TIME-VARYING NONLINEAR REGRESSION MODELS: NONPARAMETRIC ESTIMATION AND MODEL SELECTION

By Ting Zhang[1] and Wei Biao Wu[2]

Boston University and University of Chicago

This paper considers a general class of nonparametric time series regression models where the regression function can be time-dependent. We establish an asymptotic theory for estimates of the time-varying regression functions. For this general class of models, an important issue in practice is to address the necessity of modeling the regression function as nonlinear and time-varying. To tackle this, we propose an information criterion and prove its selection consistency property. The results are applied to the U.S. Treasury interest rate data.

1. Introduction. Consider the time-varying regression model

$$\text{Model I:} \quad y_i = m_i(\mathbf{x}_i) + e_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where y_i, $\mathbf{x}_i$ and e_i are the responses, the predictors and the errors, respectively, and $m_i(\cdot) = m(\cdot, i/n)$ is a time-varying regression function. Here $m : \mathbb{R}^d \times [0, 1] \to \mathbb{R}$ is a smooth function, and $i/n, i = 1, \ldots, n$, represents the time rescaled to the unit interval. Model I is very general. If $m_i(\cdot)$ is not time-varying, then (1.1) becomes

$$\text{Model II:} \quad y_i = \mu(\mathbf{x}_i) + e_i, \qquad i = 1, \ldots, n.$$

Model II has been extensively studied in the literature; see Robinson (1983), Györfi et al. (1989), Fan and Yao (2003) and Li and Racine (2007), among others. As an important example, (1.1) can be viewed as the discretized version of the nonstationary diffusion process

$$dy_t = m(y_t, t/T) \, dt + \sigma(y_t, t/T) \, d\mathbb{B}_t, \tag{1.2}$$

where $\{\mathbb{B}_s\}_{s \in \mathbb{R}}$ is a standard Brownian motion, $m(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ are, respectively, the drift and the volatility functions, which can both be time-varying, and T is the time horizon under consideration. If the functions $m(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ do not depend on time, then (1.2) becomes the stationary diffusion process

$$dy_t = \mu(y_t) \, dt + \gamma(y_t) \, d\mathbb{B}_t, \tag{1.3}$$

which relates to model II. There is a huge literature on modeling interest rates data by (1.3). For example, Vasicek (1977) considered model (1.3) with linear drift function $\mu(x) = \beta_0 + \beta_1 x$ and constant volatility $\gamma(x) \equiv \gamma$, where β_0, β_1, γ are unknown parameters. Courtadon (1982), Cox, Ingersoll and Ross (1985) and Chan et al. (1992) considered nonconstant volatility functions. Aït-Sahalia (1996), Stanton (1997) and Liu and Wu (2010) studied model (1.3) with nonlinear drift functions. See Zhao (2008) for a review. However, due to policy and societal changes, those models with static relationship between responses and predictors may not be suitable. Here we shall study estimates of time-varying regression function $m_i(\cdot)$ for model (1.1).

For model II, let $K_S(\cdot)$ be a d-dimensional kernel function

$$(1.4) \quad \tilde{T}_n(\mathbf{u}) = \frac{1}{nh_n^d} \sum_{i=1}^{n} y_i K_S\left(\frac{\mathbf{u} - \mathbf{x}_i}{h_n}\right), \qquad \tilde{f}_n(\mathbf{u}) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K_S\left(\frac{\mathbf{u} - \mathbf{x}_i}{h_n}\right),$$

where h_n be a bandwidth sequence. We can then apply the traditional Nadaraya–Watson estimate for the regression function $\mu(\cdot)$,

$$(1.5) \qquad\qquad \hat{\mu}_n(\mathbf{u}) = \frac{\tilde{T}_n(\mathbf{u})}{\tilde{f}_n(\mathbf{u})}, \qquad \mathbf{u} \in \mathbb{R}^d.$$

If the process $(\mathbf{x}_i)$ is stationary, then $\tilde{f}_n$ is the kernel density estimate of its marginal density. For stationary processes, an asymptotic theory for these nonparametric estimators has been developed by many researchers, including Robinson (1983), Castellana and Leadbetter (1986), Silverman (1986), Györfi et al. (1989), Yu (1993), Tjøstheim (1994), Wand and Jones (1995), Bosq (1996), Neumann (1998), Neumann and Kreiss (1998), Fan and Yao (2003) and Li and Racine (2007), among others. However, the case of nonstationary processes has been rarely touched. Hall, Müller and Wu (2006) considered the situation that the underlying distribution evolves with time and proposed a nonparametric time-dynamic density estimator. Assuming independence, they proved the consistency of their kernel-type estimators and applied the results to fast mode tracking. Following the spirit of Hall, Müller and Wu (2006), Vogt (2012) considered a kernel estimator of the time-varying regression model (1.1), and established its asymptotic normality and uniform bound under the classical strong mixing conditions. In Sections 3.1 and 3.2, we advance the nonparametric estimation theory for the time-varying regression model (1.1) under the framework of Draghicescu, Guillas and Wu (2009), which is convenient to use and often leads to optimal asymptotic results.

Apart from model II, model I contains another important special case: the time-varying coefficient linear regression model

$$\text{Model III:} \quad y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_i + e_i, \qquad i = 1, \dots, n,$$

where $^\top$ is the transpose and $\boldsymbol{\beta}_i = \boldsymbol{\beta}(i/n)$ for some smooth function $\boldsymbol{\beta} : [0, 1] \to \mathbb{R}^d$. The traditional linear regression model

$$\text{Model IV:} \quad y_i = \mathbf{x}_i^\top \boldsymbol{\theta} + e_i, \qquad i = 1, \dots, n,$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the regression coefficient, is a special case of model III. Estimation of $\boldsymbol{\beta}(\cdot)$ has been considered by Hoover et al. (1998), Fan and Zhang (2000a, 2000b), Huang, Wu and Zhou (2004), Ramsay and Silverman (2005), Cai (2007) and Zhou and Wu (2010), among others. The problem of distinguishing between models III and IV has been studied in the literature mainly by means of hypothesis testings; see, for example, Chow (1960), Brown, Durbin and Evans (1975), Nabeya and Tanaka (1988), Leybourne and McCabe (1989), Nyblom (1989), Ploberger, Krämer and Kontrus (1989), Andrews (1993), Davis, Huang and Yao (1995), Lin and Teräsvirta (1999) and He, Teräsvirta and González (2009). On the other hand, model IV specifies a linear relationship upon model II, and there is a huge literature on testing parametric forms of $\mu(\cdot)$; see Azzalini and Bowman (1993), González Manteiga and Cao (1993), Härdle and Mammen (1993), Zheng (1996), Dette (1999), Fan, Zhang and Zhang (2001), Zhang and Dette (2004) and Zhang and Wu (2011), among others. Nevertheless, model selection between models II and III received much less attention. Note that both of them are nested in the general model I, and they all cover the linear regression model IV. It is desirable to develop a model selection criterion. An information criterion is proposed in Section 3.3, where its consistency property is obtained.

The rest of the paper is organized as follows. Section 2 introduces the model setting. Main results are stated in Section 3 and are proved in Section 6 with some of the proofs postponed to the supplementary material [Zhang and Wu (2015)]. A simulation study is given in Section 4 along with an application to the U.S. Treasury interest rate data.

2. Model setting. For estimation of model I, temporal dynamics should be taken into consideration. Let $K_T(\cdot)$ be a temporal kernel function (kernel function for time), b_n be another sequence of bandwidths and $w_{b_n,i}(t) = K_T\{(i/n - t)/b_n\}\{S_2(t) - (t - i/n)S_1(t)\}/\{S_2(t)S_0(t) - S_1^2(t)\}$ be the local linear weights, where $S_l(t) = \sum_{j=1}^{n}(t - j/n)^l K_T\{(j/n - t)/b_n\}$, $l \in \{0, 1, 2\}$. Let $K_{S,h_n}(\cdot) = h_n^{-d} K_S(\cdot/h_n)$,

$$\hat{f}_n(\mathbf{u}, t) = \sum_{i=1}^{n} K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) w_{b_n,i}(t),$$

(2.1)

$$\hat{T}_n(\mathbf{u}, t) = \sum_{i=1}^{n} y_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) w_{b_n,i}(t),$$

we consider the time-varying kernel regression estimator

(2.2)
$$\hat{m}_n(\mathbf{u}, t) = \frac{\hat{T}_n(\mathbf{u}, t)}{\hat{f}_n(\mathbf{u}, t)}.$$

Hall, Müller and Wu (2006) proved the uniform consistency of $\hat{f}_n$ in (2.1) by assuming that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent. To allow nonstationary and dependent observations, we assume

$$(2.3) \qquad \mathbf{x}_i = \mathbf{G}(i/n; \mathcal{H}_i), \qquad \text{where } \mathcal{H}_i = (\ldots, \boldsymbol{\xi}_{i-1}, \boldsymbol{\xi}_i)$$

and $\boldsymbol{\xi}_k$, $k \in \mathbb{Z}$, are independent and identically distributed (i.i.d.) random vectors, and $\mathbf{G}$ is a measurable function such that $\mathbf{G}(t; \mathcal{H}_i)$ is well defined for each $t \in [0, 1]$. Following Draghicescu, Guillas and Wu (2009), the framework (2.3) suggests locally strict stationarity and is convenient for asymptotic study. For the error process, we assume that

$$(2.4) \qquad e_i = \sigma_i(\mathbf{x}_i)\eta_i = \sigma(\mathbf{x}_i, i/n)\eta_i,$$

where $\sigma(\cdot, \cdot) : \mathbb{R}^d \times [0, 1] \to \mathbb{R}$ is a smooth function, and (η_i) is a sequence of random variables satisfying $E(\eta_i | \mathbf{x}_i) = 0$ and $E(\eta_i^2 | \mathbf{x}_i) = 1$. At the outset (cf. Sections 3.1–3.3) we assume that η_k, $k \in \mathbb{Z}$, are i.i.d. and independent of $\mathcal{H}_j$, $j \in \mathbb{Z}$. The latter assumption can be relaxed (though technically much more tedious) to allow models with correlated errors and nonlinear autoregressive processes; see Section 3.4.

For a random vector $\mathbf{Z}$, we write $\mathbf{Z} \in \mathcal{L}^q$, $q > 0$ if $\|\mathbf{Z}\| = \{E(|\mathbf{Z}|^q)\}^{1/q} < \infty$ where $|\cdot|$ is the Euclidean vector norm, and we denote $\|\cdot\| = \|\cdot\|_2$. Let $F_1(\mathbf{u}, t | \mathcal{H}_k) = \mathrm{pr}\{\mathbf{G}(t; \mathcal{H}_{k+1}) \leq \mathbf{u} | \mathcal{H}_k\}$ be the one-step ahead predictive or conditional distribution function and $f_1(\mathbf{u}, t | \mathcal{H}_k) = \partial^d F_1(\mathbf{u}, t | \mathcal{H}_k)/\partial \mathbf{u}$ be the corresponding conditional density. Let $(\boldsymbol{\xi}'_i)$ be an i.i.d. copy of $(\boldsymbol{\xi}_j)$ and $\mathcal{H}'_k = (\ldots, \boldsymbol{\xi}_{-1}, \boldsymbol{\xi}'_0, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k)$ be the coupled shift process. We define the predictive dependence measure

$$(2.5) \qquad \psi_{k,q} = \sup_{t \in [0,1]} \sup_{\mathbf{u} \in \mathbb{R}^d} \left\| f_1(\mathbf{u}, t | \mathcal{H}_k) - f_1(\mathbf{u}, t | \mathcal{H}'_k) \right\|_q.$$

Quantity (2.5) measures the contribution of $\boldsymbol{\xi}_0$, the innovation at step 0, on the conditional or predictive distribution at step k. We shall make the following assumptions:

(A1) smoothness (third order continuous differentiability): $f, m, \sigma \in \mathcal{C}^3(\mathbb{R}^d \times [0, 1])$;

(A2) short-range dependence: $\Psi_{0,2} < \infty$, where $\Psi_{m,q} = \sum_{k=m}^{\infty} \psi_{k,q}$;

(A3) there exists a constant $c_0 < \infty$ such that almost surely,

$$\sup_{t \in [0,1]} \sup_{\mathbf{u} \in \mathbb{R}^d} \left\{ f_1(\mathbf{u}, t | \mathcal{H}_0) + \left| \partial^d f_1(\mathbf{u}, t | \mathcal{H}_0)/\partial \mathbf{u} \right| \right\} \leq c_0.$$

Condition (A3) implies that the marginal density $f(\mathbf{u}, t) = E\{f_1(\mathbf{u}, t | \mathcal{H}_0)\} \leq c_0$.

3. Main results.

3.1. *Nonparametric kernel estimation.* Throughout the paper, we assume that the kernel functions $K_S(\cdot)$ and $K_T(\cdot)$ are both symmetric and twice continuously differentiable on their support $[-1, 1]^d$ and $[-1, 1]$, respectively, and $\int_{[-1,1]^d} K_S(\mathbf{s})\, d\mathbf{s} = \int_{-1}^1 K_T(v)\, dv = 1$. Denote by "$\Rightarrow$" convergence in distribution. Theorem 3.1 provides the asymptotic normality of the time-varying kernel estimators (2.1) and (2.2), while Theorem 3.2 concerns the time-constant estimators (1.4) and (1.5).

THEOREM 3.1. *Assume* (A1)–(A3) *and* $\eta_i \in \mathcal{L}^p$, $p > 2$ *are i.i.d. Let* $(\mathbf{u}, t) \in \mathbb{R}^d \times (0, 1)$ *be a fixed point. If* $b_n \to 0$, $h_n \to 0$ *and* $nb_n h_n^d \to \infty$, *then*

$$(3.1) \qquad (nb_n h_n^d)^{1/2}\big[\hat{f}_n(\mathbf{u}, t) - E\{\hat{f}_n(\mathbf{u}, t)\}\big] \Rightarrow N\{0, f(\mathbf{u}, t)\lambda_{K_S}\lambda_{K_T}\},$$

where $\lambda_{K_T} = \int_{-1}^1 K_T(v)^2\, dv$ *and* $\lambda_{K_S} = \int_{[-1,1]^d} K_S(\mathbf{s})^2\, d\mathbf{s}$. *If in addition* $f(\mathbf{u}, t) > 0$, *then*

$$(3.2) \qquad (nb_n h_n^d)^{1/2}\bigg[\hat{m}_n(\mathbf{u}, t) - \frac{E\{\hat{T}_n(\mathbf{u}, t)\}}{E\{\hat{f}_n(\mathbf{u}, t)\}}\bigg] \Rightarrow N\bigg\{0, \frac{\sigma(\mathbf{u}, t)^2 \lambda_{K_S}\lambda_{K_T}}{f(\mathbf{u}, t)}\bigg\}.$$

Let $H_f(\mathbf{u}, t) = \{\partial^2 f(\mathbf{u}, t)/\partial u_i\, \partial u_j\}_{1 \le i, j \le d}$ be the Hessian matrix of the density function f with respect to $\mathbf{u}$. Denote $f^{(\mathbf{0},2)}(\mathbf{u}, t) = \partial^2 f(\mathbf{u}, t)/\partial t^2$, and we use the same notation for the product function $(mf)(\mathbf{u}, t) = m(\mathbf{u}, t)f(\mathbf{u}, t)$. Then for any point $(\mathbf{u}, t) \in \mathbb{R}^d \times (0, 1)$ with $f(\mathbf{u}, t) > 0$, we have

$$E\{\hat{f}_n(\mathbf{u}, t)\} = f(\mathbf{u}, t) + \frac{h_n^2}{2}\mathrm{tr}\{H_f(\mathbf{u}, t)\kappa_S\} + \frac{b_n^2}{2}f^{(\mathbf{0},2)}(\mathbf{u}, t)\kappa_T + O(b_n^3 + h_n^3),$$

where $\mathrm{tr}(\cdot)$ is the trace operator

$$\kappa_S = \int_{[-1,1]^d} K_S(\mathbf{s})\mathbf{s}\mathbf{s}^\top\, d\mathbf{s}, \qquad \kappa_T = \int_{-1}^1 K_T(v)v^2\, dv$$

and

$$\frac{E\{\hat{T}_n(\mathbf{u}, t)\}}{E\{\hat{f}_n(\mathbf{u}, t)\}} = m(\mathbf{u}, t) + \frac{h_n^2}{2f(\mathbf{u}, t)}\mathrm{tr}\big[\{H_{mf}(\mathbf{u}, t) - m(\mathbf{u}, t)H_f(\mathbf{u}, t)\}\kappa_S\big]$$

$$+ \frac{b_n^2}{2f(\mathbf{u}, t)}\{(mf)^{(\mathbf{0},2)}(\mathbf{u}, t) - m(\mathbf{u}, t)f^{(\mathbf{0},2)}(\mathbf{u}, t)\}\kappa_T$$

$$+ O(b_n^3 + h_n^3).$$

Hence (2.1) and (2.2) are consistent estimates of the local density function f and the regression function m, respectively. The asymptotic mean squared error (AMSE) optimal bandwidths satisfy $b_n \asymp n^{-1/(d+5)}$ and $h_n \asymp n^{-1/(d+5)}$. Here for positive sequences (s_n) and (r_n), we write $s_n \asymp r_n$ if $s_n/r_n + r_n/s_n$ is bounded for all large n.

THEOREM 3.2. *Assume* (A1)–(A3) *and* $\eta_i \in \mathcal{L}^p$, $p > 2$. *If* $h_n \to 0$ *and* $nh_n^d \to \infty$, *then*

$$(3.3) \qquad (nh_n^d)^{1/2}[\tilde{f}_n(\mathbf{u}) - E\{\tilde{f}_n(\mathbf{u})\}] \Rightarrow N\{0, \bar{f}(\mathbf{u})\lambda_{K_S}\}, \qquad \mathbf{u} \in \mathbb{R}^d,$$

where $\bar{f}(\mathbf{u}) = \int_0^1 f(\mathbf{u}, t) \, dt$. *If in addition* $\bar{f}(\mathbf{u}) > 0$, *then*

$$(3.4) \qquad (nh_n^d)^{1/2}\left[\hat{\mu}_n(\mathbf{u}) - \frac{E\{\tilde{T}_n(\mathbf{u})\}}{E\{\tilde{f}_n(\mathbf{u})\}}\right] \Rightarrow N\{0, \tilde{V}(\mathbf{u})\lambda_{K_S}\},$$

where, letting $\bar{m}(\mathbf{u}) = \int_0^1 m(\mathbf{u}, t) f(\mathbf{u}, t) \, dt / \bar{f}(\mathbf{u})$, *the variance function*

$$\tilde{V}(\mathbf{u}) = \bar{f}(\mathbf{u})^{-2} \int_0^1 \left[\{m(\mathbf{u}, t) - \bar{m}(\mathbf{u})\}^2 + \sigma(\mathbf{u}, t)^2\right] f(\mathbf{u}, t) \, dt.$$

For any point $\mathbf{u} \in \mathbb{R}^d$ with $\bar{f}(\mathbf{u}) > 0$, we have

$$E\{\tilde{f}_n(\mathbf{u})\} = \bar{f}(\mathbf{u}) + \frac{h_n^2}{2}\text{tr}\left\{\int_0^1 H_f(\mathbf{u}, t)\kappa_S \, dt\right\} + O(h_n^3)$$

and

$$\frac{E\{\tilde{T}_n(\mathbf{u})\}}{E\{\tilde{f}_n(\mathbf{u})\}} = \bar{m}(\mathbf{u}) + \frac{h_n^2}{2\bar{f}(\mathbf{u})}\text{tr}\left[\int_0^1 \{H_{mf}(\mathbf{u}, t) - m(\mathbf{u}, t)H_f(\mathbf{u}, t)\}\kappa_S \, dt\right]$$
$$+ O(h_n^3).$$

Therefore, (1.4) and (1.5) provide consistent estimators of $\bar{f}$ and $\bar{m}$, (weighted) temporal averages of the local density function f and the regression function m, respectively. For stationary processes, Theorem 3.2 relates to traditional results on nonparametric kernel estimators; see, for example, Robinson (1983), Bosq (1996) and Wu (2005). The AMSE optimal bandwidth for the time-constant kernel estimators (1.4) and (1.5) satisfies $h_n \asymp n^{-1/(d+4)}$.

3.2. *Uniform bounds.* For stationary or independent observations, uniform bounds for kernel estimators have been obtained by Peligrad (1992), Andrews (1995), Bosq (1996), Masry (1996), Fan and Yao (2003) and Hansen (2008), among others. Hall, Müller and Wu (2006) obtained a uniform bound for time-varying kernel density estimators for independent observations, while Vogt (2012) considered kernel regression estimators under strong mixing conditions. We shall here establish uniform bounds for the time-varying kernel estimators (2.1) and (2.2) under the locally strict stationarity framework (2.3). We need the following assumptions:

(A4) there exists a $q > 2$ such that $\Psi_{0,q} < \infty$ and $\Psi_{m,q} = O(m^{-\alpha})$ for some $\alpha > 1/2 - 1/q$;

(A5) let $\mathscr{X} \subseteq \mathbb{R}^d$ be a compact set, and assume $\inf_{t \in [0,1]} \inf_{\mathbf{u} \in \mathscr{X}} f(\mathbf{u}, t) > 0$.

THEOREM 3.3. *Assume* (A1), (A3)–(A5), $b_n \to 0$, $h_n \to 0$ *and* $nb_nh_n^d \to \infty$. (i) *If there exists* $r > r' > 0$ *such that* $\sup_{t\in[0,1]} \|\mathbf{G}(t; \mathcal{H}_0)\|_r < \infty$ *and* $n^{2/r'+2+d-q} b_n^{d-q} h_n^{d(d+q)} \to 0$, *then*

$$\sup_{t\in[0,1]} \sup_{\mathbf{u}\in\mathbb{R}^d} |\hat{f}_n(\mathbf{u}, t) - E\{\hat{f}_n(\mathbf{u}, t)\}| = O_p\left\{\frac{(\log n)^{1/2}}{(nb_nh_n^d)^{1/2}}\right\}.$$

(ii) *If* $\eta_i \in \mathcal{L}^p$ *for some* $p > 2$, *and* $n^{2+d-q} b_n^{d-q} h_n^{d(d+q)} \to 0$, *then*

$$\sup_{t\in[0,1]} \sup_{\mathbf{u}\in\mathcal{X}} \left|\hat{m}_n(\mathbf{u}, t) - \frac{E\{\hat{T}_n(\mathbf{u}, t)\}}{E\{\hat{f}_n(\mathbf{u}, t)\}}\right| = O_p\left\{\frac{(\log n)^{1/2}}{(nb_nh_n^d)^{1/2}} + \frac{n^{1/p}\log n}{nb_nh_n^d}\right\}.$$

If the bandwidths $b_n \asymp n^{-1/(d+5)}$ and $h_n \asymp n^{-1/(d+5)}$ have the optimal AMSE rate, and $\eta_i \in \mathcal{L}^p$ for some $p > (d+5)/2$, then the bound in Theorem 3.3(ii) can be simplified to $O_p\{(nb_nh_n^d)^{-1/2}(\log n)^{1/2}\}$. Theorem 3.4 provides a uniform bound for (1.4) and (1.5).

THEOREM 3.4. *Assume* (A1), (A3)–(A5), $h_n \to 0$ *and* $nh_n^d \to \infty$. (i) *If there exists* $r > r' > 0$ *such that* $\sup_{t\in[0,1]} \|\mathbf{G}(t; \mathcal{H}_0)\|_r < \infty$ *and* $n^{2/r'+2+d-q} h_n^{d(d+q)} \to 0$, *then*

$$\sup_{\mathbf{u}\in\mathbb{R}^d} |\tilde{f}_n(\mathbf{u}) - E\{\tilde{f}_n(\mathbf{u})\}| = O_p\left\{\frac{(\log n)^{1/2}}{(nh_n^d)^{1/2}}\right\}.$$

(ii) *If* $\eta_i \in \mathcal{L}^p$ *for some* $p > 2$, *and* $n^{2+d-q} h_n^{d(d+q)} \to 0$, *then*

$$\sup_{\mathbf{u}\in\mathcal{X}} \left|\hat{\mu}_n(\mathbf{u}) - \frac{E\{\tilde{T}_n(\mathbf{u})\}}{E\{\tilde{f}_n(\mathbf{u})\}}\right| = O_p\left\{\frac{(\log n)^{1/2}}{(nh_n^d)^{1/2}} + \frac{n^{1/p}\log n}{nh_n^d}\right\}.$$

If the bandwidth $h_n \asymp n^{-1/(d+4)}$ is AMSE-optimal, and $\eta_i \in \mathcal{L}^p$ for some $p > (d+4)/2$, then the bound in Theorem 3.4(ii) can be simplified to $O_p\{(nh_n^d)^{-1/2} \times (\log n)^{1/2}\}$.

3.3. *Model selection.* Model I is quite general in the sense that it does not impose any specific parametric form on the regression function and allows it to change over time. However, in practice it is useful to check whether model I can be reduced to its simpler special cases, namely models II–IV. Model selection between models II and IV, or between models III and IV, has been studied in the literature mainly by means of hypothesis testing; see references in Section 1. Nevertheless, less attention has been paid to distinguishing between models II and III. We shall here propose an information criterion that can consistently select the underlying true model among candidate models I–IV. Let $\mathcal{T} \subset (0, 1)$ be a compact

set and $\mathscr{I}_n = \{i = 1, \ldots, n | i/n \in \mathscr{T}\}$. We consider the restricted residual sum of squares for model I, which takes the form

$$\text{RSS}_n(\mathscr{X}, \mathscr{T}, \text{I}) = \sum_{i \in \mathscr{I}_n} \{y_i - \hat{m}_n(\mathbf{x}_i, i/n)\}^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}},$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. Similarly, we can define $\text{RSS}_n(\mathscr{X}, \mathscr{T}, \text{II})$, $\text{RSS}_n(\mathscr{X}, \mathscr{T}, \text{III})$ and $\text{RSS}_n(\mathscr{X}, \mathscr{T}, \text{IV})$ for models II–IV, respectively. For the simple linear regression model IV, the parameter $\boldsymbol{\theta}$ can be estimated by the least squares estimate

$$(3.5) \qquad \hat{\boldsymbol{\theta}}_n = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i y_i\right).$$

For the time-varying coefficient model III, let $K_{T,b_n}(\cdot) = b_n^{-1} K_T(\cdot/b_n)$, and we can use the kernel estimator of Priestley and Chao (1972),

$$(3.6) \quad \hat{\boldsymbol{\beta}}_n(t) = \left\{\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top K_{T,b_n}(i/n - t)\right\}^{-1} \left\{\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i y_i K_{T,b_n}(i/n - t)\right\}.$$

For a candidate model $\varrho \in \{\text{I}, \text{II}, \text{III}, \text{IV}\}$, we define the generalized information criterion

$$(3.7) \qquad \text{GIC}_{\mathscr{X}, \mathscr{T}}(\varrho) = \log\{\text{RSS}_n(\mathscr{X}, \mathscr{T}, \varrho)/n\} + \tau_n \text{DF}(\varrho),$$

where τ_n is a tuning parameter indicating the amount of penalization and $\text{DF}(\varrho)$ represents the model complexity for model $\varrho \in \{\text{I}, \text{II}, \text{III}, \text{IV}\}$ determined as follows. For the simple linear regression model IV, following the convention we set the model complexity or degree of freedom to be the number of potential predictors, namely $\text{DF}(\text{IV}) = d$. For the time-varying coefficient model III, the effective number of parameters used in kernel smoothing is b_n^{-1} for each one of the d predictors [see, e.g., Hurvich, Simonoff and Tsai (1998)], and thus we set $\text{DF}(\text{III}) = b_n^{-1} d$. Let IQR_k, $k = 1, \ldots, d$, be the componentwise interquartile ranges of $(\mathbf{x}_i)$, and motivated by the same spirit as in Hurvich, Simonoff and Tsai (1998), we set $\text{DF}(\text{II}) = (h_n^d)^{-1} \prod_{k=1}^{d}(2\text{IQR}_k)$ and $\text{DF}(\text{I}) = (b_n h_n^d)^{-1} \prod_{k=1}^{d}(2\text{IQR}_k)$, where $2\text{IQR} = 1$ for random variables having a uniform distribution on $[0, 1]$. The final model is selected by minimizing the information criterion (3.7). We shall make the following assumption:

(A6) eigenvalues of $\mathbf{M}(\mathbf{G}, t) = E\{\mathbf{G}(t; \mathcal{H}_0)\mathbf{G}(t; \mathcal{H}_0)^\top\}$ are bounded away from zero and infinity on $[0, 1]$.

In order to establish the selection consistency of (3.7), in addition to the results developed in Sections 3.1 and 3.2 regarding models I and II, we need the following conditions on estimators (3.5) and (3.6) for models IV and III, respectively:

(P1) There exists a nonrandom sequence $\boldsymbol{\theta}_n$ such that $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n = O_p(n^{-1/2})$. If model IV is correctly specified, then $\boldsymbol{\theta}_n$ can be replaced by the true value $\boldsymbol{\theta}_0$.

(P2) There exists a sequence of nonrandom functions $\boldsymbol{\beta}_n : [0, 1] \to \mathbb{R}^d$ such that

$$\sup_{t \in \mathscr{T}} |\hat{\boldsymbol{\beta}}_n(t) - \boldsymbol{\beta}_n(t)| = O_p(\phi_n),$$

where $\phi_n = (nb_n)^{-1/2}(\log n)^{1/2} + b_n^2$. If model III is correctly specified, then $\boldsymbol{\beta}_n(\cdot)$ can be replaced by the true coefficient function $\boldsymbol{\beta}_0(\cdot)$ and

$$\sup_{t \in \mathscr{T}} \left| \mathbf{M}(\mathbf{G}, t) \left\{ \hat{\boldsymbol{\beta}}_n(t) - \boldsymbol{\beta}_0(t) - \frac{\kappa_T b_n^2 \boldsymbol{\beta}_0''(t)}{2} \right\} - \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i e_i K_{T,b_n}(i/n - t) \right| = O_p(\phi_n^2),$$

where $\mathbf{x}_i e_i \in \mathcal{L}^2$, $i = 1, \ldots, n$.

REMARK 3.1. Conditions (P1) and (P2) can be verified for locally stationary processes with short-range dependence. For example, for the linear regression model IV, by Lemma 5.1 of Zhang and Wu (2012), we have $\sum_{i=1}^n \{\mathbf{x}_i \mathbf{x}_i^\top - E(\mathbf{x}_i \mathbf{x}_i^\top)\} = O_p(n^{1/2})$ and $\sum_{i=1}^n \{\mathbf{x}_i y_i - E(\mathbf{x}_i y_i)\} = O_p(n^{1/2})$. Hence we can use

$$\boldsymbol{\theta}_n = \left\{ \frac{1}{n} \sum_{i=1}^n E(\mathbf{x}_i \mathbf{x}_i^\top) \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n E(\mathbf{x}_i y_i) \right\},$$

which equals to $\boldsymbol{\theta}_0$ if $y_i = \mathbf{x}_i^\top \boldsymbol{\theta}_0 + e_i$, $i = 1, \ldots, n$. This verifies condition (P1). For the time-varying coefficient model III, by Lemma 5.3 of Zhang and Wu (2012), we have $\sup_{t \in \mathscr{T}} |n^{-1} \sum_{i=1}^n \{\mathbf{x}_i \mathbf{x}_i^\top - E(\mathbf{x}_i \mathbf{x}_i^\top)\} K_{T,b_n}(i/n - t)| = O_p(\phi_n)$ and $\sup_{t \in \mathscr{T}} |n^{-1} \sum_{i=1}^n \{\mathbf{x}_i y_i - E(\mathbf{x}_i y_i)\} K_{T,b_n}(i/n - t)| = O_p(\phi_n)$. Hence we can use

$$\boldsymbol{\beta}_n(t) = \left\{ \frac{1}{n} \sum_{i=1}^n E(\mathbf{x}_i \mathbf{x}_i^\top) K_{T,b_n}(i/n - t) \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n E(\mathbf{x}_i y_i) K_{T,b_n}(i/n - t) \right\},$$

and condition (P2) follows by the proof of Theorem 3 in Zhou and Wu (2010).

Recall that the AMSE optimal bandwidths satisfy $b_n(\mathrm{I}) \asymp n^{-1/(d+5)}$ and $h_n(\mathrm{I}) \asymp n^{-1/(d+5)}$ for model I, $h_n(\mathrm{II}) \asymp n^{-1/(d+4)}$ for model II and $b_n(\mathrm{III}) \asymp n^{-1/5}$ for model III. Theorem 3.5 provides the selection consistency of the information criterion (3.7), where the true model is denoted by ϱ_0.

THEOREM 3.5. *Assume* (A1), (A3)–(A6) *with* $q > (3d + 5)/(d + 2)$, (P1), (P2), $\eta_i \in \mathcal{L}^p$ *for some* $p > (d + 5)/2$, $i = 1, \ldots, n$, *and bandwidths with optimal AMSE rates are used for models* I–III. *If*

$$\tau_n n^{(d+1)/(d+5)} \to 0, \qquad \tau_n n^{(d+3)/(d+4)} \to \infty,$$

then for any $\varrho_1 \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}\}$ *and* $\varrho_1 \neq \varrho_0$, *we have*

$$\mathrm{pr}\{\mathrm{GIC}_{\mathscr{X}, \mathscr{T}}(\varrho_0) < \mathrm{GIC}_{\mathscr{X}, \mathscr{T}}(\varrho_1)\} \to 1.$$

3.4. *Extensions.*   Recall that in Theorems 3.1–3.5 error process (2.4) has i.i.d. η_i, which are also independent of $(\mathbf{x}_j)$. In Section 3.4.1 we allow serially correlated η_i. Section 3.4.2 concerns time-varying autoregressive processes in which (η_i) and $(\mathbf{x}_j)$ are naturally dependent.

3.4.1. *Models with serially correlated errors.*   To allow errors with serial correlation, similarly to (2.3) we assume that

$$\eta_i = L(i/n; \mathcal{J}_i), \tag{3.8}$$

where $\mathcal{J}_i = (\ldots, \zeta_{i-1}, \zeta_i)$ with ζ_k, $k \in \mathbb{Z}$, being i.i.d. random variables and independent of $\boldsymbol{\xi}_j$, $j \in \mathbb{Z}$. Therefore, (η_i) is a dependent nonstationary process that is independent of $(\mathbf{x}_j)$, and the error process $e_i = \sigma(\mathbf{x}_i, i/n)\eta_i$ can exhibit both serial correlation and heteroscedasticity; see Robinson (1983), Orbe, Ferreira and Rodriguez-Poo (2005, 2006) and references therein for similar error structures. Let ζ_i', ζ_j, $i, j \in \mathbb{Z}$, be i.i.d. and $\mathcal{J}_k' = (\ldots, \zeta_{-1}, \zeta_0', \zeta_1, \ldots, \zeta_k)$. Assume $c_{L,q} = \sup_{t \in [0,1]} \|L(t; \mathcal{J}_0)\|_q < \infty$, and define the functional dependence measure

$$\nu_{k,q} = \sup_{t \in [0,1]} \|L(t; \mathcal{J}_k) - L(t; \mathcal{J}_k')\|_q.$$

The following theorem states that the results presented in Sections 3.1–3.3 will continue to hold (except for a difference of $\log n$ on the uniform bounds) if the process (η_i) in (3.8) satisfies the geometric moment contraction (GMC) condition [Shao and Wu (2007)]. The proof is available in the supplementary material [Zhang and Wu (2015)].

THEOREM 3.6.   *Assume that the process (η_i) in (3.8) satisfies $\nu_{k,4} = O(\rho^k)$ for some $0 < \rho < 1$. Then the results of Theorems 3.1–3.5 will continue to hold except that the uniform bounds in Theorems 3.3(ii) and 3.4(ii) will be multiplied by a factor of $\log n$.*

3.4.2. *Time-varying nonlinear autoregressive models.*   In this section we shall consider the autoregressive version of (1.1),

$$y_i = m(\mathbf{x}_i, i/n) + \sigma(\mathbf{x}_i, i/n)\eta_i,$$

$$\mathbf{x}_i = (y_{i-1}, \ldots, y_{i-d})^\top, i = 1, \ldots, n, \tag{3.9}$$

where η_i are i.i.d. random variables with $E(\eta_i) = 0$ and $E(\eta_i^2) = 1$. We can view (3.9) as a time-varying or locally stationary autoregressive process, and the corresponding shift processes $\mathcal{F}_k = (\ldots, \eta_{k-1}, \eta_k)$ and $\mathcal{H}_k = \mathcal{F}_{k-1}$. We shall here present analogous versions of Theorems 3.1–3.5. Note that in this case $\mathbf{x}_i$ cannot be written in the form of (2.3). However, Proposition 3.1 implies that it can be well

approximated by a process in the form of (2.3). For each $t \in [0, 1]$, we define the process $\{y_i(t)\}_{i \in \mathbb{Z}}$ by

(3.10)
$$y_i(t) = m\{\mathbf{x}_i(t), t\} + \sigma\{\mathbf{x}_i(t), t\}\eta_i,$$

$$\mathbf{x}_i(t) = \{y_{i-1}(t), \dots, y_{i-d}(t)\}^\top.$$

LEMMA 3.1. *Assume that there exist constants* $a_1, \dots, a_d \geqq 0$ *with* $\sum_{j=1}^d a_j < 1$, *such that, for all* $\mathbf{x} = (x_1, \dots, x_d)^\top$ *and* $\mathbf{x}' = (x_1', \dots, x_d')^\top$,

$$\sup_{0 \leq t \leq 1} \left\| [m(\mathbf{x}, t) + \sigma(\mathbf{x}, t)\eta_i] - [m(\mathbf{x}', t) + \sigma(\mathbf{x}', t)\eta_i] \right\|_p$$

(3.11)

$$\leq \sum_{j=1}^d a_j |x_j - x_j'|.$$

Then (i) *the recursion* (3.10) *has a stationary solution of the form* $y_i(t) = g(t; \mathcal{F}_i)$ *which satisfies the geometric moment contraction* (GMC) *property: for some* $\rho \in (0, 1)$,

$$\sup_{0 \leq t \leq 1} \delta_i(t) = O(\rho^i), \qquad \delta_i(t) = \left\| g(t; \mathcal{F}_i) - g(t; \mathcal{F}_i') \right\|_p.$$

(ii) *If in* (3.9) *the initial values* $(y_0, y_{-1}, \dots, y_{1-d}) = \mathbf{x}_1(0)$, *then* y_i *can be written in the form* $g_i(\mathcal{F}_i)$, *where* $g_i(\cdot)$ *is a measurable function, and it also satisfies the GMC property*

(3.12) $$\sup_{i \leq n} \left\| y_i - g_i(\dots, \eta_{i-k-2}, \eta_{i-k-1}, \eta_{i-k}', \eta_{i-k+1}, \dots, \eta_i) \right\|_p = O(\rho^k).$$

Lemma 3.1(i) concerns the stationarity of the process $\{y_i(t)\}_{i \in \mathbb{Z}}$, which follows from Theorem 5.1 of Shao and Wu (2007). For (ii), denote by $\theta_k^\dagger$ the left-hand side of (3.12). Then by (3.11), $\theta_k^\dagger$ satisfies $\theta_k^\dagger \leq \sum_{j=1}^d a_j \theta_{k-j}^\dagger$, implying (3.12) via recursion.

For presentational simplicity suppose we observe $y_{1-d}, y_{2-d}, \dots, y_n$ from model (3.9) with the initial values $(y_0, y_{-1}, \dots, y_{1-d}) = \mathbf{x}_1(0)$. Estimates (2.1) and (2.2) can be computed in the same way. Proposition 3.1 implies that, for i such that $i/n \approx u$, the process $(\mathbf{x}_i)_i$ can be approximated by the stationary process $\{\mathbf{x}_i(u)\}_i$, thus suggesting local strictly stationarity. The proof is available in the supplementary material [Zhang and Wu (2015)].

PROPOSITION 3.1. *Let* $G_\eta(\mathbf{x}, t) = m(\mathbf{x}, t) + \sigma(\mathbf{x}, t)\eta$ *and* $\dot{G}_\eta(\mathbf{x}, t) = \partial G_\eta(\mathbf{x}, t)/\partial t$. *Assume* (3.11) *and*

$$\sup_{0 \leq t \leq 1} \sup_{0 \leq u \leq 1} \left\| \dot{G}_{\eta_i}\{\mathbf{x}_i(u), t\} \right\|_p < \infty.$$

Then $\|\mathbf{x}_i - \mathbf{x}_i(u)\|_p = O(n^{-1} + |u - i/n|)$.

Let $f(\mathbf{u}, t)$ be the density of $\mathbf{x}_i(t) = \{y_{i-1}(t), \ldots, y_{i-d}(t)\}$ and f_η be the density of η_i. Theorem 3.7 serves as an analogous version of Theorems 3.1–3.4, and the proof is available in the supplementary material [Zhang and Wu (2015)].

THEOREM 3.7. *Assume* (A1), (A5) *and* $\sup_w\{f_\eta(w) + |f'_\eta(w)|\} < \infty$. *Let the conditions in Lemma* 3.1 *and Proposition* 3.1 *be satisfied. Then under respective conditions in Theorems* 3.1–3.5, *the corresponding conclusions also hold, respectively.*

4. Numerical implementation.

4.1. *Bandwidth and tuning parameter selection.* Selecting bandwidths that optimize the performance of (3.7) can be quite nontrivial, and in our case, it is further complicated by the presence of dependence and nonstationarity. Assuming independence, the problem of bandwidth selection has been considered for model II by Härdle and Marron (1985), Härdle, Hall and Marron (1988), Park and Marron (1990), Ruppert, Sheather and Wand (1995), Wand and Jones (1995), Xia (1998) and Gao and Gijbels (2008), among others. Hoover et al. (1998), Fan and Zhang (2000a) and Ramsay and Silverman (2005) considered the problem for model III for longitudinal data, where multiple independent realizations are available. For the time-varying kernel density estimator (2.1) with independent observations, Hall, Müller and Wu (2006) coupled the selection of spatial and temporal bandwidths and adopted the least squares cross validation [Silverman (1986)]. Nevertheless, bandwidths selectors derived under independence can break down for dependent data [Wang (1998) and Opsomer, Wang and Yang (2001)]. We propose using the AMSE optimal bandwidths $b_n(\mathrm{I}) = c_b(\mathrm{I})n^{-1/(d+5)}$ and $h_n(\mathrm{I}) = c_h(\mathrm{I})n^{-1/(d+5)}$ for model I, $h_n(\mathrm{II}) = c_h(\mathrm{II})n^{-1/(d+4)}$ for model II and $b_n(\mathrm{III}) = c_b(\mathrm{III})n^{-1/5}$ for model III, where $0 < c_b(\mathrm{I}), c_h(\mathrm{I}), c_h(\mathrm{II}), c_b(\mathrm{III}) < \infty$ are constants. Due to the presence of both dependence and nonstationarity, estimation of these constants is difficult. Throughout this section, as a rule of thumb, we use $c_b(\mathrm{I}) = c_b(\mathrm{III}) = 1/2$ and $c_h(\mathrm{I}) = c_h(\mathrm{II}) = \prod_{k=1}^d \mathrm{IQR}_k$. Our numerical examples suggest that these simple choices have a reasonably good performance.

We shall here discuss the choice of the tuning parameter τ_n that controls the amount of penalization on models complexities. The problem has been extensively studied for the linear model IV by Akaike (1973), Mallows (1973), Schwarz (1978), Shao (1997) and Yang (2005) among others. For the generalized information criterion (3.7), given conditions in Theorem 3.5, one can choose $\tau_n = cn^{-(d+3)/(d+4)}\log n$, where $c > 0$ is a constant, which satisfies all the required conditions and thus guarantees the selection consistency. Note that the choice of c does not affect the asymptotic result, namely the proposed method will select the true model for any given $c > 0$ as long as the sample size is large enough; see Theorem 3.5. Therefore, one can simply use $c = 1$ to devise a consistent model selection procedure. As an alternative, following Fan and Li (2001) and Tibshirani

and Tibshirani (2009), we shall here consider a data-driven selector based on the K-fold cross-validation (CV). In particular, we first split the data into K parts, denoted by $\mathcal{D}_1, \ldots, \mathcal{D}_K$, then for each $k = 1, \ldots, K$, we remove the kth part from the data and use the information criterion (3.7) to select the model, based on which predictions can be made for the removed part and are denoted by $\hat{y}_i^{-k}(c)$, $i \in \mathcal{D}_k$. The selected value $\hat{c}$ is obtained by minimizing the cross-validation criterion

$$\mathrm{CV}(c) = \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \{y_i - \hat{y}_i^{-k}(c)\}^2.$$

It can be seen from the simulation results in Section 4.2 that this CV-based tuning parameter selector performs reasonably well.

4.2. *Simulation results.* We shall in this section carry out a simulation study to examine the finite-sample performance of the generalized information criterion (3.7). Let $d = 1$ and ξ_i, $i \in \mathbb{Z}$ and η_j, $j \in \mathbb{Z}$ be i.i.d. standard normal, $a(t) = (t - 1/2)^2$, $t \in [0, 1]$ and $G(t; \mathcal{H}_k) = \xi_k + \sum_{l=1}^{\infty} a(t)^l \xi_{k-l}$, $k \in \mathbb{Z}$, $t \in [0, 1]$. For the regressor and error processes with $x_i = G(i/n; \mathcal{H}_i)$ and $e_i = \sigma(x_i, i/n)\eta_i$, $i = 1, \ldots, n$, we consider model (1.1) with the following four specifications:

(a) $m(x, t) = 2.5 \sin(2\pi t) \cos(\pi x)$ and $\sigma(x, t) = \varphi|tx|/2$;
(b) $m(x, t) = \exp(x)$ and $\sigma(x, t) = \varphi t \exp(x/3)$;
(c) $m(x, t) = 5t + 4\cos(2\pi t)x$ and $\sigma(x, t) = \varphi \exp(tx/2)$;
(d) $m(x, t) = 2 + 3x$ and $\sigma(x, t) = \varphi|x/3 + t|$,

where $\varphi > 0$ is a constant indicating the noise level. Cases (a)–(d) correspond to models I–IV, respectively, and their signal-to-noise ratios (SNRs) are roughly of the same order given the same φ. The Epanechnikov kernel $K(v) = 3(1 - v^2)/4$, $v \in [-1, 1]$, is used hereafter for both the spatial and temporal kernel functions. Let $\mathcal{X} = [-2, 2]$ and $\mathcal{T} = [0.2, 0.8]$. The tuning parameter is selected by using the tenfold CV-based method described in Section 4.1. The results are summarized in Table 1 for different noise levels $\varphi \in \{1, 2, 3\}$ and sample sizes $n = 2^k \times 250$, $0 \leq k \leq 3$. For each configuration, the results are based on 1000 simulated realizations of models (a)–(d).

It can be seen from Table 1 that the proposed model selection procedure performs reasonably well as it has very high empirical probabilities of identifying the true model, even when the sample size is moderate to small. For example, if the sample size $n = 250$, which is usually considered to be small for conducting time-varying nonparametric inference, and the data are generated by model (a) with $\varphi = 1$, then 967 out of 1000 realizations are correctly identified as the time-varying nonparametric regression model I, while 33 out of 1000 realizations are under-fitted as the simple linear regression model IV. Hence, for each combination of n and φ, in the ideal case, we expect the block to have unit diagonal components and zero off-diagonal components. For each configuration, medians of the SNR are

TABLE 1
Proportions of selecting models I–IV *for different combinations of noise levels* φ, *sample sizes n and model specifications* (a)–(d) *with* 1000 *replications for each configuration. Medians of the SNR are also reported, where for each realization* $y_i = m_i(x_i) + e_i$, $i = 1, \ldots, n$, *the SNR is defined as* $\{\sum_{i=1}^{n} m_i(x_i)^2 / \sum_{i=1}^{n} e_i^2\}^{1/2}$

| | | $\varphi = 1$ | | | | | $\varphi = 2$ | | | | | $\varphi = 3$ | | | | |
| | | | Selected model | | | | | Selected model | | | | | Selected model | | | |
n	Case	SNR	I	II	III	IV	SNR	I	II	III	IV	SNR	I	II	III	IV
250	(a)	4.36	0.967	0.000	0.000	0.033	2.16	0.920	0.000	0.000	0.080	1.45	0.840	0.000	0.000	0.160
	(b)	4.09	0.116	0.882	0.000	0.002	2.04	0.119	0.857	0.000	0.024	1.36	0.132	0.784	0.002	0.082
	(c)	3.73	0.016	0.000	0.984	0.000	1.86	0.032	0.000	0.968	0.000	1.24	0.032	0.000	0.968	0.000
	(d)	5.44	0.017	0.043	0.005	0.935	2.72	0.014	0.040	0.001	0.945	1.82	0.024	0.040	0.003	0.933
500	(a)	4.29	0.985	0.000	0.000	0.015	2.15	0.945	0.000	0.000	0.055	1.44	0.896	0.000	0.000	0.104
	(b)	4.17	0.044	0.949	0.000	0.008	2.08	0.058	0.906	0.000	0.036	1.40	0.037	0.926	0.000	0.037
	(c)	3.71	0.001	0.000	0.999	0.000	1.86	0.008	0.000	0.992	0.000	1.24	0.012	0.000	0.988	0.000
	(d)	5.42	0.007	0.037	0.000	0.956	2.71	0.012	0.042	0.001	0.945	1.81	0.005	0.026	0.006	0.963
1000	(a)	4.29	0.994	0.000	0.000	0.006	2.15	0.970	0.000	0.000	0.030	1.44	0.921	0.000	0.000	0.079
	(b)	4.17	0.004	0.992	0.000	0.004	2.08	0.005	0.975	0.000	0.020	1.40	0.015	0.957	0.000	0.028
	(c)	3.71	0.000	0.000	1.000	0.000	1.86	0.001	0.000	0.999	0.000	1.24	0.004	0.000	0.996	0.000
	(d)	5.42	0.001	0.028	0.002	0.969	2.71	0.002	0.024	0.003	0.971	1.81	0.001	0.025	0.002	0.972
2000	(a)	4.29	0.999	0.000	0.000	0.001	2.15	0.979	0.000	0.000	0.021	1.44	0.948	0.000	0.000	0.052
	(b)	4.17	0.000	0.997	0.000	0.003	2.08	0.000	0.982	0.000	0.018	1.40	0.000	0.965	0.000	0.035
	(c)	3.71	0.000	0.000	1.000	0.000	1.86	0.000	0.000	1.000	0.000	1.24	0.001	0.000	0.999	0.000
	(d)	5.42	0.000	0.014	0.001	0.985	2.71	0.000	0.014	0.001	0.985	1.81	0.000	0.014	0.000	0.986
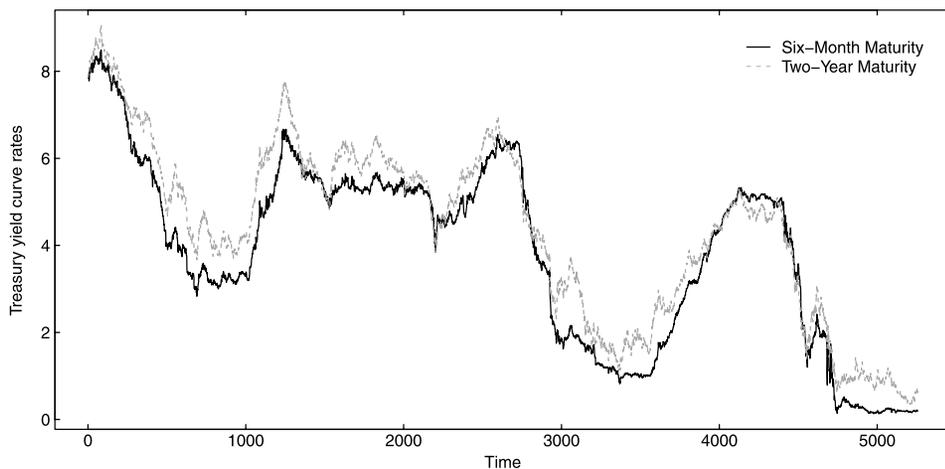
FIG. 1. *Time series plots for the U.S. daily treasury yield curve rates with six-month (*solid black*) and two-year (*dashed grey*) maturities.*

also reported, where for each realization $y_i = m_i(x_i) + e_i$, $i = 1, \ldots, n$, the SNR is defined as $\{\sum_{i=1}^{n} m_i(x_i)^2 / \sum_{i=1}^{n} e_i^2\}^{1/2}$. It can be seen that the proposed model selection procedure with the CV-based tuning parameter selector has a reasonably robust performance with respect to the noise level, and the performance improves quickly if we increase the sample size. Note that a sample size of 1000 is considered to be reasonable if one would like to conduct time-varying nonparametric inference.

4.3. *Application on modeling interest rates.* Modeling interest rates is an important problem in finance. In Black and Scholes (1973) and Merton (1974) interest rates were assumed to be constants. A popular model is the time-homogeneous diffusion process (1.3) with linear drift function; see, for example, Vasicek (1977), Courtadon (1982), Cox, Ingersoll and Ross (1985) and Chan et al. (1992). Its discretized version is given by model IV. Aït-Sahalia (1996), Stanton (1997) and Liu and Wu (2010) considered model (1.3) with nonlinear drift function, which relates to model II. We consider the daily U.S. treasury yield curve rates with six-month and two-year maturities during 01/02/1990–12/31/2010. The data can be obtained from the U.S. Department of the Treasury website at http://www.treasury.gov/. Both series contain $n = 5256$ daily rates, and their time series plots are shown in Figure 1.

We shall here model the data by the time-varying diffusion process (1.2), and apply the proposed model selection procedure to determine the forms of the drift functions. Let $x_i = r_{t_i}$ be the observation at day i. Since a year has 250 transaction days, $\Delta = t_i - t_{i-1} = 1/250$. Following Liu and Wu (2010), we consider the

TABLE 2
Results of the model selection procedure based on the generalized information criterion (3.7) *for treasury yield rates with six-month and two-year maturity periods*

Model	Six-month maturity			Two-year maturity		
	log(RSS/n)	DF	GIC	log(RSS/n)	DF	GIC
I	−6.853	69.54	−6.790	−6.126	69.54	−6.063
II	−6.824	11.10	−6.814	−6.114	11.10	−6.104
III	−6.851	22.19	−6.831	−6.126	22.19	−6.106
IV	−6.822	2.00	−6.820	−6.113	2.00	−6.111

following discretized version of (1.2):

$$y_i = r_{t_{i+1}} - r_{t_i} = \mu(x_i, i/n)\Delta + \sigma(x_i, i/n)\Delta^{1/2}\eta_i$$

(4.1)

$$\text{where } \eta_i = \frac{\mathbb{B}_{t_{i+1}} - \mathbb{B}_{t_i}}{\Delta^{1/2}}.$$

Note that η_i are i.i.d. $N\{0, 1\}$ random variables. We shall here write $\mu(x_i, i/n)\Delta$ and $\sigma(x_i, i/n)\Delta^{1/2}$ in (4.1) as $m(x_i, i/n)$ and $\sigma(x_i, i/n)$ in the sequel. Then specifications of Vasicek (1977) and Liu and Wu (2010) become models IV and II, respectively.

For the treasury yield curve rates with six-month maturity, let $\mathscr{T} = [0.2, 0.8]$, and $\mathscr{X} = [0.18, 7.89]$ which includes 95.5% of the daily rates x_i. The selected bandwidths and tuning parameter are $b_n(\text{I}) = 0.12$, $h_n(\text{I}) = 0.82$, $h_n(\text{II}) = 0.62$, $b_n(\text{III}) = 0.09$ and $\hat{\tau}_n = 0.00090$. The results are summarized in Table 2. Hence, the time-varying coefficient model III is selected, and we conclude that the treasury yield curve rates with six-month maturity should be modeled by (1.2) with $\mu(r_t, t) = \beta_0(t) + \beta_1(t)r_t$ for some smoothly varying functions $\beta_0(\cdot)$ and $\beta_1(\cdot)$, which serves as a time-varying version of Chan et al. (1992).

We then consider the treasury yield curve rates with two-year maturity. Let $\mathscr{T} = [0.2, 0.8]$ and $\mathscr{X} = [0.67, 8.16]$ which includes 95.1% of the daily rates x_i. The selected bandwidths and tuning parameter are $b_n(\text{I}) = 0.12$, $h_n(\text{I}) = 0.75$, $h_n(\text{II}) = 0.56$, $b_n(\text{III}) = 0.09$ and $\hat{\tau}_n = 0.00090$. Based on Table 2, the linear regression model IV is selected. In comparison with the results with six-month maturity, our analysis suggests that treasury yield rates with longer maturity are more stable over time.

5. Conclusion. The paper considers a time-varying nonparametric regression model, namely model I, which is able to capture time-varying and nonlinear relationships between the response variable and the explanatory variables. It includes the popular nonparametric regression model II and time-varying coefficient model III as special cases, and all of them are generalizations of the simple linear regression model IV. In comparison with existing results, the current paper

makes two major contributions. First, we develop an asymptotic theory on non-parametric estimation of the time-varying regression model (1.1) under the new framework of Draghicescu, Guillas and Wu (2009). Compared with the classical strong mixing conditions as used by Vogt (2012), the current framework is convenient to work with and often leads to optimal asymptotic results. In the proof, we use both the martingale decomposition and the m-dependence approximation techniques to obtain sharp results. Second, although the time-varying regression model I is quite general by allowing a time-varying nonlinear relationship between the response variable and the explanatory variables, it can be useful in practice to check whether it can be reduced to its simpler special cases, namely models II–IV which have been extensively used in the literature. However, existing results on model selection usually focused on distinguishing between models II and IV and between models III and IV, and much less attention has been paid to distinguishing between models II and III. Note that models II and III are both generalizations of the simple linear regression model IV but in completely different aspects, and therefore it is desirable if we can have a statistically valid method to decide which generalization (or the more general model I) should be used for a given data set. The current paper fills this gap by proposing an information criterion (3.7) in Section 3.3, which can be used to select the true model among candidate models I–IV and its selection consistency is provided by Theorem 3.5. Therefore, the current paper sheds new light on distinguishing between nonlinear and nonstationary generalizations of simple linear regression models, and the results are applied to find appropriate models for short-term and long-term interest rates.

6. Technical proofs. We shall in this section provide technical proofs for Theorems 3.1–3.5. Because of the time-varying feature and nonstationarity, the proofs are much more involved than existing ones for stationary processes. We shall here use techniques of martingale approximation and m-dependent approximation. Let $\boldsymbol{\varepsilon}_i = (\boldsymbol{\xi}_i^\top, \eta_i)^\top$ and $\mathcal{F}_i = (\ldots, \boldsymbol{\varepsilon}_{i-1}, \boldsymbol{\varepsilon}_i)$ be the corresponding shift process. We define the projection operator

$$\mathcal{P}_k \cdot := E(\cdot|\mathcal{F}_k) - E(\cdot|\mathcal{F}_{k-1}), \qquad k \in \mathbb{Z}.$$

Throughout this section, $C > 0$ denotes a constant whose value may vary from place to place. Let $\alpha_{i,n}(\mathbf{u}, t)$, $i = 1, \ldots, n$, be a triangular array of deterministic nonnegative weight functions, $(\mathbf{u}, t) \in \mathbb{R}^d \times [0, 1]$. Lemma 6.1 provides a bound for the quantity

$$Q_\alpha(\mathbf{u}, t) = \sum_{i=1}^n \{f_1(\mathbf{u}, i/n|\mathcal{F}_{i-1}) - f(\mathbf{u}, i/n)\}\alpha_{i,n}(\mathbf{u}, t),$$

and is useful for proving Theorems 3.1–3.4.

LEMMA 6.1. *Let* $A_n(\mathbf{u}, t) = \max_{1 \le i \le n} |\alpha_{i,n}(\mathbf{u}, t)|$ *and define* $\bar{A}_n(\mathbf{u}, t) = n^{-1} \sum_{i=1}^n |\alpha_{i,n}(\mathbf{u}, t)|$. *Then* $\|Q_\alpha(\mathbf{u}, t)\| \le \{nA_n(\mathbf{u}, t)\bar{A}_n(\mathbf{u}, t)\}^{1/2}\Psi_{0,2}$.

PROOF.   Since $\mathcal{P}_k Q_\alpha(\mathbf{u}, t)$, $k \in \mathbb{Z}$ form a sequence of martingale differences, we have

$$\|Q_\alpha(\mathbf{u}, t)\|^2 = \sum_{k=-\infty}^{n} \left\| \sum_{i=1}^{n} \mathcal{P}_k\{f_1(\mathbf{u}, i/n | \mathcal{F}_{i-1})\} \alpha_{i,n}(\mathbf{u}, t) \right\|^2$$

$$\leq \sum_{k=-\infty}^{n} \left\{ \sum_{i=1}^{n} \psi_{i-k-1,2} |\alpha_{i,n}(\mathbf{u}, t)| \right\}^2,$$

and the result follows by observing that $\sum_{i=1}^{n} \psi_{i-k-1,2} |\alpha_{i,n}(\mathbf{u}, t)| \leq A_n(\mathbf{u}, t) \Psi_{0,2}$ and $\sum_{i=1}^{n} \sum_{k \in \mathbb{Z}} \psi_{i-k-1,2} |\alpha_{i,n}(\mathbf{u}, t)| \leq n A_n(\mathbf{u}, t) \Psi_{0,2}$.   □

LEMMA 6.2.   *Assume* (A1)–(A3) *and* $\eta_i \in \mathcal{L}^p$, $p > 2$, $i = 1, \ldots, n$. (i) *If* $b_n \to 0$, $h_n \to 0$ *and* $nb_n h_n^d \to \infty$, *then for any* $(\mathbf{u}, t) \in \mathbb{R}^d \times (0, 1)$,

$$(nb_n h_n^d)^{1/2} [\hat{T}_n(\mathbf{u}, t) - E\{\hat{T}_n(\mathbf{u}, t)\}] \Rightarrow N[0, \{m(\mathbf{u}, t)^2 + \sigma(\mathbf{u}, t)^2\} f(\mathbf{u}, t) \lambda_K],$$

where $\lambda_K = \lambda_{K_S} \lambda_{K_T}$. (ii) *If* $h_n \to 0$ *and* $nh_n^d \to \infty$, *then for any* $\mathbf{u} \in \mathbb{R}^d$,

$$(nh_n^d)^{1/2} [\tilde{T}_n(\mathbf{u}) - E\{\tilde{T}_n(\mathbf{u})\}] \Rightarrow N\left[0, \lambda_{K_S} \int_0^1 \{m(\mathbf{u}, t)^2 + \sigma(\mathbf{u}, t)^2\} f(\mathbf{u}, t) \, dt\right].$$

PROOF.   Write

$$\hat{T}_n(\mathbf{u}, t) - E\{\hat{T}_n(\mathbf{u}, t)\} = M_n(\mathbf{u}, t) + N_n(\mathbf{u}, t),$$

where

$$M_n(\mathbf{u}, t) = \sum_{i=1}^{n} [y_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) - E\{y_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) | \mathcal{F}_{i-1}\}] w_{b_n,i}(t)$$

has summands of martingale differences, and

$$N_n(\mathbf{u}, t) = \sum_{i=1}^{n} [E\{y_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) | \mathcal{F}_{i-1}\} - E\{y_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i)\}] w_{b_n,i}(t)$$

is the remaining term. Let $\alpha_{i,n}(\mathbf{u}, t) = m(\mathbf{u}, i/n) w_{b_n,i}(t)$, and by Lemma 6.1,

$$\|N_n(\mathbf{u}, t)\| \leq \int_{[-1,1]^d} K_S(\mathbf{s}) \|Q_\alpha(\mathbf{u} - h_n \mathbf{s}, t)\| \, d\mathbf{s} = O\{(nb_n)^{-1/2}\}.$$

We apply the martingale central limit theorem on $M_n(\mathbf{u}, t)$ to show (i). Since

$$\sum_{i=1}^{n} \|[y_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) - E\{y_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) | \mathcal{F}_{i-1}\}] w_{b_n,i}(t)\|_p^p$$

$$\leq \sum_{i=1}^{n} 2^p \|y_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i)\|_p^p w_{b_n,i}(t)^p = O\{(nb_n h_n^d)^{1-p}\},$$

the Lindeberg condition is satisfied by observing that $p > 2$. Let

$$L_n(\mathbf{s}, t) = \sum_{i=1}^{n} \{m(\mathbf{s}, i/n)^2 + \sigma(\mathbf{s}, i/n)^2\}\{f_1(\mathbf{s}, i/n|\mathcal{F}_{i-1}) - f(\mathbf{s}, i/n)\}w_{b_n,i}(t)^2.$$

Then by (A1) and Lemma 6.1,

$$h_n^d \sum_{i=1}^{n} [E\{y_i^2 K_{S,h_n}(\mathbf{u} - \mathbf{x}_i)^2|\mathcal{F}_{i-1}\} - E\{y_i^2 K_{S,h_n}(\mathbf{u} - \mathbf{x}_i)^2\}]w_{b_n,i}(t)^2$$

$$= \int_{[-1,1]^d} K_S(\mathbf{s})^2 L_n(\mathbf{u} - h_n\mathbf{s}, t)\,d\mathbf{s} = O_p\{(nb_n)^{-3/2}\}.$$

Also, write $E\{y_i K_{S,h_n}(\mathbf{u}-\mathbf{x}_i)|\mathcal{F}_{i-1}\} = \int_{[-1,1]^d} m(\mathbf{u}-h_n\mathbf{s})K_S(\mathbf{s})f_1(\mathbf{u}-h_n\mathbf{s}, i/n|\mathcal{F}_{i-1})\,d\mathbf{s}$. Then we have

$$(nb_n h_n^d) \sum_{i=1}^{n} \|E\{y_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i)|\mathcal{F}_{i-1}\}\|^2 w_{b_n,i}(t)^2 = O(h_n^d),$$

and (i) follows by $(nb_n h_n^d) \sum_{i=1}^{n} E\{y_i^2 K_{S,h_n}(\mathbf{u} - \mathbf{x}_i)^2\}w_{b_n,i}(t)^2 = \{m(\mathbf{u}, t)^2 + \sigma(\mathbf{u}, t)^2\}f(\mathbf{u}, t)\lambda_{K_S}\lambda_{K_T} + o(1)$. Case (ii) can be similarly proved. □

PROOFS OF THEOREMS 3.1 AND 3.2. Letting $m \equiv 1$ and $\sigma \equiv 0$ in Lemma 6.2, (3.1) and (3.3) follow directly. For (3.2), write

$$\hat{T}_n(\mathbf{u}, t) - \hat{f}_n(\mathbf{u}, t)\frac{E\{\hat{T}_n(\mathbf{u}, t)\}}{E\{\hat{f}_n(\mathbf{u}, t)\}} = I_n + II_n,$$

where

$$I_n = [\hat{f}_n(\mathbf{u}, t) - E\{\hat{f}_n(\mathbf{u}, t)\}]\left[m(\mathbf{u}, t) - \frac{E\{\hat{T}_n(\mathbf{u}, t)\}}{E\{\hat{f}_n(\mathbf{u}, t)\}}\right] = o_p\{(nb_n h_n^d)^{-1/2}\}$$

and

$$II_n = \{\hat{T}_n(\mathbf{u}, t) - m(\mathbf{u}, t)\hat{f}_n(\mathbf{u}, t)\} - E\{\hat{T}_n(\mathbf{u}, t) - m(\mathbf{u}, t)\hat{f}_n(\mathbf{u}, t)\}.$$

Note that

$$\hat{T}_n(\mathbf{u}, t) - m(\mathbf{u}, t)\hat{f}_n(\mathbf{u}, t) = \sum_{i=1}^{n}\{y_i - m(\mathbf{u}, t)\}K_{S,h_n}(\mathbf{u} - \mathbf{x}_i)w_{b_n,i}(t),$$

by Lemma 6.2(i),

$$(nb_n h_n^d)^{1/2}II_n \Rightarrow N\{0, \sigma(\mathbf{u}, t)^2 f(\mathbf{u}, t)\lambda_{K_S}\lambda_{K_T}\}.$$

Since $\hat{f}_n(\mathbf{u}, t) \to f(\mathbf{u}, t)$ in probability, (3.2) follows by Slutsky's theorem. Case (3.4) can be similarly proved. □

PROOFS OF THEOREMS 3.3 AND 3.4. We shall first prove Theorem 3.3(i). For this, since $\sup_{t \in [0,1]} \|\mathbf{G}(t; \mathcal{H}_0)\|_r < \infty$, we have $\max_{1 \leq i \leq n} |\mathbf{x}_i| = o_p(n^{1/r'})$ for any $r' < r$. Hence, $\sup_{t \in [0,1]} \sup_{|\mathbf{u}| > n^{1/r'}} \hat{f}_n(\mathbf{u}, t) = 0$ almost surely, and $\sup_{t \in [0,1]} \sup_{|\mathbf{u}| > n^{1/r'}} E\{\hat{f}_n(\mathbf{u}, t)\} = O(n^{-1}h_n^{-d}) = o\{(nb_n h_n^d)^{-1/2}\}$. Therefore, it suffices to deal with the case in which $|\mathbf{u}| \leq n^{1/r'}$. We shall here assume that $d = 1$. Cases with higher dimensions can be similarly proved without extra essential difficulties, but they aew technically tedious. Let

$$(6.1) \quad \begin{aligned} \hat{f}_n^\circ(\mathbf{u}, t) &= \sum_{i=1}^n E\{K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) w_{b_n, i}(t) | \mathcal{F}_{i-1}\} \\ &= \sum_{i=1}^n w_{b_n, i}(t) \int K_S(\mathbf{s}) f_1(\mathbf{u} - h_n \mathbf{s}, i/n | \mathcal{F}_{i-1}) \, d\mathbf{s}. \end{aligned}$$

Observe that $K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) w_{b_n, i}(t) - E\{K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) w_{b_n, i}(t) | \mathcal{F}_{i-1}\}, i = 1, \ldots, n$, form a sequence of bounded martingale differences. By the inequality of Freedman (1975) and the proof of Theorem 2 in Wu, Huang and Huang (2010), we obtain that, for some large constant $\lambda > 0$,

$$\mathrm{pr}\Big\{ \sup_{t \in [0,1]} \sup_{|\mathbf{u}| \leq n^{1/r'}} |\hat{f}_n(\mathbf{u}, t) - \hat{f}_n^\circ(\mathbf{u}, t)| \geq \lambda (nb_n h_n)^{-1/2} (\log n)^{1/2} \Big\} = o(n^{-2}).$$

Let $\vartheta_i(\mathbf{u}) = f_1(\mathbf{u}, i/n | \mathcal{F}_{i-1}) - f(\mathbf{u}, i/n)$ and $\Theta_{l,j}(\mathbf{u}) = \sum_{i=l}^{l+j} \vartheta_i(\mathbf{u})$. By (6.1) and the proof of Lemma 5.3 in Zhang and Wu (2012), it suffices to show that for all l,

$$(6.2) \quad \mathrm{pr}\Big\{ \max_{0 \leq j \leq nb_n} \sup_{|\mathbf{u}| \leq n^{1/r'}} |\Theta_{l,j}(\mathbf{u})| \geq (h_n^{-1} nb_n \log n)^{1/2} \Big\} = o(b_n).$$

Let $\Delta = (nb_n h_n)^{-1/2} (\log n)^{1/4}$ and $\lfloor \mathbf{u} \rfloor_\Delta = \Delta \lfloor \mathbf{u}/\Delta \rfloor$. By Theorem 2(ii) in Liu, Xiao and Wu (2013), under condition (A4),

$$(6.3) \quad \begin{aligned} &\mathrm{pr}\Big\{ \max_{0 \leq j \leq nb_n} \sup_{|\mathbf{u}| \leq n^{1/r'}} |\Theta_{l,j}(\lfloor \mathbf{u} \rfloor_\Delta)| \geq (h_n^{-1} nb_n \log n)^{1/2} \Big\} \\ &= O\Big\{ \frac{nb_n \Delta^{-1} n^{1/r'}}{(h_n^{-1} nb_n \log n)^{q/2}} \Big\}. \end{aligned}$$

By (A3), $\max_{0 \leq j \leq nb_n} \sup_{|\mathbf{u}| \leq n^{1/r'}} |\Theta_{l,j}(\mathbf{u}) - \Theta_{l,j}(\lfloor \mathbf{u} \rfloor_\Delta)| = O(nb_n \Delta)$, (6.2) follows. For Theorem 3.3(ii), by Lemma 6.3, $\sup_{t \in [0,1]} \sup_{\mathbf{u} \in \mathcal{X}} |\hat{T}_n(\mathbf{u}, t) - E\{\hat{T}_n(\mathbf{u}, t)\}| = O_p\{(nb_n h_n^d)^{-1/2} (\log n)^{1/2} + (nb_n h_n^d)^{-1} (n^{1/p} \log n)\}$. Since

$$\begin{aligned} &\hat{f}_n(\mathbf{u}, t) \Big[\hat{m}_n(\mathbf{u}, t) - \frac{E\{\hat{T}_n(\mathbf{u}, t)\}}{E\{\hat{f}_n(\mathbf{u}, t)\}} \Big] \\ &= \hat{T}_n(\mathbf{u}, t) - E\{\hat{T}_n(\mathbf{u}, t)\} + E\{\hat{T}_n(\mathbf{u}, t)\} \Big[1 - \frac{\hat{f}_n(\mathbf{u}, t)}{E\{\hat{f}_n(\mathbf{u}, t)\}} \Big], \end{aligned}$$

the result follows. Theorem 3.4 can be similarly proved. $\quad\square$

Recall that $\mathscr{X} \in \mathbb{R}^d$ is a compact set. Lemma 6.3 provides uniform bounds for

$$\hat{U}(\mathbf{u}, t) = \sum_{i=1}^{n} m(\mathbf{x}_i, i/n) K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) w_{b_n,i}(t);$$

$$\hat{V}(\mathbf{u}, t) = \sum_{i=1}^{n} \sigma(\mathbf{x}_i, i/n) \eta_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) w_{b_n,i}(t);$$

$$\tilde{U}(\mathbf{u}) = n^{-1} \sum_{i=1}^{n} m(\mathbf{x}_i, i/n) K_{S,h_n}(\mathbf{u} - \mathbf{x}_i);$$

$$\tilde{V}(\mathbf{u}) = n^{-1} \sum_{i=1}^{n} \sigma(\mathbf{x}_i, i/n) \eta_i K_{S,h_n}(\mathbf{u} - \mathbf{x}_i),$$

and is useful in proving Theorems 3.3 and 3.4.

LEMMA 6.3. *Assume* (A1), (A3), (A4), $\eta_i \in \mathcal{L}^p$ *for some* $p > 2$, $i = 1, \ldots, n$, $b_n \to 0$ *and* $h_n \to 0$. *Let* $\chi_n = n^{1/p} \log n$. (i) *If* $n b_n h_n^d \to \infty$ *and* $n^{2+d-q} b_n^{d-q} h_n^{d(d+q)} \to 0$, *then*

(6.4) $$\sup_{t \in [0,1]} \sup_{\mathbf{u} \in \mathscr{X}} |\hat{U}(\mathbf{u}, t)| = O_p\{(n b_n h_n^d)^{-1/2} (\log n)^{1/2}\},$$

(6.5) $$\sup_{t \in [0,1]} \sup_{\mathbf{u} \in \mathscr{X}} |\hat{V}(\mathbf{u}, t)| = O_p\{(n b_n h_n^d)^{-1/2} (\log n)^{1/2} + (n b_n h_n^d)^{-1} \chi_n\}.$$

(ii) *If* $n h_n^d \to \infty$ *and* $n^{2+d-q} h_n^{d(d+q)} \to 0$, *then*

(6.6) $$\sup_{\mathbf{u} \in \mathscr{X}} |\tilde{U}(\mathbf{u})| = O_p\{(n h_n^d)^{-1/2} (\log n)^{1/2}\},$$

(6.7) $$\sup_{\mathbf{u} \in \mathscr{X}} |\tilde{V}(\mathbf{u})| = O_p\{(n h_n^d)^{-1/2} (\log n)^{1/2} + (n h_n^d)^{-1} \chi_n\}.$$

PROOF. The proof of (6.4) is similar to that of Theorem 3.3(i), and we shall only outline the key differences. First, the supreme in (6.4) is taken over $\mathbf{u} \in \mathscr{X}$, a compact set, instead of $\mathbb{R}^d$. Hence the truncation argument is no longer needed, and the term $\Delta^{-1} n^{1/r'}$ in (6.3) can be replaced by Δ^{-1}. Second, $E\{m(\mathbf{x}_i, i/n) K_{S,h_n}(\mathbf{u} - \mathbf{x}_i) | \mathcal{F}_{i-1}\} = \int_{[-1,1]^d} K_S(\mathbf{s}) f_1^\dagger(\mathbf{u} - h_n \mathbf{s}, i/n | \mathcal{F}_{i-1}) d\mathbf{s}$, where $f_1^\dagger(\mathbf{u}, t | \mathcal{F}_{i-1}) = m(\mathbf{u}, t) f_1(\mathbf{u}, t | \mathcal{F}_{i-1})$. By (A1), $f_1^\dagger$ satisfies condition (A3), and its predictive dependence measure is of order (2.5). Hence the proof of Theorem 3.3(i) applies. Case (6.6) can be similarly handled. For (6.5) and (6.7), we shall only provide the proof of (6.7) since (6.5) can be similarly derived. Let $\eta_i^\star = \eta_i \mathbb{1}_{\{|\eta_i| \le n^{1/p}\}}$ and $\tilde{V}^\star(\mathbf{u})$ be the counterpart of $\tilde{V}(\mathbf{u})$ with η_i therein

replaced by $\eta_i^\star$, $i = 1, \ldots, n$. Also, let $\eta_i^\dagger = \eta_i^\star - E(\eta_i^\star)$, and we can similarly define $\tilde{V}^\dagger(\mathbf{u})$. Since $\eta_i \in \mathcal{L}^p$ are i.i.d., we have $\max_{1 \le i \le n} |\eta_i| = o_p(n^{1/p})$ and $\mathrm{pr}\{\tilde{V}(\mathbf{u}) = \tilde{V}^\star(\mathbf{u})$ for all $\mathbf{u} \in \mathcal{X}\} \to 1$. In addition,

$$\tilde{V}^\star(\mathbf{u}) - \tilde{V}^\dagger(\mathbf{u}) = n^{-1} E(\eta_i^\star) \sum_{i=1}^n \sigma(\mathbf{x}_i, i/n) K_{S,h_n}(\mathbf{u} - \mathbf{x}_i).$$

Since $E(\eta_i) = 0$, we have $E(\eta_i^\star) = -E(\eta_i \mathbb{1}_{\{|\eta_i| > n^{1/p}\}}) = O(n^{1/p-1})$, and by (6.6), it suffices to show that (6.7) holds with $\tilde{V}^\dagger(\mathbf{u})$. Let $\bar{\mathcal{X}} = \{\mathbf{u} \in \mathbb{R}^d : |\mathbf{u} - \mathbf{v}| \le 1$ for some $\mathbf{v} \in \mathcal{X}\}$, $c_K = \sup_{\mathbf{v} \in [-1,1]^d} |K_S(\mathbf{v})|$, $c_1 = \mathrm{var}(\eta_i^\star)$ and $c_2 = \sup_{t \in [0,1]} \sup_{\mathbf{u} \in \bar{\mathcal{X}}} \sigma(\mathbf{u}, t)^2 < \infty$ under (A1). Recall c_0 from (A3), then $|\sigma(\mathbf{x}_i, i/n)\eta_i^\dagger K_{S,h_n}(\mathbf{u} - \mathbf{x}_i)| \le 2c_2^{1/2} c_K n^{1/p} h_n^{-d}$ and

$$E\{\sigma(\mathbf{x}_i, i/n)^2 (\eta_i^\dagger)^2 K_{S,h_n}(\mathbf{u} - \mathbf{x}_i)^2 | \mathcal{F}_{i-1}\} \le h_n^{-d} c_0 c_1 c_2 \lambda_{K_S}.$$

Let $\varpi_n = (nh_n^d)^{-1/2} (\log n)^{1/2} + (nh_n^d)^{-1}(n^{1/p} \log n)$. Applying the inequality of Freedman (1975) to $\tilde{V}^\dagger(\mathbf{u})$, we obtain that, for some large constant $\lambda > 0$,

$$\mathrm{pr}\{|\tilde{V}^\dagger(\mathbf{u})| \ge \lambda \varpi_n\}$$
$$\le 2 \exp\left(-\frac{\lambda^2 \varpi_n^2}{4c_2^{1/2} c_K \lambda n^{1/p-1} h_n^{-d} \varpi_n + 2c_0 c_1 c_2 \lambda_{K_S} n^{-1} h_n^{-d}}\right) = O(n^{-\lambda^{1/2}}),$$

and (6.7) follows by the discretization argument as in (6.3). $\quad\square$

Let $\omega_n = (nb_n h_n^d)^{-1} \log n + b_n^4 + h_n^4$, Lemmas 6.4–6.7 provide asymptotic properties of the restricted residual sum of squares for models I–IV, respectively, and are useful in proving Theorem 3.5. We shall here only provide the proof of Lemmas 6.4 and 6.5, which relate to nonparametric kernel estimation of nonlinear regression functions that have been studied in Sections 3.1 and 3.2. Lemmas 6.6 and 6.7 relate to linear models with time-varying and time-constant coefficients, and the proof is available in the supplementary material [Zhang and Wu (2015)].

LEMMA 6.4. *Assume* (A1), (A3)–(A5), $\eta_i \in \mathcal{L}^p$ *for some* $p > 2$, $i = 1, \ldots, n$, $b_n \to 0$, $h_n \to 0$ *and* $nb_n h_n^d/(\log n)^2 \to \infty$. *If* $n^{2+d-q} b_n^{d-q} h_n^{d(d+q)} \to 0$ *and* $n^{1/p-1/2} b_n^{-1/2} h_n^{-d/2} \to 0$, *then*

$$n^{-1} \mathrm{RSS}_n(\mathcal{X}, \mathcal{T}, \mathrm{I}) = n^{-1} \sum_{i \in \mathcal{I}_n} \mathbb{1}_{\{\mathbf{x}_i \in \mathcal{X}\}} e_i^2 + O_p\left\{\omega_n + \frac{b_n + h_n}{(nh_n^d)^{1/2}}\right\}.$$

PROOF. Note that one can have the decomposition

$$n^{-1} \mathrm{RSS}_n(\mathcal{X}, \mathcal{T}, \mathrm{I}) = n^{-1} \sum_{i \in \mathcal{I}_n} \mathbb{1}_{\{\mathbf{x}_i \in \mathcal{X}\}} e_i^2 + I_n - 2II_n,$$

where $I_n = n^{-1} \sum_{i \in \mathscr{I}_n} \{\hat{m}_n(\mathbf{x}_i, i/n) - m(\mathbf{x}_i, i/n)\}^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} = O_p(\omega_n)$ by Theorem 3.3, and

$$II_n = n^{-1} \sum_{i \in \mathscr{I}_n} \{\hat{m}_n(\mathbf{x}_i, i/n) - m(\mathbf{x}_i, i/n)\} \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} e_i.$$

We shall now deal with the term II_n. By Lemma 6.3(i) and Theorem 3.3, $\sup_{t \in \mathscr{T}} \sup_{\mathbf{u} \in \mathscr{X}} |\{\hat{f}_n(\mathbf{u}, t) - f(\mathbf{u}, t)\}\{\hat{m}_n(\mathbf{u}, t) - m(\mathbf{u}, t)\}| = O_p(\omega_n)$ and thus

$$\sup_{t \in \mathscr{T}} \sup_{\mathbf{u} \in \mathscr{X}} \left| \hat{m}_n(\mathbf{u}, t) - m(\mathbf{u}, t) - \frac{\hat{T}_n(\mathbf{u}, t) - m(\mathbf{u}, t)\hat{f}_n(\mathbf{u}, t)}{f(\mathbf{u}, t)} \right| = O_p(\omega_n).$$

Let $\Xi_{i,j,n} = \{m(\mathbf{x}_j, j/n) - m(\mathbf{x}_i, i/n)\}$, and we can then write

$$II_n = II_{n,L} + II_{n,Q} + O_p(\omega_n),$$

where

$$II_{n,L} = n^{-1} \sum_{i \in \mathscr{I}_n} \frac{\sum_{j=1}^n \Xi_{i,j,n} K_{S,h_n}(\mathbf{x}_i - \mathbf{x}_j) w_{b_n,j}(i/n)}{f(\mathbf{x}_i, i/n)} \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} e_i$$

and

$$II_{n,Q} = n^{-1} \sum_{i \in \mathscr{I}_n} \sum_{j=1}^n \frac{K_{S,h_n}(\mathbf{x}_i - \mathbf{x}_j) w_{b_n,j}(i/n) \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}}}{f(\mathbf{x}_i, i/n)} e_i e_j.$$

Using the orthogonality of martingale differences and Lemma 2 of Wu, Huang and Huang (2010), we have $II_{n,L} = O_p\{(nh_n^d)^{-1/2}(b_n + h_n)\}$. Also, by splitting the sum in $II_{n,Q}$ for cases with $i = j$ and $i \neq j$, one can have $II_{n,Q} = O_p\{(nb_n)^{-1} + n^{-1/2}(nb_n h_n^d)^{-1/2}\}$. Lemma 6.4 follows by $(b_n h_n^d)^{-1/2} = o\{(b_n h_n^d)^{-1}\}$. □

LEMMA 6.5. *Assume* (A1), (A3)–(A5), $\eta_i \in \mathcal{L}^p$ *for some* $p > 2$, $i = 1, \ldots, n$, $h_n \to 0$ *and* $nh_n^d \to \infty$. *If* $n^{2+d-q}h_n^{d(d+q)} \to 0$ *and* $n^{1/p-1/2}h_n^{-d/2} \to 0$, *then* (i)

$$n^{-1}\mathrm{RSS}_n(\mathscr{X}, \mathscr{T}, \mathrm{II}) = \int_{\mathscr{X}} \int_{\mathscr{T}} \{m(\mathbf{u}, t) - \bar{m}(\mathbf{u})\}^2 f(\mathbf{u}, t) \, dt \, d\mathbf{u}$$

$$+ n^{-1} \sum_{i \in \mathscr{I}_n} e_i^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} + O_p\left\{ \left(\frac{\log n}{nh_n^d} \right)^{1/2} + h_n^2 \right\}.$$

(ii) *If in addition model* II *is correctly specified, then*

$$n^{-1}\mathrm{RSS}_n(\mathscr{X}, \mathscr{T}, \mathrm{II}) = n^{-1} \sum_{i \in \mathscr{I}_n} e_i^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} + O_p\left\{ \frac{\log n}{nh_n^d} + h_n^4 + \frac{h_n}{(nh_n^d)^{1/2}} \right\}.$$

PROOF. By Theorem 3.4,

$$\mathrm{RSS}_n(\mathscr{X}, \mathscr{T}, \mathrm{II}) = \sum_{i \in \mathscr{I}_n} \left[\{y_i - \bar{m}(\mathbf{x}_i)\} - \{\hat{\mu}_n(\mathbf{x}_i) - \bar{m}(\mathbf{x}_i)\} \right]^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}}$$

$$= I_n + O_p[n\{(nh_n^d)^{-1/2}(\log n)^{1/2} + h_n^2\}],$$

where by Lemma 2 in Wu, Huang and Huang (2010),

$$I_n = \sum_{i \in \mathscr{I}_n} \left[\{y_i - m(\mathbf{x}_i, i/n)\} + \{m(\mathbf{x}_i, i/n) - \bar{m}(\mathbf{x}_i)\} \right]^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}}$$

$$= \sum_{i \in \mathscr{I}_n} \{m(\mathbf{x}_i, i/n) - \bar{m}(\mathbf{x}_i)\}^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} + \sum_{i \in \mathscr{I}_n} e_i^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} + O_p(n^{1/2}).$$

Since $\mathscr{X} \in \mathbb{R}^d$ is a compact set, by the proof of Lemma 6.2, we have

$$\sum_{i \in \mathscr{I}_n} \{m(\mathbf{x}_i, i/n) - \bar{m}(\mathbf{x}_i)\}^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}}$$

$$= \sum_{i \in \mathscr{I}_n} E\left[\{m(\mathbf{x}_i, i/n) - \bar{m}(\mathbf{x}_i)\}^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} \right] + O_p(n^{1/2})$$

$$= n \int_{\mathscr{X}} \int_{\mathscr{T}} \{m(\mathbf{u}, t) - \bar{m}(\mathbf{u})\}^2 f(\mathbf{u}, t) \, dt \, d\mathbf{u} + O(1 + n^{1/2}),$$

and (i) follows. Case (ii) follows by a similar argument as in Lemma 6.4. □

LEMMA 6.6. *Assume* (A1)–(A3), (A6), (P2) *and* $\eta_i \in \mathcal{L}^p$ *for some* $p > 2$, $i = 1, \ldots, n$. *If* $b_n \to 0$ *and* $n b_n \to \infty$, *then* (i)

$$n^{-1} \mathrm{RSS}_n(\mathscr{X}, \mathscr{T}, \mathrm{III}) = \int_{\mathscr{X}} \int_{\mathscr{T}} \{m(\mathbf{u}, t) - \mathbf{u}^\top \boldsymbol{\beta}_n(t)\}^2 f(\mathbf{u}, t) \, dt \, d\mathbf{u}$$

$$+ n^{-1} \sum_{i \in \mathscr{I}_n} e_i^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} + O_p(\phi_n).$$

(ii) *If in addition model* III *is correctly specified, then*

$$n^{-1} \mathrm{RSS}_n(\mathscr{X}, \mathscr{T}, \mathrm{III}) = n^{-1} \sum_{i \in \mathscr{I}_n} e_i^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} + O_p\left(\phi_n^2 + \frac{b_n^2}{n^{1/2}} \right).$$

LEMMA 6.7. *Assume* (A1)–(A3), (A6), (P1) *and* $\eta_i \in \mathcal{L}^p$ *for some* $p > 2$, $i = 1, \ldots, n$. *Then* (i)

$$n^{-1} \mathrm{RSS}_n(\mathscr{X}, \mathscr{T}, \mathrm{IV}) = \int_{\mathscr{X}} \int_{\mathscr{T}} \{m(\mathbf{u}, t) - \mathbf{u}^\top \boldsymbol{\theta}_n\}^2 f(\mathbf{u}, t) \, dt \, d\mathbf{u}$$

$$+ n^{-1} \sum_{i \in \mathscr{I}_n} e_i^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} + O_p(n^{-1/2}).$$

(ii) *If in addition model* IV *is correctly specified, then*

$$n^{-1} \mathrm{RSS}_n(\mathscr{X}, \mathscr{T}, \mathrm{IV}) = n^{-1} \sum_{i \in \mathscr{I}_n} e_i^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}} + O_p(n^{-1}).$$

PROOF OF THEOREM 3.5. For model I, the AMSE optimal bandwidths satisfy $b_n(\mathrm{I}) \asymp n^{-1/(d+5)}$ and $h_n(\mathrm{I}) \asymp n^{-1/(d+5)}$. By Lemma 6.4, we have

$$\log\{\mathrm{RSS}_n(\mathscr{X}, \mathscr{T}, \mathrm{I})/n\} = \log\left(n^{-1} \sum_{i \in \mathscr{I}_n} e_i^2 \mathbb{1}_{\{\mathbf{x}_i \in \mathscr{X}\}}\right) + O_p\{n^{-7/(2d+10)}\}.$$

Under the stated conditions on the tuning parameter, we have $n^{-7/(2d+10)} = o\{\tau_n(b_n h_n^d)^{-1}\}$, and thus the estimation error is dominated by $\tau_n \mathrm{DF}(\mathrm{I})$ which goes to zero as $n \to \infty$. By Lemmas 6.5–6.7, similar results can be derived for models II–IV. Note that

$$\tau_n \max\{\mathrm{DF}(\mathrm{I}), \mathrm{DF}(\mathrm{II}), \mathrm{DF}(\mathrm{III}), \mathrm{DF}(\mathrm{IV})\} = o(1),$$

which will be dominated by any model misspecification. The result follows by $\mathrm{DF}(\mathrm{IV}) < \min\{\mathrm{DF}(\mathrm{II}), \mathrm{DF}(\mathrm{III})\} \leq \max\{\mathrm{DF}(\mathrm{II}), \mathrm{DF}(\mathrm{III})\} < \mathrm{DF}(\mathrm{I})$. □

SUPPLEMENTARY MATERIAL

Additional technical proofs (DOI: [10.1214/14-AOS1299SUPP](); .pdf). This supplement contains technical proofs of Lemmas 6.6 and 6.7, Proposition 3.1 and Theorems 3.6 and 3.7.

REFERENCES

AÏT-SAHALIA, Y. (1996). Testing continuous-time models of the spot interest rate. *Rev. Finan. Stud.* **9** 385–426.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) (B. N. Petrov and F. Csaki, eds.) 267–281. Akadémiai Kiadó, Budapest. MR0483125

ANDREWS, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* **61** 821–856. MR1231678

ANDREWS, D. W. K. (1995). Nonparametric kernel estimation for semiparametric models. *Econometric Theory* **11** 560–596. MR1349935

AZZALINI, A. and BOWMAN, A. (1993). On the use of nonparametric regression for checking linear relationships. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **55** 549–557. MR1224417

BLACK, F. and SCHOLES, M. (1973). The pricing of options and corporate liabilities. *J. Polit. Economy* **81** 637–654.

BOSQ, D. (1996). *Nonparametric Statistics for Stochastic Processes*: *Estimation and Prediction*. *Lecture Notes in Statistics* **110**. Springer, New York. MR1441072

BROWN, R. L., DURBIN, J. and EVANS, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **37** 149–192. MR0378310

CAI, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *J. Econometrics* **136** 163–188. MR2328589

CASTELLANA, J. V. and LEADBETTER, M. R. (1986). On smoothed probability density estimation for stationary processes. *Stochastic Process. Appl.* **21** 179–193. MR0833950

CHAN, K. C., KAROLYI, A. G., LONGSTAFF, F. A. and SANDERS, A. B. (1992). An empirical comparison of alternative models of the short-term interest rate. *J. Finance* **47** 1209–1227.

CHOW, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica* **28** 591–605. MR0141193

COURTADON, G. (1982). The pricing of options on default-free bonds. *J. Finan. Quant. Anal.* **17** 75–100.

COX, J. C., INGERSOLL, J. E. JR. and ROSS, S. A. (1985). A theory of the term structure of interest rates. *Econometrica* **53** 385–407. MR0785475

DAVIS, R. A., HUANG, D. W. and YAO, Y.-C. (1995). Testing for a change in the parameter values and order of an autoregressive model. *Ann. Statist.* **23** 282–304. MR1331669

DETTE, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Statist.* **27** 1012–1040. MR1724039

DRAGHICESCU, D., GUILLAS, S. and WU, W. B. (2009). Quantile curve estimation and visualization for nonstationary time series. *J. Comput. Graph. Statist.* **18** 1–20. MR2511058

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and YAO, Q. (2003). *Nonlinear Time Series*: *Nonparametric and Parametric Methods*. Springer, New York. MR1964455

FAN, J. and ZHANG, J.-T. (2000a). Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 303–322. MR1749541

FAN, J. and ZHANG, W. (2000b). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Stat.* **27** 715–731. MR1804172

FAN, J., ZHANG, C. and ZHANG, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29** 153–193. MR1833962

FREEDMAN, D. A. (1975). On tail probabilities for martingales. *Ann. Probab.* **3** 100–118. MR0380971

GAO, J. and GIJBELS, I. (2008). Bandwidth selection in nonparametric kernel testing. *J. Amer. Statist. Assoc.* **103** 1584–1594. MR2504206

GONZÁLEZ MANTEIGA, W. and CAO, R. (1993). Testing the hypothesis of a general linear model using nonparametric regression estimation. *TEST* **2** 161–188. MR1265489

GYÖRFI, L., HÄRDLE, W., SARDA, P. and VIEU, P. (1989). *Nonparametric Curve Estimation from Time Series. Lecture Notes in Statistics* **60**. Springer, Berlin. MR1027837

HALL, P., MÜLLER, H.-G. and WU, P.-S. (2006). Real-time density and mode estimation with application to time-dynamic mode tracking. *J. Comput. Graph. Statist.* **15** 82–100. MR2269364

HANSEN, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748. MR2409261

HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 86–101. MR0941001

HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** 1926–1947. MR1245774

HÄRDLE, W. and MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481. MR0811503

HE, C., TERÄSVIRTA, T. and GONZÁLEZ, A. (2009). Testing parameter constancy in stationary vector autoregressive models against continuous change. *Econometric Rev.* **28** 225–245. MR2655626

HOOVER, D. R., RICE, J. A., WU, C. O. and YANG, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85** 809–822. MR1666699

HUANG, J. Z., WU, C. O. and ZHOU, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14** 763–788. MR2087972

HURVICH, C. M., SIMONOFF, J. S. and TSAI, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 271–293. MR1616041

LEYBOURNE, S. J. and MCCABE, B. P. M. (1989). On the distribution of some test statistics for coefficient constancy. *Biometrika* **76** 169–177. MR0991435

LI, Q. and RACINE, J. S. (2007). *Nonparametric Econometrics*: *Theory and Practice*. Princeton Univ. Press, Princeton, NJ. MR2283034

LIN, C.-F. J. and TERÄSVIRTA, T. (1999). Testing parameter constancy in linear models against stochastic stationary parameters. *J. Econometrics* **90** 193–213. MR1703341

LIU, W. and WU, W. B. (2010). Simultaneous nonparametric inference of time series. *Ann. Statist.* **38** 2388–2421. MR2676893

LIU, W., XIAO, H. and WU, W. B. (2013). Probability and moment inequalities under dependence. *Statist. Sinica* **23** 1257–1272. MR3114713

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MASRY, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *J. Time Series Anal.* **17** 571–599. MR1424907

MERTON, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *J. Finan. Econ.* **3** 125–144.

NABEYA, S. and TANAKA, K. (1988). Asymptotic theory of a test for the constancy of regression coefficients against the random walk alternative. *Ann. Statist.* **16** 218–235. MR0924867

NEUMANN, M. H. (1998). Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. *Ann. Statist.* **26** 2014–2048. MR1673288

NEUMANN, M. H. and KREISS, J.-P. (1998). Regression-type inference in nonparametric autoregression. *Ann. Statist.* **26** 1570–1613. MR1647701

NYBLOM, J. (1989). Testing for the constancy of parameters over time. *J. Amer. Statist. Assoc.* **84** 223–230. MR0999682

OPSOMER, J., WANG, Y. and YANG, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.* **16** 134–153. MR1861070

ORBE, S., FERREIRA, E. and RODRIGUEZ-POO, J. (2005). Nonparametric estimation of time varying parameters under shape restrictions. *J. Econometrics* **126** 53–77. MR2118278

ORBE, S., FERREIRA, E. and RODRIGUEZ-POO, J. (2006). On the estimation and testing of time varying constraints in econometric models. *Statist. Sinica* **16** 1313–1333. MR2327493

PARK, B. U. and MARRON, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85** 66–72.

PELIGRAD, M. (1992). Properties of uniform consistency of the kernel estimators of density and of regression functions under dependence assumptions. *Stochastics Stochastics Rep.* **40** 147–168. MR1275130

PLOBERGER, W., KRÄMER, W. and KONTRUS, K. (1989). A new test for structural stability in the linear regression model. *J. Econometrics* **40** 307–318. MR0994952

PRIESTLEY, M. B. and CHAO, M. T. (1972). Non-parametric function fitting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **34** 385–392. MR0331616

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

ROBINSON, P. M. (1983). Nonparametric estimators for time series. *J. Time Series Anal.* **4** 185–207. MR0732897

RUPPERT, D., SHEATHER, S. J. and WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270. MR1379468

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHAO, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. MR1466682

SHAO, X. and WU, W. B. (2007). Asymptotic spectral theory for nonlinear time series. *Ann. Statist.* **35** 1773–1801. MR2351105

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London. MR0848134

STANTON, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. *J. Finance* **52** 1973–2002.

TIBSHIRANI, R. J. and TIBSHIRANI, R. (2009). A bias correction for the minimum error rate in cross-validation. *Ann. Appl. Stat.* **3** 822–829. MR2750683

TJØSTHEIM, D. (1994). Non-linear time series: A selective review. *Scand. J. Stat.* **21** 97–130. MR1294588

VASICEK, O. (1977). An equilibrium characterization of the term structure. *J. Finan. Econ.* **5** 177–188.

VOGT, M. (2012). Nonparametric regression for locally stationary time series. *Ann. Statist.* **40** 2601–2633. MR3097614

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. *Monographs on Statistics and Applied Probability* **60**. Chapman & Hall, London. MR1319818

WANG, Y. D. (1998). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* **93** 341–348.

WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154 (electronic). MR2172215

WU, W. B., HUANG, Y. and HUANG, Y. (2010). Kernel estimation for time series: An asymptotic theory. *Stochastic Process. Appl.* **120** 2412–2431. MR2728171

XIA, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 797–811. MR1649488

YANG, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92** 937–950. MR2234196

YU, B. (1993). Density estimation in the L^{∞}-norm for dependent data with applications to the Gibbs sampler. *Ann. Statist.* **21** 711–735. MR1232514

ZHANG, C. and DETTE, H. (2004). A power comparison between nonparametric regression tests. *Statist. Probab. Lett.* **66** 289–301. MR2045474

ZHANG, T. and WU, W. B. (2011). Testing parametric assumptions of trends of a nonstationary time series. *Biometrika* **98** 599–614. MR2836409

ZHANG, T. and WU, W. B. (2012). Inference of time-varying regression models. *Ann. Statist.* **40** 1376–1402. MR3015029

ZHANG, T. and WU, W. B. (2015). Supplement to "Time-varying nonlinear regression models: Nonparametric estimation and model selection." DOI:10.1214/14-AOS1299SUPP.

ZHAO, Z. (2008). Parametric and nonparametric models and methods in financial econometrics. *Stat. Surv.* **2** 1–42. MR2520979

ZHENG, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *J. Econometrics* **75** 263–289. MR1413644

ZHOU, Z. and WU, W. B. (2010). Simultaneous inference of linear models with time varying coefficients. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 513–531. MR2758526

DEPARTMENT OF MATHEMATICS AND STATISTICS
BOSTON UNIVERSITY
BOSTON, MASSACHUSETTS 02215
USA
E-MAIL: tingz@bu.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: wbwu@galton.uchicago.edu