PENALIZED SIEVE ESTIMATION AND INFERENCE OF SEMI-NONPARAMETRIC DYNAMIC MODELS: A SELECTIVE REVIEW

By

Xiaohong Chen

May 2011

COWLES FOUNDATION DISCUSSION PAPER NO. 1804

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

http://cowles.econ.yale.edu/

Penalized Sieve Estimation and Inference of Semi-nonparametric Dynamic Models: A Selective Review*

Xiaohong Chen†
Yale University

April 2011

Abstract

In this selective review, we first provide some empirical examples that motivate the usefulness of semi-nonparametric techniques in modelling economic and financial time series. We describe popular classes of semi-nonparametric dynamic models and some temporal dependence properties. We then present penalized sieve extremum (PSE) estimation as a general method for semi-nonparametric models with cross-sectional, panel, time series, or spatial data. The method is especially powerful in estimating difficult ill-posed inverse problems such as semi-nonparametric mixtures or conditional moment restrictions. We review recent advances on inference and large sample properties of the PSE estimators, which include (1) consistency and convergence rates of the PSE estimator of the nonparametric part; (2) limiting distributions of plug-in PSE estimators of functionals that are either smooth (i.e., root-n estimable) or non-smooth (i.e., slower than root-n estimable); (3) simple criterion-based inference for plug-in PSE estimation of smooth or non-smooth functionals; and (4) root-n asymptotic normality of semiparametric two-step estimators and their consistent variance estimators. Examples from dynamic asset pricing, nonlinear spatial VAR, semiparametric GARCH, and copula-based multivariate financial models are used to illustrate the general results.

Keywords: Nonlinear time series, Temporal dependence, Tail dependence, Penalized sieve M estimation, Penalized sieve minimum distance, Semiparametric two-step, Nonlinear ill-posed inverse, Mixtures, Conditional moment restrictions, Nonparametric endogeneity, Dynamic asset pricing, Varying coefficient VAR, GARCH, Copulas, Value-at-risk.

JEL: C13, C14, C20.

*This paper was presented as an invited lecture at the World Congress of the Econometric Society, Shanghai, August 2010. It was subsequently presented as three invited graduate lectures at CEMFI, Madrid, March 2011. I thank Manuel Arellano, David Childers, Tim Christensen, Michael Jansson, Oliver Linton, Demian Pouzo and Enrique Sentana for helpful discussions and Kieran Walsh for excellent research assistance. I am grateful to Manuel Arellano, Lars Hansen and Peter Robinson for encouragement. I acknowledge financial support from the National Science Foundation via grant SES-0838161 and the Cowles Foundation.

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

Contents

1 Introduction **1**

2 Vast Classes of Semi-nonparametric Dynamic Models **3**

2.1 Motivating empirical applications . . . 3

2.2 Partial list of semi-nonparametric time series models . . . 9

2.3 Digression: nonlinearity and temporal dependence . . . 13

3 Penalized Sieve Extremum (PSE) Estimation **18**

3.1 Ill-posed versus well-posed problems and PSE estimation . . . 19

3.2 Penalized sieve M estimation . . . 23

3.3 Penalized sieve MD estimation . . . 25

4 Large Sample Properties of PSE Estimators **26**

4.1 Consistency, convergence rates of PSE estimators . . . 27

4.2 Limiting distributions and inference for PSE estimation of functionals . . . 29

4.2.1 Simultaneous penalized sieve M estimators . . . 29

4.2.2 Simultaneous penalized sieve MD estimators . . . 30

5 Semiparametric Two-step Estimation **33**

5.1 Consistent sieve estimators of Avar$(\widehat{\beta}_n)$ . . . 34

5.2 Semiparametric multi-step estimation . . . 36

6 Concluding Remarks **38**

1 Introduction

In this paper we review some recent developments in large sample theory for estimation of and inference on semi-nonparametric time series models via the method of penalized sieves. To avoid confusion, we use the same terminology as that in Chen (2007). An econometric (or statistical) model is a family of probability distributions indexed by unknown parameters. We call a model "*parametric*" if all of its unknown parameters belong to finite dimensional Euclidean spaces. We call a model "*nonparametric*" if all of its unknown parameters belong to infinite dimensional function spaces. A model is "*semiparametric*" if its parameters of interest belong to finite dimensional spaces but its nuisance parameters are in infinite dimensional spaces. Finally, a model is "*semi-nonparametric*" if it contains both finite dimensional and infinite dimensional unknown parameters of interest.

Semi-nonparametric models and methods have become popular in much theoretical and empirical work in economics. This is partly because it is often the case that economic theory suggests neither parametric functional relationships among economic variables nor particular parametric forms for error distributions. Another reason for the rising popularity of semi-nonparametric models is rapidly declining costs of collecting and analyzing large data sets. The semi-nonparametric approach is very flexible in economic structural modelling and policy and welfare analysis. Compared to parametric and semiparametric approaches, semi-nonparametrics are more robust to functional form misspecification and are better able to discover *nonlinear* economic relations. Compared to fully nonparametric methods, semi-nonparametrics suffer less from the "curse of dimensionality" and allow for more accurate estimation of structural parameters of interest.

Semi-nonparametric time series models and methods should be very useful for economic structural time series analysis. Many economic and financial time series (and panel time series) are nonlinear and non-Gaussian; see, e.g., Granger (2003). Examples include but are not restricted to (1) *nonlinear macro/financial models*: nonlinear VAR, nonlinear ARCH/GARCH, stochastic volatility (SV), diffusion, thresholding, Markov switching, copula-based Markov models, conditional value-at-risk, nonlinear duration models, nonlinear observed and/or latent factors, nonlinear spatial dependence; (2) *nonlinear dynamic asset pricing models*: endogenous default, option pricing, cash-in-advance, financial frictions; (3) *semi-nonparametric Markov decision/game models*: nonlinear pricing, dynamic contracting; (4) *semi-nonparametric dynamic program evaluations*; and (5) *DSGE models.*

As we shall illustrate in Section 2, it is very difficult to correctly specify nonlinear dynamic functional relations. Even if the nonlinear functional relation is correctly specified by chance, misspecifying distributions of nonseparable latent variables or laws of motion (LOM) generally leads to inconsistent estimates of structural parameters of interest. Among some econometricians, a common view is that for simple forecasting purposes or certain reduced form data analyses misspecifying conditional mean or other nonlinear functional relations among observed variables is not a serious problem. However, for policy and welfare analysis, it is important to uncover complicated nonlinear economic relations in dynamic structural models. Since most low frequency macro time series data sets are not large enough to allow for purely nonparametric analysis, various semi-nonparametric models and methods should be attractive to economists conducting time series structural analyses.

In this selective review, we first motivate the usefulness of semi-nonparametric techniques in modelling economic and financial time series via empirical examples. We describe popular classes of semi-nonparametric dynamic models and some temporal dependence properties. Once we move beyond the linear and Gaussian modelling framework, there are too many semi-nonparametric dynamic models to list them all. In addition to statistical specification tests, one's economic questions of interest, economic theories, empirical stylized facts, and data issues should guide one's semi-nonparametric model choice in empirical work. We then present *Penalized Sieve Extremum* (PSE) estimation as a very flexible, computable and general method for semi-nonparametric models with cross-sectional, panel, time series, or spatial data. The penalized sieve method is especially powerful in estimating difficult ill-posed inverse problems such as semi-nonparametric mixtures and semi-nonparametric conditional moment restrictions. Semi-nonparametric mixture models have been widely used to flexibly capture unobserved individual heterogeneity and/or latent state dynamic factors in labor economics, industrial organization, public economics, finance, international trade, development, etc. Semi-nonparametric conditional moment restriction models or semi-nonparametric nonlinear instrumental variables models have been widely used in asset pricing, dynamic games, and other economic models derived from agents' optimizing behaviors.

In Chen (2007), we described a very important class of PSE, *sieve extremum estimation*, as a general method for semi-nonparametric models, listed some applications of the sieve method, presented many classes of sieves (flexible combinations of simple basis functions that can approximate unknown functions well), and provided detailed large sample properties available as of 2006. Since then, the amount of empirical work applying the sieve method has been rapidly growing, and there have been theoretical advances in sieve estimation and inference.

This paper gives an update of the survey of Chen (2007). We review recent advances in inference and large sample properties of the PSE estimators, which include (1) consistency and convergence rates of the PSE estimator of the nonparametric part, allowing for difficult (nonlinear) ill-posed inverse problems; (2) limiting distributions of plug-in PSE estimators of functionals that are either smooth (i.e., root-n estimable) or non-smooth (i.e., slower than root-n estimable); and (3) simple criterion based inference and consistent variance estimators of plug-in PSE estimators of smooth or non-smooth functionals. In empirical work in economics and finance, semiparametric two-step (or multi-step) estimation procedures are commonly used. We shall describe very recent results on simple consistent estimators of the asymptotic variances of general semiparametric two-step estimators of smooth functionals when unknown functions are estimated by the penalized sieve method in the first step. Examples from semi-nonparametric consumption capital asset pricing models, varying coefficient spatial VAR, semi-nonparametric multivariate GARCH and copula-based financial models are presented.

There are already many books and review articles on semiparametric and nonparametric models and methods. For recent books see Bickel et al. (1993), Fan and Gijbels (1996), Pagan and Ullah (1999), Fan and Yao (2003), Yatchew (2003), Haerdle et al. (2004), Li and Racine (2007), Gao (2007), or Horowitz (2009), to name only a few. For recent surveys relevant to economics see all of the chapters in the book edited by Barnett, Powell and Tauchen (1991), several chapters in the Handbook of Econometrics Volume 4 edited by Engle and McFadden (1994), Handbook of Econometrics Volume 6 edited by Heckman and Leamer (2007), and some of the surveys in Advances in Econometrics, World Congress of the Econometric Society book volumes.[1] Our survey complements the existing books and review papers by focusing on the latest developments in the general method of PSE estimation and allowing for (nonlinear) ill-posed inverse problems that typically appear in semi-nonparametric structural models in econometrics.

Notation: We use the same notation as in Chen (2007). Let $|\beta|_e$ denote Euclidean norm for Euclidean parameters $\beta \in \Re^{d_\beta}$. 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. For random variables V_n and positive numbers b_n, $n \geq 1$, we define $V_n = O_P(b_n)$ as $\lim_{c\to\infty} \limsup_n P(|V_n| \geq cb_n) = 0$ and define $V_n = o_P(b_n)$ as $\lim_n P(|V_n| \geq cb_n) = 0$ for all $c > 0$. We suppose there is an underlying complete probability space, the data $\{Z_t = (Y_t', X_t')'\}_{t=1}^n$ is stationary ergodic, $Z_t \in \Re^{d_z}$, $1 \leq d_z < \infty$, and all probability calculations are done under the true probability measure P_o. Let $\mathcal{I}_t$ denote the information set up to time t and $E(\cdot| \, \mathcal{I}_t)$ denote the conditional expectation given $\mathcal{I}_t$.

2 Vast Classes of Semi-nonparametric Dynamic Models

2.1 Motivating empirical applications

In this subsection we illustrate the usefulness and flexibility of semi-nonparametric dynamic models and PSE methods by three empirical applications in macroeconomics and finance.

Example 2.1 *(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 utility function $E\{\sum_{t=0}^{\infty} \delta^t U(C_t) \mid \mathcal{I}_0\}$, where δ is the time discount factor and $U(C_t)$ is period t utility. Consumption-based asset pricing models state that for any traded asset indexed by j, with a gross

[1] These include Bierens (1987), Gallant (1987), Robinson (1994), Tauchen (1997), Florens (2003), Blundell and Powell (2003), and others.

return at time $t+1$ of $R_{j,t+1}$, the following Euler equation holds:

$$E\left[M_{t+1}R_{j,t+1} \mid \mathcal{I}_t\right] = 1, \qquad j = 1, ..., N, \tag{2.1}$$

where $M_{t+1} = \delta \frac{\partial U / \partial C_{t+1}}{\partial U / \partial C_t}$ is the intertemporal marginal rate of substitution (IMRS) in consumption and also a pricing kernel or stochastic discount factor (SDF). Different specifications of M_{t+1} imply different consumption asset pricing models. See Cochrane (2001), Singleton (2006), Hansen et al. (2007), or Hansen and Renault (2010) for many examples of M_{t+1}.

Hansen and Singleton (1982) assume 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 $M_{t+1} = \delta \left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}$ and the Euler equation becomes: $E\left(\delta_o \left(\frac{C_{t+1}}{C_t}\right)^{-\gamma_o} R_{j,t+1} - 1 \mid \mathcal{I}_t\right) = 0$, $j = 1, ..., N$. They estimate the unknown scalar parameters δ_o, γ_o using Hansen's (1982) generalized method of moment (GMM) based on the following *unconditional* moment restrictions

$$E\left(\left[\delta_o \left(\frac{C_{t+1}}{C_t}\right)^{-\gamma_o} R_{j,t+1} - 1\right] \mathbf{Z}_t\right) = 0, \quad j = 1, ..., N,$$

where the instruments $\mathbf{Z}_t$ consists of a constant, lagged consumption growth, lagged EWR (the equal weighted market return) and lagged VWR (the value weighted market return). However, this classical power utility based asset pricing model has been rejected empirically. Stock and Wright (2000) suggest it might be due to the weak instrumental variable problem.

Many finance and macro economists suspect there is misspecification due to the assumption of time separable utility in consumption. One popular theoretical fix is to let period t utility depend on habit level H_t, which is some function of current and lagged consumption; see, e.g., Constantinides (1990), Abel (1990), Campbell and Cochrane (1999). But, is habit linear or nonlinear? Is habit internal or external? Economic theories do not provide clear answers to these questions, but they are of importance for welfare and pricing implications.

In a recent paper, Chen and Ludvigson (2009) specify the SDF, M_{t+1}, to be semi-nonparametric in order to encompass different versions of the habit model. They combine the power utility specification with a nonparametric habit formation: $E\{\sum_{t=0}^{\infty} \delta^t [(C_t - H_t)^{1-\gamma} - 1]/[1-\gamma] \mid \mathcal{I}_0\}$, where $H_t = H(C_t, C_{t-1}, ..., C_{t-L})$ 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 $H_t = C_t h_o(c_t^*)$ with $h_o(\cdot)$ unknown, $0 \leq h_o(\cdot) < 1$, $h_o(\cdot)$ nondecreasing in first argument of $c_t^* = \left(\frac{C_{t-1}}{C_t}, ..., \frac{C_{t-L}}{C_t}\right)$. Then $M_{t+1} = \delta \frac{\partial U / \partial C_{t+1}}{\partial U / \partial C_t}$, where for external habit $\partial U / \partial C_t = C_t^{-\gamma}(1 - h(c_t^*))^{-\gamma}$, and for internal habit $\partial U / \partial C_t =$

$$C_t^{-\gamma}\left[(1 - h(c_t^*))^{-\gamma} - E_t\{\sum_{j=0}^{L} \delta^j \left(\frac{C_{t+j}}{C_t}\right)^{-\gamma} (1 - h(c_{t+j}^*))^{-\gamma} \frac{\partial H_{t+j}}{\partial C_t}\}\right].$$

Chen and Ludvigson (2009) apply a sieve Minimum Distance (MD) procedure with *conditional* moment restrictions:

$$E\left[M_{t+1}(\delta_o, \gamma_o, h_o(\cdot))R_{j,t+1} - 1|\mathbf{w}_t\right] = 0,\ j = 1, ..., N, \quad \mathbf{w}_t \subset \mathcal{I}_t,$$

where the unknown $h_o\left(\right)$ is approximated by a sigmoid Artificial Neural Networks (ANN) sieve,[2] and the law of motion of $\{\frac{C_t}{C_{t-1}}, R_{1,t}, ..., R_{N,t}, \mathbf{w}_t\}$ is not parametrically specified except that the data are assumed to be stationary weakly dependent. Using quarterly data from 1952:4-2001:4, some of the (statistically significant) empirical findings are: *estimated habit is nonlinear, internal habit fits the data significantly better than external habit,* estimated δ, γ are sensible, and the estimated habit generated SDF performs well in explaining cross-sectional stock returns. See Chen and Ludvigson (2009) for details.

One can easily generalize their habit formation model and modify their method to estimate many other semi-nonparametric specifications of the SDF M_{t+1} satisfying the Euler equation (2.1). See Chen, Favilukis and Ludvigson (2009) for a semiparametric estimation of a recursive preference asset pricing model.

Example 2.2 *(Semi-nonparametric spatial VAR)*: Chen and Conley (2001) present an econometric model for high-dimensional vector time series with a panel structure where there is dependence across variables as well as over time. Examples of this type of data include quarterly observations on sector-specific variables and weekly price data for many retail firms in a region. In situations like these, there are too few degrees of freedom to permit unrestricted time series estimation; restrictions are needed to make progress. In particular, Chen and Conley (2001) wish to study how an industry's sector-specific shock affects its own next period output growth and those of other industries. The data set consists of $N = 20$ industry sectors, 72:2-92:4 quarterly data of output growth $\mathbf{Y}_t = \left(Y_{1,t}, Y_{2,t}, ..., Y_{N,t}\right)'$ and inputs variables $\{s_{i,t}\}_{i=1}^{N}$. Let $D_t = (D_t(1,2), ..., D_t(1,N), D_t(2,3), ..., D_t(2,N), ..., D_t(N-1,N))'$ where $D_t(i,j) = \left|s_{i,t} - s_{j,t}\right|_e$ is the "economic distance." They propose a semi-nonparametric spatial VAR model:

$$\mathbf{Y}_{t+1} = A(D_t)\mathbf{Y}_t + Q(D_t)\epsilon_{t+1}, \quad t = 1, ..., n,$$

where $E[\epsilon_{t+1}|\mathcal{I}_t] = 0$, $E[\epsilon_t\epsilon_t'|\mathcal{I}_t] = I_N$, $\mathcal{I}_t = \sigma(\{(\mathbf{Y}_{t-l}, D_{t-l}), l \geq 0\})$, the conditional mean is $E[\mathbf{Y}_{t+1}|\mathcal{I}_t] = A(D_t)\mathbf{Y}_t$ with

$$A(D_t) = \begin{bmatrix} \alpha_1 & g_1(D_t(1,2)) & ... & g_1(D_t(1,N)) \\ g_2(D_t(2,1)) & \alpha_2 & ... & g_2(D_t(2,N)) \\ ... & ... & ... & ... \\ g_N(D_t(N,1)) & g_N(D_t(N,2)) & ... & \alpha_N \end{bmatrix},$$

[2] ANN sieves can approximate unknown nonlinear functions of high dimensional variables well; see, e.g., Chen and White (1999) and Chen (2007).

and the conditional covariance $\Sigma(D_t) = Q(D_t)Q(D_t)'$ is:

$$\Sigma(D_t) = \begin{bmatrix} \sigma_1^2 + C(0) & C(D_t(1,2)) & ... & C(D_t(1,N)) \\ C(D_t(2,1)) & \sigma_2^2 + C(0) & ... & C(D_t(2,N)) \\ ... & ... & ... & ... \\ C(D_t(N,1)) & C(D_t(N,2)) & ... & \sigma_N^2 + C(0) \end{bmatrix},$$

$C(\|\tau\|) = \int_0^\infty \exp\left(-\left[y\,|\tau|_e\right]^2\right) d\Phi(y)$, Φ an unknown bounded nondecreasing function, and $|\tau|_e^2 = \sum_{i=1}^N \tau_i^2$ for $\tau \in \Re^N$. This specification of the conditional covariance $\Sigma(D_t)$ is called "conditional isotropic". It ensures that $\Sigma(D_t)$ is always positive definite and is intuitive for modelling how a sector specific shock affects other sectors. The classic VAR models assume that $\Sigma(D_t) \equiv Q(D_t)Q(D_t)'$ is diagonal i.e., $C() = 0$, which is unable to capture how a sector specific shock affects other sectors.

Chen and Conley (2001) estimate α_j, $g_j()$, σ_j^2 and $C()$ via a simple two-step sieve Least Squares (LS) procedure, where the unknown functions $g_j()$ and $C()$ are approximated by shape-preserving cardinal B-spline wavelet sieves. One of their empirical finding is that the estimated $C()$ is *a strictly decreasing function of the economic distance and statistically significantly bounded away from zero.* See their paper for details.

One can generalize this model in many ways. For instance, the economic distance variable D_t could be endogenous in some applications. For endogenous D_t the sieve LS estimators will no longer be consistent, and one can apply sieve MD or sieve GMM procedures instead. The recent theoretical advances on sieve MD by Ai and Chen (2003) and Chen and Pouzo (2008, 2009a, 2010) can be adapted to models of spatial time series with endogeneity.

Example 2.3 *(Semi-nonparametric GARCH + residual copula models)*: Many explanations of the recent financial crisis have emphasized the role of financial frictions and collateral; see, e.g., Geanakpolos (2010) for a review. The story is that financial frictions or leverage effects amplify the impact that unexpected bad news or bad shocks (in, for example, the mortgage market) have on prices and real activity. Central to the "Leverage Cycle" theory of Geanakoplos (2010) is his assumption that bad news (or an unexpected negative return shock) increases uncertainty (volatility). Fostel and Geanakoplos (2010) provide a theoretical explanation for why bad news tends to increase volatility and good news decreases volatility. We would like to use flexible econometric models and methods to empirically recover the shapes of the "news impact curve" for individual financial series. In addition, we wish to empirically address "risk assessment" and tail dependence among shocks to different financial series, which are also important in understanding the financial crisis; see, e.g., Engle (2010).

Let $\varepsilon_{i,t}$ and $\sigma_{i,t}^2$ respectively denote the time t shock (innovation) and volatility associated with return series i. Note that the standard GARCH(1,1) model, $\sigma_{i,t}^2 = \omega_i + \gamma_i \left(\sigma_{i,t-1}\varepsilon_{i,t-1}\right)^2 + \theta_i \sigma_{i,t-1}^2$, implies a symmetric impact of shocks on subsequent volatility. We model the "news impact curve" of the i-th series via a semi-nonparametric GARCH(1,1) model: $\sigma_{i,t}^2 = \omega_i + h_i\left(\sigma_{i,t-1}\varepsilon_{i,t-1}\right) + \theta_i \sigma_{i,t-1}^2$, where the part "$\omega_i + h_i(\cdot)$" is called the "news impact curve" for series i. It represents how unexpected return shocks affect subsequent volatility. The functional form $h_i(\cdot)$ is not specified and is estimated nonparametrically from data.[3]

To accurately assess risk, our financial models must account for (i) the possibility of fat tailed marginal distributions of innovations and (ii) the dependence between shocks to different assets. The class of semiparametric copula based multivariate dynamic (SCOMDY) models proposed in Chen and Fan (2006a) can easily capture both characteristics.

In this empirical illustration, we use daily data from March 20, 2007 to December 31, 2010, and consider three series: daily excess returns on the Barclays mortgage-backed security (MBS) index (S_t^e), daily excess stock market (the daily Fama-French factor) returns (M_t^e), and daily excess returns on the Barclays bond index (B_t^e). The data on M_t^e are from the "Fama/French Factors [Daily]" dataset on the website of Kenneth French. The data on S_t^e and B_t^e are log-differences of, respectively, the total return Barclays MBS index ("MBB") and the total return Barclays bond index ("AGG"). These indexes attempt to replicate the aggregate performance of their respective sectors, MBS and investment grade bonds, in the US; see http://us.ishares.com for further details.

We propose the following multivariate semi-nonparametric time series model:

$$\begin{aligned}
\textit{MBS Market} &: \quad S_t^e = c_S + \rho_S S_{t-1}^e + \beta_S M_{t-1}^e + \sigma_{S,t}\varepsilon_{S,t} \\
\textit{Stock Market} &: \quad M_t^e = c_M + \rho_M M_{t-1}^e + \sigma_{M,t}\varepsilon_{M,t} \\
\textit{Bonds Market} &: \quad B_t^e = c_B + \rho_M B_{t-1}^e + \beta_B M_{t-1}^e + \sigma_{B,t}\varepsilon_{B,t} \\
\textit{Volatility} &: \quad \sigma_{i,t}^2 = \omega_i + \theta_i \sigma_{i,t-1}^2 + h_i\left(\sigma_{i,t-1}\varepsilon_{i,t-1}\right), \ i \in \{S, M, B\},
\end{aligned}$$

where $E(\varepsilon_{i,t}) = 0$ and $E\left(\varepsilon_{i,t}^2\right) = 1$ for $i \in \{S, M, B\}$. $\boldsymbol{\varepsilon}_t = (\varepsilon_{S,t}, \varepsilon_{M,t}, \varepsilon_{B,t})'$ are independent, identically distributed across time. $\boldsymbol{\varepsilon}_t$ has a joint distribution $F(\boldsymbol{\varepsilon}) = C(F_S(\varepsilon_S), F_M(\varepsilon_M), F_B(\varepsilon_B); \alpha)$, where $C(\cdot;\alpha) : [0,1]^3 \to [0,1]$ is a copula function[4] with unknown parameters α. In the empirical application, the marginal distributions $F_i(\cdot)$, $i \in \{S, M, B\}$, are not specified, but the copula function is assumed to be one with tail dependence, in particular the Student's t-copula $C(\mathbf{u}; \alpha)$, $\alpha = (\Sigma, v)$. Its density is

$$c(\mathbf{u}; \Sigma, v) = \frac{\Gamma\left(\frac{v+3}{2}\right)\left(\Gamma\left(\frac{v}{2}\right)\right)^2}{\sqrt{\det(\Sigma)}\left(\Gamma\left(\frac{v+1}{2}\right)\right)^3}\left(1 + \frac{\mathbf{x}'\Sigma^{-1}\mathbf{x}}{v}\right)^{-\frac{v+3}{2}} \prod_{i\in\{S,M,B\}}\left(1 + \frac{x_i^2}{v}\right)^{\frac{v+1}{2}},$$

where Σ is the correlation matrix, $\mathbf{x}' = (x_S, x_M, x_B)$, $x_i = T_v^{-1}(u_i)$, T_v is the univariate Student's t distribution with degrees of freedom v. In the t-copula case, the bivariate tail dependence between

[3] Previously, Engle and Ng (1993) used piecewise linear splines to model the Japanese stock market "news impact curve". Linton and Mammen (2005) used kernel methods to estimate "news impact curves" and applied their method to the study of S&P 500 returns.

[4] A copula function is a multivariate distribution function with uniform marginal distributions.

shocks to series i and series j is

$$\lambda_{ij} = 2T_{v+1}\left(-\frac{\sqrt{v+1}\sqrt{1-corr\,(i,j)}}{\sqrt{1+corr\,(i,j)}}\right).$$

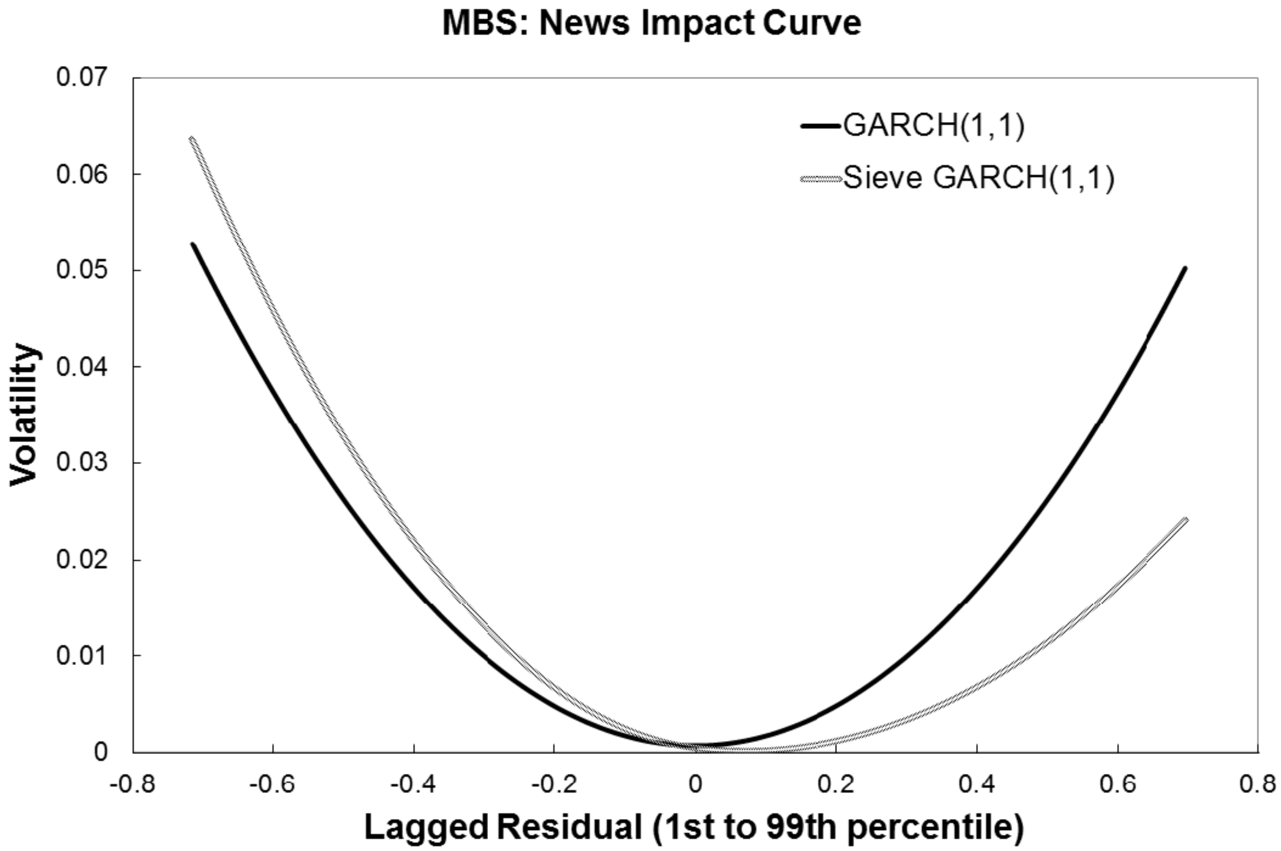
We estimate this multivariate time series model via a semi-nonparametric multi-step procedure. In the first step, we estimate each set of conditional mean and GARCH parameters via sieve quasi maximum likelihood (QMLE), where each unknown $h_i()$ is approximated via cubic B-spline sieves excluding a constant term. In the second step, we estimate each unknown marginal distribution $F_i(\cdot)$ using the empirical cdf associated with the fitted standardized residuals. In the third step, we estimate the unknown t-copula correlation matrix and degree of freedom via pseudo MLE. See Subsection 5.2 for details.

Our empirical findings are as follows. *All three estimated news impact curves exhibit the same asymmetry: bad news increases volatility more than does good news.* For mortgage-backed securities and stocks, some good news actually decreases volatility, as in Fostel and Geanakoplos (2010). As in Linton and Mammen (2005), most good news in the stock market does not have much effect on volatility. As we see in the MBS case (see figure and table below), for negative shocks Sieve-GARCH(1,1) predicts more volatility than does standard GARCH(1,1), and for positive shocks Sieve-GARCH predicts less volatility than does GARCH. For the concurrent dependence among the innovations that is described by the Student's t-copula, we find that (i) *shocks to bonds and shocks to mortgage-backed securities (MBS) are highly positively correlated*, (ii) shocks to MBS and shocks to stocks are moderately negatively correlated, (iii) shocks to bonds and shocks to stocks are moderately negatively correlated, and (iv) *shocks to MBS and shocks to bonds exhibit substantially positive tail dependence.* See the below table, in which standard errors are in parentheses.[5] Note that with estimated semi-nonparametric GARCH and residual copula dependence parameters, we could easily calculate Value-at-Risk (VaR) for a portfolio comprised of mortgage-backed securities, stocks, and bonds.

Copula Parameter Estimates

$corr\,(S,M)$	$corr\,(S,B)$	$corr\,(M,B)$	v	λ_{SM}	λ_{SB}	λ_{MB}
−.2801	.9144	−.3590	5.3903	.0137	.6110	.0097
(.0320)	(.0064)	(.0307)	(.6484)	(.0057)	(.0239)	(.0042)

[5] Additional details on conditional mean/variance estimates, their standard errors, and figures and tables are available upon request.



MBS Parameter Estimates

Model	c_S	ρ_S	β_S	ω_S	θ_S	γ_S
GARCH(1,1)	.0194 (.0057)	.0754 (.0356)	.0132 (.0051)	.0007 (.0004)	.8922 (.0252)	.1022 (.0219)
Sieve-GARCH(1,1)	.0134 (.0060)	.0734 (.0376)	.0117 (.0049)	.2369 (.1724)	.9118 (.0597)	see figure

2.2 Partial list of semi-nonparametric time series models

If we allow for nonlinear and/or non-Gaussian economic and financial time series, there are too many parametric time series models to fully list. Any one of these models can be slightly modified into various semi-nonparametric models. See, e.g., Tong (1990), Tiao and Tsay (1994), Teräsvirta, Tjϕstheim and Granger (1994), Härdle, Lütkepohl and Chen (1997), Granger (2003), Fan and Yao (2003), Fan (2005), Tsay (2005), Gao (2007), Aït-Sahalia, Hansen and Scheinkman (2009), Franke, Kreiss and Mammen (2009), Patton (2009), Linton (2009), Linton and Yan (2011), Giraitis, Leipus and Surgailis (2009) and numerous recent reviews on univariate and multivariate nonlinear/semi-nonparametric time series models. In this subsection, we mention some popular classes of such models in macro and financial econometrics and suggest ways to generate new semi-nonparametric time series models.[6]

(I) Univariate semi-nonparametric dynamic models

(I.1) Autoregressive and/or conditional heteroskedastic regression models:

$$Y_{t+1} = E[Y_{t+1}|\mathcal{I}_t] + \sqrt{Var(Y_{t+1}|\mathcal{I}_t)}\epsilon_{t+1},$$

where $E[\epsilon_{t+1}|\mathcal{I}_t] = 0$, $Var(\epsilon_{t+1}|\mathcal{I}_t) = 1$. Different specifications of conditional mean, $E[Y_{t+1}|\mathcal{I}_t]$, and/or conditional variance, $\sigma_t^2 \equiv Var(Y_{t+1}|\mathcal{I}_t)$, lead to many nonlinear time series models, such as the ARCH/GARCH models of Engle (1982) and Bollerslev (1996), the threshold model of Tong and Lim (1980) and Hansen (1996), and the smooth transition model of Granger and Teräsvirta (1993), to name only a few. If economic theories do not suggest particular nonlinear functional forms for $E[Y_{t+1}|\mathcal{I}_t]$ and/or σ_t^2, one may model these parts fully nonparametrically and estimate them from data. However, due to the "curse of dimensionality" and modest sample sizes, fully nonparametric estimation is often not practical. One could use various semi-nonparametric models, which reduce dimensionality, instead. For example, let $\{X_t, Z_t\} \subseteq \mathcal{I}_t$ where X_t and Z_t could include different Y_{t-j} for $j \geq 0$. Then $E[Y_{t+1}|\mathcal{I}_t]$ and/or σ_t^2 could be modelled in any of the following ways:

- partially linear: $X_t'\beta + h(Z_t)$; see, e.g., Engle, Granger, Rice and Weiss (1986), Robinson (1988) Haerdle, Liang and Gao (2000), Chen, Racine and Swanson (2001).

- functional coefficient: $\sum_{j=1}^q h_j(Z_t)X_{j,t}'$; see, e.g., Chen and Tsay (1993a), Cai, Fan and Yao (2000), Chen and Conley (2001), Huang and Shen (2004).

- single index: $h(X_t'\beta + Z_t'\gamma)$; see, e.g., Ichimura (1993), Wang and Yang (2009a).

- additive: $h_1(X_t) + h_2(Z_t)$; see, e.g., Stone (1985), Andrews and Whang (1990), Chen and Tsay (1993b), Mammen, Linton and Nielsen (1999), Huang and Yang (2004).

The semiparametric ARCH(∞) model, $\sigma_t^2 = \beta\sigma_{t-1}^2 + h(Y_t)$, of Engle and Ng (1993), Linton and Mammen (2005), and others is an example of a partially linear regression model for volatility. This simple model is widely used to allow for flexible "news impact curves" in finance; see, e.g., Example 2.3. The conditional mean specification in Chen and Conley (2001) could be viewed as a functional coefficient regression model; see, e.g., Example 2.2.

The methods used to model time series conditional mean and conditional variance could be easily extended to model dynamic duration (or survival) data. For instance, one can easily modify the results of Engle and Russell (1998), Zhang, Russell and Tsay (2001) and others on *Autoregressive Conditional Duration* (ACD) models to allow for more flexible semi-nonparametric specifications.

[6] Due to the lack of space and time, we describe in a relatively detailed way only a few models, ones that will be revisited in the rest of this paper.

(I.2) Transformation autoregressive regression models. As observed by Granger (2003), in order to perform economic policy evaluations and risk management, we need to model aspects of time series beyond conditional means and conditional variances. Engle and Mangenelli (2004) and Koenker and Xiao (2006) proposed *autoregressive conditional quantile regressions* to model conditional Value-at-Risk (VaR). Their models have been generalized to allow for various semi-nonparametric forms.

To allow for an internally coherent way to model conditional VaR as well as tail risk, Chen and Fan (2006b) proposed a class of *Copula-based autoregressive regressions.* They use the fact that any strictly stationary first order Markov time series $\{Y_t\}_{t=1}^n$ with continuous marginal distribution F can be equivalently characterized by a bivariate copula function, $C(u_0, u_1)$, and the marginal F. That is, the bivariate joint distribution of (Y_{t-1}, Y_t) is $F(Y_{t-1}, Y_t) = C(F(Y_{t-1}), F(Y_t))$, and the conditional density of Y_t given Y_{t-1} is $f(Y_t|Y_{t-1}) = c(F(Y_{t-1}), F(Y_t))f(Y_t)$. Leaving marginal cdf F unspecified, different parametric specifications of the copula density function, $c(u_0, u_1; \alpha)$, lead to different semiparametric transformation autoregressive regression models:

$$\Lambda_{1,\theta_1}(F(Y_t)) = \Lambda_{2,\theta_2}(F(Y_{t-1})) + \varepsilon_t, \quad E[\varepsilon_t|Y_{t-1}] = 0,$$

where $\Lambda_{1,\theta_1}(\cdot)$ is a parametric increasing function, $\Lambda_{2,\theta_2}(u) \equiv E\{\Lambda_{1,\theta_1}(F(Y_t))|F(Y_{t-1}) = u\}$, and the conditional density of ε_t given $F(Y_{t-1}) = u$ is

$$f_{\varepsilon_t|F(Y_{t-1})=u}(\varepsilon) = c(u, \Lambda_{1,\theta_1}^{-1}(\varepsilon + \Lambda_{2,\theta_2}(u)); \alpha) \div \frac{d\Lambda_{1,\theta_1}(\varepsilon + \Lambda_{2,\theta_2}(u))}{d\varepsilon}.$$

As demonstrated by Chen and Fan (2006b), Chen, Koenker and Xiao (2009), Chen, Wu and Yi (2009) and others, copula based first order Markov models are useful for modelling conditional VaR of Y_t given Y^{t-1}, which is simply the conditional quantile of Y_t given Y^{t-1}:

$$Q_q^Y(y) = F^{-1}\left(C_{2|1}^{-1}\left[q|F(y); \alpha\right]\right),$$

where $C_{2|1}[\cdot|u; \alpha] \equiv \frac{\partial}{\partial u}C(u, \cdot; \alpha)$ is the conditional distribution of $U_t \equiv F(Y_t)$ given $U_{t-1} = u$; and $C_{2|1}^{-1}\left[q|u; \alpha\right]$ is the $q-$th conditional quantile of U_t given $U_{t-1} = u$. This class of models is also useful in capturing tail dependence of the time series $\{Y_t\}$:

$$\begin{aligned}
\lim_{y \to -\infty} \Pr\left(Y_t \leq y|Y_{t-1} \leq y\right) &= \lim_{y \to -\infty} \Pr\left(F(Y_t) \leq F(y)|F(Y_{t-1}) \leq F(y)\right) = \lim_{u \to 0^+} \frac{C(u, u; \alpha)}{u}, \\
\lim_{y \to +\infty} \Pr\left(Y_t \geq y|Y_{t-1} \geq y\right) &= \lim_{u \to 1^-} \frac{1 - 2u + C(u, u; \alpha)}{1 - u},
\end{aligned}$$

provided the limits exist. See Patton (2006, 2009), Ibragimov (2009) and others for additional time series autoregressive models generated via copulas.

(I.3) Distribution-based models: There are many nonlinear time series models that directly specify flexible conditional distributions. See, e.g., Markov switching (Hamilton, 1989), hidden Markov, generalized hidden Markov, mixtures, random iterative models (Duflo, 1997), and nonlinear state space

models (Hamilton (1994), Hansen and Sargent (2007)). There are many potential ways to semi-nonparametrically relax aspects of these models.

(I.4) Discrete time data sampled from continuous-time models: Many theoretical models in macroeconomics, finance and survival analysis are presented as continuous-time stochastic processes such as stochastic volatility (Andersen, 1996), diffusions, jump-diffusions, Levy processes, continuous time Markov models (Hansen and Scheinkman, 1995), etc., while economic and financial time series data are sampled in low frequency from the underlying continous-time models. See Aït-Sahalia, Hansen and Scheinkman (2009) and others for reviews of these models.

(II) Multivariate semi-nonparametric dynamic models

All of the existing univariate nonlinear time series models can easily be generalized to multiple time series models, such as Sims' structural vector autoregression (VAR) model, Engle's vector ARCH/GARCH model and others.

In addition and, perhaps more interestingly, we may add complexity and/or flexibility in modelling multivariate economic time series by specifying comovements in various ways. Currently, there are two main approaches for modelling comovements, *factors* (e.g., Stock and Watson, 2002) and *copulas* (e.g., Embrechts, 2008). Either approach could be used to model

- concurrent comovements among multiple observed time series;
- concurrent comovements among multiple innovations;
- auto-comovements among multiple observed time series;
- auto-comovements among multiple innovations.

For example, Chen and Fan (2006a) proposed a large class of semiparametric copula based multivariate dynamic (SCOMDY) models:

$$Y_{j,t+1} = E[Y_{j,t+1}|\mathcal{I}_t] + \sqrt{Var(Y_{j,t+1}|\mathcal{I}_t)}\epsilon_{j,t+1},\ j = 1, ..., N,$$

where the innovation $\{\boldsymbol{\epsilon}_{t+1} \equiv (\epsilon_{1t+1}, \ldots, \epsilon_{Nt+1})' : t \geq 0\}$ is assumed to be i.i.d. and independent of $\mathcal{I}_t = \sigma(\{\mathbf{Y}^t, \mathbf{X}^t\})$. $E(\epsilon_{jt}) = 0$, $E(\epsilon_{jt}^2) = 1$, and each ϵ_{jt} has unknown marginal cdf $F_j(\cdot)$. $\boldsymbol{\epsilon}_t$ has joint distribution $F(\boldsymbol{\epsilon}) = C(F_1(\epsilon_1), \ldots, F_N(\epsilon_N); \alpha)$, where $C(\cdot; \alpha) : [0,1]^N \to [0,1]$ is a copula function with copula dependence parameter α.

Different specifications of $E[Y_{j,t+1}|\mathcal{I}_t]$, $Var(Y_{j,t+1}|\mathcal{I}_t)$ and $C(\cdot; \alpha)$ lead to many different examples of SCOMDY models; see, e.g., Example 2.3. These models are easy to estimate and useful for flexibly estimating conditional VaR and contagion. Recently Cherubini et al (2010) apply SCOMDY to build models of term structure of multivariate equity derivatives.

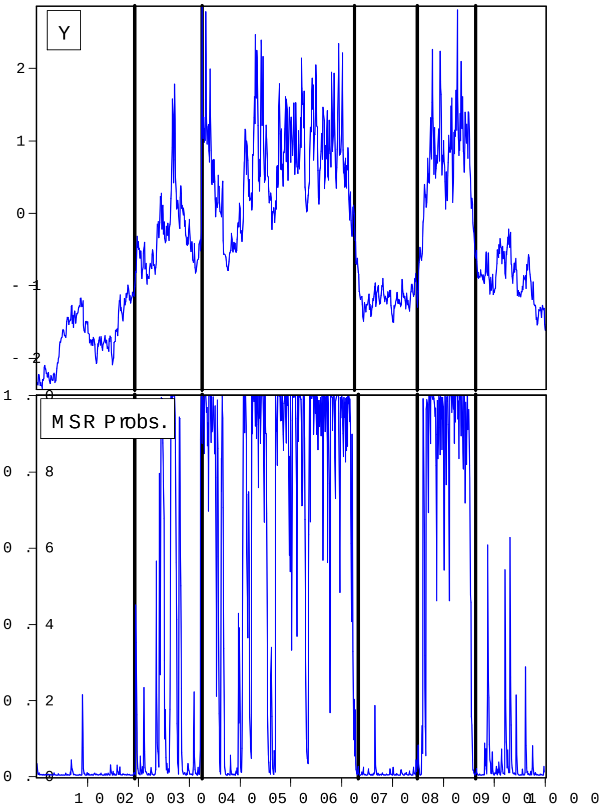
(III) Specification of semi-nonparametric dynamic models

As the reader can tell from the above descriptions, there are already many semi-nonparametric dynamic models and one can easily introduce new ones. In empirical applications in economics and finance, several different nonlinear semi-nonparametric models could generate similar empirical patterns. So which model(s) should one use? As illustrated by the three empirical examples in Subsection 2.1, the answer should depend on which question(s) the researcher wishes to address. Guidance from economic theories and empirical stylized facts could be as important as data structures and formal statistical specification tests.

For example, any strictly stationary first order Markov time series $\{Y_t\}_{t=1}^n$ with a continuous marginal distribution F can be equivalently generated using a copula function $C(u_0, u_1)$ and the marginal F as follows: (i) generate n independent random variables $\{X_t\}_{t=1}^n$ from the standard uniform distribution $U(0,1)$; (ii) let $U_1 = X_1$, $U_t = C_{2|1}^{-1}[X_t|U_{t-1}]$, and $Y_t \triangleq F^{-1}(U_t)$. In the next graph, a strictly stationary first order Markov series $\{Y_t\}_{t=1}^n$ is generated using a bivariate Clayton copula with a Student's t marginal: $C_{2|1}^{-1}[X_t|U_{t-1}] = [(X_t^{-15/16} - 1)U_{t-1}^{-15} + 1]^{-1/15}$ and $F =$ cdf of t(3). However, applying a recent structural break test of Davis et al. (2005), one will detect 5 breaks (vertical black lines). A Markov switching model also fits well. In fact, many Markov models with tail dependent copulas and fat tailed marginals will also have time series plots displaying patterns like structural breaks, Markov switching and long memory. If the researcher cares about conditional VaR or tail dependence, a copula-based Markov model is a sensible choice. See, e.g., Chen and Fan (2006b), Chen, Koenker and Xiao (2009), Bouyé and Salmon (2009), or Ibragimov and Lentzas (2009).

2.3 Digression: nonlinearity and temporal dependence

Concepts that capture temporal dependence of linear time series models (autocorrelation, long memory, fractional integration, unit roots, cointegration, etc.) are inadequate and sometimes misleading in describing temporal dependence of nonlinear time series models. For example, many researchers have asked the question, is the daily US interest rate series unit root or long memory? The answer is very likely to be yes if the interest rate is modelled as a linear process. However, the answer is very likely to be no if the interest rate is modelled as a nonlinear first order Markov process or as a discrete time realization of a continuous-time Markov diffusion process. Another example is in Chen, Hansen and Carrasco (2010). They show that a strictly stationary scalar diffusion process is always beta-mixing (see definition below); but some of the beta-mixing decay rate could be very slow, in which case some of its transformations behave like long memory (in the sense that the spectral density blows up at frequency zero in a manner like long memory in a linear time series). As a third example, any strictly stationary first order Markov time series $\{Y_t\}_{t=1}^n$ can be generated using a copula $C(u_1, u_2)$ that links Y_t and Y_{t+1} and a marginal cdf F. Ibragimov and Lentzas (2009) found in Monte Carlo studies that a Markov time


Y
2
1
0
- 1
- 2
M S R P robs.
1 . 0
0 . 8
0 . 6
0 . 4
0 . 2
0 . 0
1 0 0
2 0 0
3 0 0
4 0 0
5 0 0
6 0 0
7 0 0
8 0 0
9 0 0
1 0 0 0

series generated via a Clayton copula and a fat tailed marginal looks like long memory. Chen, Wu and Yi (2009) and Beare (2010) show that it is really beta-mixing with an exponential decay rate.

There are many different notions of temporal dependence of nonlinear time series. The ones that have been used in the econometrics literature include:

- ergodicity for Markov processes, see, e.g., Tong (1990), Meyn and Tweedie (1993);
- mixing, see, e.g., Rosenblatt (1956), Doukhan (1994), Bradley (2007);
- near epoch dependence of mixing, see, e.g., Billingsley (1968), Andrews (1984), Gallant and White (1988), Wooldridge and White (1988), Davidson (1994), Pötscher and Prucha (1997);
- physical and predictive dependence measures, see, e.g., Wu (2005, 2011);
- new weak dependence, see, e.g., Doukhan and Louhichi (1999);
- martingales, see, e.g., Hall and Heyde (1980);
- semimartingales, see, e.g., Ibragimov and Phillips (2008);
- long memory, see, e.g., Robinson (1994);
- nonlinear transformation of a unit root, or null recurrent Markov processes, see, e.g., Phillips and Park (1998), Park and Phillips (2001), Wang and Phillips (2009a), Karlsen and Tjϕstheim (2001).

In principle any of the above dependence concepts could be used for semi-nonparametric time series models. In fact, there are already published work on kernel density estimation and kernel conditional mean regression for time series data displaying any of the above dependence properties; see. e.g., Robinson (1994), Hidalgo (1997), Gao (2007) and others for long memory processes; Phillips and Park (1998), Wang and Phillips (2009a), Karlsen and Tjϕstheim (2001) and others for nonlinear and nonstationary processes. Currently all the existing papers on semi-nonparametric density and regression estimation of time series models with strong dependence rely heavily on the closed form expressions of their estimators as well as the specific model structures.

In this survey, we focus on estimation and inference of a large class of semi-nonparametric dynamic models via a general penalized sieve extremum estimation method. Although flexible, a penalized sieve extremum estimator typically does not have a close form solution for complicated semi-nonparametric models such as the empirical examples in Subsection 2.1. In the literature, the large sample properties of penalized sieve extremum estimators, especially the rates of convergence, have been established mainly using the tools from empirical process theory in probability and mathematical statistics; see, Pollard (1984), Van der Vaart and Wellner (1996), van de Geer (2000), Kosorok (2008). At this moment, empirical process theory has been well developed mainly for strictly stationary ergodic processes that satisfy various mixing conditions; see, e.g., Yu (1994), Andrews (1994a), Doukhan, Massart and Rio (1995), Chen and Shen (1998), Rio (2000). Luckily, most widely used nonlinear time series models in econometrics and finance can be shown to be beta-mixing and/or strong-mixing.

Let $\mathcal{I}_{-\infty}^{t}$ and $\mathcal{I}_{t+j}^{\infty}$ be $\sigma-$fields generated respectively by $(Y_{-\infty}, \cdots, Y_t)$ and $(Y_{t+j}, \cdots, Y_{\infty})$. Define

$$\beta(j) \equiv \sup_t E \sup\{|P(B|\mathcal{I}_{-\infty}^{t}) - P(B)| : B \in \mathcal{I}_{t+j}^{\infty}\}.$$

$$\alpha(j) \equiv \sup_t \sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{I}_{-\infty}^{t}, B \in \mathcal{I}_{t+j}^{\infty}\}.$$

$\{Y_t\}_{t=-\infty}^{\infty}$ is called *beta mixing* if $\beta(j) \to 0$ as $j \to \infty$ and is *strong mixing* if $\alpha(j) \to 0$ as $j \to \infty$.

There are alternative yet equivalent definitions of various mixing conditions for Markov processes. For a strictly stationary Markov process $\{Y_t\}_{t=0}^{\infty}$ on a set $\Omega \subseteq \mathcal{R}^d$, let $||\phi||_p^p = \int_\Omega |\phi(y)|^p dQ(y)$ and $\mathcal{T}_t\phi(y) = E[\phi(Y_t)|Y_0 = y]$. The Markov process $\{Y_t\}$ is said to be $\rho - mixing$ if

$$\rho(t) = \sup_{\phi: E[\phi(Y_t)]=0, ||\phi||_2=1} ||\mathcal{T}_t\phi||_2 \to 0 \text{ as } t \to \infty;$$

the Markov process $\{Y_t\}$ is $\alpha - mixing$ if

$$\alpha(t) = \sup_{\phi: E[\phi(Y_t)]=0, ||\phi||_\infty=1} ||\mathcal{T}_t\phi||_1 \to 0 \text{ as } t \to \infty;$$

and the Markov process $\{Y_t\}$ is $\beta - mixing$ if

$$\beta(t) = \int \sup_{0 \le \phi \le 1} \left| \mathcal{T}_t\phi(x) - \int \phi dQ \right| dQ \to 0 \text{ as } t \to \infty.$$

It is well-known that $2\alpha(t) \le \beta(t)$ and $\alpha(t) \le \rho(t)$, but $\beta(t)$ and $\rho(t)$ are not related in general. For Markov models, either $\rho(t) \equiv 1$ (strong dependence) or $\rho(t)$ decays exponentially fast, but $\beta(t)$ and $\alpha(t)$ could go to zero arbitrarily slowly. See, e.g., Bradley (2007).

The notion of $\beta - mixing$ for a Markov process is closely related to the concept called $V - ergodicity$ (in particular $1 - ergodicity$), see e.g., Meyn and Tweedie (1993). Given a Borel measurable function $V \ge 1$, the Markov process $\{Y_t\}$ is $V - ergodic$ if

$$\lim_{t \to \infty} \sup_{0 \le \phi \le V} \left| \mathcal{T}_t\phi(y) - \int \phi dQ \right| = 0 \text{ , for all } y;$$

the Markov process $\{Y_t\}$ is $V - uniformly\ ergodic$ if for all $t \ge 0$,

$$\sup_{0 \le \phi \le V} \left| \mathcal{T}_t\phi(y) - \int \phi dQ \right| \le cV(y) \exp(-\delta t)$$

for positive constants c and δ. A stationary process that is *V-uniformly ergodic* will be $\beta - mixing$ with exponential decay rate provided that $E[V(Y_t)] < \infty$. This connection is valuable because one can show that a Markov time series is beta mixing by applying the famous drift criterion (for ergodicity): There are constants $\lambda \in (0,1)$ and $d \in (0,\infty)$, a norm-like function $\Gamma() \geq 1$ and a small set $\mathbf{K}$ such that

$$E[\Gamma(Y_t)|Y_{t-1}] \leq \lambda\Gamma(Y_{t-1}) + d \times 1\{Y_{t-1} \in \mathbf{K}\}.$$

In this case, $\{Y_t\}$ is geometric ergodic and beta mixing with exponential decay rate. There is also a drift criterion for sub-geometric ergodicity or beta mixing decay at a slower than exponential rate. See, e.g., Tong (1990) and Meyn and Tweedie (1993).

Many nonlinear time series econometrics models are shown to be beta mixing or strong mixing via Tweedie's drift criterion approach. See, e.g., Tong (1990) for threshold models, Chen and Tsay (1993a, b) for functional coefficient autoregressive models and nonlinear additive ARX models, Doukhan (1994) for nonlinear ARX(p,q), Masry and Tjϕstheim (1995) for nonlinear ARCH, Yao and Attali (2000) for nonlinear AR with Markov switching, Carrasco and Chen (2002) for GARCH, stochastic volatility (SV) and autoregressive conditional duration (ACD), Chen, Hansen and Carrasco (2010) for diffusions, Chen, Wu and Yi (2009) and Beare (2010) for copula-based Markov models, and many more. In addition, a large class of generalized hidden Markov models, including, for example, nonlinear state space models, can also be shown to satisfy beta-mixing via the drift criterion. See, e.g., Carrasco and Chen (2002), Douc, Moulines, Olsson and van Handel (2011).

Most of the popular nonlinear semi-nonparametric time series models assume that innovations have positive density against Lebesgue measure, which turns out to be a crucial assumption in establishing their beta-mixing (and hence strong mixing) properties. Andrews (1984) presents a famous counter example: $Y_t = \rho Y_{t-1} + \varepsilon_t$ where $\rho \in (0, 1/2]$ and the innovation ε_t is i.i.d. Bernoulli(q), $q \in (0, 1)$. Andrews (1984) shows that this simple AR(1) process $\{Y_t\}$ with discrete innovations fails to be strong mixing but is Near Epoch Dependent (NED), which is a more general dependence concept that still satisfies central limit theorems; see, e.g., Billingsley (1968). Andrews (1984) motivates the popularity of the NED of mixing processes in econometrics. See, e.g., Wooldridge and White (1988), Wooldridge (1994) and Davidson (1994). For a stochastic sequence $\{V_t\}_{-\infty}^{+\infty}$ that is weakly dependent mixing, let $\mathcal{F}_{t-m}^{t+m} = \sigma(V_{t-m}, ..., V_{t+m})$ be such that $\{\mathcal{F}_{t-m}^{t+m}\}_{m=0}^{\infty}$ is an increasing sequence of σ-fields. If, for $p > 0$, a sequence of integrable r.v.s $\{Y_t\}_{-\infty}^{+\infty}$ satisfies

$$\left\|Y_t - E\left[Y_t|\mathcal{F}_{t-m}^{t+m}\right]\right\|_p \leq d_t v_m,$$

where $v_m \to 0$ and $\{d_t\}_{-\infty}^{+\infty}$ is a sequence of positive constants, then $\{Y_t\}_{-\infty}^{+\infty}$ is said to be near-epoch dependent in L$_p$-norm (L$_p$-NED) on $\{V_t\}_{-\infty}^{+\infty}$.

The NED dependence concept is widely used in nonlinear *parametric* time series models. However, the currently available exponential inequality associated with NED is not sufficient for establishing sharp empirical process results and hence fails to achieve the optimal rates of convergence for general penalized sieve extremum estimators for nonlinear *semi-nonparametric* models. Andrews (1991b), Chen (1995), Chen and White (1998, 2002), Lu and Linton (2007), Li, Lu and Linton (2010) have obtained some limiting distribution results for semi-nonparametrc time series models that are NED of mixing processes. These papers have established their results relying on closed form expressions or some specific properties of their estimators that general penalized sieve extremum estimators do not have.

Another useful dependence measure for strictly stationary nonlinear time series is the so-called *physical and predictive dependence* measure; see, e.g., Wu (2005, 2011). Suppose that $\{Y_t\}_{t=-\infty}^{\infty}$ is strictly stationary and can be represented as

$$Y_t = H(\dots, \varepsilon_{t-1}, \varepsilon_t) = H(\mathcal{F}_{-\infty}^{t}), \tag{2.2}$$

where ε_t, $t \in \mathbb{Z}$, are independent and identically distributed (iid) random variables, $\mathcal{F}_{-\infty}^{t} = \sigma(\dots, \varepsilon_{t-1}, \varepsilon_t)$, and H is a measurable function such that Y_t is well-defined. In (2.2), (Y_t) is causal in the sense that Y_t does not depend on the future innovations $\varepsilon_j, j > t$. Let $(\varepsilon_i^*)_{i\in\mathbb{Z}}$ be an iid copy of $(\varepsilon_i)_{i\in\mathbb{Z}}$. Hence $\varepsilon_i^*, \varepsilon_j, i, j \in \mathbb{Z}$, are iid. Let

$$Y_t^* = H(\mathcal{F}_{-\infty}^{t*}), \quad \mathcal{F}_{-\infty}^{t*} = \sigma(\dots, \varepsilon_{-1}, \varepsilon_0^*, \varepsilon_1, \dots, \varepsilon_{t-1}, \varepsilon_t).$$

Assume $\|Y_t\|_p := (\mathbb{E}|Y|^p)^{1/p} < \infty$ for $p > 0$. For $t \geq 0$ define the physical dependence measure

$$\delta_p(t) = \|Y_t - Y_t^*\|_p,$$

and the predictive dependence measure ($p \geq 1$)

$$\theta_p(t) = \|\mathbb{E}(Y_t|\mathcal{F}_{-\infty}^{0}) - \mathbb{E}(Y_t|\mathcal{F}_{-\infty}^{-1})\|_p, \text{ or } \omega_p(t) = \|\mathbb{E}(Y_t|\mathcal{F}_{-\infty}^{0}) - \mathbb{E}(Y_t|\mathcal{F}_{-\infty}^{0*})\|_p.$$

The process (Y_t) defined in (2.2) is p-stable if $\sum_{j=0}^{\infty} \delta_p(j) < \infty$, and is weakly p-stable if $\sum_{t=0}^{\infty} \theta_p(t) < \infty$ (or equivalently if $\sum_{t=0}^{\infty} \omega_p(t) < \infty$). It is a special case of NED processes and allows for the famous example of Andrews (1984). Wu and his co-authors have shown that many nonlinear time series models satisfy these dependence measures and are developing limiting theorems and empirical process results for strictly stationary time series models that can be represented as (2.2).

3 Penalized Sieve Extremum (PSE) Estimation

A semi-nonparametric structural model specifies a family of probability distributions of $\{Z_t\}_{t=1}^{n}$ up to some finite dimensional Euclidean parameter β and some unknown functions h. Let $\theta = (\beta, h) \in \Theta = B \times \mathcal{H}$ be an infinite dimensional parameter space endowed with a (pseudo-) metric d. There is a population criterion function $Q : \Theta \to \Re$, which is maximized at a (pseudo-) true parameter

$\theta_o = (\beta_o, h_o) \in \Theta$. The choice of $Q(\cdot)$ and the existence of θ_o are suggested by the identification of the semi-nonparametric model.

Let $\widehat{Q}_n : \Theta \rightarrow \Re$ be an empirical criterion, which is a jointly measurable function of θ and data $\{Z_t\}_{t=1}^n$ and converges to Q in some sense (to be more precise later) as $n \rightarrow \infty$. One general way to estimate θ_o is by maximizing $\widehat{Q}_n$ over Θ. In particular, an approximate *extremum* estimator $\hat{\theta}_n$ satisfies

$$\widehat{Q}_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} \widehat{Q}_n(\theta) - O_P(\eta_n), \quad \text{with } \eta_n \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Examples of the criterion function $\widehat{Q}_n()$ include ML, MD, GMM, GEL and many more. See Amemiya (1985), Newey and McFadden (1994), White (1994) and others.

It is well known that if the following two conditions are satisfied,

- (IU condition) $\theta_o \in \Theta$ is said to satisfy "*identifiable uniqueness*" if

$$Q(\theta_o) > \sup_{\theta \in \Theta : d(\theta_o, \theta) \geq \varepsilon} Q(\theta) \quad \text{for all } \varepsilon > 0.$$

- (ULLN condition) $\sup_{\theta \in \Theta} \left| \widehat{Q}_n(\theta) - Q(\theta) \right| = o_p(1)$.

then the approximate extremum estimator $\hat{\theta}_n$ is consistent for θ_o, i.e., $d(\hat{\theta}_n, \theta_o) = o_p(1)$.

3.1 Ill-posed versus well-posed problems and PSE estimation

When Θ is infinite dimensional and possibly not compact with respect to the (pseudo-) metric d, maximizing $\widehat{Q}_n$ over Θ may not be well-defined; and even if a maximizer $\arg\sup_{\theta \in \Theta} \widehat{Q}_n(\theta)$ exists, it is generally difficult to compute. Even if one is able to compute $\hat{\theta}_n = \arg\sup_{\theta \in \Theta} \widehat{Q}_n(\theta)$, it may be inconsistent for θ_o; and even if consistent, it may converge to θ_o at a very slow convergence rate. These difficulties arise because the problem of optimization over an infinite dimensional non-compact space may no longer be well-posed.

Following Chen (2007), we say the optimization problem is

- *well-posed* if for all sequences $\{\theta_k\}$ in Θ with $Q(\theta_o) - Q(\theta_k) \rightarrow 0$ then $d(\theta_o, \theta_k) \rightarrow 0$;
- *ill-posed* (or *not well-posed*) if there exists a sequence $\{\theta_k\}$ in Θ with $Q(\theta_o) - Q(\theta_k) \rightarrow 0$ but $d(\theta_o, \theta_k) \nrightarrow 0$.

Therefore, the semi-nonparametric problem becomes ill-posed whenever the "identifiable uniqueness" condition fails. It is clear that "identifiable uniqueness" fails if θ_o is not point identified (i.e., if $Q()$ is maximized at more than one point in Θ). Even if $Q()$ is uniquely maximized at $\theta_o \in \Theta$ and is upper semicontinuous in (Θ, d), the "identifiable uniqueness" condition may still fail if Θ is not compact in d, which is typically the case in semi-nonparametric mixture models and semi-nonparametric conditional moment restriction problems.

Example 3.1 *(Semi-nonparametric mixture models):* data $\{Z_t = (Y_t, X_t')'\}_{t=1}^n$ are assumed drawn from a semi-nonparametric mixture density $f(Y_t|X_t;\beta_o,h_o) = \int_0^1 f(Y_t|X_t;\beta_o,u)h_o(u)du$, where $\beta_o \in B$, a compact subset in $\Re^{d_\beta}$, and $h_o \in \mathcal{H}$, a space of Lipschitz continuous probability density functions over $[0,1]$. It is clear that $\theta_o = (\beta_o, h_o) \in \arg\sup_{\beta\in B,h\in\mathcal{H}} Q(\beta,h)$, where

$$Q(\beta,h) = E\left(\log\left\{\int_0^1 f(Y_t|X_t;\beta,u)h(u)du\right\}\right).$$

Without any restriction on the parametric functional form of $f(Y_t|X_t;\beta,u)$, $\theta_o \in \Theta = B\times\mathcal{H}$ is not point identified. Even if we impose restrictions on $f(Y_t|X_t;\beta,u)$ so that $Q(\theta)$ is uniquely maximized at θ_o, the "identifiable uniqueness" condition still fails and the problem is ill-posed for (Θ,d) when $d(\theta,\theta_o) = |\beta-\beta_o|_e + d_H(h,h_o)$ for $d_H(h,h_o) = \sup_{u\in[0,1]}|h(u)-h_o(u)|$ or $\int_0^1|h(u)-h_o(u)|\,du$.

Example 3.2 *(Single index instrumental variables regression):* data $\{Z_t = (Y_{1t}, Y_{2t}', Y_{3t}, X_t')'\}_{t=1}^n$ are assumed to satisfy $E[Y_{1t} - h_o(Y_{2t}'\beta_o + Y_{3t})|X_t] = 0$ almost surely, where $\beta_o \in B$, a compact subset in $\Re^{d_\beta}$, and $h_o \in \mathcal{H}$, a space of increasing functions with continuous derivatives over $\Re$. It is clear that $\theta_o = (\beta_o,h_o) \in \arg\sup_{\beta\in B,h\in\mathcal{H}} Q(\beta,h)$, where

$$Q(\beta,h) = -E\left(\left\{E[Y_{1t} - h(Y_{2t}'\beta + Y_{3t})|X_t]\right\}^2\right).$$

Recently Chen, Chernozhukov, Lee and Newey (2011) provided sufficient conditions for local identification of $\theta_o = (\beta_o,h_o)$. However, even if we assume that $Q(\theta)$ is uniquely maximized at $\theta_o \in \Theta = B\times\mathcal{H}$, the problem is still ill-posed for (Θ,d) when $d(\theta,\theta_o) = |\beta-\beta_o|_e + d_H(h,h_o)$ for $d_H(h,h_o) = \sup_u |h(u)-h_o(u)|$ or $\sqrt{E[|h(Y_{2t}'\beta_o + Y_{3t}) - h_o(Y_{2t}'\beta_o+Y_{3t})|^2]}$. This example is a special case of semi-nonparametric conditional moment restrictions (3.7) (see below for further discussion).

Whether or not 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 $\Theta_{k(n)}$, called *sieves* by Grenander (1981), which are less complex but dense in Θ. Popular sieves are typically compact, non-decreasing ($\Theta_k \subseteq \Theta_{k+1} \subseteq \cdots$) and are such that $\Theta \subseteq cl\left(\cup_k \Theta_k\right)$, that is, for any $\theta\in\Theta$ there exists an element $\pi_{k(n)}\theta$ in $\Theta_{k(n)}$ satisfying $d(\theta,\pi_{k(n)}\theta) \to 0$ as $n\to\infty$, where we may interpret $\pi_{k(n)}$ as a projection mapping from Θ to $\Theta_{k(n)}$.

Like the method of sieves, the method of penalization (or regularization) is a general approach for solving possibly ill-posed, infinite dimensional optimization problems. This method estimates θ_o by maximizing $\left\{\widehat{Q}_n(\theta) - \lambda_n Pen(\theta)\right\}$, a penalized criterion, over the entire infinite dimensional parameter space Θ, where $\lambda_n > 0$ is a penalization parameter such that $\lambda_n \to 0$ as $n\to\infty$ and the penalty $Pen() > 0$ is typically chosen such that $\{\theta\in\Theta : Pen(\theta) \le M\}$ is compact in d for all $M\in(0,\infty)$.

Let $\Theta = B \times \mathcal{H}$ be an infinite dimensional space endowed with a (pseudo-) metric d, where for any $\theta_j = (\beta_j, h_j) \in \Theta$, $j = 1, 2$, $d(\theta_1, \theta_2) = |\beta_1 - \beta_2|_e + d_H(h_1, h_2)$ with Euclidean distance $|\cdot|_e$ on B and a pseudo-metric $d_H(h_1, h_2)$ on $\mathcal{H}$. We assume that B is a compact subset in $\Re^{d_\beta}$ but that the function space $\mathcal{H}$ may not be compact in d_H. We introduce a class of approximate *penalized sieve extremum* (PSE) estimators, $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{h}_n) \in \Theta_{k(n)} = B \times \mathcal{H}_{k(n)}$, defined by:

$$\left\{\widehat{Q}_n(\hat{\theta}_n) - \lambda_n \widehat{P}_n(\widehat{h}_n)\right\} \geq \sup_{\theta \in B \times \mathcal{H}_{k(n)}} \left\{\widehat{Q}_n(\theta) - \lambda_n \widehat{P}_n(h)\right\} - O_P(\eta_n), \tag{3.1}$$

where $\{\eta_n\}_{n=1}^{\infty}$ is a sequence of positive real values such that $\eta_n = o(1)$; $\mathcal{H}_{k(n)}$ is a sieve parameter space whose complexity (denoted $k(n) \equiv \dim(\mathcal{H}_{k(n)})$) grows with sample size n and becomes dense in the original function space $\mathcal{H}$ under the (pseudo-) metric d_H; $\lambda_n \geq 0$ is a penalization parameter such that $\lambda_n \to 0$ as $n \to \infty$; and the penalty $\widehat{P}_n() \geq 0$, which is an empirical analog of a non-random penalty function $Pen : \mathcal{H} \to [0, +\infty)$, is jointly measurable in h and the data $\{Z_t\}_{t=1}^{n}$.

The sieve space $\mathcal{H}_{k(n)}$ in the definition of the PSE (3.1) could be finite dimensional, infinite dimensional, compact or non-compact (in d_H). Commonly used finite-dimensional linear sieves (also called *series*) take the form:

$$\mathcal{H}_{k(n)} = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot) \right\}, \quad k(n) < \infty, \ k(n) \to \infty \text{ slowly as } n \to \infty, \tag{3.2}$$

where $\{q_k\}_{k=1}^{\infty}$ is a sequence of known basis functions of a Banach space $(\mathbf{H}, d_H)$ such as wavelets, splines, Fourier series, Hermite polynomial series, Power series, Chebychev series, etc. Linear sieves with constraints, which are commonly used, can be expressed as:

$$\mathcal{H}_{k(n)} = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot), \ R_n(h) \leq B_n \right\}, \quad B_n \to \infty \text{ slowly as } n \to \infty, \tag{3.3}$$

where the constraint $R_n(h) \leq B_n$ reflects prior information about $h_0 \in \mathcal{H}$ such as smoothness properties. The sieve space $\mathcal{H}_{k(n)}$ in (3.3) is finite dimensional and compact (in d_H) if and only if $k(n) < \infty$ and $\mathcal{H}_{k(n)}$ is closed and bounded; it is infinite dimensional and compact (in d_H) if and only if $k(n) = \infty$ and $\mathcal{H}_{k(n)}$ is closed and totally bounded. For example, $\mathcal{H}_{k(n)} = \left\{h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} a_k q_k(\cdot), \ \|h\|_H \leq \log(n)\right\}$ is compact if $k(n) < \infty$, but it is not compact (in d_H) if $k(n) = \infty$. Linear sieves (or series) are widely used in econometrics. But, to approximate $h(\cdot)$ that depends on a high dimensional variable, nonlinear sieves such as Neural Networks, radial basis, ridgetlets, mixtures of some known distributions (or densities) or others could be more useful. See Chen (2007), DeVore and Lorentz (1993) and the references therein for additional examples of linear and nonlinear sieves.

The penalty function $Pen()$ is typically convex and/or *lower semicompact* (i.e., the set $\{h \in \mathcal{H} : Pen(h) \leq M\}$ is compact in $(\mathbf{H}, d_H)$ for all $M \in [0, \infty)$) and reflects prior information about $h_0 \in \mathcal{H}$.

For instance, when $\mathcal{H} \subseteq L^p(d\mu)$, $1 \leq p < \infty$, a commonly used penalty function is $\widehat{P}_n(h) = ||h||^p_{L^p(d\mu)}$ for a known measure $d\mu$, or $\widehat{P}_n(h) = ||h||^p_{L^p(d\widehat{\mu})}$ for an empirical measure $d\widehat{\mu}$ when $d\mu$ is unknown. When $\mathcal{H}$ is a mixed weighted Sobolev space $\{h : ||h||^2_{L^2(d\mu)} + ||\nabla^r h||^p_{L^p(leb)} < \infty\}$, $1 \leq p < \infty$, $r \geq 1$, we can let $|| \cdot ||_H$ be the $L^2(d\mu)-$norm, and $\widehat{P}_n(h) = ||h||^2_{L^2(d\widehat{\mu})} + ||\nabla^k h||^p_{L^p(leb)}$ or $\widehat{P}_n(h) = ||\nabla^k h||^p_{L^p(leb)}$ for some $k \in [1, r]$.

Our definition of PSE (3.1) includes both the method of sieves and the method of penalization (or regularization) as special cases. In particular, when $\lambda_n \widehat{P}_n() = 0$, the (approximate) PSE (3.1) becomes the solution to:

$$\widehat{Q}_n(\hat{\theta}_n) \geq \sup_{\theta \in B \times \mathcal{H}_{k(n)}} \widehat{Q}_n(\theta) - O_P(\eta_n), \quad \text{with } \eta_n \to 0 \text{ as } n \to \infty, \tag{3.4}$$

which is the original (approximate) *sieve extremum estimator* defined in Chen (2007). When $\lambda_n \widehat{P}_n() > 0$, $\widehat{P}_n() = Pen()$ and $\mathcal{H}_{k(n)} = \mathcal{H}$ (i.e., $k(n) = \infty$), the (approximate) PSE (3.1) becomes the solution to:

$$\left\{\widehat{Q}_n(\hat{\theta}_n) - \lambda_n Pen(\widehat{h}_n)\right\} \geq \sup_{\theta \in B \times \mathcal{H}} \left\{\widehat{Q}_n(\theta) - \lambda_n Pen(h)\right\}, \tag{3.5}$$

which is a *function space penalized (or regularized) extremum estimator.*

Which method should one use?

(1) Both the sieve method (3.4) and the function space penalization method (3.5) are quite flexible. A researcher has to make similar choices in applying either method. For the sieve method (3.4), one must choose the sieve space $\mathcal{H}_{k(n)}$ (and, for a given finite dimensional sieve, the number of sieve terms $k(n)$). For the penalization method (3.5), one must choose the penalty function $Pen(\cdot)$ and the regularization parameter λ_n. Both the choices of $\mathcal{H}_{k(n)}$ and $Pen(h)$ should be guided by prior information about smoothness and/or shape properties of the unknown function h as well as computational issues. In general, the smoothing parameters ($k(n)$ and λ_n) could be chosen via cross validation.

(2) From a theoretical point of view, sieve extremum estimators (3.4) and function space penalized extremum estimators (3.5) have similar large sample properties. For example, with an optimal choice of sieve number of terms $k(n)$ for the nonparametric part the sieve estimator $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{h}_n)$ defined in (3.4) can simultaneously achieve root-n asymptotic normality of the smooth functional part ($\widehat{\beta}_n$) and the optimal convergence rate for the nonparametric part ($\widehat{h}_n$). Likewise, with an optimal choice of the regularization parameter λ_n for the nonparametric part the penalization estimator $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{h}_n)$ defined in (3.5) can simultaneously achieve root-n asymptotic normality of $\widehat{\beta}_n$ and the optimal convergence rate for $\widehat{h}_n$. See Section 4 for details.

(3) The sieve extremum estimator (3.4) with finite dimensional sieves is much easier to compute. Once the unknown functions are approximated by finite dimensional sieves, the implementation of the sieve extremum estimation (3.4) is the same as parametric nonlinear extremum estimation. Also, with the sieve method it is easy to impose shape restrictions on unknown functions, such as monotonicity, concavity, additivity, non-negativity and other restrictions. In the numerical implementation of functional space penalized estimation (3.5), one typically expands the unknown function $h()$ in terms of infinite dimensional linear sieves, $h(\cdot) = \sum_{k=1}^{\infty} a_k q_k(\cdot)$, and then penalizes the sieve coefficients $\{a_k : k \geq 1\}$. See, e.g., Donoho, et al. (1995) for regularized wavelets and Eliers and Marx (1996) for penalized splines. This makes the penalized estimation very similar to penalized sieve estimation.

(4) When the problem is ill-posed (or when the identifiable uniqueness condition fails), in terms of finite sample performance as well as conditions for asymptotic optimal rate of convergence, it is better to use the sieve extremum estimator (3.4) with finite dimensional compact sieves such as (3.3) or the PSE estimator (3.1) with high but finite dimensional linear sieves (series) (3.2). See, e.g., Newey and Powell (2003), Ai and Chen (2003), Blundell, Chen and Kristensen (2007), Chen and Pouzo (2008, 2009a). This motivates us to present the more general penalized sieve method.

3.2 Penalized sieve M estimation

(Penalized) sieve M estimation is a special case of (penalized) sieve extremum estimation when $\widehat{Q}_n(\theta)$ in (3.1) is

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

where $l : \Theta \times \Re^{d_z} \to \Re$ is the criterion based on a single observation. In estimating $\theta_o = \arg\sup_{\theta\in\Theta} E[l(\theta, Z_t)]$, this is a natural procedure. It is called (penalized) sieve minimum contrast estimation in statistics.

Different choices of the criterion $l(\theta, Z_t)$ yield special cases of (penalized) sieve M estimation. Examples include (penalized) sieve Maximum Likelihood (ML), (penalized) sieve Quasi Maximum Likelihood (QML), (penalized) sieve Least Squares (LS), (penalized) sieve Generalized Least Squares (GLS), (penalized) sieve Quantile Regression (QR), and many others.

In econometrics, the SNP estimator proposed by Gallant and Nychka (1987) is really a special case of sieve MLE using Hermite polynomial sieves to approximate an unknown density. Heckman and Singer's (1984) nonparametric MLE (NPMLE) is simply a sieve MLE using a first order spline sieve to approximate the latent heterogeneity distribution.

Example 3.1 continued *(Semi-nonparametric mixture models)*: Recall that

$$\theta_o = (\beta_o, h_o) \in \arg\sup_{\beta\in B, h\in\mathcal{H}} E\left(\log\left\{\int_0^1 f(Y_t|X_t;\beta,u)h(u)du\right\}\right).$$

We can estimate θ_o via a sieve MLE $\widehat{\theta} = (\widehat{\beta}, \widehat{h}) \in B \times \mathcal{H}_{k(n)}$, which solves

$$\sup_{\theta\in B\times\mathcal{H}_{k(n)}} \frac{1}{n}\sum_{t=1}^{n}\log\left\{\int_0^1 f(Y_t|X_t;\beta,u)h(u)du\right\},$$

where the sieve space $\mathcal{H}_{k(n)}$ could be mixtures of Bernstein densities (see, e.g., Ghosal (2001)):

$$\mathcal{H}_{k(n)} = \left\{ h(u) = k(n) \sum_{j=1}^{k(n)} a_{j,k(n)} \left(\begin{array}{c} k(n)-1 \\ j-1 \end{array} \right) u^{j-1}(1-u)^{k(n)-j} : a_{j,k(n)} \geq 0, \sum_{j=1}^{k(n)} a_{j,k(n)} = 1 \right\}.$$

We could also estimate θ_o via a penalized MLE $\widehat{\theta} = (\widehat{\beta}, \widehat{h}) \in B \times \mathcal{H}$, which solves

$$\sup_{\theta \in B \times \mathcal{H}} \left\{ \frac{1}{n} \sum_{t=1}^{n} \log \left\{ \int_0^1 f(Y_t | X_t; \beta, u) h(u) du \right\} - \lambda_n Pen(h) \right\},$$

where $\lambda_n > 0$ and $\lambda_n \to 0$ slowly as $n \to \infty$. The penalty could be, for example, $Pen(h) = \int [\nabla h(u)]^2 du$ or $\int |\nabla h(u)| \, du$. See, e.g., Eggermont and LaRiccia (2001) for other penalties.

In empirical work, most people use finite dimensional sieve M estimation without any penalty. In terms of practical implementation of functional space penalized M estimation, there are two popular approaches. The first one is the smoothing spline approach; see, e.g., Wahba (1990), Koenker, Ng and Portnoy (1994) and Gu (2002). The second approach is to expand the unknown function $h()$ in terms of infinite dimensional linear sieves, $h(\cdot) = \sum_{k=1}^{\infty} a_k q_k(\cdot)$, and then penalizes the sieve coefficients $\{a_k : k \geq 1\}$; see, e.g., Donoho, et al. (1995) and Eliers and Marx (1996).[7]

An important special case of sieve M estimation in econometrics is 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 $\Theta_{k(n)}$.

Example 3.3 *(Series LS estimation):* $Y_t = \theta_o(X_t) + \varepsilon_t$, $E[\varepsilon_t | X_t] = 0$. Let $\{p_j(X), j = 1, 2, ...\}$ be a sequence of known basis functions that can approximate any $\theta \in \Theta$ well. $p^{k_n}(X) = (p_1(X), ..., p_{k_n}(X))'$. Then $\Theta_{k(n)} = \{h : h(x) = p^{k_n}(x)'A : A \in \Re^{k_n}\}$, with $k_n \to \infty$ slowly as $n \to \infty$, is a finite-dimensional linear sieve for Θ. And $\widehat{\theta}$ is a sieve (or series) LS estimator of $\theta_o = \arg\inf_{\theta \in \Theta} E\left([Y_t - \theta(X_t)]^2\right)$:

$$\widehat{\theta} = \arg \max_{\theta \in \Theta_{k(n)}} \frac{-1}{n} \sum_{t=1}^{n} [Y_t - \theta(X_t)]^2 = p^{k_n}(\cdot)'(P'P)^{-} \sum_{t=1}^{n} p^{k_n}(X_t) Y_t.$$

Partial list of empirical applications of sieve M estimation to economic time series models: Engle et al. (1986) forecast electricity demand using a partially linear spline LS regression. Engle and Gonzalez-Rivera (1991) apply sieve MLE to estimate ARCH models where the unknown density of the standardized innovation is approximated by a first order spline sieve. Chen and Conley (2001) apply a simple two-step sieve LS procedure to estimate a spatial temporal model where both the unknown conditional mean and unknown conditional covariance are approximated by shape-preserving cardinal B-spline wavelet sieves. Engle and Rangel (2007) propose a new Spline GARCH model to measure

[7] For example, a function penalty $Pen(h) = \int |\nabla h(u)| \, du$ would become ℓ_1 penalty on first difference of wavelet or spline sieve coefficients, which looks like a LASSO penalty for high-dimensional sparse models (see, e.g., Van de Geer (2008), Belloni and Chernozhukov (2011)).

unconditional volatility and have applied it to equity markets for 50 countries for up to 50 years of daily data. Audrino and Bühlmann (2009) leave the entire volatility process unspecified and approximated by B-spline sieves. White (1990) and Granger and Teräsvirta (1993) suggest nonparametric LS forecasting via sigmoid ANN sieves. Hutchinson et al. (1994) apply radial basis ANN to option pricing. Chen et al. (2001) use partially linear ANN and ridgelet sieves to forecast US inflation. McCaffrey et al. (1992) estimate the Lyapunov exponent of a chaotic system via ANN sieves. Phillips (1998) applies orthonormal bases to analyze spurious regressions. See Fan and Yao (2003) and Gao (2008) for additional applications to financial time series models.

3.3 Penalized sieve MD estimation

(Penalized) sieve MD estimation is a special case of (penalized) sieve extremum estimation in which $-\widehat{Q}_n(\theta)$ in (3.1) can be expressed as some distance from zero.

One typical minimum distance criterion takes the following quadratic form:

$$\widehat{Q}_n(\theta) = -n^{-1}\sum_{t=1}^{n}\widehat{m}(X_t,\theta)'\{\widehat{\Sigma}(X_t)\}^{-1}\widehat{m}(X_t,\theta), \tag{3.6}$$

where $\widehat{m}(X_t,\theta)$ is a consistently estimated vector-valued function $m(X_t,\theta)$ of fixed finite dimension, and $\widehat{\Sigma}(X_t)$ is a consistently estimated weighting matrix $\Sigma(X_t)$ that is introduced for efficiency. This is a natural procedure for estimating $\theta_o = \arg\inf_{\theta\in\Theta} E\left[m(X_t,\theta)'\{\Sigma(X_t)\}^{-1}m(X_t,\theta)\right]$.

We can apply the (penalized) sieve MD procedure to estimate models belonging to the class of semi-nonparametric conditional moment restrictions

$$E[\rho(Z_t,\beta_o,h_o(\cdot))|X_t] = 0, \tag{3.7}$$

where the difference $\rho(Z_t,\beta,h(\cdot))-\rho(Z_t,\beta_o,h_o(\cdot))$ depends on the endogenous variables Y_t. In particular, $\widehat{m}(X_t,\theta)$ could be any nonparametric estimator, such as a kernel, local linear regression or series LS estimator, of the conditional mean function $m(X_t,\theta) = E[\rho(Z_t,\theta)|X_t]$ with $\theta = (\beta,h)$. For example, a series LS estimator is

$$\widehat{m}(X_t,\theta) = p^{J_n}(X_t)'(P'P)^{-}\sum_{i=1}^{n}p^{J_n}(X_i)\rho(Z_i,\theta), \tag{3.8}$$

where $\{p_j()\}_{j=1}^{\infty}$ is a sequence of known basis functions that can approximate any square integrable function of X well, J_n is the number of approximating terms such that $J_n \to \infty$ slowly as $n \to \infty$, $p^{J_n}(X) = (p_1(X),...,p_{J_n}(X))'$, $P = (p^{J_n}(X_1),...,p^{J_n}(X_n))'$, and $(P'P)^{-}$ is the generalized inverse of the matrix $P'P$. See, e.g., Newey and Powell (1989, 2003), Ai and Chen (1999, 2003), Chen and Pouzo (2008, 2009a) and others for more details and applications of this estimator.

Another typical minimum distance criterion is the following *sieve GMM*:

$$\widehat{Q}_n(\theta) = -\widehat{g}_n(\theta)'\widehat{W}\widehat{g}_n(\theta) \tag{3.9}$$

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

$$E[\rho(Z_t, \theta_o)p_j(X_t)] = 0, \ j = 1, 2, ..., J_n, \tag{3.10}$$

where $\{p_j(X), j = 1, 2, ..., J_n\}$ is a sequence of known basis functions that can approximate any real-valued square integrable function of X well as $J_n \to \infty$. Let $p^{J_n}(X) = (p_1(X), ..., p_{J_n}(X))'$. It is now obvious that the semi-nonparametric conditional moment restrictions (3.7) can be estimated via the sieve GMM criterion (3.9) using $\widehat{g}_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} \rho(Z_t, \theta) \otimes p^{J_n}(X_t)$.

Partial list of empirical applications of sieve MD estimation to economic time series models: Chen and Ludvigson (2009) apply the sieve MD procedure to estimate a semi-nonparametric habit formation consumption asset pricing model where the unknown habit function is approximated via a sigmoid ANN sieve. Chen et al. (1998) employ a shape-preserving spline-wavelet sieve to estimate the eigenfunctions of a fully nonparametric scalar diffusion model from discrete-time low-frequency observations. Gallant and Tauchen (1989) and Gallant et al. (1991) employ Hermite polynomial sieves to study asset pricing and foreign exchange rates. Gallant and Tauchen (1996, 2004) use a combination of Hermite polynomial sieves and the simulated method of moments to solve many complicated asset pricing models with latent factors, and their methods have been widely applied in empirical finance. Bansal and Viswanathan (1993), Bansal et al. (1993) and Chapman (1997) consider sieve approximation of the whole stochastic discount factor (or pricing kernel) as a function of a few macroeconomic factors.

4 Large Sample Properties of PSE Estimators

The general theory on large sample properties of PSE estimation of unknown functions is technically involved and relies on the theory of empirical processes. Chen (2007) presented a detailed review on large sample properties of sieve extremum estimators that were available as of 2006. Since then, there have been additional convergence rate results for sieve M estimators, and there have been rapid advances on convergence rates of penalized sieve MD estimators for nonparametric conditional moment restriction models, a large class of nonparametric nonlinear (possibly ill-posed inverse) problems with unknown operators. Perhaps more importantly, there have been some recent developments on limiting distributions of plug-in PSE estimators of functionals that may or may not be root-n estimable and on simple inference methods.

4.1 Consistency, convergence rates of PSE estimators

In Chen (2007, theorem 3.1) we provide a consistency theorem for approximate sieve extremum estimators that allows for possibly ill-posed semi-nonparametric problems. Chen and Pouzo (2008) present a slightly more general consistency theorem for approximate PSE estimators, allowing for ill-posed problems and noncompact parameter spaces.

(I) Convergence rates of penalized sieve M estimators

Let $\theta_o = (\beta_o, h_o) = \arg\sup_{(\beta,h)\in B\times\mathcal{H}} E[l(\beta, h, Z_t)]$. Let $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{h}_n)$ be either the approximate sieve M estimator (with $\eta_n = o(1)$):

$$\frac{1}{n}\sum_{t=1}^{n} l(\hat{\theta}_n, Z_t) \geq \sup_{\theta\in B\times\mathcal{H}_{k(n)}} \frac{1}{n}\sum_{t=1}^{n} l(\theta, Z_t) - O_P(\eta_n),$$

or the approximate functional space penalized M estimator (with $\eta_n = o(1)$):

$$\frac{1}{n}\sum_{t=1}^{n} l(\hat{\theta}_n, Z_t) - \lambda_n Pen(\widehat{h}_n) \geq \sup_{\theta\in B\times\mathcal{H}} \left\{\frac{1}{n}\sum_{t=1}^{n} l(\theta, Z_t) - \lambda_n Pen(h)\right\} - O_P(\eta_n).$$

There are many results on convergence rates of sieve M estimators of unknown functions for i.i.d. data; see Chen (2007) for a detailed review and the references therein. Chen and Shen (1998) obtain the convergence rate for general sieve M estimation with stationary beta-mixing data; their convergence rate is the same as if the data were iid. Huang (2002) derives the convergence rate for a polynomial spline series LS estimator for weakly dependent strong mixing time series data. Both papers establish the convergence rates under a metric $\|\theta - \theta_o\| \asymp \sqrt{E\left[l(\theta_o, Z_t) - l(\theta, Z_t)\right]}$. For series LS regression Example 3.3, there are also some results on the convergence rate under the sup-norm $\|\theta - \theta_o\|_\infty = \sup_{x\in\mathcal{X}} |\theta(x) - \theta_o(x)|$ for iid data; see, e.g., Stone (1982), Newey (1997), de Jong (2002), and Song (2008). One could easily extend these sup-norm convergence rate results for iid data to series LS regression for strong mixing dependent data.

There are also many convergence rate results for functional space penalized M estimators for i.i.d. data; see, e.g., Shen (1997), van de Geer (2000) and the references therein. Chen (1997) established the convergence rates of the functional space penalized M estimators for weakly dependent data such as uniform mixing, beta mixing and strong mixing. Her convergence rate for uniform mixing and beta mixing time series can achieve the optimal rates of general penalized M estimators for iid data.[8]

The optimal rates of convergence are achieved by choosing the smoothing parameters, $k(n)$ for sieve M estimation and λ_n for penalized M estimation, to balance the bias and the complexity of the nonparametric models (or, roughly, the standard errors in nonparametric regression models). There are many theoretical results on data driven choices of smoothing parameters ($k(n)$ or λ_n) in nonparametric

[8] Chen (1997) has never been submitted for any journal publication because the author feels that function space penalized M estimation is not as practical as sieve M estimation.

M estimation of h_o. See Arlot and Celisse (2010), Hansen and Racine (2010), Leeb and Pötscher (2009), Ruppert et al. (2003), Barron et al. (1999), Shen and Ye (2002), Li (1987), Andrews (1991a), Hurvich et al. (1998), Stone et al. (1997), Coppejans and Gallant (2002), Phillips and Ploberger (2003) and others. In practice, cross-validation (CV) and small sample corrected AIC have been used; see, e.g., Ichimura and Todd (2007) for a recent review on implementation of series M estimators.

(II) Convergence rates of penalized sieve MD estimators

Many structural econometric models belong to the class of semi-nonparametric conditional moment restrictions (3.7). Recently, there has been a lot of work on identification and estimation of two important examples of this class of models. The first example is the nonparametric instrumental variables regression (NPIV):

$$E[Y_{1t} - h_0(Y_{2t})|X_t] = 0;$$

see, e.g., Newey and Powell (2003), Hall and Horowitz (2005), Blundell, Chen and Kristensen (2007), Carrasco, Florens and Renault (2007), Chen and Reiss (2010), Horowitz (2011), Darolles, Fan, Florens and Renault (2011) and others. The second example is the nonparametric quantile instrumental variables regression (NPQIV):

$$E[1\{Y_{1t} \leq h_0(Y_{2t})\}|X_t] = \gamma \in (0,1);$$

see, e.g., Chernozhukov and Hansen (2005), Chernozhukov, Imbens and Newey (2007), Horowitz and Lee (2007), Chen and Pouzo (2008, 2009a), Chernozhukov, Gagliardini and Scaillet (2010) and others. Most asset pricing models also imply the conditional moment restrictions (3.7); see, e.g., Example 2.1 (habit based asset pricing) and Chen and Pouzo (2009b, 2010) for nonparametric pricing of endogenous default risk.

Chen, Chernozhukov, Lee and Newey (2011) provide some sufficient conditions for identification of this class of models (3.7). Chen and Pouzo (2008, 2009a) propose a class of penalized sieve MD estimators $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{h}_n) \in \Theta_{k(n)} = B \times \mathcal{H}_{k(n)}$ defined as:

$$\widehat{\theta}_n = \arg \inf_{(\beta,h)\in B\times\mathcal{H}_{k(n)}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(X_i,\beta,h)'\widehat{\Sigma}(X_i)^{-1}\widehat{m}(X_i,\beta,h) + \lambda_n \widehat{P}_n(h) \right\}.$$

See their papers for a detailed study of consistency and the rate of convergence of this class of estimators, which allows for nonlinear ill-posed inverse problems such as the partially linear quantile IV regression $E[1\{Y_{3t} \leq Y_{1t}'\beta_o + h_o(Y_{2t})\}|X_t] = \gamma$. Horowitz (2010) considered a data-driven way to select sieve number of terms $k(n)$ in a sieve estimation of the NPIV model $E[Y_{1t} - h_0(Y_{2t})|X_t] = 0$. There is little work on model selection for the penalized sieve MD estimation for the general model (3.7).

4.2 Limiting distributions and inference for PSE estimation of functionals

Recall that a semi-nonparametric (or semiparametric) model consists of two sets of parameters $\theta = (\beta, h)$, where β is a vector of finite dimensional parameters of interest, and h is a vector of infinite dimensional parameters of interest (or nuisance parameters). In many economic applications, we are interested in conducting inference on a real valued functional $\phi : \Theta \rightarrow \Re$. Examples include $\phi(\theta_o) = \lambda' \beta_o$ for $0 \neq \lambda \in \Re^{d_\beta}$ and $(h_o(y_1^*), ..., h_o(y_d^*))\, \lambda$ for $0 \neq \lambda \in \Re^d$, where $h_o()$ is a real valued function.

A functional can be classified into three categories:

- either (a) $\phi(\theta_o)$ can be estimated at a $\sqrt{n}-$rate, (i.e., $\phi(\theta_o)$ is a regular functional, a smooth functional or a bounded functional); see van der Vaart (1991), Newey (1990) and Bickel et al. (1993);
- or (b) $\phi(\theta_o)$ can be best estimated at a slower than $\sqrt{n}-$rate, (i.e., $\phi(\theta_o)$ is a non-smooth functional or an unbounded functional);
- or (c) $\phi(\theta_o)$ can be estimated at a faster than $\sqrt{n}-$rate, typically at an $n-$rate such as in settings with structural breaks, parameters at the boundary, unit roots, etc.

Let $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{h}_n)$ be a consistent estimator of $\theta_o = (\beta_o, h_o)$ that is identified by a semi-nonparametric (or semiparametric) model. Then $\phi(\widehat{\theta}_n)$ is a simple plug-in estimator of the functional of interest $\phi(\theta_o)$. There are many general theoretical results on $\sqrt{n}-$asymptotic normality and semiparametric efficiency of various plug-in estimators of smooth functionals (category (a)); see, e.g. Chen (2007) for a recent review and the references therein. Also, there are some recent developments in estimation and inference of non-smooth functionals (category (b)). However, there is not yet well developed general theory on faster than $\sqrt{n}-$rate of functionals (category (c)) in semi-nonparametric models.

In this section we briefly survey recent results for categories (a) and (b) in which $\widehat{\theta}_n$ is estimated via the method of penalized sieve extremum estimation.

4.2.1 Simultaneous penalized sieve M estimators

(I) Smooth (or regular) functional case

For i.i.d data and when $\phi(\theta_o)$ is a smooth functional, there are many general theory papers about the $\sqrt{n}-$asymptotic normality of simultaneous sieve M estimators $\phi(\widehat{\theta}_n)$ of $\phi(\theta_o)$. See, e.g., Wong and Severini (1991) on profile nonparametric MLE, Shen (1997) on sieve MLE, Murphy and Van der Vaart (2000) on profile nonparametric MLE, van de Geer (2000) on semiparametric penalized M estimation, Shen (2002) on Bayesian sieve MLE, to name only a few. There are also several general theory papers on inference for smooth functionals; see, e.g., Murphy and Van der Vaart (2000) on the profiled nonparametric likelihood ratio, Shen and Shi (2005) on the sieve likelihood ratio, Cheng and Kosorok (2009) on the profile sampler, Cheng and Huang (2010) on the bootstrap of profile nonparametric M estimation, Kleijn and Bickel (2010) on semiparametric Bayesian Bernstein-Von Mises theorem, to name only a few.

For weakly dependent time series data, Chen and Shen (1998) and Chen (1997) respectively establish $\sqrt{n}$ asymptotic normality of sieve M estimation and penalized M estimation of $\phi(\theta_o)$. They also show that sieve MLE and penalized MLE are asymptotically efficient. One advantage of this (penalized) sieve M estimation of $\phi(\theta_o)$ is that the optimal choice of smoothing parameter for nonparametric part can also lead to root-n asymptotic normality of $\phi(\widehat{\theta}_n)$. Recently Chen, Liao and Sun (2011) provide a simple consistent estimator of the asymptotic variance of the sieve M estimator $\phi(\widehat{\theta}_n)$ of $\phi(\theta_o)$.

(II) Possibly non-smooth functional case

When $\phi(\theta_o)$ is a non-smooth functional such as $(h_o(y_1^*), ..., h_o(y_d^*))\,\lambda$ for $0 \neq \lambda \in \Re^d$, where $h_o()$ is a real valued function, there are not many general theory papers about the limiting distributions and inference for simultaneous sieve M estimators $\phi(\widehat{\theta}_n)$ of $\phi(\theta_o)$.

For i.i.d. data, Wang and Yang (2009b) provide uniform confidence bands for first order polynomial spline LS regression, Krivobokova, Kneib and Claeskens (2010) and Koenker (2010) respectively obtain uniform confidence bands for penalized spline LS regression and additive penalized quantile regression. Chen, Chernozhukov and Liao (2010) obtain uniform confidence bands for sieve M estimators of unknown functions $h_o()$. Their work extends earlier results (Newey (1997), Huang (2003) and others) on pointwise normality of series LS estimators or series density estimators.

For weakly dependent strongly mixing data, Yang and his co-authors have recently established some uniform confidence bands for a first order polynomial spline LS regression estimator; see, e.g., Song and Yang (2009, 2010), Wang and Yang (2010). For NED time series data, Andrews (1991b) obtained the pointwise limiting distribution of a series LS regression estimator. For beta mixing time series data, Chen, Liao and Sun (2011) derive the limiting distributions of sieve M estimators $\phi(\widehat{\theta}_n)$ of possibly non-smooth functionals $\phi(\theta_o)$, and provide a simple consistent estimator of the variance.

(III) Partially identified case

The above results all rely on the assumption that $\theta_o = (\beta_o, h_o)$ is the unique maximizer of $E[l(\beta, h, Z_t)]$ over $\Theta = B \times \mathcal{H}$. In many semi-nonparametric mixture models, such as structural search models, models with latent heterogeneity and state dependence, or dynamic discrete choice models with unspecified initial distributions, it is impossible to check whether the parameter of interest β_o is point identified or not. Recently, Chen, Tamer and Torgovitsky (2010) provided a simple weighted bootstrap method for inference for sieve MLE of β in partially identified semiparametric models.

4.2.2 Simultaneous penalized sieve MD estimators

In Subsection 3.1, we mentioned the existing results on identification of $\theta_o = (\beta_o, h_o)$ in the semi-nonparametric conditional moment model (3.7) and the consistency and the convergence rate of penalized sieve MD estimators $\widehat{\theta} \equiv (\widehat{\beta}, \widehat{h})$ of $\theta_o = (\beta_o, h_o)$. In this subsection we briefly describe the recent advances on asymptotic properties of the plug-in PSMD estimator $\phi(\widehat{\theta})$ of any real-valued functional $\phi(\theta_o)$.

(I) Smooth (or regular) functional case

Chamberlain (1992) and Ai and Chen (1999, 2003) derive the semiparametric efficiency bound for β_o satisfying the conditional moment restriction (3.7). For iid data and for the particular real-valued smooth functional $\phi(\theta_o) = \lambda' \beta_o$ that is identified by the model (3.7), Ai and Chen (1999, 2003) establish the $\sqrt{n}-$asymptotic normality of the simultaneous sieve MD estimator $\widehat{\beta}$ of β_o. Although the asymptotic variance of $\widehat{\beta}$ in general does not have a closed-form expression, they provide a simple consistent sieve estimator of the asymptotic covariance of $\widehat{\beta}$. They also show that the optimally weighted sieve MD estimator of β_o achieves the semiparametric efficiency bound of β_o.

Ai and Chen (1999, 2003) establish their results under the assumption that the generalized residual functions $\rho(Z, \beta, h(\cdot))$ are pointwise differentiable in $\theta_o = (\beta_o, h_o)$. In particular, their simple consistent asymptotic variance estimator of $\widehat{\beta}$ hinges on the continuous pointwise differentiability of the residual functions $\rho(Z; \beta, h(\cdot))$ in $\theta_o = (\beta_o, h_o)$. Chen and Pouzo (2009a) relax these assumptions and generalize Ai and Chen's results in several major ways. *First*, they show that, for the general semi-nonparametric conditional moment restrictions (3.7) with nonparametric endogeneity, the PSMD estimator $\widehat{\theta} \equiv (\widehat{\beta}, \widehat{h})$ can simultaneously achieve root-n asymptotic normality of $\widehat{\beta}$ and the optimal convergence rate of $\hat{h}$ (in strong norm $||\cdot||_H$), allowing for possibly nonsmooth residuals and/or a possibly noncompact (in $||\cdot||_H$) function space ($\mathcal{H}$) or noncompact sieve spaces ($\mathcal{H}_{k(n)}$). *Second*, Chen and Pouzo (2009a) show that a simple weighted bootstrap procedure can consistently estimate the limiting distribution of the PSMD $\widehat{\beta}$, even when the residual functions $\rho(Z; \beta, h(\cdot))$ could be non-smooth in $\theta_o = (\beta_o, h_o)$. This is the case in a partially linear quantile IV regression example $E[1\{Y_3 \leq Y_1' \beta_o + h_o(Y_2)\}|X] = \gamma \in (0, 1)$. They propose a weighted bootstrap to consistently approximate the confidence region. *Third*, Chen and Pouzo (2009a) show that their optimally weighted PSMD procedure achieves the semiparametric efficiency bound of β_o under nonsmooth residuals. *Fourth*, Chen and Pouzo (2009a) show that the profiled optimally weighted PSMD criterion is asymptotically chi-square distributed. This leads to an alternative confidence region construction method which involves inverting the profiled optimally weighted criterion function. This should be easier to compute than the weighted bootstrap. *Finally*, all the general theoretical results are established in terms of any nonparametric estimator of the conditional mean functions $E[\rho(Z; \beta, h)|X = \cdot]$. They also provide low level sufficient conditions in terms of the series least squares (LS) estimator of $E[\rho(Z; \beta, h)|X = \cdot]$.

For i.i.d. data, Ai and Chen (2007) consider an extension of (3.7) to a more general semiparametric conditional moment restriction with a different information set:

$$E[\rho_j(Z,\beta_o,h_o())|X_j]=0, \quad j=1,2,...,J \tag{3.11}$$

with finite J. Here $Z=(Y',X')'\in\mathcal{Z}$ denotes all the random variables, and $X_j\in\mathcal{X}_j$ denotes the conditioning variables used in the j^{th} equation $\rho_j(Z,\beta,h)$ for $j=1,...,J$. X_j is either equal to a subset of X or a degenerate random variable; and if X_j is degenerate, the conditional expectation $E[\rho_j(Z,\beta,h)|X_j]$ is the same as the unconditional expectation $E[\rho_j(Z,\beta,h)]$. There are many applications where different equations may require different sets of instruments. The semiparametric hedonic price system where some explanatory variables in some equations are correlated with the errors in other equations is one such example. Another example is the simultaneous equations model with measurement error in some exogenous variables or some omitted variables correlated with what would otherwise be exogenous variables. A semiparametric panel data model where some variables that are uncorrelated with the error in a given time period are correlated with the errors in previous periods is a third example. The triangular simultaneous equations system studied in Newey, Powell and Vella (1999), the dynamic panel sample selection model, and semiparametric game models with incomplete information also fit the general framework (3.11). Moreover, Ai and Chen (2007) allow for the possibility of misspecification, which is when

$$E[\sum_{j=1}^{J}\{E[\rho_j(Z,\beta,h())|X_j]\}^2]>0 \quad \text{for all} \quad \theta=(\beta,h)\in\Theta=B\times\mathcal{H}.$$

Let $m(X,\theta)\equiv(m_1(X_1,\theta),...,m_J(X_J,\theta))'$ with $m_j(X_j,\theta)\equiv E\{\rho_j(Z,\theta)|X_j\}$ and $\Sigma(X)$ be a $J\times J-$ positive definite weighting matrix. They assume that $\theta_*=(\beta_*,h_*)\in\Theta$ is the unique solution to $\inf_{\theta\in\Theta}E\{m(X,\theta)'\Sigma(X)^{-1}m(X,\theta)\}$. Clearly $m(X,\theta_*)=0$ if and only if the semiparametric conditional moment restriction model (3.11) is correctly specified, and in this case $\theta_*=\theta_o$.

For the general model (3.11) allowing for misspecification and for iid data, Ai and Chen (2007) propose a modified sieve MD estimator $\widehat{\theta}=(\widehat{\beta},\widehat{h})$ for $\theta_*=(\beta_*,h_*)$ and derive the asymptotic properties of $\widehat{\theta}$. Under low-level sufficient conditions, they show that: (i) $\widehat{\theta}$ converges to the pseudo-true value θ_* in probability; (ii) the plug-in sieve MD estimator $\phi(\widehat{\theta})$ of smooth functionals $\phi(\theta_*)$, including the estimators of β_* and the average derivative of h_*, are $\sqrt{n}-$asymptotically normally distributed; and (iii) the estimators for the asymptotic covariances of $\phi(\widehat{\theta})$ of smooth functionals are consistent and easy to compute. To the best of our knowledge, these results in Ai and Chen (2007) are the first to allow researchers to perform asymptotically valid tests of various hypotheses on the smooth functionals $\phi(\theta_*)$ regardless of whether model (3.11) is correctly specified or not.

(II) Possibly non-smooth functional case

For the semi-nonparametric conditional moment restrictions (3.7) with nonparametric endogeneity, it is in general difficult to check whether a real-valued functional $\phi(\theta_o)$ is a smooth (or regular) functional or not, since the problem could be a nonlinear ill-posed inverse problem with unknown operators. Recently, Chen and Pouzo (2010) established asymptotic normality of the plug-in PSMD estimator $\phi(\widehat{\theta})$ of a functional $\phi(\theta_o)$ that could be non-smooth (or slower than root-n estimable). They also provide two ways to construct asymptotically valid confidence sets for $\phi(\widehat{\theta})$. The first one is by inverting the optimally weighted criterion function. The second one is based on weighted bootstrap and is valid even for non-optimally weighted criterion functions. The authors are currently working on time series extensions.

(III) Partially identified case

The above results all rely on the assumption that $E\{E[\rho(Z,\beta,h)|X]'\Sigma(X)^{-1}E[\rho(Z,\beta,h)|X]\}$ is uniquely minimized at $\theta_o = (\beta_o, h_o) \in \Theta$. For the special case of NPIV model: $E[\rho(Z,\theta_o)|X] = E[Y_{1t} - h_0(Y_{2t})|X_t] = 0$, Santos (2010) considers how to construct confidence sets for $\phi(\theta_o)$ without imposing point identification. Currently we are working on a simple weighted bootstrap procedure for inference for the profiled, continuously updated optimally weighted penalized sieve MD estimator of β_0 when the model $E[\rho(Z,\beta,h)|X] = 0$ may have multiple solutions.

5 Semiparametric Two-step Estimation

For a semi-nonparametric model, $\theta_o \in \Theta$ consists of two parts $\theta_o = (\beta_o, h_o) \in \Theta = B \times \mathcal{H}$, where B denotes a finite dimensional compact parameter space and $\mathcal{H}$ denotes an infinite dimensional parameter space. In complicated empirical work, it is often difficult to jointly estimate $(\beta_o, h_o) = \arg\sup_{(\beta,h)\in B\times\mathcal{H}} Q(\beta,h)$. For an arbitrary $\beta \in B$, let

$$h_*(\cdot,\beta) = \arg\sup_{h\in\mathcal{H}} Q_1(\beta,h),\ \ \beta_o = \arg\max_{\beta\in B} Q_2(\beta, h_*(\cdot,\beta)),\ \ h_o = h_*(\cdot,\beta_o).$$

A computationally attractive alternative method is the **semiparametric two-step procedure**:

- Step 1: for an arbitrarily fixed $\beta \in B$, estimate the unknown $h_*(\cdot,\beta)$ using some nonparametric estimator $\widetilde{h}(\cdot,\beta)$, say, using a sieve extremum estimator $\widetilde{h}(\cdot,\beta) = \arg\max_{h\in\mathcal{H}_{k(n)}} \widehat{Q}_{1,n}(\beta,h)$;
- Step 2: estimate the unknown β_o by plugging in the estimated $h(\cdot)$ and using an existing nonlinear extremum procedure, say, $\widehat{\beta}_n = \arg\max_{\beta\in B} \widehat{Q}_{2,n}(\beta,\widetilde{h}(\cdot,\beta))$. Then $\widehat{h}_n(\cdot) = \widetilde{h}(\cdot,\widehat{\beta}_n)$.

We call $\widehat{\beta}_n$ a **semiparametric two-step M estimator** if

$$\widehat{Q}_{2,n}(\beta,\widetilde{h}(\cdot,\beta)) = \frac{1}{n}\sum_{t=1}^{n} l_2(\beta,\widetilde{h}(\cdot,\beta),Z_t)$$

and $Q_2(\beta,h_*(\cdot,\beta)) = E[l_2(\beta,h_*(\cdot,\beta),Z_t)]$ is maximized at $\beta = \beta_o \in B$.

We call $\widehat{\beta}_n$ a **semiparametric two-step GMM estimator** if

$$\widehat{Q}_{2,n}(\beta, \widetilde{h}(\cdot,\beta)) = -M_n(\beta, \widetilde{h}(\cdot,\beta))'W_n M_n(\beta, \widetilde{h}(\cdot,\beta))$$

and $Q_2(\beta, h_*(\cdot,\beta)) = -M(\beta, h_*(\cdot,\beta))'WM(\beta, h_*(\cdot,\beta))$, where $d_M \geq d_\beta$ and $M(\beta, h_*(\cdot,\beta)) = 0$ at $\beta = \beta_o \in B$. $M_n : B \times \mathcal{H} \to \Re^{d_M}$ is a random vector-valued function depending on the data $\{Z_t\}_{t=1}^n$, such that $M_n(\beta, h_*(\cdot,\beta))'WM_n(\beta, h_*(\cdot,\beta))$ is close to $M(\beta, h_*(\cdot,\beta))'WM(\beta, h_*(\cdot,\beta))$ for a symmetric matrix W. W_n is a possibly random weighting matrix such that $W_n - W = o_P(1)$.

The **(approximate) profile sieve extremum estimation procedure** is a special case of the semiparametric two-step procedure in which both steps use the same criterion function:

Step 1: for an arbitrarily fixed value $\beta \in B$, compute $\widehat{Q}_n(\beta, \widetilde{h}(\cdot,\beta)) \geq \sup_{h\in\mathcal{H}_{k(n)}} \widehat{Q}_n(\beta, h) - O_P(\eta_n)$ with $\eta_n = o(1)$;

Step 2: estimate β_o by $\widehat{\beta}_n$ solving $\widehat{Q}_n(\widehat{\beta}, \widetilde{h}(\cdot,\widehat{\beta})) \geq \max_{\beta\in B} \widehat{Q}_n(\beta, \widetilde{h}(\cdot,\beta)) - O_P(\eta_n)$, and then estimate h_o by $\widehat{h}_n = \widetilde{h}(\cdot, \widehat{\beta}_n)$.

Depending on the specific structure of a semi-nonparametric model, the (approximate) profile sieve extremum estimation procedure may be easier to compute. Nevertheless, the profile sieve extremum estimation is numerically equivalent to joint (or simultaneous) sieve extremum estimation of (β_o, h_o) by solving $\widehat{Q}_n(\widehat{\beta}_n, \widehat{h}_n) \geq \sup_{\beta\in B, h\in\mathcal{H}_{k(n)}} \widehat{Q}_n(\beta, h) - O_P(\eta_n)$.

Compared to a joint estimation procedure (i.e., simultaneous estimation of all the unknown parameters (β_o, h_o)), semiparametric two-step procedures are easier to compute, and with them it is easier to establish consistency and root-n asymptotic normality of smooth functionals (β). However, there are two main drawbacks of semiparametric two-step procedures. First, they are not semiparametrically efficient in general. Second, it is difficult to derive the asymptotic variance of $\widehat{\beta}_n$, Avar$(\widehat{\beta}_n)$, in closed form. Hence, it is difficult to provide consistent estimators of Avar$(\widehat{\beta}_n)$.

5.1 Consistent sieve estimators of Avar($\widehat{\beta}_n$)

There are many general theory papers on consistency and root-n asymptotic normality of semiparametric two-step estimators $\widehat{\beta}_n$ of smooth functionals β for various semiparametric models under various assumptions. See, e.g., Andrews (1994b), Newey (1994), Newey and McFadden (1994), Pakes and Olley (1995), Chen, Linton and van Keilegom (2003), Chen (2007), Ai and Chen (2007), and Ichimura and Lee (2010), to name a few. The results in Chen (2007, theorem 4.1 and lemma 4.2) allow for time series beta mixing processes. Ai and Chen (2007) and Ichimura and Lee (2010) allow for misspecified semiparametric models.

As we already mentioned, for complicated semiparametric models, it is difficult to derive the asymptotic variance of $\widehat{\beta}_n$, Avar$(\widehat{\beta}_n)$, in closed form, and hence it is difficult to provide consistent estimators of Avar($\widehat{\beta}_n$). For example, for the semiparametric two-step GMM estimator

$$\widehat{\beta}_n = \arg\min_{\beta \in B} M_n(\beta, \widetilde{h}(\cdot, \beta))' W_n M_n(\beta, \widetilde{h}(\cdot, \beta)), \tag{5.1}$$

Chen, Linton and van Keilegom (2003) and Chen (2007) establish root-n asymptotic normality under mild regularity conditions, allowing the unknown functions $h_o(\cdot) = h_*(\cdot, \beta_o)$ to depend on endogenous variables and to be estimated by any consistent nonparametric estimator $\widetilde{h}(\cdot, \beta)$ in the first step. Let $\Gamma_1 \equiv \Gamma_1(\beta_o, h_o)$, where $\Gamma_1(\beta, h_o)$ is the ordinary partial derivative of $M(\beta, h_o)$ in β, and let $\Gamma_2(\beta_o, h_o)[h - h_o] = \lim_{\tau \to 0}[M(\beta_o, h_o + \tau(h - h_o)) - M(\beta_o, h_o)]/\tau$ be the pathwise derivative of $M(\beta_o, h)$ in direction $[h - h_o]$. They show that

$$\sqrt{n}(\widehat{\beta} - \beta_o) \xrightarrow{d} \mathcal{N}[0, (\Gamma_1' W \Gamma_1)^{-1} \Gamma_1' W V_1 W \Gamma_1 (\Gamma_1' W \Gamma_1)^{-1}],$$

where $\Gamma_1' W \Gamma_1$ is nonsingular, $W = p\lim W_n$, and the finite matrix V_1 is such that

$$\sqrt{n}\{M_n(\beta_o, h_o) + \Gamma_2(\beta_o, h_o)[\widetilde{h}(\cdot, \beta_o) - h_o]\} \xrightarrow{d} \mathcal{N}[0, V_1].$$

To compute a consistent estimator of Avar($\widehat{\beta}_n$)= $(\Gamma_1' W \Gamma_1)^{-1} \Gamma_1' W V_1 W \Gamma_1 (\Gamma_1' W \Gamma_1)^{-1}$, one typically needs to estimate V_1 consistently. Unfortunately, without any information about the first step nonparametric estimator $\widetilde{h}(\cdot, \beta)$ it is generally very difficult to provide any consistent estimator of V_1. For complicated semi-nonparametric problems, say when there are several unknown functions or when unknown functions depend on endogenous variables, there is no closed form expression for V_1. Hence, it is difficult to estimate it consistently. This is why for iid data, Chen, Linton and van Keilegom (2003) suggest constructing an asymptotically valid confidence set for β via a nonparametric bootstrap. But, nonparametric bootstrap procedures are computationally intensive and work less well for semi-nonparametric time series models.

For i.i.d. data, Ai and Chen (2007) provide a consistent sieve estimator of the $Avar(\widehat{\beta}_n)$ for their modified sieve MD estimator for the general semiparametric conditional moment restrictions (3.11) with different information sets, where the unknown functions $h(\cdot)$ may depend on endogenous variables and the model (3.11) may not be correctly specified. A special case of their model is the so-called plug-in problem: $h_*() = \arg\inf_{h \in \mathcal{H}} E\left[(E[\rho_1(Z_t, h(\cdot))|X_{1t}])^2\right]$, $E[\rho_2(Z, \beta_o, h_*())] = 0$ with $\dim(\rho_2) = \dim(\beta)$. For this special case, their joint modified sieve MD estimation is equivalent to semiparametric two-step estimation where the first step is a sieve MD estimation of $h_*()$, $\widetilde{h}(\cdot) = \arg\inf_{h \in \mathcal{H}_{k(n)}} \frac{1}{n} \sum_{t=1}^{n} \left(\widehat{E}[\rho_1(Z, h(\cdot))|X_{1t}]\right)^2$ with $\widehat{E}[\rho_1(Z, h(\cdot))|X_{1t}]$ a series LS estimator of $E[\rho_1(Z_t, h(\cdot))|X_{1t}]$, and the second step is a method of moments estimation using $M_n(\beta, \widetilde{h}(\cdot)) = \frac{1}{n} \sum_{i=1}^{n} \rho_2(Z_i, \beta, \widetilde{h}(\cdot))$ in (5.1).

Newey (1984), Murphy and Topel (1985), Newey and McFadden (1994) and others present a general formula for computing the consistent asymptotic covariance matrix of the second stage estimator $\widehat{\beta}_n$ in a parametric two-step estimation framework. For iid data, Ackerberg, Chen and Hahn (2010) show that in a large class of semiparametric models, one can greatly simplify the estimation of $Avar(\widehat{\beta}_n)$, provided that the first stage unknown function h is estimated by a sieve (or series) method. They show, by extending earlier work of Newey (1994), that the consistent estimate of the *semiparametric* $Avar(\widehat{\beta}_n)$ using the method of Ai and Chen (2007) is *numerically identical* to the estimate of the *parametric* asymptotic variance using the standard parametric two-step framework of Murphy and Topel (1985).

For weakly dependent time series data, Chen, Hahn and Liao (2011) first propose a consistent sieve estimator of the $Avar(\widehat{\beta}_n)$ for a semiparametric two-step GMM estimator (5.1) when the first step unknown function is estimated via sieve M estimation. They then show that this consistent estimate of the *semiparametric* $Avar(\widehat{\beta}_n)$ is *numerically identical* to the estimate of the *parametric* asymptotic variance using the standard parametric two-step framework for time series data. These results greatly simplify the computation of semiparametric standard errors of semiparametric two-step GMM estimators for time series models.

5.2 Semiparametric multi-step estimation

In empirical work using complicated semiparametric models arising from dynamic games, Markov decisions, models with latent state variables, auctions, multivariate nonlinear time series with GARCH errors, and others, applied researchers sometimes have to perform the estimation of all the parameters of interest in multiple steps. Since it is already difficult to compute standard errors for semiparametric two-step estimators, it seems it would be a daunting task to characterize the asymptotic variance for the final step estimator $\widehat{\beta}_n$ in a multi-step procedure and provide consistent estimates of $Avar(\widehat{\beta}_n)$.

For i.i.d. data, Hahn and Ridder (2010) provide a characterization of the asymptotic variance for a class of semiparametric three-step estimators $\widehat{\beta}_n$, but they do not provide consistent estimates of $Avar(\widehat{\beta}_n)$. We conjecture that if the first or second step nonparametric parts are estimated by finite dimensional sieves, the results of Ackerberg, Chen and Hahn (2010) and Chen, Hahn and Liao (2011) can be generalized to the setting of semiparametric three-step estimation. This is a subject of ongoing research.

In specific applications, one could use the special properties of a semi-nonparametric model to characterize the asymptotic variances and to compute standard errors. We conclude this section with such an example.

Example 2.3 continued *(Semi-nonparametric GARCH + residual copula models)*: We estimate all the parameters and functions of interest by a simple three-step sieve M estimation procedure.

Step 1: For each series i, we perform sieve QMLE of the conditional mean and the semi-nonparametric GARCH(1,1) parameters as if the standardized innovation $\varepsilon_{i,t}$ were standard normal. Since the parameters associated with each series are estimated separately, we will suppress subscripts for now and let

Y_t denote any of the three return processes (S_t^e, M_t^e, B_t^e):

$$\begin{aligned} Y_t &= c + \rho Y_{t-1} + \beta M_{t-1}^e + \sigma_{Y,t}\varepsilon_{Y,t}, \\ \sigma_{Y,t}^2 &= \omega + \theta\sigma_{Y,t-1}^2 + h(\sigma_{Y,t-1}\varepsilon_{Y,t-1}), \end{aligned}$$

where $\beta = 0$ for the stock market process (M_t^e). We approximate each unknown function $h()$ (suppressing asset subscripts for now) via $h_{k(n)}()$, which is a 5 term cubic B-spline sieve or a 3rd order polynomial spline sieve excluding a constant term.

Let $\varphi = \left(c, \rho, \beta, \omega, \theta, h_{k(n)}\right)'$. We estimate φ via sieve QMLE $\widetilde{\varphi}$:

$$\widetilde{\varphi} = \arg\max_{\varphi} \frac{-1}{2n} \sum_{t=1}^{n} \left(\frac{\left(Y_t - c - \rho Y_{t-1} - \beta M_{t-1}^e\right)^2}{\sigma_{Y,t}^2(\varphi)} + \log \sigma_{Y,t}^2(\varphi) \right),$$

where given φ, $\sigma_{Y,t}^2(\varphi) = \omega + \theta\sigma_{Y,t-1}^2 + h_{k(n)}(\sigma_{i,t-1}\varepsilon_{i,t-1})$ is defined recursively (letting $\sigma_{Y,0}^2(\varphi)$ be the sample variance of Y_t).[9]

Step 2: estimation of the marginal distributions of standardized innovations. From Step 1, we can compute the fitted residual as:

$$\widetilde{\varepsilon}_{Y,t} = \frac{Y_t - \widetilde{c} - \widetilde{\rho} Y_{t-1} - \widetilde{\beta} M_{t-1}^e}{\sigma_{Y,t}(\widetilde{\varphi})}.$$

Given $\widetilde{\varphi}$ from Step 1**,** we estimate each F_i with the rescaled empirical distribution of $\widetilde{\varepsilon}_{i,t}$**:**

$$\widetilde{F}_{ni}(x) = \frac{1}{n+1} \sum_{t=1}^{n} 1\left(\widetilde{\varepsilon}_{i,t} \le x\right).$$

Step 3: estimation of copula parameters. We estimate α, the vector of copula dependence parameters, via pseudo MLE:

$$\widehat{\alpha} = \arg\max_{\alpha} \frac{1}{n} \sum_{t=1}^{n} \log c\left(\widetilde{F}_{nS}(\widetilde{\varepsilon}_{S,t}), \widetilde{F}_{nM}(\widetilde{\varepsilon}_{M,t}), \widetilde{F}_{nB}(\widetilde{\varepsilon}_{B,t}); \alpha\right).$$

Asymptotic properties and Inference: By applying existing results for GARCH models, one can show that each series is stationary beta mixing with an exponential decay rate. Step 1 estimation is a special case of sieve M estimation. By applying Chen and Shen (1998) for sieve M estimation with beta mixing data, we obtain root-n asymptotic normality of conditional mean and GARCH parameters as well as the optimal rate of convergence for the unknown function $h()$. By applying Chen, Liao and Sun (2011) for sieve M estimation with time series data, we can easily compute simple consistent variance estimators of

[9] We use Matlab to perform the QMLE computations. OLS estimates provide initial values for the conditional mean parameters. Standard GARCH(1,1) estimates provide initial values for the volatility parameters. Initial spline sieve coefficients are chosen so that the initial news impact curve matches the standard quadratic GARCH(1,1) estimate. Given these initial values, we first use the derivative-free, unconstrained "fminsearch" optimization function. We use the output of this step to initialize the derivative-based, constrained optimization routine "fmincon." Nonlinear constraints ensure positive volatility estimates.

sieve QMLEs of finite dimensional parameters as well as the pointwise confidence bands for $\widetilde{h}()$. Steps 2 and 3 follow directly from Chen and Fan (2006b) and Chan et al (2009). A surprising result established in Chen and Fan (2006b) and Chan et al (2009) is that the first step estimation of conditional mean and conditional variance of their parametric GARCH(p,q) model only affects the asymptotic variance of the second step rescaled empirical cdf $\widetilde{F}_{ni}()$ of the standardized innovations; the first step estimation does not affect the asymptotic variance of the final step pseudo MLE of copula dependence parameters. The only difference between our Example 2.3 and theirs is that our first step is a sieve GARCH(1,1) instead of a parametric GARCH(p,q). But we can adapt their results to obtain root-n asymptotic normality of the copula dependence parameter estimator $\widehat{\alpha}$ in step 3 as well as a simple consistent estimator of its asymptotic variance.

6 Concluding Remarks

In this selective review, we demonstrate the usefulness of semi-nonparametric models and methods for nonlinear economic and financial time series data. We briefly discuss a large class of flexible semi-nonparametric time series models and some of their temporal dependence properties. We present a general Penalized Sieve Extremum (PSE) estimation method that is very powerful and easy to compute for virtually all the semi-nonparametric problems. We review some recent large sample theory (consistency, convergence rate, limiting distribution) for penalized sieve M estimation for weakly dependent time series models. The method and results can be easily adapted to treat semi-nonparametric panel time series models and spatial models. We also present recent advances on large sample properties (consistency, convergence rate, limiting distribution) of penalized sieve MD estimation for cross sectional and panel data semi-nonparametric structural models, allowing for difficult (nonlinear) ill-posed inverse problems such as nonparametric instrumental variables problems. Some of these results can be easily extended to weakly dependent time series data and spatially dependent data. Recent advances in simple criterion based inference and consistent sieve estimation of asymptotic variances are also presented.

There are many unsolved issues in the study of semi-nonparametric dynamic models. For example, in empirical work it is difficult both to decide which class of semi-nonparametric nonlinear time series models to use and how many lagged dependent variables to include. It is also difficult to provide simple restrictions on the parameter spaces that are necessary and sufficient for particular temporal dependence properties. Also, estimation procedures originally designed for cross sectional semi-nonparametric models might have quite different performance in a time series context. For example, the non-stationary nonparametric instrumental variables example of Wang and Phillips (2009b) has properties which are quite different from those in the corresponding cross-sectional data case. As another example, the first order strictly stationary Markov process generated via Clayton copula and a fat tailed marginal distribution is beta mixing with an exponential decay rate, and hence the popular two-step pseudo maximum likelihood estimator of the copula dependence parameter originally proposed for bivariate iid data is still consistent and root-n asymptotically normally distributed. However, although this estimator performs well for bivariate iid data, it works terribly for time series with strong tail dependence. In particularly, it severely underestimates the tail dependence and hence underestimates the tail risk; see, Chen, Wu and Yi (2009).

There are also many open questions in the method of penalized sieve extremum estimation and its applications to economic semi-nonparametric time series models. We conclude this survey by listing a few of them. First, we need to establish large sample properties of PSE estimators for strongly dependent and nonstationary data. Second, it will be very fruitful to combine the PSE method with simulation based methods for semi-nonparametric dynamic models with nonlinear, non-Gaussian latent structures. Third, we need to design procedures that are robust to the lack of point identification and/or weak identification in complicated semi-nonparametric dynamic models. Recent theoretical work by Chernozhukov, Hong and Tamer (2007), Andrews and Cheng (2010), Andrews and Shi (2010), Chernozhukov, Lee and Rosen (2009) and others could be extended to semi-nonparametric settings. Fourth, there is little work on data-driven choices of smoothing parameters in penalized sieve MD estimation. Fifth, although for PSE estimators the optimal smoothing parameter choices that lead to nonparametric optimal rates of convergence could also lead to root-n asymptotic normality of smooth functionals, we need to investigate data driven methods of choosing smoothing parameters for plug-in PSE estimation of non-smooth functionals.

References

[1] Abel, A. (1990) “Asset Prices Under Habit Formation and Catching-up With Joneses”, *American Economic Review Papers and Proceedings*, 80, 38-42.

[2] Ackerberg, D., X. Chen, and J. Hahn (2010) “A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators”, *Review of Economics and Statistics*, Forthcoming.

[3] Ai, C., and X. Chen (2003) “Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions”, *Econometrica*, 71, 1795-1843. Working paper version, 1999.

[4] Ai, C., and X. Chen (2007) “Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables”, *Journal of Econometrics*, 141, 5-43.

[5] Aït-Sahalia, Y., L. Hansen and J. Scheinkman (2009) “Operator Methods for Continuous-Time Markov Processes”, in Y. Aït-Sahalia and L.P. Hansen (eds.), *Handbook of Financial Econometrics*. Amsterdam: North-Holland.

[6] Amemiya, T. (1985) *Advanced Econometrics*. Cambridge: Harvard University Press.

[7] Andersen, T.G. (1996) "Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility", *Journal of Finance*, 51, 169-204.

[8] Andrews, D. (1984) "Non-Strong Mixing Autoregressive Processes", *Journal of Applied Probability*, 21, 930-934.

[9] 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.

[10] Andrews, D. (1991b) "An Empirical Process Central Limit Theorem for Dependent Non-identically Distributed Random Variables", *Journal of Multivariate Analysis*, 38, 187-203.

[11] Andrews, D. (1994a) "Empirical process method in econometrics", in R.F. Engle III and D.F. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. North-Holland, Amsterdam.

[12] Andrews, D. (1994b) "Asymptotics for Semi-parametric Econometric Models via Stochastic Equicontinuity", *Econometrica,* 62, 43-72.

[13] Andrews, D. and X. Cheng (2010) "Estimation and Inference with Weak, Semi-strong, and Strong Identification", Cowles Foundation Discussion Paper No. 1773.

[14] Andrews, D. and X. Shi (2010) "Inference Based on Conditional Moment Inequalities", Cowles Foundation Discussion Paper No. 1761.

[15] Andrews, D. and Y.-J. Whang (1990) "Additive Interactive Regression Models: Circumvention of the Curse of Dimensionality", *Econometric Theory*, 6, 466-479.

[16] Arlot, S. and A. Celisse (2010) "A Survey of Cross-Validation Procedures for Model Selection", *Statistics Surveys*, 4, 40-79.

[17] Audrino, F. and P. Buehlmann (2009) "Splines for Financial Volatility", *Journal of the Royal Statistical Society*, 71, 655-670.

[18] Bansal, R., D. Hsieh and S. Viswanathan (1993) "A New Approach to International Arbitrage Pricing", *The Journal of Finance*, 48, 1719-1747.

[19] Bansal, R. and S. Viswanathan (1993) "No Arbitrage and Arbitrage Pricing: A New Approach", *The Journal of Finance*, 48(4), 1231-1262.

[20] Barnett, W.A., J. Powell and G. Tauchen (1991) *Non-parametric and Semi-parametric Methods in Econometrics and Statistics*. New York: Cambridge University Press.

[21] Barron, A., L. Birgé, P. Massart (1999) "Risk bounds for model selection via penalization", *Probab. Theory Related Fields*, 113, 301-413.

[22] Beare, B.K. (2010) "Copulas and Temporal Dependence", *Econometrica*, 78, 395-410.

[23] Belloni, A. and V. Chernozhukov (2011) "L1-Penalized Quantile Regression in High-Dimensional Sparse Models," *Annals of Statistics,* forthcoming.

[24] Bickel, P.J., C.A.J. Klaassen, Y. Ritov and J.A. Wellner (1993) *Efficient and adaptive estimation for semiparametric models*. Baltimore: The John Hopkins University Press.

[25] Bierens, H. J. (1987) "Kernel Estimators of Regression Functions", in T. F. Bewley (ed.), *Advances in Econometrics: Fifth World Congress*, vol. 1. Cambridge University Press.

[26] Billingsley, P. (1968) *Convergence of Probability Measures*. New York: Wiley.

[27] Blundell, R., X. Chen and D. Kristensen (2007) "Semi-nonparametric IV estimation of shape invariant Engel curves", *Econometrica*, 75, 1613-1669.

[28] Blundell, R. and J.L. Powell (2003) "Endogeneity in Nonparametric and Semiparametric Regression Models", in M. Dewatripont, L.P. Hansen and S.J. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Vol. 2. Cambridge, UK: Cambridge University Press.

[29] Bollerslev, T. (1986) "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, 31, 307-327.

[30] Bouyé, E. and M. Salmon (2009) "Copula Quantile Regressions and Tail Area Dynamic Dependence in Forex Markets", *The European Journal of Finance*, Vol. 15, Issue 7 and 8, 721-750.

[31] Bradley, R.C. (2007) *Introduction to Strong Mixing Conditions*, vols. 1-3. Heber City: Kendrick Press.

[32] Cai, Z., J. Fan and Q. Yao (2000) "Functional-coefficient Regression Models for Nonlinear Time Series", *Journal of American Statistical Association*, 95, 941-956.

[33] Campbell, J. and J. Cochrane (1999) "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior", *Journal of Political Economy*, 107, 205-251.

[34] Carrasco, M. and X. Chen (2002) "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models", *Econometric Theory*, 18, 17-39.

[35] Carrasco, M., J.-P. Florens and E. Renault (2007) "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization", in J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, vol. 6. Amsterdam: North-Holland.

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

[37] Chan, N., J. Chen, X. Chen, Y. Fan and L. Peng (2009) "Statistical Inference for Multivariate Residual Copula of Garch Models", *Statistica Sinica*, 19, 53-70.

[38] Chapman, D. (1997) “Approximating the Asset Pricing Kernel”, *Journal of Finance,* 52(4), 1383-1410.

[39] Chen, R. and R. Tsay (1993a) “Functional-coefficient Autoregressive Models”, *Journal of American Statistical Association*, 88, 298-308.

[40] Chen, R. and R. Tsay (1993b) “Nonlinear additive ARX Models”, *Journal of American Statistical Association*, 88, 955-967.

[41] Chen, X. (1995) “Nonparametric Recursive Moment Estimation with Dependent Data”, University of Chicago, unpublished working paper.

[42] Chen, X. (1997) “Rate and Normality of Penalized Extremum Estimates with Time Series Observations”, University of Chicago, unpublished working paper.

[43] Chen, X. (2007) “Large Sample Sieve Estimation of Semi-Nonparametric Models”, in J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, vol. 6B. Amsterdam: North-Holland.

[44] Chen, X., V. Chernozhukov, S. Lee and W. Newey (2011) “Identification in Semiparametric and Nonparametric Conditional Moment Models”, Yale, MIT and UCL, unpublished working Paper.

[45] Chen, X., V. Chernozhukov and Z. Liao (2010) “On Uniform Confidence Bands for Sieve M estimators of unknown functions”, Yale and MIT, unpublished working paper.

[46] Chen, X. and T. Conley (2001) “A New Semiparametric Spatial Model for Panel Time Series”, *Journal of Econometrics,* 105, 59-83.

[47] Chen, X., and Y. Fan (2006a): “Estimation and Model Selection of Semiparametric Copula-based Multivariate Dynamic Models under Copula Misspecification”, *Journal of Econometrics*, 135, 125-154.

[48] Chen, X., and Y. Fan (2006b): “Estimation of copula-based semiparametric time series models”, *Journal of Econometrics*, 130, 307–335.

[49] Chen, X., J. Favilukis and S. Ludvigson (2009) “On Estimation of Economic Models with Recursive Preferences”, Yale, LSE and NYU, unpublished working paper.

[50] Chen, X., J. Hahn and Z. Liao (2011) “Simple Estimation of Asymptotic Variance for Semiparametric Two-step Estimators with Weakly Dependent Data”, Yale and UCLA, unpublished working Paper.

[51] Chen, X., L.P. Hansen and M. Carrasco (2010) “Nonlinearity and Temporal Dependence”, *Journal of Econometrics*, 155, 155-169.

[52] Chen, X., L.P. Hansen and J. Scheinkman (1998) “Shape-preserving Estimation of Diffusions”, University of Chicago, unpublished working Paper.

[53] Chen, X., R. Koenker, and Z. Xiao (2009) "Copula-Based Nonlinear Quantile Autoregression", *the Econometrics Journal*, vol. 12, 50-67.

[54] Chen, X., Z. Liao and Y. Sun (2011) "On Inference of Sieve M-estimation of functionals with Weakly Dependent Data", Yale and UCSD, unpublished working Paper.

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

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

[57] Chen, X., and D. Pouzo (2008) "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals", *Cowles Foundation Discussion Paper*, No. 1650R.

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

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

[60] Chen, X. and D. Pouzo (2010) "On Inference of PSMD Estimators of Functionals of Nonparametric Conditional Moment Restrictions", Yale and UC Berkeley, unpublished working paper.

[61] Chen, X., D. Pouzo, and E. Tamer (2009) "Estimation and Inference of Partially Identified Semi-nonparametric Conditional Moment Models", working paper.

[62] Chen, X., J. Racine and N. Swanson (2001) "Semiparametric ARX Neural Network Models with an Application to Forecasting Inflation", *IEEE Tran. Neural Networks,* 12, 674-683.

[63] Chen, X. and M. Reiß (2010) "On Rate Optimality for Ill-Posed Inverse Problems in Econometrics", *Econometric Theory*, forthcoming.

[64] Chen, X. and X. Shen (1998) "Sieve Extremum Estimates for Weakly Dependent Data", *Econometrica*, 66, 289-314.

[65] Chen, X., E. Tamer and A. Torgovitsky (2010) "Sensitivity Analysis in Partially Identified Semiparametric Likelihood Models", Yale and Northwestern, unpublished working paper.

[66] Chen, X. and H. White (1998) "Central Limit and Functional Central Limit Theorems for Hilbert-Valued Dependent Heterogeneous Arrays with Applications", *Econometric Theory*, 260-284.

[67] Chen, X. and H. White (1999) "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators", *IEEE Tran. Information Theory,* 45, 682-691.

[68] Chen, X. and H. White (2002) "Asymptotic Properties of Some Projection-based Robbins-Monro Procedures in a Hilbert Space", *Studies in Nonlinear Dynamics and Econometrics*, vol. 6, issue 1, article 1.

[69] Chen, X., W. Wu and Y. Yi (2009) "Efficient estimation of copula-based semiparametric Markov models", *Annals of Statistics*, 2009, 37(6B), 4214-4253.

[70] Cheng, G. and J. Huang (2010) "Bootstrap consistency for general semiparametric M-estimation", *The Annals of Statistics*, 38, 5, 2884-2915.

[71] Cheng, G. and M.R. Kosorok (2009) "The penalized profile sampler", *Journal of Multivariate Analysis*, 100, 345-362.

[72] Chernozhukov, V., P. Gagliardini and O. Scaillet (2010) "Nonparametric instrumental variable estimation of quantile structural effects", Working Paper.

[73] Chernozhukov, V., and C. Hansen (2005) "An IV Model of Quantile Treatment Effects", *Econometrica,* 73, 245-261.

[74] Chernozhukov, V., H. Hong and E. Tamer (2007) "Estimation and Inference on Identified Parameter Sets", *Econometrica*, 75, 5, 1243-1284.

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

[76] Chernozhukov, V., S. Lee, and A. Rosen (2009) "Interesection Bounds: Estimation and Inference", Working Paper.

[77] Cherubini U., F. Gobbi, S. Mulinacci, and S. Romagnoli (2010) "On the Term Structure of Multivariate Equity Derivatives", Working Paper.

[78] Cochrane, J. (2001) *Asset Pricing.* Princeton: Princeton University Press.

[79] Constantinides, G. (1990) "Habit-formation: A Resolution of the Equity Premium Puzzle", *Journal of Political Economy*, 98, 519-543.

[80] Coppejans, M. and A.R. Gallant (2002) "Cross-validated SNP density estimates", *Journal of Econometrics,* 110, 27-65.

[81] Darolles, S, Y. Fan, J.-P. Florens, and E. Renault (2011) "Nonparametric Instrumental Regression", *Econometrica,* forthcoming.

[82] Davidson, J. (1994) *Stochastic Limit Theory: An Introduction for Econometricians.* Oxford: Oxford University Press.

[83] Davis, R.A., Lee, T., and Rodriguez-Yam, G. (2005) "Structural Break Estimation for Nonstationary Time Series Signals", Proceedings of IEEE/SP 13th Workshop on Statistical Signal Processing. Bordeaux, France (July 2005).

[84] de Jong, R. (2002) "A Note on 'Convergence rates and asymptotic normality for series estimators:' Uniform convergence rates", *Journal of Econometrics,* 111, 1-9.

[85] DeVore, R.A. and G. G. Lorentz (1993) *Constructive Approximation.* Springer-Verlag, Berlin.

[86] Donald, S. and W. Newey (2001) "Choosing the Number of Instruments", *Econometrica*, 69, 1161-1191.

[87] Donoho, D. L., I. M. Johnstone, G. Kerkyacharian and D. Picard (1995) "Wavelet Shrinkage: Asymptopia?" *Journal of the Royal Statistical Society, Series B*, 57, 301-369.

[88] Douc, R., E. Moulines, J. Olsson and R. van Handel (2011) "Consistency of the Maximum Likelihood Estimator for General Hidden Markov Models", *the Annals of Statistics*, 39, 474-513.

[89] Doukhan, P., P. Massart and E. Rio (1995) "Invariance Principles for Absolutely Regular Empirical Processes," *Ann. Inst. Henri Poincaré - Probabilités et Statistiques*, 31, 393-427.

[90] Doukhan, P. (1994) *Mixing: Properites and Examples*, New York: Springer-Verlag.

[91] Doukhan, P. and S. Louhichi (1999) "A new weak dependence condition and applications to moment inequalities", *Stochastic Processes and their Applications*, 84, 313-342.

[92] Duflo, M. (1997) *Random Iterative Models.* Heidelberg: Springer-Verlag.

[93] Eliers, P. and Marx, B. (1996) "Flexible smoothing with B-splines and penalties (with Discussion)", *Statistical Science*, 89, 89-121.

[94] Embrechts, P. (2008) "Copulas: A personal view," forthcoming in *Journal of Risk and Insurance.*

[95] Engle, R. (1982) "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom inflation", *Econometrica*, 50, 987-1007.

[96] Engle, R.F. (2010) "Long Term Skewness and Systemic Risk", Presidential Address SoFiE, 2009.

[97] Engle, R. and G. Gonzalez-Rivera (1991) "Semiparametric ARCH Models", *Journal of Business and Economic Statistics*, 9, 345-359.

[98] Engle, R., C. Granger, J. Rice and A. Weiss (1986) "Semiparametric Estimates of the Relation between Weather and Electricity Sales", *Journal of the American Statistical Association*, 81, 310-320.

[99] Engle, R. and S. Manganelli (2004) "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles", *Journal of Business and Economic Statistics*, Vol. 22, 4, 367-381.

[100] Engle, R.F. and D.F. McFadden (1994) *The Handbook of Econometrics*, vol. 4. Amsterdam: North-Holland.

[101] Engle, R.F and V. Ng (1993) "Measuring and Testing the Impact of News On Volatility", *Journal of Finance*, 48, 1749-1778.

[102] Engle, R.F. and J.G. Rangel (2007) "The Spline-GARCH Model for Unconditional Volatility and its Global Macroeconomic Causes", *Review of Financial Studies.*

[103] Engle, R.F and J.R. Russell (1998) "Autoregressive conditional duration: A new model for irregularly spaced transaction data", *Econometrica*, 66, 1127-1162.

[104] Fan, J. (2005) "A selective overview of nonparametric methods in financial econometrics", *Statistical Science* 20, 317-357.

[105] Fan, J. and I. Gijbels (1996) *Local Polynomial Modelling and Its Applications.* London: Chapman and Hall.

[106] Fan, J. and Y. Wang (2007) Multi-scale Jump and Volatility Analysis for High-Frequency Financial Data. *Journal of the American Statistical Association* 102, 1349-1362.

[107] Fan, J. and Q. Yao (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods.* New York: Springer-Verlag.

[108] Florens, J.-P. (2003) "Inverse Problems and Structural Econometrics: The Example of Instrumental Variables", in M. Dewatripont, L.P. Hansen and S.J. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications - Eight World Congress*, Econometric Society Monographs, Vol. 36. Cambridge University Press.

[109] Fostel, A. and J. Geanakoplos (2010) "Why Does Bad News Increase Volatility and Decrease Leverage", Cowles Foundation Discussion Paper No. 1762.

[110] Franke, J., J.P. Kreiss and E. Mammen (2009) "Nonparametric Modeling in Financial Time Series", in T. Mikosch, J.P. Kreiss, R.A. Davis and T.G. Andersen (eds.), *Handbook of Financial Time Series.* New York: Springer.

[111] Gallant, A.R. (1987) "Identification and Consistency in Seminonparametric Regression", in T. F. Bewley (ed.), *Advances in Econometrics: Fifth World Congress*, vol. 1. Cambridge University Press.

[112] Gallant, A.R. and D. Nychka (1987) "Semi-non-parametric maximum likelihood estimation", *Econometrica*, 55, 363-390.

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

[114] Gallant, A.R. and G. Tauchen (1996) "Which Moments to Match?" *Econometric Theory*, 12, 657-681.

[115] Gallant, A.R. and G. Tauchen (2004) "EMM: A Program for Efficient Method of Moments Estimation, Version 2.0 User's Guide", Working paper, Duke University.

[116] Gallant, A.R. and H. White (1988) *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models.* Oxford: Basil Blackwell.

[117] Gallant, A.R., D. Hsieh and G. Tauchen (1991) "On Fitting a Recalcitrant Series: The Pound/Dollar Exchange Rate, 1974-83", in Barnett, W.A., J. Powell and G. Tauchen (eds.), *Nonparametric and Semi-parametric Methods in Econometrics and Statistics,* 199-240, Cambridge: Cambridge University Press.

[118] Gao, J. (2007), *Nonlinear Time Series: Semiparametric and Nonparametric Methods*. London: Chapman & Hall/CRC.

[119] Geanakoplos, J. (2010) "The Leverage Cycle", in D.Acemoglu, K. Rogoff, and M. Woodford (eds.), *NBER Macro-economics Annual 2009*, vol. 24, University of Chicago Press, Chicago, 2010, pp. 1-65.

[120] Ghosal, S. (2001) "Convergence Rates for Density Estimation with Bernstein Polynomials," *Annals of Statistics*, 29, 1264-1280.

[121] Giraitis, L., R. Leipus and D. Surgailis (2008) "ARCH(∞) models and long-memory properties," in T.G. Andersen, R.A. Davis, J.-P. Kreiss and T. Mikosch (eds.), *Handbook of Financial Time Series*. New York: Springer.

[122] Granger, C.W.J. (2003) "Time series concepts for conditional distributions", *Oxford Bulletin of Economics and Statistics*, 65, supplement 689-701.

[123] Granger, C.W.J., and T. Teräsvirta (1993) *Modelling nonlinear economic relationships.* New York: Oxford.

[124] Grenander, U. (1981) *Abstract Inference*, New York: Wiley Series.

[125] Gu, C. (2002) *Smoothing Spline ANOVA Models*, New York: Springer.

[126] Haerdle, W., H. Liang and J. Gao (2000) *Partially Linear Models.* Heidelberg: Physica Verlag.

[127] Haerdle, W., H. Luetkepohl, and R. Chen (1997) "A Review of Nonparametric Time Series Analysis", *International Statistical Review*, 65, 49-72.

[128] Haerdle, W., M. Mueller, S. Sperlich and A. Werwatz (2004) *Nonparametric and Semiparametric Models.* New York: Springer.

[129] Hahn, J. and G. Ridder (2010) "The Asymptotic Variance of Semi-parametric Estimators with Generated Regressors", UCLA and USC, working paper.

[130] Hall, P. and C.C. Heyde (1980) *Martingale Limit Theory and Its Application.* Boston: Academic Press.

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

[132] Hamilton, J.D. (1989) "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle", *Econometrica*, 57, 357-384.

[133] Hamilton, J.D. (1994) “State-Space Models”, in R.F. Engle III and D.F. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. Amsterdam: North-Holland.

[134] Hansen, B. (1996) “Inference in TAR models”, *Studies in Nonlinear Dynamics and Econometrics*, Vol. 2, 1.

[135] Hansen, B. and J. Racine (2010) “Jackknife Model Averaging”, University of Wisconsin, unpublished working paper.

[136] Hansen, L.P. (1982) “Large Sample Properties of Generalized Method of Moments Estimators”, *Econometrica*, 50, 1029-1054.

[137] Hansen L.P., J. Heaton, J. Lee and N. Roussanov (2007) “Intertemporal Substitution and Risk Aversion”, in J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, vol. 6. Amsterdam: North-Holland.

[138] Hansen, L.P. and E. Renault (2010) “Pricing Kernels and Stochastic Discount Factors”, *Encyclopedia of Quantitative Finance*, Chapter 19-009, Wiley Press.

[139] Hansen, L.P. and T.J. Sargent (2007) “Robust Estimation and Control Without Commitment,” *Journal of Economic Theory*, 136, 1-27.

[140] Hansen, L.P. and J.A. Scheinkman (1995) “Back To the Future: Generating Moment Implications for Continuous Time Markov-Processes”, *Econometrica*, 63, 767- 804.

[141] Hansen, L.P. and K. Singleton (1982) “Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models”, *Econometrica*, 50, 1269-86.

[142] Heckman, J.J. and E.E. Leamer (2007) *The Handbook of Econometrics*, vol. 6. Amsterdam: North-Holland.

[143] Heckman, J. and B. Singer (1984) “A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data”, *Econometrica*, 68, 839-874.

[144] Hidalgo, J. (1997) “Non-parametric Estimation with Strongly Dependent Multivariate Time Series,” *Journal of Time Series Analysis*, 18, 95-122.

[145] Horowitz, J. (2009) *Semiparametric and Nonparametric Methods in Econometrics.* New York: Springer-Verlag.

[146] Horowitz, J. (2010) “Adaptive Nonparametric Instrumental Variables Estimation: Empirical Choice of the Regularization Parameter,” Northwestern, unpublished working paper.

[147] Horowitz, J. (2011) “Applied Nonparametric Instrumental Variables Estimation”, *Econometrica*, 79, 347–394.

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

[149] Huang, J. (2002) “The use of polynomial splines in nonlinear time series modeling”, University of Pennsylvania, unpublished working paper.

[150] Huang, J. (2003) “Local asymptotics for polynomial spline regression”, *The Annals of Statistics*, 31, 1600-1635.

[151] Huang, J. and H. Shen (2004) “Functional Coefficient Regression Models for Nonlinear Time Series: a Polynomial Spline Approach”, *Scandinavian Journal of Statistics*, 31, 515-534.

[152] Huang, J. and L. Yang (2004) “Identification of Non-Linear Additive Autoregressive Models”, *Journal of Royal Statistical Society, Series B*, 66, p. 463-477.

[153] Hurvich, C., J. Simonoff and C. Tsai (1998) “Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion”, *Journal of the Royal Statistical Society, Series B*, 60, 271-293.

[154] Hutchinson, J., A. Lo and T. Poggio (1994) “A non-parametric approach to pricing and hedging derivative securities via learning networks”, *Journal of Finance*, 3, 851-889.

[155] Ibragimov, R. (2009) “Copula-based characterizations for higher-order Markov processes”, *Econometric Theory*, 25, 819-846.

[156] Ibragimov, R. and G. Lentzas (2009) “Copulas and long memory”, Harvard Institute of Economic Research Discussion Paper No. 2160.

[157] Ibragimov, R. and P.C.B Phillips (2008) “Regression asymptotics using martingale convergence methods”, *Econometric Theory*, 24, 888-947.

[158] Ichimura, H. (1993) “Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models”, *Journal of Econometrics,* 58, 71-120.

[159] Ichimura, H. and S. Lee (2010) “Characterization of the Asymptotic Distribution of Semiparametric M-Estimators”, *Journal of Econometrics,* 58, 71-120.

[160] Ichimura, H. and P. Todd (2007) “Implementing Nonparametric and Semiparametric Estimators”, in J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, vol. 6B. Amsterdam: North-Holland.

[161] Imbens, G., W. Newey and G. Ridder (2005) “Mean-squared-error Calculations for Average Treatment Effects”, manuscript, UC Berkeley.

[162] Karlsen, H. and D. Tjϕstheim (2001) “Nonparametric Estimation in Null Recurrent Time Series”, *The Annals of Statistics*, 29, 372-416.

[163] Kleijn, B. and P. Bickel (2010) “The semiparametric Bernstein-Von Mises theorem,” UC Berkeley, unpublished working paper.

[164] Koenker, R. (2010) "Additive models for quantile regression: model selection and confidence bandaids," UIUC, unpublished working paper.

[165] Koenker, R. and Z. Xiao (2006) "Quantile Autoregression", *Journal of the American Statistical Association*, 101, 980-990.

[166] Kosorok, M. (2008) *Introduction to Empirical Processes and Semiparametric Inference.* New York: Springer.

[167] Krivobokova, T., T. Kneib, and G. Claeskens (2010) "Simultaneous Confidence Bands for Penalized Spline Estimators," *J. of Am. Stat. Assoc.*, forthcoming.

[168] Leeb, H. and B. Potscher (2009) "Model Selection", in T.G. Andersen, R.A. Davis, J.-P. Kreiss and T. Mikosch (eds.), *Handbook of Financial Time Series.* New York: Springer.

[169] Li, K. (1987) "Asymptotic Optimality for C_p, C_L, Cross-validation, and Generalized Cross-validation: Discrete Index Set", *Annals of Statistics* 15, 958-975.

[170] Li, D., Z. Lu and O. Linton (2010) "Local Linear Fitting under Near Epoch Dependence: Uniform Consistency with Convergence Rates", Discussion paper, London School of Economics.

[171] Li, Q. and J. Racine (2007) *Nonparametric Econometrics Theory and Practice.* Princeton: Princeton University Press.

[172] Linton, O. (2009) "Semiparametric and nonparametric ARCH modelling", in T.G. Andersen, R.A. Davis, J.-P. Kreiss and T. Mikosch (eds.), *Handbook of Financial Time Series.* New York: Springer.

[173] Linton, O. and E. Mammen (2005) "Estimating Semiparametric ARCH(∞) Models by Kernel Smoothing Methods", *Econometrica,* 73, 771-836.

[174] Linton, O. and Y. Yan (2011) "Semi- and Nonparametric ARCH Processes", *Journal of Probability and Statistics*, forthcoming.

[175] Lu, Z. and O. Linton (2007) "Local linear fitting under near epoch dependence", *Econometric Theory,* 23, 37-70.

[176] Mammen, E., O. Linton and J. Nielsen (1999) "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions", *The Annals of Statistics*, 27, 1443-1490.

[177] Masry, E. and D. Tjϕstheim (1995) "Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality", *Econometric Theory*, 11, 258-289.

[178] McCaffrey, D., S. Ellner, A.R. Gallant, and D. Nychka (1992) "Estimating the Lyapunov Exponent of a Chaotic System with Nonparametric Regression", *Journal of the American Statistical Association*, 87, 682-695.

[179] Meyn, S.P. and R.L. Tweedie (1993) *Markov chains and Stochastic Stability.* London: Springer-Verlag.

[180] Murphy, K. and R. Topel (1985) "Estimation and inference in two step econometric models", *Journal of Business and Economic Statistics*, 3, 370-9.

[181] Murphy, S. and A. van der Vaart (2000) "On Profile Likelihood", *Journal of the American Statistical Association*, 95, 449-465.

[182] Newey, W.K. (1984) "A Method of Moments Interpretation of Sequential Estimators", *Economics Letters* 14, 201-206.

[183] Newey, W.K. (1990) "Semiparametric Efficiency Bounds", *Journal of Applied Econometrics*, 5, 99-135.

[184] Newey, W.K. (1994) "The Asymptotic Variance of Semiparametric Estimators", *Econometrica,* 62, 1349-1382.

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

[186] Newey, W.K. and D. F. McFadden (1994) "Large sample estimation and hypothesis testing", in R.F. Engle III and D.F. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. Amsterdam: North-Holland.

[187] Newey, W.K. and J.L Powell (2003) "Instrumental Variable Estimation of Nonparametric Models", *Econometrica,* 71, 1565-1578. Working paper version, 1989.

[188] Newey, W.K., J.L. Powell and F. Vella (1999) "Nonparametric Estimation of Triangular Simultaneous Equations Models", *Econometrica,* 67, 565-603.

[189] Pagan, A. and A. Ullah (1999) *Nonparametric Econometrics*, Cambridge University Press.

[190] Pakes, A. and S. Olley (1995) "A Limit Theorem for A Smooth Class of Semiparametric Estimators", *Journal of Econometrics*, 65, 295-332.

[191] Park, J. and P. Phillips (2001) "Nonlinear Regressions with Integrated Time Series", *Econometrica,* 69, 117-161.

[192] Patton, A. (2006) "Modeling Asymmetric Exchange Rate Dependence", *International Economic Review*, 47, 527-56.

[193] Patton, A. (2009) "Copula-Based Models for Financial Time Series", in T.G. Andersen, R.A. Davis, J.-P. Kreiss and T. Mikosch (eds.), *Handbook of Financial Time Series.* Springer Verlag.

[194] Phillips, P.C.B. (1998) "New Tools for Understanding Spurious Regressions", *Econometrica,* 66, 1299-1325.

[195] Phillips, P.C.B. and J. Park (1998) "Nonstationary density estimation and kernel autoregression", Cowles Foundation Discussion Paper, No. 1181, Yale University.

[196] Phillips, P.C.B. and W. Ploberger (2003) "An Introduction to Best Empirical Models when the Parameter Space is Infinite Dimensional", *Oxford Bulletin of Economics and Statistics*, 65, 877-890.

[197] Pollard, D. (1984) *Convergence of Statistical Processes*. Springer-Verlag, New York.

[198] Pötscher, B. M. and I.R. Prucha (1997) *Dynamic Nonlinear Econometric Models: Asymptotic Theory.* Berlin: Springer-verlag.

[199] Rio, E. (2000) *Théorie asymptotique des processes aléatoires faiblement dépendants*. Mathématiques & Applications, 31, Berlin: Springer-Verlag.

[200] Robinson, P. (1988) "Root-N-Consistent Semiparametric Regression", *Econometrica,* 56, 931-954.

[201] Robinson, P. (1994) "Time series with strong dependence", C. Sims (eds.), *Advances in Econometrics*, Sixth World Congress, Vol. 1. Cambridge: Cambridge University Press.

[202] Rosenblatt, M. (1956) "A central limit theorem and a strong mixing condition", *Proc. Natl. Acad. Sci. USA*, 42, 43–47.

[203] Ruppert, D., M. Wand and R. Carroll (2003) *Semiparametric Regression*, Cambridge: Cambridge University Press.

[204] Santos, A. (2010) "Inference in Nonparametric Instrumental Variables with Partial Identification", UCSD unpublished working paper.

[205] Shen, X. (1997) "On Methods of Sieves and Penalization", *The Annals of Statistics,* 25, 2555-2591.

[206] Shen, X. (2002) "Asymptotic normality of semiparametric and nonparametric posterior distributions," *Journal of the American Statistical Association* 97, 222-235.

[207] Shen, X. and J. Shi (2005) "Sieve Likelihood ratio inference on general parameter space", *Science in China*, 48, 67-78.

[208] Shen, X. and J. Ye (2002) "Adaptive Model Selection", *Journal of American Statistical Association* 97, 210-221.

[209] Singleton, K. (2006) *Empirical Dynamic Asset Pricing.* Princeton, New Jersey: Princeton University Press.

[210] Song, K. (2008) "Uniform Convergence of Series Estimators Over Function Spaces", *Econometric Theory,* 24, 1463-1499.

[211] Song, Q. and L. Yang (2009) "Spline confidence bands for variance function", *Journal of Nonparametric Statistics,* 21, 589-609.

[212] Song, Q. and L. Yang (2010) “Oracally efficient spline smoothing of nonlinear additive autoregression model with simultaneous confidence band”, *Journal of Multivariate Analysis,* 101, 2008-2025.

[213] Stock, J., and M. Watson (2002) “Macroeconomic Forecasting Using Diffusion Indexes,” *Journal of Business and Economic Statistics* 20, 147–162.

[214] Stock, J. and J. Wright (2000), “GMM with Weak Identification”, *Econometrica,* 51, 1055-1096.

[215] Stone, C.J. (1982) “Optimal global rates of convergence for nonparametric regression”, *The Annals of Statistics,* 10, 1040-1053.

[216] Stone, C.J. (1985) “Additive regression and other nonparametric models”, *The Annals of Statistics,* 13, 689-705.

[217] Stone, C. J., M.H. Hansen, C. Kooperberg and Y.K. Truong (1997) “Polynomial splines and their tensor products in extended linear modeling”, *The Annals of Statistics*, 25, 1371-1425.

[218] Tauchen, G. (1997) “New Minimum Chi-Square Methods in Empirical Finance”, in D. Kreps and K. Wallis (eds.), *Advances in Econometrics, Seventh World Congress.* Cambridge, UK: Cambridge University Press.

[219] Teräsvirta, T., D. Tjϕstheim and C.W.J. Granger (1994) “Aspects of Modelling Nonlinear Time Series”, in R.F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4. Amsterdam: North-Holland.

[220] Tiao, G.C. and R.S. Tsay (1994) “Some Advances in Non-linear and Adaptive Modelling in Time-series”, *Journal of Forecasting*, 13, 109-131.

[221] Tong, H. (1990) *Non-linear Time Series: A Dynamical System Approach*, Oxford: Oxford University Press.

[222] Tong, H. and K.S. Lim (1980) “Threshold Autoregressions, Limit Cycles and Data”, *Journal of the Royal Statistical Society*, 42, 245-92.

[223] Tsay, R. (2005) *Analysis of Financial Time Series*, 2nd Edition. New York: John Wiley and Sons.

[224] Van de Geer, S. (2000) *Empirical Processes in M-estimation*, Cambridge University Press.

[225] Van de Geer, S. (2008) “High-dimensional generalized linear models and the Lasso,” *The Annals of Statistics,* 36, 614–645.

[226] Van der Vaart, A. (1991) “On Differentiable Functionals”, *The Annals of Statistics,* 19, 178-204.

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

[228] Wahba, G. (1990) *Spline Models for Observational Data,* CBMS-NSF Regional Conference Series, Philadelphia.

[229] Wang, L. and L. Yang (2009a) “Spline Estimation of Single-index Models”, *Statistica Sinica*, 19, 765-783

[230] Wang, J. and L. Yang (2009b) “Polynomial spline confidence bands for regression curves”, *Statistica Sinica*, 19, 325-342.

[231] Wang, L. and L. Yang (2010) “Simultaneous confidence bands for time series prediction function,” *Journal of Nonparametric Statistics* 22, 999-1018.

[232] Wang, Q. and P.C.B Phillips (2009a) “Asymptotic Theory for Local Time Density Estimation and Nonparametric Cointegrating Regression”, *Econometric Theory*, 25(3), 710-738.

[233] Wang, Q. and P.C.B Phillips (2009b) “Structural Nonparametric Cointegrating Regression”, *Econometrica*, 77, 1901-1948.

[234] White, H. (1990) “Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings”, *Neural Networks*, 3, 535-550.

[235] White, H. (1994) *Estimation, Inference and Specification Analysis*, Cambridge University Press.

[236] Wong, W.H. and T. Severini (1991) “On Maximum Likelihood Estimation in Infinite Dimensional Parameter Spaces”, *The Annals of Statistics,* 19, 603-632.

[237] Wooldridge, J. (1994) “Estimation and Inference for Dependent Processes”, in R.F. Engle III and D.F. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. Amsterdam: North-Holland.

[238] Wooldridge, J. and H. White (1988) “Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes,” *Econometric Theory*, 4, 210-230.

[239] Wu, W. (2005) “Nonlinear system theory: Another look at dependence”, *Proc. Natl. Acad. Sci. USA*, 102, 14150-14154.

[240] Wu, W. (2011) “Asymptotic Theory for Stationary Processes”, University of Chicago, working paper.

[241] Yao, J. and J. Attali (2000) “On Stability of Nonlinear AR Processes with Markov Switching”, *Advances in Applied Probability*, 32, 394-407.

[242] Yatchew, A. (2003) *Semiparametric Regression for the Applied Econometrician.* New York: Cambridge University Press.

[243] Yu, B. (1994) “Rates of Convergence for Empirical Processes of Stationary Mixing Sequences,” *The Annals of Probability*, 22, 94–116.

[244] Zhang, M.Y., J.R. Russell, and R.S. Tsay (2001) “A Nonlinear Autoregressive Conditional Duration Model with Applications to Financial Transaction Data”, *Journal of Econometrics*, 104, 179-207.