)) \left[ \tilde{\Sigma}_\alpha(X_i, \alpha(t)) \right]^{-1} \tilde{m}(X_i, \alpha(t)).
\]

By definition of \( \tilde{\alpha}, \) and a Taylor expansion around \( t = 0 \) up to second order, we obtain

\[
0 \leq \tilde{L}_n(\tilde{\alpha}) - \tilde{L}_n(\tilde{\alpha} \pm \varepsilon_n v_{n0}) = -\frac{d \tilde{L}_n(\alpha(t))}{dt} \bigg|_{t=0} - \frac{1}{2} \frac{d^2 \tilde{L}_n(\alpha(t))}{dt^2} \bigg|_{t=0}
\]

for some \( s \in [0, 1], \) where

\[
-\frac{d \tilde{L}_n(\alpha(t))}{dt} \bigg|_{t=0} = \frac{\pm \varepsilon_n}{n} \sum_{i=1}^{n} \left\{ \frac{d \tilde{m}(X_i, \tilde{\alpha})}{d \alpha} \right\}_{v_{n0}} \left[ \tilde{\Sigma}_\alpha(X_i, \tilde{\alpha}) \right]^{-1} \tilde{m}(X_i, \tilde{\alpha})
\]

\[
+ \frac{\pm \varepsilon_n}{2n} \sum_{i=1}^{n} \left\{ \tilde{m}(X_i, \tilde{\alpha}) \right\}^{\prime} \left\{ \frac{d^2 \tilde{\Sigma}_\alpha(X_i, \tilde{\alpha})}{d \alpha} \right\}_{v_{n0}} \tilde{m}(X_i, \tilde{\alpha}),
\]

\[
-\frac{d^2 \tilde{L}_n(\alpha(t))}{dt^2} \bigg|_{t=0} = \frac{\varepsilon_n^2}{n} \sum_{i=1}^{n} \left\{ \frac{d^2 \tilde{m}(X_i, \alpha(s))}{d \alpha d \alpha} \right\}_{v_{n0}, v_{n0}} \left[ \tilde{\Sigma}_\alpha(X_i, \alpha(s)) \right]^{-1} \tilde{m}(X_i, \alpha(s))
\]

\[
+ \frac{\varepsilon_n^2}{n} \sum_{i=1}^{n} \left\{ \frac{d \tilde{m}(X_i, \alpha(s))}{d \alpha} \right\}_{v_{n0}} \left[ \tilde{\Sigma}_\alpha(X_i, \alpha(s)) \right]^{-1} \left\{ \frac{d \tilde{m}(X_i, \alpha(s))}{d \alpha} \right\}_{v_{n0}}
\]

\[
+ \frac{2\varepsilon_n^2}{n} \sum_{i=1}^{n} \left\{ \frac{d \tilde{m}(X_i, \alpha(s))}{d \alpha} \right\}_{v_{n0}} \left[ \tilde{\Sigma}_\alpha(X_i, \alpha(s))^{-1} \right]_{v_{n0}} \tilde{m}(X_i, \alpha(s))
\]

\[
+ \frac{\varepsilon_n^2}{2n} \sum_{i=1}^{n} \left\{ \tilde{m}(X_i, \alpha(s)) \right\}^{\prime} \left\{ \frac{d^2 \tilde{\Sigma}_\alpha(X_i, \alpha(s))}{d \alpha d \alpha} \right\}_{v_{n0}, v_{n0}} \tilde{m}(X_i, \alpha(s)).
\]
MODELS WITH CONDITIONAL MOMENT RESTRICTIONS

where \( \alpha(s) = \alpha + s\epsilon_s v_{no} \in \mathbb{N}_{on} \). Applying Corollaries A.2, C.2, and Assumptions 3.6(ii), 4.3(i), 4.6, and 6.3, all second derivative terms are \( O_p(\epsilon_s^2) \). Moreover, since \( \epsilon_s = o(n^{-1/2}) > 0 \), we obtain uniformly over \( \alpha(s) \in \mathbb{N}_{on} \):

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \hat{\alpha})}{d\alpha} [v_{no}] \right\}' \left[ \hat{\Sigma}_s(X_i, \hat{\alpha}) \right]^{-1} \hat{m}(X_i, \hat{\alpha}) = o_p(n^{-1/2}).
\]

Write

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \hat{\alpha})}{d\alpha} [v_{no}] \right\}' \left[ \hat{\Sigma}_s(X_i, \hat{\alpha}) \right]^{-1} \hat{m}(X_i, \hat{\alpha})
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \hat{\alpha})}{d\alpha} [v_{no}] - \frac{d\hat{m}(X_i, \hat{\alpha})}{d\alpha} [v_{no}] \right\}' \left[ \hat{\Sigma}_s(X_i, \hat{\alpha}) \right]^{-1} \hat{m}(X_i, \hat{\alpha})
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \hat{\alpha})}{d\alpha} [v_{no}] \right\}' \left[ (\hat{\Sigma}_s(X_i, \hat{\alpha}) - \Sigma_0(X_i, \hat{\alpha}))^{-1} \right] \hat{m}(X_i, \hat{\alpha})
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \hat{\alpha})}{d\alpha} [v_{no}] \right\}' \left[ \Sigma_0(X_i, \hat{\alpha}) \right]^{-1} \hat{m}(X_i, \hat{\alpha}).
\]

Applying Corollaries A.2, C.1(i), and E.1, the first two terms on the right-hand side are \( o_p(n^{-1/2}) \).

Denote

\[ g(X, \hat{\alpha}, v_o) = \left\{ \frac{d\hat{m}(X, \hat{\alpha})}{d\alpha} [v_{no}] \right\}' \left[ \hat{\Sigma}_s(X, \hat{\alpha}) \right]^{-1}. \]

Using proofs similar to those of Corollary C.1(ii), we can show

\[
\frac{1}{n} \sum_{i=1}^{n} \| g(X_i, \hat{\alpha}, v_o) - g(X_i, \alpha_o, v_o) \|_E^2 = o_p(n^{-1/2}) \quad \text{and}
\]

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \alpha_o)}{d\alpha} [v_{no}] \right\}' \left[ \Sigma_0(X_i, \alpha_o) \right]^{-1} \hat{m}(X_i, \hat{\alpha}) = o_p(n^{-1/2}).
\]

Furthermore, by Assumption 4.2(i) and Corollary A.2, we can replace \( v_{no} \) with \( v_o \) to obtain

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \alpha_o)}{d\alpha} [v_o] \right\}' \left[ \Sigma_0(X_i, \alpha_o) \right]^{-1} \hat{m}(X_i, \hat{\alpha}) = o_p(n^{-1/2}).
\]

By Corollary C.3(ii) and (iii),

\[
\sqrt{n} \lambda' (\tilde{\theta} - \theta_o) = - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{d\hat{m}(X_i, \alpha_o)}{d\alpha} [v_o] \right\}' \left[ \Sigma_0(X_i) \right]^{-1} \rho(Z_i, \alpha_o) + o_p(1).
\]

Theorem 6.2 follows from applying a standard CLT for i.i.d. data. \( Q.E.D. \)

**Proof of Proposition 6.2:** We verify that all assumptions of Theorem 6.2 are satisfied. Conditions 6.2.1(ii) and 6.2.2(i) imply Assumption 5.1. Condition 6.2.1(i) implies Assumption 6.1. Conditions 3.2.6 and 6.2.1(ii) imply

\[
E \left\{ \left\| \frac{d\hat{m}(X, \alpha)}{d\alpha} [v_{no}] - \frac{d\hat{m}(X, \alpha_o)}{d\alpha} [v_{no}] \right\|_E^2 \right\} = o(n^{-1/2}).
\]
Assumption 4.1(iii), conditions 3.2.1(iii), 3.2.3, and 6.2.1(ii) imply
\[
E\left(\frac{d m(X, \alpha_0)}{d \alpha} \right) = E\left(\frac{d m(X, \alpha)}{d \alpha} \right) = E\left(\frac{d m(X, \alpha_0)}{d \alpha} - \frac{d m(X, \alpha)}{d \alpha} \right) = \sum \left( \begin{array}{c} \Sigma^{-1}(X, \alpha) - \Sigma^{-1}(X, \alpha_0) \end{array} \right)^2
\]
\[
\leq \text{const.} E\left(\left(\Sigma(X, \alpha) - \Sigma(X, \alpha_0)\right)^2\right).
\]

By the mean value theorem and condition 3.2.2(ii), for all \( \alpha \in \mathcal{A} \),
\[
\Sigma(X, \alpha) - \Sigma(X, \alpha_0)
\]
\[
= E\left(\left(\begin{array}{c} F \left( Y_2^2 \theta_0 + \sum_{i=1}^{q} h_{ij}(X_i) \right) - F \left( Y_2^2 \theta + \sum_{i=1}^{q} h_i(X_i) \right) \end{array} \right)^2 \right) = E\left(\left(\left( Y_2^2 \theta - \theta_0 \right) + \sum_{j=1}^{q} \left( h_j(X_j) - h_{ij}(X_j) \right) \right)^2 \right) \leq \text{const.} \left( \left\| \theta - \theta_0 \right\|_2^2 + E\left( \left( \sum_{j=1}^{q} \left( h_j(X_j) - h_{ij}(X_j) \right) \right)^2 \right) \right) = o(n^{-1/2}).
\]

Thus Assumption 6.2 is satisfied. Assumption 6.3 follows from condition 6.2.2.

**REFERENCES**


Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions
Chunrong Ai; Xiaohong Chen
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28200311%2971%3A6%3C1795%3AEEOMWC%3E2.0.CO%3B2-H

This article references the following linked citations. If you are trying to access articles from an off-campus location, you may be required to first logon via your library web site to access JSTOR. Please visit your library's website or contact a librarian to learn about options for remote access to JSTOR.

[Footnotes]

2 Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity
Donald W. K. Andrews
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199401%2962%3A1%3C43%3AAFSEMV%3E2.0.CO%3B2-0

2 The Asymptotic Variance of Semiparametric Estimators
Whitney K. Newey
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199411%2962%3A6%3C1349%3ATAVOSE%3E2.0.CO%3B2-F

2 Sieve Extremum Estimates for Weakly Dependent Data
Xiaohong Chen; Xiaotong Shen
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199803%2966%3A2%3C289%3ASEEFWD%3E2.0.CO%3B2-2

2 Estimation of Semiparametric Models When the Criterion Function Is Not Smooth
Xiaohong Chen; Oliver Linton; Ingrid Van Keilegom
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28200309%2971%3A5%3C1591%3AOSMWT%3E2.0.CO%3B2-A

NOTE: The reference numbering from the original has been maintained in this citation list.
3 Root-N-Consistent Semiparametric Regression
P. M. Robinson
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28198807%2956%3A4%3C931%3ARSR%3E2.0.CO%3B2-3

3 An Efficient Semiparametric Estimator for Binary Response Models
Roger W. Klein; Richard H. Spady
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199303%2961%3A2%3C387%3AAESEFB%3E2.0.CO%3B2-G

References

A Semiparametric Maximum Likelihood Estimator
Chunrong Ai
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199707%2965%3A4%3C933%3AASMLE%3E2.0.CO%3B2-0

Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity
Donald W. K. Andrews
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199401%2962%3A1%3C43%3AAESEM%3E2.0.CO%3B2-0

Efficiency Bounds for Semiparametric Regression
Gary Chamberlain
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199205%2960%3A3%3C567%3AEBFSR%3E2.0.CO%3B2-2

NOTE: The reference numbering from the original has been maintained in this citation list.
Sieve Extremum Estimates for Weakly Dependent Data
Xiaohong Chen; Xiaotong Shen
Stable URL: [http://links.jstor.org/sici?sici=0012-9682%28199803%2966%3A2%3C289%3ASEEFWD%3E2.0.CO%3B2-2](http://links.jstor.org/sici?sici=0012-9682%28199803%2966%3A2%3C289%3ASEEFWD%3E2.0.CO%3B2-2)

Estimation of Semiparametric Models When the Criterion Function Is Not Smooth
Xiaohong Chen; Oliver Linton; Ingrid Van Keilegom
Stable URL: [http://links.jstor.org/sici?sici=0012-9682%28200309%2971%3A5%3C1591%3AESMWT%3E2.0.CO%3B2-A](http://links.jstor.org/sici?sici=0012-9682%28200309%2971%3A5%3C1591%3AESMWT%3E2.0.CO%3B2-A)

Convergence Rates of SNP Density Estimators
Victor M. Fenton; A. Ronald Gallant
Stable URL: [http://links.jstor.org/sici?sici=0012-9682%28199605%2964%3A3%3C719%3ACROSDE%3E2.0.CO%3B2-O](http://links.jstor.org/sici?sici=0012-9682%28199605%2964%3A3%3C719%3ACROSDE%3E2.0.CO%3B2-O)

Semi-Nonparametric Maximum Likelihood Estimation
A. Ronald Gallant; Douglas W. Nychka
Stable URL: [http://links.jstor.org/sici?sici=0012-9682%28198703%2955%3A2%3C363%3ASNMLE%3E2.0.CO%3B2-R](http://links.jstor.org/sici?sici=0012-9682%28198703%2955%3A2%3C363%3ASNMLE%3E2.0.CO%3B2-R)

Large Sample Properties of Generalized Method of Moments Estimators
Lars Peter Hansen
Stable URL: [http://links.jstor.org/sici?sici=0012-9682%28198207%2950%3A4%3C1029%3AGMOM%3E2.0.CO%3B2-O](http://links.jstor.org/sici?sici=0012-9682%28198207%2950%3A4%3C1029%3AGMOM%3E2.0.CO%3B2-O)

Finite-Sample Properties of Some Alternative GMM Estimators
Lars Peter Hansen; John Heaton; Amir Yaron
Stable URL: [http://links.jstor.org/sici?sici=0735-0015%28199607%2914%3A3%3C262%3AFOSG%3E2.0.CO%3B2-P](http://links.jstor.org/sici?sici=0735-0015%28199607%2914%3A3%3C262%3AFOSG%3E2.0.CO%3B2-P)

**NOTE:** The reference numbering from the original has been maintained in this citation list.
An Efficient Semiparametric Estimator for Binary Response Models
Roger W. Klein; Richard H. Spady
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199303%2961%3A2%3C387%3AAESEFB%3E2.0.CO%3B2-G

Efficient Instrumental Variables Estimation of Nonlinear Models
Whitney K. Newey
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199007%2958%3A4%3C809%3AEIVEON%3E2.0.CO%3B2-Y

Semiparametric Efficiency Bounds
Whitney K. Newey
Stable URL: http://links.jstor.org/sici?sici=0883-7252%28199004%2F06%295%3A2%3C99%3ASEB%3E2.0.CO%3B2-G

The Asymptotic Variance of Semiparametric Estimators
Whitney K. Newey
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199411%2962%3A6%3C1349%3ATAVOSE%3E2.0.CO%3B2-F

Instrumental Variable Estimation of Nonparametric Models
Whitney K. Newey; James L. Powell
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28200309%2971%3A5%3C1565%3AIVEON%3E2.0.CO%3B2-S

Nonparametric Estimation of Triangular Simultaneous Equations Models
Whitney K. Newey; James L. Powell; Francis Vella
Stable URL: http://links.jstor.org/sici?sici=0012-9682%28199905%2967%3A3%3C565%3ANOTSE%3E2.0.CO%3B2-Q

NOTE: The reference numbering from the original has been maintained in this citation list.
Semiparametric Estimation of Index Coefficients
James L. Powell; James H. Stock; Thomas M. Stoker
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198911%2957%3A6%3C1403%3ASEOIC%3E2.0.CO%3B2-R

Root-N-Consistent Semiparametric Regression
P. M. Robinson
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198807%2956%3A4%3C931%3ARSR%3E2.0.CO%3B2-3

Profile Likelihood and Conditionally Parametric Models
Thomas A. Severini; Wing Hung Wong
Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199212%2920%3A4%3C1768%3APLACPM%3E2.0.CO%3B2-H

On Methods of Sieves and Penalization
Xiaotong Shen
Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199712%2925%3A6%3C2555%3AOMOSAP%3E2.0.CO%3B2-C

Convergence Rate of Sieve Estimates
Xiaotong Shen; Wing Hung Wong
Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199406%2922%3A2%3C580%3ACROSE%3E2.0.CO%3B2-8

On Differentiable Functionals
Aad Van Der Vaart
Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199103%2919%3A1%3C178%3AODE%3E2.0.CO%3B2-O

NOTE: The reference numbering from the original has been maintained in this citation list.
On Maximum Likelihood Estimation in Infinite Dimensional Parameter Spaces
Wing Hung Wong; Thomas A. Severini
Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199106%2919%3A2%3C603%3AOMLEII%3E2.0.CO%3B2-7

NOTE: The reference numbering from the original has been maintained in this citation list.