Chapter 76

LARGE SAMPLE SIEVE ESTIMATION OF SEMI-NONPARAMETRIC MODELS*

XIAOHONG CHEN

Department of Economics, Yale University, Box 208281, New Haven, CT 06520, USA
e-mail: *xiaohong.chen@yale.edu*

Contents

* The author thanks C. Ai, J. Heckman, B. Honore, J. Huang, G. Imbens, R. Matzkin, W. Newey, J. Powell and H. White for valuable suggestions, J. Huang for showing his work on concave extended linear models, and two anonymous referees for critical comments that lead to thorough revisions. She also thanks K. Hyndman, A. Ingster, M. Kredler, D. Pouzo and R. Sela for proof-reading, M. Garibotti, D. Pouzo and V. Tsyrennikov for simulations and other PhD students who went through earlier versions used as the lecture notes for *Topics in Econometrics* during the Fall 2002, Fall 2003, Spring 2005 and Fall 2005 sessions at New York University. The author acknowledges financial support from the National Science Foundation and the C.V. Starr Center at NYU. Any errors or omissions are the responsibility of the author.

Handbook of Econometrics, Volume 6B

DOI: 10.1016/S1573-4412(07)06076-X

Abstract

Often researchers find parametric models restrictive and sensitive to deviations from the parametric specifications; semi-nonparametric models are more flexible and robust, but lead to other complications such as introducing infinite-dimensional parameter spaces that may not be compact and the optimization problem may no longer be well-posed. The method of sieves provides one way to tackle such difficulties by optimizing an empirical criterion over a sequence of approximating parameter spaces (i.e., sieves); the sieves are less complex but are dense in the original space and the resulting optimization problem becomes well-posed. With different choices of criteria and sieves, the method of sieves is very flexible in estimating complicated semi-nonparametric models with (or without) endogeneity and latent heterogeneity. It can easily incorporate prior information and constraints, often derived from economic theory, such as monotonicity, convexity, additivity, multiplicity, exclusion and nonnegativity. It can simultaneously estimate the parametric and nonparametric parts in semi-nonparametric models, typically with optimal convergence rates for both parts.

This chapter describes estimation of semi-nonparametric econometric models via the method of sieves. We present some general results on the large sample properties of the sieve estimates, including consistency of the sieve extremum estimates, convergence rates of the sieve M-estimates, pointwise normality of series estimates of regression functions, root-n asymptotic normality and efficiency of sieve estimates of smooth functionals of infinite-dimensional parameters. Examples are used to illustrate the general results.

Keywords

sieve extremum estimation, series, sieve minimum distance, semiparametric two-step estimation, endogeneity in semi-nonparametric models

JEL classification: C13, C14, C20

1. Introduction

Semiparametric and nonparametric modelling techniques have grown increasingly popular in both theoretical and applied econometrics.[1] This is partly because economic theory seldom suggests any parametric functional relationships among economic variables, nor does it suggest particular parametric forms for error distributions. An additional reason for the growing popularity of semi-nonparametric models is the declining computational cost of collecting and analyzing large economic data sets. All of the chapters in the book edited by Barnett, Powell and Tauchen (1991) and several chapters[2] in the Handbook of Econometrics Volume 4 edited by Engle and McFadden (1994) have already reviewed the work in semiparametric and nonparametric econometrics that has been conducted up to the mid-1990s. More recently, Horowitz (1998) has provided a comprehensive treatment of four leading classes of semiparametric econometric models estimated via the kernel method. Pagan and Ullah (1999), Härdle et al. (2004) and Li and Racine (2007) have surveyed the most well-known existing theoretical and empirical work on the estimation and testing of semiparametric and nonparametric econometric models via the methods of kernel, local linear regression and series. This chapter will review some recent developments in large sample theory on estimation of semi-nonparametric models via the *method of sieves* [Grenander (1981)].

Semi-nonparametric models involve unknown parameters that lie in infinite-dimensional parameter spaces; hence it can be computationally difficult to estimate such models using finite samples. Moreover, even if one could solve the problem of optimizing a sample criterion over an infinite-dimensional parameter space, the resulting estimator may have undesirable large sample properties such as inconsistency and/or a very slow rate of convergence; this is because the problem of optimization over an infinite-dimensional noncompact space may no longer be well-posed. To resolve this problem, the method of sieves optimizes a criterion function over a sequence of significantly less complex, and often finite-dimensional, parameter spaces, which we call *sieves*. To ensure consistency of the method, we require that the complexity of sieves increases with the sample size so that in the limit the sieves are dense in the original parameter space.[3]

The infinite-dimensional unknown parameter in a nonparametric or semiparametric model can often be viewed as a member of some function space with certain regularities (e.g., having bounded second derivatives, monotone, concave). Thus, many deterministic approximation results developed in mathematics and computer science can be used to

[1] In this chapter, an econometric model is termed "*parametric*" if all of its parameters are in finite-dimensional parameter spaces; a model is "*nonparametric*" if all of its parameters are in infinite-dimensional parameter spaces; a model is "*semiparametric*" if its parameters of interests are in finite-dimensional spaces but its nuisance parameters are in infinite-dimensional spaces; a model is "*semi-nonparametric*" if it contains both finite-dimensional and infinite-dimensional unknown parameters of interests.

[2] See the ones written by Newey and McFadden (1994), Andrews (1994a), Powell (1994), Härdle and Linton (1994), Matzkin (1994), Manski (1994) and others.

[3] These terms will become much clearer in the next two sections.

suggest sieves that provide good and computable approximations to an unknown function. For example, the sieves or approximating spaces can be constructed using linear spans of power series, Fourier series, splines or many other basis functions; see e.g. Judd (1998, Chapters 6 and 12) for numerical implementation of such sieves for problems in economics and finance. Since these approximating spaces can often be characterized by a finite number of "parameters", a nonparametric or semiparametric estimation problem is often reduced to a parametric one when the method of sieves is implemented. However, to obtain the desired theoretical properties of the estimator, it is necessary that the number of parameters increase slowly with the sample size. It is this feature that gives the sieve method its added flexibility and robustness over classical parametric methods which assume fixed, finite-dimensional parameter spaces.

One attractive feature of the method of sieves is that it is easy to implement. The sieve method is particularly convenient when the unknown functions enter the criterion function (or moment condition) nonlinearly, satisfy some known restrictions such as monotonicity, concavity, additivity, multiplicity and exclusion, or when the error distribution has known tail behavior such as fat tails. With different choices of criteria and sieves, the method of sieves provides a flexible and computationally feasible approach to estimate complicated semi-nonparametric models with (or without) constraints, endogeneity and latent heterogeneity. Moreover, it can simultaneously estimate the parametric and nonparametric components in semi-nonparametric models, and can often achieve optimal convergence rates for both parts. We shall demonstrate these with some examples in the subsequent sections.

Although the method of sieves is easy to implement and the sieve estimators typically have desirable large sample properties, its theoretical properties cannot be justified by applying the classical theory for parametric models. Any appropriate large sample theory for the sieve method should not only account for the approximation errors, which arise because we replace the original parameter space with the simpler sieve space, but also control for the complexity of the sieve parameter spaces, which increases with the sample size. Consequently, the large sample properties of the sieve method are in general difficult to derive, which may partly explain why currently there are fewer econometric applications using such techniques than those using the kernel method. However, we should mention that the sieve estimation method admits, as special cases, many standard estimation methods (such as series-based method) in econometrics. As a result, some large sample results appear in the literature in papers that do not mention the word "sieve" at all.

In this chapter we shall present some general results on large sample estimation theory using the method of sieves and illustrate how to apply these results with examples. Instead of presenting the current sieve estimation theory at its greatest generality, we have chosen to review results that are relatively accessible but general enough to cover most semi-nonparametric econometric applications. References are given for the results that are not presented in detail.

The rest of this chapter is organized as follows. In Section 2, we first present several examples of semi-nonparametric econometric models. We then define the sieve extremum estimation and its special cases including sieve M-estimation, sieve maximum likelihood estimation (MLE), sieve generalized least squares (GLS), sieve minimum distance (MD) and others. The various criterion functions are illustrated using examples. In addition, we introduce the popular *series* estimators as the sieve M-estimators obtained when the criterion functions are concave and the sieve spaces are finite-dimensional linear.[4] We then review typical function spaces and sieve spaces used in econometrics, and conclude this section with a small Monte Carlo study to demonstrate the implementation of the sieve extremum estimation.[5] Section 3 focuses on the large sample properties of sieve estimation of infinite-dimensional unknown parameters. We first provide a new consistency theorem for general sieve extremum estimation where the original parameter space may not be compact and the problem may not be well-posed. This theorem implies consistency of sieve M-estimators and of sieve MD-estimators in two remarks. We then present a convergence rate result for sieve M-estimators and illustrate how to apply the result with some examples. We also review the convergence rate and the pointwise asymptotic normality results for the series estimators. In Section 4, we present general results on $\sqrt{n}$-asymptotic normality of sieve estimators of smooth functionals of unknown infinite-dimensional parameters, where n denotes the sample size. Here we first discuss the popular two-step semiparametric procedures in which the first step unknown functions could be estimated by any nonparametric procedures such as kernel, local linear regression and sieve methods, and the second step unknown parametric components are estimated by the generalized method of moments (GMM). The theorem on $\sqrt{n}$-asymptotic normality of the second step GMM estimator is a slight refinement of the existing ones in the semiparametric literature. We then review the $\sqrt{n}$-asymptotic normality of the sieve M-estimation of smooth functionals of unknown functions, as well as the semiparametric efficiency of the sieve MLE. Finally we present the recent theory on the sieve MD estimation for the parametric components in semi-nonparametric conditional moment models where the unknown functions could depend on endogenous variables. Section 5 points out additional topics on statistical inference via the method of sieves that are not reviewed here due to the lack of space.

Throughout this chapter, we assume that there is an underlying complete probability space, the data $\{Z_t = (Y_t', X_t')': t \geqslant 1\}$ are strictly stationary ergodic,[6] and all probability calculations are done under the true probability measure P_o. For random variables V_n and positive numbers b_n, $n \geqslant 1$, we define $V_n = O_P(b_n)$ as $\lim_{c\to\infty} \limsup_n P(|V_n| \geqslant$

[4] We note that this definition of series estimators differs slightly from those in the current econometrics literature.

[5] See the chapter by Ichimura and Todd (2007) for more details on the implementation of semi-nonparametric estimators.

[6] In this chapter, the notation $'$ denotes the transpose of a vector. See Hansen (1982), White (1984) or Wooldridge (1994) for the definition of a strictly stationary ergodic process. We make this assumption to simplify the presentation. See White and Wooldridge (1991) on sieve extremum estimation for general dependent heterogeneous processes.

$cb_n) = 0$, and define $V_n = o_P(b_n)$ as $\lim_n P(|V_n| \geqslant cb_n) = 0$ for all $c > 0$. The notation $\text{plim}_{n\to\infty} V_n = 0$ also means that $V_n = o_P(1)$ (i.e., V_n converges to 0 in probability). Similarly $V_n = o_{a.s.}(1)$ means that V_n converges to 0 almost surely. For two sequences of positive numbers b_{1n} and b_{2n}, the notation $b_{1n} \asymp b_{2n}$ means that the ratio b_{1n}/b_{2n} is bounded below and above by positive constants that are independent of n.

2. Sieve estimation: Examples, definitions, sieves

As alluded to in the introduction, the method of sieves consists of two key ingredients: a criterion function and sieve parameter spaces (a sequence of approximating spaces). Both the criterion functions and the sieve spaces can be very flexible. In particular, almost all of the classical criterion functions stated in Newey and McFadden (1994), so long as they still allow for identification, can be used as criterion functions in the method of sieve estimation. Therefore, the main new ingredient is the choice of sieve parameter spaces, which will be discussed in this section.

2.1. Empirical examples of semi-nonparametric econometric models

It is impossible to list all of the existing and potential semi-nonparametric models and their empirical applications in econometrics. In this subsection we present three empirical examples as illustration; additional ones can be found in Manski (1994), Powell (1994), Matzkin (1994), Horowitz (1998), Pagan and Ullah (1999), Blundell and Powell (2003) and other surveys on this topic.

EXAMPLE 2.1 *(Single spell duration models with unobserved heterogeneity).* Classical single spell duration models in search unemployment [Flinn and Heckman (1982)], job turnover [Jovanovic (1979)], labor supply [Heckman and Willis (1977)] and others often suggest a functional form for the structural duration distribution conditional on individual heterogeneity. More specifically, let $G(\tau|u, x)$ be the structural distribution function of duration T conditional on a scalar of unobserved heterogeneity $U = u$ and a vector of observed heterogeneity $X = x$. The distribution of observed duration given $X = x$ is

$$F(\tau|x) = \int G(\tau|u, x)\,\mathrm{d}h(u),$$

where the unobserved heterogeneity U is modelled as a random factor with distribution function $h(\cdot)$. An i.i.d. sample of observations $\{T_i, X_i\}_{i=1}^n$ allows us to recover the true $F(\tau|x)$ uniquely. Theoretical models often imply parametric functional forms of G up to unknown finite-dimensional parameters β. Denote $g(\cdot|\beta, u, x)$ as the probability density function of $G(\cdot|\beta, u, x)$. Conventional parametric MLE method assumes that the unobserved heterogeneity follows some known distribution h_γ up to some unknown finite-dimensional parameters γ. Under this assumption it then estimates the unknown parameters β, γ by $\arg\max_{\beta,\gamma} \frac{1}{n}\sum_{i=1}^n \log\{\int g(T_i|\beta, u, X_i)\,\mathrm{d}h_\gamma(u)\}$.

Heckman and Singer (1984) point out that both theoretical and empirical examples indicate that the parametric MLE estimates of structural parameters β in these duration models are inconsistent if the distribution of the unobserved heterogeneity is misspecified. Instead, they propose the following semi-nonparametric single spell duration model

$$F(\tau|\beta, h, x) = \int G(\tau|\beta, u, x)\,\mathrm{d}h(u), \tag{2.1}$$

where the distribution h of unobserved heterogeneity is left unspecified. Heckman and Singer (1984) establish the identification of (β', h), and propose a sieve MLE method to estimate (β', h) jointly. They also show that their estimator is consistent.

The Heckman–Singer model is a typical example of a broad class of semi-nonparametric models that specify the (conditional) distribution associated with the observed economic variables semi-nonparametrically, where the specific semi-nonparametric form can be derived from independence of errors and regressors such as in discrete choice models, transformation models, sample selection models, mixture models, random censoring, nonlinear measurement errors and others. More generally, one could consider semi-nonparametric models based on quantile independence, symmetry or other qualitative restrictions on distributions. See Horowitz (1998), Manski (1994), Powell (1994) and Bickel et al. (1993) for examples.

EXAMPLE 2.2 *(Shape-invariant system of Engel curves).* Blundell, Browning and Crawford (2003) have shown that a system of Engel curves that satisfies Slutsky's symmetry condition and allows for demographic effects on budget shares in a given year must take the following form:

$$Y_{1\ell i} = h_{1\ell}\big(Y_{2i} - h_0(X_{1i})\big) + h_{2\ell}(X_{1i}) + \varepsilon_{\ell i}, \quad \ell = 1, \ldots, N,$$

where $Y_{1\ell i}$ is the ith household budget share on ℓth goods, Y_{2i} is the ith household log-total nondurable expenditure, X_{1i} is a vector of the ith household demographic variables that affect the household's nondurable consumption. Note that $h_0(X_{1i})$ is common among all the goods and is called an "equivalence scale" in the consumer demand literature. Citing strong empirical evidence and many existing works, Blundell, Browning and Crawford (2003) have argued that popular parametric linear and quadratic forms for $h_{1\ell}(\cdot)$ are inadequate, and that consumer demand theory only suggests the purely nonparametric specification:

$$\begin{aligned} &E\big[Y_{1\ell i} - \big\{h_{1\ell}\big(Y_{2i} - h_0(X_{1i})\big) + h_{2\ell}(X_{1i})\big\}\big|X_{1i}, Y_{2i}\big] \\ &\quad = E[\varepsilon_{\ell i}|X_{1i}, Y_{2i}] = 0, \end{aligned} \tag{2.2}$$

where $h_{1\ell}$, $h_{2\ell}$ and h_0 are all unknown functions. For the identification of all these unknown functions $\theta = (h_0, h_{11}, \ldots, h_{1N}, h_{21}, \ldots, h_{2N})'$ satisfying (2.2), it suffices to assume that at least one of $h_{1\ell}$, $\ell = 1, \ldots, N$, is nonlinear and that $h_{2\ell}(x_1^*) = 0$, $\ell = 1, \ldots, N$, for some x_1^* in the support of X_1.

Unfortunately, when X_{1i} contains too many household demographic variables (say when $\dim(X_{1i}) \geqslant 3$), the fully nonparametric specification (2.2) cannot lead to precise estimates of the unknown functions $h_0, h_{21}, \ldots, h_{2N}$ due to the so-called "curse of dimensionality". Therefore, applied researchers must impose more structure on the model. Using the British family expenditure survey (FES) data, Blundell, Duncan and Pendakur (1998) found the following semi-nonparametric specification to be reasonable:

$$E\big[Y_{1\ell i} - \big\{h_{1\ell}\big(Y_{2i} - g\big(X_{1i}'\beta_1\big)\big) + X_{1i}'\beta_{2\ell}\big\}\big|X_{1i}, Y_{2i}\big] = 0, \tag{2.3}$$

where $h_{1\ell}$, $\ell = 1, \ldots, N$, are still unknown functions, but now $h_0(X_{1i}) = g(X_{1i}'\beta_1)$ and $h_{2\ell}(X_{1i}) = X_{1i}'\beta_{2\ell}$ are known up to unknown finite-dimensional parameters β_1 and $\beta_{2\ell}$. Here the parameters of interest are $\theta = (\beta_1', \beta_{21}', \ldots, \beta_{2N}', h_{11}, \ldots, h_{1N})'$. This semi-nonparametric specification has been estimated by Blundell, Duncan and Pendakur (1998) using the kernel method and Blundell, Chen and Kristensen (2007) using the sieve method.

Both the specifications (2.2) and (2.3) assume that the total nondurable expenditure Y_{2i} is exogenous. However, this assumption has been rejected empirically. Noting the endogeneity of total nondurable expenditure, Blundell, Chen and Kristensen (2007) considered the following semi-nonparametric instrumental variables (IV) regression:

$$E\big[Y_{1\ell i} - \big\{h_{1\ell}\big(Y_{2i} - g\big(X_{1i}'\beta_1\big)\big) + X_{1i}'\beta_{2\ell}\big\}\big|X_{1i}, X_{2i}\big] = 0, \tag{2.4}$$

where the parameters of interest are still $\theta = (\beta_1', \beta_{21}', \ldots, \beta_{2N}', h_{11}, \ldots, h_{1N})'$, and X_{2i} is the gross earnings of the head of the ith household which is used as an instrument for the total nondurable expenditure Y_{2i}. They estimated this model via the sieve method and their empirical findings demonstrate the importance of accounting for the endogenous total expenditure semi-nonparametrically.

EXAMPLE 2.3 *(Consumption-based asset pricing models).* A standard consumption-based asset pricing model assumes that at time zero a representative agent maximizes the expected present value of the total utility function $E_0\{\sum_{t=0}^{\infty} \delta^t u(C_t)\}$, where δ is the time discount factor and $u(C_t)$ is period t's utility. The consumption-based asset pricing model comes from the first-order conditions of a representative agent's optimal consumption choice problem. These first-order conditions place restrictions on the joint distribution of the intertemporal marginal rate of substitution in consumption and asset returns. They imply that for any traded asset indexed by ℓ, with a gross return at time $t+1$ of $R_{\ell,t+1}$, the following Euler equation holds:

$$E(M_{t+1}R_{\ell,t+1}|\mathbf{w}_t) = 1, \quad \ell = 1, \ldots, N, \tag{2.5}$$

where M_{t+1} is the intertemporal marginal rate of substitution in consumption, and $E(\cdot|\mathbf{w}_t)$ denotes the conditional expectation given the information set at time t (which is the sigma-field generated by $\mathbf{w}_t$). More generally, any nonnegative random variable M_{t+1} satisfying Equation (2.5) is called a stochastic discount factor (SDF); see Hansen and Richard (1987) and Cochrane (2001).

Hansen and Singleton (1982) have assumed that the period t utility takes the power specification $u(C_t) = [(C_t)^{1-\gamma} - 1]/[1-\gamma]$, where γ is the curvature parameter of the utility function at each period, which implies that the SDF takes the form $M_{t+1} = \delta(\frac{C_{t+1}}{C_t})^{-\gamma}$ and the Euler equation becomes:

$$E\left(\delta_o\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma_o} R_{\ell,t+1} - 1\middle|\mathbf{w}_t\right) = 0, \quad \ell = 1, \ldots, N, \tag{2.6}$$

where the unknown scalar parameters δ_o, γ_o can be estimated by Hansen's (1982) generalized method of moment (GMM). However, this classical power utility-based asset pricing model (2.6) has been rejected empirically.

Many subsequent papers have tried to relax the model (2.6) to fit the data better by introducing durable goods, habit formation or a nonseparable preference specification. The first class of papers proposes various parametric forms of the SDF, M_{t+1}, that are more flexible than $M_{t+1} = \delta(\frac{C_{t+1}}{C_t})^{-\gamma}$; see e.g. Eichenbaum and Hansen (1990), Constantinides (1990), Campbell and Cochrane (1999). The second class of papers has made the SDF, M_{t+1}, a purely nonparametric function of a few state variables; see e.g. Gallant and Tauchen (1989), Newey and Powell (1989) and Bansal and Viswanathan (1993). Recently, Chen and Ludvigson (2003) have specified the SDF, M_{t+1}, to be semi-nonparametric in order to incorporate some preference parameters. In particular, they combine the power utility specification with a nonparametric internal habit formation: $E_0\{\sum_{t=0}^{\infty} \delta^t[(C_t - H_t)^{1-\gamma} - 1]/[1-\gamma]\}$, where $H_t = H(C_t, C_{t-1}, \ldots, C_{t-L})$ is the period t habit level. Here $H(\cdot)$ is a homogeneous of degree one unknown function of current and past consumption, and can be rewritten as $H(C_t, C_{t-1}, \ldots, C_{t-L}) = C_t h_o(\frac{C_{t-1}}{C_t}, \ldots, \frac{C_{t-L}}{C_t})$ with $h_o(\cdot)$ unknown. It is obvious that one needs to impose $0 \leqslant h_o(\cdot) < 1$ so that $0 \leqslant H_t < C_t$. The following external habit specification is a special case of their model:

$$E\left(\delta_o\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma_o} \frac{\left(1 - h_o\left(\frac{C_t}{C_{t+1}}, \ldots, \frac{C_{t+1-L}}{C_{t+1}}\right)\right)^{-\gamma_o}}{\left(1 - h_o\left(\frac{C_{t-1}}{C_t}, \ldots, \frac{C_{t-L}}{C_t}\right)\right)^{-\gamma_o}} R_{\ell,t+1} - 1\middle|\mathbf{w}_t\right) = 0, \tag{2.7}$$

for $\ell = 1, \ldots, N$, where $\gamma_o > 0$, $\delta_o > 0$ are unknown scalar preference parameters, $h_o(\cdot) \in [0, 1)$ is an unknown function and $H_{t+1} = C_{t+1} h_o(\frac{C_t}{C_{t+1}}, \ldots, \frac{C_{t+1-L}}{C_{t+1}})$ is the habit level at time $t+1$. Chen and Ludvigson (2003) have applied the sieve method to estimate this model and its generalization which allows for internal habit formation of unknown form. Their empirical findings, using quarterly data, are in favor of flexible nonlinear internal habit formation.

Semi-nonparametric conditional moment models. We note that Examples 2.2 and 2.3 and many other economic models imply semi-nonparametric conditional moment restrictions of the form

$$E\left[\rho(Z_t; \theta_o)\middle|X_t\right] = 0, \quad \theta_o \equiv \left(\beta_o', h_o'\right)', \tag{2.8}$$

where $\rho(\cdot;\cdot)$ is a column vector of residual functions whose functional forms are known up to unknown parameters, $\theta \equiv (\beta', h')'$, and $\{Z_t' = (Y_t', X_t')\}_{t=1}^n$ is the data where Y_t is a vector of endogenous variables and X_t is a vector of conditioning variables. Here $E[\rho(Z_t, \theta)|X_t]$ denotes the conditional expectation of $\rho(Z_t, \theta)$ given X_t, and the true conditional distribution of Y_t given X_t is unspecified (and is treated as a nuisance function). The parameters of interest $\theta_o \equiv (\beta_o', h_o')'$ contain a vector of finite-dimensional unknown parameters β_o and a vector of infinite-dimensional unknown functions $h_o(\cdot) = (h_{o1}(\cdot), \ldots, h_{oq}(\cdot))'$, where the arguments of $h_{oj}(\cdot)$ could depend on Y, X, known index function $\delta_j(Z, \beta_o)$ up to unknown β_o, other unknown function $h_{ok}(\cdot)$ for $k \neq j$, or could also depend on unobserved random variables. Motivated by the asset pricing and rational expectations models, Hansen (1982, 1985) studied the conditional moment restriction $E[\rho(Z_t; \beta_o)|X_t] = 0$ (i.e., without unknown h_o) for stationary ergodic time series data (where typically $Z_t' = (Y_t', X_t')$ and X_t includes lagged Y_t and other pre-determined variables known at time t). Chamberlain (1992), Newey and Powell (2003), Ai and Chen (2003) and Chen and Pouzo (2006) studied the general case $E[\rho(Z_t; \beta_o, h_o)|X_t] = 0$ for i.i.d. data.

The semi-nonparametric conditional moment models given by (2.8) can be classified into two broad subclasses. The first subclass consists of *models without endogeneity* in the sense that $\rho(Z_t, \theta) - \rho(Z_t, \theta_o)$ does not depend on any endogenous variables (Y_t); hence the true parameter θ_o can be identified as the unique maximizer of $Q(\theta) = -E[\rho(Z_t, \theta)'\{\Sigma(X_t)\}^{-1}\rho(Z_t, \theta)]$, where $\Sigma(X_t)$ is a positive definite weighting matrix. The second subclass consists of *models with endogeneity* in the sense that $\rho(Z_t, \theta) - \rho(Z_t, \theta_o)$ does depend on endogenous variables (Y_t). Here the true parameter θ_o can be identified as the unique maximizer of

$$Q(\theta) = -E\big[m(X_t, \theta)'\big\{\Sigma(X_t)\big\}^{-1} m(X_t, \theta)\big] \quad \text{with } m(X_t, \theta) \equiv E\big[\rho(Z_t, \theta)\big|X_t\big].$$

Although the second subclass includes the first subclass as a special case, when θ contains unknown functions, it is much easier to derive asymptotic properties for various nonparametric estimators of θ identified by the conditional moment models belonging to the first subclass. The first subclass includes, as special cases, many semi-nonparametric regression models that have been well studied in econometrics. For example, it includes the specifications (2.2) and (2.3) of Example 2.2, the partially linear regression $E[Y_i - X_{1i}'\beta_o - h_o(X_{2i})|X_{1i}, X_{2i}] = 0$ of Engle et al. (1986) and Robinson (1988), the index regression $E[Y_i - h_o(X_i'\beta_o)|X_i] = 0$ of Powell, Stock and Stoker (1989), Ichimura (1993) and Klein and Spady (1993), the varying coefficient model $E[Y_i - \sum_{j=1}^q h_{oj}(D_{ji})X_{ji}|(D_{ki}, X_{ki}), k = 1, \ldots, q] = 0$ of Chen and Tsay (1993), Cai, Fan and Yao (2000) and Chen and Conley (2001), and the additive model with a known link (F) function $E[Y_i - F(\sum_{j=1}^q h_{oj}(X_{ji}))|X_{1i}, \ldots, X_{qi}] = 0$ of Horowitz and Mammen (2004).

The second subclass includes, as special cases, the specification (2.4) of Example 2.2, Example 2.3, semi-nonparametric asset pricing and rational expectation models, and simultaneous equations with flexible parameterization. A leading, yet difficult example of this subclass, is the purely nonparametric instrumental variables (IV) regression

$E[Y_{1i} - h_o(Y_{2i})|X_i] = 0$ studied by Newey and Powell (2003), Darolles, Florens and Renault (2002), Blundell, Chen and Kristensen (2007), Hall and Horowitz (2005) and Carrasco, Florens and Renault (2006). A more difficult example is the nonparametric IV quantile regression $E[1\{Y_{1i} \leqslant h_o(Y_{2i})\} - \gamma|X_i] = 0$ for some known $\gamma \in (0, 1)$ considered by Chernozhukov, Imbens and Newey (2007), Horowitz and Lee (2007) and Chen and Pouzo (2006). See Blundell and Powell (2003), Florens (2003), Newey and Powell (1989), Carrasco, Florens and Renault (2006) and Chen and Pouzo (2006) for additional examples.

2.2. *Definition of sieve extremum estimation*

2.2.1. *Ill-posed versus well-posed problem, sieve extremum estimation*

Let Θ be an infinite-dimensional parameter space endowed with a (pseudo-) metric d. A typical semi-nonparametric econometric model specifies that there is a population criterion function $Q : \Theta \to \mathcal{R}$, which is uniquely maximized at a (pseudo-) true parameter $\theta_o \in \Theta$.[7] The choice of $Q(\cdot)$ and the existence of θ_o are suggested by the identification of an econometric model. The (pseudo-) true parameter $\theta_o \in \Theta$ is unknown but is related to a joint probability measure $P_o(z_1, \ldots, z_n)$, from which a sample of size n observations $\{Z_t\}_{t=1}^n$, $Z_t \in \mathcal{R}^{d_z}$, $1 \leqslant d_z < \infty$, is available. Let $\widehat{Q}_n : \Theta \to \mathcal{R}$ be an empirical criterion, which is a measurable function of the data $\{Z_t\}_{t=1}^n$ for all $\theta \in \Theta$, and converges to Q in some sense (to be more precise in Subsection 3.1) as the sample size $n \to \infty$. One general way to estimate θ_o is by maximizing $\widehat{Q}_n$ over Θ; the maximizer, $\arg\sup_{\theta\in\Theta} \widehat{Q}_n(\theta)$, assuming it exists, is then called the *extremum* estimate. See e.g. Amemiya (1985, Chapter 4), Gallant and White (1988b), Newey and McFadden (1994) and White (1994).

When Θ is infinite-dimensional and possibly not compact with respect to the (pseudo-) metric d,[8] maximizing $\widehat{Q}_n$ over Θ may not be well-defined; or even if a maximizer $\arg\sup_{\theta\in\Theta} \widehat{Q}_n(\theta)$ exists, it is generally difficult to compute, and may have undesirable large sample properties such as inconsistency and/or a very slow rate of convergence. These difficulties arise because the problem of optimization over an infinite-dimensional noncompact space may no longer be well-posed. Throughout this chapter, we say the optimization problem is *well-posed*, if for all sequences $\{\theta_k\}$ in Θ such that $Q(\theta_o) - Q(\theta_k) \to 0$, then $d(\theta_o, \theta_k) \to 0$; is *ill-posed* (or *not well-posed*) if there exists a sequence $\{\theta_k\}$ in Θ such that $Q(\theta_o) - Q(\theta_k) \to 0$ but $d(\theta_o, \theta_k) \nrightarrow 0$.[9] For a given

[7] Although we often call θ_o the "true" parameter in this survey chapter, it in fact could be a pseudo-true parameter value, depending on the specification of the econometrics model and the choice of Q. See Ai and Chen (2007) for estimation of misspecified semi-nonparametric models.

[8] In an infinite-dimensional metric space $(\mathcal{H}, d)$, a compact set is a d-closed and totally bounded set. (A set is totally bounded if for any $\varepsilon > 0$, there exist finitely many open balls with radius ε that cover the set.) A d-closed and bounded set is compact only in a finite-dimensional Euclidean space.

[9] See Carrasco, Florens and Renault (2006) and Vapnik (1998) for surveys on ill-posed inverse problems in linear nonparametric models.

semi-nonparametric model, suppose the criterion $Q(\theta)$ and the space Θ are chosen such that $Q(\theta)$ is uniquely maximized at θ_o in Θ. Then whether the problem is ill-posed or well-posed depends on the choice of the pseudo-metric d. This is because different metrics on an infinite-dimensional space Θ may not be equivalent to each other.[10] In particular, it is likely that some standard norm (say $\|\theta_o - \theta\|_s$) on Θ is not continuous in $Q(\theta_o) - Q(\theta)$ and the problem is ill-posed under $\|\cdot\|_s$, but there is another pseudo-metric (say $\|\theta_o - \theta\|_w$) on Θ that is continuous in $Q(\theta_o) - Q(\theta)$, hence the problem becomes well-posed under this $\|\cdot\|_w$; such a pseudo-metric is typically weaker than $\|\cdot\|_s$ (i.e., $\|\theta_o - \theta\|_s \to 0$ implies $\|\theta_o - \theta\|_w \to 0$). See Ai and Chen (2003, 2007) for more discussions.[11]

No matter whether the semi-nonparametric problems are well-posed or ill-posed, the method of sieves provides one general approach to resolve the difficulties associated with maximizing $\widehat{Q}_n$ over an infinite-dimensional space Θ by maximizing $\widehat{Q}_n$ over a sequence of approximating spaces Θ_n, called *sieves* by Grenander (1981), which are less complex but are dense in Θ. Popular sieves are typically compact, nondecreasing ($\Theta_n \subseteq \Theta_{n+1} \subseteq \cdots \subseteq \Theta$) and are such that for any $\theta \in \Theta$ there exists an element $\pi_n\theta$ in Θ_n satisfying $d(\theta, \pi_n\theta) \to 0$ as $n \to \infty$, where the notation π_n can be regarded as a projection mapping from Θ to Θ_n.

An *approximate sieve extremum estimate*, denoted by $\hat{\theta}_n$, is defined as an approximate maximizer of $\widehat{Q}_n(\theta)$ over the sieve space Θ_n, i.e.,

$$\widehat{Q}_n(\hat{\theta}_n) \geqslant \sup_{\theta \in \Theta_n} \widehat{Q}_n(\theta) - O_P(\eta_n), \quad \text{with } \eta_n \to 0 \text{ as } n \to \infty. \tag{2.9}$$

When $\eta_n = 0$, we call $\hat{\theta}_n$ in (2.9) the *exact* sieve extremum estimate.[12] The sieve extremum estimation method clearly includes the standard extremum estimation method by setting $\Theta_n = \Theta$ for all n.

REMARK 2.1. Following White and Wooldridge (1991, Theorem 2.2), one can show that $\hat{\theta}_n$ in (2.9) is well defined and measurable under the following mild sufficient conditions: (i) $\widehat{Q}_n(\theta)$ is a measurable function of the data $\{Z_t\}_{t=1}^n$ for all $\theta \in \Theta_n$; (ii) for any data $\{Z_t\}_{t=1}^n$, $\widehat{Q}_n(\theta)$ is upper semicontinuous on Θ_n under the metric $d(\cdot,\cdot)$; and (iii) the sieve space Θ_n is compact under the metric $d(\cdot,\cdot)$. Therefore, in the rest of this chapter we assume that $\hat{\theta}_n$ in (2.9) exists and is measurable.

For a semi-nonparametric econometric model, $\theta_o \in \Theta$ can be decomposed into two parts $\theta_o = (\beta_o', h_o')' \in B \times \mathcal{H}$, where B denotes a finite-dimensional compact parameter space, and $\mathcal{H}$ an infinite-dimensional parameter space. In this case, a natural sieve

[10] This is in contrast to the fact that all the norms are equivalent on a finite-dimensional Euclidean space.

[11] The use of a weaker pseudo-metric enables Ai and Chen (2003) to obtain root-n normality of $\hat{\beta}$ for β_o identified via the model $E[\rho(Z_t; \beta_o, h_o)|X_t] = 0$, even when $h_o()$ is a function of the endogenous variable Y and the estimation problem may be ill-posed under the standard mean squared error metric $\sqrt{E[h(Y) - h_o(Y)]^2}$.

[12] Since the complexity of the sieve space Θ_n increases with the sample size, it is obvious that the maximization of $\widehat{Q}_n(\theta)$ over Θ_n need not be exact and the approximate maximizer $\hat{\theta}_n$ in (2.9) will be enough for consistency; see the consistency theorem in Subsection 3.1.

space will be $\Theta_n = B \times \mathcal{H}_n$ with $\mathcal{H}_n$ being a sieve for $\mathcal{H}$, and the resulting estimate $\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n)$ in (2.9) will sometimes be called a simultaneous (or joint) sieve extremum estimate. For a semi-nonparametric model, we can also estimate the parameters of interest (β_o, h_o) by the *approximate profile sieve extremum estimation* that consists of two steps:

Step 1. For an arbitrarily fixed value $\beta \in B$, compute

$$\widehat{Q}_n\big(\beta, \tilde{h}(\beta)\big) \geqslant \sup_{h \in \mathcal{H}_n} \widehat{Q}_n(\beta, h) - O_P(\eta_n)$$

with $\eta_n = o(1)$;

Step 2. Estimate β_o by $\hat{\beta}_n$ solving $\widehat{Q}_n(\hat{\beta}, \tilde{h}(\hat{\beta})) \geqslant \max_{\beta \in B} \widehat{Q}_n(\beta, \tilde{h}(\beta)) - O_P(\eta_n)$, and then estimate h_o by $\hat{h}_n = \tilde{h}(\hat{\beta}_n)$.

Depending on the specific structure of a semi-nonparametric model, the profile sieve extremum estimation procedure may be easier to compute.

2.2.2. *Sieve M-estimation*

When $\widehat{Q}_n(\theta)$ can be expressed as a sample average of the form

$$\sup_{\theta \in \Theta_n} \widehat{Q}_n(\theta) = \sup_{\theta \in \Theta_n} \frac{1}{n} \sum_{t=1}^{n} l(\theta, Z_t),$$

with $l : \Theta \times \mathcal{R}^{d_z} \to \mathcal{R}$ being the criterion based on a single observation, we also call the $\hat{\theta}_n$ solving (2.9) as an approximate *sieve maximum-likelihood-like* (M-) estimate.[13] This includes sieve maximum likelihood estimation (MLE), sieve least squares (LS), sieve generalized least squares (GLS) and sieve quantile regression as special cases.

EXAMPLE 2.1 *(Continued)*. Heckman and Singer (1984) estimated the unknown true parameters $\theta_o = (\beta_o', h_o)' \in \Theta$ in their semiparametric specification, (2.1), of Example 2.1 by the sieve MLE:

$$\sup_{\theta \in \Theta_n} \widehat{Q}_n(\theta) = \sup_{\beta \in B,\, h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \log\left(\int g(T_i | \beta, u, X_i)\, \mathrm{d}h(u) \right),$$

where as $n \to \infty$, the sieve space, $\mathcal{H}_n$, becomes dense in the space of probability distribution functions over $\mathcal{R}$.

[13] Our definition follows that in Newey and McFadden (1994). Some statisticians such as Birgé and Massart (1998) call this a sieve minimum contrast estimate.

EXAMPLE 2.2 *(Continued)*. The nonparametric exogenous expenditure specification (2.2) of Example 2.2 can be estimated by the sieve nonlinear LS:

$$\sup_{\theta\in\Theta_n} \widehat{Q}_n(\theta) = \sup_{h\in\mathcal{H}_n} \frac{-1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{N}\bigl[Y_{1\ell i} - \bigl\{h_{1\ell}\bigl(Y_{2i} - h_0(X_{1i})\bigr) + h_{2\ell}(X_{1i})\bigr\}\bigr]^2,$$

with $\theta = h = (h_0, h_{11}, \ldots, h_{1N}, h_{21}, \ldots, h_{2N})'$ the unknown parameters and $\Theta_n = \mathcal{H}_n = \mathcal{H}_{0,n} \times \prod_{\ell=1}^{N} \mathcal{H}_{1\ell,n} \times \prod_{\ell=1}^{N} \mathcal{H}_{2\ell,n}$ the sieve space,[14] where we impose the identification condition $h_{2\ell}(x_1^*) = 0$ on the sieve space $\mathcal{H}_{2\ell,n}$ for $\ell = 1, \ldots, N$. The semi-nonparametric exogenous expenditure specification (2.3) of Example 2.2 can be also estimated by the sieve nonlinear LS:

$$\sup_{\theta\in\Theta_n} \widehat{Q}_n(\theta) = \sup_{\beta\in B,\, h\in\mathcal{H}_n} \frac{-1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{N}\bigl[Y_{1\ell i} - \bigl\{h_{1\ell}\bigl(Y_{2i} - g\bigl(X_{1i}'\beta_1\bigr)\bigr) + X_{1i}'\beta_{2\ell}\bigr\}\bigr]^2,$$

with $\theta = (\beta', h')' = (\beta_1', \beta_{21}', \ldots, \beta_{2N}', h_{11}, \ldots, h_{1N})'$ the unknown parameters and $\Theta_n = B \times \mathcal{H}_n = B_1 \times \prod_{\ell=1}^{N} B_{2\ell} \times \prod_{\ell=1}^{N} \mathcal{H}_{1\ell,n}$ the sieve space.

More generally, we can apply the sieve GLS criterion

$$\sup_{\theta\in\Theta_n} \widehat{Q}_n(\theta) = \sup_{\theta\in\Theta_n} \frac{-1}{n}\sum_{i=1}^{n}\rho(Z_i,\theta)'\bigl\{\Sigma(X_i)\bigr\}^{-1}\rho(Z_i,\theta)$$

to estimate all the models belonging to the first subclass of the conditional moment restrictions (2.8) where $\rho(Z_i, \theta) - \rho(Z_i, \theta_o)$ does not depend on endogenous variables Y_i, here $\Sigma(X_i)$ is a positive definite weighting matrix function such as the identity matrix. See Remark 4.3 in Subsection 4.3 for optimally weighted version of this procedure.

2.2.3. Series estimation, concave extended linear models

In this chapter, we call a special case of sieve M-estimation *series* estimation, which is sieve M-estimation with *concave* criterion functions $\widehat{Q}_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} l(\theta, Z_t)$ and *finite-dimensional linear* sieve spaces Θ_n. We say the criterion is concave if $\widehat{Q}_n(\tau\theta_1 + (1-\tau)\theta_2) \geqslant \tau\widehat{Q}_n(\theta_1) + (1-\tau)\widehat{Q}_n(\theta_2)$ for any $\theta_1, \theta_2 \in \Theta$ and any scalar $\tau \in (0, 1)$. Of course this definition only makes sense when the parameter space Θ is convex (i.e., for any $\theta_1, \theta_2 \in \Theta$, we have $\tau\theta_1 + (1-\tau)\theta_2 \in \Theta$ for any scalar $\tau \in (0, 1)$). We say a sieve Θ_n is finite-dimensional linear if it is a linear span of finitely many known basis functions; see Subsection 2.3.1 for examples.

Although our definition of series estimation may differ from those in the current econometrics literature, it is closely related to the definition of the sieve M-estimation of "*concave extended linear models*" in the statistics literature; see e.g. Hansen (1994), Stone et al. (1997), and Huang (2001). Consider a $\mathcal{Z}$-valued random variable Z, where

[14] Throughout this chapter $\prod_{\ell=1}^{N} \mathcal{H}_{\ell,n}$ denotes a Cartesian product $\mathcal{H}_{1,n} \times \cdots \times \mathcal{H}_{N,n}$.

$\mathcal{Z}$ is an arbitrary set. The probability density $p_o(z)$ of Z depends on a true but unknown parameter θ_o. All the concave extended linear models have three common ingredients: (1) a (possibly infinite-dimensional) linear parameter space Θ; (2) the criterion evaluated at a single observation is concave; that is, given any $\theta_1, \theta_2 \in \Theta$, $l(\tau\theta_1 + (1-\tau)\theta_2, z) \geqslant \tau l(\theta_1, z) + (1-\tau)l(\theta_2, z)$ for any scalar $\tau \in (0, 1)$ and any value $z \in \mathcal{Z}$; (3) the population criterion $Q(\theta) = E[l(\theta, Z)]$ is strictly concave; that is, given any two essentially different functions $\theta_1, \theta_2 \in \Theta$, $E[l(\tau\theta_1 + (1-\tau)\theta_2, Z)] > \tau E[l(\theta_1, Z)] + (1-\tau)E[l(\theta_2, Z)]$ for any scalar $\tau \in (0, 1)$.

The sieve M-estimation of a concave extended linear model can be implemented by maximizing $\widehat{Q}_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} l(\theta, Z_t)$ over a finite-dimensional linear sieve space Θ_n without any constraints. The resulting estimator is called a series estimator in this paper. Therefore, for the same concave criterion function, a sieve M-estimator is a series estimator if the sieve spaces Θ_n are finite-dimensional linear (such as the ones listed in Subsections 2.3.1 and 2.3.2), but is not a series estimator if the sieve spaces Θ_n are not finite-dimensional linear (such as the ones listed in Subsections 2.3.3 and 2.3.4). Although this definition of a series estimator might look restrictive, it will make the descriptions of large sample properties much easier in Section 3.

For series estimation, concavity of the criterion function plays a central role. In particular, the sieve spaces used in estimation are not required to be compact and can be any unrestricted finite-dimensional linear spaces. Such sieves not only make it easy to compute the estimators, but also make it convenient to discuss orthogonal projections and functional analysis of variance (ANOVA) decompositions (such as additivity) in the nonparametric multivariate regression framework; see e.g. Stone (1985, 1986), Andrews and Whang (1990), Huang (1998a).

In order to apply the series estimation to a semi-nonparametric model, one needs to first find a concave criterion function that identifies the unknown parameters of interest. We now present several such examples.

EXAMPLE 2.4 *(Multivariate LS regression).* We consider the estimation of an unknown multivariate conditional mean function $\theta_o(\cdot) = h_o(\cdot) = E(Y|X = \cdot)$. Here $Z = (Y, X)$, Y is a scalar, X has support $\mathcal{X}$ that is a bounded subset of $\mathcal{R}^d$, $d \geqslant 1$. Suppose $h_o \in \Theta$, where Θ is a linear subspace of the space of functions h with $E[h(X)^2] < \infty$. Let $l(h, Z) = -[Y - h(X)]^2$ and $Q(\theta) = -E\{[Y - h(X)]^2\}$; then both are concave in h and Q is strictly concave in $h \in \Theta$.

Let $\{p_j(X), j = 1, 2, \ldots\}$ denote a sequence of known basis functions that can approximate any real-valued square integrable functions of X well; see Subsection 2.3.1 or Newey (1997) for specific examples of such basis functions. Then

$$\Theta_n = \mathcal{H}_n = \left\{h : \mathcal{X} \to \mathcal{R},\ h(x) = \sum_{j=1}^{k_n} a_j p_j(x) \colon a_1, \ldots, a_{k_n} \in \mathcal{R}\right\}, \tag{2.10}$$

with $\dim(\Theta_n) = k_n \to \infty$ slowly as $n \to \infty$, is a finite-dimensional linear sieve for Θ, and $\hat{h} = \arg\max_{h \in \mathcal{H}_n} \frac{-1}{n}\sum_{t=1}^{n}[Y_t - h(X_t)]^2$ is a series estimator of the conditional mean $h_o(\cdot) = E(Y|X = \cdot)$. Moreover, this series estimator $\hat{h}$ has a simple closed-form expression:

$$\hat{h}(x) = p^{k_n}(x)'(P'P)^{-} \sum_{i=1}^{n} p^{k_n}(X_i)Y_i, \quad x \in \mathcal{X}, \tag{2.11}$$

with $p^{k_n}(X) = (p_1(X), \ldots, p_{k_n}(X))'$, $P = (p^{k_n}(X_1), \ldots, p^{k_n}(X_n))'$ and $(P'P)^{-}$ the Moore–Penrose generalized inverse. The estimator $\hat{h}$ given in (2.11) will be called a series LS estimator or a linear sieve LS estimator.

EXAMPLE 2.5 *(Multivariate quantile regression).* Let $\alpha \in (0, 1)$. We consider the estimation of an unknown multivariate αth quantile function $\theta_o(\cdot) = h_o(\cdot)$ such that $E[1\{Y \leqslant h_o(X)\}|X] = \alpha$. Here $Z = (Y, X)$, X has support $\mathcal{X}$ that is a bounded subset of $\mathcal{R}^d$, $d \geqslant 1$. Suppose $h_o \in \Theta$, where Θ is a linear subspace of the space of functions h with $E[h(X)^2] < \infty$. Let $l(h, Z) = [1\{Y \leqslant h(X)\} - \alpha][Y - h(X)]$,[15] and $Q(\theta) = E\{[1\{Y \leqslant h(X)\} - \alpha][Y - h(X)]\}$, then both are concave in h and Q is strictly concave in $h \in \Theta$.

Let $\Theta_n = \mathcal{H}_n$ be a finite-dimensional linear sieve such as the one given in (2.10). Then $\hat{h} = \arg\max_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{t=1}^{n} [1\{Y_t \leqslant h(X_t)\} - \alpha][Y_t - h(X_t)]$ is a series estimator of the conditional quantile function h_o.

EXAMPLE 2.6 *(Log-density estimation).* Let f_o be the true unknown positive probability density of Z on $\mathcal{Z}$ and suppose that we want to estimate the log-density, $\log f_o$. Since $\log f_o$ is subject to the nonlinear constraint $\int_{\mathcal{Z}} \exp\{\log f_o(z)\}\,dz = 1$, it is more convenient to write $\log f_o = h_o - \log \int_{\mathcal{Z}} \exp h_o(z)\,dz$, and treat h_o as an unknown function in some linear space. Since $\log f_o = [h_o + c] - \log \int_{\mathcal{Z}} \exp[h_o(z) + c]\,dz$ for any constant c, we need some location normalization to ensure the identification of h_o. By imposing a linear constraint such as $\int_{\mathcal{Z}} h(z)\,dz = 0$ (or $h(z^*) = 0$ for a fixed $z^* \in \mathcal{Z}$), we can determine h uniquely and make the mapping $h \mapsto \log f$ one-to-one. Therefore, we assume $h_o \in \Theta$, where Θ is a linear subspace of the space of real-valued functions h with $E[h(Z)^2] < \infty$ and $\int_{\mathcal{Z}} h(z)\,dz = 0$. The log-likelihood evaluated at a single observation Z is given by $l(h, Z) = h(Z) - \log \int_{\mathcal{Z}} \exp h(z)\,dz$. Stone (1990) has shown that $l(h, Z)$ is concave and $Q(\theta) = E\{h(Z) - \log \int_{\mathcal{Z}} \exp h(z)\,dz\}$ is strictly concave in $h \in \Theta$.

Let $\{p_j(Z), j = 1, 2, \ldots\}$ denote a sequence of known basis functions that can approximate any real-valued square integrable functions of Z well. Then

$$\Theta_n = \mathcal{H}_n = \left\{h : \mathcal{Z} \to \mathcal{R},\ h(z) = \sum_{j=1}^{k_n} a_j p_j(z) \colon \int_{\mathcal{Z}} h(z)\,dz = 0,\ a_1, \ldots, a_{k_n} \in \mathcal{R}\right\},$$

[15] This is a "check" function in Koenker and Bassett (1978).

with $\dim(\Theta_n) = k_n \to \infty$ slowly as $n \to \infty$, is a finite-dimensional linear sieve for Θ, and

$$\hat{h} = \arg\max_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \left[h(Z_i) - \log \int_{\mathcal{Z}} \exp h(z)\, \mathrm{d}z \right]$$

is a series estimator of the log-density function h_o.

It is easy to see that log-conditional density and log-spectral density estimation can be carried out in the same way; see e.g. Stone (1994) and Kooperberg, Stone and Truong (1995b).

EXAMPLE 2.7 *(Estimation of conditional hazard function).* Consider a positive survival time T, a positive censoring time C, the observed time $Y = \min(T, C)$ and an $\mathcal{X}$-valued random vector X of covariates. Let $Z = (X', Y, 1(T \leqslant C))'$ denote a single observation. Suppose T and C are conditionally independent given X, and that $\Pr(C \leqslant \tau_0) = 1$ for a known positive constant τ_0. Let $f_o(\tau|x)$ and $F_o(\tau|x)$, $\tau > 0$, be the true unknown conditional density function and conditional distribution function, respectively, of T given $X = x$. Then the ratio $f_o(\tau|x)/[1 - F_o(\tau|x)]$, $\tau > 0$, is called the conditional hazard function of T given $X = x$. We want to estimate the log-conditional hazard function $h_o(\tau, x) = \log\{f_o(\tau|x)/[1 - F_o(\tau|x)]\}$. Since the likelihood at a single observation Z equals

$$\begin{aligned}
&\big[f(Y|X)\big]^{1(T \leqslant C)} \big[1 - F(Y|X)\big]^{1(T > C)} \\
&\quad = \big[\exp\big\{h(Y, X)\big\}\big]^{1(T \leqslant C)} \exp\left(-\int_0^Y \exp\big\{h(\tau, X)\big\}\, \mathrm{d}\tau\right),
\end{aligned}$$

the log-likelihood evaluated at a single observation is given by

$$l(h, Z) = 1(T \leqslant C) h(Y, X) - \int_0^Y \exp\big\{h(\tau, X)\big\}\, \mathrm{d}\tau.$$

Kooperberg, Stone and Truong (1995a) showed that the $l(h, Z)$ is concave in h and $Q(\theta) = E\{l(h, Z)\}$ is strictly concave in h.

Suppose $h_o \in \Theta$, where Θ is a linear subspace of the space of real-valued functions h with $E[h(Y, X)^2] < \infty$. Let $\{p_j(Y, X),\ j = 1, 2, \ldots\}$ denote a sequence of known basis functions that can approximate any real-valued square integrable functions of (Y, X) well. Then

$$\begin{aligned}
\Theta_n &= \mathcal{H}_n \\
&= \left\{ h : (0, \tau_0] \times \mathcal{X} \to \mathcal{R},\ h(\tau, x) = \sum_{j=1}^{k_n} a_j p_j(\tau, x) \colon a_1, \ldots, a_{k_n} \in \mathcal{R} \right\},
\end{aligned}$$

with $\dim(\Theta_n) = k_n \to \infty$ slowly as $n \to \infty$, is a finite-dimensional linear sieve for Θ, and

$$\hat{h} = \arg\max_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \left[1(T_i \leqslant C_i) h(Y_i, X_i) - \int_0^{Y_i} \exp\left\{h(\tau, X_i)\right\} \mathrm{d}\tau \right]$$

is a series estimator of the log-conditional hazard function h_o.

Finally, we should point out that not all semi-nonparametric M-estimation problems can be reparameterized into series estimation problems. For example, the nonparametric exogenous expenditure specification (2.2) of Example 2.2 does not belong to the concave extended linear models, since, in this specification, the unknown function $h_0(X_1)$ enters the other unknown functions $h_{1\ell}(Y_2 - h_0(X_1))$, $\ell = 1, \ldots, L$, nonlinearly as an argument. Nevertheless, as described in the previous subsection, this model can still be estimated by the general sieve M-estimation method.

2.2.4. Sieve MD estimation

When $-\widehat{Q}_n(\theta)$ can be expressed as a quadratic distance from zero, we call the $\hat{\theta}_n$ solving (2.9) an approximate *sieve minimum distance* (MD) estimate.

One typical quadratic form is

$$\sup_{\theta \in \Theta_n} \widehat{Q}_n(\theta) = \sup_{\theta \in \Theta_n} -\frac{1}{n} \sum_{t=1}^{n} \hat{m}(X_t, \theta)' \left\{\widehat{\Sigma}(X_t)\right\}^{-1} \hat{m}(X_t, \theta) \tag{2.12}$$

with $\hat{m}(X_t, \theta_o) \to 0$ in probability. Here $\hat{m}(X_t, \theta)$ is a nonparametrically estimated moment restriction function of fixed, finite dimension, and $\widehat{\Sigma}(X_t)$ is a possibly nonparametrically estimated weighting matrix of the same dimension as that of $\hat{m}(X_t, \theta)$. The weighting matrix, $\widehat{\Sigma}$, is introduced for the purpose of efficiency,[16] and $\widehat{\Sigma}(X_t) \to \Sigma(X_t)$ in probability, where $\Sigma(X_t)$ is a positive definite matrix (of the same fixed, finite dimension as that of $\widehat{\Sigma}(X_t)$). We can apply the sieve MD criterion, (2.12), to estimate all the models belonging to the conditional moment restrictions $E[\rho(Z, \theta_o)|X] = 0$, regardless of whether or not $\rho(Z_t, \theta) - \rho(Z_t, \theta_o)$ depends on endogenous variables Y_t. In particular, $\hat{m}(X_t, \theta)$ could be any nonparametric estimate of the conditional mean function $m(X_t, \theta) = E[\rho(Z, \theta)|X = X_t]$; see e.g. Newey and Powell (1989, 2003) and Ai and Chen (1999, 2003).

Another typical quadratic form is the sieve GMM criterion

$$\sup_{\theta \in \Theta_n} \widehat{Q}_n(\theta) = \sup_{\theta \in \Theta_n} -\hat{g}_n(\theta)' \widehat{W} \hat{g}_n(\theta) \tag{2.13}$$

[16] See Ai and Chen (2003) or Subsection 4.3 for details on semiparametric efficiency.

with $\hat{g}_n(\theta_o) \to 0$ in probability. Here $\hat{g}_n(\theta)$ is a sample average of some unconditional moment conditions of increasing dimension, and $\widehat{W}$ is a possibly random weighting matrix of the same increasing dimension as that of $\hat{g}_n(\theta)$. As above, the weighting matrix $\widehat{W}$ is introduced for the purpose of efficiency, and $\widehat{W} - W_n \to 0$ in probability, with W_n being a positive definite matrix (of the same increasing dimension as that of $\widehat{W}$). Note that $E[\rho(Z, \theta_o)|X] = 0$ if and only if the following increasing number of unconditional moment restrictions hold:

$$E\big[\rho(Z_t, \theta_o) p_{0j}(X_t)\big] = 0, \quad j = 1, 2, \ldots, k_{m,n}, \tag{2.14}$$

where $\{p_{0j}(X),\ j = 1, 2, \ldots, k_{m,n}\}$ is a sequence of known basis functions that can approximate any real-valued square integrable functions of X well as $k_{m,n} \to \infty$. Let $p^{k_{m,n}}(X) = (p_{01}(X), \ldots, p_{0k_{m,n}}(X))'$. It is now obvious that the conditional moment restrictions (2.8) $E[\rho(Z, \theta_o)|X] = 0$ can be estimated via the sieve GMM criterion (2.13) using $\hat{g}_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} \rho(Z_t, \theta) \otimes p^{k_{m,n}}(X_t)$.

Not only it is possible for both the sieve MD, (2.12), and the sieve GMM, (2.13), to estimate all the models belonging to the conditional moment restrictions (2.8), but they are also very closely related. For example, when applying the sieve MD (2.12) procedure, we could use the series LS estimator (2.15) as an estimator of the conditional mean function $m(X, \theta) = E[\rho(Z, \theta)|X]$:

$$\hat{m}(X, \theta) = \sum_{j=1}^{n} \rho(Z_j, \theta) p^{k_{m,n}}(X_j)'(P'P)^{-} p^{k_{m,n}}(X), \tag{2.15}$$

with $P = (p^{k_{m,n}}(X_1), \ldots, p^{k_{m,n}}(X_n))'$ where $k_{m,n} \to \infty$ slowly as $n \to \infty$, and $(P'P)^{-}$ the Moore–Penrose inverse. The resulting sieve MD (2.12) with identity weighting $\widehat{\Sigma}(X_t) = I$ will become the following sieve GMM (2.13):

$$\min_{\theta\in\Theta_n}\left(\sum_{i=1}^{n} \rho(Z_i, \theta) \otimes p^{k_{m,n}}(X_i)\right)'\big(I \otimes (P'P)^{-}\big)\left(\sum_{i=1}^{n} \rho(Z_i, \theta) \otimes p^{k_{m,n}}(X_i)\right), \tag{2.16}$$

where $\otimes$ denotes the Kronecker product; see Ai and Chen (2003) for details.

EXAMPLE 2.2 *(Continued)*. The semi-nonparametric endogenous expenditure specification (2.4) of Example 2.2 can be estimated by the sieve MD (2.12), with $\hat{m}(X_i, \theta) = (\hat{m}_1(X_i, \theta), \ldots, \hat{m}_N(X_i, \theta))'$,

$$\begin{aligned}\hat{m}_\ell(X_i, \theta) \\ = \sum_{j=1}^{n}\big[Y_{1\ell j} - \big\{h_{1\ell}\big(Y_{2j} - g\big(X_{1j}'\beta_1\big)\big) + X_{1j}'\beta_{2\ell}\big\}\big] p^{k_{m,n}}(X_j)'(P'P)^{-} p^{k_{m,n}}(X_i),\end{aligned}$$

where $\theta = (\beta', h')' = (\beta_1', \beta_{21}', \ldots, \beta_{2N}', h_{11}, \ldots, h_{1N})'$ is the vector of unknown parameters, and $\Theta_n = B \times \mathcal{H}_n = B_1 \times \prod_{\ell=1}^{N} B_{2\ell} \times \prod_{\ell=1}^{N} \mathcal{H}_{1\ell,n}$ is the sieve space; see Blundell, Chen and Kristensen (2007) for details.

EXAMPLE 2.3 *(Continued)*. The semi-nonparametric external habit specification (2.7) of Example 2.3 can be estimated by the sieve GMM criterion (2.16), with $\rho(Z_t,\theta) = (\rho_1(Z_t,\theta),\dots,\rho_N(Z_t,\theta))'$,

$$\rho_\ell(Z_t,\theta) = \delta\left(\frac{C_t}{C_{t+1}}\right)^{\gamma} \frac{\left(1 - h\left(\frac{C_t}{C_{t+1}},\dots,\frac{C_{t+1-L}}{C_{t+1}}\right)\right)^{-\gamma}}{\left(1 - h\left(\frac{C_{t-1}}{C_t},\dots,\frac{C_{t-L}}{C_t}\right)\right)^{-\gamma}} R_{\ell,t+1} - 1,$$
$$\ell = 1,\dots,N,$$
$$Z_t = \left(\frac{C_t}{C_{t+1}},\dots,\frac{C_{t+1-L}}{C_{t+1}},\frac{C_{t-1}}{C_t},\dots,\frac{C_{t-L}}{C_t}, R_{1,t+1},\dots,R_{N,t+1}, X_t\right),$$
$$X_t = \mathbf{w}_t,$$

where $\theta = (\beta', h)' = (\delta,\gamma,h)'$ is the vector of unknown parameters, and $\Theta_n = B \times \mathcal{H}_n = B_\delta \times B_\gamma \times \mathcal{H}_n$ is the sieve space, here $0 \leqslant h < 1$ is imposed on the sieve space $\mathcal{H}_n$. Obviously, this model (2.7) can also be estimated by the sieve MD (2.12), with $\hat{m}(X_t,\theta) = \hat{m}(\mathbf{w}_t,\theta)$ being a nonparametric estimator such as the series LS estimator (2.15) of $E[\rho(Z_t,\theta)|X_t = \mathbf{w}_t]$; see Chen and Ludvigson (2003) for details.[17]

2.3. Typical function spaces and sieve spaces

Here we will present some commonly used sieves whose approximation properties are already known in the mathematical literature on approximation theory.

2.3.1. Typical smoothness classes and (finite-dimensional) linear sieves

We first review the most popular smoothness classes of functions used in the nonparametric estimation literature; see e.g. Stone (1982, 1994), Robinson (1988), Newey (1997) and Horowitz (1998). Suppose for the moment that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ is the Cartesian product of compact intervals $\mathcal{X}_1,\dots,\mathcal{X}_d$. Let $0 < \gamma \leqslant 1$. A real-valued function h on $\mathcal{X}$ is said to satisfy a Hölder condition with exponent γ if there is a positive number c such that $|h(x)-h(y)| \leqslant c|x-y|_e^\gamma$ for all $x, y \in \mathcal{X}$; here $|x|_e = (\sum_{l=1}^d x_l^2)^{1/2}$ is the Euclidean norm of $x = (x_1,\dots,x_d) \in \mathcal{X}$. Given a d-tuple $\alpha = (\alpha_1,\dots,\alpha_d)$ of nonnegative integers, set $[\alpha] = \alpha_1 + \cdots + \alpha_d$ and let D^α denote the differential operator defined by

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1}\dots\partial x_d^{\alpha_d}}.$$

[17] There are also semi-nonparametric recursive method of moment procedures that enable us to estimate nonlinear time series models with latent variables. See e.g. Chen and White (1998, 2002), Pastorello, Patilea and Renault (2003) and Linton and Mammen (2005).

Let m be a nonnegative integer and set $p = m + \gamma$. A real-valued function h on $\mathcal{X}$ is said to be *p-smooth* if it is m times continuously differentiable on $\mathcal{X}$ and $D^\alpha h$ satisfies a Hölder condition with exponent γ for all α with $[\alpha] = m$.

Denote the class of all p-smooth real-valued functions on $\mathcal{X}$ by $\Lambda^p(\mathcal{X})$ (called a Hölder class), and the space of all m-times continuously differentiable real-valued functions on $\mathcal{X}$ by $C^m(\mathcal{X})$. Define a Hölder ball with smoothness $p = m + \gamma$ as

$$\Lambda_c^p(\mathcal{X}) = \left\{ h \in C^m(\mathcal{X}) \colon \sup_{[\alpha] \leqslant m} \sup_{x \in \mathcal{X}} \left| D^\alpha h(x) \right| \leqslant c, \right.$$
$$\left. \sup_{[\alpha]=m} \sup_{x, y \in \mathcal{X}, x \neq y} \frac{|D^\alpha h(x) - D^\alpha h(y)|}{|x - y|_e^\gamma} \leqslant c \right\}.$$

The Hölder (or p-smooth) class of functions are popular in econometrics because a *p-smooth* function can be approximated well by various linear sieves.

A sieve is called a "(finite-dimensional) linear sieve" if it is a linear span of finitely many known basis functions. Linear sieves, including power series, Fourier series, splines and wavelets, form a large class of sieves useful for sieve extremum estimation. We now provide some examples of commonly used linear sieves for univariate functions with support $\mathcal{X} = [0, 1]$.

Polynomials. Let $\mathrm{Pol}(J_n)$ denote the space of polynomials on [0, 1] of degree J_n or less; that is,

$$\mathrm{Pol}(J_n) = \left\{ \sum_{k=0}^{J_n} a_k x^k,\ x \in [0, 1] \colon a_k \in \mathcal{R} \right\}.$$

Trigonometric polynomials. Let $\mathrm{TriPol}(J_n)$ denote the space of trigonometric polynomials on [0, 1] of degree J_n or less; that is,

$$\mathrm{TriPol}(J_n) = \left\{ a_0 + \sum_{k=1}^{J_n} \left[a_k \cos(2k\pi x) + b_k \sin(2k\pi x) \right],\ x \in [0, 1] \colon a_k, b_k \in \mathcal{R} \right\}.$$

Let $\mathrm{CosPol}(J_n)$ denote the space of cosine polynomials on [0, 1] of degree J_n or less; that is,

$$\mathrm{CosPol}(J_n) = \left\{ a_0 + \sum_{k=1}^{J_n} a_k \cos(k\pi x),\ x \in [0, 1] \colon a_k \in \mathcal{R} \right\}.$$

Let $\mathrm{SinPol}(J_n)$ denote the space of sine polynomials on [0, 1] of degree J_n or less; that is,

$$\mathrm{SinPol}(J_n) = \left\{ \sum_{k=1}^{J_n} a_k \sin(k\pi x),\ x \in [0, 1] \colon a_k \in \mathcal{R} \right\}.$$

We note that the classical trigonometric sieve, $\mathrm{TriPol}(J_n)$, is well suited for approximating periodic functions on [0, 1], while the cosine sieve, $\mathrm{CosPol}(J_n)$, is well suited for approximating aperiodic functions on [0, 1] and the sine sieve, $\mathrm{SinPol}(J_n)$, can approximate functions vanishing at the boundary points (i.e., when $h(0) = h(1) = 0$).

Univariate splines. Let J_n be a positive integer, and let $t_0, t_1, \ldots, t_{J_n}, t_{J_n+1}$ be real numbers with $0 = t_0 < t_1 < \cdots < t_{J_n} < t_{J_n+1} = 1$. Partition [0, 1] into $J_n + 1$ subintervals $I_j = [t_j, t_{j+1})$, $j = 0, \ldots, J_n - 1$, and $I_{J_n} = [t_{J_n}, t_{J_n+1}]$. We assume that the knots $t_1, \ldots, t_{J_n}$ have bounded mesh ratio:

$$\frac{\max_{0\leqslant j\leqslant J_n}(t_{j+1} - t_j)}{\min_{0\leqslant j\leqslant J_n}(t_{j+1} - t_j)} \leqslant c \quad \text{for some constant } c > 0. \tag{2.17}$$

Let $r \geqslant 1$ be an integer. A function on [0, 1] is a *spline of order* r, equivalently, *of degree* $m \equiv r - 1$, *with knots* $t_1, \ldots, t_{J_n}$ if the following hold: (i) it is a polynomial of degree m or less on each interval I_j, $j = 0, \ldots, J_n$; and (ii) (for $m \geqslant 1$) it is $(m - 1)$-times continuously differentiable on [0, 1]. Such spline functions constitute a linear space of dimension $J_n + r$. For detailed discussions of univariate splines; see de Boor (1978) and Schumaker (1981). For a fixed integer $r \geqslant 1$, we let $\mathrm{Spl}(r, J_n)$ denote the space of splines of order r (or of degree $m \equiv r - 1$) with J_n knots satisfying (2.17). Since

$$\mathrm{Spl}(r, J_n) = \left\{\sum_{k=0}^{r-1} a_k x^k + \sum_{j=1}^{J_n} b_j\bigl[\max\{x - t_j, 0\}\bigr]^{r-1},\ x \in [0, 1]\colon a_k, b_j \in \mathcal{R}\right\},$$

we also call $\mathrm{Spl}(r, J_n)$ the polynomial spline sieve of degree $m \equiv r - 1$.

In this chapter, $L_2(\mathcal{X}, leb)$ denotes the space of real-valued functions h such that $\int_{\mathcal{X}} |h(x)|^2\,\mathrm{d}x < \infty$.

Wavelets. Let $m \geqslant 0$ be an integer. A real-valued function ψ is called a "mother wavelet" of degree m if it satisfies the following: (i) $\int_{\mathcal{R}} x^k\psi(x)\,\mathrm{d}x = 0$ for $0 \leqslant k \leqslant m$; (ii) ψ and all its derivatives up to order m decrease rapidly as $|x| \to \infty$; (iii) $\{2^{j/2}\psi(2^j x - k)\colon\ j, k \in \mathbb{Z}\}$ forms a Riesz basis of $L_2(\mathcal{R}, leb)$, in the sense that the linear span of $\{2^{j/2}\psi(2^j x - k)\colon\ j, k \in \mathbb{Z}\}$ is dense in $L_2(\mathcal{R}, leb)$ and there exist positive constants $c_1 \leqslant c_2 < \infty$ such that

$$\begin{aligned} c_1 \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} |a_{jk}|^2 &\leqslant \left\| \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} a_{jk}2^{j/2}\psi\bigl(2^j x - k\bigr)\right\|^2_{L_2(\mathcal{R},leb)} \\ &\leqslant c_2 \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} |a_{jk}|^2 \end{aligned}$$

for all doubly bi-infinite square-summable sequences $\{a_{jk}\colon\ j, k \in \mathbb{Z}\}$.

A scaling function ϕ is called a "father wavelet" of degree m if it satisfies the following: (i) $\int_{\mathcal{R}} \phi(x)\,\mathrm{d}x = 1$; (ii) ϕ and all its derivatives up to order m decrease rapidly as $|x| \to \infty$; (iii) $\{\phi(x-k)\colon k \in \mathbb{Z}\}$ forms a Riesz basis for a closed subspace of $L_2(\mathcal{R}, leb)$.

Orthogonal wavelets. Given an integer $m \geqslant 0$, there exist a father wavelet ϕ of degree m and a mother wavelet ψ of degree m, both compactly supported, such that for any integer $j_0 \geqslant 0$, any function g in $L_2(\mathcal{R}, leb)$ has the following wavelet m-regular multiresolution expansion:

$$g(x) = \sum_{k=-\infty}^{\infty} a_{j_0 k}\phi_{j_0 k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=-\infty}^{\infty} b_{jk}\psi_{jk}(x), \quad x \in \mathcal{R},$$

where

$$a_{jk} = \int_{\mathcal{R}} g(x)\phi_{jk}(x)\,\mathrm{d}x, \quad \phi_{jk}(x) = 2^{j/2}\phi\big(2^j x - k\big),\ x \in \mathcal{R},$$
$$b_{jk} = \int_{\mathcal{R}} g(x)\psi_{jk}(x)\,\mathrm{d}x, \quad \psi_{jk}(x) = 2^{j/2}\psi\big(2^j x - k\big),\ x \in \mathcal{R},$$

and $\{\phi_{j_0 k},\ k \in \mathbb{Z};\ \psi_{jk},\ j \geqslant j_0,\ k \in \mathbb{Z}\}$ is an orthonormal[18] basis of $L_2(\mathcal{R}, leb)$; see Meyer (1992, Theorem 3.3).

For $j \geqslant 0$ and $0 \leqslant k \leqslant 2^j - 1$, denote the periodized wavelets on $[0, 1]$ by

$$\phi^*_{jk}(x) = 2^{j/2}\sum_{l\in\mathbb{Z}}\phi\big(2^j x + 2^j l - k\big),$$
$$\psi^*_{jk}(x) = 2^{j/2}\sum_{l\in\mathbb{Z}}\psi\big(2^j x + 2^j l - k\big), \quad x \in [0, 1].$$

For $j_0 \geqslant 0$, the collection $\{\phi^*_{j_0 k},\ k = 0, \ldots, 2^{j_0} - 1;\ \psi^*_{jk},\ j \geqslant j_0,\ k = 0, \ldots, 2^j - 1\}$ is an orthonormal basis of $L_2([0, 1], leb)$ [see Daubechies (1992)]. We consider the finite-dimensional linear space spanned by this wavelet basis. For an integer $J_n > j_0$, set

$$\mathrm{Wav}\big(m, 2^{J_n}\big) = \Bigg\{\sum_{k=0}^{2^{j_0}-1}\alpha_{j_0 k}\phi^*_{j_0 k}(x) + \sum_{j=j_0}^{J_n-1}\sum_{k=0}^{2^j-1}\beta_{jk}\psi^*_{jk}(x),$$
$$x \in [0, 1]\colon \alpha_{j_0 k}, \beta_{jk} \in \mathcal{R}\Bigg\}$$

or, equivalently [see Meyer (1992)],

$$\mathrm{Wav}\big(m, 2^{J_n}\big) = \Bigg\{\sum_{k=0}^{2^{J_n}-1}\alpha_k\phi^*_{J_n k}(x),\ x \in [0, 1]\colon \alpha_k \in \mathcal{R}\Bigg\}.$$

[18] I.e., $\int_{\mathcal{R}}\psi_{jk}(x)\psi_{jk}(x)\,\mathrm{d}x = 1$ and $\int_{\mathcal{R}}\psi_{jk}(x)\psi_{j'k'}(x)\,\mathrm{d}x = 0$ for $j \neq j'$ or $k \neq k'$; also $\int_{\mathcal{R}}\phi_{j_0 k}(x)\phi_{j_0 k}(x)\,\mathrm{d}x = 1$ and $\int_{\mathcal{R}}\phi_{j_0 k}(x)\phi_{j_0 k'}(x)\,\mathrm{d}x = 0$ for $k \neq k'$; in addition $\int_{\mathcal{R}}\phi_{j_0 k}(x)\psi_{jk'}(x)\,\mathrm{d}x = 0$ for $j \geqslant j_0$.

Tensor product spaces. Let $\mathcal{U}_\ell$, $1 \leqslant \ell \leqslant d$, be compact sets in Euclidean spaces and $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_d$ be their Cartesian product. Let $\mathbb{G}_\ell$ be a linear space of functions on $\mathcal{U}_\ell$ for $1 \leqslant \ell \leqslant d$, each of which can be any of the sieve spaces described above, among others. The tensor product, $\mathbb{G}$, of $\mathbb{G}_1, \ldots, \mathbb{G}_d$ is defined as the space of functions on $\mathcal{U}$ spanned by the functions $\prod_{\ell=1}^d g_\ell(x_\ell)$, where $g_\ell \in \mathbb{G}_\ell$ for $1 \leqslant \ell \leqslant d$. We note that $\dim(\mathbb{G}) = \prod_{\ell=1}^d \dim(\mathbb{G}_\ell)$. Tensor-product construction is a standard way to generate linear sieves of multivariate functions from linear sieves of univariate functions.

Linear sieves are attractive because of their simplicity and ease of implementation. Moreover, linear sieves can approximate functions in a Hölder space, $\Lambda^p(\mathcal{X})$, well. In the following we let θ denote a real-valued function with a bounded domain $\mathcal{X} \subset \mathcal{R}^d$, $\|\theta\|_\infty \equiv \sup_{x\in\mathcal{X}} |\theta(x)|$ denote its L_∞ norm, and $\|\theta\|_{2,leb} \equiv \{\int_{\mathcal{X}} [\theta(x)]^2\,\mathrm{d}x/\operatorname{vol}(\mathcal{X})\}^{1/2}$ be the scaled L_2 norm relative to the Lebesgue measure of $\mathcal{X}$. Define the sieve approximation errors to $\theta_o \in \Lambda^p(\mathcal{X})$ in $L_\infty(\mathcal{X}, leb)$-norm and $L_2(\mathcal{X}, leb)$-norm as

$$\rho_{\infty n} \equiv \inf_{g\in\Theta_n} \|g - \theta_o\|_\infty \quad \text{and} \quad \rho_{2n} \equiv \inf_{g\in\Theta_n} \|g - \theta_o\|_{2,leb}.$$

It is obvious that $\rho_{2n} \leqslant \rho_{\infty n}$. For a multivariate function $\theta_o \in \Theta = \Lambda^p([0,1]^d)$, we consider the tensor product linear sieve space Θ_n, which is constructed as a tensor product space of some commonly used univariate linear approximating spaces $\Theta_{n1}, \ldots, \Theta_{nd}$. Let $\dim(\Theta_n) = k_n$ and $[p]$ be the biggest integer satisfying $[p] < p$. Then we have the following tensor product sieve approximation error rates for $\theta_o \in \Lambda^p([0,1]^d)$:

Polynomials. If each $\Theta_{n\ell} = \mathrm{Pol}(J_n)$, then $\rho_{\infty n} = O(J_n^{-p}) = O(k_n^{-p/d})$ [see e.g. Section 5.3.2 of Timan (1963)].

Trigonometric polynomials. If θ_o can be extended to a periodic function, and if each $\Theta_{n\ell} = \mathrm{TriPol}(J_n)$, then $\rho_{\infty n} = O(J_n^{-p}) = O(k_n^{-p/d})$ [see e.g. Section 5.3.1 of Timan (1963)].

Splines. If each $\Theta_{n\ell} = \mathrm{Spl}(r, J_n)$ with $r \geqslant [p]+1$, then $\rho_{\infty n} = O(J_n^{-p}) = O(k_n^{-p/d})$ [see (13.69) and Theorem 12.8 of Schumaker (1981)].

Orthogonal wavelets. If each $\Theta_{n\ell} = \mathrm{Wav}(m, 2^{J_n})$ with $m > p$, then $\rho_{\infty n} = O(2^{-pJ_n}) = O(k_n^{-p/d})$ [see Proposition 2.5 of Meyer (1992)].

2.3.2. *Weighted smoothness classes and (finite-dimensional) linear sieves*

In semi-nonparametric econometric applications, sometimes the parameters of interest are functions with unbounded supports. Here we present two finite-dimensional linear sieves that can approximate functions with unbounded supports well. In the following we let $L_p(\mathcal{X}, \omega)$, $1 \leqslant p < \infty$, denote the space of real-valued functions h such that $\int_{\mathcal{X}} |h(x)|^p \omega(x)\,\mathrm{d}x < \infty$ for a smooth weight function $\omega : \mathcal{X} \mapsto (0, \infty)$.

Hermite polynomials. Hermite polynomial series $\{H_k\colon k = 1, 2, \ldots\}$ is an orthonormal basis of $L_2(\mathcal{R}, \omega)$ with $\omega(x) = \exp\{-x^2\}$. It can be obtained by applying the Gram–Schmidt procedure to the polynomial series $\{x^{k-1}\colon k = 1, 2, \ldots\}$ under the inner product $\langle f, g\rangle_\omega = \int_{\mathcal{R}} f(x)g(x)\exp\{-x^2\}\,\mathrm{d}x$. That is, $H_1(x) = 1/\sqrt{\int_{\mathcal{R}}\exp\{-x^2\}\,\mathrm{d}x} = \pi^{-1/4}$, and for all $k \geqslant 2$,

$$H_k(x) = \frac{x^{k-1} - \sum_{j=1}^{k-1}\langle x^{k-1}, H_j\rangle_\omega H_j(x)}{\sqrt{\int_{\mathcal{R}}[x^{k-1} - \sum_{j=1}^{k-1}\langle x^{k-1}, H_j\rangle_\omega H_j(x)]^2\exp\{-x^2\}\,\mathrm{d}x}}.$$

Let $\mathrm{HPol}(J_n)$ denote the space of Hermite polynomials on $\mathcal{R}$ of degree J_n or less:

$$\mathrm{HPol}(J_n) = \left\{\sum_{k=1}^{J_n+1} a_k H_k(x)\exp\left\{-\frac{x^2}{2}\right\},\ x \in \mathcal{R}\colon a_k \in \mathcal{R}\right\}.$$

Then any function in $L_2(\mathcal{R}, leb)$ can be approximated by the $\mathrm{HPol}(J_n)$ sieve as $J_n \to \infty$.

When the $\mathrm{HPol}(J_n)$ sieve is used to approximate an unknown $\sqrt{\theta_o}$, where θ_o is a probability density function over $\mathcal{R}$, the corresponding sieve maximum likelihood estimation is also called SNP in econometrics; see e.g. Gallant and Nychka (1987), Gallant and Tauchen (1989) and Coppejans and Gallant (2002).

Laguerre polynomials. Laguerre polynomial series $\{L_k\colon k = 1, 2, \ldots\}$ is an orthonormal basis of $L_2([0, \infty), \omega)$ with $\omega(x) = \exp\{-x\}$. It can be obtained by applying the Gram–Schmidt procedure to the polynomial series $\{x^{k-1}\colon k = 1, 2, \ldots\}$ under the inner product $\langle f, g\rangle_\omega = \int_0^\infty f(x)g(x)\exp\{-x\}\,\mathrm{d}x$. Let $\mathrm{LPol}(J_n)$ denote the space of Laguerre polynomials on $[0, \infty)$ of degree J_n or less:

$$\mathrm{LPol}(J_n) = \left\{\sum_{k=1}^{J_n+1} a_k L_k(x)\exp\left\{-\frac{x}{2}\right\},\ x \in [0, \infty)\colon a_k \in \mathcal{R}\right\}.$$

Then any function in $L_2([0, \infty), leb)$ can be approximated by the $\mathrm{LPol}(J_n)$ sieve as $J_n \to \infty$.

2.3.3. Other smoothness classes and (finite-dimensional) nonlinear sieves

Nonlinear sieves can also be used for sieve extremum estimation. A popular class of nonlinear sieves in econometrics is single hidden layer feedforward Artificial Neural Networks (ANN). Here we present three typical forms of ANNs; see Hornik et al. (1994) for additional ones.

Sigmoid ANN. Define

$$\mathrm{sANN}(k_n) = \left\{\sum_{j=1}^{k_n}\alpha_j S\big(\gamma_j' x + \gamma_{0,j}\big)\colon \gamma_j \in \mathcal{R}^d,\ \alpha_j, \gamma_{0,j} \in \mathcal{R}\right\},$$

where $S: \mathcal{R} \to \mathcal{R}$ is a sigmoid activation function, i.e., a bounded nondecreasing function such that $\lim_{u \to -\infty} S(u) = 0$ and $\lim_{u \to \infty} S(u) = 1$. Some popular sigmoid activation functions include

- Heaviside $S(u) = 1\{u \geqslant 0\}$;
- logistic $S(u) = 1/(1 + \exp\{-u\})$;
- hyperbolic tangent $S(u) = (\exp\{u\} - \exp\{-u\})/(\exp\{u\} + \exp\{-u\})$;
- Gaussian sigmoid $S(u) = (2\pi)^{-1/2} \int_{-\infty}^{u} \exp(-y^2/2)\,\mathrm{d}y$;
- cosine squasher $S(u) = \frac{1+\cos(u+3\pi/2)}{2} 1\{|u| \leqslant \pi/2\} + 1\{u > \pi/2\}$.

Let $\mathcal{X}$ be a compact set in $\mathcal{R}^d$, and $C(\mathcal{X})$ be the space of continuous functions mapping from $\mathcal{X}$ to $\mathcal{R}$. Gallant and White (1988a) first established that the sANN sieve with the cosine squasher activation function is dense in $C(\mathcal{X})$ under the sup-norm. Cybenko (1990) and Hornik, Stinchcombe and White (1989) show that the sANN(k_n), with any sigmoid activation function, is dense in $C(\mathcal{X})$ under the sup-norm.

Let $\mathcal{H} = \{h \in L_2(\mathcal{X}, leb)\colon \int_{\mathcal{R}^d} |w||\tilde{h}(w)|\,\mathrm{d}w < \infty\}$. This means $h \in \mathcal{H}$ if and only if it is square integrable and its Fourier transform $\tilde{h}$ has finite first moment, where $\tilde{h}(w) \equiv \int \exp(-iwx)h(x)\,\mathrm{d}x$ is the Fourier transform of h. Barron (1993) established that for any $h_o \in \mathcal{H}$, the sANN(k_n) sieve approximation error rate in $L_2(\mathcal{X}, leb)$-norm ρ_{2n} is no slower than $O([k_n]^{-1/2})$, which was later improved to $O([k_n]^{-1/2-1/(2d)})$ in Makovoz (1996) for the sANN(k_n) with the Heaviside sigmoid function, and to $O([k_n]^{-1/2-1/(d+1)})$ in Chen and White (1999) for the sANN(k_n) with general sigmoid function.

General ANN. Define

$$\text{gANN}(k_n) = \left\{ \sum_{j=1}^{2^r k_n} \alpha_j \left[\max\{|\gamma_j|_e, 1\}\right]^{-m} \psi\left(\gamma_j' x + \gamma_{0,j}\right)\colon \gamma_j \in \mathcal{R}^d,\ \alpha_j, \gamma_{0,j} \in \mathcal{R} \right\},$$

where $\psi: \mathcal{R} \to \mathcal{R}$ is any activation function but not a polynomial with fixed degree. In particular, we often let ψ be a smooth function in a Hölder space $\Lambda^m(\mathcal{R})$ and satisfy $0 < \int_{\mathcal{R}} |D^r \psi(x)|\,\mathrm{d}x < \infty$ for some $r \geqslant m$. This includes all the above sigmoid activation functions as special cases (with $m = 0$ and $r = 1$); see Hornik et al. (1994) for additional examples.

Let

$$\mathcal{H} = \left\{ h \in L_2(\mathcal{X}, \mu)\colon h(x) = \int \exp(ia'x)\,\mathrm{d}\sigma_h(a), \right.$$
$$\left. \int_{\mathcal{R}^d} \left[\max\{|a|_e, 1\}\right]^{m+1} \mathrm{d}|\sigma_h|_{\mathrm{tv}}(a) < \infty \right\},$$

where σ_h is a complex-valued measure, and $|\sigma_h|_{\mathrm{tv}}$ denotes the total variation of σ_h. Let $W_2^m(\mathcal{X}, \mu)$ be the weighted Sobolev space of functions, where functions as well as all their partial derivatives (up to mth order) are $L_2(\mathcal{X}, \mu)$-integrable for a finite measure μ. It is known that a function in $\mathcal{H}$ also belongs to $W_2^m(\mathcal{X},\mu)$. Denote $\|h\|_{m,\mu} = \{\int h(x)^2\,\mathrm{d}\mu(x) + \int |D^m h(x)|_e^2\,\mathrm{d}\mu(x)\}^{1/2}$ as the weighted Sobolev norm. Hornik et al. (1994) established that for any $h_o \in \mathcal{H}$, the gANN(k_n) sieve approximation error rate in the weighted Sobolev norm ($\|\cdot\|_{m,\mu}$) is no slower than $O([k_n]^{-1/2})$, which was later improved to $O([k_n]^{-1/2-1/(d+1)})$ in Chen and White (1999).

Gaussian radial basis ANN. Let $\mathcal{X} = \mathcal{R}^d$. Define

$$\mathrm{rbANN}(k_n) = \left\{\alpha_0 + \sum_{j=1}^{k_n} \alpha_j G\left(\frac{\{(x-\gamma_j)'(x-\gamma_j)\}^{1/2}}{\sigma_j}\right):\ \gamma_j \in \mathcal{R}^d,\right.$$
$$\left.\alpha_j, \sigma_j \in \mathcal{R},\ \sigma_j > 0\right\},$$

where G is the standard Gaussian density function. Let $W_1^m(\mathcal{X})$ be the Sobolev space of functions, where functions as well as all their partial derivatives (up to mth order) are $L_1(\mathcal{X}, leb)$-integrable. Meyer (1992) shows that rbANN(k_n) is dense in the smoothness class $W_1^m(\mathcal{X})$. Girosi (1994) established that for any $h_o \in \mathcal{H}$, the rbANN(k_n) sieve approximation error rate in $L_2(\mathcal{X}, leb)$-norm ρ_{2n} is no slower than $O([k_n]^{-1/2})$, which was later improved to $O([k_n]^{-1/2-1/(d+1)})$ in Chen, Racine and Swanson (2001).

Additional examples of nonlinear sieves include spline sieves with data-driven choices of knot locations (or free-knot splines), and wavelet sieves with thresholding. Nonlinear sieves are more flexible and may enjoy better approximation properties than linear sieves; see e.g. Chen and Shen (1998) for the comparison of linear vs. nonlinear sieves.

2.3.4. Infinite-dimensional (nonlinear) sieves and method of penalization

Most commonly used sieve spaces are finite-dimensional truncated series such as those listed above. However, the general theory on sieve extremum estimation can also allow for infinite-dimensional sieve spaces. For example, consider the smoothness class $\Theta = \Lambda^p(\mathcal{X})$ with $\mathcal{X} = [0,1]$, $p > 1/2$. It is well known that any function $\theta \in \Theta$ can be expressed as an infinite Fourier series $\theta(x) = \sum_{k=1}^{\infty}[a_k\cos(kx) + b_k\sin(kx)]$, and its derivative with fractional power $\gamma \in (0, p]$ can also be defined in terms of Fourier series:

$$\theta^{(\gamma)}(x) = \sum_{k=1}^{\infty} k^{\gamma}\left[\left(a_k\cos\frac{\pi\gamma}{2} + b_k\sin\frac{\pi\gamma}{2}\right)\cos(kx)\right.$$
$$\left. + \left(b_k\cos\frac{\pi\gamma}{2} - a_k\sin\frac{\pi\gamma}{2}\right)\sin(kx)\right].$$

Similarly, any function $\theta \in \Theta = \Lambda^p(\mathcal{X})$ and its fractional derivatives can be expressed as infinite series of splines and wavelets; see e.g. Meyer (1992). Let $\mathrm{pen}(\theta) =$

$(\int_{\mathcal{X}} |\theta^{(p)}(x)|^q \, \mathrm{d}x)^{1/q}$ for $p > 1/2$ and some integer $q \geqslant 1$. Then we can take the sieves to be $\Theta_n = \{\theta \in \Theta\colon \operatorname{pen}(\theta) \leqslant b_n\}$ with $b_n \to \infty$ as $n \to \infty$ arbitrarily slowly; see e.g. Shen (1997). The choice of q is typically related to the criterion function $\widehat{Q}_n(\theta)$, such as $q = 2$ for conditional mean regression [Wahba (1990)], $q = 1$ [Koenker, Ng and Portnoy (1994)] and total variation norm [Koenker and Mizera (2003)] for quantile regressions.

More generally, if the parameter space Θ is a typical function space such as a Hölder, Sobolev or Besov space, then any function $\theta \in \Theta$ can be expressed as infinite series of some known Riesz basis $\{B_k(\cdot)\}_{k=1}^{\infty}$. An infinite-dimensional sieve space could take the form:

$$\Theta_n = \left\{\theta \in \Theta\colon \theta(\cdot) = \sum_{k=1}^{\infty} a_k B_k(\cdot),\ \operatorname{pen}(\theta) \leqslant b_n\right\} \quad \text{with } b_n \to \infty \text{ slowly}, \tag{2.18}$$

where $\operatorname{pen}(\theta)$ is a smoothness (or roughness) penalty term.

REMARK 2.2. When $\widehat{Q}_n(\theta)$ is concave and $\operatorname{pen}(\theta)$ is convex, the sieve extremum estimation, $\sup_{\theta\in\Theta_n} \widehat{Q}_n(\theta)$ with Θ_n given in (2.18), becomes equivalent to the *penalized extremum estimation*

$$\max_{\theta\in\Theta}\left\{\widehat{Q}_n(\theta) - \lambda_n \operatorname{pen}(\theta)\right\} \tag{2.19}$$

where the Lagrange multiplier λ_n is chosen such that the solution satisfies $\operatorname{pen}(\hat{\theta}) = b_n$. See e.g. Eggermont and LaRiccia (2001, Subsection 1.6).

2.3.5. *Shape-preserving sieves*

There are many sieves that can preserve the shape, such as nonnegativity, monotonicity and convexity, of the unknown function to be approximated. See e.g. DeVore (1977a, 1977b) on shape-preserving spline and polynomial sieves, Anastassiou and Yu (1992a, 1992b) and Dechevsky and Penev (1997) on shape-preserving wavelet sieves. Here we mention one of such shape-preserving sieves.

Cardinal B-spline wavelets. The cardinal B-spline of order $r \geqslant 1$ is given by

$$B_r(x) = \frac{1}{(r-1)!}\sum_{j=0}^{r}(-1)^j \binom{r}{j}\left[\max(0, x-j)\right]^{r-1}, \tag{2.20}$$

which has support $[0, r]$, is symmetric at $r/2$ and is a piecewise polynomial of highest degree $r - 1$. It satisfies $B_r(x) \geqslant 0$, $\sum_{k=-\infty}^{+\infty} B_r(x-k) = 1$ for all $x \in \mathcal{R}$, which is crucial to preserve the shape of the unknown function to be approximated. Its derivative satisfies $\frac{\partial}{\partial x} B_r(x) = B_{r-1}(x) - B_{r-1}(x-1)$. See Chui (1992, Chapter 4) for a recursive construction of cardinal B-splines and their properties.

We can construct a cardinal B-spline wavelet basis for the space $L_2(\mathcal{R}, leb)$ as follows. Let $\phi_r(x) = B_r(x)$ be the father wavelet (or the scaling function). Then there is a "unique" mother wavelet function ψ_r with minimum support $[0, 2r-1]$ and is given by

$$\psi_r(x) = \sum_{\ell=0}^{3r-2} q_\ell B_r(2x-\ell), \quad q_\ell = (-1)^\ell 2^{1-r} \sum_{j=0}^{r} \binom{r}{j} B_{2r}(\ell+1-j).$$

Let

$$\phi_{r,jk}(x) = 2^{j/2} B_r\left(2^j x - k\right), \quad \psi_{r,jk}(x) = 2^{j/2}\psi_r\left(2^j x - k\right), \quad x \in \mathcal{R}.$$

Then for an integer $j_0 \geqslant 0$, $\{\phi_{r,j_0k},\ k \in \mathbb{Z};\ \psi_{r,jk},\ j \geqslant j_0,\ k \in \mathbb{Z}\}$ is a Riesz basis of $L_2(\mathcal{R}, leb)$. Moreover, any function g in $L_2(\mathcal{R}, leb)$ has the following spline-wavelet $m = r-1$ regular multiresolution expansion:

$$g(x) = \sum_{k=-\infty}^{\infty} a_{j_0k} 2^{j_0/2} B_r\left(2^{j_0}x - k\right) + \sum_{j=j_0}^{\infty} \sum_{k=-\infty}^{\infty} b_{jk}\psi_{r,jk}(x), \quad x \in \mathcal{R},$$

see Chui (1992, Chapter 6). For an integer $J_n > j_0 = 0$, set

$$\begin{aligned}\mathrm{SplWav}\left(r-1, 2^{J_n}\right) = \Bigg\{ & \sum_{k=-\infty}^{\infty} a_{0k} B_r(x-k) \\ & + \sum_{j=0}^{J_n-1} \sum_{k=-\infty}^{\infty} \beta_{jk}\psi_{r,jk}(x),\ x \in \mathcal{R}\colon a_{0k}, \beta_{jk} \in \mathcal{R} \Bigg\}\end{aligned}$$

or, equivalently,[19]

$$\mathrm{SplWav}\left(r-1, 2^{J_n}\right) = \left\{ \sum_{k=-\infty}^{\infty} \alpha_k 2^{J_n/2} B_r\left(2^{J_n}x - k\right),\ x \in \mathcal{R}\colon \alpha_k \in \mathcal{R} \right\}.$$

Any nondecreasing continuous function on $\mathcal{R}$ can be approximated well by the $\mathrm{SplWav}(r-1, 2^{J_n})$ sieve with nondecreasing sequence $\{\alpha_k\}$ (i.e., $\alpha_k \leqslant \alpha_{k+1}$). In particular, let

$$\begin{aligned}\mathrm{MSplWav}\left(r-1, 2^{J_n}\right) = \Bigg\{ & g(x) = \sum_{k=-\infty}^{\infty} \alpha_k 2^{J_n/2} B_r\left(2^{J_n}x - k + \frac{r}{2}\right)\colon \\ & \alpha_k \leqslant \alpha_{k+1} \Bigg\}\end{aligned}$$

[19] See Chen, Hansen and Scheinkman (1998) for the approximation property of this sieve for twice differentiable functions on $\mathcal{R}$.

denote the monotone spline wavelet sieve. Then for any bounded nondecreasing continuous function θ_o on $\mathcal{R}$, the MSplWav$(r-1, 2^{J_n})$, $r \geqslant 1$, sieve approximation error rate in sup-norm is $O(2^{-J_n})$; for any bounded nondecreasing continuously differentiable function θ_o on $\mathcal{R}$, the MSplWav$(r-1, 2^{J_n})$, $r \geqslant 2$, sieve approximation error rate in sup-norm is $O(2^{-2J_n})$; see e.g. Anastassiou and Yu (1992a).

2.3.6. *Choice of a sieve space*

The choice of a sieve space $\Theta_n = B \times \mathcal{H}_n$ depends on how well it approximates $\Theta = B \times \mathcal{H}$ and how easily one can compute $\max_{\theta \in \Theta_n} \widehat{Q}_n(\theta)$.

In general, it will be easier to compute $\max_{\theta \in \Theta_n} \widehat{Q}_n(\theta)$ when the sieve space, $\Theta_n = B \times \mathcal{H}_n$, is an unconstrained finite-dimensional linear space. Moreover, if the criterion function, $\widehat{Q}_n(\theta)$, is concave, one can choose such a linear sieve, just as in the series estimation of a concave extended linear model described in Subsection 2.2.2.

However, the ease of computation should not be the only concern when one decides which sieve to use in practice. This is because the large sample performance of a sieve estimate also depends on the approximation properties of the chosen sieve. Unfortunately, a finite-dimensional linear sieve does not always possess better approximation properties than some nonlinear sieves. For example, let us consider the estimation of a multivariate conditional mean function $h_o(\cdot) = E[Y_t|X_t = \cdot] \in \Theta$. Let Θ_n be a sieve space. Then $\hat{\theta} = \hat{h} = \arg\max_{h \in \Theta_n} \frac{-1}{n} \sum_{t=1}^{n} [Y_t - h(X_t)]^2$ is a sieve M-estimator of h_o. If $\Theta = \Lambda^p([0,1]^d)$ is the space of p-smooth functions with $p > d/2$, then one can take Θ_n to be any of the finite-dimensional linear sieve space in Subsection 2.3.1, and the resulting estimator $\hat{h}$ is a series estimator. However, if $\Theta = W_1^1([0,1]^d)$ as defined in Subsection 2.3.3, then it is better to choose the sieve space, Θ_n, to be the nonlinear Gaussian radial basis ANN in Subsection 2.3.3; the resulting estimator is still a sieve M-estimator but not a series estimator. See Section 3 for additional examples.

How well a sieve, Θ_n, approximates Θ often depends on the support, the smoothness, the shape restrictions of functions in Θ and the structure, such as additivity, nonnegativity, exclusion restrictions, imposed by the econometric model. For example, a Hermite polynomial sieve can approximate a multivariate unknown smooth density with unbounded supports and relatively thin tails well, but a power series sieve and a Fourier series sieve cannot. This is why Gallant and Nychka (1987) considered Hermite polynomial sieve MLE since they wanted to approximate multivariate densities that are smooth, have unbounded supports and include the multivariate normal density as a special case. As another example, a first-order monotone spline sieve can approximate any bounded monotone but nondifferentiable function well, and a third-order cardinal B-spline wavelet sieve can approximate any bounded monotone differentiable function well. In Example 2.1, Heckman and Singer (1984, pp. 300 and 301) did not want to impose any assumptions on the distribution function $h(\cdot)$ of the latent random factor, hence they applied a first-order monotone spline sieve to approximate it. In their estimation of the first eigenfunction of the conditional expectation operator associated with a fully nonparametric scalar diffusion model, Chen, Hansen and Scheinkman (1998) applied a shape-preserving third order cardinal B-spline wavelet sieve to approximate the unknown first eigenfunction, since the first eigenfunction is known to be monotone and twice continuously differentiable. As a final example, in their sieve MD estimation of the semi-nonparametric external habit model (2.7) of Example 2.3, Chen and Ludvigson (2003) used the sANN sieve with logistic activation function to approximate the unknown habit function $H(C_t, C_{t-1}, \ldots, C_{t-L}) = C_t h(\frac{C_{t-1}}{C_t}, \ldots, \frac{C_{t-L}}{C_t})$. This is partly because when $L \geqslant 3$, the unknown smooth function $h : \mathcal{R}^L \to [0, 1)$ can be approximated by a sANN sieve well, and partly because it is very easy to impose the habit constraint $0 \leqslant H(C_t, C_{t-1}, \ldots, C_{t-L}) < C_t$ when $h(\frac{C_{t-1}}{C_t}, \ldots, \frac{C_{t-L}}{C_t})$ is approximated by the sANN sieve with logistic activation function.

For a sieve estimate to be consistent with a fast rate of convergence, it is important to choose sieves with good approximation error rates as well as controlled complexity.[20] Nevertheless, for econometric applications where the only prior information on the unknown functions is their smoothness and supports, the choice of a sieve space is not important, as long as the chosen sieve space has the desired approximation error rate.

2.4. A small Monte Carlo study

To illustrate how to implement the sieve extremum estimation, we present a small Monte Carlo simulation carried out using Matlab and Fortran. The true model is: $Y_1 = X_1\beta_o + h_{o1}(Y_2) + h_{o2}(X_2) + U$ with $\beta_o = 1$, $h_{o1}(Y_2) = 1/[1 + \exp\{-Y_2\}]$ and $h_{o2}(X_2) = \log(1 + X_2)$. We assume that Y_2 is endogenous and $Y_2 = X_1 + X_2 + X_3 + R \times U + e$ with either $R = 0.9$ (strong correlation) or 0.1 (weak correlation). Suppose that the regressors X_1, X_2, X_3 are independent and uniformly distributed over $[0, 1]$, and that e is independent of (X, U) and normally distributed with mean zero and variance 0.1. (We have also tried $E[e^2] = 0.05, 0.25$, the simulation results share very similar patterns to the ones when $E[e^2] = 0.1$, hence are not reported here.) Conditional on $X = (X_1, X_2, X_3)'$, U is normally distributed with mean zero and variance $(X_1^2 + X_2^2 + X_3^2)/3$. Let $Z = (Y_1, Y_2, X')'$. A random sample of $n = 1000$ data $\{Z_i\}_{i=1}^n$ is generated from this design. An econometrician observes the simulated data $\{Z_i\}_{i=1}^n$, and wants to estimate $\theta_o = (\beta_o, h_{o1}, h_{o2})'$, obeying the conditional moment restriction:

$$E\big[Y_{1i} - \big\{X_{1i}\beta_o + h_{o1}(Y_{2i}) + h_{o2}(X_{2i})\big\}\big|X_i\big] = 0. \tag{2.21}$$

This model is a generalization of the partially linear IV regression $E[Y_1 - \{X_1\beta_o + h_{o1}(Y_2)\}|X] = 0$ example of Ai and Chen (2003) to a partially additive IV regression. Since $h_{o1}(Y_2)$ is an unknown function of the endogenous variable Y_2, both examples belong to the so-called ill-posed inverse problems.

Let $\rho(Z, \theta) = Y_1 - \{X_1\beta + h_1(Y_2) + h_2(X_2)\}$ with $\theta = (\beta, h_1, h_2)'$. We say that the parameters $\theta_o = (\beta_o, h_{o1}, h_{o2})'$ are identified if $E[\rho(Z, \theta)|X] = 0$ only when $\theta = \theta_o$.

[20] This will become clear from the large sample theory discussed later in Section 3.

As a sufficient condition for the identification of θ_o, we assume that $\mathrm{Var}(X_1) > 0$, $h_1(y_2)$ is a bounded function with $\sup_{y_2} |h_1(y_2)| \leqslant 1$ and that $h_2(x_2)$ satisfies $h_2(0.5) = \log(3/2)$. In particular, we assume that $\theta_o = (\beta_o, h_{o1}, h_{o2})' \in \Theta = B \times \mathcal{H}_1 \times \mathcal{H}_2$ with B a compact interval in $\mathcal{R}$, $\mathcal{H}_1 = \{h_1 \in C^2(\mathcal{R})\colon \sup_{y_2} |h_1(y_2)| \leqslant 1, \int [D^2 h_1(y_2)]^2 \,\mathrm{d}y_2 < \infty\}$ and $\mathcal{H}_2 = \{h_2 \in C^2([0,1])\colon h_2(0.5) = \log(3/2), \int [D^2 h_2(x_2)]^2 \,\mathrm{d}x_2 < \infty\}$.

Since this model (2.21) fits into the second subclass of the conditional moment restrictions (2.8) with $E[\rho(Z, \theta_o)|X] = 0$, we can apply the sieve MD criterion (2.12) to estimate $\theta_o = (\beta_o, h_{o1}, h_{o2})$. We take $\Theta_n = B \times \mathcal{H}_{1n} \times \mathcal{H}_{2n}$ as the sieve space, where

$$\mathcal{H}_{1n} = \left\{ h_1(y_2) = \Pi_1' B^{k_{1,n}}(y_2)\colon \int \left[D^2 h_1(y_2)\right]^2 \mathrm{d}y_2 \leqslant c_1 \log n \right\},$$

$B^{k_{1,n}}(y_2)$ is either a polynomial spline basis with equally spaced (according to empirical quantile of Y_2) knots, or a 3rd order cardinal B-spline basis, or a Hermite polynomial basis,[21] and $\dim(\Pi_1) = k_{1,n}$ is the number of unknown sieve coefficient of h_1. Similarly,

$$\mathcal{H}_{2n} = \left\{ h_2(x_2) = \Pi_2' B^{k_{2,n}}(x_2)\colon \int \left[D^2 h_2(x_2)\right]^2 \mathrm{d}x_2 \leqslant c_2 \log n, \right.$$
$$\left. h_2(0.5) = \log(3/2) \right\},$$

$B^{k_{2,n}}(x_2)$ is either a polynomial spline basis with equally spaced (according to empirical quantile of X_2) knots, or a 3rd order cardinal B-spline basis, and $\dim(\Pi_2) = k_{2,n}$ is the number of unknown sieve coefficients of h_2. In the Monte Carlo study, we have tried $k_{1,n} = 4, 5, 6, 8$ and $k_{2,n} = 4, 5, 6$.

As an illustration, we only consider the sieve MD estimation (2.12) using the identity weighting $\widehat{\Sigma}(X) = I$,[22] and the series LS estimator as the $\hat{m}(X, \theta)$ for the conditional mean function $E[\rho(Z, \theta)|X]$, thus the criterion becomes

$$\min_{\beta \in B,\, h_1 \in \mathcal{H}_{1n},\, h_2 \in \mathcal{H}_{2n}} \frac{1}{n} \sum_{i=1}^{n} \left\{\hat{m}(X_i, \theta)\right\}^2, \quad \text{with}$$

$$\hat{m}(X, \theta) = \sum_{j=1}^{n} \left[Y_{1j} - \left\{X_{1j}\beta + h_1(Y_{2j}) \right.\right.$$
$$\left.\left. + h_2(X_{2j})\right\}\right] p^{k_{m,n}}(X_j)'(P'P)^{-} p^{k_{m,n}}(X),$$

where in the simulation $p^{k_{m,n}}(X)$ is taken to be the 4th degree polynomial spline sieve, with basis $\{1, X_1, X_1^2, X_1^3, X_1^4, [\max(X_1 - 0.5, 0)]^4, X_2, X_2^2, X_2^3, X_2^4, [\max(X_2 - 0.5, 0)]^4, X_3, X_3^2, X_3^3, X_3^4, [\max(X_3 - 0.1, 0)]^4, [\max(X_3 - 0.25, 0)]^4, [\max(X_3 -$

[21] See Blundell, Chen and Kristensen (2007) for a more detailed description on the choice of $\mathcal{H}_{1n}$.

[22] See Subsection 4.3 or Ai and Chen (2003) for the sieve MD procedure with the optimal weighting matrix.

$0.5, 0)]^4$, $[\max(X_3 - 0.75, 0)]^4$, $[\max(X_3 - 0.90, 0)]^4$, X_1X_3, X_2X_3, $X_1[\max(X_3 - 0.25, 0)]^4$, $X_2[\max(X_3 - 0.25, 0)]^4$, $X_1[\max(X_3 - 0.75, 0)]^4$, $X_2[\max(X_3 - 0.75, 0)]^4\}$. We note that the above criterion is equivalent to a constrained 2 Stage Least Squares (2SLS) with $k_{m,n} = 26$ instruments and $\dim(\Theta_n) = 1 + k_{1,n} + k_{2,n}$ $(< k_{m,n})$ unknown parameters:

$$\min_{\beta \in B,\, h_1 \in \mathcal{H}_{1n},\, h_2 \in \mathcal{H}_{2n}} [\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi]' P(P'P)^- P' [\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi],$$

where $\mathbf{Y}_1 = (Y_{11}, \ldots, Y_{1n})'$, $\mathbf{X}_1 = (X_{11}, \ldots, X_{1n})'$, $\Pi = (\Pi_1', \Pi_2')'$, $\mathbf{B}_1 = (B^{k_{1,n}}(Y_{21}), \ldots, B^{k_{1,n}}(Y_{2n}))'$, $\mathbf{B}_2 = (B^{k_{2,n}}(X_{21}), \ldots, B^{k_{2,n}}(X_{2n}))'$ and $\mathbf{B} = (\mathbf{B}_1', \mathbf{B}_2')'$.

Since $\rho(Z, \theta)$ is linear in $\theta = (\beta, h_1, h_2)'$, the joint sieve MD estimation is equivalent to the profile sieve MD estimation for this model. We can first compute a profile sieve estimator for $h_1(y_2) + h_2(x_2)$. That is, for any fixed β, we compute the sieve coefficients Π by minimizing $\sum_{i=1}^n \{\hat{m}(X_i, \theta)\}^2$ subject to the smoothness constraints imposed on the functions h_1 and h_2:

$$\min_{\Pi:\, \int [D^2 h_\ell(y)]^2\, \mathrm{d}y \leqslant c_\ell \log n,\, \ell = 1,2} [\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi]' P(P'P)^- P' [\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi] \tag{2.22}$$

for some upper bounds $c_\ell > 0$, $\ell = 1, 2$. Let $\widetilde{\Pi}(\beta)$ be the solution to (2.22) and $\tilde{h}_1(y_2; \beta) + \tilde{h}_2(x_2; \beta) = (B^{k_{1,n}}(y_2)', B^{k_{2,n}}(x_2)')\widetilde{\Pi}(\beta)$ be the profile sieve estimator of $h_1(y_2) + h_2(x_2)$. Next, we estimate β by $\hat{\beta}_{\mathrm{iv}}$ which solves the following 2SLS problem:

$$\min_{\beta} \big[\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\widetilde{\Pi}(\beta)\big]' P(P'P)^- P' \big[\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\widetilde{\Pi}(\beta)\big]. \tag{2.23}$$

Finally we estimate $h_{o1}(y_2) + h_{o2}(x_2)$ by

$$\hat{h}_1(y_2) + \hat{h}_2(x_2) = \big(B^{k_{1,n}}(y_2)', B^{k_{2,n}}(x_2)'\big)\widetilde{\Pi}(\hat{\beta}_{\mathrm{iv}}),$$

and then estimate h_{o1} and h_{o2} by imposing the location constraint $h_2(0.5) = \log(3/2)$:

$$\begin{aligned} \hat{h}_{2,\mathrm{iv}}(x_2) &= B^{k_{2,n}}(x_2)'\widetilde{\Pi}_2(\hat{\beta}_{\mathrm{iv}}) - B^{k_{2,n}}(0.5)'\widetilde{\Pi}_2(\hat{\beta}_{\mathrm{iv}}) + \log(3/2), \\ \hat{h}_{1,\mathrm{iv}}(y_2) &= B^{k_{1,n}}(y_2)'\widetilde{\Pi}_1(\hat{\beta}_{\mathrm{iv}}) + B^{k_{2,n}}(0.5)'\widetilde{\Pi}_2(\hat{\beta}_{\mathrm{iv}}) - \log(3/2). \end{aligned}$$

We note that although this model (2.21) belongs to the nasty ill-posed inverse problem, the above profile sieve MD procedure is very easy to compute, and in fact, $\hat{\beta}_{\mathrm{iv}}$ and $\widetilde{\Pi}(\hat{\beta}_{\mathrm{iv}})$ have closed form solutions. To see this, we note that (2.22) is equivalent to

$$\begin{aligned} &\min_{\Pi, \lambda_\ell} (\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi)' P(P'P)^- P' (\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi) \\ &\quad + \sum_{\ell=1}^{2} \lambda_\ell \big\{\Pi_\ell' C_\ell \Pi_\ell - c_\ell \log n\big\}, \end{aligned}$$

where for $\ell = 1, 2$, $C_\ell = \int [D^2 B^{k_{\ell,n}}(y)][D^2 B^{k_{\ell,n}}(y)]'\, \mathrm{d}y$, $\Pi_\ell' C_\ell \Pi_\ell = \int [D^2 h_\ell(y)]^2\, \mathrm{d}y$ and $\lambda_\ell \geqslant 0$ is the Lagrange multiplier. However, we do not want to specify the upper bounds $c_\ell > 0, \ell = 1, 2$, instead we choose some small values as the penalization weights λ_1, λ_2, and solve the following problems:

$$\min_{\Pi}(\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi)' P(P'P)^- P'(\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi) + \sum_{\ell=1}^{2} \lambda_\ell \Pi_\ell' C_\ell \Pi_\ell. \quad (2.24)$$

Denote $C(\lambda_1, \lambda_2) = \left[\begin{smallmatrix} \lambda_1 C_1 & 0 \\ 0 & \lambda_2 C_2 \end{smallmatrix}\right]$ as the smoothness penalization matrix. The minimization problem (2.24) has a simple closed form solution:

$$\begin{aligned} \widetilde{\Pi}(\beta) &= \left(\mathbf{B}'P(P'P)^- P'\mathbf{B} + C(\lambda_1, \lambda_2)\right)^- \mathbf{B}'P(P'P)^- P'[\mathbf{Y}_1 - \mathbf{X}_1\beta] \\ &= W[\mathbf{Y}_1 - \mathbf{X}_1\beta], \end{aligned}$$

with $W = (\mathbf{B}'P(P'P)^- P'\mathbf{B} + C(\lambda_1, \lambda_2))^- \mathbf{B}'P(P'P)^- P'$. Substituting the solution $\widetilde{\Pi}(\beta)$ into the 2SLS problem (2.23), we obtain

$$\begin{aligned} \hat{\beta}_{\mathrm{iv}} &= \left[\mathbf{X}_1'(I - \mathbf{B}W)' P(P'P)^- P'(I - \mathbf{B}W)\mathbf{X}_1\right]^{-1} \mathbf{X}_1' \\ &\quad \times (I - \mathbf{B}W)' P(P'P)^- P'(I - \mathbf{B}W)\mathbf{Y}_1, \end{aligned}$$

and $\widetilde{\Pi}(\hat{\beta}_{\mathrm{iv}}) = W[\mathbf{Y}_1 - \mathbf{X}_1\hat{\beta}_{\mathrm{iv}}]$.

Table 1
Different endogeneity, Spl(3, 2) for h_2, $k_{2n} = 5$, $\lambda_2 = 0.0001$

R	β	SE(β)	IBias^2(h_1)	IMSE(h_1)	IBias^2(h_2)	IMSE(h_2)
		Spl(3, 2)	$k_{1n} = 5$	$\lambda_1 = 0.005$		
0.0	1.0081	0.0909	0.0003	0.0427	0.0000	0.0026
0.1	1.0021	0.0907	0.0003	0.0446	0.0000	0.0026
0.9	0.9404	0.0947	0.0148	0.0926	0.0003	0.0030
		Spl(3, 1)	$k_{1n} = 4$	$\lambda_1 = 0.001$		
0.0	1.0076	0.0891	0.0002	0.0225	0.0000	0.0025
0.1	1.0010	0.0886	0.0002	0.0229	0.0000	0.0025
0.9	0.9398	0.0941	0.0160	0.0623	0.0003	0.0029
		HPol(4)	$k_{1n} = 5$	$\lambda_1 = 0.005$		
0.0	1.0089	0.0906	0.0003	0.0395	0.0000	0.0026
0.1	1.0029	0.0901	0.0003	0.0397	0.0000	0.0026
0.9	0.9418	0.0948	0.0121	0.0830	0.0003	0.0030
		HPol(3)	$k_{1n} = 4$	$\lambda_1 = 0.001$		
0.0	1.0078	0.0890	0.0002	0.0202	0.0000	0.0025
0.1	1.0012	0.0885	0.0002	0.0205	0.0000	0.0025
0.9	0.9401	0.0941	0.0112	0.0546	0.0003	0.0029

Table 2
Different penalization levels and sieve terms, $R = 0.9$

(λ_1, λ_2)	β	SE(β)	IBias$^2(h_1)$	IMSE(h_1)	IBias$^2(h_2)$	IMSE(h_2)
Spl(3, 1) for h_1 and h_2, $k_{1n} = k_{2n} = 4$						
(0.001, 0.0)	0.9366	0.0941	0.0176	0.0612	0.0003	0.0018
(0.05, 0.001)	0.9324	0.0867	0.0185	0.0568	0.0003	0.0016
Spl(3, 3) for h_1 and h_2, $k_{1n} = k_{2n} = 6$						
(0.001, 0.0)	0.9451	0.0984	0.0124	0.1594	0.0003	0.0032
(0.05, 0.001)	0.9441	0.0954	0.0125	0.0720	0.0003	0.0028

For $R = 0.9$, 0.1 and 0.0, a sample of 1000 data points were generated according to the above design. The sieve MD procedure was applied to the data with identity weighting matrix $\widehat{\Sigma}(X) = I$ and the penalization weights $\lambda_1 = 0.005$ (or 0.001) and $\lambda_2 = 0.0001$ (or 0) for simplicity. The estimated coefficients were recorded. Then, a new sample of 1000 data points were drawn and the estimated coefficients were computed again. This procedure was repeated 400 times. The mean (M) and standard error (SE) of the β_o estimator across the 400 simulations are reported in Tables 1–2. To evaluate the performance of the sieve MD estimators of the nonparametric components $h_{o1}(Y_2)$ and $h_{o2}(X_2)$, we report their integrated squared biases (IBias^2) and the integrated mean squared errors (IMSE) across the 400 simulations in Tables 1–2.[23] Table 1 summarizes the performance of the estimators across different degrees of endogeneity and different sieves for $h_1(Y_2)$. Table 2 summarizes the sensitivity of the estimators (under $R = 0.9$) to different sieve number of terms and penalization parameters for both $h_1(Y_2)$ and $h_2(X_2)$. We also plot the estimated functions $h_{o1}(Y_2)$ and $h_{o2}(X_2)$ corresponding to the strong correlation case ($R = 0.9$) in Figure 1, where the solid lines represent the true functions and the dashed (or dotted) lines denote the sieve MD (or sieve IV) estimates.

Tables 1–2 and Figure 1 indicate that even under strong correlation, the sieve MD estimates of β_o and $h_{o2}(X_2)$ perform well. We find that the sieve IV estimates of β_o and $h_{o2}(X_2)$ are not sensitive to the choices of the penalization parameters λ_1, λ_2, nor to the choices of sieve bases for $h_{o1}(Y_2)$. The sieve IV estimate of $h_{o1}(Y_2)$ is also not very sensitive to the choices of sieve bases, although it is slightly more sensitive to the penalization parameter λ_1 under strong correlation. Since under strong correlation, the

[23] The IBias$^2(h_1)$ and IMSE(h_1) in Table 1 are calculated as follows. Let $\hat{h}_i$ be the estimate of h_{o1} from the ith simulated data set, and $\bar{h}(y) = \sum_{i=1}^{400} \hat{h}_i(y)/400$ be the pointwise average across 200 simulations. We calculate the pointwise squared bias as $[\bar{h}(y) - h_{o1}(y)]^2$, and the pointwise variance as $400^{-1} \sum_{i=1}^{400} [\hat{h}_i(y) - \bar{h}(y)]^2$. The integrated squared bias is calculated by numerically integrating the pointwise squared bias from $\underline{y}$ to $\bar{y}$ which are respectively the 2.5th and 97.5th empirical percentiles of Y_2; The integrated MSE are computed in a similar way.

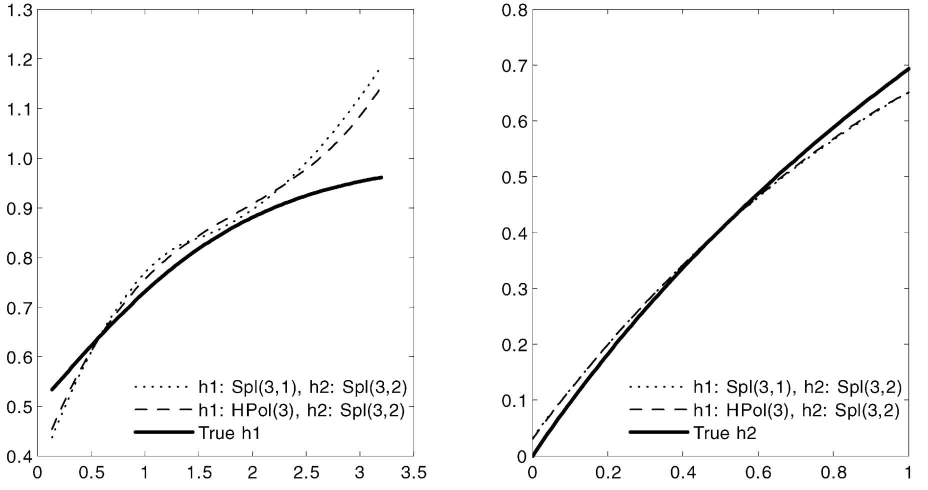

Figure 1. True and estimated functions with $R = 0.9$, $\lambda_1 = 0.001$, $\lambda_2 = 0.0001$.

estimation of $h_{o1}(Y_2)$ is a nasty ill-posed inverse problem, as the penalization parameter λ_1 gets smaller, the integrated squared bias of $h_{o1}()$ does not change much but the integrated variance of $h_{o1}()$ increases more. The additional Monte Carlo results for other sieve bases such as 3rd order cardinal B-splines and for different combinations of sieve number of terms and penalization levels share similar patterns to the ones reported here. These findings are also consistent with the more detailed Monte Carlo studies in Blundell, Chen and Kristensen (2007).

2.5. An incomplete list of sieve applications in econometrics

We conclude this section by listing a few applications of the sieve extremum estimation in econometrics.[24] Most of the existing applications are done in microeconometrics. Elbadawi, Gallant and Souza (1983) studied Fourier series LS estimation of demand elasticity. Cosslett (1983) proposed nonparametric ML estimation of a binary choice model. Heckman and Singer (1984) considered sieve ML estimation of a duration model where the unknown error distribution is approximated by a first-order spline. Their estimation procedure was also applied in Cameron and Heckman (1998) to a life-cycle schooling problem. Duncan (1986) used spline sieve MLE in estimating a censored regression. Hausman and Newey (1995) considered power series and spline series LS estimation of consumer surplus. Hahn (1998) and Imbens, Newey and Ridder (2005)

[24] Although restricting our attention to economic applications only, it is still impossible to mention all the existing applications of sieve methods in econometrics. Any omissions reflect my lack of awareness and are purely unintentional.

used power series and splines in the two-step efficient estimation of the average treatment effect models. Newey, Powell and Vella (1999), and Pinkse (2000) considered series estimation of a triangular system of simultaneous equations. To estimate semiparametric generalizations of Heckman's (1979) sample selection model, Gallant and Nychka (1987) proposed the Hermite polynomial sieve MLE, while Newey (1988) and Das, Newey and Vella (2003) applied the series LS estimation method. Recently, Newey (2001) used the sieve MD procedure to estimate a nonlinear measurement error model. Blundell, Chen and Kristensen (2007) considered a profile sieve MD procedure to estimate shape-invariant Engel curves with nonparametric endogenous expenditure. Coppejans (2001) proposed sieve ML estimation of a binary choice model. Khan (2005) considered a sieve LS estimation of a probit binary choice model with unknown heteroskedasticity. Hirano, Imbens and Ridder (2003) proposed a sieve logistic regression to estimate propensity score for treatment effect models. Mahajan (2004) estimated a semiparametric single index model with binary misclassified regressors via sieve MLE. Chen, Fan and Tsyrennikov (2006) studied sieve MLE of semi-nonparametric multivariate copula models. Chen, Hong and Tamer (2005) made use of spline sieves to estimate nonlinear nonclassical measurement error models with an auxiliary sample. Their estimation procedure was shown in Chen, Hong and Tarozzi (2007) to be semiparametrically efficient for general nonlinear GMM models of nonclassical measurement errors, missing data and treatment effects. Hu and Schennach (2006) apply sieve MLE to estimate a nonlinear nonclassical measurement error model with instruments. Brendstrup and Paarsch (2004) applied Hermite and Laguerre polynomial sieve MLE to estimate sequential asymmetric English auctions. Bierens (in press) and Bierens and Carvalho (in press) applied Legendre polynomial sieve MLE respectively to estimate an interval-censored mixed proportional hazard model and a competing risks model of recidivism.

There have also been many applications of the method of sieves in time series econometrics. Engle et al. (1986) forecasted electricity demand using a partially linear spline regression. Engle and Gonzalez-Rivera (1991) applied sieve MLE to estimate ARCH models where the unknown density of the standardized innovation is approximated by a first order spline sieve. Gallant and Tauchen (1989) and Gallant, Hsieh and Tauchen (1991) employed Hermite polynomial sieve MLE to study asset pricing and foreign exchange rates. Gallant and Tauchen (1996, 2004) have proposed the combinations of Hermite polynomial sieve and simulated method of moments to effectively solve many complicated asset pricing models with latent factors, and their methods have been widely applied in empirical finance. Bansal and Viswanathan (1993), Bansal, Hsieh and Viswanathan (1993) and Chapman (1997) considered sieve approximation of the whole stochastic discount factor (or pricing kernel) as a function of a few macroeconomic factors. White (1990) and Granger and Teräsvirta (1993) suggested nonparametric LS forecasting via sigmoid ANN sieve. Hutchinson, Lo and Poggio (1994) applied radial basis ANN to option pricing. Chen, Racine and Swanson (2001) used partially linear ANN and ridgelet sieves to forecast US inflation. McCaffrey et al. (1992) estimated the Lyapunov exponent of a chaotic system via ANN sieves.[25] Chen and Ludvigson (2003) employed a sigmoid ANN sieve to estimate the unknown habit function in a consumption asset pricing model. Polk, Thompson and Vuolteenaho (2003) applied sigmoid ANN to compute conditional quantile in testing stock return predictability. Chen, Hansen and Scheinkman (1998) employed a shape-preserving spline-wavelet sieve to estimate the eigenfunctions of a fully nonparametric scalar diffusion model from discrete-time low-frequency observations. Chen and Conley (2001) made use of the same sieve to estimate a spatial temporal model with flexible conditional mean and conditional covariance. Phillips (1998) applied orthonormal basis to analyze spurious regressions. Engle and Rangel (2004) proposed a new Spline GARCH model to measure unconditional volatility and have applied it to equity markets for 50 countries for up to 50 years of daily data. See Fan and Yao (2003) for additional applications to financial time series models.

3. Large sample properties of sieve estimation of unknown functions

We already know that the sieve method is very general and easily implementable. In this section, we shall first establish that, under mild regularity conditions, the sieve extremum estimation will consistently estimate both finite-dimensional and infinite-dimensional unknown parameters. However, for econometric and statistical inference, one would like to know how accurate a consistent sieve estimator might be given a finite data set and what its limiting distribution is. Unfortunately there does not yet exist a general theory of pointwise limiting distribution for a sieve extremum estimator of an unknown function. There are a few results on pointwise limiting distribution for series estimators of densities and LS regression functions, which we shall review at the end of this section. However, all is not lost. We do have a well developed theory on $\sqrt{n}$-asymptotic normality of sieve estimators of smooth functionals[26] of unknown functions.

As we shall see in Section 4, in order to derive $\sqrt{n}$-asymptotic normality and semiparametric efficiency of sieve estimators of parametric components in a semi-nonparametric model, the sieve estimators of the nonparametric components should converge to the true unknown functions at rates faster than $n^{-1/4}$ under certain metric. This motivates the importance of establishing rates of convergence for sieve estimators of unknown functions even when the unknown functions are nuisance parameters (i.e., not the parameters of interest). Moreover, when an unknown function is also a parameter of interest in a nonparametric or a semi-nonparametric model, the convergence rate

[25] Their work is closely related to the estimation of derivative of a multivariate unknown regression function via ANN sieves in Gallant and White (1992). Shintani and Linton (2004) proposed a nonparametric test of chaos via ANN sieves.

[26] See Section 4 for the definition of a "smooth functional". Here it suffices to know that regular finite-dimensional parameters and average derivatives of unknown functions are examples of smooth functionals.

will provide useful information on the accuracy of a sieve estimator for a given finite sample size. Unfortunately, to date there is no unified theory on rates of convergence for the general sieve extremum estimators of unknown functions either.[27] Nevertheless, the theory on convergence rates of sieve M-estimators is by now well developed.

In this section we first provide a new consistency theorem on general sieve extremum estimation in Subsection 3.1. We then review the existing results on convergence rates and pointwise limiting distributions for sieve M-estimators of unknown functions. We begin this discussion with a survey of the convergence rate results for general sieve M-estimators of unknown functions in Subsection 3.2 and illustrate how to verify the technical conditions assumed for the general result with two examples. Although series estimation is a special case of sieve M-estimation, due to its special properties (i.e., concave criterion and finite-dimensional linear sieve space), the convergence rate of a series estimator can be derived under alternative sufficient conditions, which will be reviewed in Subsection 3.3. Subsection 3.4 presents the existing results on the pointwise normality of the series estimator in the special case of a LS regression function.

3.1. *Consistency of sieve extremum estimators*

For an infinite-dimensional, possibly noncompact parameter space Θ, Geman and Hwang (1982) obtained the consistency of sieve MLE with i.i.d. data; White and Wooldridge (1991) obtained the consistency of sieve extremum estimates with dependent and heterogeneous data. For an infinite-dimensional, compact parameter space Θ, Gallant (1987) and Gallant and Nychka (1987) derived the consistency of sieve M-estimates; Newey and Powell (2003) and Chernozhukov, Imbens and Newey (2007) established the consistency of sieve MD estimates. In the following, we present a new consistency theorem for approximate sieve extremum estimates that allows for noncompact infinite-dimensional Θ and is applicable to ill-posed semi-nonparametric problems.[28]

Let $d(\cdot,\cdot)$ be a (pseudo) metric on Θ. In particular, when $\Theta = B \times \mathcal{H}$ where B is a subset of some Euclidean space and $\mathcal{H}$ is a subset of some normed function space, we

[27] To the best of our knowledge, currently there is one unpublished paper [Chen and Pouzo (2006)] that derives the convergence rates for the sieve MD estimates $\hat{\theta}_n$ of $\theta_o = (\beta_o, h_o)$ satisfying the semi-nonparametric conditional moment models $E[\rho(Z, \beta_o, h_o(\cdot))|X] = 0$, where the unknown $h_o(\cdot)$ could depend on the endogenous variables Y or latent variables. Earlier, Ai and Chen (2003) obtained a faster than $n^{-1/4}$ convergence rate under a weaker metric. There are also a few papers on convergence rates of sieve MD estimate of h_o in specific models; see e.g. Blundell, Chen and Kristensen (2007) and Hall and Horowitz (2005) for the model $E[Y_1 - h_o(Y_2)|X] = 0$. Van der Vaart and Wellner (1996, Theorem 3.4.1) stated an abstract rate result for sieve extremum estimation. However, their conditions rule out ill-posed semi-nonparametric problems, and require a maximal inequality with rate for the process $\sqrt{n}(\widehat{Q}_n - Q)$, which is currently not available for a general criterion $\widehat{Q}_n$. Hence, it is fair to say that a general theory on rates of convergence for sieve extremum estimators is currently lacking.

[28] Based on a recent theorem of Stinchcombe (2002), the consistency of sieve extremum estimates is a generic property.

can use $d(\theta, \tilde{\theta}) = |\beta - \tilde{\beta}|_e + \|h - \tilde{h}\|_{\mathcal{H}}$, where $|\cdot|_e$ denotes the Euclidean norm, and $\|\cdot\|_{\mathcal{H}}$ is a norm imposed on the function space $\mathcal{H}$. For example, if $\mathcal{H} = C^m(\mathcal{X})$ with a bounded $\mathcal{X}$, we could take $\|h\|_{\mathcal{H}}$ to be $\|h\|_\infty$ or $\|h\|_{2,leb}$.

CONDITION 3.1 *(Identification).*

(i) $Q(\theta_o) > -\infty$, and if $Q(\theta_o) = +\infty$ then $Q(\theta) < +\infty$ for all $\theta \in \Theta_k \setminus \{\theta_o\}$ for all $k \geqslant 1$;

(ii) there are a nonincreasing positive function $\delta()$ and a positive function $g()$ such that for all $\varepsilon > 0$ and for all $k \geqslant 1$,

$$Q(\theta_o) - \sup_{\{\theta \in \Theta_k:\, d(\theta, \theta_o) \geqslant \varepsilon\}} Q(\theta) \geqslant \delta(k) g(\varepsilon) > 0.$$

CONDITION 3.2 *(Sieve spaces).* $\Theta_k \subseteq \Theta_{k+1} \subseteq \Theta$ for all $k \geqslant 1$; and there exists a sequence $\pi_k \theta_o \in \Theta_k$ such that $d(\theta_o, \pi_k \theta_o) \to 0$ as $k \to \infty$.

CONDITION 3.3 *(Continuity).*

(i) For each $k \geqslant 1$, $Q(\theta)$ is upper semicontinuous on Θ_k under the metric $d(\cdot,\cdot)$;

(ii) $|Q(\theta_o) - Q(\pi_{k(n)}\theta_o)| = o(\delta(k(n)))$.

CONDITION 3.4 *(Compact sieve space).* The sieve spaces, Θ_k, are compact under $d(\cdot,\cdot)$.

CONDITION 3.5 *(Uniform convergence over sieves).*

(i) For all $k \geqslant 1$, $\operatorname{plim}_{n\to\infty} \sup_{\theta \in \Theta_k} |\widehat{Q}_n(\theta) - Q(\theta)| = 0$;

(ii) $\hat{c}(k(n)) = o_P(\delta(k(n)))$ where $\hat{c}(k(n)) \equiv \sup_{\theta \in \Theta_{k(n)}} |\widehat{Q}_n(\theta) - Q(\theta)|$;

(iii) $\eta_{k(n)} = o(\delta(k(n)))$.

THEOREM 3.1. *Let $\hat{\theta}_n$ be the approximate sieve extremum estimator defined by (2.9). If Conditions 3.1–3.5 hold, then $d(\hat{\theta}_n, \theta_o) = o_P(1)$.*

PROOF. By Remark 2.1, $\hat{\theta}_n$ is well defined and measurable. For all $\varepsilon > 0$, under Conditions 3.3(i) and 3.4, $\sup_{\{\theta \in \Theta_{k(n)}:\, d(\theta,\theta_o) \geqslant \varepsilon\}} Q(\theta)$ exists. By definition, we have for all $\varepsilon > 0$,

$$\begin{aligned}
\Pr\bigl(d(\hat{\theta}_n, \theta_o) > \varepsilon\bigr) &\leqslant \Pr\Bigl(\sup_{\{\theta \in \Theta_{k(n)}:\, d(\theta,\theta_o) \geqslant \varepsilon\}} \widehat{Q}_n(\theta) \geqslant \widehat{Q}_n(\pi_{k(n)}\theta_o) - O(\eta_{k(n)})\Bigr) \\
&\leqslant P_1 + P_2,
\end{aligned}$$

where

$$\begin{aligned}
P_1 &\equiv \Pr\Bigl(\sup_{\{\theta \in \Theta_{k(n)}:\, d(\theta,\theta_o) \geqslant \varepsilon\}} \bigl|\widehat{Q}_n(\theta) - Q(\theta)\bigr| > \hat{\nu}\bigl(k(n)\bigr)\Bigr) \\
&\leqslant \Pr\Bigl(\sup_{\theta \in \Theta_{k(n)}} \bigl|\widehat{Q}_n(\theta) - Q(\theta)\bigr| > \hat{\nu}\bigl(k(n)\bigr)\Bigr),
\end{aligned}$$

and

$$P_2 \equiv \Pr\Big(\sup_{\{\theta\in\Theta_{k(n)}:\, d(\theta,\theta_o)\geqslant\varepsilon\}} Q(\theta) \geqslant Q(\pi_{k(n)}\theta_o) - 2\hat{\nu}\big(k(n)\big) - O(\eta_{k(n)})\Big)$$

$$= \Pr\Big(2\hat{\nu}\big(k(n)\big) + \big\{Q(\theta_o) - Q(\pi_{k(n)}\theta_o)\big\} + O(\eta_{k(n)}) \geqslant Q(\theta_o) - \sup_{\{\theta\in\Theta_{k(n)}:\, d(\theta,\theta_o)\geqslant\varepsilon\}} Q(\theta)\Big).$$

Choosing $\hat{\nu}(k(n)) = \hat{c}(k(n))$ it follows that the $P_1 = 0$ by definition of $\hat{c}(k(n))$ and Condition 3.5(i), and $P_2 \leqslant \Pr[2\hat{c}(k(n)) + \{Q(\theta_o) - Q(\pi_{k(n)}\theta_o)\} + O(\eta_{k(n)}) \geqslant \delta(k(n))g(\varepsilon)] \to 0$ by Conditions 3.1 and 3.5(ii). □

REMARK 3.1. (1) Theorem 3.1 is applicable to both well-posed and ill-posed semi-nonparametric models. When the problem (such as the nonparametric IV regression $E[Y_1 - h_o(Y_2)|X] = 0$) is ill-posed, one may have $\liminf_k \delta(k) = 0$, which is still allowed by Conditions 3.1(ii), 3.3(ii) and 3.5(ii)(iii). See Chen and Pouzo (2006) for alternative general consistency theorems for sieve extremum estimates that allow for ill-posed problems.

(2) If $\liminf_k \delta(k) > 0$, then Condition 3.5(iii) is automatically satisfied with $\eta_{k(n)} = o(1)$, Condition 3.5(ii) is implied by Condition 3.5(i), and Condition 3.3(ii) is implied by Condition 3.2 and Condition 3.3(ii)$'$:

CONDITION 3.3(ii)$'$. $Q(\theta)$ is continuous at θ_o in Θ.

(3) Theorem 3.1 is an extension of Corollary 2.6 of White and Wooldridge (1991). Their corollary implies $d(\hat{\theta}_n, \theta_o) = o_P(1)$ under Conditions 3.4, 3.5(i) and Conditions 3.1$'$, 3.2$'$ and 3.3$'$:

CONDITION 3.1$'$.
(i) $Q(\theta)$ is continuous at θ_o in Θ, $Q(\theta_o) > -\infty$;
(ii) for all $\varepsilon > 0$, $Q(\theta_o) > \sup_{\{\theta\in\Theta:\, d(\theta,\theta_o)\geqslant\varepsilon\}} Q(\theta)$.

CONDITION 3.2$'$. $\Theta_k \subseteq \Theta_{k+1} \subseteq \Theta$ for all $k \geqslant 1$; and for any $\theta \in \Theta$ there exists $\pi_k\theta \in \Theta_k$ such that $d(\theta, \pi_k\theta) \to 0$ as $k \to \infty$.

CONDITION 3.3$'$. For each $k \geqslant 1$,
(i) $\widehat{Q}_n(\theta)$ is a measurable function of the data $\{Z_t\}_{t=1}^n$ for all $\theta \in \Theta_k$; and
(ii) for any data $\{Z_t\}_{t=1}^n$, $\widehat{Q}_n(\theta)$ is upper semicontinuous on Θ_k under the metric $d(\cdot,\cdot)$.

We note that under Condition 3.2, Condition 3.1$'$(ii) implies that Condition 3.1(ii) is satisfied with $\delta(k) = const. > 0$, hence Remark 3.1(2) is applicable and $d(\hat{\theta}_n, \theta_o) =$

$o_P(1)$. Unfortunately, Condition 3.1$'$(ii) may fail to be satisfied in some ill-posed semi-nonparametric models when Θ is a noncompact infinite-dimensional parameter space.

(4) Condition 3.1$'$ is satisfied by Condition 3.1$''$:

CONDITION 3.1$''$.

(i) Θ is compact under $d(\cdot,\cdot)$, and $Q(\theta)$ is upper semicontinuous on Θ under $d(\cdot,\cdot)$;
(ii) $Q(\theta)$ is uniquely maximized at θ_o in Θ, $Q(\theta_o) > -\infty$.

As a consequence of Theorem 3.1, we obtain: $d(\hat{\theta}_n, \theta_o) = o_P(1)$ under Conditions 3.1$''$, 3.2, 3.4 and 3.5(i). This result is very similar to Lemmas A.1 in Newey and Powell (2003) and Chernozhukov, Imbens and Newey (2007).

REMARK 3.2. If $\hat{\theta}_n$ satisfies $\widehat{Q}_n(\hat{\theta}_n) \geqslant \sup_{\theta\in\Theta_n} \widehat{Q}_n(\theta) - O_{a.s.}(\eta_n)$, then $d(\hat{\theta}_n, \theta_o) = o_{a.s.}(1)$ under Conditions 3.1–3.4 and Condition 3.5$''$:

CONDITION 3.5$''$.

(i) For all $k \geqslant 1$, $\sup_{\theta\in\Theta_k} |\widehat{Q}_n(\theta) - Q(\theta)| = o_{a.s.}(1)$;
(ii) $\hat{c}(k(n)) = o_{a.s.}(\delta(k(n)))$;
(iii) $\eta_{k(n)} = o(\delta(k(n)))$.

This extends Gallant's (1987) theorem to almost sure convergence of approximate sieve extremum estimates, allowing for noncompact infinite-dimensional Θ and for ill-posed semi-nonparametric models.

Note that when $\Theta_k = \Theta$ is compact, the conditions for Theorem 3.1 become the standard assumptions imposed for consistency of parametric extremum estimation in Newey and McFadden (1994) and White (1994). For semi-nonparametric models, the entire parameter space Θ contains infinite-dimensional unknown functions and is generally noncompact. Nevertheless, one can easily construct compact approximating parameter spaces (sieves) Θ_k. Moreover, it is relatively easy to verify the uniform convergence over compact sieve spaces,[29] while "$\text{plim}_{n\to\infty} \sup_{\theta\in\Theta} |\widehat{Q}_n(\theta) - Q(\theta)| = 0$" may fail when the space Θ is too "large" or too "complex".

We now review some notions of complexity of a function class. Let $L_r(P_o)$, $r \in [1,\infty)$, denote the space of real-valued random variables with finite rth moments and $\|\cdot\|_r$ denote the $L_r(P_o)$-norm. Let $\mathcal{F}_n = \{g(\theta,\cdot)\colon \theta \in \Theta_n\}$ be a class of real-valued, $L_r(P_o)$-measurable functions indexed by $\theta \in \Theta_n$. One notion of complexity of the class $\mathcal{F}_n$ is the $L_r(P_o)$-*covering numbers without bracketing*, which is the minimal number of w-balls $\{\{f\colon \|f - g_j\|_r \leqslant w\},\ \|g_j\|_r < \infty,\ j = 1,\dots,N\}$ that cover $\mathcal{F}_n$, denoted

[29] One could modify the proof of Corollary 2.2 in Newey (1991) or the proof of Lemma 1 in Andrews (1992) to provide sufficient conditions for Condition 3.5(i) in terms of Conditions 3.3(i) and 3.4 and the pointwise convergence over Θ_k.

as $N(w, \mathcal{F}_n, \|\cdot\|_r)$. Likewise, we can define $N(w, \mathcal{F}_n, \|\cdot\|_{n,r})$ as the $L_r(P_n)$-(random) covering numbers without bracketing, where $\|\cdot\|_{n,r}$ denotes the $L_r(P_n)$-norm and P_n denotes the empirical measure of a random sample $\{Z_i\}_{i=1}^n$. Sometimes the covering numbers of $\mathcal{F}_n$ can grow to infinity very fast as n grows; it is then more convenient to measure the complexity of $\mathcal{F}_n$ using the notion of $L_r(P_o)$-*metric entropy without bracketing*, $H(w, \mathcal{F}_n, \|\cdot\|_r) \equiv \log(N(w, \mathcal{F}_n, \|\cdot\|_r))$, and the $L_r(P_n)$-(random) metric entropy without bracketing, $H(w, \mathcal{F}_n, \|\cdot\|_{n,r}) \equiv \log(N(w, \mathcal{F}_n, \|\cdot\|_{n,r}))$. Detailed discussions of metric entropy can be found in Pollard (1984), Andrews (1994a), van der Vaart and Wellner (1996) and van de Geer (2000).

When the function class Θ is too complex in terms of its metric entropy being too large, then the uniform convergence over the entire parameter space Θ may fail, but the uniform convergence over a sieve space Θ_n (i.e., Condition 3.5(i)) can still be satisfied. For example, when $\widehat{Q}_n(\theta) = n^{-1}\sum_{t=1}^n l(\theta, Z_t)$ and $\{Z_t\}_{t=1}^n$ is i.i.d., $E\{\sup_{\theta\in\Theta_n} |l(\theta, Z_t)|\} < \infty$, then Condition 3.5(i) is satisfied if and only if $H(w, \{l(\theta,\cdot)\colon \theta \in \Theta_n\}, \|\cdot\|_{n,1}) = o_P(n)$ for all $w > 0$; see Pollard (1984). When the space Θ is infinite-dimensional and not totally bounded, $H(w, \{l(\theta,\cdot)\colon \theta \in \Theta\}, \|\cdot\|_{n,1}) = O_P(n)$ may occur; hence $\sup_{\theta\in\Theta} |\widehat{Q}_n(\theta) - Q(\theta)| \neq o_P(1)$. For such a case, the extremum estimator obtained by maximizing over the entire parameter space Θ, $\arg\sup_{\theta\in\Theta} \widehat{Q}_n(\theta)$, may fail to exist or be inconsistent.

Conditions 3.1–3.4 of Theorem 3.1 are basic regularity conditions; one can provide more primitive sufficient assumptions for Condition 3.5 in specific applications. In the next remarks we present simple consistency results for sieve M-estimators and sieve MD-estimators. Let $N(w, \Theta_n, d)$ denote the minimal number of w-radius balls (under the metric d) that cover the sieve space Θ_n.

REMARK 3.3 *(Consistency of sieve M-estimator $\hat{\theta}_n = \arg\sup_{\theta\in\Theta_n} n^{-1}\sum_{t=1}^n l(\theta, Z_t) - o_P(1)$).* Suppose that Conditions 3.2 and 3.4 hold, that Condition 3.1 is satisfied with $Q(\theta) = E\{l(\theta, Z_t)\}$ and $\liminf_{k(n)} \delta(k(n)) > 0$, and that $E\{l(\theta, Z_t)\}$ is continuous at $\theta = \theta_o \in \Theta$. Then $d(\hat{\theta}_n, \theta_o) = o_P(1)$ under the following Condition 3.5M:

CONDITION 3.5M.

(i) $\{Z_t\}_{t=1}^n$ is i.i.d., $E\{\sup_{\theta\in\Theta_n} |l(\theta, Z_t)|\}$ is bounded;

(ii) there are a finite $s > 0$ and a random variable $U(Z_t)$ with $E\{U(Z_t)\} < \infty$ such that $\sup_{\theta,\theta'\in\Theta_n\colon d(\theta,\theta')\leqslant\delta} |l(\theta, Z_t) - l(\theta', Z_t)| \leqslant \delta^s U(Z_t)$;

(iii) $\log N(\delta^{1/s}, \Theta_n, d) = o(n)$ for all $\delta > 0$.

Remark 3.3 is a direct consequence of Theorem 3.1 and Pollard's (1984) Theorem II.24. This is because Condition 3.5M(i) and (ii) imply $H(w, \{l(\theta,\cdot)\colon \theta \in \Theta_n\}, \|\cdot\|_{n,1}) \leqslant \log N(\delta^{1/s}, \Theta_n, d)$, hence Condition 3.5M implies Condition 3.5(i). See White and Wooldridge (1991, Theorem 2.5) and Ai and Chen (2007, Lemma A.1) for more general sufficient assumptions for Condition 3.5.

REMARK 3.4 *(Consistency of sieve MD-estimator $\hat{\theta}_n = \arg\inf_{\theta\in\Theta_n} \frac{1}{n}\sum_{t=1}^{n} \hat{m}(X_t,\theta)' \times \{\widehat{\Sigma}(X_t)\}^{-1}\hat{m}(X_t,\theta) + o_P(1)$).* Suppose that Conditions 3.2 and 3.4 hold, that $m(X_t,\theta) \equiv E\{\rho(Z_t,\theta)|X_t\} = 0$ only when $\theta = \theta_o \in \Theta$, that for all X_t, $m(X_t,\theta)$ is continuous in θ_o under the metric $d(\cdot,\cdot)$, and that $\liminf_{k(n)} \delta(k(n)) > 0$. Then $d(\hat{\theta}_n,\theta_o) = o_P(1)$ under the following Condition 3.5MD:

CONDITION 3.5MD.
- (i) $\{Z_t\}_{t=1}^{n}$ is i.i.d., $E\{\sup_{\theta\in\Theta_n} |m(X_t,\theta)'m(X_t,\theta)|\}$ is bounded;
- (ii) there are a finite $s > 0$ and a $U(X_t)$ with $E\{[U(X_t)]^2\} < \infty$ such that $\sup_{\theta,\theta'\in\Theta_n:\, d(\theta,\theta')\leqslant\delta} |m(X_t,\theta) - m(X_t,\theta')| \leqslant \delta^s U(X_t)$;
- (iii) $\log N(\delta^{1/s},\Theta_n,d) = o(n)$ for all $\delta > 0$;
- (iv) uniformly over X_t, $\widehat{\Sigma}(X_t) = \Sigma(X_t) + o_P(1)$ for a positive definite and finite $\Sigma(X_t)$;
- (v) $\frac{1}{n}\sum_{i=1}^{n} |\hat{m}(X_i,\theta) - m(X_i,\theta)|^2 = o_P(1)$ uniformly over $\theta \in \Theta_n$.

See Chen and Pouzo (2006) for a proof of Remark 3.4; they also provide sufficient conditions for the consistency of sieve MD-estimator $\hat{\theta}_n$ without imposing $\liminf_{k(n)} \delta(k(n)) > 0$. Also see Newey and Powell (2003) and Ai and Chen (1999, 2003, 2007) for primitive sufficient conditions for Condition 3.5MD(iv) and (v) where $\widehat{\Sigma}(X_t)$ and $\hat{m}(X_t,\theta)$ are kernel or series estimates of $\Sigma(X_t)$ and $m(X_t,\theta)$, respectively.

Finally, Theorem 3.1 is also applicable to derive convergence of sieve extremum estimates to some pseudo-true values in misspecified semi-nonparametric models; see Lemma 3.1 of Ai and Chen (2007) for such an application.

3.2. Convergence rates of sieve M-estimators

There are many results on convergence rates of sieve M-estimators of unknown functions. For i.i.d. data, Van de Geer (1995) obtained the rate for sieve LS regression. Shen and Wong (1994), and Birgé and Massart (1998) derived the rates for general sieve M-estimation. Van de Geer (1993) and Wong and Shen (1995) obtained the rates for sieve MLE. For time series data, Chen and Shen (1998) derived the rate for sieve M-estimation of stationary beta-mixing models.[30] The general theory on convergence rates is technically involved and relies on the theory of empirical processes. In this section we present a simple version of the rate results for sieve M-estimation whose conditions are easy to verify. However, readers who are interested in the most general theory on convergence rates of sieve M-estimates are encouraged to read the papers by Shen and Wong (1994), Wong and Shen (1995) and Birgé and Massart (1998).

[30] It is impossible to mention here all the existing results on convergence rates of sieve M-estimates. There are many papers on convergence rates of particular sieves, such as the work on polynomial spline regression and density estimation by Stone and his collaborators, see Subsection 3.3 for details; the work on wavelets by Donoho, Johnstone and others [see e.g., Donoho et al. (1995)]; the work on neural networks by Barron (1993), White (1990) and others.

Recall $\theta_o \in \Theta$ and that the approximate sieve M-estimate $\hat{\theta}_n$ solves:

$$n^{-1}\sum_{t=1}^{n} l(\hat{\theta}_n, Z_t) \geqslant \sup_{\theta\in\Theta_n} n^{-1}\sum_{t=1}^{n} l(\theta, Z_t) - O_P\left(\varepsilon_n^2\right) \quad \text{with } \varepsilon_n \to 0. \tag{3.1}$$

Let $d(\theta_o, \theta)$ be a (pseudo-) metric on Θ such that $d(\theta_o, \hat{\theta}_n) = o_P(1)$. Let $K(\theta_o, \theta) \equiv E(l(\theta_o, Z_t) - l(\theta, Z_t))$.[31] Let $\|\theta_o - \theta\|$ be a metric on Θ such that $\|\theta_o - \theta\| \leqslant const.d(\theta_o, \theta)$ for all $\theta \in \Theta$, and $\|\theta_o - \theta\| \asymp K^{1/2}(\theta_o, \theta)$ for $\theta \in \Theta$ with $d(\theta_o, \theta) = o(1)$. We shall give a convergence rate for sieve estimate $\hat{\theta}_n$ under $\|\theta_o - \theta\|$, and thus automatically give an upper bound on $\bar{d}(\theta_o, \hat{\theta}_n)$, where $\bar{d}$ is any other metric on Θ satisfying $\bar{d}(\theta_o, \theta) \leqslant const.K^{1/2}(\theta_o, \theta)$.

In order for $\hat{\theta}_n$ to converge to θ_o at a fast rate under the metric $\|\theta_o - \hat{\theta}_n\|$, not only does the sieve approximation error rate, $\|\theta_o - \pi_n\theta_o\|$, have to approach zero suitably fast, but additionally, the sieve space, Θ_n, must not be too complex. We have already introduced $L_r(P_o)$-covering numbers (metric entropy) without bracketing as a complexity measure of a class $\mathcal{F}_n = \{g(\theta, \cdot)\colon \theta \in \Theta_n\}$, we now consider another measure of complexity. Let $\mathcal{L}_r$ be the completion of $\mathcal{F}_n$ under the norm $\|\cdot\|_r$. For any given $w > 0$, if there exists a collection of functions (brackets) $\{g_1^l, g_1^u, \ldots, g_N^l, g_N^u\} \subset \mathcal{L}_r$ such that $\max_{1\leqslant j\leqslant N}\|g_j^u - g_j^l\|_r \leqslant w$ and for any $g \in \mathcal{F}_n$, there exists $j \in \{1, \ldots, N\}$ with $g_j^l \leqslant g \leqslant g_j^u$ a.e.-P_o, then the minimal number of such brackets, $N_{[]}(w, \mathcal{F}_n, \|\cdot\|_r) \equiv \min(N\colon \{g_1^l, g_1^u, \ldots, g_N^l, g_N^u\})$, is called the *$L_r(P_o)$-covering numbers with bracketing*. Likewise, $H_{[]}(w, \mathcal{F}_n, \|\cdot\|_r) \equiv \log(N_{[]}(w, \mathcal{F}_n, \|\cdot\|_r))$ is called the *$L_r(P_o)$-metric entropy with bracketing* of the class $\mathcal{F}_n$. See Pollard (1984), Andrews (1994a), Van der Vaart and Wellner (1996) and Van de Geer (2000) for more details.

We now present a result of Chen and Shen (1996) for i.i.d. data; see Chen and Shen (1998) for the stationary beta-mixing case and Chen and White (1999) for the stationary uniform-mixing case.[32]

CONDITION 3.6. $\{Z_t\}_{t=1}^n$ is an i.i.d. or m-dependent sequence.

CONDITION 3.7. There is $C_1 > 0$ such that for all small $\varepsilon > 0$,

$$\sup_{\{\theta\in\Theta_n\colon \|\theta_o-\theta\|\leqslant\varepsilon\}} \operatorname{Var}\big(l(\theta, Z_t) - l(\theta_o, Z_t)\big) \leqslant C_1\varepsilon^2.$$

CONDITION 3.8. For any $\delta > 0$, there exists a constant $s \in (0, 2)$ such that

$$\sup_{\{\theta\in\Theta_n\colon \|\theta_o-\theta\|\leqslant\delta\}} \big|l(\theta, Z_t) - l(\theta_o, Z_t)\big| \leqslant \delta^s U(Z_t),$$

with $E([U(Z_t)]^\gamma) \leqslant C_2$ for some $\gamma \geqslant 2$.

[31] If the criterion is a log-likelihood, then $K(\theta_o, \theta)$ is simply the Kullback–Leibler information.

[32] See Fan and Yao (2003) for description of various nonparametric methods applied to nonlinear time series models.

Conditions 3.6 and 3.7 imply that, within a neighborhood of θ_o,

$$\operatorname{Var}\left(n^{-1/2}\sum_{t=1}^{n}\bigl(l(\theta, Z_t) - l(\theta_o, Z_t)\bigr)\right)$$

behaves like $\|\theta_o - \theta\|^2$. Condition 3.8 implies that, when restricting to a local neighborhood of θ_o, $l(\theta, Z_t)$ is "continuous" at θ_o with respect to a metric $\|\theta_o - \theta\|$, which is locally equivalent to $K^{1/2}$. Conditions 3.7 and 3.8 are usually easily verifiable by exploiting the specific form of the criterion function.

Denote $\mathcal{F}_n = \{l(\theta, Z_t) - l(\theta_o, Z_t)\colon \|\theta_o - \theta\| \leqslant \delta,\ \theta \in \Theta_n\}$, and for some constant $b > 0$, let[33]

$$\delta_n = \inf\left\{\delta \in (0,1)\colon \frac{1}{\sqrt{n}\delta^2}\int_{b\delta^2}^{\delta}\sqrt{H_{[\,]}\bigl(w, \mathcal{F}_n, \|\cdot\|_2\bigr)}\,\mathrm{d}w \leqslant const.\right\}.$$

To calculate δ_n, an upper bound on $H_{[\,]}(w, \mathcal{F}_n, \|\cdot\|_2)$ is often enough, and, fortunately for us, much of the work has already been done. For instance, according to Lemma 2.1 of Ossiander (1987) we have that, $H_{[\,]}(w, \mathcal{F}_n, \|\cdot\|_2) \leqslant H(w, \mathcal{F}_n, \|\cdot\|_\infty)$. Moreover, Condition 3.8 implies that

$$H_{[\,]}\bigl(w, \mathcal{F}_n, \|\cdot\|_2\bigr) \leqslant \log N\bigl(w^{1/s}, \Theta_n, \|\cdot\|\bigr).$$

For finite-dimensional linear sieves such as those listed in Subsection 2.3.1 we have $\log N(\epsilon, \Theta_n, \|\cdot\|) \leqslant const. \dim(\Theta_n)\log(\frac{1}{\epsilon})$ [see e.g. Chen and Shen (1998)]; and for neural network and ridgelet nonlinear sieves we have $\log N(\epsilon, \Theta_n, \|\cdot\|) \leqslant const. \dim(\Theta_n)\log(\frac{\dim(\Theta_n)}{\epsilon})$ [see e.g. Chen and White (1999)].

THEOREM 3.2. *Let $\hat{\theta}_n$ be the approximate sieve M-estimator defined by* (3.1). *If Conditions* 3.6–3.8 *hold, then*

$$\|\theta_o - \hat{\theta}_n\| = O_P(\varepsilon_n), \quad \text{with } \varepsilon_n = \max\bigl\{\delta_n, \|\theta_o - \pi_n\theta_o\|\bigr\}.$$

We note that δ_n increases with the complexity of the sieve Θ_n and can be interpreted as a measure of the standard deviation term, while the deterministic approximation error $\|\theta_o - \pi_n\theta_o\|$ decreases with the complexity of the sieve Θ_n and is a measure of the bias. The best convergence rate can be obtained by choosing the complexity of the sieve Θ_n such that $\delta_n \asymp \|\theta_o - \pi_n\theta_o\|$.

Chen and Shen (1998) have demonstrated how to apply the time series version of this theorem with three examples: first, they considered a multivariate nonparametric regression with either a neural network sieve, a wavelet sieve or a spline sieve; second, a partially additive time series model via spline and Fourier series sieves; and third,

[33] There is a typo in Chen and Shen (1998, p. 297), where the "sup" in the definition of δ_n should be replaced by the "inf". Nevertheless, all the other calculations of δ_n in Chen and Shen (1998) are correct.

a transformation model with an unknown link via a monotone spline sieve. Chen and White (1999) considered a time series nonparametric conditional quantile regression via neural network sieve and multivariate conditional density estimation via neural network sieve. Chen and Conley (2001) applied this theorem to a varying coefficient VAR model with a flexible spatial conditional covariance. In the following we illustrate the verification of the conditions of Theorem 3.2 with two examples.

3.2.1. *Example: Additive mean regression with a monotone constraint*

Suppose that the i.i.d. data $\{Y_t, X_t' = (X_{1t}, \dots, X_{qt})\}_{t=1}^n$ are generated according to

$$Y_t = h_{o1}(X_{1t}) + \cdots + h_{oq}(X_{qt}) + e_t, \qquad E[e_t|X_t] = 0.$$

Let $\theta_o = (h_{o1}, \dots, h_{oq})' \in \Theta = \mathcal{H}$ be the parameters of interest with $\mathcal{H} = \mathcal{H}^1 \times \cdots \times \mathcal{H}^q$ to be specified in Assumption 3.1. For simplicity, we assume that $\dim(X_j) = 1$ for $j = 1, \dots, q$, $\dim(X) = q$ and $\dim(Y) = 1$. We estimate the regression function $\theta_o(X) = \sum_{j=1}^q h_{oj}(X_{jt})$ by maximizing over a sieve $\Theta_n = \mathcal{H}_n$ the criterion $\widehat{Q}_n(\theta) = n^{-1} \sum_{t=1}^n l(\theta, Z_t)$, where $l(\theta, Z_t) = -(1/2)[Y_t - \sum_{j=1}^q h_j(X_{jt})]^2$ and $Z_t = (Y_t, X_t')'$. Let $\|\theta - \theta_o\|^2 = E(\theta(X_t) - \theta_o(X_t))^2 = E\{\sum_{j=1}^q [h_j(X_{jt}) - h_{oj}(X_{jt})]\}^2$.

ASSUMPTION 3.1.
(i) $h_{o1} \in \mathcal{H}^1 = C([b_{11}, b_{21}]) \cap \{h\colon \text{nondecreasing}\}$;
(ii) for $j = 2, \dots, q$, $h_{oj} \in \mathcal{H}^j = \Lambda_{c_j}^{p_j}([b_{1j}, b_{2j}])$ with $p_j > 1/2$; and $h_{oj}(x_j^*) = 0$ for some known $x_j^* \in (b_{1j}, b_{2j})$.

ASSUMPTION 3.2. $\sigma^2(X) \equiv E[e^2|X]$ is bounded.

Assumption 3.1(ii) is sufficient for identification, and Assumption 3.2 is a simple regularity condition that has been imposed in many papers; see e.g. Newey (1997).

The sieve will be chosen to have the form $\mathcal{H}_n = \mathcal{H}_n^1 \times \cdots \times \mathcal{H}_n^q$. First we let $\mathcal{H}_n^1$ be a shape-preserving sieve such as the monotone spline wavelet sieve MSplWav$(r_1 - 1, 2^{J_{1n}})$ with $r_1 \geqslant 1$ and $k_{1n} = 2^{J_{1n}}$ in Subsection 2.3.5. For $j = 2, \dots, q$, we let $\mathcal{H}_n^j = \{h_j \in \Theta_{jn}\colon h_j(x_j^*) = 0,\ \|h_j\|_\infty \leqslant c_j\}$ where Θ_{jn} can be any of the finite-dimensional linear sieve examples in Subsection 2.3.1 such as $\Theta_{jn} = \mathrm{Pol}(k_{jn})$ or TriPol(k_{jn}) or Spl(r_j, k_{jn}) with $r_j \geqslant [p_j] + 1$, or Wav$(m_j, 2^{J_{jn}})$ with $m_j > p_j$ and $k_{jn} = 2^{J_{jn}}$.

In the following result we denote $p_1 = 1$ and $p = \min\{p_1, p_2, \dots, p_q\}$.

PROPOSITION 3.3. *Let $\hat{\theta}_n$ be the sieve M-estimate. Suppose that Assumptions 3.1 and 3.2 hold. Let $k_{jn} = O(n^{1/(2p_j+1)})$ for $j = 1, \dots, q$. Then $\|\hat{\theta}_n - \theta_o\| = O_P(n^{-p/(2p+1)})$ with $p = \min\{p_1, \dots, p_q\}$.*

PROOF. Theorem 3.2 is readily applicable to prove this result. It is easy to see that $K(\theta_o, \theta) \asymp \|\theta - \theta_o\|^2$. Condition 3.6 is assumed. Now we check Conditions 3.7 and 3.8.

Since $l(\theta, Z_t) - l(\theta_o, Z_t) = (\theta - \theta_o)[e_t + (\theta_o - \theta)/2]$, we have

$$\begin{aligned}E\big[l(\theta, Z_t) - l(\theta_o, Z_t)\big]^2 &\leqslant 2E\big(\sigma^2(X_t)\big[\theta_o(X_t) - \theta(X_t)\big]^2\big) \\ &\quad + (1/2)E\big(\big[\theta_o(X_t) - \theta(X_t)\big]^4\big) \\ &\leqslant const.\|\theta - \theta_o\|^2 + (1/2)E\big(\big[\theta_o(X_t) - \theta(X_t)\big]^4\big).\end{aligned}$$

By Theorem 1 of Gabushin (1967) when p is an integer and Lemma 2 in Chen and Shen (1998) for any $p > 0$, we have $\|\theta - \theta_o\|_\infty \leqslant c\|\theta - \theta_o\|^{2p/(2p+1)}$. Hence

$$\begin{aligned}E\big(\big[\theta_o(X_t) - \theta(X_t)\big]^4\big) &\leqslant \sup_x\big[\theta(x) - \theta_o(x)\big]^2 E\big(\big[\theta_o(X_t) - \theta(X_t)\big]^2\big) \\ &\leqslant C\|\theta - \theta_o\|^{2(1+[2p/(2p+1)])}.\end{aligned}$$

So Condition 3.7 is satisfied for all $\varepsilon \leqslant 1$. On the other hand,

$$\big|l(\theta, Z_t) - l(\theta_o, Z_t)\big| \leqslant \|\theta - \theta_o\|_\infty\big[|e_t| + \big(\|\theta_o\|_\infty + \|\theta\|_\infty\big)/2\big] \quad \text{a.s.}$$

Using Lemma 2 in Chen and Shen (1998) we see that Condition 3.8 is then satisfied with $s = 2p/(2p+1)$, $U(Z_t) = |e_t| + const.$ and $\gamma = 2$.

To apply Theorem 3.2, it remains to compute the deterministic approximation error rate $\|\theta_o - \pi_n\theta_o\|$ and the metric entropy with bracketing $H_{[\,]}(w, \mathcal{F}_n, \|\cdot\|_2)$ of the class $\mathcal{F}_n = \{l(\theta, Z_t) - l(\theta_o, Z_t)\colon\ \|\theta - \theta_o\| \leqslant \delta,\ \theta \in \Theta_n\}$. By definition, $\|\theta_o - \pi_n\theta_o\| \leqslant const.\max\{\|h_{oj} - \pi_n h_{oj}\|_\infty\colon\ j = 1, \ldots, q\}$. Let $C = \sqrt{E\{U(Z_t)^2\}}$, then for all $0 < \frac{w}{C} \leqslant \delta < 1$, $H_{[\,]}(w, \mathcal{F}_n, \|\cdot\|_2) \leqslant \sum_{j=1}^q \log N(\frac{w}{C}, \mathcal{H}_n^j, \|\cdot\|_\infty)$.

The final bit of calculation now depends on the choice of sieves. First, $\|h_{o1} - \pi_n h_{o1}\|_\infty = O((k_{1n})^{-1})$ by Anastassiou and Yu (1992a); and for $j = 2, \ldots, q$, $\mathcal{H}^j = \Lambda_{c_j}^{p_j}$, $\|h_{oj} - \pi_n h_{oj}\|_\infty = O((k_{jn})^{-p_j})$ by Lorentz (1966). Second, for all $j = 1, 2, \ldots, q$, $\log N(\frac{w}{C}, \mathcal{H}_n^j, \|\cdot\|_\infty) \leqslant const. \times k_{jn} \times \log(1 + \frac{4c_j}{w})$ by Lemma 2.5 in van de Geer (2000). Hence δ_n solves

$$\begin{aligned}&\frac{1}{\sqrt{n}\delta_n^2}\int_{b\delta_n^2}^{\delta_n}\sqrt{H_{[\,]}\big(w, \mathcal{F}_n, \|\cdot\|_2\big)}\,\mathrm{d}w \\ &\quad\leqslant \frac{1}{\sqrt{n}\delta_n^2}\max_{j=1,\ldots,q}\int_{b\delta_n^2}^{\delta_n}\sqrt{k_{jn} \times \log\left(1 + \frac{4c_j}{w}\right)}\,\mathrm{d}w \\ &\quad\leqslant \frac{1}{\sqrt{n}\delta_n^2}\max_{j=1,\ldots,q}\sqrt{k_{jn}} \times \delta_n \leqslant const.\end{aligned}$$

and the solution is $\delta_n \asymp \max_{j=1,\ldots,q}\sqrt{\frac{k_{jn}}{n}}$. By Theorem 3.2, $\|\hat{\theta}_n - \theta_o\| = O_P(\max_{j=1,\ldots,q}\{(k_{jn})^{-p_j}, \delta_n\})$. With the choice of $k_{jn} = O(n^{1/(2p_j+1)})$ for $j = 1, \ldots, q$, we obtain $\|\hat{\theta}_n - \theta_o\| = O_P(n^{-p/(2p+1)})$ with $p = \min\{p_1, \ldots, p_q\} > 0.5$. This immediately implies $\|\hat{h}_j - h_{oj}\|_2 = O_P(n^{-p/(2p+1)})$ for $j = 1, \ldots, q$. □

REMARK 3.5. (1) Since the parameter space $\mathcal{H} = \mathcal{H}^1 \times \cdots \times \mathcal{H}^q$ specified in Assumption 3.1 is compact with respect to the norm $\|\cdot\|$, we can take the original parameter space $\mathcal{H}$ as the sieve space $\mathcal{H}_n$. Applying Theorem 3.2 again, note that the approximation error $\|\pi_n\theta_o - \theta_o\| = 0$, we have $\|\hat{\theta}_n - \theta_o\| = O_P(\delta_n)$, where δ_n solves:

$$\begin{aligned}
&\frac{1}{\sqrt{n}\delta_n^2}\int_{b\delta_n^2}^{\delta_n}\sqrt{\sum_{j=1}^{q}\log N\big(w, \mathcal{H}^j, \|\cdot\|_\infty\big)}\,\mathrm{d}w \\
&\quad\leqslant \frac{1}{\sqrt{n}\delta_n^2}\int_{b\delta_n^2}^{\delta_n}\sqrt{\sum_{j=1}^{q}\left(\frac{c_j}{w}\right)^{1/p_j}}\,\mathrm{d}w \quad \text{by Birman and Solomjak (1967)} \\
&\quad\leqslant \frac{1}{\sqrt{n}\delta_n^2}\max_{j=1,\ldots,q} const.(\delta_n)^{1-\frac{1}{2p_j}} \leqslant const.
\end{aligned}$$

which is satisfied if $\delta_n = O(n^{-p/(2p+1)})$ with $p = \min\{p_1, \ldots, p_q\} > 0.5$. However, it is unclear how one can implement such an optimization over the entire parameter space $\mathcal{H}$ given a finite data set.

(2) Suppose that in Proposition 3.3 we replace Assumption 3.1(i) by $h_{o1} \in \Lambda_{c_1}^{p_1}([b_{11}, b_{21}])$ and let $\mathcal{H}_n^1 = \mathrm{Pol}(k_{1n})$, or $\mathrm{TriPol}(k_{1n})$, or $\mathrm{Spl}(r_1, k_{1n})$ with $r_1 \geqslant [p_1]+1$, or $\mathrm{Wav}(m_1, 2^{J_{1n}})$ with $m_1 > p_1, 2^{J_{1n}} = k_{1n}$. Let $p = \min\{p_1, \ldots, p_q\} > 0.5$. Then we have $\|\hat{h}_j - h_{oj}\|_2 = O_P(n^{-p/(2p+1)})$ for $j = 1, \ldots, q$. Further, let $\|D^m\hat{h}_j - D^m h_{oj}\|_2 = \{E[D^m\hat{h}_j(X_{jt}) - D^m h_{oj}(X_{jt})]^2\}^{1/2}$ for an integer $m \geqslant 1$. If $p > m \geqslant 1$ then $\|D^m\hat{h}_j - D^m h_{oj}\|_2 = O_P(k_{jn}^{-(p-m)}) = O_P(n^{-(p-m)/(2p+1)})$ for $j = 1, \ldots, q$. This convergence rate achieves the optimal one derived in Stone (1982).

3.2.2. *Example: Multivariate quantile regression*

Suppose that the i.i.d. data $\{Y_t, X_t\}_{t=1}^n$ are generated according to

$$Y_t = \theta_o(X_t) + e_t, \qquad P[e_t \leqslant 0|X_t] = \alpha \in (0,1),$$

where $X_t \in \mathcal{X} = \mathcal{R}^d$, $d \geqslant 1$. We estimate the conditional quantile function $\theta_o(\cdot)$ by maximizing over Θ_n the criterion $\widehat{Q}_n(\theta) = n^{-1}\sum_{t=1}^n l(\theta, Y_t, X_t)$, where $l(\theta, Y_t, X_t) = \{1(Y_t < \theta(X_t)) - \alpha\}[Y_t - \theta(X_t)]$. Let $\|\theta - \theta_o\|^2 = E(\theta(X_t) - \theta_o(X_t))^2$ and $W_1^1(\mathcal{X})$ be the Sobolev space defined in Subsection 2.3.3.

ASSUMPTION 3.3. $\theta_o \in \Theta = W_1^1(\mathcal{X})$.

ASSUMPTION 3.4. Let $f_{e|X}$ be the conditional density of e_t given X_t satisfying $0 < \inf_{x\in\mathcal{X}} f_{e|X=x}(0) \leqslant \sup_{x\in\mathcal{X}} f_{e|X=x}(0) < \infty$ and $\sup_{x\in\mathcal{X}} |f_{e|X=x}(z) - f_{e|X=x}(0)| \to 0$ as $|z| \to 0$.

It is known that the tensor product of finite-dimensional linear sieves such as those in Subsection 2.3.1 will not be able to approximate functions in $W_1^m(\mathcal{X})$, $m \geqslant 1$, well, hence the sieve convergence rates based on those linear sieves will be slower than those based on nonlinear sieves; see e.g. Chen and Shen (1998, Proposition 1, Case 1.3(ii)) for such an example. For time series regression models, Chen and White (1999), Chen, Racine and Swanson (2001) have shown that neural network sieves lead to faster convergence rates for functions in $W_1^m(\mathcal{X})$. Thus we consider the following Gaussian radial basis ANN sieve Θ_n for the unknown $\theta_o \in W_1^1(\mathcal{X})$:

$$\Theta_n = \left\{\alpha_0 + \sum_{j=1}^{k_n} \alpha_j G\left(\frac{\{(x-\gamma_j)'(x-\gamma_j)\}^{1/2}}{\sigma_j}\right),\right.$$
$$\left.\sum_{j=0}^{k_n} |\alpha_j| \leqslant c_0,\ |\gamma_j| \leqslant c_1,\ 0 < \sigma_j \leqslant c_2\right\},$$

where G is the standard Gaussian density function.

PROPOSITION 3.4. *Let $\hat{\theta}_n$ be the sieve M-estimate. Suppose that Assumptions 3.3 and 3.4 hold. Let $k_n^{2(1+1/(d+1))}\log(k_n) = O(n)$. Then*

$$\|\hat{\theta}_n - \theta_o\| = O_P\big([n/\log n]^{-(1+2/(d+1))/[4(1+1/(d+1))]}\big).$$

PROOF. Theorem 3.2 is readily applicable to prove this result. Condition 3.6 is directly assumed. By the above assumptions on conditional density $f_{e|X}$, it is easy to check that $K(\theta_o, \theta) \asymp E(\theta(X_t) - \theta_o(X_t))^2$; see Chen and White (1999, pp. 686–687) for details. Now let us check Conditions 3.7 and 3.8. Note that $|l(\theta, Y_t, X_t) - l(\theta_o, Y_t, X_t)| \leqslant \max(\alpha, 1-\alpha)|\theta(X_t) - \theta_o(X_t)|$, we have

$$\begin{aligned}\operatorname{Var}\big(l(\theta, Y_t, X_t) - l(\theta_o, Y_t, X_t)\big) &\leqslant E\big[l(\theta, Y_t, X_t) - l(\theta_o, Y_t, X_t)\big]^2 \\ &\leqslant E\big[\theta(X_t) - \theta_o(X_t)\big]^2,\end{aligned}$$

and thus Condition 3.7 is satisfied. Moreover, we have

$$\sup_{\{\theta\in\Theta_n:\ \|\theta-\theta_o\|\leqslant\delta\}} \big|l(\theta, Y_t, X_t) - l(\theta_o, Y_t, X_t)\big| \leqslant \sup_{\{\theta\in\Theta_n:\ \|\theta-\theta_o\|\leqslant\delta\}} \big|\theta(X_t) - \theta_o(X_t)\big|,$$

and $\|\theta - \theta_o\|_\infty \leqslant c\|\theta - \theta_o\|^{2/3}$ by Theorem 1 of Gabushin (1967). Hence, Condition 3.8 is satisfied with $s = 2/3$, $U(X_t) \equiv c$.

Now by results in Chen, Racine and Swanson (2001),

$$\|\theta_o - \pi_n\theta_o\| \leqslant const.(k_n)^{-1/2-1/(d+1)}$$

and $\log N(w, \Theta_n, \|\cdot\|_\infty) \leqslant const.k_n \log(\frac{k_n}{w})$. With $k_n^{2(1+1/(d+1))}\log(k_n) = O(n)$, it is easy to see that $\|\hat{\theta}_n - \theta_o\| = O_P([n/\log n]^{-(1+2/(d+1))/[4(1+1/(d+1))]})$ by applying Theorem 3.2. □

3.3. Convergence rates of series estimators

In this subsection we present the convergence rate of the series estimators for the concave extended linear models. Recall that in this framework, the parameter space, Θ, is a linear space which is often a subspace of the space of square integrable functions, the sample criterion function $\widehat{Q}_n(\theta) = n^{-1}\sum_{i=1}^n l(\theta, Z_i)$ is concave in $\theta \in \Theta$ almost surely and the population criterion function $Q(\theta) = E[l(\theta, Z_i)]$ is strictly concave in $\theta \in \Theta$. The results reported here are largely based on those of Huang (1998a, 2001) and Newey (1997).

Throughout this subsection, $\{Z_i\}_{i=1}^n$ is i.i.d. and θ denotes a real-valued function with a bounded domain, $\mathcal{X} \subset \mathcal{R}^d$. We use $\|\hat{\theta} - \theta_o\|$ to measure the discrepancy between $\hat{\theta}$ and θ_o.

CONDITION 3.9. $\|\theta\| \asymp \|\theta\|_{2,leb}$ for any Lebesgue square-integrable function θ.

In the multivariate LS regression of Example 2.4, $\theta_o(X) = E[Y|X]$, a natural choice for the norm is $\|\theta\| = \|\theta\|_2 = \{E[\theta(X)^2]\}^{1/2}$. If the density of X is bounded away from zero and infinity, then Condition 3.9 is satisfied. In general a natural choice of the norm, $\|\cdot\|$, will depend on the specific application and on the data generating process.

We impose the following condition on the linear sieve space.

CONDITION 3.10. The finite-dimensional linear sieve space, Θ_n, is theoretically identifiable in the sense that any $\theta \in \Theta_n$ with $\|\theta\| = 0$ implies that $\theta(u) = 0$ everywhere.

Under Condition 3.9, Condition 3.10 is trivially satisfied by commonly used linear approximation spaces such as those given in Subsection 2.3.1.

CONDITION 3.11. $\theta_o = \arg\max_{\Theta} E[l(\theta, Z)]$ satisfies $\|\theta_o\|_\infty \leqslant K_o < \infty$.

CONDITION 3.12. For any bounded functions $\theta_1, \theta_2 \in \Theta$, $E[l(\theta_1 + \tau(\theta_2 - \theta_1), Z)]$ is twice continuously differentiable with respect to $\tau \in [0, 1]$. For any constant $0 < K < \infty$, $\frac{\partial^2}{\partial\tau^2} E[l(\theta_1 + \tau(\theta_2 - \theta_1), Z)] \asymp -\|\theta_2 - \theta_1\|^2$ for $\theta_1, \theta_2 \in \Theta$ with $\|\theta_1\|_\infty \leqslant K$ and $\|\theta_2\|_\infty \leqslant K$ and $0 \leqslant \tau \leqslant 1$.

Given the above conditions, we can define $\bar{\theta}_n \equiv \arg\max_{\theta \in \Theta_n} E[l(\theta, Z)]$, and it is easy to see that $\|\bar{\theta}_n - \theta_o\| \asymp \inf_{\theta \in \Theta_n} \|\theta - \theta_o\|$.

CONDITION 3.13. For any pair of functions $\theta_1, \theta_2 \in \Theta_n$, $l(\theta_1 + \tau(\theta_2 - \theta_1), Z)$ is twice continuously differentiable with respect to τ. Moreover,

(i)

$$\sup_{g \in \Theta_n} \frac{|\frac{\partial}{\partial\tau} l(\bar{\theta}_n + \tau g, Z)|_{\tau=0}|}{\|g\|} = O_P\left(\sqrt{\frac{\dim(\Theta_n)}{n}}\right);$$

(ii) for any constant $0 < K < \infty$, there is a $c > 0$ such that $\frac{\partial^2}{\partial \tau^2} l(\theta_1 + \tau(\theta_2 - \theta_1), Z) \leqslant -c\|\theta_2 - \theta_1\|^2$ for any $\theta_1, \theta_2 \in \Theta_n$ with $\|\theta_1\|_\infty \leqslant K$ and $\|\theta_2\|_\infty \leqslant K$ and $0 \leqslant \tau \leqslant 1$, except on an event whose probability tends to zero as $n \to \infty$.

Denote $k_n = \dim(\Theta_n)$, $A_n \equiv \sup_{\theta \in \Theta_n, \|\theta\|_{2,leb} \neq 0}(\|\theta\|_\infty / \|\theta\|_{2,leb})$ and $\rho_{2n} \equiv \inf_{\theta \in \Theta_n} \|\theta - \theta_o\|_{2,leb}$. Under Conditions 3.9–3.11, we have $\rho_{2n} \asymp \inf_{\theta \in \Theta_n} \|\theta - \theta_o\|$. The following result is a special case of Huang (2001) for the sieve estimator of a concave extended linear model.

THEOREM 3.5. *Suppose Conditions 3.9–3.13 hold. Let $\lim_{n\to\infty} A_n \rho_{2n} = 0$ and $\lim_{n\to\infty} A_n^2 k_n / n = 0$. Then the series estimator, $\hat{\theta}$, exists uniquely with probability approaching one as $n \to \infty$, and*

$$\|\hat{\theta} - \theta_o\| = O_P\left(\sqrt{\frac{k_n}{n}} + \rho_{2n}\right).$$

This theorem could be regarded as a special case of Theorem 3.2 by taking $\delta_n \asymp \sqrt{\frac{k_n}{n}}$ and $\|\pi_n \theta_o - \theta_o\| \asymp \rho_{2n}$. To see this, first note that under Conditions 3.9–3.11 there is an essentially unique element $\pi_n \theta_o \in \Theta_n$ such that $\|\pi_n \theta_o - \theta_o\| = \inf_{\theta \in \Theta_n} \|\theta - \theta_o\|$, and $\|\pi_n \theta_o - \theta_o\| \asymp \|\pi_n \theta_o - \theta_o\|_{2,leb} \asymp \rho_{2n}$, which is the approximation error rate. Second, within the framework of concave extended linear models, for a finite-dimensional linear sieve Θ_n we have $\log N(w, \Theta_n, \|\cdot\|_\infty) \leqslant const.k_n \log(\frac{1}{w})$, hence $\delta_n \asymp \sqrt{\frac{k_n}{n}}$.

The constant $A_n \geqslant 1$ is a measure of irregularity of the finite-dimensional linear sieve space, Θ_n. Since we require that Θ_n be theoretically identifiable and functions in Θ_n be bounded, A_n is finite. In fact, let $\{\phi_j, j = 1, \ldots, k_n\}$ be an orthonormal basis of Θ_n relative to the theoretical inner product. Then, by the Cauchy–Schwarz inequality, $A_n \leqslant \{\sum_{j=1}^{k_n} \|\phi_j\|_\infty^2\}^{1/2} < \infty$. It is obvious that $\|\theta\|_\infty \leqslant A_n \|\theta\|_{2,leb}$ for all $\theta \in \Theta_n$. The linear sieve spaces are usually chosen to be among commonly used approximating spaces such as those described in Subsection 2.3.1 and the associated constant A_n is readily obtained by using results in the approximation theory literature. Here are some examples.

Polynomials. If $\Theta_n = \mathrm{Pol}(J_n)$ and $\mathcal{X} = [0, 1]$, then $A_n \asymp J_n$ [see Theorem 4.2.6 of DeVore and Lorentz (1993)].

Trigonometric polynomials. If $\Theta_n = \mathrm{TriPol}(J_n)$ and $\mathcal{X} = [0, 1]$, then $A_n \asymp J_n^{1/2}$ [see Theorem 4.2.6 of DeVore and Lorentz (1993)].

Univariate splines. If $\Theta_n = \mathrm{Spl}(r, J_n)$ and $\mathcal{X} = [0, 1]$, then $A_n \asymp J_n^{1/2}$ [see Theorem 5.1.2 of DeVore and Lorentz (1993)].

Orthogonal wavelets. If $\Theta_n = \mathrm{Wav}(m, 2^{J_n})$ and $\mathcal{X} = [0, 1]$, then $A_n \asymp 2^{J_n/2}$ [see Lemma 2.8 of Meyer (1992)].

Tensor product spaces. Let Θ_n be the tensor product of $\Theta_{n1}, \ldots, \Theta_{nd}$. The constant A_n associated with the tensor product linear sieve space, Θ_n, can be determined from the corresponding constants for its components. Set

$$a_{n\ell} = \sup_{\theta \in \Theta_{n\ell},\, \|\theta\|_{2,leb} \neq 0} \left(\|\theta\|_\infty / \|\theta\|_{2,leb}\right)$$

for $1 \leqslant \ell \leqslant d$. It is shown in Huang (1998a) that $A_n \leqslant const. \prod_{\ell=1}^{d} a_{n\ell}$.

We conclude this subsection with an application to the multivariate LS regression of Example 2.4.

ASSUMPTION 3.5.

(i) X has a compact support $\mathcal{X}$ and has a density that is bounded away from zero and infinity on $\mathcal{X}$, where $\mathcal{X} \subset \mathcal{R}^d$ is a Cartesian product of compact intervals $\mathcal{X}_1, \ldots, \mathcal{X}_d$;
(ii) $\mathrm{Var}(Y|X = \cdot)$ is bounded on $\mathcal{X}$;
(iii) $h_o(\cdot) = E[Y|X = \cdot] \in \Lambda^p(\mathcal{X})$ with $p > d/2$.

Theorem 3.5 can treat a general finite-dimensional linear sieve space Θ_n. For simplicity, however, we consider here only the case when the sieve space, Θ_n, in Example 2.4 is constructed as a tensor product space of some commonly used univariate linear approximating spaces $\Theta_{n1}, \ldots, \Theta_{nd}$. Then $k_n = \dim(\Theta_n) = \prod_{\ell=1}^{d} \dim(\Theta_{n\ell})$.

PROPOSITION 3.6. *Suppose Assumption 3.5 holds. Let $\hat{h}_n$ be the series estimate of h_o in Example 2.4, with the sieve, Θ_n, being the tensor-product of the univariate sieve spaces $\Theta_{n1}, \ldots, \Theta_{nd}$. For $\ell = 1, \ldots, d$,*

- *if $\Theta_{n\ell} = \mathrm{Pol}(J_n)$, $p > d$ and $J_n^{3d}/n \to 0$, then $\|\hat{h}_n - h_o\| = O_P(\sqrt{J_n^d/n} + J_n^{-p})$;*
- *if $\Theta_{n\ell} = \mathrm{TriPol}(J_n)$, $p > d/2$ and $J_n^{2d}/n \to 0$, then $\|\hat{h}_n - h_o\| = O_P(\sqrt{J_n^d/n} + J_n^{-p})$;*
- *if $\Theta_{n\ell} = \mathrm{Spl}(r, J_n)$ with $r \geqslant [p] + 1$, $p > d/2$ and $J_n^{2d}/n \to 0$, then $\|\hat{h}_n - h_o\| = O_P(\sqrt{J_n^d/n} + J_n^{-p})$.*

Let $J_n = O(n^{1/(2p+d)})$, then $\|\hat{h}_n - h_o\| = O_P(n^{-p/(2p+d)})$.

We note that this proposition can also be obtained as a direct consequence of Theorem 1 in Newey (1997).[34] The choice of $J_n \asymp n^{1/(2p+d)}$ balances the variance (J_n^d/n) and the squared bias (J_n^{-2p}) trade-off: $J_n^d/n \asymp J_n^{-2p}$. The resulting rate of convergence

[34] Proposition 3.6 is about the convergence rates under $\|\cdot\|_2$-norm for LS regressions. There are also a few results on the convergence rates under $\|\cdot\|_\infty$-norm for LS regressions; see e.g. Stone (1982), Newey (1997) and de Jong (2002).

$n^{-2p/(2p+d)}$ is actually optimal in the context of regression and density estimations: no estimate has a faster rate of convergence uniformly over the class of p-smooth functions [Stone (1982)]. The rate of convergence depends on two quantities: the specified smoothness p of the target function θ_o and the dimension d of the domain on which the target function is defined. Note the dependence of the rate of convergence on the dimension d: given the smoothness p, the larger the dimension, the slower the rate of convergence; moreover, the rate of convergence tends to zero as the dimension tends to infinity. This provides a mathematical description of a phenomenon commonly known as the "curse of dimensionality". Imposing additivity on an unknown multivariate function can imply faster rates of convergence of the corresponding estimate; see Subsection 3.2.1, Stone (1985, 1986), Andrews and Whang (1990), Huang (1998b) and Huang et al. (2000).

3.4. Pointwise asymptotic normality of series LS estimators

To date, we have a relatively complete theory on the rates of convergence for sieve M-estimators. The corresponding asymptotic distribution theory, however, is incomplete and requires much future work. All of the currently available results are for series estimators of densities and the LS regression functions. Asymptotic normality of the series LS estimators has been studied in Andrews (1991b), Gallant and Souza (1991), Newey (1994b, 1997), Zhou, Shen and Wolfe (1998), and Huang (2003). Stone (1990) and Strawderman and Tsiatis (1996) have given asymptotic normality results for polynomial spline estimators in the context of density estimation and hazard estimation, respectively.[35]

We focus on Example 2.4 throughout this subsection. That is, we assume that the data $\{Z_i = (Y_i, X_i')'\}_{i=1}^n$ are i.i.d., and the parameter of interest, $\theta_o(\cdot) = h_o(\cdot) = E[Y|X = \cdot]$, is a real-valued regression function with a bounded domain $\mathcal{X} \subset \mathcal{R}^d$.

3.4.1. Asymptotic normality of the spline series LS estimator

Here we present a result by Huang (2003) on the pointwise asymptotic normality of the spline series LS estimator.

ASSUMPTION 3.6.

(i) $\mathrm{Var}(Y|X = \cdot)$ is bounded away from zero on $\mathcal{X}$;

(ii)

$$\sup_{x\in\mathcal{X}} E\big[\big\{Y - h_o(X)\big\}^2 \times 1\big(\big|Y - h_o(X)\big| > \lambda\big)\big|X = x\big] \to 0 \quad \text{as } \lambda \to \infty.$$

[35] See Portnoy (1997) for a closely related result on the asymptotic normality for smoothing spline quantile estimators.

In the following, $\Phi(\cdot)$ denotes the standard normal distribution function, and $\mathrm{SD}(\hat{h}(x)|X_1, \ldots, X_n) = \{\mathrm{Var}(\hat{h}(x)|X_1, \ldots, X_n)\}^{1/2}$.

THEOREM 3.7. *[See Huang (2003).] Suppose Assumptions 3.5 and 3.6 hold. Let $\hat{h}_n$ be the series estimate of h_o in Example 2.4, with the sieve, Θ_n, being the tensor-product of the univariate spline sieve spaces $\Theta_{n\ell} = \mathrm{Spl}(r, J_n)$, $r \geqslant [p] + 1$, $1 \leqslant \ell \leqslant d$. If $\lim_{n\to\infty} J_n^d \log n/n = 0$ and $\lim_{n\to\infty} J_n/n^{1/(2p+d)} = \infty$, then*

$$\Pr\big(\hat{h}(x) - h_o(x) \leqslant t \times \mathrm{SD}\big(\hat{h}(x)\big|X_1, \ldots, X_n\big)\big) \to \Phi(t), \quad t \in \mathcal{R}.$$

Asymptotic distribution results such as Theorem 3.7 can be used to construct asymptotic confidence intervals. Let $\widehat{\mathrm{SD}}(\hat{h}(x)|X_1, \ldots, X_n)$ be a consistent estimate of $\mathrm{SD}(\hat{h}(x)|X_1, \ldots, X_n)$; see Andrews (1991b) and Newey (1997) for such an estimate. Let $\hat{h}_\alpha^l(x) = \hat{h}(x) - z_{1-\alpha/2}\widehat{\mathrm{SD}}(\hat{h}(x)|X_1, \ldots, X_n)$ and $\hat{h}_\alpha^u(x) = \hat{h}(x) + z_{1-\alpha/2}\widehat{\mathrm{SD}}(\hat{h}(x)|X_1, \ldots, X_n)$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution. If the conditions of Theorem 3.7 hold, then $[\hat{h}_\alpha^l(x), \hat{h}_\alpha^u(x)]$ is an asymptotic $1-\alpha$ confidence interval of $h_o(x)$; that is, $\lim_{n\to\infty} P(\hat{h}_\alpha^l(x) \leqslant h_o(x) \leqslant \hat{h}_\alpha^u(x)) = 1 - \alpha$.

Recall that for the tensor product spline sieve Θ_n, $k_n = \dim(\Theta_n) \asymp J_n^d$. If $h_o(\cdot)$ is p-smooth, then the tensor product spline sieve has the bias order $J_n^{-p} \asymp k_n^{-p/d}$. The condition $\lim_{n\to\infty} J_n/n^{1/(2p+d)} = \infty$ in Theorem 3.7 implies that the bias term is asymptotically negligible relative to the standard deviation of the estimate. Such a condition, $\lim_{n\to\infty} k_n/n^{d/(2p+d)} = \infty$, is usually called undersmoothing (or overfitting); that is, the total number of sieve parameters (k_n) required for undersmoothing is more than what is required to achieve Stone's (1982) optimal rate of convergence.

3.4.2. *Asymptotic normality of functionals of series LS estimator*

We now review the asymptotic normality results in Newey (1997) for any series estimation of functionals of the LS regression function. Let $a : \Theta \to \mathcal{R}$ be a functional, and we want to estimate $a(h_o)$, where $h_o(\cdot) = E[Y|X = \cdot] \in \Theta$. Recall that $\hat{h}(\cdot) = p^{k_n}(\cdot)'(P'P)^- \sum_{i=1}^n p^{k_n}(X_i)Y_i$ is the series LS estimator of $h_o(\cdot)$, with $p^{k_n}(X)$ being the finite-dimensional linear sieve (2.10), see Example 2.4. Then $a(\hat{h})$ will be a natural estimator for $a(h_o)$.

Let $s \geqslant 0$ be an integer, and define a strong norm on Θ as $\|h\|_{s,\infty} = \max_{[\gamma]\leqslant s} \sup_{x\in\mathcal{X}} |D^\gamma h(x)|$. Also, let $\zeta_0(k_n) \equiv \sup_{x\in\mathcal{X}} |p^{k_n}(x)|_e$, $\zeta_s(k_n) \equiv \max_{[\gamma]\leqslant s} \sup_{x\in\mathcal{X}} |D^\gamma p^{k_n}(x)|_e$, where $|\cdot|_e$ is the Euclidean norm.

ASSUMPTION 3.7.

(i) $\mathrm{Var}(Y|X = \cdot)$ is bounded away from zero on $\mathcal{X}$; $\sup_{x\in\mathcal{X}} E[\{Y - h_o(X)\}^4|X = x] < \infty$;

(ii) the smallest eigenvalue of $E[p^{k_n}(X)p^{k_n}(X)']$ is bounded away from zero uniformly in k_n;

(iii) for an integer $s \geqslant 0$ there are $\alpha > 0$, $\beta^*_{k_n}$ such that $\inf_{g \in \Theta_n} \|g - h_o\|_{s,\infty} = \|p^{k_n}(\cdot)'\beta^*_{k_n} - h_o(\cdot)\|_{s,\infty} = O(k_n^{-\alpha})$.

ASSUMPTION 3.8. Either

(i) $\lim_{n\to\infty} k_n\{\zeta_0(k_n)\}^2/n = 0$, and $a(h)$ is linear in $h \in \Theta$; or

(ii) for s as in Assumption 3.7, $\lim_{n\to\infty} k_n^2\{\zeta_s(k_n)\}^4/n = 0$, and there exists a function $D(h; \tilde{h})$ that is linear in $h \in \Theta$ such that for some $c_1, c_2, \varepsilon > 0$ and for all $\tilde{h}, \bar{h}$ with $\|\tilde{h} - h_o\|_{s,\infty} < \varepsilon$, $\|\bar{h} - h_o\|_{s,\infty} < \varepsilon$, it is true that

$$\left|a(h) - a(\tilde{h}) - D(h - \tilde{h}; \tilde{h})\right| \leqslant c_1\left\{\|h - \tilde{h}\|_{s,\infty}\right\}^2; \quad \text{and}$$
$$\left|D(h; \tilde{h}) - D(h; \bar{h})\right| \leqslant c_2\|h\|_{s,\infty}\|\tilde{h} - \bar{h}\|_{s,\infty}.$$

ASSUMPTION 3.9.

(i) There is a positive constant c such that $|D(h; h_o)| \leqslant c\|h\|_{s,\infty}$ for s from Assumption 3.7;

(ii) there is an $h_n \in \Theta_n$ such that $E[h_n(X)^2] \to 0$ but $D(h_n; h_o)$ is bounded away from zero.

Assumption 3.7(iii) is a condition on the sieve approximation error under the strong norm $\|h\|_{s,\infty}$. Assumption 3.8 implies that $a(h)$ is Frechet differentiable in h with respect to the norm $\|h\|_{s,\infty}$. Assumption 3.9 says that the derivative $D(h; h_o)$ is continuous in the norm $\|h\|_{s,\infty}$ but not in the mean-square norm $\|h\|_2 = \{E[h(X)^2]\}^{1/2}$. The lack of mean-square continuity will imply that the estimator $a(\hat{h})$ is not $\sqrt{n}$-consistent for $a(h_o)$; see Newey (1997) for detailed discussions. In the following we denote $\Sigma = E[p^{k_n}(X)p^{k_n}(X)' \operatorname{Var}(Y|X)]$,

$$A = \left.\frac{\partial a(p^{k_n}(X)'\beta)}{\partial \beta}\right|_{\beta^*_{k_n}} \quad \text{and}$$
$$V_{k_n} = A'\left\{E\left[p^{k_n}(X)p^{k_n}(X)'\right]\right\}^{-1}\Sigma\left\{E\left[p^{k_n}(X)p^{k_n}(X)'\right]\right\}^{-1}A.$$

We let $\xrightarrow{d}$ denote convergence in distribution and $\mathcal{N}(0, 1)$ denote a scalar random variable drawn from a standard normal distribution.

THEOREM 3.8. *[See Newey (1997).] Suppose Assumptions 3.5(i)(ii), 3.7–3.9 hold. Let $\hat{h}_n$ be the series estimate of h_o in Example 2.4, with the sieve Θ_n being the linear sieve (2.10). If $\lim_{n\to\infty}\sqrt{n}k_n^{-\alpha} = 0$, then*

$$\sqrt{\frac{n}{V_{k_n}}}\left(a(\hat{h}) - a(h_o)\right) \xrightarrow{d} \mathcal{N}(0, 1).$$

We note that for the linear functional $a(h_o) = h_o(x)$, this theorem implies pointwise asymptotic normality of any series LS estimators $\hat{h}(x)$ satisfying Assumptions 3.5(i)(ii), 3.7, 3.8(i) and 3.9(ii). When we specialize this theorem further to the tensor product spline series estimator of $h_o(x)$, then Assumption 3.8(i) requires $\lim_{n\to\infty} k_n^2/n = 0$, which is stronger than the condition $\lim_{n\to\infty} k_n \log n/n = 0$ in Theorem 3.7. However, Theorem 3.7 is applicable only to the spline series LS estimator, while the results by Newey (1994b, 1997) are much more general.

The normality results reported in this section are only valid for i.i.d. data; see Andrews (1991b) for asymptotic normality of linear functionals of the series LS estimators based on time series dependent observations.

4. Large sample properties of sieve estimation of parametric parts in semiparametric models

In the general sieve extremum estimation framework of Section 2, a model typically contains a parameter vector $\theta = (\beta, h)$, where β is a vector of finite-dimensional parameters and h is a vector of infinite-dimensional parameters. When both β and h are parameters of interest we call the model "semi-nonparametric". When h is a vector of nuisance parameters, then, following Powell (1994) and others, we will call the model "semiparametric".

For weakly dependent observations, semiparametric models can be classified into two categories: (i) β cannot be estimated at a $\sqrt{n}$-rate, i.e., β has zero information bound; see van der Vaart (1991); and (ii) β can be estimated at a $\sqrt{n}$-rate. Models belonging to category (i) should be correctly viewed as nonparametric. However, since these models can still be estimated by the method of sieves, the general sieve convergence rate results can be applied to derive slower than $\sqrt{n}$-rates for the sieve estimates of β. To date there is little research about whether or not the sieve estimate of β can reach the optimal convergence rate and what its limiting distribution is. It is worth mentioning that for Heckman and Singer's (1984) model, Ishwaran (1996a) established that the β-parameters cannot be estimated at $\sqrt{n}$-rate, while Ishwaran (1996b) constructed another estimator of β that converges at the optimal rate but is not a sieve MLE. Prior to the work of Ishwaran (1996a, 1996b), Honoré (1990, 1994) proposed a clever estimator of β that is not a sieve MLE either and computed its convergence rate. It is still an open question whether or not Heckman and Singer's (1984) sieve MLE estimator could reach Ishwaran's optimal rate.[36]

There is a large literature on semiparametric estimation of β for models belonging to category (ii); see Bickel et al. (1993), Newey and McFadden (1994), Powell (1994),

[36] There are other important results in econometrics about specific models belonging to category (i). For example, Manski (1985) proposed a maximum score estimator of a binary choice model with zero median restriction; Kim and Pollard (1990) derived the $n^{1/3}$ consistency of Manski's estimator; Horowitz (1992) proposed a smoothed maximum score estimator for Manski's model, and proved that his smoothed estimator converges faster than $n^{1/3}$ and is asymptotically normal; Andrews and Schafgans (1998) proposed a slower than $\sqrt{n}$ rate kernel estimator of the intercept in Heckman's sample selection model; Honoré and Kyriazidou (2000) developed a slower than $\sqrt{n}$ rate kernel estimator of a discrete choice dynamic panel data model. See Powell (1994), Horowitz (1998), Pagan and Ullah (1999) for more examples.

Horowitz (1998) and Pagan and Ullah (1999) for reviews. Most of these results are derived using the so-called two-step procedure: Step one estimates h nonparametrically by $\hat{h}$, while step two estimates β via either M-estimation, GMM or more generally, MD-estimation with the unknown h replaced by $\hat{h}$. A few general results deal with the simultaneous estimation of β and h. For example, the sieve simultaneous procedure jointly estimates β and h by maximizing a sample criterion function $\widehat{Q}_n(\beta, h)$ over the sieve parameter space $\Theta_n = B \times \mathcal{H}_n$. The earlier applications of sieve MLE in econometrics, such as the papers by Duncan (1986) and Gallant and Nychka (1987) took this approach.

In Subsection 4.1 we review existing theory on the $\sqrt{n}$-asymptotic normality of the two-step estimates of β. In Subsection 4.2, we present recent advances on the $\sqrt{n}$-asymptotic normality and efficiency of the sieve simultaneous M-estimates of β. In Subsection 4.3, we mention the $\sqrt{n}$-asymptotic normality and efficiency of the simultaneous sieve MD estimates of β.

4.1. Semiparametric two-step estimators

There are several general theory papers in econometrics about the semiparametric two-step procedure. Andrews (1994b) proposed the MINPIN estimator of β, which is the extremum estimator of β where the empirical criterion function depends on the first step nonparametric estimator of h. Andrews (1994b) also provided a set of relatively high level conditions to ensure the $\sqrt{n}$-normality of his MINPIN estimator of β. Ichimura and Lee (2006) presented a set of relatively low level conditions to ensure the $\sqrt{n}$-normality of the semiparametric two-step M-estimator of β. Newey (1994a), Pakes and Olley (1995), and Chen, Linton and van Keilegom (2003) have studied the properties of the semiparametric two-step GMM estimator of β. In addition to providing a general way to compute the asymptotic variance of the second step β estimate, Newey (1994a) showed that the second stage estimation of β and its asymptotic variance do not depend on the particular choice of the nonparametric estimation technique in the first step, but only depend on the convergence rate of the first step estimation.

4.1.1. Asymptotic normality

In the following we state two results which are slight modifications of those in Chen, Linton and van Keilegom (2003), in which the empirical criterion function can be nonsmooth with respect to both β and h. Let $M : B \times \mathcal{H} \mapsto \mathcal{R}^{d_m}$ be a nonrandom, vector-valued measurable function, where B is a compact subset in $\mathcal{R}^{d_\beta}$ with $d_m \geqslant d_\beta$. The identifying assumption is that $M(\beta, h_o(\cdot, \beta)) = 0$ at $\beta = \beta_o \in B$ and $M(\beta, h_o(\cdot, \beta)) \neq 0$ for all $\beta \neq \beta_o$. We denote $\beta_o \in B$ and $h_o \in \mathcal{H}$ as the true unknown finite- and infinite-dimensional parameters, where the function $h_o \in \mathcal{H}$ can depend on the parameters β and the data Z. We usually suppress the arguments of the function h_o for notational convenience; thus: $(\beta, h) \equiv (\beta, h(\cdot, \beta))$, $(\beta, h_o) \equiv (\beta, h_o(\cdot, \beta))$ and $(\beta_o, h_o) \equiv (\beta_o, h_o(\cdot, \beta_o))$. We assume that $\mathcal{H}$ is a vector space of functions endowed with a pseudo-metric $\|\cdot\|_{\mathcal{H}}$, which is a sup-norm metric with respect to the β-argument and a pseudo-metric with respect to all the other arguments. Suppose that there also exists a random vector-valued function $M_n: B \times \mathcal{H} \to \mathcal{R}^{d_m}$ depending on the data $\{Z_i\colon i = 1, \ldots, n\}$, such that $M_n(\beta, h_o)' W M_n(\beta, h_o)$ is close to $M(\beta, h_o)' W M(\beta, h_o)$ for some symmetric positive-definite matrix W. Suppose that for each β there is an initial nonparametric estimator $\hat{h}(.)$ for $h_o(.)$. Denote W_n as a possibly random weighting matrix such that $W_n - W = o_P(1)$. Then β_o can be estimated by $\hat{\beta}$, which solves the sample minimum distance problem[37]:

$$\min_{\beta \in B} M_n(\beta, \hat{h})' W_n M_n(\beta, \hat{h}). \tag{4.1}$$

For any $\beta \in B$, we say that $M(\beta, h)$ is pathwise differentiable at h in the direction $[\bar{h} - h]$ if $\{h + \tau(\bar{h} - h)\colon \tau \in [0, 1]\} \subset \mathcal{H}$ and $\lim_{\tau \to 0}[M(\beta, h + \tau(\bar{h} - h)) - M(\beta, h)]/\tau$ exists; we denote the limit by $\Gamma_2(\beta, h)[\bar{h} - h]$.

THEOREM 4.1. *Suppose that $\beta_o \in \operatorname{int}(B)$ satisfies $M(\beta_o, h_o) = 0$, that $\hat{\beta} - \beta_o = o_P(1)$, $W_n - W = o_P(1)$, and that*:

(4.1.1) $\|M_n(\hat{\beta}, \hat{h})\| = \inf_{\|\beta - \beta_o\| \leqslant \delta_n} \|M_n(\beta, \hat{h})\| + o_P(1/\sqrt{n})$ *for some positive sequence* $\delta_n = o(1)$.

(4.1.2) (i) *The ordinary partial derivative $\Gamma_1(\beta, h_o)$ in β of $M(\beta, h_o)$ exists in a neighborhood of β_o, and is continuous at $\beta = \beta_o$*; (ii) *the matrix $\Gamma_1 \equiv \Gamma_1(\beta_o, h_o)$ is such that $\Gamma_1' W \Gamma_1$ is nonsingular.*

(4.1.3) *The pathwise derivative $\Gamma_2(\beta, h_o)[h - h_o]$ of $M(\beta, h_o)$ exists in all directions $[h - h_o]$ and satisfies*:

$$\big\|\Gamma_2(\beta, h_o)[h - h_o] - \Gamma_2(\beta_o, h_o)[h - h_o]\big\| \leqslant \|\beta - \beta_o\| \times o(1)$$

for all β with $\|\beta - \beta_o\| = o(1)$, all h with $\|h - h_o\|_{\mathcal{H}} = o(1)$. Either

(4.1.4) $\|M(\beta, \hat{h}) - M(\beta, h_o) - \Gamma_2(\beta, h_o)[\hat{h} - h_o]\| = o_P(n^{-1/2})$ *for all β with* $\|\beta - \beta_o\| = o(1)$; *or*

(4.1.4)$'$ (i) *there are some constants $c \geqslant 0$, $\epsilon \in (0, 1]$ such that*

$$\big\|M(\beta, h) - M(\beta, h_o) - \Gamma_2(\beta, h_o)[h - h_o]\big\| \leqslant c\|h - h_o\|_{\mathcal{H}}^{1+\epsilon}$$

for all β with $\|\beta - \beta_o\| = o(1)$ and all h with $\|h - h_o\|_{\mathcal{H}} = o(1)$; *and* (ii) $c\|\hat{h} - h_o\|_{\mathcal{H}}^{1+\epsilon} = o_P(n^{-1/2})$.

(4.1.5) *For all sequences of positive numbers $\{\delta_n\}$ with $\delta_n = o(1)$,*

$$\sup_{\|\beta - \beta_o\| < \delta_n,\, \|h - h_o\|_{\mathcal{H}} < \delta_n} \frac{\|M_n(\beta, h) - M(\beta, h) - M_n(\beta_o, h_o)\|}{n^{-1/2} + \|M_n(\beta, h)\| + \|M(\beta, h)\|} = o_P(1).$$

(4.1.6) *For some finite matrix V_1, $\sqrt{n}\{M_n(\beta_o, h_o) + \Gamma_2(\beta_o, h_o)[\hat{h} - h_o]\} \xrightarrow{d} \mathcal{N}[0, V_1]$. Then $\sqrt{n}(\hat{\beta} - \beta_o) \xrightarrow{d} \mathcal{N}[0, (\Gamma_1' W \Gamma_1)^{-1} \Gamma_1' W V_1 W \Gamma_1 (\Gamma_1' W \Gamma_1)^{-1}]$.*

[37] See Theorem 1 in Chen, Linton and van Keilegom (2003) for the consistency property of $\hat{\beta} - \beta_o = o_P(1)$.

REMARK 4.1. This theorem can be established by following the proof of Theorem 2 in Chen, Linton and van Keilegom (2003). Note that condition (4.1.4) is implied by condition (4.1.4)$'$, while condition (4.1.4)$'$ with $\epsilon = 1$ becomes the one imposed in Newey (1994a) and Chen, Linton and van Keilegom (2003). When $M(\beta, h)$ is highly nonlinear in h and/or when the argument "$\cdot$" of $h(\cdot, \beta)$ has unbounded support, then condition (4.1.4)$'$(i) with $\epsilon = 1$ may fail to hold, but condition (4.1.4)$'$ with $0 < \epsilon < 1$ is typically satisfied. See Chen, Hong and Tarozzi (2007) for such an example in the two-step GMM estimation for nonclassical measurement error models and missing data problems. Of course a smaller ϵ has to be compensated by a faster rate of convergence of $\hat{h}$ to h_o in condition (4.1.4)$'$(ii) $\|\hat{h} - h_o\|_{\mathcal{H}} = o_P(n^{-1/2(1+\epsilon)})$. In the extreme case when $\|\hat{h} - h_o\|_{\mathcal{H}} = O_P(n^{-1/2})$, which is the case if h_o is a probability distribution function, then condition (4.1.4) is implied by condition

(4.1.4)$''$ (i) $\|M(\beta, h) - M(\beta, h_o) - \Gamma_2(\beta, h_o)[h - h_o]\| = \|h - h_o\|_{\mathcal{H}} \times o(1)$ *for all* β *with* $\|\beta - \beta_o\| = o(1)$ *and all* h *with* $\|h - h_o\|_{\mathcal{H}} = o(1)$; *and* (ii) $\|\hat{h} - h_o\|_{\mathcal{H}} = O_P(n^{-1/2})$.

Many econometric models correspond to $M(\beta, h) = E[m(Z_i, \beta, h)]$, $M_n(\beta, h) = n^{-1}\sum_{i=1}^{n} m(Z_i, \beta, h)$, where $m : \mathcal{R}^{d_z} \times B \times \mathcal{H} \to \mathcal{R}^{d_m}$ is a measurable, vector-valued function such that $E[m(Z_i, \beta, h_o(\cdot, \beta))] = 0$ if and only if $\beta = \beta_o \in B$, a subset of $\mathcal{R}^{d_\beta}$. In this situation, Theorem 3 in Chen, Linton and van Keilegom (2003) provides a set of easily-verifiable sufficient conditions for the stochastic equicontinuity condition (4.1.5) with i.i.d. data $\{Z_i\}$. The following lemma extends their result to strictly stationary processes. Let $\mathcal{F} = \{m(z, \beta, h)\colon \beta \in B,\ h \in \mathcal{H}\}$ denote the class of measurable functions indexed by (β, h), and $H_{[\,]}(w, \mathcal{F}, \|\cdot\|_r)$ be the $L_r(P_o)$-metric entropy with bracketing of the class $\mathcal{F}$.

LEMMA 4.2. *Suppose that* $\{Z_t\colon t \geqslant 1\}$ *is strictly stationary, that* $M(\beta, h) = E[m(Z_t, \beta, h)]$ *and* $M_n(\beta, h) = n^{-1}\sum_{i=1}^{n} m(Z_t, \beta, h)$, *and that the arguments of the* $h(\cdot)$ *in* $m(Z_t, \beta, h(\cdot))$ *only depend on* β *and finitely many* Z_t. *Suppose that each component* m_j *of* $m = (m_1, \ldots, m_{d_m})'$ *satisfies*:

(4.2.1) $m_j(\cdot, \beta, h)$ *is locally uniformly* $L_r(P_o)$*-continuous with respect to* β, h *in the sense*:

$$\left(E\left[\sup_{(\beta', h')\colon \|\beta' - \beta\| < \delta,\ \|h' - h\|_{\mathcal{H}} < \delta} \left|m_{lcj}(Z, \beta', h') - m_{lcj}(Z, \beta, h)\right|^r\right]\right)^{1/r} \leqslant K_j \delta^{s_j}$$

for all $(\beta, h) \in B \times \mathcal{H}$, *all small positive value* $\delta = o(1)$, *and for some constants* $s_j \in (0, 1]$, $K_j > 0$ *and* $r \geqslant 1$.

Then: (i) $H_{[\,]}(w, \mathcal{F}_j, \|\cdot\|_r) \leqslant \log N([\frac{\varepsilon}{2K_j}]^{1/s_j}, B, \|\cdot\|) + \log N([\frac{\varepsilon}{2K_j}]^{1/s_j}, \mathcal{H}, \|\cdot\|_{\mathcal{H}})$ *for* $j = 1, \ldots, d_m$.

Furthermore, suppose that

(4.2.2) B *is a compact subset of* $\mathcal{R}^{d_\beta}$, *and* $\int_0^\infty \sqrt{\log N(\varepsilon^{1/s_j}, \mathcal{H}, \|\cdot\|_{\mathcal{H}})}\, \mathrm{d}\varepsilon < \infty$ *for* $j = 1, \ldots, d_m$.

(4.2.3) *Either* $\{Z_t\}_{t=1}^n$ *is i.i.d. and* (4.2.1) *holds with* $r \geqslant 2$, *or* $\{Z_t\}_{t=1}^n$ *is beta-mixing with a mixing decay rate satisfying* $\sum_{t=1}^{\infty} t^{2/(r-2)}\beta_t < \infty$ *for some* $r > 2$, *and* (4.2.1) *holds with* $r > 2$.

Then: (ii) *for all positive* δ_n *with* $\delta_n = o(1)$,

$$\sup_{\|\beta-\beta_o\|<\delta_n,\, \|h-h_o\|_{\mathcal{H}}<\delta_n} \big\| M_n(\beta,h) - M(\beta,h) - \{M_n(\beta_o,h_o) - M(\beta_o,h_o)\}\big\| = o_P\big(n^{-1/2}\big). \tag{4.2}$$

PROOF. Result (i) is already derived in the proof of Theorem 3 in Chen, Linton and van Keilegom (2003). Result (ii) for i.i.d. case is Theorem 3 of Chen, Linton and van Keilegom (2003). Now for stationary beta-mixing case, conditions (4.2.1)–(4.2.3) imply that $\int_0^\infty \sqrt{H_{[\,]}(w,\mathcal{F},\|\cdot\|_r)}\,\mathrm{d}w < \infty$ with $r > 2$. This and $\sum_{t=1}^{\infty} t^{2/(r-2)}\beta_t < \infty$ imply that all the assumptions in Doukhan, Massart and Rio (1995) for the Donsker theorem on stationary beta-mixing are satisfied, which in turn implies the stochastic equicontinuity (4.2) result (ii). □

Both Theorem 3 in Chen, Linton and van Keilegom (2003) and Lemma 4.2 are extensions of the "type II class" and "type IV class" defined in Andrews (1994a) from $\beta \in B$ to $(\beta,h) \in B \times \mathcal{H}$. Condition (4.2.1) allows for discontinuous moment functions in (β,h) such as sign and indicator functions of (β,h).

Given the results of Newey (1994a), Chen, Linton and van Keilegom (2003) and Theorem 4.1, the choice of estimation of h in the first step should mainly depend on the ease of implementation. Recently, for the partially linear quantile regression $Y_t = X_{0t}'\beta_o + h_o(X_{1t}) + e_t$, $P[e_t \leqslant 0|X_t] = \alpha \in (0,1)$, Lee (2003) proposed a two-step, $\sqrt{n}$ asymptotically normal and efficient estimator of β, where the first step involved a high-dimensional kernel quantile regression of Y_t on $X = (X_0', X_1')'$. Chen, Linton and van Keilegom (2003) considered a modification of Lee's model to a partially linear quantile regression with some endogenous regressors, and proposed another $\sqrt{n}$ asymptotically normal estimator of β by two-step GMM where the first step nonparametric estimation only involves $h(X_{1t})$. We can extend their models further to a partially additive quantile regression:

$$Y_t = X_{0t}'\beta_o + \sum_{j=1}^{q} h_{oj}(X_{jt}) + e_t, \qquad P[e_t \leqslant 0|X_t] = \alpha \in (0,1).$$

If $h_{o1},\dots,h_{oq}$ were known, then β_o could be estimated based on the moment restriction $E[m(Z_i,\beta,h_o)] = 0$ iff $\beta = \beta_o$ with $m(Z_i,\beta,h_o) = X_0\{\alpha - 1(Y \leqslant X_{0t}'\beta + \sum_{j=1}^{q} h_{oj}(X_{jt})\}$. Clearly, to estimate β by semiparametric two-step GMM using the sample moment $n^{-1}\sum_{i=1}^{n} m(Z_i,\beta,\hat{h})$, it would be much easier if $\hat{h} = (\hat{h}_1,\dots,\hat{h}_q)$ were a sieve estimate, say obtained by $\max_{h\in\mathcal{H}_n} \widehat{Q}_n(\beta,h) = n^{-1}\sum_{t=1}^{n} l(\beta,h,Y_t,X_t)$ for any arbitrarily fixed β, where

$$l(\beta, h, Y_t, X_t) = \left\{1\left(Y_t < X_{0t}'\beta + \sum_{j=1}^{q} h_j(X_{jt})\right) - \alpha\right\}\left[Y_t - X_{0t}'\beta - \sum_{j=1}^{q} h_j(X_{jt})\right],$$

and $\mathcal{H}_n = \mathcal{H}_n^1 \times \cdots \times \mathcal{H}_n^q$ as in Subsection 3.2.1, rather than a high-dimensional kernel quantile regression. Andrews (1994b), Newey (1994a, 1994b), Newey, Powell and Vella (1999) and Das, Newey and Vella (2003) have made the same recommendation in the context of two-step estimation with an additive LS regression in the first step.

There is also a large literature on the general theory of efficient estimation of β via various two-step procedures. For instance, the profile MLE estimation of β [which can be viewed as an important subclass of Andrews' (1994b) MINPIN procedure] can lead to efficient estimation of β; see e.g. Severini and Wong (1992), Ai (1997) and Murphy and van der Vaart (2000). Other two-step procedures which lead to the efficient estimation of β include those based on the efficient score equation approach; see Bickel et al. (1993) and Newey (1990a), and the optimally weighted GMM approach; see Newey (1990a, 1990b, 1993). See Powell (1994) and Pagan and Ullah (1999) for other examples. Clearly, these efficient procedures can be combined with a first step nonparametric estimation of h via the method of sieves.

4.2. Sieve simultaneous M-estimation

There are few general theory papers about the sieve simultaneous M-estimation of β and h; see Wong and Severini (1991), Shen (1997), Chen and Shen (1998). This procedure jointly estimates β and h by maximizing a sample criterion function $\widehat{Q}_n(\beta, h)$ over the sieve parameter space $\Theta_n = B \times \mathcal{H}_n$, where $\widehat{Q}_n(\beta, h)$ takes a sample average form $\frac{1}{n}\sum_{i=1}^{n} l(\beta, h, Z_i)$. Wong and Severini (1991) established $\sqrt{n}$-asymptotic normality and efficiency of smooth functionals of nonparametric MLE with parameter space $\Theta_n \equiv \Theta = B \times \mathcal{H}$. Shen (1997) extended their results to sieve MLE and to allow for highly curved (nonlinear) least favorable directions. Chen and Shen (1998) extend the result of Shen (1997) to general sieve M-estimation with stationary weakly dependent data.

4.2.1. Asymptotic normality of smooth functionals of sieve M-estimators

Let $\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n) = \arg\max_{(\beta,h)\in B\times\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^{n} l(\beta, h, Z_i)$ denote the sieve M-estimate of $\theta_o = (\beta_o, h_o)$. In this subsection we present a simple $\sqrt{n}$-asymptotic normality theorem for the plug-in estimate of a smooth functional of θ_o. See Shen (1997) and Chen and Shen (1998) for the general version.

Suppose that $\Theta = B \times \mathcal{H}$ is convex in θ_o so that $\theta_o + \tau[\theta - \theta_o] \in \Theta$ for all small $\tau \in [0, 1]$ and for all fixed $\theta \in \Theta$. Suppose that the directional derivative

$$\frac{\partial l(\theta_o, z)}{\partial\theta}[\theta - \theta_o] \equiv \lim_{\tau\to 0} \frac{l(\theta_o + \tau[\theta - \theta_o], z) - l(\theta_o, z)}{\tau}$$

is well defined for almost all z in the support of Z.

Let $\Theta = B \times \mathcal{H}$ be equipped with a norm $\|\cdot\|$. Suppose the functional of interest, $f : \Theta \to \mathcal{R}$, is smooth in the sense that

$$\frac{\partial f(\theta_o)}{\partial \theta}[\theta - \theta_o] \equiv \lim_{\tau \to 0} \frac{f(\theta_o + \tau[\theta - \theta_o]) - f(\theta_o)}{\tau}$$

is well defined and

$$\left\| \frac{\partial f(\theta_o)}{\partial \theta} \right\| \equiv \sup_{\{\theta \in \Theta:\ \|\theta - \theta_o\| > 0\}} \frac{|\frac{\partial f(\theta_o)}{\partial \theta}[\theta - \theta_o]|}{\|\theta - \theta_o\|} < \infty.$$

Next, suppose that $\|\cdot\|$ induces an inner product $\langle \cdot, \cdot \rangle$ on the completion of the space spanned by $\Theta - \theta_o$, denoted as $\bar{V}$. By the Riesz representation theorem, there exists $v^* \in \bar{V}$ such that, for any $\theta \in \Theta$,

$$\frac{\partial f(\theta_o)}{\partial \theta}[\theta - \theta_o] = \langle \theta - \theta_o, v^* \rangle \quad \text{iff} \quad \left\| \frac{\partial f(\theta_o)}{\partial \theta} \right\| < \infty.$$

Suppose that the sieve M-estimate $\hat{\theta}_n$ converges to θ_o at a rate faster than δ_n (i.e., $\|\hat{\theta}_n - \theta_o\| = o_P(\delta_n)$). Let ε_n denote any sequence satisfying $\varepsilon_n = o(n^{-1/2})$, and $\mu_n(g(Z)) = \frac{1}{n}\sum_{t=1}^{n}\{g(Z_t) - E(g(Z_t))\}$ denote the empirical process indexed by the function g. Recall that $K(\theta_o, \theta) \equiv E[l(\theta_o, Z_i) - l(\theta, Z_i)]$.

CONDITION 4.1.

(i) There is $\omega > 0$ such that $|f(\theta) - f(\theta_o) - \frac{\partial f(\theta_o)}{\partial \theta}[\theta - \theta_o]| = O(\|\theta - \theta_o\|^{\omega})$ uniformly in $\theta \in \Theta_n$ with $\|\theta - \theta_o\| = o(1)$;

(ii) $\|\frac{\partial f(\theta_o)}{\partial \theta}\| < \infty$;

(iii) there is $\pi_n v^* \in \Theta_n$ such that $\|\pi_n v^* - v^*\| \times \|\hat{\theta}_n - \theta_o\| = o_P(n^{-1/2})$.

CONDITION 4.2.

$$\sup_{\{\theta \in \Theta_n:\ \|\theta - \theta_o\| \leqslant \delta_n\}} \mu_n\left(l(\theta, Z) - l(\theta \pm \varepsilon_n \pi_n v^*, Z) - \frac{\partial l(\theta_o, Z)}{\partial \theta}[\pm \varepsilon_n \pi_n v^*] \right) = O_P\left(\varepsilon_n^2\right).$$

CONDITION 4.3. $K(\theta_o, \hat{\theta}_n) - K(\theta_o, \hat{\theta}_n \pm \varepsilon_n \pi_n v^*) = \pm \varepsilon_n \langle \hat{\theta}_n - \theta_o, \pi_n v^* \rangle + o(n^{-1})$.

CONDITION 4.4.

(i) $\mu_n(\frac{\partial l(\theta_o, Z)}{\partial \theta}[\pi_n v^* - v^*]) = o_P(n^{-1/2})$;

(ii) $E\{\frac{\partial l(\theta_o, Z)}{\partial \theta}[\pi_n v^*]\} = o(n^{-1/2})$.

CONDITION 4.5. $n^{1/2}\mu_n(\frac{\partial l(\theta_o, Z)}{\partial \theta}[v^*]) \xrightarrow{d} \mathcal{N}(0, \sigma_{v^*}^2)$, with $\sigma_{v^*}^2 > 0$.

We note that for classical nonlinear M-estimation such as those reviewed in Newey and McFadden (1994), Conditions 4.1(i)(ii), 4.2, 4.3 and 4.5 are still required (albeit in slightly different expressions), while Conditions 4.1(iii) and 4.4 are automatically satisfied since $\pi_n v^* = v^*$ for the standard nonlinear M-estimation. Note that for i.i.d. data Condition 4.5 is satisfied whenever $\sigma_{v^*}^2 = \mathrm{Var}(\frac{\partial l(\theta_o, Z)}{\partial \theta}[v^*]) > 0$. If $l(\theta, Z)$ is also pathwise differentiable in $\theta \in \Theta_n$ with $\|\theta - \theta_o\| = o(1)$, then Conditions 4.2 and 4.3 are implied by Conditions 4.2′ and 4.3′ respectively, where

CONDITION 4.2′.

$$\sup_{\{\bar{\theta} \in \Theta_n \colon \|\bar{\theta} - \theta_o\| \leqslant \delta_n\}} \mu_n \left(\frac{\partial l(\bar{\theta}, Z)}{\partial \theta}[\pi_n v^*] - \frac{\partial l(\theta_o, Z)}{\partial \theta}[\pi_n v^*] \right) = o_P\left(n^{-1/2}\right).$$

CONDITION 4.3′. $E\{\frac{\partial l(\hat{\theta}_n, Z)}{\partial \theta}[\pi_n v^*]\} = \langle \hat{\theta}_n - \theta_o, \pi_n v^* \rangle + o(n^{-1/2})$.

Condition 4.2 (or 4.2′) can be verified by applying Lemma 4.2. Condition 4.3 (or 4.3′) can be verified when a Hilbert norm $\|\theta - \theta_o\|$ is chosen.

Conditions 4.2–4.4 may need to be modified when the parameter space Θ is not convex; see Shen (1997) and Chen and Shen (1998) for the needed modification.

THEOREM 4.3. *Suppose Conditions 4.1–4.5 hold, and $\|\hat{\theta}_n - \theta_o\|^{\omega} = o_P(n^{-1/2})$. Then, for the sieve M-estimate $\hat{\theta}_n$, $n^{1/2}(f(\hat{\theta}_n) - f(\theta_o)) \xrightarrow{d} \mathcal{N}(0, \sigma_{v^*}^2)$.*

The proof of Theorem 4.3 follows trivially from those in Shen (1997) and Ai and Chen (1999). In applications, one needs to specify a Hilbert norm $\|\theta - \theta_o\|$ in order to compute the representer v^*. Wong and Severini (1991) and Shen (1997) have used the Fisher norm, $\|\theta - \theta_o\|^2 = E\{\frac{\partial l(\theta_o, Z_i)}{\partial \theta}[\theta - \theta_o]\}^2$, for the sieve MLE procedure. Ai and Chen (1999, 2003) have introduced a Fisher-like norm for their sieve MD and sieve GLS procedures. In the next subsection we specialize Theorem 4.3 to derive root-n asymptotic normality of parametric parts in sieve GLS problems.

4.2.2. Asymptotic normality of sieve GLS

Recall that for all the models belonging to the first subclass of the conditional moment restrictions (2.8), $E\{\rho(Z, \theta_o)|X\} = 0$, where $\rho(Z, \theta) - \rho(Z, \theta_o)$ does not depend on endogenous variables Y, we can estimate $\theta_o = (\beta_o, h_o) \in B \times \mathcal{H}$ by the sieve GLS procedure:

$$\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n) = \arg \min_{(\beta, h) \in B \times \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \rho(Z_i, \beta, h)' \Sigma(X_i)^{-1} \rho(Z_i, \beta, h),$$

where $\Sigma(X_i)$ is a positive definite weighting matrix. When $\Sigma(X_i)$ is known such as the identity matrix, this belongs to the sieve M-estimation with $l(\theta, Z_i) = -\rho(Z_i, \theta)' \Sigma(X_i)^{-1} \rho(Z_i, \theta)/2$. See Subsection 4.3 and Remark 4.3 for estimation of the optimal weighting matrix $\Sigma_o(X_i) \equiv \mathrm{Var}\{\rho(Z_i, \theta_o)|X_i\}$.

We now apply Theorem 4.3 to derive root-n asymptotic normality of the sieve GLS estimator $\hat{\beta}_n$. Define the norm $\|\theta-\theta_o\|^2 = E\{(\frac{\partial\rho(Z_i,\theta_o)}{\partial\theta}[\theta-\theta_o])\Sigma(X_i)^{-1}(\frac{\partial\rho(Z_i,\theta_o)}{\partial\theta}[\theta-\theta_o])\}$. For $j=1,\ldots,d_\beta$, let

$$D_{w_j}(X) = \frac{\partial\rho(Z,\beta,h_o(\cdot))}{\partial\beta_j}\bigg|_{\beta=\beta_o} - \frac{\partial\rho(X,\beta_o,h_o(\cdot)+\tau w_j(\cdot))}{\partial\tau}\bigg|_{\tau=0}$$
$$= \frac{\partial\rho(Z,\theta_o)}{\partial\beta_j} - \frac{\partial\rho(Z,\theta_o)}{\partial h}[w_j],$$

$w=(w_1,\ldots,w_{d_\beta})$, and $D_w(X) = (D_{w_1}(X),\ldots,D_{w_{d_\beta}}(X)) = \frac{\partial\rho(Z,\theta_o)}{\partial\beta'} - \frac{\partial\rho(Z,\theta_o)}{\partial h}[w]$ be a $(d_\rho\times d_\beta)$-matrix valued measurable function of X. Let $w^* = (w_1^*,\ldots,w_{d_\beta}^*)$, where for $j=1,\ldots,d_\beta$, w_j^* solves

$$E\{D_{w_j^*}(X)'\Sigma(X)^{-1}D_{w_j^*}(X)\} = \inf_{w_j} E\{D_{w_j}(X)'\Sigma(X)^{-1}D_{w_j}(X)\}.$$

Denote $D_{w^*}(X) = \frac{\partial\rho(Z,\theta_o)}{\partial\beta'} - \frac{\partial\rho(Z,\theta_o)}{\partial h}[w^*]$. Let

$$v_\beta^* = \left(E\{D_{w^*}(X)'\Sigma(X)^{-1}D_{w^*}(X)\}\right)^{-1}\lambda,$$

$v_h^* = -w^* v_\beta^*$ and $v^* = (v_\beta^*, v_h^*)$.

ASSUMPTION 4.1.
(i) $\beta_o \in \mathrm{int}(B)$;
(ii) $E[D_{w^*}(X)'\Sigma(X)^{-1}D_{w^*}(X)]$ is positive definite;
(iii) there is $\pi_n v^* \in \Theta_n$ such that $\|\pi_n v^* - v^*\| \times \|\hat{\theta}_n - \theta_o\| = o_P(n^{-1/2})$.

ASSUMPTION 4.2.
(i) $\Sigma(X)$ and $\Sigma_o(X) \equiv \mathrm{Var}\{\rho(Z_i,\theta_o)|X\}$ are positive definite and bounded uniform over X;
(ii) $\rho(Z,\theta)$ is twice continuously pathwise differentiable with respect to $\theta\in\Theta$ with $\|\theta-\theta_o\| = o(1)$;
(iii) Conditions 4.2′ and 4.3′ are satisfied with $\frac{\partial l(\bar{\theta},Z)}{\partial\theta}[\pi_n v^*] = -\rho(Z,\bar{\theta})' \times \Sigma(X)^{-1}\{\frac{\partial\rho(Z,\bar{\theta})}{\partial\theta}[\pi_n v^*]\}$ for all $\bar{\theta}\in\Theta_n$ with $\|\bar{\theta}-\theta_o\| = o(1)$;
(iv) $\{Z_i\}_{i=1}^n$ is i.i.d., $E\{\rho(Z,\theta_o)|X\} = 0$, $E\{\rho(Z,\theta)-\rho(Z,\theta_o)|X\} = \rho(Z,\theta) - \rho(Z,\theta_o)$ for all $\theta\in\Theta$.

PROPOSITION 4.4. *Let $\hat{\theta}_n$ be the sieve GLS estimate. Suppose Assumptions 4.1–4.2 hold. Then $n^{1/2}(\hat{\beta}_n-\beta_o) \xrightarrow{d} \mathcal{N}(0, V_1^{-1}V_2V_1^{-1})$ where*

$$V_1 = E\left[D_{w^*}(X)'\Sigma(X)^{-1}D_{w^*}(X)\right],$$
$$V_2 = E\left[D_{w^*}(X)'\Sigma(X)^{-1}\Sigma_o(X)\Sigma(X)^{-1}D_{w^*}(X)\right].$$

PROOF. Let $f(\theta) = \lambda'\beta$, where λ is an arbitrary unit vector in $\mathcal{R}^{d_\beta}$. Clearly, Condition 4.1(i) is satisfied with $\frac{\partial f(\theta_o)}{\partial\theta}[\theta - \theta_o] = (\beta - \beta_o)'\lambda$ and $\omega = \infty$. In addition, under Assumption 4.1(i)(ii), we have $v^* = (v_\beta^*, v_h^*)$ and

$$\|v^*\|^2 = \sup_{\{\theta\in\Theta:\ \|\theta-\theta_o\|>0\}} \frac{\{(\beta-\beta_o)'\lambda\}^2}{\|\theta-\theta_o\|^2}$$
$$= \lambda'\bigl(E\bigl\{D_{w^*}(X)'\Sigma(X)^{-1}D_{w^*}(X)\bigr\}\bigr)^{-1}\lambda < \infty.$$

Thus Condition 4.1 is implied by Assumption 4.1. Note that

$$\frac{\partial l(\theta_o, Z)}{\partial\theta}[\theta-\theta_o]$$
$$= -\rho(Z,\theta_o)'\Sigma(X)^{-1}\left(\frac{\partial\rho(Z,\theta_o)}{\partial\beta'}(\beta-\beta_o) + \frac{\partial\rho(Z,\theta_o)}{\partial h}[h-h_o]\right),$$

we have

$$E\left\{\frac{\partial l(\theta_o, Z)}{\partial\theta}[\pi_n v^*]\right\}$$
$$= -E\left\{\rho(Z,\theta_o)'\Sigma(X)^{-1}\left(\frac{\partial\rho(Z,\theta_o)}{\partial\beta'}\bigl(v_\beta^*\bigr) + \frac{\partial\rho(Z,\theta_o)}{\partial h}\bigl[\pi_n v_h^*\bigr]\right)\right\} = 0,$$

hence Condition 4.4(ii) is automatically satisfied. Since

$$\frac{1}{n}\sum_{t=1}^{n}\frac{\partial l(\theta_o, Z_t)}{\partial\theta}[\pi_n v^* - v^*]$$
$$= \frac{-1}{n}\sum_{t=1}^{n}\rho(Z_t,\theta_o)'\Sigma(X_t)^{-1}\left(\frac{\partial\rho(Z_t,\theta_o)}{\partial h}\bigl[\pi_n v_h^* - v_h^*\bigr]\right),$$

by Chebyshev inequality and Assumptions 4.1(iii) and 4.2(i), we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l(\theta_o, Z_i)}{\partial\theta}[\pi_n v^* - v^*] = o_P\bigl(n^{-1/2}\bigr),$$

hence Condition 4.4(i) is satisfied. Since data are i.i.d. and under Assumptions 4.1(ii) and 4.2(i),

$$\sigma_{v^*}^2 = \mathrm{Var}\left\{\frac{\partial l(\theta_o, Z)}{\partial\theta}[v^*]\right\}$$
$$= \mathrm{Var}\left\{\rho(Z,\theta_o)'\Sigma(X)^{-1}\left(\frac{\partial\rho(Z,\theta_o)}{\partial\beta'} - \frac{\partial\rho(Z,\theta_o)}{\partial h}[w^*]\right)\bigl(v_\beta^*\bigr)\right\}$$
$$= \bigl(v_\beta^*\bigr)'E\bigl\{D_{w^*}(X)'\Sigma(X)^{-1}\Sigma_o(X)\Sigma(X)^{-1}D_{w^*}(X)\bigr\}\bigl(v_\beta^*\bigr)$$
$$= \lambda' V_1^{-1}V_2V_1^{-1}\lambda > 0,$$

Condition 4.5 is satisfied. By Theorem 4.3, we obtain, for any arbitrary unit vector $\lambda \in \mathcal{R}^{d_\beta}$, $n^{1/2}\lambda'(\hat{\beta}_n - \beta_o) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{v^*})$. Hence $\sqrt{n}(\hat{\beta}_n - \beta_o) \xrightarrow{d} \mathcal{N}(0, V_1^{-1}V_2V_1^{-1})$. □

REMARK 4.2. The asymptotic variance, $V_1^{-1}V_2V_1^{-1}$, of the sieve GLS estimator $\hat{\beta}_n$ can be consistently estimated by $\widehat{V}_1^{-1}\widehat{V}_2\widehat{V}_1^{-1}$, where

$$\widehat{V}_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial\beta'} - \frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial h}[\hat{w}]\right)' \times \Sigma(X_i)^{-1}\left(\frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial\beta'} - \frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial h}[\hat{w}]\right),$$

$$\widehat{V}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial\beta'} - \frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial h}[\hat{w}]\right)' \times \Sigma(X_i)^{-1}\widehat{\Sigma}_o(X_i)\Sigma(X_i)^{-1}\left(\frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial\beta'} - \frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial h}[\hat{w}]\right),$$

$\hat{w} = (\hat{w}_1, \ldots, \hat{w}_{d_\beta})$ solves the following sieve minimization problem: for $j = 1, \ldots, d_\beta$,

$$\min_{w_j \in \mathcal{H}_n} \sum_{i=1}^{n}\left(\frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial\beta_j} - \frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial h}[w_j]\right)' \times \left[\Sigma(X_i)\right]^{-1}\left(\frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial\beta_j} - \frac{\partial\rho(Z_i,\hat{\theta}_n)}{\partial h}[w_j]\right),$$

and $\widehat{\Sigma}_o(X_i)$ can be any consistent nonparametric estimator of $\Sigma_o(X_i)$; see Ai and Chen (1999) for kernel estimator and Ai and Chen (2003, 2007) for series LS estimator of $\Sigma_o(X_i)$.

4.2.3. *Example: Partially additive mean regression with a monotone constraint*

Suppose that the i.i.d. data $\{Y_t, X_t' = (X_{0t}', X_{1t}, \ldots, X_{qt})\}_{t=1}^{n}$ are generated according to

$$Y_i = X_{0i}'\beta_o + h_{o1}(X_{1i}) + \cdots + h_{oq}(X_{qi}) + e_i, \qquad E[e_i|X_i] = 0.$$

Let $\theta_o = (\beta_o', h_{o1}, \ldots, h_{oq})' \in \Theta = B \times \mathcal{H}$ be the parameters of interests, where B is a compact subset of $\mathcal{R}^{d_\beta}$ and $\mathcal{H}$ is the same as that in Subsection 3.2.1. Since $h_{o1}(\cdot)$ can have a constant we assume that X_0 does not contain the constant regressor, $\dim(X_0) = d_\beta$, $\dim(X_j) = 1$ for $j = 1, \ldots, q$, $\dim(X) = d_\beta + q$, and $\dim(Y) = 1$. We estimate the regression function $\theta_o(X) = X_{0t}'\beta_o + \sum_{j=1}^{q} h_{oj}(X_{jt})$ by maximizing over $\Theta_n = B \times \mathcal{H}_n$ the criterion $\widehat{Q}_n(\theta) = n^{-1}\sum_{t=1}^{n} l(\theta, Y_t, X_t)$, where $l(\theta, Y_t, X_t) =$

$-\frac{1}{2}[Y_t - X_{0t}'\beta - \sum_{j=1}^q h_j(X_{jt})]^2$. Let $\|\theta-\theta_o\|^2 = E\{X_{0t}'(\beta-\beta_o)+\sum_{j=1}^q [h_j(X_{jt}) - h_{oj}(X_{jt})]\}^2$.

Note that $D_{w^*}(X)' = X_0 - \sum_{k=1}^q w^{*k}(X_k)$, where $w^{*k}(X_k)$, $k = 1,\dots,q$, solves

$$\inf_{w^k, k=1,\dots,q:\ E[|X_0-\sum_{k=1}^q w^k(X_k)|_e^2]>0} E\left[\left(X_0 - \sum_{k=1}^q w^k(X_k)\right)\left(X_0 - \sum_{k=1}^q w^k(X_k)\right)'\right].$$

PROPOSITION 4.5. *Suppose that Assumption 3.1 and the following hold*:

(i) $\beta_o \in \operatorname{int}(B)$;

(ii) $\Sigma_o(X)$ *is positive and bounded*;

(iii) $E[X_0X_0']$ *is positive definite*; $E[D_{w^*}(X)'D_{w^*}(X)]$ *is positive definite*;

(iv) *each element of* w^{*j} *belongs to the Hölder space* Λ^{m_j} *with* $m_j > 1/2$ *for* $j = 1,\dots,q$.

Let $k_{jn} = O(n^{1/(2p_j+1)})$ *for* $j = 1,\dots,q$. *Then* $n^{1/2}(\hat\beta_n - \beta_o) \xrightarrow{d} \mathcal{N}(0, V_1^{-1}V_2V_1^{-1})$ *where* $V_1 = E[D_{w^*}(X)'D_{w^*}(X)]$, $V_2 = E[D_{w^*}(X)'\Sigma_o(X)D_{w^*}(X)]$.

PROOF. We obtain the result by applying Proposition 4.4. Let $\Theta_n = B\times\mathcal{H}_n$ and $\mathcal{H}_n = \mathcal{H}_n^1\times\cdots\times\mathcal{H}_n^q$, where $\mathcal{H}_n^j$, $j = 1,2,\dots,q$, are the same as those in Subsection 3.2.1. By the same proof as that for Proposition 3.3, we have $\|\hat\theta_n - \theta_o\| = O_P(n^{-p/(2p+1)})$ provided that $p = \min\{p_1,\dots,p_q\} > 0.5$. This and assumption (iv) imply Assumption 4.1(iii). Condition 4.3′ is trivially satisfied given the definition of the metric $\|\cdot\|$. It remains to verify Condition 4.2′:

$$\mu_n\left(\left\{X_0'[v_\beta^*] + \sum_{j=1}^q [\pi_n v_{h_j}^*(X_j)]\right\}\left\{X_0'[\beta-\beta_o] + \sum_{j=1}^q [h_j(X_j) - h_{oj}(X_j)]\right\}\right)$$
$$= o_P(n^{-1/2}),$$

uniformly over $\theta\in\Theta_n$ with $\|\theta-\theta_o\| \leqslant \delta_n = O(n^{-p/(2p+1)})$. Applying Theorem 3 in Chen, Linton and van Keilegom (2003) (or Lemma 4.2 for i.i.d. case), assumptions (i)–(iv) and Assumption 3.1 ($h_j \in \mathcal{H}^j = \Lambda_c^{m_j}$ with $m_j > 1/2$ for all $j = 1,\dots,q$) imply Condition 4.2′; also see van der Vaart and Wellner (1996). □

Notice that for the well-known partially linear regression model $Y_i = X_{0i}'\beta_o + h_{o1}(X_{1i}) + e_i$, $E[e_i|X_i] = 0$, we can explicitly solve for $D_{w^*}(X)' \equiv X_0 - w^{*1}(X_1)$ with $w^{*1}(X_1) = E\{X_0|X_1\}$. Hence assumption (iv) will be satisfied if $E\{X_0|X_1\}$ is smooth enough. See Remark 4.3 for semiparametric efficient estimation of β_o.

4.2.4. *Efficiency of sieve MLE*

Wong (1992), and Wong and Severini (1991) established asymptotic efficiency of plug-in nonparametric MLE estimates of smooth functionals. Shen (1997) extended their results to sieve MLE. We review the results of Wong (1992) and Shen (1997) in this subsection. Related work can be found in Begun et al. (1983), Ibragimov and Hasminskii (1991), Bickel et al. (1993).

Here the criterion is $\widehat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(Z_i,\theta)$, where $l(Z_i,\theta) = \log p(Z_i,\theta)$ is a log-likelihood evaluated at the single observation Z_i. We use the Fisher norm: $\|\theta-\theta_o\|^2 = E\{\frac{\partial \log p(Z_i,\theta_o)}{\partial\theta}[\theta-\theta_o]\}^2$. Recall that a probability family $\{P_\theta\colon \theta\in\Theta\}$ is *locally asymptotically normal* (LAN) at θ_o, if (1) for any g in the linear span of $\Theta-\theta_o$, $\theta_o + tn^{-1/2}g\in\Theta$ for all small $t\geqslant 0$, and (2)

$$\frac{\mathrm{d}P_{\theta_o+n^{-1/2}g}}{\mathrm{d}P_{\theta_o}}(Z_1,\dots,Z_n) = \exp\left\{\Sigma_n(g) - \frac{1}{2}\|g\|^2 + R_n(\theta_o,g)\right\},$$

where $\Sigma_n(g)$ is linear in g, $\Sigma_n(g)\xrightarrow{d}\mathcal{N}(0,\|g\|^2)$ and $\operatorname{plim}_{n\to\infty} R_n(\theta_o,g)=0$ (both limits are under the true probability measure $P_o = P_{\theta_o}$); see e.g. LeCam (1960).

To avoid the "super-efficiency" phenomenon, certain conditions on the estimates are required. In estimating a smooth functional in the infinite-dimensional case, Wong (1992, p. 58) defines the class of *pathwise regular* estimates in the sense of Bahadur (1964). An estimate $T_n(Z_1,\dots,Z_n)$ of $f(\theta_o)$ is *pathwise regular* if for any real number $\tau>0$ and any g in the linear span of $\Theta-\theta_o$, we have

$$\limsup_{n\to\infty} P_{\theta_{n,\tau}}\big(T_n < f(\theta_{n,\tau})\big) \leqslant \liminf_{n\to\infty} P_{\theta_{n,-\tau}}\big(T_n < f(\theta_{n,-\tau})\big),$$

where $\theta_{n,\tau} = \theta_o + n^{-1/2}\tau g$.

THEOREM 4.6. *[See Wong (1992), Shen (1997).] In addition to LAN, suppose the functional $f:\Theta\to\mathcal{R}$ is Frechet-differentiable at θ_o with $0<\|\frac{\partial f(\theta_o)}{\partial\theta}\|<\infty$. Then for any pathwise regular estimate T_n of $f(\theta_o)$, and any real number $\tau>0$,*

$$\limsup_{n\to\infty} P_o\big(\sqrt{n}\big|T_n - f(\theta_o)\big|\leqslant\tau\big) \leqslant P_o\left(\left|\mathcal{N}\left(0,\left\|\frac{\partial f(\theta_o)}{\partial\theta}\right\|^2\right)\right|\leqslant\tau\right)$$

where $\mathcal{N}(0,\|\frac{\partial f(\theta_o)}{\partial\theta}\|^2)$ is a scalar random variable drawn from a normal distribution with mean 0 and variance $\|\frac{\partial f(\theta_o)}{\partial\theta}\|^2$.

THEOREM 4.7. *[See Shen (1997).] In addition to the conditions to ensure $n^{1/2}(f(\hat\theta_n) - f(\theta_o))\xrightarrow{\mathrm{p}_{\theta_o}}\mathcal{N}(0,\sigma_{v^*}^2)$ with $\sigma_{v^*}^2 = \|\frac{\partial f(\theta_o)}{\partial\theta}\|^2$, if LAN holds, then for the plug-in sieve MLE estimates of $f(\theta)$, any real number $\tau>0$, and any g in the linear span of $\Theta-\theta_o$,*

$$n^{1/2}\big(f(\hat\theta_n) - f(\theta_{n,\tau})\big)\xrightarrow{\mathrm{p}_{\theta_{n,\tau}}}\mathcal{N}\big(0,\sigma_{v^*}^2\big),$$

where $\theta_{n,\tau} = \theta_o + n^{-1/2}\tau g$. Here $\xrightarrow{\mathrm{p}_\theta}$ means convergence in distribution under probability measure P_θ.

4.3. Sieve simultaneous MD estimation: Normality and efficiency

As we mentioned in Section 2.1, most structural econometric models belong to the semiparametric conditional moment framework: $E[\rho(Z, \beta_o, h_o(\cdot))|X] = 0$, where the difference $\rho(Z, \beta, h(\cdot)) - \rho(Z, \beta_o, h_o(\cdot))$ does depend on the endogenous variables Y. There are even fewer general theory papers on the sieve simultaneous MD estimation of β_o and h_o for this class of models; see Newey and Powell (1989, 2003) and Ai and Chen (1999, 2003). The sieve simultaneous MD procedure jointly estimates β_o and h_o by minimizing a sample quadratic form $\frac{1}{n}\sum_{i=1}^{n}\hat{m}(X_i, \beta, h)'[\widehat{\Sigma}(X_i)]^{-1}\hat{m}(X_i, \beta, h)$ over the sieve parameter space $\Theta_n = B \times \mathcal{H}_n$, where $\hat{m}(X_i, \beta, h)$ is any nonparametric estimator of the conditional mean function $m(X, \beta, h) \equiv E[\rho(Z, \beta, h(\cdot))|X]$, $\widehat{\Sigma}(X) \to \Sigma(X)$ in probability and $\Sigma(X)$ is a positive definite weighting matrix. Ai and Chen (1999, 2003) established the $\sqrt{n}$-asymptotic normality of this sieve MD estimator $\hat{\beta}$ of β_o.

For semiparametric efficient estimation of β_o, Ai and Chen (1999) proposed the three-step optimally weighted sieve MD procedure:

Step 1. Obtain an initial consistent sieve MD estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n)$ by

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

where $\hat{m}(X_i, \theta)$ is any nonparametric estimator of the conditional mean function $m(X, \theta) \equiv E[\rho(Z, \beta, h(\cdot))|X]$.

Step 2. Obtain a consistent estimator $\widehat{\Sigma}_o(X)$ of the optimal weighting matrix $\Sigma_o(X) \equiv \mathrm{Var}[\rho(Z, \beta_o, h_o(\cdot))|X]$ using $\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n)$ and any nonparametric regression procedures (such as kernel, nearest-neighbor or series LS estimation).

Step 3. Obtain the optimally weighted estimator $\tilde{\theta}_n = (\tilde{\beta}_n, \tilde{h}_n)$ by solving

$$\min_{\theta=(\beta,h)\in B\times\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^{n}\hat{m}(X_i, \theta)'\big[\widehat{\Sigma}_o(X_i)\big]^{-1}\hat{m}(X_i, \theta).$$

As an alternative way to efficiently estimate β_o, Ai and Chen (2003) proposed the locally continuously updated sieve MD procedure:

Step 1. Obtain an initial consistent sieve MD estimator $\hat{\theta}_n$ by

$$\min_{\theta\in B\times\mathcal{H}_n} \sum_{i=1}^{n}\hat{m}(X_i, \theta)'\hat{m}(X_i, \theta),$$

where $\hat{m}(X_i, \theta)$ is the series LS estimator (2.15) of $m(X, \theta) \equiv E[\rho(Z, \beta, h(\cdot))|X]$.

Step 2. Obtain the optimally weighted sieve MD estimator $\tilde{\theta}_n = (\tilde{\beta}_n, \tilde{h}_n)$ by

$$\min_{\theta=(\beta,h)\in N_{on}} \frac{1}{n}\sum_{i=1}^{n}\hat{m}(X_i, \theta)'\big[\widehat{\Sigma}_o(X_i, \theta)\big]^{-1}\hat{m}(X_i, \theta),$$

where N_{on} is a shrinking neighborhood of $\theta_o = (\beta_o, h_o)$ within the sieve space $B \times \mathcal{H}_n$, and $\widehat{\Sigma}_o(X_i, \theta)$ is any nonparametric estimator of the conditional variance function $\Sigma_o(X, \theta) \equiv \mathrm{Var}[\rho(Z, \beta, h(\cdot))|X]$. To compute this Step 2 one could use $\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n)$ from Step 1 as a starting point.

While Ai and Chen (1999) consider kernel estimation of the conditional mean $m(\cdot, \theta)$ and the conditional variance $\Sigma_o(\cdot, \theta)$, Ai and Chen (2003) propose series LS estimation of $m(\cdot, \theta)$ and $\Sigma_o(\cdot, \theta)$. Let $\{p_{0j}(X), j = 1, 2, \ldots, k_{m,n}\}$ be a sequence of known basis functions that can approximate any real-valued square integrable functions of X well as $k_{m,n} \to \infty$, $p^{k_{m,n}}(X) = (p_{01}(X), \ldots, p_{0k_{m,n}}(X))'$ and $P = (p^{k_{m,n}}(X_1), \ldots, p^{k_{m,n}}(X_n))'$. Then a series LS estimator of the conditional variance $\Sigma_o(X, \theta) \equiv \mathrm{Var}[\rho(Z, \theta)|X]$ is

$$\widehat{\Sigma}_o(X, \theta) \equiv \sum_{i=1}^{n} \rho(Z_i, \theta)\rho(Z_i, \theta)' p^{k_{m,n}}(X_i)'(P'P)^{-1} p^{k_{m,n}}(X).$$

Also, $\Sigma_o(X) = \mathrm{Var}[\rho(Z, \theta_o)|X]$ can be simply estimated by $\widehat{\Sigma}_o(X) \equiv \widehat{\Sigma}_o(X, \hat{\theta}_n)$.

We state the following result on semiparametric efficient estimation of β_o for the class of conditional moment restrictions $E[\rho(Z, \beta_o, h_o(\cdot))|X] = 0$; see Ai and Chen (2003) for details. For $j = 1, \ldots, d_\beta$, let

$$\begin{aligned} D_{w_j}(X) &\equiv \left.\frac{\partial E\{\rho(Z, \beta, h_o(\cdot))|X\}}{\partial \beta_j}\right|_{\beta=\beta_o} - \left.\frac{\partial E\{\rho(X, \beta_o, h_o(\cdot) + \tau w_j(\cdot))|X\}}{\partial \tau}\right|_{\tau=0} \\ &\equiv \frac{\partial m(X, \theta_o)}{\partial \beta_j} - \frac{\partial m(X, \theta_o)}{\partial h}[w_j], \\ E\big\{D_{w_{oj}}(X)'\Sigma_o(X)^{-1} D_{w_{oj}}(X)\big\} &= \inf_{w_j} E\big\{D_{w_j}(X)'\Sigma_o(X)^{-1} D_{w_j}(X)\big\}, \end{aligned}$$

$w_o = (w_{o1}, \ldots, w_{od_\beta})$, and $D_{w_o}(X) \equiv (D_{w_{o1}}(X), \ldots, D_{w_{od_\beta}}(X))$ be a $(d_\rho \times d_\beta)$-matrix valued measurable function of X.

THEOREM 4.8. *Let $\tilde{\beta}_n$ be either the three-step optimally weighted sieve MD estimator or the two-step locally continuously updated sieve MD estimator. Under the conditions stated in Ai and Chen (2003, Theorems 6.1 and 6.2), $\tilde{\beta}_n$ is semiparametric efficient and satisfies $\sqrt{n}(\tilde{\beta}_n - \beta_o) \xrightarrow{d} \mathcal{N}(0, V_o^{-1})$, with*

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

Ai and Chen (2003) also provide a simple consistent estimator, $\widehat{V}_o^{-1}$, for the asymptotic variance V_o^{-1} of $\tilde{\beta}_n$, where

$$\begin{aligned} \widehat{V}_o = \frac{1}{n}\sum_{i=1}^{n} &\left(\frac{\partial \hat{m}(X_i, \tilde{\theta}_n)}{\partial \beta'} - \frac{\partial \hat{m}(X_i, \tilde{\theta}_n)}{\partial h}[\hat{w}_o]\right)' \\ &\times \big\{\widehat{\Sigma}_o(X_i)\big\}^{-1}\left(\frac{\partial \hat{m}(X_i, \tilde{\theta}_n)}{\partial \beta'} - \frac{\partial \hat{m}(X_i, \tilde{\theta}_n)}{\partial h}[\hat{w}_o]\right), \end{aligned}$$

$\hat{w}_o = (\hat{w}_{o1}, \ldots, \hat{w}_{od_\beta})$ solves the following sieve minimization problem:

$$\min_{w_j \in \mathcal{H}_n} \sum_{i=1}^{n} \left(\frac{\partial \hat{m}(X_i, \tilde{\theta}_n)}{\partial \beta_j} - \frac{\partial \hat{m}(X_i, \tilde{\theta}_n)}{\partial h}[w_j] \right)' \left[\widehat{\Sigma}_o(X_i) \right]^{-1} \times \left(\frac{\partial \hat{m}(X_i, \tilde{\theta}_n)}{\partial \beta_j} - \frac{\partial \hat{m}(X_i, \tilde{\theta}_n)}{\partial h}[w_j] \right)$$

for $j = 1, \ldots, d_\beta$, and

$$\frac{\partial \hat{m}(X, \theta)}{\partial \beta_j} - \frac{\partial \hat{m}(X, \theta)}{\partial h}[w_j] \equiv \sum_{i=1}^{n} \left(\frac{\partial \rho(Z_i, \theta)}{\partial \beta_j} - \frac{\partial \rho(Z_i, \theta)}{\partial h}[w_j] \right) p^{k_{m,n}}(X_i)'(P'P)^{-1} p^{k_{m,n}}(X).$$

REMARK 4.3. (1) Recently, Chen and Pouzo (2006) have extended the root-n normality and efficiency results of Ai and Chen (2003) to allow that the generalized residual functions $\rho(Z, \beta, h(\cdot))$ are not pointwise continuous in $\theta = (\beta, h)$.

(2) The three-step optimally weighted sieve MD leads to semiparametric efficient estimation of β_o for the model $E[\rho(Z, \beta_o, h_o(\cdot))|X] = 0$ regardless of whether $\rho(Z, \beta, h(\cdot)) - \rho(Z, \beta_o, h_o(\cdot))$ depends on the endogenous variables Y or not. However, when $\rho(Z, \beta, h(\cdot)) - \rho(Z, \beta_o, h_o(\cdot))$ does not depend on Y, to obtain an efficient estimator of β_o one can also apply the following simpler three-step sieve GLS procedure as suggested in Ai and Chen (1999):

Step 1. Obtain an initial consistent sieve GLS estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n)$ by

$$\min_{(\beta,h) \in B \times \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \rho\big(Z_i, \beta, h(\cdot)\big)' \rho\big(Z_i, \beta, h(\cdot)\big).$$

Step 2. Obtain a consistent estimator $\widehat{\Sigma}_o(X)$ of $\Sigma_o(X) = \mathrm{Var}[\rho(Z, \theta_o)|X]$ using $\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n)$ and any nonparametric regression procedures such as $\widehat{\Sigma}_o(X) = \widehat{\Sigma}_o(X, \hat{\theta}_n)$.

Step 3. Obtain the optimally weighted GLS estimator $\tilde{\theta}_n = (\tilde{\beta}_n, \tilde{h}_n)$ by solving

$$\min_{(\beta,h) \in B \times \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \rho\big(Z_i, \beta, h(\cdot)\big)' \big[\widehat{\Sigma}_o(X_i)\big]^{-1} \rho\big(Z_i, \beta, h(\cdot)\big).$$

That is, for all the models belonging to the first subclass of the conditional moment restrictions (2.8), $E\{\rho(Z, \beta_o, h_o)|X\} = 0$, where $\rho(Z, \theta) - \rho(Z, \theta_o)$ does not depend on endogenous variables Y, the simple three-step sieve GLS estimator $\tilde{\beta}_n$ also satisfies $\sqrt{n}(\tilde{\beta}_n - \beta_o) \xrightarrow{d} \mathcal{N}(0, V_o^{-1})$. Of course, the following continuously updated sieve GLS procedure will also lead to semiparametric efficient estimation of β_o:

$$
\begin{aligned}
&(\tilde{\beta}_{\mathrm{cgls}}, \tilde{h}_{\mathrm{cgls}}) \\
&\quad = \arg \min_{(\beta,h)\in B\times\mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \rho\big(Z_i, \beta, h(\cdot)\big)' \big[\widehat{\Sigma}_o\big(X_i, \beta, h(\cdot)\big)\big]^{-1} \rho\big(Z_i, \beta, h(\cdot)\big).
\end{aligned}
$$

For the conditional moment restriction (without unknown function h_o), $E[\rho(Z, \beta_o)| X] = 0$, there are many alternative efficient estimation procedures for β_o, including the empirical likelihood of Donald, Imbens and Newey (2003), the generalized empirical likelihood (GEL) of Newey and Smith (2004), the kernel-based empirical likelihood of Kitamura, Tripathi and Ahn (2004), the continuously updated minimum distance procedure or the Euclidean conditional empirical likelihood of Antoine, Bonnal and Renault (2007), among others. It seems that one could extend their results to the more general conditional moment framework $E[\rho(Z, \beta_o, h_o())|X] = 0$, where the unknown function $h_o()$ is approximated by a sieve. In fact, Zhang and Gijbels (2003) have already considered the sieve empirical likelihood procedure for the special case $E[\rho(Z, \beta_o, h_o(X))|X] = 0$ where h_o is a function of conditioning variable X only; See Otsu (2005) for the general case.

Recently Ai and Chen (2007, 2004) have considered the semiparametric conditional moment framework $E[\rho_j(Z, \beta_o, h_o())|X_j] = 0$ for $j = 1, \ldots, J$ with finite J, where each conditional moment has its own conditioning set X_j that could differ across equations. This extension would be useful to estimating semiparametric structure models with incomplete information.

5. Concluding remarks

In this chapter, we have surveyed some recent large sample results on nonparametric and semiparametric estimation of econometric models via the method of sieves. We have restricted our attention to general consistency and convergence rates of sieve estimation of unknown functions and $\sqrt{n}$-asymptotic normality of sieve estimation of smooth functionals. Examples were used to illustrate the general sieve estimation theory. It is our hope that the examples adequately depicted the general sieve extremum estimation approach and its versatility. We conclude this chapter by pointing out additional topics on the method of sieves that have not been reviewed for lack of time and space.

First, although there is still lack of general theory on testing via the sieve method, there are some consistent specification tests using the method of sieves. For example, Hong and White (1995) tested a parametric regression model using series LS estimators; Hart (1997) presented many consistent tests using series estimators; Stinchcombe and White (1998) tested a parametric conditional moment restriction $E[\rho(Z, \beta_o)| X] = 0$ using neural network sieves and Li, Hsiao and Zinn (2003) tested semiparametric/nonparametric regression models using spline series estimators. Most recently Song (2005) proposed consistent tests of semi-nonparametric regression models via conditional martingale transforms where the unknown functions are estimated by the method of sieves. Additional references include Wooldridge (1992), Bierens (1990), Bierens and Ploberger (1997) and de Jong (1996). Also in principle, all of the existing test results based on kernel or local linear regression methods such as those in Robinson (1989), Fan and Li (1996), Lavergne and Vuong (1996), Chen and Fan (1999), Fan and Linton (1999), Aït-Sahalia, Bickel and Stoker (2001), Horowitz and Spokoiny (2001) and Fan, Zhang and Zhang (2001) can be done using the method of sieves.

Second, we have not touched on the issue of data-driven selection of sieve spaces. In practice, many existing model selection methods such as cross-validation (CV), generalized CV and AIC have been used in the current context due to the connection of the method of sieves with the parametric models; see the survey chapter by Ichimura and Todd (2007) on implementation details of semi-nonparametric estimators including series estimators, and the review by Stone et al. (1997) and Ruppert, Wand and Carroll (2003) on model selection with spline sieves for extended linear models. There are a few papers in statistics including Barron, Birgé and Massart (1999) and Shen and Ye (2002) that address data-driven selection among different sieve bases. There are many results on data-driven selection of the number of terms for a given sieve basis; see e.g. Li (1987), Andrews (1991a), Hurvich, Simonoff and Tsai (1998), Donald and Newey (2001), Coppejans and Gallant (2002), Phillips and Ploberger (2003), Fan and Peng (2004) and Imbens, Newey and Ridder (2005). In particular, Andrews (1991a) establishes the asymptotic optimality of CV as a method to select series terms for nonparametric least square regressions with heteroskedastic errors. Imbens, Newey and Ridder (2005) establishes a similar result for semiparametrically efficient estimation of average treatment effect parameters with a first step series estimation of conditional means. It would be very useful to extend their results to handle a more general class of semi-nonparametric models estimated via the method of sieves.

Third, so far there is little research on the higher order refinements of the large sample properties of the semiparametric efficient sieve estimators. Many authors, including Linton (1995) and Heckman et al. (1998), have pointed out that the first-order asymptotics of semiparametric procedures could be misleading and unhelpful. For the case of kernel estimators, some papers such as Robinson (1995), Linton (1995, 2001), Nishiyama and Robinson (2000, 2005), Xiao and Linton (2001) and Ichimura and Linton (2002) have obtained higher order refinements. It would be useful to extend these results to semiparametric efficient estimators using the method of sieves.

Finally, given the relative ease of implementation of the sieve method, but the general difficulty of deriving its large sample properties, it might be fruitful to combine the sieve method with the kernel or the local linear regression methods [see e.g. Fan and Gijbels (1996)]. Recent papers by Horowitz and Mammen (2004) and Horowitz and Lee (2005) have demonstrated the usefulness of this combination.

References

Ai, C. (1997). "A semiparametric maximum likelihood estimator". Econometrica 65, 933–964.

Ai, C., Chen, X. (2003). "Efficient estimation of models with conditional moment restrictions containing unknown functions". Econometrica 71, 1795–1843. Working paper version, 1999.

Ai, C., Chen, X. (1999). "Efficient sieve minimum distance estimation of semiparametric conditional moment models". Manuscript. London School of Economics.
Ai, C., Chen, X. (2004). "On efficient sequential estimation of semi-nonparametric moment models". Working paper. New York University.
Ai, C., Chen, X. (2007). "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables". Journal of Econometrics. In press.
Aït-Sahalia, Y., Bickel, P., Stoker, T. (2001). "Goodness-of-fit tests for kernel regression with an application to option implied volatilities". Journal of Econometrics 105, 363–412.
Amemiya, T. (1985). Advanced Econometrics. Harvard University Press, Cambridge.
Anastassiou, G., Yu, X. (1992a). "Monotone and probabilistic wavelet approximation". Stochastic Analysis and Applications 10, 251–264.
Anastassiou, G., Yu, X. (1992b). "Convex and convex-probabilistic wavelet approximation". Stochastic Analysis and Applications 10, 507–521.
Andrews, D. (1991a). "Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors". Journal of Econometrics 47, 359–377.
Andrews, D. (1991b). "Asymptotic normality of series estimators for nonparametric and semiparametric regression models". Econometrica 59, 307–345.
Andrews, D. (1992). "Generic uniform convergence". Econometric Theory, 241–257.
Andrews, D. (1994a). "Empirical process method in econometrics". In: Engle III, R.F., McFadden, D.F. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.
Andrews, D. (1994b). "Asymptotics for semi-parametric econometric models via stochastic equicontinuity". Econometrica 62, 43–72.
Andrews, D., Schafgans, M. (1998). "Semiparametric estimation of the intercept of a sample selection model". Review of Economic Studies 65, 497–517.
Andrews, D., Whang, Y. (1990). "Additive interactive regression models: Circumvention of the curse of dimensionality". Econometric Theory 6, 466–479.
Antoine, B., Bonnal, H., Renault, E. (2007). "On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood". Journal of Econometrics 138, 488–512.
Bahadur, R.R. (1964). "On Fisher's bound for asymptotic variances". Ann. Math. Statist. 35, 1545–1552.
Bansal, R., Viswanathan, S. (1993). "No arbitrage and arbitrage pricing: A new approach". The Journal of Finance 48 (4), 1231–1262.
Bansal, R., Hsieh, D., Viswanathan, S. (1993). "A new approach to international arbitrage pricing". The Journal of Finance 48, 1719–1747.
Barnett, W.A., Powell, J., Tauchen, G. (1991). Non-parametric and Semi-parametric Methods in Econometrics and Statistics. Cambridge University Press, New York.
Barron, A.R. (1993). "Universal approximation bounds for superpositions of a sigmoidal function". IEEE Trans. Information Theory 39, 930–945.
Barron, A., Birgé, L., Massart, P. (1999). "Risk bounds for model selection via penalization". Probab. Theory Related Fields 113, 301–413.
Begun, J., Hall, W., Huang, W., Wellner, J.A. (1983). "Information and asymptotic efficiency in parametric-nonparametric models". The Annals of Statistics 11, 432–452.
Bickel, P.J., Klaassen, C.A.J., Ritov, Y., Wellner, J.A. (1993). Efficient and Adaptive Estimation for Semiparametric Models. The John Hopkins University Press, Baltimore.
Bierens, H. (1990). "A consistent conditional moment test of functional form". Econometrica 58, 1443–1458.
Bierens, H. (2006). "Semi-nonparametric interval-censored mixed proportional hazard models: Identification and consistency results". Econometric Theory. In press.
Bierens, H., Carvalho, J. (2006). "Semi-nonparametric competing risks analysis of recidivism". Journal of Applied Econometrics. In press.
Bierens, H., Ploberger, W. (1997). "Asymptotic theory of integrated conditional moment tests". Econometrica 65, 1129–1151.

Birgé, L., Massart, P. (1998). "Minimum contrast estimators on sieves: Exponential bounds and rates of convergence". Bernoulli 4, 329–375.

Birman, M., Solomjak, M. (1967). "Piece-wise polynomial approximations of functions in the class W_p^α". Mathematics of the USSR Sbornik 73, 295–317.

Blundell, R., Powell, J. (2003). "Endogeneity in nonparametric and semiparametric regression models". In: Dewatripont, M., Hansen, L.P., Turnovsky, S. (Eds.), Advances in Economics and Econometrics: Theory and Applications, vol. 2. Cambridge University Press, Cambridge, pp. 312–357.

Blundell, R., Browning, M., Crawford, I. (2003). "Non-parametric Engel curves and revealed preference". Econometrica 71, 205–240.

Blundell, R., Chen, X., Kristensen, D. (2007). "Semi-nonparametric IV estimation of shape-invariant Engel curves". Econometrica. In press.

Blundell, R., Duncan, A., Pendakur, K. (1998). "Semiparametric estimation and consumer demand". Journal of Applied Econometrics 13, 435–461.

Brendstrup, B., Paarsch, H. (2004). "Identification and estimation in sequential, asymmetric, English auctions". Manuscript, University of Iowa.

Cai, Z., Fan, J., Yao, Q. (2000). "Functional-coefficient regression models for nonlinear time series". Journal of American Statistical Association 95, 941–956.

Cameron, S., Heckman, J. (1998). "Life cycle schooling and dynamic selection bias". Journal of Political Economy 106, 262–333.

Campbell, J., Cochrane, J. (1999). "By force of habit: A consumption-based explanation of aggregate stock market behavior". Journal of Political Economy 107, 205–251.

Carrasco, M., Florens, J.-P., Renault, E. (2006). "Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization". In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics, vol. 6. North-Holland, Amsterdam.

Chamberlain, G. (1992). "Efficiency bounds for semiparametric regression". Econometrica 60, 567–596.

Chapman, D. (1997). "Approximating the asset pricing kernel". The Journal of Finance 52 (4), 1383–1410.

Chen, X., Conley, T. (2001). "A new semiparametric spatial model for panel time series". Journal of Econometrics 105, 59–83.

Chen, X., Fan, Y. (1999). "Consistent hypothesis testing in semiparametric and nonparametric models for econometric time series". Journal of Econometrics 91, 373–401.

Chen, X., Ludvigson, S. (2003). "Land of addicts? An empirical investigation of habit-based asset pricing models". Manuscript. New York University.

Chen, X., Pouzo, D. (2006). "Efficient estimation of semi-nonparametric conditional moment models with possibly nonsmooth moments". Manuscript. New York University.

Chen, X., Shen, X. (1996). "Asymptotic properties of sieve extremum estimates for weakly dependent data with applications". Manuscript. University of Chicago.

Chen, X., Shen, X. (1998). "Sieve extremum estimates for weakly dependent data". Econometrica 66, 289–314.

Chen, R., Tsay, R. (1993). "Functional-coefficient autoregressive models". Journal of American Statistical Association 88, 298–308.

Chen, X., White, H. (1998). "Nonparametric adaptive learning with feedback". Journal of Economic Theory 82, 190–222.

Chen, X., White, H. (1999). "Improved rates and asymptotic normality for nonparametric neural network estimators". IEEE Tran. Information Theory 45, 682–691.

Chen, X., White, H. (2002). "Asymptotic properties of some projection-based Robbins–Monro procedures in a Hilbert space". Studies in Nonlinear Dynamics and Econometrics 6 (1). Article 1.

Chen, X., Fan, Y., Tsyrennikov, V. (2006). "Efficient estimation of semiparametric multivariate copula models". Journal of the American Statistical Association 101, 1228–1240.

Chen, X., Hansen, L.P., Scheinkman, J. (1998). "Shape-preserving estimation of diffusions". Manuscript. University of Chicago.

Chen, X., Hong, H., Tamer, E. (2005). "Measurement error models with auxiliary data". Review of Economic Studies 72, 343–366.

Chen, X., Hong, H., Tarozzi, A. (2007). "Semiparametric efficiency in GMM models of nonclassical measurement errors, missing data and treatment effects". The Annals of Statistics. In press.
Chen, X., Linton, O., van Keilegom, I. (2003). "Estimation of semiparametric models when the criterion function is not smooth". Econometrica 71, 1591–1608.
Chen, X., Racine, J., Swanson, N. (2001). "Semiparametric ARX neural network models with an application to forecasting inflation". IEEE Tran. Neural Networks 12, 674–683.
Chernozhukov, V., Imbens, G., Newey, W. (2007). "Instrumental variable identification and estimation of nonseparable models via quantile conditions". Journal of Econometrics 139, 4–14.
Chui, C. (1992). An Introduction to Wavelets. Academic Press, Inc., San Diego.
Cochrane, J. (2001). Asset Pricing. Princeton University Press, Princeton, NJ.
Constantinides, G. (1990). "Habit-formation: A resolution of the equity premium puzzle". Journal of Political Economy 98, 519–543.
Coppejans, M. (2001). "Estimation of the binary response model using a mixture of distributions estimator (MOD)". Journal of Econometrics 102, 231–261.
Coppejans, M., Gallant, A.R. (2002). "Cross-validated SNP density estimates". Journal of Econometrics 110, 27–65.
Cosslett, S. (1983). "Distribution-free maximum likelihood estimation of the binary choice model". Econometrica 51, 765–782.
Cybenko, G. (1990). "Approximation by superpositions of a sigmoid function". Mathematics of Control, Signals and Systems 2, 303–314.
Darolles, S., Florens, J.-P., Renault, E. (2002). "Nonparametric instrumental regression". Mimeo. GREMAQ, University of Toulouse.
Das, M., Newey, W.K., Vella, F. (2003). "Nonparametric estimation of sample selection models". Review of Economic Studies 70, 33–58.
Daubechies, I. (1992). Ten Lectures on Wavelets. SIAM, Philadelphia.
de Boor, C. (1978). A Practical Guide to Splines. Springer-Verlag, New York.
Dechevsky, L., Penev, S. (1997). "On shape-preserving probabilistic wavelet approximators". Stochastic Analysis and Applications 15, 187–215.
de Jong, R. (1996). "The Bierens test under data dependence". Journal of Econometrics 72, 1–32.
de Jong, R. (2002). "A note on 'Convergence rates and asymptotic normality for series estimators': Uniform convergence rates". Journal of Econometrics 111, 1–9.
DeVore, R.A. (1977a). "Monotone approximation by splines". SIAM Journal on Mathematical Analysis 8, 891–905.
DeVore, R.A. (1977b). "Monotone approximation by polynomials". SIAM Journal on Mathematical Analysis 8, 906–921.
DeVore, R.A., Lorentz, G.G. (1993). Constructive Approximation. Springer-Verlag, Berlin.
Donald, S., Newey, W. (2001). "Choosing the number of instruments". Econometrica 69, 1161–1191.
Donald, S., Imbens, G., Newey, W. (2003). "Empirical likelihood estimation and consistent tests with conditional moment restrictions". Journal of Econometrics 117, 55–93.
Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D. (1995). "Wavelet shrinkage: Asymptopia?". Journal of the Royal Statistical Society, Series B 57, 301–369.
Doukhan, P., Massart, P., Rio, E. (1995). "Invariance principles for absolutely regular empirical processes". Ann. Inst. Henri Poincaré – Probabilités et Statistiques 31, 393–427.
Duncan, G.M. (1986). "A semiparametric censored regression estimator". Journal of Econometrics 32, 5–34.
Eggermont, P., LaRiccia, V. (2001). Maximum Penalized Likelihood Estimation, Volume I: Density Estimation. Springer, New York.
Eichenbaum, M., Hansen, L.P. (1990). "Estimating models with intertemporal substitution using aggregate time series data". Journal of Business and Economic Statistics 8, 53–69.
Elbadawi, I., Gallant, A.R., Souza, G. (1983). "An elasticity can be estimated consistently without a prior knowledge of functional form". Econometrica 51, 1731–1751.
Engle, R., Gonzalez-Rivera, G. (1991). "Semiparametric ARCH models". Journal of Business and Economic Statistics 9, 345–359.

Engle, R.F., McFadden, D.L. (Eds.) (1994). Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.
Engle, R., Rangel, G. (2004). "The spline GARCH model for unconditional volatility and its global macroeconomic causes". Working paper. New York University.
Engle, R., Granger, C., Rice, J., Weiss, A. (1986). "Semiparametric estimates of the relation between weather and electricity sales". Journal of the American Statistical Association 81, 310–320.
Fan, J., Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. Chapman and Hall, London.
Fan, Y., Li, Q. (1996). "Consistent model specification tests: Omitted variables, parametric and semiparametric functional forms". Econometrica 64, 865–890.
Fan, Y., Linton, O. (1999). "Some higher order theory for a consistent nonparametric model specification test". Working paper LSE.
Fan, J., Peng, H. (2004). "On non-concave penalized likelihood with diverging number of parameters". The Annals of Statistics 32, 928–961.
Fan, J., Yao, Q. (2003). Nonlinear Time Series: Nonparametric and Parametric Methods. Springer-Verlag, New York.
Fan, J., Zhang, C., Zhang, J. (2001). "Generalized likelihood ratio statistics and Wilks phenomenon". The Annals of Statistics 29, 153–193.
Flinn, C., Heckman, J. (1982). "New methods for analyzing structural models of labor force dynamics". Journal of Econometrics 18, 115–168.
Florens, J.P. (2003). "Inverse problems and structural econometrics: The example of instrumental variables". In: Dewatripont, M., Hansen, L.P., Turnovsky, S. (Eds.), Advances in Economics and Econometrics: Theory and Applications, vol. 2. Cambridge University Press, Cambridge, pp. 284–311.
Gabushin, O. (1967). "Inequalities for norms of functions and their derivatives in the L_p metric". Matematicheskie Zametki 1, 291–298.
Gallant, A.R. (1987). "Identification and consistency in seminonparametric regression". In: Bewley, T.F. (Ed.), Advances in Econometrics, vol. I. Cambridge University Press, pp. 145–170.
Gallant, A.R., Nychka, D. (1987). "Semi-non-parametric maximum likelihood estimation". Econometrica 55, 363–390.
Gallant, A.R., Souza, G. (1991). "On the asymptotic normality of Fourier flexible form estimates". Journal of Econometrics 50, 329–353.
Gallant, A.R., Tauchen, G. (1989). "Semiparametric estimation of conditional constrained heterogenous processes: Asset pricing applications". Econometrica 57, 1091–1120.
Gallant, A.R., Tauchen, G. (1996). "Which moments to match?". Econometric Theory 12, 657–681.
Gallant, A.R., Tauchen, G. (2004). "EMM: A program for efficient method of moments estimation, Version 2.0 User's Guide". Working paper. Duke University.
Gallant, A.R., White, H. (1988a). "There exists a neural network that does not make avoidable mistakes". In: Proceedings of the IEEE 1988 International Conference on Neural Networks, vol. 1. IEEE, New York, pp. 657–664.
Gallant, A.R., White, H. (1988b). A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models. Basil Blackwell, Oxford.
Gallant, A.R., White, H. (1992). "On learning the derivatives of an unknown mapping with multilayer feedforward networks". Neural Networks 5, 129–138.
Gallant, A.R., Hsieh, D., Tauchen, G. (1991). "On fitting a recalcitrant series: The pound/dollar exchange rate, 1974–83". In: Barnett, W.A., Powell, J., Tauchen, G. (Eds.), Non-parametric and Semi-parametric Methods in Econometrics and Statistics. Cambridge University Press, Cambridge, pp. 199–240.
Geman, S., Hwang, C. (1982). "Nonparametric maximum likelihood estimation by the method of sieves". The Annals of Statistics 10, 401–414.
Girosi, F. (1994). "Regularization theory, radial basis functions and networks". In: Cherkassky, V., Friedman, J.H., Wechsler, H. (Eds.), From Statistics to Neural Networks. Theory and Pattern Recognition Applications. Springer-Verlag, Berlin.
Granger, C.W.J., Teräsvirta, T. (1993). Modelling Nonlinear Economic Relationships. Oxford University Press, New York.

Grenander, U. (1981). Abstract Inference. Wiley Series, New York.
Härdle, W., Linton, O. (1994). "Applied nonparametric methods". In: Engle III, R.F., McFadden, D.F. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.
Härdle, W., Mueller, M., Sperlich, S., Werwatz, A. (2004). Nonparametric and Semiparametric Models. Springer, New York.
Hahn, J. (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects". Econometrica 66, 315–332.
Hall, P., Horowitz, J. (2005). "Nonparametric methods for inference in the presence of instrumental variables". The Annals of Statistics 33, 2904–2929.
Hansen, L.P. (1982). "Large sample properties of generalized method of moments estimators". Econometrica 50, 1029–1054.
Hansen, L.P. (1985). "A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators". Journal of Econometrics 30, 203–238.
Hansen, M.H. (1994). "Extended linear models, multivariate splines, and ANOVA". PhD Dissertation. Department of Statistics, University of California at Berkeley.
Hansen, L.P., Richard, S. (1987). "The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models". Econometrica 55, 587–613.
Hansen, L.P., Singleton, K. (1982). "Generalized instrumental variables estimation of nonlinear rational expectations models". Econometrica 50, 1269–1286.
Hart, J. (1997). Nonparametric Smoothing and Lack-of-Fit Tests. Springer-Verlag, New York.
Hausman, J., Newey, W. (1995). "Nonparametric estimation of exact consumer surplus and deadweight loss". Econometrica 63, 1445–1467.
Heckman, J. (1979). "Sample selection bias as a specification error". Econometrica 47, 153–161.
Heckman, J., Singer, B. (1984). "A method for minimizing the impact of distributional assumptions in econometric models for duration data". Econometrica 68, 839–874.
Heckman, J., Willis, R. (1977). "A beta logistic model for the analysis of sequential labor force participation of married women". Journal of Political Economy 85, 27–58.
Heckman, J., Ichimura, H., Smith, J., Todd, P. (1998). "Characterization of selection bias using experimental data". Econometrica 66, 1017–1098.
Hirano, K., Imbens, G., Ridder, G. (2003). "Efficient estimation of average treatment effects using the estimated propensity score". Econometrica 71, 1161–1189.
Hong, Y., White, H. (1995). "Consistent specification testing via nonparametric series regression". Econometrica 63, 1133–1159.
Honoré, B. (1990). "Simple estimation of a duration model with unobserved heterogeneity". Econometrica 58, 453–473.
Honoré, B. (1994). "A note on the rate of convergence of estimators of mixtures of Weibulls". Manuscript. Northwestern University.
Honoré, B., Kyriazidou, E. (2000). "Panel data discrete choice models with lagged dependent variables". Econometrica 68, 839–874.
Hornik, K., Stinchcombe, M., White, H. (1989). "Multilayer feedforward networks are universal approximators". Neural Networks 2, 359–366.
Hornik, K., Stinchcombe, M., White, H., Auer, P. (1994). "Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives". Neural Computation 6, 1262–1275.
Horowitz, J. (1992). "A smoothed maximum score estimator for the binary response model". Econometrica 60, 505–531.
Horowitz, J. (1998). Semiparametric Methods in Econometrics. Springer-Verlag, New York.
Horowitz, J., Lee, S. (2005). "Nonparametric estimation of an additive quantile regression model". Journal of the American Statistical Association 100, 1238–1249.
Horowitz, J., Lee, S. (2007). "Nonparametric instrumental variables estimation of a quantile regression model". Econometrica 75, 1191–1208.
Horowitz, J., Mammen, E. (2004). "Nonparametric estimation of an additive model with a link function". The Annals of Statistics 32, 2412–2443.

Horowitz, J., Spokoiny, V. (2001). "An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative". Econometrica 69, 599–631.

Hu, Y., Schennach, S. (2006). "Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions using instruments". Working paper. University of Texas, Austin.

Huang, J.Z. (1998a). "Projection estimation in multiple regression with application to functional ANOVA models". The Annals of Statistics 26, 242–272.

Huang, J.Z. (1998b). "Functional ANOVA models for generalized regression". Journal of Multivariate Analysis 67, 49–71.

Huang, J.Z. (2001). "Concave extended linear modeling: A theoretical synthesis". Statistica Sinica 11, 173–197.

Huang, J.Z. (2003). "Local asymptotics for polynomial spline regression". The Annals of Statistics 31, 1600–1635.

Huang, J.Z., Kooperberg, C., Stone, C.J., Truong, Y.K. (2000). "Functional ANOVA modeling for proportional hazards regression". The Annals of Statistics 28, 960–999.

Hurvich, C., Simonoff, J., Tsai, C. (1998). "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion". Journal of the Royal Statistical Society, Series B 60, 271–293.

Hutchinson, J., Lo, A., Poggio, T. (1994). "A non-parametric approach to pricing and hedging derivative securities via learning networks". The Journal of Finance 3, 851–889.

Ibragimov, I.A., Hasminskii, R.Z. (1991). "Asymptotically normal families of distributions and efficient estimation". The Annals of Statistics 19, 1681–1724.

Ichimura, H. (1993). "Semiparametric least squares (SLS), and weighted SLS estimation of single index models". Journal of Econometrics 58, 71–120.

Ichimura, H., Lee, S. (2006). "Characterization of the asymptotic distribution of semiparametric M-estimators". Manuscript. UCL.

Ichimura, H., Linton, O. (2002). "Asymptotic expansions for some semiparametric program evaluation estimators". Working paper IFS and LSE.

Ichimura, H., Todd, P. (2007). "Implementing nonparametric and semiparametric estimators". In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier. Chapter 74.

Imbens, G., Newey, W., Ridder, G. (2005). "Mean-squared-error calculations for average treatment effects". Manuscript. UC Berkeley.

Ishwaran, H. (1996a). "Identification and rates of estimation for scale parameters in location mixture models". The Annals of Statistics 24, 1560–1571.

Ishwaran, H. (1996b). "Uniform rates of estimation in the semiparametric Weibull mixture models". The Annals of Statistics 24, 1572–1585.

Jovanovic, B. (1979). "Job matching and the theory of turnover". Journal of Political Economy 87, 972–990.

Judd, K. (1998). Numerical Method in Economics. MIT University Press.

Khan, S. (2005). "An alternative approach to semiparametric estimation of heteroskedastic binary response models". Manuscript. University of Rochester.

Kim, J., Pollard, D. (1990). "Cube root asymptotics". The Annals of Statistics 18, 191–219.

Kitamura, Y., Tripathi, G., Ahn, H. (2004). "Empirical likelihood-based inference in conditional moment restriction models". Econometrica 72, 1667–1714.

Klein, R., Spady, R. (1993). "An efficient semiparametric estimator for binary response models". Econometrica 61, 387–421.

Koenker, R., Bassett, G. (1978). "Regression quantiles". Econometrica 46, 33–50.

Koenker, R., Mizera, I. (2003). "Penalized triograms: Total variation regularization for bivariate smoothing". Journal of the Royal Statistical Society, Series B 66, 145–163.

Koenker, R., Ng, P., Portnoy, S. (1994). "Quantile smoothing splines". Biometrika 81, 673–680.

Kooperberg, C., Stone, C.J., Truong, Y.K. (1995a). "Hazard regression". Journal of the American Statistical Association 90, 78–94.

Kooperberg, C., Stone, C.J., Truong, Y.K. (1995b). "Rate of convergence for logspline spectral density estimation". Journal of Time Series Analysis 16, 389–401.

Lavergne, P., Vuong, Q. (1996). "Nonparametric selection of regressors: The nonnested case". Econometrica 64, 207–219.
LeCam, L. (1960). "Local asymptotically normal families of distributions". Univ. California Publications in Statist. 3, 37–98.
Lee, S. (2003). "Efficient semiparametric estimation of a partially linear quantile regression model". Econometric Theory 19, 1–31.
Li, K. (1987). "Asymptotic optimality for C_p, C_L cross-validation, and generalized cross-validation: Discrete index set". The Annals of Statistics 15, 958–975.
Li, Q., Racine, J. (2007). Nonparametric Econometrics Theory and Practice. Princeton University Press. In press.
Li, Q., Hsiao, C., Zinn, J. (2003). "Consistent specification tests for semiparametric/nonparametric models based on series estimation methods". Journal of Econometrics 112, 295–325.
Linton, O. (1995). "Second order approximation in the partially linear regression model". Econometrica 63, 1079–1112.
Linton, O. (2001). "Edgeworth approximations for semiparametric instrumental variable estimators and test statistics". Journal of Econometrics 106, 325–368.
Linton, O., Mammen, E. (2005). "Estimating semiparametric ARCH(∞) models by kernel smoothing methods". Econometrica 73, 771–836.
Lorentz, G. (1966). Approximation of Functions. Holt, New York.
Mahajan, A. (2004). "Identification and estimation of single index models with misclassified regressors". Manuscript. Stanford University.
Makovoz, Y. (1996). "Random approximants and neural networks". Journal of Approximation Theory 85, 98–109.
Manski, C. (1985). "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator". Journal of Econometrics 27, 313–334.
Manski, C. (1994). "Analog estimation of econometric models". In: Engle III, R.F., McFadden, D.F. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.
Matzkin, R.L. (1994). "Restrictions of economic theory in nonparametric methods". In: Engle III, R.F., McFadden, D.F. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.
McCaffrey, D., Ellner, S., Gallant, A., Nychka, D. (1992). "Estimating the Lyapunov exponent of a chaotic system with nonparametric regression". Journal of the American Statistical Association 87, 682–695.
Meyer, Y. (1992). Ondelettes et operateurs I: Ondelettes. Hermann, Paris.
Murphy, S., van der Vaart, A. (2000). "On profile likelihood". Journal of the American Statistical Association 95, 449–465.
Newey, W.K. (1990a). "Semiparametric efficiency bounds". Journal of Applied Econometrics 5, 99–135.
Newey, W.K. (1990b). "Efficient instrumental variables estimation of nonlinear models". Econometrica 58, 809–837.
Newey, W.K. (1991). "Uniform convergence in probability and stochastic equicontinuity". Econometrica 59, 1161–1167.
Newey, W.K. (1993). "Efficient estimation of models with conditional moment restrictions". In: Maddala, G.S., Rao, C.R., Vinod, H.D. (Eds.), Handbook of Statistics, vol. 11. North-Holland, Amsterdam.
Newey, W.K. (1994a). "The asymptotic variance of semiparametric estimators". Econometrica 62, 1349–1382.
Newey, W.K. (1994b). "Series estimation of regression functionals". Econometric Theory 10, 1–28.
Newey, W.K. (1997). "Convergence rates and asymptotic normality for series estimators". Journal of Econometrics 79, 147–168.
Newey, W.K. (2001). "Flexible simulated moment estimation of nonlinear errors in variables models". Review of Economics and Statistics 83, 616–627.
Newey, W.K. (1988). "Two step series estimation of sample selection models". Manuscript. MIT Department of Economics.
Newey, W.K., McFadden, D.L. (1994). "Large sample estimation and hypothesis testing". In: Engle III, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.

Newey, W.K., Powell, J.L. (1989). “Nonparametric instrumental variable estimation”. Manuscript. Princeton University.
Newey, W.K., Powell, J.L. (2003). “Instrumental variable estimation of nonparametric models”. Econometrica 71, 1565–1578. Working paper version, 1989.
Newey, W.K., Smith, R. (2004). “Higher order properties of GMM and generalized empirical likelihood estimators”. Econometrica 72, 219–256.
Newey, W.K., Powell, J.L., Vella, F. (1999). “Nonparametric estimation of triangular simultaneous equations models”. Econometrica 67, 565–603.
Nishiyama, Y., Robinson, P.M. (2000). “Edgeworth expansions for semiparametric averaged derivatives”. Econometrica 68, 931–980.
Nishiyama, Y., Robinson, P.M. (2005). “The bootstrap and the Edgeworth correction for semiparametric averaged derivatives”. Econometrica 73, 903–980.
Ossiander, M. (1987). “A central limit theorem under metric entropy with L_2 bracketing”. The Annals of Probability 15, 897–919.
Otsu, T. (2005). “Sieve conditional empirical likelihood estimation of semiparametric models”. Manuscript. Yale University.
Pagan, A., Ullah, A. (1999). Nonparametric Econometrics. Cambridge University Press.
Pakes, A., Olley, G.S. (1995). “A limit theorem for a smooth class of semiparametric estimators”. Journal of Econometrics 65, 295–332.
Pastorello, S., Patilea, V., Renault, E. (2003). “Iterative and recursive estimation in structural non-adaptive models”. Journal of Business & Economic Statistics 21, 449–509.
Phillips, P.C.B. (1998). “New tools for understanding spurious regressions”. Econometrica 66, 1299–1325.
Phillips, P.C.B., Ploberger, W. (2003). “An introduction to best empirical models when the parameter space is infinite-dimensional”. Oxford Bulletin of Economics and Statistics 65, 877–890.
Pinkse, J. (2000). “Nonparametric two-step regression estimation when regressors and errors are dependent”. Canadian Journal of Statistics 28, 289–300.
Polk, C., Thompson, T.S., Vuolteenaho, T. (2003). “New forecasts of the equity premium”. Manuscript. Harvard University.
Pollard, D. (1984). Convergence of Statistical Processes. Springer-Verlag, New York.
Portnoy, S. (1997). “Local asymptotics for quantile smoothing splines”. The Annals of Statistics 25, 387–413.
Powell, J. (1994). “Estimation of semiparametric models”. In: Engle III, R.F., McFadden, D.F. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.
Powell, J., Stock, J., Stoker, T. (1989). “Semiparametric estimation of index coefficients”. Econometrica 57, 1403–1430.
Robinson, P. (1988). “Root-N-consistent semiparametric regression”. Econometrica 56, 931–954.
Robinson, P. (1989). “Hypothesis testing in semiparametric and nonparametric models for econometric time series”. Review of Economic Studies 56, 511–534.
Robinson, P. (1995). “The normal approximation for semiparametric averaged derivatives”. Econometrica 63, 667–680.
Ruppert, D., Wand, M., Carroll, R. (2003). Semiparametric Regression. Cambridge University Press, Cambridge.
Schumaker, L. (1981). Spline Functions: Basic Theory. John Wiley & Sons, New York.
Severini, T., Wong, W.H. (1992). “Profile likelihood and conditionally parametric models”. The Annals of Statistics 20, 1768–1802.
Shen, X. (1997). “On methods of sieves and penalization”. The Annals of Statistics 25, 2555–2591.
Shen, X., Wong, W. (1994). “Convergence rate of sieve estimates”. The Annals of Statistics 22, 580–615.
Shen, X., Ye, J. (2002). “Adaptive model selection”. Journal of the American Statistical Association 97, 210–221.
Shintani, M., Linton, O. (2004). “Nonparametric neural network estimation of Lyapunov exponents and a direct test for chaos”. Journal of Econometrics 120, 1–34.
Song, K. (2005). “Testing semiparametric conditional moment restrictions using conditional martingale transforms”. Manuscript. Yale University, Department of Economics.

Stinchcombe, M. (2002). “Some genericity analyses in nonparametric econometrics”. Manuscript, University of Texas, Austin, Department of Economics.
Stinchcombe, M., White, H. (1998). “Consistent specification testing with nuisance parameters present only under the alternative”. Econometric Theory 14, 295–325.
Stone, C.J. (1982). “Optimal global rates of convergence for nonparametric regression”. The Annals of Statistics 10, 1040–1053.
Stone, C.J. (1985). “Additive regression and other nonparametric models”. The Annals of Statistics 13, 689–705.
Stone, C.J. (1986). “The dimensionality reduction principle for generalized additive models”. The Annals of Statistics 14, 590–606.
Stone, C.J. (1990). “Large-sample inference for log-spline models”. The Annals of Statistics 18, 717–741.
Stone, C.J. (1994). “The use of polynomial splines and their tensor products in multivariate function estimation (with discussion)”. The Annals of Statistics 22, 118–184.
Stone, C.J., Hansen, M., Kooperberg, C., Truong, Y.K. (1997). “Polynomial splines and their tensor products in extended linear modeling (with discussion)”. The Annals of Statistics 25, 1371–1470.
Strawderman, R.L., Tsiatis, A.A. (1996). “On the asymptotic properties of a flexible hazard estimator”. The Annals of Statistics 24, 41–63.
Timan, A.F. (1963). Theory of Approximation of Functions of a Real Variable. MacMillan, New York.
Van de Geer, S. (1993). “Hellinger-consistency of certain nonparametric maximum likelihood estimators”. The Annals of Statistics 21, 14–44.
Van de Geer, S. (1995). “The method of sieves and minimum contrast estimators”. Mathematical Methods of Statistics 4, 20–38.
Van de Geer, S. (2000). Empirical Processes in M-estimation. Cambridge University Press.
Van der Vaart, A. (1991). “On differentiable functionals”. The Annals of Statistics 19, 178–204.
Van der Vaart, A., Wellner, J. (1996). Weak Convergence and Empirical Processes: With Applications to Statistics. Springer-Verlag, New York.
Vapnik, V. (1998). Statistical Learning Theory. Wiley–Interscience, New York.
Wahba, G. (1990). Spline Models for Observational Data. CBMS–NSF Regional Conference Series. Philadelphia.
White, H. (1984). Asymptotic Theory for Econometricians. Academic Press.
White, H. (1990). “Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings”. Neural Networks 3, 535–550.
White, H. (1994). Estimation, Inference and Specification Analysis. Cambridge University Press.
White, H., Wooldridge, J. (1991). “Some results on sieve estimation with dependent observations”. In: Barnett, W.A., Powell, J., Tauchen, G. (Eds.), Non-parametric and Semi-parametric Methods in Econometrics and Statistics. Cambridge University Press, Cambridge, pp. 459–493.
Wong, W.H. (1992). “On asymptotic efficiency in estimation theory”. Statistica Sinica 2, 47–68.
Wong, W.H., Severini, T. (1991). “On maximum likelihood estimation in infinite dimensional parameter spaces”. The Annals of Statistics 19, 603–632.
Wong, W.H., Shen, X. (1995). “Probability inequalities for likelihood ratios and convergence rates for sieve MLE’s”. The Annals of Statistics 23, 339–362.
Wooldridge, J. (1992). “A test for functional form against nonparametric alternatives”. Econometric Theory 8, 452–475.
Wooldridge, J. (1994). “Estimation and inference for dependent processes”. In: Engle III, R.F., McFadden, D.F. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.
Xiao, Z., Linton, O. (2001). “Second order approximation for an adaptive estimator in a linear regression”. Econometric Theory 17, 984–1024.
Zhang, J., Gijbels, I. (2003). “Sieve empirical likelihood and extensions of the generalized least squares”. Scandinavian Journal of Statistics 30, 1–24.
Zhou, S., Shen, X., Wolfe, D.A. (1998). “Local asymptotics for regression splines and confidence regions”. The Annals of Statistics 26, 1760–1782.