Sieve Extremum Estimates for Weakly Dependent Data
Author(s): Xiaohong Chen and Xiaotong Shen
Source: *Econometrica*, Vol. 66, No. 2 (Mar., 1998), pp. 289-314
Published by: The Econometric Society
Stable URL: http://www.jstor.org/stable/2998559
Accessed: 10-06-2016 01:20 UTC

Econometrica, Vol. 66, No. 2 (March, 1998), 289–314

SIEVE EXTREMUM ESTIMATES FOR WEAKLY DEPENDENT DATA

BY XIAOHONG CHEN AND XIAOTONG SHEN[1]

Many non/semi-parametric time series estimates may be regarded as different forms of *sieve extremum estimates*. For stationary β-*mixing* observations, we obtain convergence rates of sieve extremum estimates and root-n asymptotic normality of "plug-in" sieve extremum estimates of smooth functionals. As applications to time series models, we give convergence rates for nonparametric ARX(p, q) regression via neural networks, splines, and wavelets; root-n asymptotic normality for partial linear additive AR(p) models, and monotone transformation AR(1) models.

KEYWORDS: Sieve extremum estimates, β-mixing, rate and normality, neural networks, wavelets, shape-preserving splines.

1. INTRODUCTION

MANY NON/SEMI-PARAMETRIC METHODS such as Fourier series, orthogonal polynomials, splines, neural networks, and wavelets can be regarded as special cases of the method of *sieve extremum estimation* (Grenander (1981)). The sieve method optimizes an empirical criterion over a sequence of approximating parameter spaces (a sieve) which is dense in the underlying (possibly infinite-dimensional) parameter space Θ. This method is very flexible with different choices of criteria and sieves. For example, it can incorporate economic restrictions such as nonnegativity, monotonicity, and convexity into estimation. Unlike the standard infinite-dimensional maximum likelihood (ML) estimation method, which may yield inconsistency and slow rates of convergence,[2] the sieve method can achieve optimal rates of convergence.

Several authors have studied asymptotic theories of general sieve extremum estimates.[3] For independently (and mostly identically) distributed observations, Geman and Hwang (1982) establish consistency of the sieve estimates; Shen and Wong (1994) and Birgé and Massart (1994) obtain convergence rates of the sieve estimates; Shen (1997) develops asymptotic normality and efficiency of the

[1] The authors are indebted to a co-editor, two anonymous referees, and especially to Lars Hansen for detailed constructive written comments. They are also grateful to Andrew Barron, John Curran, Yanqin Fan, Yuichi Kitamura, and Whitney Newey for helpful suggestions. The authors are responsible for remaining shortcomings. Research of the second author is supported in part by a grant from NSF.

[2] See, e.g., Grenander (1981) for examples of nonexistence and inconsistency, and Shen and Wong (1994) and Birgé and Massart (1994) for examples of slow rates of convergence of the standard (infinite-dimensional) ML estimates.

[3] There are many papers on asymptotic properties of specific sieve estimates in econometrics; see, e.g., Andrews (1991) for series, Gallant and Nychka (1987) and Fenton and Gallant (1996) for semi-nonparametric ML density estimation via Hermite series, and White (1990) for neural networks, etc.

"plug-in" sieve ML estimates of smooth functionals.[4] For weakly dependent data, White and Wooldridge (1991) establish consistency of sieve estremum estimates. To apply the sieve method in econometric time series models, one needs asymptotic results beyond consistency for dependent data.

In this paper, we provide a general theory on the convergence rate of sieve extremum estimates and root-n asymptotic normality of "plug-in" sieve estimates of smooth functionals[5] for time series stationary β-mixing observations. Under a set of sufficient conditions similar to those in Shen and Wong (1994) for i.i.d. data, we achieve the same convergence rates for β-mixing data as those for i.i.d. data. When specializing in time series regression problems, our theory yields convergence rates for many different sieves including Fourier series, polynomials, splines, neural networks, and wavelets. In addition, when the underlying parameter space Θ is similar to the one studied by Stone (1982), the rates agree with the optimal rates obtained by Stone (1982) for i.i.d. data. We obtain the rate result for β-mixing observations by first establishing a tight uniform exponential probability inequality for empirical processes indexed by a general class of functions. Such an inequality also allows us to establish stochastic equicontinuity with desired rate, which in turn allows us to extend Shen's (1997) root-n asymptotic normality result for i.i.d. data to β-mixing data. The normality result applies to essentially any plug-in sieve estimates of smooth functionals in time series models. In particular, it is applicable when the nonparametric sieve extremum estimate does not satisfy an infinite-dimensional score equation, or when the true parameter is not an interior point of the parameter space, which often occurs in constrained optimization problems.

Our approach is parallel to Wong and Severini's (1991) on nonsieve ML estimation for i.i.d. data, but it differs from approaches such as Andrews' (1994) and Newey's (1994) commonly used in the semiparametric econometrics literature. The latter two papers assume that there exist some nonparametric estimates and that the finite-dimensional score equations satisfy some stochastic equicontinuity conditions. Andrews' (1994) gives general sufficient conditions for root-n normality of the parametric components; Newey (1994) gives a general procedure to compute the asymptotic variance of the parametric components. We provide a theory for time series models when the parameters of interest include both parametric and nonparametric components. In particular, our results allow for the joint estimation of the parametric and nonparametric components in semiparametric models with and without constraints.

Because many macroeconometric and financial econometrics problems can be formulated as different forms of sieve extremum estimation problems, and because many time series models such as nonlinear ARX, nonlinear ARCH, and diffusion models may generate stationary β-mixing observations, the theory developed here is widely applicable. To keep the paper to a reasonable length,

[4] Wong and Severini (1991) and Birgé and Massart (1993) establish asymptotic properties of standard (nonsieve) (infinite-dimensional) ML estimates with independent observations.

[5] It is simply $f(\hat{\theta}_n)$ when $f(\theta_0)$ is a smooth functional of θ_0 and $\hat{\theta}_n$ is a sieve estimate of θ_0.

we only present regression applications. Example 1 is a nonparametric ARX(p, q) regression model. We give primitive sufficient conditions to ensure stationary β-mixing. We estimate the conditional mean function by three different types of sieves: neural networks, splines, and wavelets. We observe that there is not a universal "best" sieve in terms of convergence rates, because the rates depend upon prior assumptions about the parameter space Θ to which the true θ_0 belongs. For example, we obtain a new neural network convergence rate which is faster than the rates based on other sieves such as orthogonal series, splines, and wavelets in a high dimension problem; however the other sieves could be dense in parameter spaces larger than that of neural networks. Examples 2 and 3 illustrate that root-n normality of the plug-in sieve estimate $f(\hat{\theta}_n)$ can be estimated when the sieve convergence rate is not too slow. Example 2 is a partial linear additive AR(p) regression model without/with monotone constraint. Example 3 is a monotone transformation model with a lagged endogenous regressor. Here we assume that the unknown transformation function is continuous and monotone but could have kinks.

The organization of the paper is as follows. Section 2 defines the sieve extremum estimates and discusses the concept of β-mixing for stationary dependent observations. In addition, we get three regression examples to illustrate our general results. Section 3 presents the convergence rate theorem for sieve extremum estimates with dependent observations. Section 4 provides the root-n asymptotic normality theorem for plug-in sieve extremum estimates of any smooth functionals. Section 5 applies the main theorems to the examples in Section 2. We state the convergence rates and root-n asymptotic normality under relatively primitive sufficient conditions. Section 6 is a brief summary. The Appendix contains all the technical proofs.

2. DEFINITIONS AND EXAMPLES

2.1. *Method of Sieves and Mixing Processes*

We first introduce some notation. Let $\{Y_t\}_{t=1}^n$ be a p-dimensional stationary process with a marginal probability measure $P_0(\cdot)$. Let Θ be a parameter space (possibly infinite dimensional), containing the true parameter θ_0. Here $\theta_0 \in \Theta$ is defined as $E[L_n(\theta_0)] \geq E[L_n(\theta)]$ for all $\theta \in \Theta$, where $E[\cdot]$ is the expectation under P_0, and $L_n(\theta) \equiv n^{-1} \sum_{t=1}^n l(\theta, Y_t)$ with $l: \Theta \times \mathscr{R}^p \to \mathscr{R}$ the empirical criterion based on a single observation. Let $f: \Theta \to \mathscr{R}$ be a known functional. In many economic problems both θ_0 and $f(\theta_0)$ are of interest. For example, in a semiparametric model, when $\theta_0 = (\alpha_0, \eta_0)$, $f(\theta_0)$ may be the parametric component α_0; see e.g., Heckman and Singer (1984) for the mixed hazard model, Robinson (1988) for the partial linear model, and Ichimura (1993) for the single index model. In a stationary Markov model, when θ_0 is the conditional density, $f(\theta_0)$ may be certain conditional moments; see, e.g., Gallant, Rossi, and Tauchen (1993).

One approach to estimating θ_0 is to maximize L_n over Θ, and the maximizer $\hat{\theta}_n$ is called an extremum (Amemiya (1985)) estimate. When Θ is infinite-dimensional, optimization over the entire parameter space is difficult. Instead, optimization is often restricted to a sequence of approximating spaces Θ_n, a sieve according to Grenander (1981), such that $\{\Theta_n\}$ is dense in Θ as $n \to \infty$ (i.e., for any $\theta \in \Theta$, there exists $\pi_n \theta \in \Theta_n$ such that $d(\theta, \pi_n \theta) \to 0$ as $n \to \infty$, where d is a pseudo-distance). An *approximate sieve extremum* estimate $\hat{\theta}_n$ is defined as an approximate maximizer of $L_n(\theta)$ over Θ_n, i.e.,

$$L_n\left(\hat{\theta}_n\right) \geq \sup_{\theta \in \Theta_n} L_n(\theta) - O(\varepsilon_n^2), \tag{2.1}$$

where $\varepsilon_n \to 0$ as $n \to \infty$. For the *exact* estimate, $\varepsilon_n = 0$. The $f(\hat{\theta}_n)$ is defined as the *plug-in* sieve estimate for $f(\theta_0)$. The sieve extremum estimation method includes the standard extremum estimation method by setting $\Theta_n = \Theta$ for all n.

In this paper, we study asymptotic properties of sieve extremum estimates with stationary time series observations. A stochastic sequence $\{Y_t\}_{=-\infty}^{\infty}$ is called *β-mixing* if $\beta(j) \to 0$ as $j \to \infty$, where

$$\beta(j) \equiv \sup_t E \sup\left\{|P(B \mid \mathcal{F}_{-\infty}^t) - P(B)| : B \in \mathcal{F}_{t+j}^{\infty}\right\},$$

and $\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+j}^{\infty}$ are the σ-fields generated respectively by $(Y_{-\infty}, \ldots, Y_t)$ and $(Y_{t+j}, \ldots, Y_{\infty})$. Davydov (1973) gives the following equivalent definition of β-mixing for a stationary homogeneous Markov process $\{Y_t\}_{t=0}^{\infty}$:

$$\beta(j) \equiv \int |P^j(x, \cdot) - \nu_0(\cdot)| \, \nu_0(dx) \to 0 \quad \text{as } j \to \infty,$$

where $|\cdot|$ is the total variation of a signed measure; ν_0 is the stationary marginal distribution of Y_0, and $P^j(x, A) \equiv \Pr(Y_j \in A \mid Y_0 = x)$, the j-step transition kernel.

Many nonlinear Markov processes have been shown to satisfy stationary β-mixing with exponential decay rates (i.e., $\beta(j) \leq \beta_0 \exp(-cj)$ for some $\beta_0, c > 0$); see, e.g., Doukhan (1994) for nonlinear ARX(p, q) models and certain nonlinear ARCH models and Meyn and Tweedie (1992) for general sufficient conditions.

2.2. *Regression Examples*

We now present three examples to illustrate the usefulness of our main results in Sections 3 and 4. We discuss additional applications and different types of sieves in Section 5.

EXAMPLE 1 (Nonparametric ARX(p, q) Model): Suppose $\{Y_t\}_{t=1}^n$ is generated according to

$$Y_t = \theta_0(Y_{t-1}, \ldots, Y_{t-p}, X_t, \ldots, X_{t-q+1}) + e_t, \tag{2.2}$$

where $\{X_t\}$ and $\{e_t\}$ are independent with $E[e_t] = 0$; $\{X_t \in \mathscr{R}^k\}$ and $\{e_t \in \mathscr{R}\}$ are both i.i.d. for simplicity. The function $\theta_0 : \mathscr{R}^p \times \mathscr{R}^{kq} \to \mathscr{R}$ is the parameter of interest, where $p, k, q \geq 1$ are fixed known integers. $\{Y_t\}$ is β-mixing under certain conditions on θ_0, $\{X_t\}$, and $\{e_t\}$ to be specified in Section 5.

Denote $Z_t = (Y_{t-1}, \ldots, Y_{t-p}, X_t, \ldots, X_{t-q+1}) \in \mathscr{R}^d$, with $d \equiv p + kq$, and denote $\|\theta - \theta_0\|^2 = E[\theta(Z_t) - \theta_0(Z_t)]^2$. Here θ_0 is estimated using $l(\theta, Y_t, Z_t) = -\frac{1}{2}[Y_t - \theta(Z_t)]^2$.

Case 1.1 (Sigmoid neural networks): Suppose $\theta_0 \in \Theta$, the space of functions which have finite first absolute moments of the Fourier magnitude distributions, i.e.,

$$\Theta = \left\{ \theta : \mathscr{R}^d \to \mathscr{R} : \theta(z) = \int \exp(ia^T z) d\mu_\theta(a), \right.$$
$$\left. \|\mu_\theta\|_1 \equiv \int [\max(|a|, 1)] d|\mu_\theta|(a) \leq C \right\}, \tag{2.3}$$

where μ_θ is a complex-valued measure on $\mathscr{R}^d$, $|\mu_\theta|$ denotes the total variation of μ_θ, and $|a| \equiv \sum_{i=1}^d |a_i|$ for $a^T = (a_1, \ldots, a_d) \in \mathscr{R}$. See Section 5 for discussions about this space.

Suppose Z_t has the (unknown) distribution F_0 with a compact support in $\mathscr{R}^d$, and we estimate θ_0 using the following neural network sieve:

$$\Theta_n = \left\{ \theta \in \Theta : \theta(z) = b_0 + \sum_{j=1}^{r_n} b_j \psi\left(a_j^T z + a_{0,j}\right), \right.$$
$$\left. \sum_{j=0}^{r_n} |b_j| \leq c_n, \max_{1 \leq j \leq r_n} \sum_{i=0}^{d} |a_{i,j}| \leq \tilde{c}_n \right\}, \tag{2.4}$$

where ψ is a *sigmoid* function, i.e., a bounded measurable function on $\mathscr{R}$ with $\psi(u) \to 1$ as $u \to \infty$; and $\psi(u) \to 0$ as $u \to -\infty$. Cybenko (1989) established the denseness of $\bigcup_n \Theta_n$ in Θ. Barron (1993) obtained a deterministic root mean square approximation error rate $\|\theta - \pi_n \theta\| = O((r_n)^{-1/2})$. Recently Makovoz (1996) improved Barron's result to $\|\theta - \pi_n \theta\| = O((r_n)^{-1/2 - 1/(2d)})$.

Denote $\hat{\theta}_n = \operatorname{argmax}_{\Theta_n} L_n(\theta)$ with Θ_n given by (2.4). By applying our convergence rate result for β-mixing (Theorem 1 in Section 3) and Makovoz's (1996) approximation rate, we obtain a neural network sieve convergence rate $\|\hat{\theta}_n - \theta_0\| = O_p([n/\log n]^{-(1+1/d)/[4(1+1/2d)]})$. Previously by using the method of minimum description length and Barron's (1993) approximation rate, a convergence rate of $O_p([n/\log n]^{-1/4})$ has been obtained by Barron (1994) for i.i.d. data, and by Modha and Masry (1996) for m-dependent data.

By the normality result in Section 4, our neural network sieve convergence rate is fast enough to ensure root-n asymptotic normality of plug-in estimates of smooth functionals of the ARX(p, q) regression function θ_0, say linear combinations of partial average derivatives, regardless of how big the dimension d is.

In Section 5, we study the ARX(p, q) model in (2.2) using multivariate splines and wavelets, and obtain rates that differ from this neural network rate.

Both partial linear regression and additive regression models have been studied intensively in econometrics; see e.g., Robinson (1988), Andrews (1991, 1994), Newey (1994), etc. Here we use the method of sieves to illustrate our results on root-n normality of the parametric components with/without monotone constraints.

EXAMPLE 2 (Semiparametric Additive AR(p) Regression): Suppose the time series data $\{Y_t\}_{t=1}^n$ is generated according to[6]

$$(2.5)\qquad Y_t = \sum_{i=1}^{p_1} Y_{t-i}\alpha_{0,i} + \sum_{i=p_1+1}^{p} \eta_{0,i-p_1}(Y_{t-i}) + e_t,$$

$$E[e_t \mid Y_{t-i}, 1 \le i \le p] = 0,$$

$$E\left[e_t^2 \mid Y_{t-i}, 1 \le i \le p\right] = \sigma^2(Y_{t-i}, 1 \le i \le p),$$

where $\{Y_t\}$ is a stationary β-mixing sequence. The parameters of interest are $\alpha^T \equiv (\alpha_1, \ldots, \alpha_{p_1})$ and $\eta^T \equiv (\eta_1, \ldots, \eta_{p-p_1})$, where $\alpha_i \in \mathscr{R}$ and $\eta_j: \mathscr{R} \to \mathscr{R}$. Let

$$\theta = (\alpha, \eta), \quad Z_t^T \equiv (Y_{t-1}, \ldots, Y_{t-p_1}),$$

$$\eta^T(X_t) \equiv \left(\eta_1(Y_{t-(p_1+1)}), \ldots, \eta_{p-p_1}(Y_{t-p})\right), \quad \text{and}$$

$$\theta(Z_t, X_t) = Z_t^T\alpha + \eta(X_t).$$

Let

$$l(\theta, Y_t, Z_t, X_t) = -0.5\left(Y_t - (Z_t^T\alpha + \eta(X_t))\right)^2, \quad \text{and}$$

$$\|\theta - \theta_0\|^2 = E\left[(\alpha - \alpha_0)^T Z_t + (\eta(X_t) - \eta_0(X_t))\right]^2.$$

Let $\theta_0 \in \Theta = A \times D$, where $A = (-1, 1)^{p_1}$, and $D = \prod_{i=1}^{p-p_1} D_i$, $D_i = \Lambda^{m_i}$, a Hölder space with smoothness $m_i \equiv s_i + \gamma_i$, $s_i = 0, 1, \ldots,$ an integer, and $0 < \gamma_i \le 1$:

$$(2.6)\qquad D_i = \left\{ \eta \in C^{s_i}([b_1, b_2]): \right.$$

$$\left. \sup_{x, y \in [b_1, b_2], x \ne y} \frac{|\eta^{(s_i)}(x) - \eta^{(s_i)}(y)|}{|x - y|^{\gamma_i}} \le c_\eta < \infty \right\}, \quad m_i > 1/2.$$

(Since c_η depends on η, the space D_i is not compact.)

[6] The choice of which lags enter linearly or nonlinearly is flexible here. In fact, one can also replace this example by a semiparametric additive ARX(p,q) model.

Case 2.1 (Fourier Series without Monotone Constraint): Consider a sieve: $\Theta_n = A \times D_n$, with $D_n = \prod_{i=1}^{p-p_1} D_{n,i}$:

$$D_{n,i} = \Bigg\{ \eta_i(x) = \sum_{j=1}^{r_{n,i}} \big(a_{j,i}\cos(2\pi jx) + b_{j,i}\sin(2\pi jx)\big), \sum_{j=1}^{r_{n,i}} j^{2q_i}\big(a_{j,i}^2 + b_{j,i}^2\big) \le c_{n,i}^2 \Bigg\}, \tag{2.7}$$

where q_i is a constant arbitrarily close to m_i with $m_i > q_i > 0.5$; $r_{n,i}$ and $c_{n,i}$ will be specified later. Many other truncated series can also be used to approximate D_i. For each fixed $i = 1,\dots,p-p_1$, the denseness of $\bigcup_n D_{n,i}$ in D_i and the determininstic approximation rate $\|\eta - \pi_n\eta\|_\infty = O((r_{n,i})^{-m_i})$ can be found in Lorentz (1966).

Case 2.2 (Spline with Monotone Constraint): Suppose $\{Y_t\}$ is generated according to (2.5) with increasing functions $\eta_{0,j}$ for $j = 1,\dots,p-p_1$. That is, $\theta_0 \in \Theta = A \times D'$ with $D' = \prod_{i=1}^{p-p_1} D_i'$, where

$$D_i' = \{\eta \in D_i \text{ in (2.6)}: \eta' \ge 0\}, \quad s_i \ge 1, \quad m_i > 1. \tag{2.8}$$

Consider a sieve: $\Theta_n = A \times D_n'$, with $D_n' = \prod_{i=1}^{p-p_i} D_{n,i}$, where each $D_{n,i}$ is a shape-preserving sieve such as Schoenberg cardinal B-splines (DeVore and Lorentz (1993, Chapter 13)): Consider a set of knots $(b_1 < x_1 < \cdots < x_{r_n} < b_2)$ with the auxiliary knots $x_{-k+1} = \cdots = x_0 = b_1$ and $x_{r_n+1} = \cdots = x_{r_n+k} = b_2$ such that

$$0 < c_4 r_n^{-1} \le \min_{-k+1 \le j \le r_n} x_j - x_{j-1} \le \max_{-k+1 \le j \le r_n} x_j - x_{j-1} \le c_5 r_n^{-1} < \infty. \tag{2.9}$$

Let $N_{j,k}(x)$, $j \in \{-k+1,\dots,r_n\}$, with the supports $[x_j, x_{j+k}]$ be the sequence of cardinal B-splines with order k.

$$D_{n,i} = \Bigg\{ \eta_i(x) = \sum_{j=-m_i}^{r_n} a_j N_{j,m_i}(x), -c_{n,i} \le a_{-m_i} \le a_{-m_i+1} \le \cdots \le a_{r_n} \le c_{n,i} \Bigg\}, \tag{2.10}$$

where $r_n, c_{n,i} \to \infty$ as $n \to \infty$ at some orders to be specified later. For each fixed $i = 1,\dots,p-p_1$, the denseness of $\bigcup_n D_{n,i}$ in D_i' and the deterministic approximation rate $\|\eta - \pi_n\eta\|_\infty = O((r_n)^{-m_i})$ can be found in DeVore (1977).

A semi-parametric sieve extremum estimate $\hat\theta_n = (\hat\alpha_n, \hat\eta_n)$ is the maximizer of $L_n(\theta)$ over $A \times D_n$. By Theorem 1 in Section 3, $\|\hat\eta_n - \eta_0\| = O_P(n^{-m/(2m+1)})$ with $m = \min\{m_1,\dots,m_{p-p_1}\}$; by Theorem 2 in Section 4, $n^{1/2}(\hat\alpha_n - \alpha_0) \xrightarrow{P\theta_0} \mathcal{N}(0, \Omega_*)$, with Ω_* given in Section 5.

In the above example, we require that the monotone functions have at least one time continuous derivatives. Such an assumption is not necessary. In the

next example, we estimate a monotone transformation model when the unknown transformation function η_0 is only continuous and increasing; hence it can have kinks.[7]

EXAMPLE (Monotone Transformation Model): Suppose that $\{Y_t\}$ is generated according to

$$\eta_0(Y_t) = \alpha_0 Y_{t-1} + e_t, \quad \eta_0(\cdot) \text{ increasing}, \tag{2.11}$$

where $\{e_t\}$ is i.i.d. with $E[e_t] = 0$, $E[e_t^2] = 1$, e_t is absolutely continuous with respect to Lebesgue measure in real line, $\alpha_0 \in A$ (a compact set in $\mathscr{R}$), and $\eta_0(x)$ is monotone in x. Suppose that $\{Y_t\}$ has bounded support $[b_1, b_2]$. Here $\{Y_t\}$ is stationary β-mixing under some additional conditions on η_0 in Section 5.

Let $\theta_0 = (\alpha_0, \eta_0(\cdot)) \in \Theta = A \times D$, where

$$D = \left\{ \eta \in C([b_1, b_2]) : \infty > c_2 \geq \sup_x \eta'(x) \geq \inf_x \eta'(x) \geq c_1 > 0 \right\}, \tag{2.12}$$

where η' is the right derivative(s) at the kink point(s) of η. To estimate η and α, consider the following criterion: $l(\theta, Y) = -\frac{1}{2}(\eta(Y_t) - \alpha Y_{t-1})^2 + \log \eta'(Y_t)$, (which is log-likelihood when e_t is distributed as $\mathscr{N}(0,1)$). Let $2\|\theta - \theta_0\|^2$ be

$$\begin{aligned} &E[e_t([\alpha - \alpha_0]Y_{t-1} - [\eta(Y_t) - \eta_0(Y_t)]) + (\eta'(Y_t) - \eta_0'(Y_t))/\eta_0'(Y_t)]^2 \\ &\quad = E[(\alpha - \alpha_0)Y_{t-1} - (\eta(Y_t) - \eta_0(Y_t))]^2 \\ &\qquad + E[(\eta'(Y_t) - \eta_0'(Y_t))/\eta_0'(Y_t)]^2. \end{aligned}$$

Let $\Theta_n = A \times D_n$ with D_n a sieve based on step functions:

$$D_n = \left\{ \eta \in D : \eta(x) = \sum_{i=1}^{r_n} d_i 1(x_{i-1} \leq x < x_i), c_3 < d_1 < \cdots < d_{r_n} < c_4 \right\}, \tag{2.13}$$

where $(b_1 = x_0 < x_1 < \cdots < x_{r_n} = b_2)$ is the set of the knot locations satisfying restriction (2.9). For each fixed knot location, the coefficients $(\alpha, d_1, \ldots, d_{r_n})$ are estimated by maximizing $l(\theta, Y_t)$ over Θ_n with η' in $l(\theta, Y_t)$ being replaced by the corresponding finite differences. By Theorem 1 in Section 3, $\|\hat{\eta}_n - \eta_0\| = O_p(n^{-1/3})$ when r_n increases with n at a rate of $n^{1/3}$. By Theorem 2 in Section 4, $n^{1/2}(\hat{\alpha}_n - \alpha_0) \xrightarrow{P\theta_0} \mathscr{N}(0, \sigma_*^2)$, with σ_*^2 given in Section 5.

3. RATE OF CONVERGENCE

In the following, all probability calculations are made with respect to the true probability P_0. Let

$$K(\theta_0, \theta) \equiv n^{-1} \sum_{t=1}^{n} E(l(\theta_0, Y_t) - l(\theta, Y_t)) = E[l(\theta_0, Y_t) - l(\theta, Y_t)].$$

[7] When η_0 is several times continuously differentiable, Horowitz's (1996) method is applicable to estimate η as well as the error distribution. Robinson (1991) gives a semiparametric efficient estimate of the parametric part α_0 when η_0 is known but the error distribution is unknown.

Here $K(\theta_0, \theta)$ is the average Kullback-Leibler information when the criterion is a log-likelihood. Let d (sometimes we use $\|\cdot\|$) be an equivalent pseudo-metric to $K^{1/2}$ on Θ, i.e., there exist constants $c_1, c_2 > 0$ such that $c_1 K^{1/2}(\theta_0, \theta) \leq d(\theta_0, \theta) \leq c_2 K^{1/2}(\theta_0, \theta)$ for any $\theta \in \Theta$. In this paper, we study the convergence rate of the sieve estimate $\hat{\theta}_n$ under d, which automatically provides an upper bound on $d'(\theta_0, \hat{\theta}_n)$ for any pseudo-metric d' that is weaker than $K^{1/2}$.

We apply the L_2 metric entropy with bracketing to measure the size of a space. Let $\mathscr{F} = \{h(\theta, \cdot): \theta \in \Theta\}$ be a class of L_2 measurable functions (from $\mathscr{R}^p$ to $\mathscr{R}$) indexed by Θ, such that $E[h(\theta, Y)]^2 < \infty$ for all $\theta \in \Theta$. Denote $\|\cdot\|_2$ as the L_2-norm on $\mathscr{F}$, i.e., $\|h(\theta_1, Y) - h(\theta_2, Y)\|_2 = [E(h(\theta_1, Y) - h(\theta_2, Y))^2]^{1/2}$ for any $h \in \mathscr{F}$. Let $\mathscr{L}_2$ be the completion of $\mathscr{F}$ under $\|\cdot\|_2$. For any given $w > 0$, if there exists $S(w, N) = \{h_1^l, h_1^u, \ldots, h_N^l, h_N^u\} \subset \mathscr{L}_2$ such that $\max_{1 \leq j \leq N} \|h_j^u - h_j^l\|_2 \leq w$, and if for any $h \in \mathscr{F}$, there exists a $j \in \{1, \ldots, N\}$ with $h_j^l \leq h \leq h_j^u$ a.e. $-P$, then $H(w, \mathscr{F}) = \log(\min\{N: S(w, N)\})$ is defined as the bracketing L_2 metric entropy of the space $\mathscr{F}$; see, e.g., Pollard (1984) for a more detailed discussion.

Now we provide a set of sufficient conditions to obtain rates of convergence for sieve estimates with stationary observations.

Condition A

A.1. $\{Y_t\}_{t=1}^n$ *is a stationary* β*-mixing sequence with* $\beta(j) \leq \beta_0 j^{-\zeta}$ *for some* $\beta_0 > 0$, $\zeta = \gamma - 2 > 2$.

A.2. *For all small* $\varepsilon > 0$,

$$\sup_{\{\theta \in \Theta_n: d(\theta_0, \theta) \leq \varepsilon\}} \operatorname{var}(l(\theta, Y_t) - l(\theta_0, Y_t)) \leq C_1 \varepsilon^2.$$

A.3 *Let* $\mathscr{F}_n = \{l(\theta, Y_t) - l(\theta_0, Y): d(\theta_0, \theta) \leq \delta, \theta \in \Theta_n\}$. *There exists* $\delta_n \in (0, 1)$ *such that*

$$\delta_n = \sup\left\{\delta > 0: \delta^{-2} \int_{b\delta^2}^{a\delta} H^{1/2}(w, \mathscr{F}_n) dw \leq C_2 n^{1/2}\right\}.$$

A.4 *For any* $\delta > 0$, *there exist a constant* $s \in (0, 2)$ *and a measurable function* $U_n(\cdot)$ *such that*

$$\sup_{\{\theta \in \Theta_n: d(\theta_0, \theta) \leq \delta\}} |l(\theta, Y_t) - l(\theta_0, Y_t)| \leq \delta^s U_n(Y_t),$$

with $\sup_n E[U_n(Y_t)]^\gamma \leq C_3$ *for* $\gamma > 2$.

THEOREM 1—Rate: *If conditions* A.1–A.4 *hold, then* $d(\hat{\theta}_n, \theta_0) = O_p(\max(\delta_n, d(\theta_0, \pi_n \theta_0)))$.

REMARK 1: (a) The constants C_i $(i = 1, 2, 3)$ in Conditions A.2–A.4 can be allowed to depend on n. They will not affect the rates of convergence as long as they grow to infinity slowly with n. Sometimes it is useful to compute an initial

global rate for $d(\hat{\theta}_n, \theta_0)$, which allows us to restrict θ to the local neighborhood of θ_0 with C_1, C_2, and C_3 independent of n. Theorem 1 can again be applied to obtain an improved rate. Since this aspect of the theory has been illustrated in Shen and Wong (1994), and the same procedure applies to time series data, we do not repeat it here.

(b) The assumption on ζ in Condition A.1 assumes that $\gamma > 4$. In Theorem 3 in Appendix A, we state another sufficient condition on ζ without $\gamma > 4$. Moreover, from the proofs of Theorems 1 and 3, it is clear that, when $\{Y_t\}$ is stationary β-mixing with $\beta(j) \le \beta_0 \exp(-cj)$ for some β_0, $c > 0$, one only needs $\gamma > 2$ as stated in Condition A.4.

(c) Condition A.4 implies Condition A.2 if $s = 1$. Condition A.4 is often satisfied with $s = 1$ for finite-dimensional problems and $0 < s \le 1$ for infinite-dimensional problems. It is often satisfied by the interpolation relationship between L_∞ norm and L_2 norm in a function space (either Θ or Θ_n); see, e.g., Lemma 2 or the verification of Proposition 1 in Appendix B. When restricting θ to a local neighborhood of θ_0, A.4 implies that $l(\theta, Y_t)$ is "continuous" at θ_0 with respect to a metric d which is equivalent to $K^{1/2}$; by Condition A.2, $d^2(\theta, \theta_0)$ behaves locally as the variance metric. Consequently this is not a strong mode of continuity requirement on $l(\theta, Y_t)$.

(d) In applications, the final convergence rate is often obtained by setting $\delta_n \equiv d(\theta_0, \pi_n \theta_0)$. The deterministic approximation error rate $d(\theta_0, \pi_n \theta_0)$ is often available in the literature of approximation theory. To calculate δ_n from Condition A.3, an upper bound of $H(w, \mathcal{F}_n)$ suffices. For instance, L_∞ metric entropy can be used as an upper bound according to Ossiander (1987).

(e) Theorem 1 yields the same convergence rates for β-mixing observations as those in Shen and Wong (1994) for i.i.d. data. Under Conditions A.2–A.4, Chen and Shen (1996) also obtain convergence rates for strong mixing sequences; however, the rates are slower.

(f) When the criterion function $L_n(\theta)$ is uniformly bounded almost-surely, Chen and Shen (1996) have the following uniform large deviation probability bound with uniform (ϕ-) mixing data (hereafter P^* denotes the outer probability measure, and $\mu_n(g) \equiv n^{-1} \sum_{t=1}^n [g(\theta, Y_t) - E_0 g(\theta, Y_t)]$ denotes the empirical process induced by g):

COROLLARY 1: *Suppose A.2, A.3, and* $\sup_{\theta \in \Theta} \sup_t l(\theta, Y_t) \le C_3 < \infty$ *hold. Suppose* $\{Y_t\}$ *is a stationary uniform mixing sequence with* $\phi(j) \le \phi_0 j^{-\zeta}$ *for some* $\zeta > 1$. *Then, there exist constants* $c, C > 0$ *such that for any* $x \ge 1$ *and any integer* n,

$$P^*\left(\sup_{\{d(\theta_0, \theta) \ge x\varepsilon_n, \theta \in \Theta_n\}} \mu_n(l(\theta) - l(\theta_0)) \ge (x\varepsilon_n)^2/2 \right) \le 4c \exp(-Cn\varepsilon_n^2 x^2),$$

where $\varepsilon_n = \max(\delta_n, d(\theta_0, \pi_n \theta_0))$. *Hence,*

$$P\left(d\left(\hat{\theta}_n, \theta_0\right) \ge x\varepsilon_n \right) \le 4c \exp(-Cn\varepsilon_n^2 x^2).$$

4. ASYMPTOTIC NORMALITY

In this section, we derive root-n asymptotic normality for any plug-in sieve estimate $f(\hat{\theta}_n)$, which extends the result of Shen (1997) from the i.i.d case to the dependent case.

Let Θ equip with a norm $\|\cdot\|$. Suppose for all $\theta \in \Theta$ and all $y, l(\theta, y) - l(\theta_0, y)$ can be approximated by $l'_{\theta_0}[\theta - \theta_0, y]$ such that $l'_{\theta_0}[\theta - \theta_0, Y_t] - E_0 l'_{\theta_0}[\theta - \theta_0, Y_t]$ is linear in $\theta - \theta_0$. Denote the remainder of the approximation as:

$$r[\theta - \theta_0, y] \equiv l(\theta, y) - l(\theta_0, y) - l'_{\theta_0}[\theta - \theta_0, y], \tag{4.1}$$

where $l'_{\theta_0}[\theta - \theta_0, y]$ is defined as $\lim_{\tau \to 0}[(l(\theta(\theta_0, \tau), y) - l(\theta_0, y))/\tau]$, and $\theta(\theta_0, \tau) \in \Theta$ is a path in τ connecting θ_0 and θ such that $\theta(\theta_0, 0) = \theta_0$ and $\theta(\theta_0, 1) = \theta$. When $\theta_0 + \tau[\theta - \theta_0] \in \Theta$ for any $\tau \in [0, 1]$ and for all fixed $\theta \in \Theta$, we can let $\theta(\theta_0, \tau) = \theta_0 + \tau[\theta - \theta_0]$, which implies that $l'_{\theta_0}[\theta - \theta_0, y]$ is the directional derivative of l at θ_0, and is linear in $\theta - \theta_0$. Sometimes $\theta(\theta_0, \tau)$ is nonlinear in τ, which is useful in the constrained optimization problems, in which case $l'_{\theta_0}[\theta - \theta_0, y]$ may not be linear in $\theta - \theta_0$.

Suppose the functional of interest f satisfies: for any $\theta \in \Theta_n$,

$$\left| f(\theta) - f(\theta_0) - f'_{\theta_0}[\theta - \theta_0] \right| \le O\left(\|\theta - \theta_0\|^{\omega} \right), \tag{4.2}$$

uniformly in θ as $\|\theta - \theta_0\| \to 0$,

for some $\omega > 0$, and $f'_{\theta_0}[\theta - \theta_0]$ is assumed to be linear in $[\theta - \theta_0]$ such that $\|f'_{\theta_0}\| < \infty$, where

$$f'_{\theta_0}[\theta - \theta_0] \equiv \lim_{\tau \to 0} [(f(\theta(\theta_0, \tau)) - f(\theta_0))/\tau]$$

and

$$\|f'_{\theta_0}\| \equiv \sup_{\{\theta \in \Theta : \|\theta - \theta_0\| > 0\}} \frac{|f'_{\theta_0}[\theta - \theta_0]|}{\|\theta - \theta_0\|}.$$

Suppose that $\|\cdot\|$ induces an inner product $\langle \cdot, \cdot \rangle$ on the completion of the space spanned by $\Theta - \theta_0$, denoted as $\bar{V}$. By the Riesz Representation Theorem, there exists $v^* \in \bar{V}$ such that, for any $\theta \in \Theta$, $f'_{\theta_0}[\theta - \theta_0] = \langle \theta - \theta_0, v^* \rangle$.

Suppose that the sieve estimate $\hat{\theta}_n$ has a convergence rate $\|\hat{\theta}_n - \theta_0\| = o_p(\delta_n)$. For any $\theta \in \{\theta \in \Theta_n : \|\theta - \theta_0\| \le \delta_n\}$, we consider a local alternative value $\theta^*(\theta, \varepsilon_n) = (1 - \varepsilon_n)\theta + \varepsilon_n(u^* + \theta_0) \in \bar{V}$, where $u^* \equiv \pm v^*$ and $\varepsilon_n = o(n^{-1/2})$. Let $P_n(\cdot)$ be a projection of $\bar{V}$ to Θ_n. $P_n(u)$ can be chosen as $\pi_n u$ if $u \in \Theta \subset \bar{V}$.

The following conditions are sufficient for deriving asymptotic normality of $f(\hat{\theta}_n)$.

Condition B

B.1. *For* $r[\cdot, \cdot]$ *given in* (4.1),

$$\sup_{\{\theta \in \Theta_n : \|\theta - \theta_0\| \le \delta_n\}} \mu_n(r[\theta - \theta_0, Y] - r[P_n(\theta^*(\theta, \varepsilon_n)) - \theta_0, Y]) = O_p(\varepsilon_n^2).$$

B.2 $$\sup_{\{\theta\in\Theta_n:0<\|\theta-\theta_0\|\le\delta_n\}} [K(\theta_0, P_n\theta^*(\theta,\varepsilon_n)) - K(\theta_0,\theta)] - \tfrac{1}{2}\left[\|\theta^*(\theta,\varepsilon_n)-\theta_0\|^2 - \|\theta-\theta_0\|^2\right] = O(\varepsilon_n^2).$$

B.3 $$\sup_{\{\theta\in\Theta_n:0<\|\theta-\theta_0\|\le\delta_n\}} \|\theta^*(\theta,\varepsilon_n) - P_n(\theta^*(\theta,\varepsilon_n))\| = O(\delta_n^{-1}\varepsilon_n^2);$$

in addition,

$$\sup_{\{\theta\in\Theta_n:\|\theta-\theta_0\|\le\delta_n\}} \mu_n\left(l'_{\theta_0}[\theta^*(\theta,\varepsilon_n) - P_n(\theta^*(\theta,\varepsilon_n)), Y]\right) = O_p(\varepsilon_n^2).$$

B.4 $$\sup_{\{\theta\in\Theta_n:\|\theta-\theta_0\|\le\delta_n\}} \mu_n\left(l'_{\theta_0}[\theta-\theta_0, Y]\right) = O_p(\varepsilon_n).$$

B.5 $$n^{1/2}\mu_n\left(l'_{\theta_0}[v^*, Y]\right) \xrightarrow{P_{\theta_0}} \mathcal{N}(0,\sigma_{v^*}^2),$$

$$\text{with}\quad \sigma_{v^*}^2 = \lim_{n\to\infty} n^{-1}\operatorname{var}_0\left(\sum_{t=1}^{n} l'_{\theta_0}[v^*, Y_t]\right) > 0.$$

THEOREM 2—Normality: *Suppose conditions* B.1–B.5 *hold, and f satisfies* (4.2) *with* $\|\hat\theta_n-\theta_0\|^\omega = o_p(n^{-1/2})$. *Then, for the sieve estimate* $\hat\theta_n$, $n^{1/2}(f(\hat\theta_n) - f(\theta_0)) \xrightarrow{P_{\theta_0}} \mathcal{N}(0,\sigma_{v^*}^2)$.

REMARK 2: (a) Conditions B.1–B.4 are the same as those in Shen (1997) used for obtaining the asymptotic normality for the i.i.d case. Condition B.1 specifies a linear approximation to the empirical criterion function within a small neighborhood of θ_0. Condition B.2 characterizes the local quadratic behavior of the expected value of the criterion difference. When Θ is infinite-dimensional, there may not exist any interior points with respect to $\|\cdot\|$ and $\hat\theta_n$ is often on the boundary of Θ_n. Thus the score function specified by the directional derivative evaluated at $\hat\theta_n$ may not be zero. Conditions B.3 and B.4 are generalizations of the usual assumption that θ_0 is an interior point of Θ, and $\hat\theta_n$ is a solution to the score equation in the finite-dimensional case.

(b) Condition B.5 only requires a classical finite-dimensional CLT. It is implied by many commonly used assumptions such as strong mixing condition with finite qth moments ($q>2$) and mixing coefficients $\alpha(i)$ satisfying $\sum_{i=1}^{\infty}[\alpha(i)]^{(q-2)/q} < \infty$. In particular, Condition A.1 implies Condition B.5 and

$$\sigma_{v^*}^2 = \operatorname{var}_0\left(l'_{\theta_0}[v^*, Y_1]\right) + 2\sum_{j=1}^{\infty}\operatorname{cov}_0\left(l'_{\theta_0}[v^*, Y_1], l'_{\theta_0}[v^*, Y_j]\right).$$

(c) To verify Conditions B.1, B.3(ii), and B.4, we may apply Theorem 3 in Appendix A to establish the convergence rates. The latter involves verifying Conditions A.1–A.4 for $g(\theta, Y_t)$ instead of $l(\theta, Y_t) - l(\theta_0, Y_t)$, and obtains a uniform large deviation probability bound for empirical process $\mu_n(g)$ indexed by a general class of functions $\{g\}$ with β-mixing data.

5. APPLICATIONS

EXAMPLE 1 CONTINUED: Tensor-product construction is a standard way to generate sieves for multi-dimensional spaces from sieves for one-dimensional spaces. For example, let $\{\phi_i\}_{i=0}^{\infty}$ be an orthonormal basis for $L_2[0,1]$; then the tensor products $\{\phi_{\vec{i}}(x) = \prod_{j=1}^{d} \phi_{i_j}(x_j): \vec{i} = (i_1, \ldots, i_d) \in \{0,1,\ldots\}^d\}$ become an orthonormal basis for $L_2[0,1]^d$. Here we present two kinds of tensor-product sieves.

Case 1.2 (Tensor-product Spline): Suppose $\theta_0 \in \Theta = W_2^m([b_1, b_2]^d)$, a Sobolev space which consists of functions with L_2-integrable derivatives up to integer order $m \geq 1$. Consider the following multivariate B-spline sieve:

$$\Theta_n = \left\{ \theta \in \Theta: \theta(u_1, \ldots, u_d) = \sum_{|\vec{i}|=1}^{r_n+m+1} a_{\vec{i}} \prod_{j=1}^{d} \phi_{i_j}(u_j), \right. \tag{5.1}$$

$$\left. \max_{|\vec{i}|} |a_{\vec{i}}| \leq c_n, \sum_{|\vec{i}|} a_{\vec{i}}^2 |\vec{i}|^{2m} \leq \tilde{c}_n \right\},$$

where $\vec{i} \equiv \{i_1, \ldots, i_d) \in \{1, \ldots, r_n + m + 1\}^d$, $|\vec{i}| \equiv \max_{1 \leq j \leq d} i_j$. $\{\phi_i\}$ are B-spline bases of order $m+1$ on $[b_1, b_2]$ with ϕ_i supported on $[x_i, x_{i+m+2}]$, and $(b_1 \equiv x_0 < x_1 < \cdots < x_{r_{n+(m+1)}} \equiv b_2)$ is the knot location satisfying restriction (2.9); $c_n, \tilde{c}_n$ go to ∞ arbitrarily slowly with n.

Case 1.3 (Tensor-product Wavelet): Suppose $\theta_0 \in \Theta = B_{p,\infty}^m(C)$ $(m > 0, 1 \leq p \leq \infty)$, a Besov ball which consists of functions in $L_p(\mathscr{R}^d)$ with $\|\theta\|_{B_{p,\infty}^m} < C$. There are several equivalent definitions of a Besov space; we follow that of Meyer (1990) in terms of wavelet coefficients. Suppose there exists a multivariate scaling function ϕ which has a compact support and continuous derivatives up to order $r \geq m$. Then there also exist $2^d - 1$ associated wavelets ψ^{ε}, $\varepsilon \in \{1, \ldots, 2^d - 1\}$, such that for a given j_0 integer, $\{\phi_{j_0,i}, \psi_{j,k}^{\varepsilon}: j \geq j_0, (i,k) \in \mathscr{Z}^{2d}, \varepsilon \in \{1, \ldots, 2^d - 1\}\}$ is an orthonormal basis of $L_2(\mathscr{R}^d)$, where $\psi_{j,i}(\cdot) = 2^{j/2}\psi(2^j \cdot - i)$ for $i \in \mathscr{Z}^d$. From Meyer (1990, Chapter 3, Sections 9 and 10), we have for any $\theta \in L_2(\mathscr{R}^d)$,

$$\theta = \sum_{i \in \mathscr{Z}^d} a_{j_0,i} \phi_{j_0,i} + \sum_{j \geq j_0} \sum_{i \in \mathscr{Z}^d} \sum_{1 \leq \varepsilon \leq 2^d - 1} c_{j,i}^{\varepsilon} \psi_{j,i}^{\varepsilon}.$$

Now for $m > 0$, $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, $\theta \in B_{p,q}^m(\mathscr{R}^d)$ if and only if (Meyer (1990, Ch. 6, Proposition 7, p. 200))

$$\|\theta\|_{B_{p,q}^m} \equiv \|a_{0,\cdot}\|_p + \left[\sum_{j \geq 0} \left(2^{j(m+d/2-d/p)} \|c_{j,\cdot}^{\cdot}\|_p \right)^q \right]^{1/q} < \infty$$

with the convention for $q = \infty$:

$$\|\theta\|_{B^m_{p,\infty}} \equiv \|a_{0,\cdot}\|_p + \sup_{j \geq 0} 2^{j(m+d/2-d/p)} \|c^{\cdot}_{j,\cdot}\|_p < \infty$$

where

$$\|a_{0,\cdot}\|_p \equiv \left[\sum_{i \in \mathcal{Z}^d} |a_{0,i}|^p \right]^{1/p}, \qquad \|c^{\cdot}_{j,\cdot}\|_p \equiv \left[\sum_{i \in \mathcal{Z}^d} \sum_{1 \leq \varepsilon \leq 2^d - 1} |c^{\varepsilon}_{j,i}|^p \right]^{1/p}.$$

Consider the wavelet sieve:

$$\Theta_n = \left\{ \theta \in L_2\left([b_1, b_2]^d\right): \right.$$

$$\left. \theta(x) = \sum_{j=0}^{J_n} \sum_{i \in \mathcal{I}_n} \sum_{1 \leq \varepsilon \leq 2^d - 1} c^{\varepsilon}_{j,i} \psi^{\varepsilon}_{j,i}(x), \|\theta\|_{B^m_{p,\infty}} < C \right\}, \tag{5.2}$$

where $\mathcal{I}_n \subset \mathcal{Z}^d$ is the set of localization parameters depending on the support of the wavelets as well as the dilation parameter j. Let $[b_1, b_2]^d$ be the support of ϕ. For each fixed $j, i \equiv (i_1, \ldots, i_d) \in \mathcal{I}_n$ provided that for any $k \in \{1, \ldots, d\}$, for any $t \in \{1, \ldots, n\}$, we need $b_1 \leq 2^j Z_{k,t} - i_k \leq b_2$.

PROPOSITION 1: *Suppose that* $\{Y_t\}$ *is generated according to* (2.2) *and the followings hold*:

(5.1.1) $\{X_t\}$ *and* $\{e_t\}$ *are independent.* $\{X_t\}$ *is i.i.d. with marginal distribution* F_X *on* $\mathcal{R}^k$. $\{e_t\}$ *is i.i.d. with marginal distribution* F_e *on* $\mathcal{R}$.

(5.1.2) F_e *is absolutely continuous with respect to the Lebesgue measure on* $\mathcal{R}$.

(5.1.3) *There exist constants* c, $z_0 > 0$, $a_1, \ldots, a_p \geq 0$, *and a locally bounded and measurable function* $h: \mathcal{R}^k \to [0, \infty)$ *such that* $|\theta_0(z)| \leq \sum_{i=1}^p a_i |y_i| + \sum_{j=1}^q h(x_j) - c$ *if* $|z| > z_0$; *and* $\sup_{z: |z| \leq z_0} |\theta_0(z)| < \infty$ *for* $z = (y_1, \ldots, y_p, x_1, \ldots, x_q) \in \mathcal{R}^p \times \mathcal{R}^{kq}$, *where*

$$|z| = \max(|y_1|, \ldots, |y_p|, |x_1|, \ldots, |x_q|).$$

(5.1.4) $Eh(X_1) + E|e_1|^{2+\zeta} < \infty$ *for some* $\zeta > 0$.

(5.1.5) *The polynomial* $P(u) = u^p - a_1 u^{p-1} - \cdots - a_p$ *has a unique real root* $\rho \in [0, 1)$.

(5.1.6) *The starting point* Y_0 *is drawn from the invariant distribution.*

Case 1.1 (*Neural networks*): *Suppose* $\theta_0 \in \Theta$ *given in* (2.3). *Let* $\hat{\theta}_n$ *be the sigmoid neural network sieve* (2.4) *estimate, with* $|\psi(v_1) - \psi(v_2)| \leq C |v_1 - v_2|$ *for any* $v_1, v_2 \in \mathcal{R}$. *Let* $c_n = const.$, $\tilde{c}_n = const.$, *and* $r_n^{2+1/d} \log r_n = O(n)$. *Then* $\|\hat{\theta}_n - \theta_0\| = O_P([n/\log n]^{-(1+1/d)/[4(1+1/(2d))]})$.

Case 1.2 (*Spline*): *Suppose* $\theta_0 \in \Theta = W_2^m([b_1, b_2]^d)$. *Let* $\hat{\theta}_n$ *be the spline sieve* (5.1) *estimate. Let* $r_n = O(n^{1/(2m+d)})$ *and* $c_n, \tilde{c}_n = O(\log(n))$. *Then* $\|\hat{\theta}_n - \theta_0\| = O_P(n^{-m/(2m+d)})$.

Case 1.3 (Wavelet): *Suppose* $\theta_0 \in \Theta = B^m_{p,\infty}(C)$. *Let* $\hat{\theta}_n$ *be the wavelet sieve* (5.2) *estimate.*

(i) *For* $p \geq 2$, $m > 0$, *let* $2^{J_n} = O(n^{1/(2m+d)})$. *Then* $\|\hat{\theta}_n - \theta_0\| = O_P(n^{-m/(2m+d)})$.

(ii) *For* $1 \leq p < 2$ *and* $m > d(1/p - 1/2)$, *let* $2^J = O(n^{1/[d+2(m-d(1/p-1/2))]})$. *Then* $\|\hat{\theta}_n - \theta_0\| = O_P(n^{-(m-d(1/p-1/2))/[d+2(m-d(1/p-1/2))]})$.

REMARK 3: (a) The Besov space $B^m_{p,q}$ includes many classical spaces as special cases: for example, $B^m_{\infty,\infty} \equiv \Lambda^m$, the Hölder space, and $B^m_{2,2} \equiv W^m_2$ the Hilbert-Sobolev space with integer m. For fixed m, p, the Besov space $B^m_{p,q}$ gets larger with increasing q; for fixed m, q, the space $B^m_{p,q}$ gets larger with decreasing p. $B^m_{2,2} \subset B^m_{2,\infty} \subset B^m_{p,\infty}$ for $p < 2$. Note here that $\hat{\theta}_n$ based on the wavelet sieve (5.2) cannot achieve Stone's (1982) optimal convergence rate $O_p(n^{-m/(2m+d)})$ when the underlying space is too large ($\Theta = B^m_{p,\infty}$ with $1 \leq p < 2$). In fact, it is known that no linear estimates can attain the optimal convergence rate (in L_2 norm) for $\theta_0 \in B^m_{p,q}$ with $1 \leq p < 2$, $1 \leq q \leq \infty$, a space consisting of inhomogeneous functions. Nonlinear procedures such as wavelet shrinkage estimates (see, e.g., Donoho, Johnstone, Kerkyacharian, and Picard (1995)) are needed to achieve faster rates.

(b) There is no universal "best" sieve in terms of convergence rate, since the rate depends on the parameter space to which θ_0 belongs. For instance, for some large d and small m, the nonlinear sigmoid neural network sieve (2.4) estimate has a faster rate $O_p([n/\log n]^{-(1+1/d)/[4(1+1/(2d))]})$ as compared to the rate $O_p(n^{-m/(2m+d)})$ achievable by many linear estimates based on kernels, orthogonal series, splines, and wavelets. However, this is because $\theta_0 \in \Theta$ as given in (2.3) for the neural networks whereas $\theta_0 \in W^m_2(\mathscr{X})$ ($\mathscr{X}$ a compact subset in $\mathscr{R}^d$) for the other cases; and the smoothness conditions imposed on θ_0 by these two spaces are not directly comparable. On the one hand, Barron (1993) points out that Θ in (2.3) includes functions which are linear combinations of Gaussian densities with different means and different variances, that is, the space of Gaussian radial basis functions (or "hump algebra" according to Meyer (1990, Chapter 6)) $B^1_{1,1}$ is a subset of Θ in (2.3). From (a), we know that these linear estimates yield slower rates (in L_2 norm) than nonlinear estimates when $\theta_0 \in B^1_{1,1}$. On the other hand, when $\theta_0 \in B^m_{2,2} \equiv W^m_2$, these linear estimates achieve Stone's (1982) optimal rate $O_p(n^{-m/(2m+d)})$ while neural network sieves may not. For example, for i.i.d. observations, when $\theta_0 \in W^m_2(\mathscr{X})$ ($m \geq d$), McCaffrey and Gallant (1994) obtain the rate $O_p(n^{-m/(2m+d+5)})$ for a cosine neural network sieve, slower than the rate $O_p(n^{-m/(2m+d)})$ ($m \geq d+1$) attained by a standard linear Fourier series sieve.

PROPOSITION 2: *Suppose* $\{Y_t\}$ *is generated according to* (2.5), *and is stationary* β*-mixing satisfying Condition* A.1. *Suppose* $\{e_t\}$ *has common marginal density,* $E[\sigma^4(Y_{t-i}, 1 \leq i \leq p)] < \infty$ *and* $E[|e_t|^{2+\zeta}] < \infty$ *for some* $\zeta > 2$. *Let* $r_{n,i} =$

$O(n^{1/(2m_i+1)})$ *for* $i = 1, \ldots, p - p_1$. *Then*:

(i) $\|\hat{\eta}_n - \eta_0\| = O_P(n^{-m/(2m+1)})$ *with* $m \equiv \min_{1 \le i \le p-p_i} m_i$, *in either of the two cases*:

Case 2.1 (*Fourier series without monotone constraint*): $\theta_0 \in \Theta$ *given in* (2.6), $\hat{\theta}_n$ *the Fourier series sieve* (2.7) *estimate*, $m > (1 + \sqrt{5})/4$ *and* $c_{n,i} = O(\log(n))$.

Case 2.2 (*Spline with monotone constraint*): $\theta_0 \in \Theta$ *given in* (2.8), $\hat{\theta}_n$ *the spline sieve* (2.10) *estimate*, $m \ge 1$ *and* $c_{n,i} = O(\log(n))$.

(ii) *Let* $W_t^T \equiv (Y_{t-1}, \ldots, Y_{t-p_1}) - E[(Y_{t-1}, \ldots, Y_{t-p_1}) \mid Y_{t-p_1-1}, \ldots, Y_{t-p}]$. *If* $E[W_t W_t^T]$ *is positive definite, and* $E[|W_t|^{2+\zeta}] < \infty$ *for some* $\zeta > 0$, *then* $n^{1/2}(\hat{\alpha}_n - \alpha_0) \xrightarrow{P\theta_0} \mathcal{N}(0, \Omega_*)$, *with*

$$\Omega_* = \Sigma^{-1}\left(E\left[e_1^2 W_1 W_1^T\right] + \sum_{j=2}^{\infty} E\left[e_1 e_j W_1 W_j^T + e_j e_1 W_j W_1^T\right]\right)\Sigma^{-1},$$

$$\Sigma = E\left[W_1 W_1^T\right].$$

(iii) *If in addition*, $E(e_t \mid Y_i, i \le t-1) = 0$ *and* $E(e_t^2 \mid Y_i, i \le t-1) = \sigma^2$, *then* $\Omega_* = \sigma^2 \Sigma^{-1}$.

REMARK 4: (a) By applying results of Doukhan (1994, Sec. 2.4) or the results of Meyn and Tweedie (1992), one can again impose more primitive conditions to ensure that $\{Y_t\}$ is β-mixing.

(b) There exist many other shape-preserving sieves besides the Cardinal B-splines; see DeVore and Lorentz (1993) for more examples.

EXAMPLE 3 (Monotone Transformation Model): Define $\phi(x, e) \equiv \eta^{-1}(\alpha x + e)$. Define $\phi_e(x) \equiv \phi(x, e)$, and $\phi_e^k \equiv \phi_e \circ \cdots \circ \phi_e$ (k times).

PROPOSITION 3: *Suppose* $\theta_0 = (\alpha_0, \eta_0(\cdot)) \in \Theta = A \times D$, *where* A *is an open subset of* $\mathcal{R}$ *and* $D = \{\eta(\cdot): c_2 \ge \eta'(\cdot) \ge c_1\}$, *with* $c_i > 0$ *constants. In addition suppose the following hold*:

(5.3.1) *There exists* $e \in \mathcal{R}$ *such that* ϕ_e^k *is Lipschitzian with order strictly less than* 1.

(5.3.2) *The sequence* $\phi_{e_n} \circ \cdots \circ \phi_{e_{n-k}}(x)$ *converges* (*as* $k \to \infty$) *in mean of order* s *to a limit independent of* x *for some real number* $s > 0$.

(5.3.3) $P(\forall n \ge 0, Y_n \in [b_1, b_2]) = 1$. Y_0 *has marginal distribution* ν_0. $\{e_t\}$ *is i.i.d. with continuous density*, $Ee_t = 0$, $Ee_t^2 = 1$.

(5.3.4) *There exists a nonnegative measurable function* G (*the Lyapounov function*) *on* $[b_1, b_2]$, *a compact subset* K *and positive constants* $c_1, c_2 > 0$ *and* $0 < \rho < 1$ *such that* (i) $\nu_0(K \cap [b_1, b_2]) > 0$, (ii) *for any* $x \in K \cap [b_1, b_2]$, $E[G(Y_n) \mid Y_{n-1} = x] \le \rho G(x) - c_1$; (iii) *for any* $x \in K \cap [b_1, b_2]$, $E[G(Y_n) \mid Y_{n-1} = x] \le c_2$.

Then: (i) $\|\hat{\eta}_n - \eta_0\| = O_P(n^{-1/3})$; (ii) $n^{1/2}(\hat{\alpha}_n - \alpha_0) \xrightarrow{P\theta_0} \mathcal{N}(0, \sigma_*^2)$ *with*

$$\sigma_*^2 = \mathrm{var}_0\big(l'_{\theta_0}[v^*, Y_1]\big) + 2\sum_{j=2}^{\infty} \mathrm{cov}_0\big(l'_{\theta_0}[v^*, Y_1], l'_{\theta_0}[v^*, Y_j]\big),$$

where v^* *is given by* (b.7) *in Appendix B.*

6. SUMMARY

In this paper, we provide a general theory on convergence rate of sieve extremum estimates and root-n asymptotic normality of plug-in sieve estimates for stationary β-mixing observations. Although the main theorems are stated under sets of "high-level" sufficient conditions, they are widely applicable, and especially useful in non/semi-parametric time series models with economic constraints. As illustration, we present several kinds of sieves in three regression models, and the rates and asymptotic normality are derived under relatively primitive sufficient conditions.

The research may be extended to time series observations which allow for greater temporal dependence and heterogeneity, such as *near epoch dependent* (NED) functions of mixing processes. However, in order to obtain some reasonable convergence rates for these processes, we need to modify the L_2 bracketing metric entropy concept to allow for heterogeneous observations, and to develop a Bernstein-type inequality for NED similar to Lemma 1 for β-mixing.

Dept. of Economics, University of Chicago, 1126 East 59th St., Chicago, IL 60637, U.S.A.;

and

Dept. of Statistics, The Ohio State University, 1958 Neil Ave., Columbus, OH 43210, U.S.A.

Manuscript received June, 1996; final revision received May, 1997.

APPENDIX A: PROOFS OF THEOREMS

LEMMA 1 (β-mixing): *Let* $\{Y_t\}$ *be a stationary* β*-mixing sequence with* $a_{n2}\,\beta(a_{n1}) \to 0$ *for any integer pair* (a_{n1}, a_{n2}) *with* $a_{n2} = [n/(2a_{n1})] \to \infty$, $a_{n1} \to \infty$. *Suppose* $\sup_{f\in\mathcal{F}} f(Y_t) \le T$, $\sup_{f\in\mathcal{F}} n^{-1}\,\mathrm{var}(\sum_{t=1}^{n} f(Y_t)) \le \sigma^2$. *Suppose for any* $0 < \xi < 1$,

(a.1) $$M \le \xi\sigma^2/4,$$

(a.2) $$\int_{\xi M/32}^{\sigma T^{1/2}} H^{1/2}(w, \mathcal{F})\,dw \le Mn^{1/2}\xi^{3/2}/2^{10},$$

and

(a.3) $$(M/3) > (2T/a_{n2}).$$

Then for any $M > 0$ *satisfying* (a.1)–(a.3),

$$P^*\left(\sup_{f\in\mathscr{F}} n^{-1} \sum_{t=1}^{n} (f(Y_t) - Ef(Y_t)) \geq M \right) \leq 6\exp\left(-(1-\xi)\frac{nM^2}{9\sigma^2(1 + a_{n1}T\xi/12)} \right) + 2(a_{n2} - 1)\beta(a_{n1}).$$

PROOF OF LEMMA 1: Consider the following independent blocks (IB) construction: For any integer pair (a_{n1}, a_{n2}), with $a_{n2} = [n/(2a_{n1})]$, divide $(Y_1, \ldots, Y_n)$ into $2a_{n2}$ blocks with length a_{n1}, and the remaining block of length $n - 2a_{n2}a_{n1}$. Denote

$$H_{1,j} = \{i: 2(j-1)a_{n1} + 1 \leq i \leq (2j-1)a_{n1}\} \quad \text{and}$$

$$H_{2,j} = \{i: (2j-1)a_{n1} + 1 \leq i \leq 2ja_{n1}\},$$

for $j = 1, \ldots, a_{n2}$ and $R = \{i: 2a_{n2}a_{n1} + 1 \leq i \leq n\}$. We now construct a random sequence $\{X_t: t = 1, \ldots, n\}$, which is independent of $\{Y_t: t = 1, \ldots, n\}$ and has independent blocks such that each block has the same (joint) distribution as the corresponding block of the original Y-sequence:

$$\mathscr{L}(X_1, \ldots, X_n) = \mathscr{L}(Y_i: i \in H_{1,1}) \times \mathscr{L}(Y_i: i \in H_{2,1}) \times \mathscr{L}(Y_i: i \in H_{1,2}) \times \cdots.$$

(Note that $\{X_t: t = 1, \ldots, n\}$ is not an i.i.d. sequence, since elements within a single block $\{X_i: i \in H_{1,j}\}$ could be correlated.) By Eberlein (1984), we have the following: for any measurable set A,

$$\begin{aligned} (a.4) \qquad & |P[(X_1, \ldots, X_{a_{n1}}, X_{2a_{n1}+1}, \ldots, X_{3a_{n1}}, \ldots, X_{2(a_{n2}-1)a_{n1}+1}, \ldots, X_{2a_{n2}a_{n1}}) \in A] \\ & - P[(Y_1, \ldots, Y_{a_{n1}}, Y_{2a_{n1}+1}, \ldots, Y_{3a_{n1}}, \ldots, Y_{2(a_{n2}-1)a_{n1}+1}, \ldots, Y_{2a_{n2}a_{n1}}) \in A]| \\ & \leq (a_{n2} - 1)\beta(a_{n1}). \end{aligned}$$

Define

$$V_{1j,f} = \sum_{t\in H_{1,j}} (f(Y_t) - Ef(Y_t)) = \sum_{t=2(j-1)a_{n1}+1}^{(2j-1)a_{n1}} (f(Y_t) - Ef(Y_t))$$

and $V_{2j,f} = \sum_{t\in H_{2,j}} (f(Y_t) - Ef(Y_t))$. Define $U_{1j,f} = \sum_{t\in H_{1,j}} (f(X_t) - Ef(X_t))$ and $U_{2j,f} = \sum_{t\in H_{2,j}} (f(X_t) - Ef(X_t))$. Then

$$n^{-1} \sum_{t=1}^{n} (f(Y_t) - Ef(Y_t)) = n^{-1} \sum_{j=1}^{a_{n2}} V_{1j,f} + n^{-1} \sum_{j=1}^{a_{n2}} V_{2j,f} + n^{-1} \sum_{t\in R} (f(Y_t) - Ef(Y_t)).$$

By stationarity, we have for any $M > 0$,

$$(a.5) \qquad P^*\left(\sup_{f\in\mathscr{F}} n^{-1} \sum_{t=1}^{n} (f(Y_t) - Ef(Y_t)) \geq M \right) \leq 2 \times P_5 + P_6,$$

with

$$P_5 \equiv P^*\left(\sup_{f\in\mathscr{F}} n^{-1} \sum_{j=1}^{a_{n2}} V_{1j,f} \geq M/3 \right),$$

$$P_6 \equiv P^*\left(\sup_{f\in\mathscr{F}} n^{-1} \sum_{t\in R} (f(Y_t) - Ef(Y_t)) \geq M/3 \right).$$

By (a.3) and the definition of a_{n1}, a_{n2}, we have

$$\sup_{f\in\mathscr{F}} n^{-1}\sum_{t\in R}|f(Y_t)-Ef(Y_t)| \le 2T(n-2a_{n2}a_{n1})/n \le 2T/a_{n2} < M/3.$$

Thus $P_6 \equiv 0$.

In (a.4), let $A = \{\sup_{f\in\mathscr{F}} n^{-1}\sum_{j=1}^{q_{n2}} V_{1j,f} \ge M/2\}$; then we have

$$P_5 \le P^*\left(\sup_{f\in\mathscr{F}} n^{-1}\sum_{j=1}^{a_{n2}} U_{1j,f} \ge (M/3)\right) + (a_{n2}-1)\beta(a_{n1}). \tag{a.6}$$

Note that $\{U_{1j,f}: j=1,\dots,a_{n2}\}$ are i.i.d. with mean zero, $\operatorname{var}(U_{1j,f}) \le a_{n1}\sigma^2$, and $\sup_{f\in\mathscr{F}} U_{1j,f} \le a_{n1}T$. By Lemma 1 of Chen and Shen (1996) or a reformulation of Theorem 3 of Shen and Wong (1994),

$$\begin{aligned} P^*\left(\sup_{f\in\mathscr{F}} n^{-1}\sum_{j=1}^{a_{n2}} U_{1j,f} \ge (M/3)\right) &= P^*\left(\sup_{f\in\mathscr{F}} (a_{n2})^{-1}\sum_{j=1}^{a_{n2}} U_{1j,f} \ge (n/a_{n2})(M/3)\right) \\ &\le 3\exp\left(-(1-\xi)\frac{nM^2}{9\sigma^2(1+a_{n1}T\xi/12)}\right). \end{aligned} \tag{a.7}$$

The result then follows from (a.5)–(a.7). *Q.E.D.*

THEOREM 3: *Suppose Conditions* A.1–A.4 *hold, together with either* (i) $2 < \gamma/2 \le \zeta \le \gamma - 2$; *or* (ii) $\zeta > \max\{\gamma/2, \gamma-2, \gamma(6-2s-\gamma(2-s))/((2+s)\gamma-4)\}$. *Then there exist constants* $c', c, C, C_4 > 0$ *such that for any* $x \ge 1$ *and integer* n,

$$\begin{aligned} &P^*\left(\sup_{\{d(\theta_0,\theta)\ge x\varepsilon_n,\,\theta\in\Theta_n\}} \mu_n(l(\theta)-l(\theta_0)) \ge (x\varepsilon_n)^2/2\right) \\ &\quad\le 7\exp(-Cnx^{2-s}\varepsilon_n^2) + C_4 n^{-\gamma/2}x^{-(2-s)\gamma}\varepsilon_n^{-(2-s)\gamma}, \end{aligned}$$

where $\varepsilon_n = \max(\delta_n, d(\theta_0, \pi_n\theta_0))$.

PROOF OF THEOREM 3:

$$\begin{aligned} &P^*\left(\sup_{\{d(\theta_0,\theta)\ge x\varepsilon_n,\,\theta\in\Theta_n\}} \mu_n(l(\theta)-l(\theta_0)) \ge 0.5(x\varepsilon_n)^2\right) \\ &\quad\le P^*\left(\sup_{\{d(\theta_0,\theta)\ge x\varepsilon_n,\,\theta\in\Theta_n\}} \mu_n(l(\theta)-l(\theta_0))I(U(Y_t)\le B_n) \ge (x\varepsilon_n)^2/4\right) \\ &\qquad + P^*\left(\sup_{\{d(\theta_0,\theta)\ge x\varepsilon_n,\,\theta\in\Theta_n\}} \mu_n(l(\theta)-l(\theta_0))I(U(Y_t)\ge B_n) \ge (x\varepsilon_n)^2/4\right) \\ &\quad\le \sum_{k=0}^{\infty} P^*\left(\sup_{A_{n,k}} \mu_n(l(\theta)-l(\theta_0))I(U(Y_t)\le B_n) \ge (2^{k-2}x\varepsilon_n)^2\right) \\ &\qquad + \sum_{k=0}^{\infty} P^*\left(\sup_{A_{n,k}} \mu_n(l(\theta)-l(\theta_0))I(U(Y_t)\ge B_n) \ge (2^{k-2}x\varepsilon_n)^2\right) \\ &\quad\equiv P_3 + P_4, \end{aligned}$$

where the truncation sequence $B_n \to \infty$ such that $a_{n1}B_n\varepsilon_n^s \ge (2C_3)^{1/(\gamma-1)}$. For each $k = 0, 1, 2, \dots$, let

$$A_{n,k} \equiv \{\theta\in\Theta_n: 2^{k-1}x\varepsilon_n \le d(\theta_0,\theta) \le 2^k x\varepsilon_n\}.$$

We now apply Lemma 1 to bound P_3: Let $M_k = (2^{k-1}x\varepsilon_n)^2$, $T_k = \min((2^k x\varepsilon_n)^s B_n, 8/C_1)$, and $\sigma_k^2 = C_1(2^k x\varepsilon_n)^2 T_k$. Then by A.2, we have $\sup_{A_{n,k}} \operatorname{var}(l(\theta, Y_t) - l(\theta_0, Y_t)) \le \sigma_k^2$; by A.4 we have $\sup_{A_{n,k}} |l(\theta, Y_t) - l(\theta_0, Y_t)| \le T_k$. (a.1) is satisfied by $M_k \le \xi\sigma_k^2/4$ for $\xi = 1/2$. (a.2) is also satisfied by A.3:

$$M_k^{-1} \int_{M_k/64}^{\sigma_k T_k^{1/2}} H^{1/2}(w, A_{n,k})dw \le cn^{1/2}.$$

If we choose $a_{n2k} = [n/2a_{n1k}] = [7T_k/M_k]$, then (a.3) is satisfied. Given the value of T_k and M_k, such a choice is valid, since $a_{n2k} \to \infty$ and $a_{n1k} \to \infty$ as $n \to \infty$. Applying Lemma 1, we get $P_3 \le P_{3,I} + P_{3,II}$ with

$$P_{3,I} \le \sum_{k=0}^{\infty} 6\exp\left(-\frac{nM_k^2}{18\sigma_k^2(1 + a_{n1k}T_k/24)}\right) \le \sum_{k=0}^{\infty} 6\exp\left(-\frac{n(2^{k-1}x\varepsilon_n)^2}{(2^k x)^s(\varepsilon_n^s B_n a_{n1k})}\right).$$

Since $\varepsilon_n^s B_n a_{n1k} \ge (2C_3)^{1/(\gamma-1)}$,

$$P_{3,I} \le 6\sum_{k=0}^{\infty} \exp\left(-C'n(2^k x)^{2-s}\varepsilon_n^2\right) \le 7\exp(-Cnx^{2-s}\varepsilon_n^2).$$

By A.1, $\beta(j) \le \beta_0 j^{-\zeta}$ for some $\beta_0 > 0$ and $\zeta > 1$; thus we have

$$\begin{aligned} P_{3,II} &\le 2\sum_{k=0}^{\infty} a_{n2k}\beta(a_{n1k}) \le 2\beta_0 \sum_{k=0}^{\infty} a_{n2k}(a_{n1k})^{-\zeta} \\ &\le \beta_0 14^{1+\zeta} n^{-\zeta} \sum_{k=0}^{\infty} (T_k/M_k)^{1+\zeta} \le Cn^{-\zeta} \sum_{k=0}^{\infty} \left((2^k x\varepsilon_n)^{s-2} B_n\right)^{1+\zeta} \\ &\le C' x^{(s-2)(1+\zeta)} \varepsilon_n^{(s-2)(1+\zeta)} B_n^{1+\zeta} n^{-\zeta} \le C_4 n^{-\gamma/2} x^{-(2-s)\gamma} (\varepsilon_n)^{-(2-s)\gamma} \end{aligned}$$

where the last inequality follows from the relation between ζ, γ, and s in the assumption.

To bound P_4, note that by A.1, A.4 with $0 < s < 2$, and $x > 1$, $\varepsilon_n^{(2-s)/(\gamma-1)} B_n \ge 1$. Then we have

$$E[U(Y_t)1(U(Y_t) \ge B_n)] \le c_3^{\gamma} B_n^{1-\gamma} \le 0.5(2^{k-1}x\varepsilon_n)^{2-s}.$$

By Condition A.4, $\sup_{A_{n,k}} |l(\theta, Y_t) - l(\theta_0, Y_t)| \le (2^k x\varepsilon_n)^s U(Y_t)$, where $0 < s < 2$. Hence

$$\begin{aligned} P_4 &\le \sum_{k=0}^{\infty} P\left(2n^{-1}\sum_{t=1}^{n} (2^k x\varepsilon_n)^s U(Y_t)1(U(Y_t) \ge B_n) \ge (2^{k-1}x\epsilon_n)^2\right) \\ &= \sum_{k=0}^{\infty} P\left(\sum_{t=1}^{n} U(Y_t)1(U(Y_t) \ge B_n) \ge 2^{-(s+1)} n(2^{k-1}x\varepsilon_n)^{2-s}\right) \\ &\le \sum_{k=0}^{\infty} P\left(\sum_{t=1}^{n} [U(Y_t)I(U(Y_t) \ge B_n) - E(U(Y_t)I(U(Y_t) \ge B_n))] \ge 2^{-1} n(2^{k-1}x\varepsilon_n)^{2-s}\right) \\ &\le \sum_{k=0}^{\infty} E\left[\left|\sum_{t=1}^{n} [U(Y_t)I(U(Y_t) \ge B_n) - E(U(Y_t)I(U(Y_t) \ge B_n))]\right|^{\gamma}\right] \\ &\quad \times 2^{\gamma}\left[n(2^{k-1}x\varepsilon_n)^{2-s}\right]^{-\gamma}. \end{aligned}$$

Given Conditions A.1 and A.4 with $\gamma > 2$, $\{U(Y_t)1(U(Y_t) \geq B_n) - E(U(Y_t)1(U(Y_t) \geq B_n))\}$ is a β-mixing sequence satisfying conditions in the moment inequality of Yokoyama (1980). Thus,

$$P_4 \leq \sum_{k=0}^{\infty} Cn^{\gamma/2} 2^{\gamma} \left[n(2^{k-1} x \varepsilon_n)^{2-s} \right]^{-\gamma} \leq C_4 n^{-\gamma/2} x^{-(2-s)\gamma} (\varepsilon_n)^{-(2-s)\gamma}.$$

This completes the proof. *Q.E.D.*

PROOF OF THEOREM 1: By definitions of $\hat{\theta}_n$ and the fact that $\pi_n \theta_0 \in \Theta_n$, we have

$$\begin{aligned}
&P(d(\theta_0, \hat{\theta}_n) \geq x\varepsilon_n) \\
&\quad \leq P^* \left(\sup_{\{d(\theta_0, \theta) \geq x\varepsilon_n, \theta \in \Theta_n\}} L_n(\theta) - L_n(\theta_0) \geq L_n(\hat{\theta}_n) - L_n(\theta_0) \right) \\
&\quad \leq P^* \left(\sup_{\{d(\theta_0, \theta) \geq x\varepsilon_n, \theta \in \Theta_n\}} L_n(\theta) - L_n(\theta_0) \geq L_n(\pi_n \theta_0) - L_n(\theta_0) - O(\varepsilon_n^2) \right) \\
&\quad \leq P_1 + P_2,
\end{aligned}$$

where

$$\begin{aligned}
P_2 &\equiv P(L_n(\theta_0) - L_n(\pi_n \theta_0) \geq (x\varepsilon_n)^2/2 - O(\varepsilon_n^2)), \\
P_1 &\equiv P^* \left(\sup_{\{d(\theta_0, \theta) \geq x\varepsilon_n, \theta \in \Theta_n\}} (L_n(\theta) - L_n(\theta_0)) \geq -(x\varepsilon_n)^2/2 \right) \\
&= P^* \left(\sup_{\{d(\theta_0, \theta) \geq x\varepsilon_n, \theta \in \Theta_n\}} \mu_n(l(\theta) - l(\theta_0)) \right. \\
&\qquad \left. \geq \inf_{\{d(\theta_0, \theta) \geq x\varepsilon_n, \theta \in \Theta_n\}} K(\theta_0, \theta) - 0.5(x\varepsilon_n)^2 \right) \\
&\leq P^* \left(\sup_{\{d(\theta_0, \theta) \geq x\varepsilon_n, \theta \in \Theta_n\}} \mu_n(l(\theta) - l(\theta)) \geq 0.5(x\varepsilon_n)^2 \right).
\end{aligned}$$

By Theorem 3, we have

$$P_1 \leq 4c_\kappa \exp(-Cn\varepsilon_n^2 x^{2-s}) + C_4 n^{-\gamma/2} \varepsilon_n^{-(2-s)\gamma} x^{-(2-s)\gamma}.$$

Now we bound P_2: Since $E[L_n(\theta_0) - L_n(\pi_n \theta_0)] = K(\theta_0, \pi_n \theta_0) \leq \varepsilon_n^2$ and $x \geq 1$,

$$\begin{aligned}
P_2 &\leq P([L_n(\theta_0) - L_n(\pi_n \theta_0)] - E[L_n(\theta_0) - L_n(\pi_n \theta_0)] \geq (x\varepsilon_n)^2/4) \\
&\leq E \left[\left| \sum_{t=1}^{n} ([l(\theta_0, Y_t) - l(\pi_n \theta_0, Y_t)] \right. \right. \\
&\qquad \left. \left. - E[l(\theta_0, Y_t) - l(\pi_n \theta_0, Y_t)]) \right|^{\gamma} \right] [n(x\varepsilon_n/2)^2]^{-\gamma}.
\end{aligned}$$

Denote $W_t = l(\theta_0, Y_t) - l(\pi_n \theta_0, Y_t) - E[l(\theta_0, Y_t) - l(\pi_n \theta_0, Y_t)]$. Conditions A.1 and A.4 imply that $\{W_t\}$ is stationary β-mixing, $EW_t = 0$, $E|W_t|^{\gamma} \leq C_3 2^{\gamma} (d(\theta_0, \pi_n \theta_0))^{\gamma s} \leq C' \varepsilon_n^{\gamma s}$. Now by the moment inequality of Yokoyama (1980) for β-mixing, we have

$$P_2 \leq C_5 n^{-\gamma/2} x^{-2\gamma} \varepsilon_n^{-(2-s)\gamma} \leq C_5 n^{-\gamma/2} x^{-(2-s)\gamma} \varepsilon_n^{-(2-s)\gamma},$$

where the second inequality holds because $x \geq 1$, $0 < s < 2$. This completes the proof. *Q.E.D.*

PROOF OF THEOREM 2: The proof is almost the same as that for CLT with the i.i.d. case in Shen (1997), except that we use Condition B.5 for the dependent case. *Q.E.D.*

APPENDIX B: PROOFS OF PROPOSITIONS

LEMMA 2: *Denote* $m = s + \gamma$ *with* $s = 0, 1, \ldots$ *integer, and* $0 < \gamma \leq 1$. *Define the Hölder space*

$$\Lambda^m = \left\{ f: [a,b]^d \to \mathscr{R}, \|f\|_H = \sup_{x, y \in [a,b]^d, x \neq y} \frac{|f^{(s)}(x) - f^{(s)}(y)|}{|x - y|^\gamma} \leq L \right\},$$

where $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in [a,b]^d$, *and* $\|\cdot\|_H$ *is the Hölder norm. Then,* $\|f\|_{\sup} \leq 2\|f\|_2^c L^{1-c}$, *where*

$$c = \frac{2(s+\gamma)}{2(s+\gamma)+d}.$$

PROOF: The proof is based on a modification of the result of Gabushin (1967). For simplicity, we only prove the case in which $f(a) = f(b) = 0$ and $s = 0$. For any $\delta > 0$ and

$$x \in [a,b]^d \cap \prod_{i=1}^d \left(x_i - \frac{\delta}{2}, x_i + \frac{\delta}{2} \right),$$

there exists

$$x^* \in [a,b]^d \cap \prod_{i=1}^d \left(x_i - \frac{\delta}{2}, x_i + \frac{\delta}{2} \right)$$

such that

$$|f(x^*)| = \min_{x \in [a,b]^d \cap \prod_{i=1}^d (x_i - (\delta/2), x_i + (\delta/2))} |f(x)|.$$

Then, it can be seen that

$$\begin{aligned} |f(x)| &\leq |f(x^*)| + \delta^{s+\gamma} \|f\|_H \\ &\leq \delta^{-d/2} \|f\|_2 + \delta^{(s+\gamma)d} L. \end{aligned}$$

By choosing $\delta = (\|f\|_2 / L)^{1/(s+\gamma+(d/2))}$, we obtain the desired result. *Q.E.D.*

PROOF OF PROPOSITION 1: By Theorem 7 in Doukhan (1994, Section 2.4.2.1, p. 102), Assumptions (5.1.1)–(5.1.6) imply that $\{Y_t\}$ is stationary β-mixing with exponential decay rate, thus satisfying Condition A.1. For A.2 and A.4, note that $l(\theta, Y_t, Z_t) - l(\theta_0, Y_t, Z_t) = (\theta - \theta_0)[e_t + (\theta_0 - \theta)/2]$. Assumption (5.1.4) implies that $Ee_t^2 \equiv \sigma^2 < \infty$; thus

$$E[l(\theta, Y_t, Z_t) - l(\theta_0, Y_t, Z_t)]^2 = 2\sigma^2 E([\theta_0(Z_t) - \theta(Z_t)]^2) + 0.5E([\theta_0(Z_t) - \theta(Z_t)]^4).$$

Note that

$$E([\theta_0(Z_t) - \theta(Z_t)]^4) \leq E(\sup[\theta(Z_t) - \theta_0(Z_t)]^2) E([\theta_0(Z_t) - \theta(Z_t)]^2),$$

and for Case 1.1, $\theta \in \Theta_n$ in (2.4), $\|\theta\|_{\sup} \leq c$; for Case 1.2, $\theta \in \Theta_n$ in (5.1), $\|\theta\|_{\sup} \leq c_n$; for Case 1.3, $\theta \in \Theta_n$ in (5.2), $\|\theta\|_{\sup} \leq c$. Note that when $c_n \to \infty$ arbitrarily slowly, by Remark 1(a), the result holds as if $c_n = \textit{constant}$. Hence Condition A.2 is satisfied for all small $\varepsilon \leq 1$. For Condition A.4,

$$|l(\theta, Y_t, Z_t) - l(\theta_0, Y_t, Z_t)| \leq \|\theta - \theta_0\|_{\sup} [|e_t| + (\|\theta_0\|_{\sup} + \|\theta\|_{\sup})/2].$$

Condition A.4 is then satisfied with $s = 2m/(2m+d)$ and $U(Y_t) = |e_t| + c_n$ with $m = 1$ for Case 1.1. Here we have used the relationship between the sup-norm and the L_2 norm as stated in Lemma 2.

Case 1.1 (Sigmoid neural networks): By Theorem 3 of Makovoz (1996), for any $\theta \in \Theta$, there exists $\pi_n\theta \in \Theta_n$ such that

$$\|\pi_n\theta - \theta\| \le \text{const.}(r_n)^{-1/2-1/(2d)}. \tag{b.1}$$

After some calculations, we have $H(w, \mathscr{F}_n) \le r_n c_n(d+1)\log(r_n c_n(d+1)/w)$. Note that $c_n = \text{const.}$ can be chosen to satisfy Condition A.3; it suffices to choose δ_n such that

$$\delta_n^{-1}(r_n \log r_n)^{1/2} \le \text{const.} n^{1/2}. \tag{b.2}$$

When $\delta_n = \|\pi_n\theta_0 - \theta_0\|$ in (b.1) and (b.2), Theorem 1 yields $r_n^{2+1/d} \log r_n = O(n)$, and $\|\hat{\theta}_n - \theta_0\| = O_p([n/\log n]^{-(1+1/d)/[4(1+1/(2d))]})$.

Case 1.2 (Spine): By Theorems 12.8 and 13.24 of Schumaker (1981), we have the approximation error $\|\theta - \pi_n\theta\| = O(r_n^{-m})$. For A.3, note that $H(w, \mathscr{F}_n) \le c(r_n + m + 1)^d \log(\delta/w)$ for $0 < w < \delta$ and some constant $c > 0$; hence we get $\delta_n = O([(r_n + m + 1)^d/n]^{1/2})$. Now Theorem 1 implies that $r_n = O(n^{1/(2m+d)})$ and the convergence rate $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-m/(2m+d)})$.

Case 1.3 (Wavelet): By Proposition 7 of Meyer (1990, Ch. 6, p. 200), we have for any $\theta \in \Theta$, that there exists $\pi_n\theta \in \Theta_n$ such that $\|\theta - \pi_n\theta\| \le \text{const.} 2^{-Jm'}$, where $m' = m > 0$ if $p \ge 2$; and $m' = m - d(1/p - 1/2) > 0$ if $1 \le p < 2$. For A.3, by Lemma 2.1 of Ossiander (1987), we have $H(w, \mathscr{F}_n) \le \text{const.} 2^{dJ} \log(C/w)$;

$$\delta^{-2}\int_{\delta^2}^{\delta} H^{1/2}(w, \mathscr{F}_n)dw \le (2^{dJ_n})^{1/2}\delta^{-2}\int_{\delta^2}^{\delta}[\log(C/w)]dw \le (2^{dJ_n})^{1/2}\delta^{-1} \le n^{1/2}.$$

Thus A.3 is satisfied with δ_n such that $\delta_n^{-1} 2^{(dJ_n)/2} = O(n^{1/2})$. Now the convergence rates follow from Theorem 1 by setting $\delta_n = \|\theta - \pi_n\theta\|$. *Q.E.D.*

PROOF OF PROPOSITION 2: (i) Condition A.1 is assumed. Conditions A.2 and A.4 can be verified as those in Example 1 or Shen and Wong's (1994) Example 3.

Case 2.1 (Fourier series without monotone constraint): Condition A.3 is satisfied with $\delta_n = \text{const.}(r_n/n)^{1/2}$ by the same calculation as in Shen and Wong's (1994) Example 3. Lorentz (1966) gives the deterministic approximation error $\|\pi_n\eta - \eta\| \le \sup_{x, 1 \le i \le p-p_1} |\pi_n\eta_i(x) - \eta_i(x)| \le cr_n^{-m}$.

Case 2.2 (Spline with monotone constraint): A.3 is satisfied with $\delta_n = O([(r_n + m + 1)/n]^{1/2})$ by the same calculation as that for Case 1.2 in Example 1. By DeVore (1977) for shape-preserving spline, we have the deterministic approximation error $\|\theta - \pi_n\theta\| = O(r_n^{-m})$.

Now Theorem 1 implies the convergence rate $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-m/(2m+1)})$ with $r_n = O(n^{1/(2m+1)})$ for both cases.

(ii) Let $f(\theta) = \lambda^T\alpha$, where λ is an arbitrary unit vector in $\mathscr{R}^{p_1}$. Clearly, (4.2) is satisfied with $f'_{\theta_0}[\theta - \theta_0] = (\alpha - \alpha_0)^T\lambda$ and $\omega = \infty$. In addition,

$$\begin{aligned}\|v^*\|^2 &= \sup_{\{\theta \in \Theta : \|\theta - \theta_0\| > 0\}} \frac{((\alpha - \alpha_0)^T\lambda)^2}{\|\theta - \theta_0\|^2} \\ &= \sup_{\{(b,h): b \ne 0, h \in D, \|Z + h(X)\| > 0\}} \frac{(b^T\lambda)^2}{b^T E[(Z_t + h(X_t))(Z_t + h(X_t))^T]b} \\ &= \lambda^T\Sigma^{-1}\lambda,\end{aligned}$$

where $\Sigma \equiv E[(Z_t - E(Z_t \mid X_t))(Z_t - E(Z_t \mid X_t))^T]$ is positive definite. Thus

$$v^* = \left(\Sigma^{-1}\lambda, -(\Sigma^{-1}\lambda)^T E(Z_t \mid X_t) \right).$$

Furthermore, $v^* \in \bar{V}$ if the conditional density of $Z_t \mid X_t$ is smooth enough.

Condition B.2 is satisfied. To check conditions B.1, B.3–B.5, note that $\{\theta(Z_t, X_t)\}_{t=1}^n$ is a stationary β-mixing process with coefficients satisfying Condition A.1. Choose $P_n(\theta^*(\theta, \varepsilon_n)) = (1 - \varepsilon_n)\theta + \varepsilon_n(\pi_n u^* + \pi_n \theta_0)$ for any $\theta \in \Theta_n$, and $u^* = \pm v^* \in \Theta$. Because

$$l'_{\theta_0}[\theta - \theta_0, Y_t, Z_t, X_t] = -[Z_t^T(\alpha - \alpha_0) + (\eta(X_t) - \eta_0(X_t))][Y_t - (Z_t^T\alpha_0 + \eta_0(X_t))],$$

$r[\theta - \theta_0, y, z, x] = -(\theta - \theta_0)^2(y, z, x)$. To verify Condition B.1, it suffices to calculate

$$\text{(b.3)} \qquad \sup_{\{\theta \in \Theta : \|\theta - \theta_0\| \le \delta_n\}} \mu_n((\theta - \pi_n\theta_0)(\pi_n u^* + \pi_n\theta_0 - \theta_0)) = O_p(\varepsilon_n),$$

and

$$\text{(b.4)} \qquad \sup_{\{\theta \in \Theta : \|\theta - \theta_0\| \le \delta_n\}} \mu_n((\theta - \theta_0)^2) = O_p(\varepsilon_n).$$

By Theorem 3, we obtain that the convergence rates of the empirical processes in (b.3) and (b.4) are both $O_P(n^{-1}r_n)$. Thus Condition B.1 is satisfied if

$$\text{(b.5)} \qquad n^{-1}r_n < n^{-1/2}.$$

Since

$$\|P_n(\theta^*(\theta, \varepsilon_n)) - \theta^*(\theta, \varepsilon_n)\| = \varepsilon_n \|\pi_n(u^* + \theta_0) - (u^* + \theta_0)\| \le \varepsilon_n r_n^{-m},$$

condition B.3 holds if

$$\text{(b.6)} \qquad \varepsilon_n r_n^{-m} = O(\max(\varepsilon_n^2\delta_n^{-1}, \varepsilon_n)).$$

Both (b.5) anf (b.6) are satisfied if $r_n = O(n^{1/(2m+1)})$ and $m > 1/2$. Condition B.4 is also satisfied with this choice of r_n and m. Note that $W_t \equiv Z_t - E(Z_t \mid X_t)$. We have

$$l'_{\theta_0}[v^*, Y_t, Z_t, X_t] = -\left[(\Sigma^{-1}\lambda)^T Z_t - (\Sigma^{-1}\lambda)^T E(Z_t \mid X_t) \right] e_t = (\Sigma^{-1}\lambda)^T W_t e_t.$$

Using the assumptions on $\{Y_t, Z_t, X_t\}$, we know that $\{l'_{\theta_0}[v^*, Y_t, Z_t, X_t]\}_{t=1}^n$ is a β-mixing satisfying Remark 2(b). This implies that the sequence satisfies Condition B.5. By Theorem 2, we obtain for any arbitrary unit vector $\lambda \in \mathscr{R}^{p_1}$,

$$n^{1/2}\lambda^T(\hat{\alpha}_n - \alpha_0) \xrightarrow{P\theta_0} \mathscr{N}(0, \sigma_*^2), \qquad \sigma_*^2 \equiv \lim_{n \to \infty} n^{-1} E\left[\sum_{t=1}^n (\Sigma^{-1}\lambda)^T W_t e_t \right]^2.$$

Consequently, $n^{1/2}(\hat{\alpha}_n - \alpha_0) \xrightarrow{P\theta_0} \mathscr{N}(0, \Omega_*)$, with

$$\Omega_* = \Sigma^{-1}\left(E[e_1^2 W_1 W_1^T] + \sum_{j=2}^{\infty} E[e_1 e_j W_1 W_j^T + e_j e_1 W_j W_1^T] \right) \Sigma^{-1}. \qquad Q.E.D.$$

PROOF OF PROPOSITION 3: Conditions (5.3.1)–(5.3.4) imply that $\{Y_t\}$ is stationary β-mixing with exponential decay. Thus Condition A.1 is satisfied. Conditions A.2 can be verified by the Taylor expansion, the norm relationship $\|\eta\|_{\sup} \le c_4^{-1} r_n^{1/2} \|\eta\|$ for any $\eta \in \Theta_n$, and the deterministic approximation error $\|\eta_0 - \pi_n\eta_0\|_{\sup} = O(r_n^{-1})$. Condition A.4 can be verified with $s = 2/3$ using the assumption on tail distribution of e_t and again the norm relationship $\|\eta - \eta_0\|_{\sup} \le C r_n^{1/2} \|\eta - \pi_n\eta_0\| + O(r_n^{-1})$ for any $\eta \in \Theta_n$. Condition A.3 is fulfilled with $\delta_n = n^{-1/2} r_n^{1/2}$ since $H(w, \mathscr{T}_n) \le$ const.$(r_n + 1)\log(1/w)$ for small $w > 0$. Furthermore the approximation error is $\|\eta_0 - \pi_n\eta_0\| = O(r_n^{-1})$. By Theorem 1, the convergence rate is $\|\hat{\eta}_n - \eta_0\| = O_p(\max(n^{-1/2} r_n^{1/2}, r_n^{-1})) = O_p(n^{-1/3})$ when $r_n = O(n^{1/3})$.

Let $f(\theta)=\alpha$. Easily, $f'_{\theta_0}[\theta-\theta_0]=\alpha-\alpha_0$ with $\omega=\infty$. As for Proposition 2, we have $v^*=I_{\theta_0}^{-1}(1,-h^*)$, where

$$\text{(b.7)}\qquad I_{\theta_0}=\inf_{\{\pm h\in D-\{\eta_0\}\}} E[e_t(Y_{t-1}-h(Y_t))+h'(Y_t)/\eta_0'(Y_t)]^2$$

$$=E[e_t(Y_{t-1}-h^*(Y_t))+(h^*)'(Y_t)/\eta_0'(Y_t)]^2.$$

To see $I_{\theta_0}>0$, note that $I_{\theta_0}=0$ implies that $e_t(Y_{t-1}-h^*(Y_t))+(h^*)'(Y_t)/\eta_0'(Y_t)=0$ with probability one, i.e., $h(\cdot)$ is the solution of the following differential equation: $h'(Y_t)=\eta_0'(Y_t)e_t(h(Y_t)-Y_{t-1})$ almost sure, which is impossible given the assumption of e_t. The solution of the above infinite dimensional optimization exists but does not appear to have a simple explicit expression. In practice, such an infinite dimensional problem can be approximated by a finite dimensional optimization problem based on data. The approximated h^* can be calculated numerically.

By Theorem 3, the convergence rate of the corresponding empirical process in Conditions B.1 is $O_p(n^{-1/2}r_n n^{-1})=O_p(\varepsilon_n^2)$. Condition B.2 can be verified by applying a Taylor expansion and using the norm relationship. The approximation error in Condition B.3 is bounded by $O(n^{-1/2}r_n^{1/2})$. The second condition of B.3 and Condition B.4 can be verified as in Condition B.1. Conditions B.1–B.4 are satisfied when $r_n=O(n^{1/3})$. Because

$$l'_{\theta_0}[\theta-\theta_0]=e_t([\alpha-\alpha_0]Y_{t-1}-[\eta(Y_t)-\eta_0(Y_t)])+[\eta'(Y_t)-\eta_0'(Y_t)]/\eta_0'(Y_t),$$

$\{l'_{\theta_0}[v^*,Y_t]\}$ is stationary β-mixing satisfying Condition B.5. By Theorem 2, $(\hat\alpha_n-\alpha_0)\xrightarrow{P\theta_0}\mathcal{N}(0,\sigma_*^2)$, with

$$\sigma_*^2=\operatorname{var}_0(l'_{\theta_0}[v^*,Y_1])+2\sum_{j=2}^{\infty}\operatorname{cov}(l'_{\theta_0}[v^*,Y_1],l'_{\theta_0}[v^*,Y_j]). \qquad Q.E.D.$$

REFERENCES

AMEMIYA, T. (1985): *Advanced Econometrics*. Cambridge: Harvard University Press.

ANDREWS, D. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semi-parametric Regression Models," *Econometrica*, 59, 307–346.

——— (1994): "Asymptotics for Semi-parametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62, 43–72.

BARRON, A. (1993): "Universal Approximation Bounds for Superpositions of a Sigmoidal Function," *IEEE Transactions on Information Theory*, 39, 930–945.

——— (1994): "Approximation and Estimation Bounds for Artificial Neural Networks," *Machine Learning*, 14, 115–133.

BIRGÉ, L., AND P. MASSART (1993): "Rate of Convergence for Minimum Contrast Estimators," *Probability Theory and Related Fields*, 97, 113–150.

——— (1994): "Minimum Contrast Estimators on Sieves," Technical Report, Universite Paris-Sud.

CHEN, X., AND X. SHEN (1996): "Asymptotic Properties of Sieve Extremum Estimates for Weakly Dependent Data with Applications," Technical Report, University of Chicago.

CYBENKO, G. (1989): "Approximation by Superpositions of a Sigmoid Function," *Mathematics of Control, Signals, and Systems*, 2, 303–314.

DAVYDOV, Y. A. (1973): "Mixing Conditions for Markov Chains," *Theory of Probability and Its Applications*, 18, 312–328.

DEVORE, R. A. (1977): "Monotone Approximation by Splines," *SIAM Journal on Mathematical Analysis*, 8, 891–905.

DEVORE, R. A., AND G. LORENTZ (1993): *Constructive Approximation*. New York: Springer-Verlag.

DONOHO, D. L., I. M. JOHNSTONE, G. KERKYACHARIAN, AND D. PICARD (1995): "Wavelet Shrinkage: Asymptopia?" *Journal of the Royal Statistical Society, Series B*, 57, 301–369.

DOUKHAN, P. (1994): *Mixing: Properties and Examples*. New York: Springer-Verlag.

EBERLEIN, E. (1984): "Weak Convergence of Partial Sums of Absolutely Regular Sequences," *Statistics and Probability Letters*, 2, 291–293.

FENTON, V., AND A. R. GALLANT (1996): "Convergence Rate of SNP Density Estimators," *Econometrica*, 64, 719–727.

GABUSHIN, V. N. (1967): "Inequalities for Norms of Functions and their Derivatives in the L_p Metric," *Matematicheskie Zametki*, 1, 291–298.

GALLANT, A. R., AND D. NYCHKA (1987): "Semi-non-parametric Maximum Likelihood Estimation," *Econometrica*, 55, 363–390.

GALLANT, A. R., P. E. ROSSI, AND G. TAUCHEN (1993): "Nonlinear Dynamic Structures," *Econometrica*, 61, 871–907.

GEMAN, S., AND C. HWANG (1982): "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *The Annals of Statistics*, 10, 401–414.

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

HECKMAN, J. J., AND B. SINGER (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, 271–320.

HOROWITZ, J. L. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64, 103–138.

ICHIMURA, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58, 71–120.

LORENTZ, G. (1966): *Approximation of Functions*. New York: Holt.

MAKOVOZ, Y. (1996): "Random Approximants and Neural Networks," *Journal of Approximation Theory*, 85, 98–109.

MCCAFFREY, D. F., AND A. R. GALLANT (1994): "Convergence Rates for Single Hidden Layer Feedforward Networks," *Neural Networks*, 7, 147–158.

MEYER, Y. (1990): *Ondelettes et Operateurs I: Ondelettes*. Paris: Hermann. (English translation, 1992, Cambridge University Press.)

MEYN, S. P., AND R. L. TWEEDIE (1992): "Stability of Markovian Processes I: Criteria for Discrete-time Chains," *Advances in Applied Probability*, 24, 542–574.

MODHA, D. S., AND E. MASRY (1996): "Minimum Complexity Regression Estimation with Weakly Dependent Observations," *IEEE Transactions on Information Theory*, 42, 2133–2145.

NEWEY, W. (1994): "The Asymptotic Variance of Semi-Parametric Estimators," *Econometrica*, 62, 1349–1382.

OSSIANDER, M. (1987): "A Central Limit Theorem under Metric Entropy with L_2 Bracketing," *The Annals of Probability*, 15, 897–919.

POLLARD, D. (1984): *Convergence of Stochastic Processes*. New York: Springer-Verlag.

ROBINSON, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

——— (1991): "Best Nonlinear Three-stage Least Squares Estimation of Certain Econometric Models," *Econometrica*, 59, 755–786.

SCHUMAKER, L. (1981): *Spline Functions: Basic Theory*. New York: John Wiley & Sons.

SHEN, X. (1997): "On Methods of Sieves and Penalization," forthcoming in *The Annals of Statistics*.

SHEN, X, AND W. H. WONG (1994): "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22, 580–615.

STONE, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053.

WHITE, H. (1990): "Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings," *Neural Networks*, 3, 535–550.

WHITE, H., AND J. WOOLDRIDGE (1991): "Some Results on Sieve Estimation with Dependent Observations," in *Non-parametric and Semi-parametric Methods in Econometrics and Statistics*, ed. by. W. A. Barnett, J. Powell, and G. Tauchen. New York: Cambridge University Press.

WONG, W. H., AND T. A. SEVERINI (1991): "On Maximum Likelihood Estimation in Infinite Dimensional Parameter Spaces," *The Annals of Statistics*, 19, 603–632.

YOKOYAMA, R. (1980): "Moment Bounds for Stationary Mixing Sequences," *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 52, 45–57.