THE UNIVERSITY OF ADELAIDE AUSTRALIA

SUB CRUCE LUMEN

The University of Adelaide
School of Economics

Research Paper No. 2009-01
January 2009

Bandwidth Selection in Nonparametric Kernel Testing

Jiti Gao and Irene Gijbels

Bandwidth Selection in Nonparametric Kernel Testing [1]

By Jiti Gao and Irène Gijbels

The University of Adelaide and Katholieke Universiteit Leuven

Abstract. We propose a sound approach to bandwidth selection in nonparametric kernel testing. The main idea is to find an Edgeworth expansion of the asymptotic distribution of the test concerned. Due to the involvement of a kernel bandwidth in the leading term of the Edgeworth expansion, we are able to establish closed–form expressions to explicitly represent the leading terms of both the size and power functions and then determine how the bandwidth should be chosen according to certain requirements for both the size and power functions. For example, when a significance level is given, we can choose the bandwidth such that the power function is maximized while the size function is controlled by the significance level. Both asymptotic theory and methodology are established. In addition, we develop an easy implementation procedure for the practical realization of the established methodology and illustrate this on two simulated examples and a real data example.

Keywords: Choice of bandwidth parameter, Edgeworth expansion, nonparametric kernel testing, power function, size function.

[1]Jiti Gao is Professor of Economics, School of Economics, The University of Adelaide, Adelaide SA 5005, Australia (E–mail: *jiti.gao@adelaide.edu.au*). Irene Gijbels is Professor of Statistics, Department of Mathematics, University of Leuven, Celestijnenlaan 200B, B-3001, Leuven, Belgium (E–mail: *irene.gijbels@wis.kuleuven.be*). The first author would like to thank Jiying Yin for his excellent computing assistance and the Australian Research Council Discovery Grants under Grant Numbers: DP0558602 and DP0879088 for the financial support. The second author thanks the School of Mathematics and Statistics at the University of Western Australia for its kind hospitality and support. The second author also gratefully acknowledges the Research Fund K.U.Leuven (GOA/2007/4) and the Flemish Science Foundation, FWO Belgium (Project G.0328.08) for financial support.

1. Introduction

Consider a nonparametric regression model of the form

$$Y_i = m(X_i) + e_i, \; i = 1, 2, \ldots, n, \tag{1.1}$$

where $\{X_i\}$ is a sequence of strictly stationary time series variables, $\{e_i\}$ is a sequence of independent and identically distributed (i.i.d.) errors with $E[e_1] = 0$ and $0 < E[e_1^2] = \sigma^2 < \infty$, $m(\cdot)$ is an unknown function defined over $I\!R^d$ for $d \geq 1$, and n is the number of observations. We assume that $\{X_i\}$ and $\{e_j\}$ are independent for all $1 \leq i \leq j \leq n$.

To avoid the so–called "curse of dimensionality" problem, we mainly consider the case of $1 \leq d \leq 3$ in this paper. For the case of $d \geq 4$, various dimension reduction estimation and specification methods have been discussed extensively in several monographs, such as Fan and Gijbels (1996), Hart (1997), Fan and Yao (2003), Gao (2007), and Li and Racine (2007).

There is a vast literature on testing a parametric regression model (null hypothesis) versus a nonparametric model, especially for the case of i.i.d. X_i's (random or fixed design case). Many goodness-of-fit testing procedures are based on evaluating a distance between a parametric estimate of the regression function m (assuming the null hypothesis is true) and a nonparametric estimate of that function. Among the popular choices for a nonparametric kernel estimator for m are the Nadaraya-Watson estimator, the Gasser-Müller estimator and a local linear (polynomial) estimator. Earlier papers following this approach of evaluating such a distance include Härdle and Mammen (1993), Weihrather (1993) and González-Manteiga and Cao (1993), among others. Härdle and Mammen (1993) consider a weighted L_2-distance between a parametric estimator and a nonparametric Nadaraya-Watson estimator of the regression function. The asymptotic distribution of their test statistic under the null hypothesis depends on the unknown error variance (the conditional error variance function). Weihrather (1993) instead uses a Gasser-Müller nonparametric estimator in the fixed design regression case, divides by an estimator of the error variance and considers a discretized version of the L_2-distance. González-Manteiga and Cao (1993) also consider the fixed design regression case but rely on minimum distance estimation of the parametric model, seeking for minimizing a weighted L_2-type distance between the parametric model and a pilot nonparametric estimator.

Another approach to the same testing problem is introduced in Dette (1999) who focusses on the integrated conditional variance function, and uses as a test statistic the difference of a parametric estimator and a nonparametric (Nadaraya-Watson based) estimator of this integrated variance. It is shown that this estimator (asymptotically) corresponds to test statistics based on a weighted L_2-distance between a parametric and nonparametric estimator of the regression function, as in the above mentioned papers, using an appropriate weight function in defining the L_2-distance. Dette (1999) studies the asymptotic distribution of the test statistic under fixed alternatives. Such kind of alternatives are to be distinguished from the so-called sequences of local alternatives, where the difference between the regression function under the alternative and the one under the null hypothesis depends on the sample size n and decreases with n. The latter setup is the one considered in our study.

The above papers and several more recent goodness-of-fit tests (see for example Zhang and Dette (2004) and references therein) have in common that they rely on nonparametric kernel type regression estimators and that the resulting test statistics are of a similar form (at least in first-order asymptotics), and all depend on a bandwidth parameter. The choice of the bandwidth parameter in such goodness-of-fit testing procedures is the main concern in the present paper. Roughly speaking one can distinguish in the literature two approaches to deal with this bandwidth parameter choice in nonparametric and semiparametric kernel methods used for constructing model specification tests for the mean function of model (1.1). A first approach is to use an estimation-based optimal bandwidth value, such as a cross–validation bandwidth. A second approach is to consider a set of suitable values for the bandwidth and proceed further from there.

Existing studies based on the first approach include Härdle and Mammen (1993) for testing nonparametric regression with i.i.d. designs and errors, Hjellvik and Tjøstheim (1995), and Hjellvik, Yao and Tjøstheim (1998) for testing linearity in dependent time series cases, Li (1999) for specification testing in econometric time series cases, Chen, Härdle and Li (2003) for using empirical likelihood–based tests, Juhl and Xiao (2005) for testing structural change in nonparametric time series regression, and others. As pointed out in the literature, such choices cannot be justified in both theory and practice since estimation–based optimal values may not be optimal for testing purposes.

Nonparametric tests involving the second approach of choising either a set of suitable bandwidth values for the kernel case or a sequence of positive integers for the smoothing spline case include Fan (1996), Fan, Zhang and Zhang (2001), and Horowitz and Spokoiny (2001). The practical implementation of choosing such sets or sequences is however problematic. This is probably why Horowitz and Spokoiny (2001) develop their theoretical results based on a set of suitable bandwidths on the one hand, but choose their practical bandwidth values based on the assessment of the power function of their test on the other hand. Apart from using such test statistics based on nonparametric kernel, nonparametric series, spline smoothing and wavelet methods, there are test statistics constructed and studied based on empirical distributions. Such studies have recently been summarized in Zhu (2005).

To the best of our knowledge, the idea of choosing the appropriate smoothing parameter such that the size of the test under consideration is preserved while maximizing the power against a given alternative was only first explored analytically by Kulasekera and Wang (1997), in which the authors propose using a nonparametric kernel test to check whether the mean functions of two data sets can be identical in a nonparametric fixed design setting. In some other closely related studies, various discussions have been given on the comparison of power values of the same test at different bandwidths or different tests at the same bandwidth. Such studies include Hart (1997), Hjellvik, Yao and Tjøstheim (1998), Hunsberger and Follmann (2001), and Zhang and Dette (2004). The last paper compares three main types of nonparametric kernel tests proposed in Härdle and Mammen (1993), Zheng (1996), and Fan, Zhang and Zhang (2001).

On the issue of size correction, there have recently been some studies. For example, Fan and Linton (2003) develop an Edgeworth expansion for the size function of their test and then propose using corrected asymptotic critical values to improve the small–medium sample size properties of the size of their test. Some other related studies include Nishiyama and Robinson (2000), Horowitz (2003), Nishiyama and Robinson (2005), who develop some useful Edgeworth expansions for bootstrap distributions of partial–sum type of tests for improving the size performance.

The current paper is motivated by such existing studies, especially by Kulasekera and Wang (1997), Fan and Linton (2003), Dette and Spreckelsen (2004), and Zhang and Dette (2004), to develop a solid theory to support a power function–based bandwidth selection procedure such that the power of the proposed test is maximized while the size is under control when using nonparametric kernel testing in parametric specification of a nonparametric regression model of the form (1.1) associated with the hypothesis form of (1.2) below.

To state the main results of this paper, we introduce some notational details. The main interest of this paper is to test a parametric null hypothesis of the form

$$\begin{aligned} \mathcal{H}_0: \ m(x) &= m_{\theta_0}(x) \quad \text{versus a sequence of alternatives of the form} \\ \mathcal{H}_1: \ m(x) &= m_{\theta_1}(x) + \Delta_n(x) \ \ \text{for all } x \in I\!\!R^d, \end{aligned} \tag{1.2}$$

where both $\theta_0, \ \theta_1 \in \Theta$ are unknown parameters and Θ is a parameter space of $I\!\!R^p$, and $\Delta_n(x)$ is a sequence of nonparametrically unknown functions over $I\!\!R^d$. With $\Delta_n(x)$ not being equal to zero, the function $m_{\theta_1}(x)$ in $\mathcal{H}_1$ is in fact the projection of the true function on the null model.

Note that $m(x)$ under $\mathcal{H}_1$ in (1.2) is semiparametric when $\{\Delta_n(x)\}$ is unknown nonparametrically. Note also that instead of requiring (1.2) for all $x \in I\!\!R^d$, it may be assumed that (1.2) holds with probability one for $x = X_i$. Some first–order asymptotic properties for both the size and power functions of a nonparametric kernel test for the case where $\Delta_n(\cdot) \equiv \Delta(\cdot)$, corresponding to a class of fixed alternatives (not depending on n), have already been discussed in the literature, such as Dette and Spreckelsen (2004). This paper focuses on studying higher–order asymptotic properties of such kernel tests for the case where $\{\Delta_n(\cdot)\}$ is a sequence of local alternatives in the sense that $\lim_{n\to\infty} \Delta_n(x) = 0$ for all $x \in I\!\!R^d$.

Let $K(\cdot)$ be the probability kernel density function and h be the bandwidth involved in the construction of a nonparametric kernel test statistic denoted by $\widehat{T}_n(h)$. To implement the kernel test in practice, we propose a new bootstrap simulation procedure to approximate the $1-\alpha$ quantile of the distribution of the kernel test by a bootstrap simulated critical value l_α. Let $\alpha_n(h) = P\left(\widehat{T}_n(h) > l_\alpha | \mathcal{H}_0\right)$ and $\beta_n(h) = P\left(\widehat{T}_n(h) > l_\alpha | \mathcal{H}_1\right)$ be the respective size and power functions. In Theorem 2.2 we show that

$$\alpha_n(h) = 1 - \Phi(l_\alpha - s_n) - \kappa_n \left(1 - (l_\alpha - s_n)^2\right) \phi(l_\alpha - s_n) + o\left(\sqrt{h^d}\right), \quad (1.3)$$

$$\beta_n(h) = 1 - \Phi(l_\alpha - r_n) - \kappa_n \left(1 - (l_\alpha - r_n)^2\right) \phi(l_\alpha - r_n) + o\left(\sqrt{h^d}\right), \quad (1.4)$$

where $s_n = p_1 \sqrt{h^d}$, $r_n = p_2\ n\ \delta_n^2 \sqrt{h^d}$, $\kappa_n = p_3 \sqrt{h^d}$, and $\Phi(\cdot)$ and $\phi(\cdot)$ denote respectively the cumulative distribution and density function of the standard Normal random variable, in which all p_i's are positive constants and $\delta_n^2 = \int \Delta_n^2(x)\pi^2(x)dx$ with $\pi(\cdot)$ being the marginal density function of $\{X_i\}$.

Our aim is to choose a bandwidth $h_{\rm ew}$ such that $\beta_n(h_{\rm ew}) = \max_{h\in H_n(\alpha)} \beta_n(h)$ with $H_n(\alpha) = \{h :\ \alpha - c_{\rm min} < \alpha_n(h) < \alpha + c_{\rm min}\}$ for some small $0 < c_{\rm min} < \alpha$. Our detailed study in Section 3 shows that $h_{\rm ew}$ is proportional to $(n\ \delta_n^2)^{-\frac{3}{2d}}$. Such established relationship between δ_n and $h_{\rm ew}$ shows us that the choice of an optimal rate of $h_{\rm ew}$ depends on that of an order of δ_n.

If δ_n is chosen proportional to $n^{-\frac{d+12}{6(d+4)}}$ for a sequence of local alternatives under $\mathcal{H}_1$, then the optimal rate of h_{ew} is proportional to $n^{-\frac{1}{d+4}}$, which is the order of a nonparametric cross–validation estimation–based bandwidth frequently used for testing purposes. When considering a sequence of local alternatives with $\delta_n = O\left(n^{-\frac{1}{2}}\sqrt{\text{loglog}n}\right)$ being chosen as the optimal rate for testing in this kind of kernel testing (Horowitz and Spokoiny 2001), the optimal rate of h_{ew} is proportional to $(\text{loglog}n)^{-\frac{3}{2d}}$.

The rest of the paper is organised as follows. Section 2 points out that existing nonparametric kernel tests can be decomposed with quadratic forms of $\{e_i\}$ as leading terms in the decomposition. This motivates the discussion about establishing Edgeworth expansions for such quadratic forms. In Section 3, we apply the Edgeworth expansions to study both the size and power functions of a representative kernel test. Section 4 presents several examples of implementation. Some concluding remarks are made in Section 5. Mathematical assumptions and proofs are provided in the appendix.

2. Nonparametric kernel testing

As mentioned in the introductory section, various authors have discussed and studied nonparametric kernel test statistics based on a (weighted) L_2–distance function between a nonparametric kernel estimator and a parametric counterpart of the mean function. It can be shown that the leading term of each of these nonparametric kernel test statistics is of a quadratic form (see, for example, Chen, Härdle and Li 2003)

$$P_n(h) = \sum_{i=1}^{n}\sum_{j=1}^{n} e_i \; w(X_i)L_h(X_i - X_j)w(X_j) \; e_j, \tag{2.1}$$

where $L_h(\cdot) = \frac{1}{n\sqrt{h^d}} L\left(\frac{\cdot}{h}\right)$, $L(x) = \int K(y)K(x+y)dy$, and $w(\cdot)$ is a suitable weight function probably depending on either $\pi(\cdot)$, $\sigma^2(\cdot)$ or both, in which $K(\cdot)$ is a probability kernel function, h is a bandwidth parameter and both are involved in a nonparametric kernel estimation of $m(\cdot)$.

In this paper, we concentrate on a second group of nonparametric kernel test statistics using a different distance function. Rewrite model (1.1) into a notational version of the form under $\mathcal{H}_0$

$$Y = m_{\theta_0}(X) + e, \tag{2.2}$$

where X is assumed to be random and θ_0 is the true value of θ under $\mathcal{H}_0$. Obviously, $E[e|X] = 0$ under $\mathcal{H}_0$. Existing studies (Zheng 1996; Li and Wang 1998; Li 1999; Fan and Linton 2003; Dette and Spreckelsen 2004; Juhl and Xiao 2005) propose using a distance function of the form

$$E\left[eE\left(e|X\right)\pi(X)\right] = E\left[\left(E^2(e|X)\right)\pi(X)\right], \tag{2.3}$$

where $\pi(\cdot)$ is the marginal density function of X.

This suggests using a normalized kernel–based sample analogue of (2.3) of the form

$$T_n(h) = \frac{1}{n\sqrt{h^d}\ \sigma_n}\sum_{i=1}^{n}\sum_{j=1,\neq i}^{n} e_i\ K\left(\frac{X_i - X_j}{h}\right)\ e_j, \tag{2.4}$$

where $\sigma_n^2 = 2\mu_2^2\ \nu_2 \int K^2(u)du$ with $\mu_k = E[e_1^k]$ for $k \geq 1$ and $\nu_l = E[\pi^l(X_1)]$ for $l \geq 1$.

It can easily be seen that $T_n(h)$ is the leading term of the following quadratic form

$$Q_n(h) = \frac{1}{n\sqrt{h^d}\ \sigma_n}\sum_{i=1}^{n}\sum_{j=1}^{n} e_i\ K\left(\frac{X_i - X_j}{h}\right)\ e_j. \tag{2.5}$$

In summary, both equations (2.1) and (2.5) can be generally written as

$$R_n(h) = \sum_{i=1}^{n}\sum_{j=1}^{n} e_i\ \phi_n(X_i, X_j)\ e_j, \tag{2.6}$$

where $\phi_n(\cdot,\cdot)$ may depend on n, the bandwidth h and the kernel function K.

Thus, it is of general interest to study asymptotic distributions and their Edgeworth expansions for quadratic forms of type (2.6). To present the main ideas of establishing Edgeworth expansions for such quadratic forms, we focus on $T_n(h)$ in the rest of this paper. This is because the main technology for establishing an Edgeworth expansion for the asymptotic distribution of each of such tests is the same as that for $T_n(h)$.

Since $T_n(h)$ involves some unknown quantities, we estimate it by a stochastically normalized version of the form

$$\widehat{T}_n(h) = \frac{\sum_{i=1}^n \sum_{j=1,\neq i}^n \widehat{e}_i \ K\left(\frac{X_i - X_j}{h}\right) \ \widehat{e}_j}{n\sqrt{h^d} \ \widehat{\sigma}_n}, \tag{2.7}$$

where $\widehat{e}_i = Y_i - m_{\widehat{\theta}}(X_i)$ and $\widehat{\sigma}_n^2 = 2\widehat{\mu}_2^2 \ \widehat{\nu}_2 \int K^2(u)du$ with $\widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n \widehat{e}_i^2$ and $\widehat{\nu}_2 = \frac{1}{n}\sum_{i=1}^n \widehat{\pi}^2(X_i)$, in which $\widehat{\theta}$ is a $\sqrt{n}$–consistent estimator of θ_0 under $\mathcal{H}_0$ and $\widehat{\pi}(x) = \frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)$ is the conventional nonparametric kernel density estimator with $\widehat{b}_{\rm cv}$ being a bandwidth parameter chosen by cross–validation (see for example Silverman 1986).

Similarly to existing results (Li 1999), it may be shown that for each given h

$$\widehat{T}_n(h) = T_n(h) + o_P\left(\sqrt{h^d}\right). \tag{2.8}$$

Thus, we may use the distribution of $\widehat{T}_n(h)$ to approximate that of $T_n(h)$. Let l_α^e $(0 < \alpha < 1)$ be the $1-\alpha$ quantile of the exact finite–sample distribution of $\widehat{T}_n(h)$. Because l_α^e may not be evaluated in practice, we therefore suggest choosing either a non–random approximate α–level critical value, l_α, or a stochastic approximate α–level critical value, l_α^* by using the following simulation procedure:

- We generate $Y_i^* = m_{\widehat{\theta}}(X_i) + \sqrt{\widehat{\mu}_2} \ e_i^*$ for $1 \le i \le n$, where $\{e_i^*\}$ is a sequence of i.i.d. random samples drawn from a pre-specified distribution, such as $N(0,1)$. Use the data set $\{(X_i, Y_i^*) : i = 1, 2, \ldots, n\}$ to estimate $\widehat{\theta}$ by $\widehat{\theta}^*$ and compute $\widehat{T}_n(h)$. Let l_α be the $1-\alpha$ quantile of the distribution of

$$\widehat{T}_n^*(h) = \frac{\sum_{i=1}^n \sum_{j=1,\neq i}^n \widehat{e}_i^* \ K\left(\frac{X_i - X_j}{h}\right) \ \widehat{e}_j^*}{n\sqrt{h^d} \ \widehat{\sigma}_n^*}, \tag{2.9}$$

  where $\widehat{e}_i^* = Y_i^* - m_{\widehat{\theta}^*}(X_i)$ and $\widehat{\sigma}_n^{*2} = 2\widehat{\mu}_2^{*2} \ \widehat{\nu}_2 \int K^2(u)du$ with $\widehat{\mu}_2^* = \frac{1}{n}\sum_{i=1}^n \widehat{e}_i^{*2}$.

  In the simulation process, the original sample $\mathcal{X}_n = (X_1, \cdots, X_n)$ acts in the resampling as a fixed design even when $\{X_i\}$ is a sequence of random regressors.

- Repeat the above step M times and produce M versions of $\widehat{T}_n^*(h)$ denoted by $\widehat{T}_{n,m}^*(h)$ for $m = 1, 2, \ldots, M$. Use the M values of $\widehat{T}_{n,m}^*(h)$ to construct their empirical distribution function. The bootstrap distribution of $\widehat{T}_n^*(h)$ given $\mathcal{W}_n = \{(X_i, Y_i) : 1 \leq i \leq n\}$ is defined by $P^*\left(\widehat{T}_n^*(h) \leq x\right) = P\left(\widehat{T}_n^*(h) \leq x | \mathcal{W}_n\right)$. Let l_α^* $(0 < \alpha < 1)$ satisfy $P^*\left(\widehat{T}_n^*(h) \geq l_\alpha^*\right) = \alpha$ and then estimate l_α by l_α^*.

Note that both $l_\alpha = l_\alpha(h)$ and $l_\alpha^* = l_\alpha^*(h)$ depend on h. It should be pointed out that the choice of a pre–specified distribution does not have much impact on both the theoretical and practical results. In addition, we may also use a wild bootstrap procedure to generate a sequence of resamples for $\{e_i^*\}$.

Note also that the above simulation is based on the so–called regression bootstrap simulation procedure discussed in the literature, such as Li and Wang (1998), Franke, Kreiss and Mammen (2002), and Li and Racine (2007). When $X_i = Y_{i-1}$, we may also use a recursive simulation procedure, commonly-used in the literature. See for example, Hjellvik and Tjøstheim (1995), and Franke, Kreiss and Mammen (2002).

Since the choice of a simulation procedure does not affect the establishment of our theory, our main results are established based on the proposed simulation procedure. We now have the following results in Theorems 2.1 and 2.2; their proofs are provided in the appendix.

Theorem 2.1. *Suppose that Assumptions A.1 and A.2 listed in the appendix hold. Then under* $\mathcal{H}_0$

$$\sup_{x \in R^1} \left| P^*(\widehat{T}_n^*(h) \leq x) - P(\widehat{T}_n(h) \leq x) \right| = O\left(\sqrt{h^d}\right) \tag{2.10}$$

holds in probability with respect to the joint distribution of $\mathcal{W}_n$, *and*

$$P\left(\widehat{T}_n(h) > l_\alpha^*\right) = \alpha + O\left(\sqrt{h^d}\right). \tag{2.11}$$

For an equivalent test, Li and Wang (1998) establish some results weaker than (2.10). Fan and Linton (2003) consider some higher–order approximations to the size function of the test discussed in Li and Wang (1998).

For each h we define the following size and power functions

$$\alpha_n(h) = P\left(\widehat{T}_n(h) > l_\alpha | \mathcal{H}_0\right) \quad \text{and} \quad \beta_n(h) = P\left(\widehat{T}_n(h) > l_\alpha | \mathcal{H}_1\right). \tag{2.12}$$

Correspondingly, we define $(\alpha_n^*(h), \beta_n^*(h))$ with l_α replaced by l_α^*.

Before we discuss how to choose an optimal bandwidth in Section 3, we give Edgeworth expansions of both the size and power functions in Theorem 2.2 below. In order to express the Edgeworth expansions, we need to introduce the following notation. Let

$$\kappa_n = \frac{\sqrt{h^d}\left(\frac{\mu_3^2 K^2(0)}{nh^d} + \frac{4\mu_2^3\nu_3}{3}K^{(3)}(0)\right)}{\sigma_n^3}, \tag{2.13}$$

where $\nu_l = E[\pi^l(X_1)] = \int \pi^{l+1}(x)dx$, and $K^{(3)}(\cdot)$ is the three–time convolution of $K(\cdot)$ with itself.

Theorem 2.2. (i) *Suppose that Assumptions A.1 and A.2 listed in the appendix hold. Then*

$$\begin{aligned} \alpha_n(h) &= 1 - \Phi(l_\alpha - s_n) - \kappa_n\ (1-(l_\alpha - s_n)^2)\ \phi(l_\alpha - s_n) + o\left(\sqrt{h^d}\right), & (2.14)\\ \alpha_n^*(h) &= 1 - \Phi(l_\alpha^* - s_n) - \kappa_n\ (1-(l_\alpha^* - s_n)^2)\ \phi(l_\alpha^* - s_n) + o\left(\sqrt{h^d}\right) & (2.15)\end{aligned}$$

hold in probability with respect to the joint distribution of $\mathcal{W}_n$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the probability distribution and density functions of $N(0,1)$, respectively, and $s_n = C_0(m)\sqrt{h^d}$ with

$$C_0(m) = \frac{\int \left(\frac{\partial m_{\theta_0}(x)}{\partial\theta}\right)^\tau \left(E\left[\left(\frac{m_{\theta_0}(X_1)}{\partial\theta}\right)\left(\frac{m_{\theta_0}(X_1)}{\partial\theta}\right)^\tau\right]\right)^{-1}\left(\frac{m_{\theta_0}(x)}{\partial\theta}\right)\pi^2(x)dx}{\sqrt{2\nu_2\ \int K^2(v)dv}}.$$

(ii) *Suppose that Assumptions A.1–A.3 listed in the appendix hold. Then the following equations hold in probability with respect to the joint distribution of $\mathcal{W}_n$:*

$$\begin{aligned} \beta_n(h) &= 1 - \Phi(l_\alpha - r_n) - \kappa_n\ (1-(l_\alpha - r_n)^2)\ \phi(l_\alpha - r_n) + o\left(\sqrt{h^d}\right), & (2.16)\\ \beta_n^*(h) &= 1 - \Phi(l_\alpha^* - r_n) - \kappa_n\ (1-(l_\alpha^* - r_n)^2)\ \phi(l_\alpha^* - r_n) + o\left(\sqrt{h^d}\right), & (2.17)\end{aligned}$$

where $r_n = n\ C_n^2\ \sqrt{h^d}$, *in which*

$$C_n^2 = \frac{\int \Delta_n^2(x)\pi^2(x)dx}{\sigma^2\sqrt{2\nu_2\ \int K^2(v)dv}}. \tag{2.18}$$

Assumption A.2 implies that the random quantity $C_0(m)$ is bounded in probability. As expected, the rate of r_n depends on the form of $\Delta_n(\cdot)$.

To simplify the following expressions, let z_α be the $1-\alpha$ quantile of the standard normal distribution and $d_j = (z_\alpha^2 - 1)c_j$ for $j = 1, 2$, where

$$c_1 = \frac{4K^{(3)}(0)\mu_2^3\nu_3}{3\sigma_n^3} \quad \text{and} \quad c_2 = \frac{\mu_3^2 K^2(0)}{\sigma_n^3}. \tag{2.19}$$

Let $d_0 = d_1 - C_0(m)$. A corollary of Theorem 2.2 is given in Theorem 2.3 below.

Theorem 2.3. *Suppose that the conditions of Theorem 2.2(i) hold. Then under* $\mathcal{H}_0$

$$l_\alpha \approx z_\alpha + d_0\sqrt{h^d} + d_2\ \frac{1}{n\sqrt{h^d}} \quad \text{in probability,} \tag{2.20}$$

$$l_\alpha^* \approx z_\alpha + d_0\sqrt{h^d} + d_2\ \frac{1}{n\sqrt{h^d}} \quad \text{in probability.} \tag{2.21}$$

Theorem 2.3 shows that the size distortion of the proposed test is $d_0\sqrt{h^d} + d_2\ \frac{1}{n\sqrt{h^d}}$ when using the standard asymptotic normality in practice. A similar result has been obtained by Fan and Linton (2003). We show in addition that the bootstrap simulated critical value is approximated explicitly by $z_\alpha + d_0\sqrt{h^d} + d_2\ \frac{1}{n\sqrt{h^d}}$.

As the main objective of this paper, Section 3 below proposes a suitable selection criterion for the choice of h such that while the size function is appropriately controlled, the power function is maximized at such h. A closed–form expression of the power function–based optimal bandwidth is given.

3. Power function–based bandwidth choice

We now employ the Edgeworth expansions established in Section 2 to choose a suitable bandwidth such that the power function $\beta_n(h)$ is maximized while the size function $\alpha_n(h)$ is controlled by a significance level. We thus define

$$h_{\rm ew} = \arg\max_{h \in H_n(\alpha)} \beta_n(h) \quad \text{with} \quad H_n(\alpha) = \{h: \ \alpha - c_{\min} < \alpha_n(h) < \alpha + c_{\min}\} \tag{3.1}$$

for some arbitrarily small $c_{\min} > 0$.

We now start to discuss how to solve the optimization problem (3.1). It follows from (2.13) and (2.19) that

$$\kappa_n = \frac{\sqrt{h^d}\left(\frac{\mu_3^2 K^2(0)}{nh^d} + \frac{4\mu_2^3 \nu_3}{3} K^{(3)}(0)\right)}{\sigma_n^3} = c_1 \sqrt{h^d} + c_2 \frac{1}{n\sqrt{h^d}}. \tag{3.2}$$

Let $x = \sqrt{h^d}$. We rewrite κ_n as $\kappa_n = c_1\, x + c_2\, n^{-1}\, x^{-1}$. Let $\gamma_n = (z_\alpha^2 - 1)\kappa_n$,

$$l_\alpha - r_n \approx z_\alpha + \gamma_n - r_n = z_\alpha + \left(d_1 - n\, C_n^2\right) x + d_4\, x^{-1} \equiv z_\alpha + d_3\, x + d_4\, x^{-1}, \tag{3.3}$$

$$l_\alpha - s_n \approx z_\alpha + \gamma_n - s_n \approx z_\alpha + (d_1 - C_0(m))\, x + d_4\, x^{-1} = z_\alpha + d_0\, x + d_4\, x^{-1}, \tag{3.4}$$

where $d_0 = d_1 - C_0(m)$, $d_1 = (z_\alpha^2 - 1)c_1$, $d_3 = d_1 - n\, C_n^2$ and $d_4 = c_2\, (z_\alpha^2 - 1)\, n^{-1}$. Note that $\lim_{n\to\infty} d_4 = 0$. Since Assumption A.3 implies that $\lim_{n\to\infty} n\, C_n^2 = +\infty$, we thus have

$$\lim_{n\to\infty} d_3 = -\infty \quad \text{when} \quad \lim_{n\to\infty} n\, C_n^2 = +\infty. \tag{3.5}$$

Due to this, we treat d_3 as a sufficiently large negative value when $n\, C_n^2$ is viewed as a sufficiently large positive value in the finite–sample analysis of this section.

Ignoring the higher–order terms (i.e. terms of order $o(x + n^{-1}x^{-1})$ or smaller), we now re–write the power and size functions $\beta_n(h)$ and $\alpha_n(h)$ simply as functions of $x = \sqrt{h^d}$ as follows:

$$\begin{aligned}
\beta_n(h) &\approx 1 - \Phi(l_\alpha - r_n) - \kappa_n\, (1 - (l_\alpha - r_n)^2)\, \phi(l_\alpha - r_n) \\
&\approx 1 - \Phi(z_\alpha + d_3 x + d_4 x^{-1}) - \left(c_1 x + c_2 n^{-1} x^{-1}\right) \\
&\times \left(1 - (z_\alpha + d_3 x + d_4 x^{-1})^2\right) \phi\left(z_\alpha + d_3 x + d_4 x^{-1}\right) \equiv \beta(x), \qquad (3.6) \\
\alpha_n(h) &\approx 1 - \Phi(l_\alpha - s_n) - \kappa_n\, (1 - (l_\alpha - s_n)^2)\, \phi(l_\alpha - s_n)
\end{aligned}$$

$$
\begin{aligned}
&\approx \quad 1-\Phi(z_\alpha+d_0x+d_4x^{-1})-\left(c_1x+c_2n^{-1}x^{-1}\right)\\
&\times \quad \left(1-(z_\alpha+d_0x+d_4x^{-1})^2\right)\phi\left(z_\alpha+d_0x+d_4x^{-1}\right)\equiv\alpha(x). \qquad (3.7)
\end{aligned}
$$

Our objective is then to find $x_{\rm ew}=\sqrt{h_{\rm ew}^d}$ such that

$$
x_{\rm ew}=\arg\max_{x\in H_n(\alpha)}\beta(x) \quad \text{with} \quad H_n(\alpha)=\{x:\ \alpha-c_{\min}<\alpha(x)<\alpha+c_{\min}\}, \qquad (3.8)
$$

where $c_{\min}$ is chosen as $c_{\min}=\frac{\alpha}{10}$ for example. Finding roots of $\beta'(x)=0$ implies that the leading order of the unique real root of the equation is given approximately by

$$
h_{\rm ew}=x_{\rm ew}^{\frac{2}{d}}=a_1^{-\frac{1}{2d}}t_n^{-\frac{3}{2d}}, \qquad (3.9)
$$

where $t_n=n\ C_n^2$, $a_1=\frac{\sqrt{2}K^{(3)}(0)}{3\left(\sqrt{\int K^2(u)du}\right)^3}\cdot c(\pi)$ with $c(\pi)=\frac{\int\pi^3(x)dx}{\left(\sqrt{\int\pi^2(x)dx}\right)^3}$, in which C_n^2 is as defined in Theorem 2.2(ii).

It can also be shown that $h_{\rm ew}$ is the maximizer of the power function $\beta_n(h)$ at $h=h_{\rm ew}$ such that

$$
\beta_n''(x)\big|_{x=\sqrt{h_{\rm ew}^d}}<0\ , \qquad (3.10)
$$

at least for sufficiently large n. Detailed derivations of (3.9) and (3.10) are given in Appendix B below.

Furthermore, the choice of $h_{\rm ew}$ satisfies both Assumptions A.1(v) and A.3 that

$$
\lim_{n\to\infty}n\ h_{\rm ew}^d=+\infty \quad \text{and} \quad \lim_{n\to\infty}n\ \sqrt{h_{\rm ew}^d}\ C_n^2=+\infty.
$$

This implies that the choice of $h_{\rm ew}$ is valid to ensure $\lim_{n\to\infty}\beta_n(h_{\rm ew})=1$.

When both $\sigma^2=\mu_2=E[e_1^2]$ and the marginal density function $\pi(\cdot)$ of $\{X_i\}$ are unknown in practice, we propose using an estimated version of $h_{\rm ew}$ as follows:

$$
\widehat{h}_{\rm ew}=\widehat{a}_1^{-\frac{1}{2d}}\widehat{t}_n^{-\frac{3}{2d}}, \qquad (3.11)
$$

where

$$
\begin{aligned}
\widehat{t}_n &= n\ \widehat{C}_n^2 \quad \text{with} \quad \widehat{C}_n^2=\frac{\frac{1}{n}\sum_{i=1}^n\widehat{\Delta}_n^2(X_i)\widehat{\pi}(X_i)}{\widehat{\mu}_2\sqrt{2\widehat{\nu}_2\ \int K^2(v)dv}} \quad \text{and}\\
\widehat{a}_1 &= \frac{\sqrt{2}K^{(3)}(0)}{3\left(\sqrt{\int K^2(u)du}\right)^3}\ \widehat{c}(\pi) \quad \text{with} \quad \widehat{c}(\pi)=\frac{\frac{1}{n}\sum_{i=1}^n\widehat{\pi}^2(X_i)}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^n\widehat{\pi}(X_i)}\right)^3},
\end{aligned}
$$

in which $\widehat{\mu}_2$, $\widehat{\nu}_2$ and $\widehat{\pi}(\cdot)$ are as defined in (2.7), and $\widehat{\Delta}_n(x)$ is given by

$$\widehat{\Delta}_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\left(Y_i - m_{\widehat{\theta}}(X_i)\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)}$$

with $\widehat{\theta}$ and $\widehat{b}_{\rm cv}$ being the same as in (2.7).

Note also that $\widehat{h}_{\rm ew}$ provides an optimal bandwidth irrespectively of whether one works under the null hypothesis $\mathcal{H}_0$ or under the alternative hypothesis $\mathcal{H}_1$. In other words, it can be used for computing not only the power under an alternative $\mathcal{H}_1$, but also the size under $\mathcal{H}_0$ in each case. Detailed discussion about this is given in Appendix B below.

We conclude this section by summarizing the above discussion into the following proposition; its proof is given in Appendix B below.

Proposition 3.1. *Suppose that Assumptions A.1–A.3 listed in the appendix hold. Additionally, suppose that $\Delta_n(x)$ is continuously differentiable such that*

$$\lim_{n\to\infty} \sup_{x\in D_\pi} \frac{||\Delta_n'(x)||}{|\Delta_n(x)|} \leq C < \infty \quad \textit{and} \quad \lim_{n\to\infty} \inf_{x\in I\!R^d} |\Delta_n(x)| \sqrt{n\widehat{b}_{\rm cv}^d} = \infty \textit{ in probability}$$

for some $C > 0$, where $D_\pi = \{x \in I\!R^d : \pi(x) > 0\}$ and $|| \cdot ||^2$ denotes the Euclidean norm. Then

$$\lim_{n\to\infty} \frac{\beta_n(\widehat{h}_{\rm ew})}{\beta_n(h_{\rm ew})} = 1 \quad \text{in probability.} \tag{3.12}$$

As pointed out in the introduction, implementation of each of existing nonparametric kernel tests involves either a single bandwidth chosen optimally for estimation purposes or a set of bandwidth values. The proposed $\widehat{h}_{\rm ew}$ is chosen optimally for testing purposes. Section 4 below shows how to implement the proposed test based on our bandwidth in practice and compares the finite–sample performance of the proposed choice with that of some closely relevant alternatives in the literature.

4. Examples of implementation

This section presents two simulated examples and one real data example to illustrate the proposed theory and methods in Sections 2 and 3 as well as to make comparisons with some closely relevant alternatives in the literature. Simulated example 4.1 below discusses the finite–sample performance of the proposed test $\widehat{T}_n(\widehat{h}_{\rm ew})$ with that of the alternative version where the test is coupled with a cross–validation (CV) bandwidth choice. Simulated example 4.2 below compares our test with some of the commonly used tests in the literature. Example 4.3 provides a real data example to show that the proposed test makes a clear difference. In the following finite–sample study in Examples 4.1–4.3 below, we consider the case where $\Delta_n(x) = c_n\ \Delta(x)$, in which $\{c_n\}$ is a sequence of positive real numbers satisfying $\lim_{n\to\infty} c_n = 0$ and $\Delta(x)$ is an unknown function not depending on n.

Example 4.1. Consider a nonparametric time series regression model of the form

$$Y_i = \theta_1 X_{i1} + \theta_2 X_{i2} + c_n(X_{i1}^2 + X_{i2}^2) + e_i,\ 1 \le i \le n, \tag{4.1}$$

where $\{e_i\}$ is a sequence of Normal errors and both X_{i1} and X_{i2} are time series variables generated by

$$X_{i1} = \alpha X_{i-1,1} + u_i \quad \text{and} \quad X_{i2} = \beta X_{i-1,2} + v_i,\ 1 \le i \le n \tag{4.2}$$

with $\{u_i\}$ and $\{v_i\}$ being i.i.d. random errors generated independently from Normal distributions as below.

Under $\mathcal{H}_0$, we generate a sequence of observations $\{Y_i\}$ with $\theta_1 = \theta_2 = 1$ as the true parameters, i.e.

$$\mathcal{H}_0 :\ Y_i = X_{i1} + X_{i2} + e_i, \tag{4.3}$$

where $\{e_i\}$ is a sequence of independent and identically distributed random errors generated from $N(0,1)$, and $\{X_{i1}\}$ and $\{X_{i2}\}$ are independently generated from

$$X_{i1} = 0.5X_{i-1,1} + u_i \ \text{ and } \ X_{i2} = 0.5X_{i-1,2} + v_i,\ 1 \le i \le n \tag{4.4}$$

with $X_{01} = X_{02} = 0$ and $\{u_i\}$ and $\{v_i\}$ are sequences of independent and identically distributed random errors and generated independently from a $N(0,1)$.

Under $\mathcal{H}_1$, we are interested in two alternative models of the form

$$\mathcal{H}_1: \; Y_i = X_{i1} + X_{i2} + c_n(X_{i1}^2 + X_{i2}^2) + e_i, \quad e_i \sim N(0,1) \tag{4.5}$$

with c_n being chosen as either $c_{1n} = n^{-\frac{1}{2}}\sqrt{\text{loglog}(n)}$ or $c_{2n} = n^{-\frac{7}{18}}$.

In the testing procedure, the parameters θ_1 and θ_2 in the parametric model are estimated as discussed in Sections 1 and 2.

The reasoning for the above choice of c_{jn} is as follows. The rate of $c_{1n} = n^{-\frac{1}{2}}\sqrt{\text{loglog}(n)}$ should be an optimal rate of testing in this kind of nonparametric kernel testing problem as discussed in Horowitz and Spokoiny (2001). The rate of $c_{2n} = n^{-\frac{7}{18}}$ implies that the optimal bandwidth $\widehat{h}_{\text{ew}}$ in (B.43) with $d = 2$ is proportional to $n^{-\frac{1}{6}}$.

Throughout this example, we choose $K(\cdot)$ as the standard normal density function. Let $\widehat{h}_{\text{cv}}$ be chosen by a cross–validation criterion of the form

$$\widehat{h}_{\text{cv}} = \arg\min_{h \in H_{\text{cv}}} \frac{1}{n}\sum_{i=1}^n \left(Y_i - \widehat{m}_{-i}(X_{i1}, X_{i2}; h)\right)^2 \text{ with } H_{\text{cv}} = \left[n^{-1}, n^{\frac{1}{6}}\right] \tag{4.6}$$

in which

$$\widehat{m}_{-i}(X_{i1}, X_{i2}; h) = \frac{\sum_{l=1,\neq i}^n K\left(\frac{X_{l1}-X_{i1}}{h}\right) K\left(\frac{X_{l2}-X_{i2}}{h}\right) Y_l}{\sum_{l=1,\neq i}^n K\left(\frac{X_{l1}-X_{i1}}{h}\right) K\left(\frac{X_{l2}-X_{i2}}{h}\right)}.$$

Let $\widehat{h}_{0\text{test}}$ be the corresponding version of $\widehat{h}_{\text{ew}}$ in (B.43) and $\widehat{h}_{0\text{cv}}$ be the corresponding version of $\widehat{h}_{\text{cv}}$ in (4.6) both computed under $\mathcal{H}_0$. Since $\{Y_i\}$ under $\mathcal{H}_1$ depends on the choice of c_n, thus the computing of both $\widehat{h}_{\text{ew}}$ of (B.43) and $\widehat{h}_{\text{cv}}$ of (4.6) under $\mathcal{H}_1$ depend on the choice of c_n. Let $\widehat{h}_{j\text{test}}$ be the corresponding versions of $\widehat{h}_{\text{ew}}$ in (B.43) and $\widehat{h}_{j\text{cv}}$ be the corresponding versions of $\widehat{h}_{\text{cv}}$ in (4.6) with $c_n = c_{jn}$ for $j = 1, 2$.

In order to compare the size and power properties of $\widehat{T}_n(h)$ with the most relevant alternatives, we introduce the following simplified notation: for $j = 1, 2$,

$$\begin{aligned} \alpha_{01} &= P\left(\widehat{T}_n\left(\widehat{h}_{0\text{cv}}\right) > l_\alpha^*\left(\widehat{h}_{0\text{cv}}\right) |\mathcal{H}_0\right), \quad \beta_{j1} = P\left(\widehat{T}_n\left(\widehat{h}_{j\text{cv}}\right) > l_\alpha^*\left(\widehat{h}_{0\text{cv}}\right) |\mathcal{H}_1\right), \\ \alpha_{02} &= P\left(\widehat{T}_n\left(\widehat{h}_{0\text{test}}\right) > l_\alpha^*\left(\widehat{h}_{0\text{test}}\right) |\mathcal{H}_0\right), \quad \beta_{j2} = P\left(\widehat{T}_n\left(\widehat{h}_{j\text{test}}\right) > l_\alpha^*\left(\widehat{h}_{0\text{test}}\right) |\mathcal{H}_1\right). \end{aligned}$$

We consider cases where the number of replications of each of the sample versions of α_{0k} and β_{jk} for $j, k = 1, 2$ was $M = 1000$, each with $B = 250$ number of bootstrapping resamples, and the simulations were done for the cases of $n = 250$, 500 and 750. The detailed results at the 1%, 5% and 10% significance level are given in Tables 4.1–4.3, respectively.

Table 4.1. Simulated size and power values at the 1% significance level

Sample Size	Null Hypothesis Is True		Null Hypothesis Is False			
n	α_{01}	α_{02}	β_{11}	β_{21}	β_{12}	β_{22}
250	0.012	0.016	0.212	0.239	0.294	0.272
500	0.018	0.014	0.270	0.303	0.318	0.334
750	0.014	0.008	0.310	0.367	0.408	0.422

Table 4.2. Simulated size and power values at the 5% significance level

Sample Size	Null Hypothesis Is True		Null Hypothesis Is False			
n	α_{01}	α_{02}	β_{11}	β_{21}	β_{12}	β_{22}
250	0.054	0.046	0.514	0.522	0.656	0.658
500	0.052	0.058	0.572	0.564	0.690	0.730
750	0.046	0.052	0.648	0.658	0.820	0.812

Table 4.3. Simulated size and power values at the 10% significance level

Sample Size	Null Hypothesis Is True		Null Hypothesis Is False			
n	α_{01}	α_{02}	β_{11}	β_{21}	β_{12}	β_{22}
250	0.116	0.110	0.696	0.764	0.884	0.909
500	0.104	0.090	0.744	0.817	0.860	0.934
750	0.108	0.090	0.844	0.895	0.946	0.968

Tables 4.1–4.3 report comprehensive simulation results for both the sizes and power values of the proposed tests for models (4.3) and (4.4). Column 2 in each of Tables 4.1–4.3 shows that while the sizes for the test based on $\widehat{h}_{0\mathrm{cv}}$ are comparable with these given in column 3 based on $\widehat{h}_{0\mathrm{test}}$, the power values of the test based on $\widehat{h}_{j\mathrm{test}}$ in columns 6 and 7 are always greater than these given in columns 4 and 5 based on $\widehat{h}_{j\mathrm{cv}}$. This is not surprising, because the theory shows that each of $\widehat{h}_{j\mathrm{test}}$ is chosen such that the resulting power function is maximized while the corresponding size function is under control by the significance level.

In addition, the test based on $\widehat{h}_{2\mathrm{test}}$ is almost uniformly more powerful than the best based on $\widehat{h}_{1\mathrm{test}}$, which is the second most powerful test. This is basically because $\widehat{h}_{2\mathrm{test}}$ is based on considering $\mathcal{H}_1$ with $c_{2n} = n^{-\frac{7}{18}}$, which goes to zero slower than $c_{1n} = n^{-\frac{1}{2}}\sqrt{\log\log(n)}$, and hence the distance between the alternative and the null is biggest in the former case (and therefore easier to detect). Meanwhile, the last columns of Tables 4.1–4.3 show that the test based on the bandwidth $\widehat{h}_{2\mathrm{test}}$ is still a powerful test even though the bandwidth is proportional to $n^{-\frac{1}{6}}$, which is the same as the optimal bandwidth based on a cross–validation estimation method. This shows that whether an estimation–based optimal bandwidth may be used for testing depends on whether the bandwidth is chosen optimally for testing purposes.

We finally want to stress that the proposed test based on either $\widehat{h}_{1\mathrm{test}}$ or $\widehat{h}_{2\mathrm{test}}$ has not only stable sizes even at a small sample size of $n = 250$, but also reasonable power values even when the 'distance' between the null and the alternative has been made deliberately close at the rate of $\sqrt{n^{-1}\ \mathrm{loglog}(n)} = 0.060$ for $n = 500$ for example. We can expect that the test would have bigger power values when the 'distance' is made wider. Overall, Tables 4.1–4.3 show that the established theory and methodology is workable in the small and medium–sample case.

Example 4.1 discusses the small and medium–sample comparison results for the proposed test with either testing–based optimal bandwidth or estimation–based (CV) bandwidth. Example 4.2 below considers comparing the small and medium–sample performance of the proposed test associated with the optimal bandwidth with some closely related nonparametric tests available in both the econometrics and statistics literature.

Example 4.2. Consider a linear model of the form

$$Y_i = \alpha_0 + \beta_0\ X_i + e_i,\ 1 \le i \le n = 250, \tag{4.7}$$

where $\{X_i\}$ is a sequence of independent random variables sampled from $N(0, 25)$ distribution truncated at its 5th and 95th percentiles, and $\{e_i\}$ is sampled from one of the three distributions: (i) $e_i \sim N(0, 4)$; (ii) a mixture of Normals in which $\{e_i\}$ is sampled from $N(0, 1.56)$ with probability 0.9 and from $N(0, 25)$ with probability 0.1; and (iii) the Type I extreme value distribution scaled to have a variance of 4. The mixture distribution is leptokurtic with a variance of 0.39, and the Type I extreme value distribution is asymmetrical.

This is the same example as used in Horowitz and Spokoiny (2001) for the comparison with some of the commonly used tests in the literature, such as the Andrews' test proposed in Andrews (1997), the HM test proposed in Härdle and Mammen (1993), the HS test proposed in Horowitz and Spokoiny (2001) and the empirical likelihood (EL) test proposed in Chen, Härdle and Li (2003).

To compute the sizes of the test, choose $\alpha_0 = \beta_0 = 1$ as the true parameters and then generate $\{Y_i\}$ from $Y_i = 1 + X_i + e_i$ under $\mathcal{H}_0$, and generate $\{Y_i\}$ from $Y_i = 1 + X_i + \frac{5}{\tau}\ \phi\left(\frac{X_i}{\tau}\right) + e_i$ under $\mathcal{H}_1$, where $\tau = 1$ or 0.25, and $\phi(\cdot)$ is the density function of the standard normal distribution.

The kernel function used here is $K(x) = \frac{15}{16}\ (1-x^2)^2\ I(|x| \le 1)$. Choose $c_n = 5\tau^{-1}$ and $\Delta(x) = \phi(x\ \tau^{-1})$ for the corresponding forms in (1.2). For $j = 1, 2$, let $c_{jn} = 5\tau_j^{-1}$ and $\Delta_j(x) = \phi(x\ \tau_j^{-1})$ with $\tau_1 = 1$ and $\tau_2 = 0.25$. Let $\widehat{h}_{i\mathrm{new}}$ be the corresponding version of $\widehat{h}_{\mathrm{ew}}$ of (B.43) based on $(c_{jn}, \Delta_j(x))$ for $j = 1, 2$.

In order to make a fair comparison, we use the same number of the bootstrap resamples of $M = 99$, the same number of replications of $M = 1000$ under $\mathcal{H}_0$ and

$M = 250$ under $\mathcal{H}_1$ as in Table 1 of Horowitz and Spokoiny (2001). In Table 4.4 below, we add the size and power values to the last two columns for both the EL test and the proposed test–$\widehat{T}_n\left(\widehat{h}_{i\text{new}}\right)$ of this paper. The other parts of the table are obtained and tabulated similarly to Table 1 of Horowitz and Spokoiny (2001).

Table 4.4. Simulated size and power values at the 5% significance level

		Probability of Rejecting Null Hypothesis				
Distribution	τ	Andrews Test	HM Test	HS Test	EL Test	$\widehat{T}_n\left(\widehat{h}_{\text{new}}\right)$ Test
		Null Hypothesis Is True				
Normal		0.057	0.060	0.066	0.053	0.049
Mixture		0.053	0.053	0.048	0.055	0.052
Extreme		0.063	0.057	0.055	0.057	0.052
		Null Hypothesis Is False				
Normal	1.0	0.680	0.752	0.792	0.900	0.907
Mixture	1.0	0.692	0.736	0.835	0.905	1.000
Extreme	1.0	0.600	0.760	0.820	0.924	0.935
Normal	0.25	0.536	0.770	0.924	0.929	0.993
Mixture	0.25	0.592	0.704	0.922	0.986	0.999
Extreme	0.25	0.604	0.696	0.968	0.989	0.989

Table 4.4 shows that the proposed test has better power properties than any of the commonly used tests, while the size values are comparable with those of the competitors. The results further support the power–based bandwidth selection procedure proposed in Sections 2 and 3.

As discussed in the supplemental material, the proposed theory and methodology for model (1.1) can be applied to an extended model of the form

$$Y_i = m(X_i) + e_i \quad \text{with} \quad e_i = \sigma(X_i)\ \epsilon_i,\ 1 \leq i \leq n, \tag{4.8}$$

where $\sigma(\cdot)$ satisfying $\inf_{x \in I\!R^d} \sigma(x) > 0$ is unknown nonparametrically and $\{\epsilon_i\}$ is a sequence of i.i.d. random errors with zero mean and finite variance. In addition, $\{\epsilon_i\}$ and $\{X_j\}$ are assumed to be independent for all $1 \le j \le i \le n$. A special case of model (4.8) is discussed in Example 4.3 below.

Example 4.3. This example examines the high frequency seven–day Eurodollar deposit rate sampled daily from 1 June 1973 to 25 February 1995. This provides us with $n = 5505$ observations. Let $\{X_i : i = 1, 2, \cdots, n = 5505\}$ be the set of Eurodollar deposit rate data. Figures 4.1 and 4.2 below plot the data values and the conventional nonparametric kernel density estimator

$$\widehat{\pi}(x) = \frac{1}{n\widetilde{h}_{\text{cv}}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\widetilde{h}_{\text{cv}}}\right)$$

respectively, where $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $\widetilde{h}_{\text{cv}}$ is the conventional normal–reference based bandwidth given by

$$\widetilde{h}_{\text{cv}} = 1.06 \cdot n^{-\frac{1}{5}} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2} \ \ \text{with} \ \ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i. \qquad (4.9)$$

Note that $\widehat{b}_{\text{cv}}$ of (2.7), $\widehat{h}_{\text{cv}}$ of (4.6) and $\widetilde{h}_{\text{cv}}$ of (4.9) are normally different from each other. In the case where $\{X_i\}$ follows an autoregressive model, they can be chosen the same. Thus, they are chosen the same in this example.

It has been assumed in the literature (see, for example, Aït–Sahalia 1996; Fan and Zheng 2003; Arapis and Gao 2006) that the Eurodollar data set $\{X_i\}$ may be modeled by a nonlinear time series model of the form

$$Y_i = \mu(X_i) + \sigma(X_i)\ \epsilon_i, \ 1 \le i \le n, \qquad (4.10)$$

where $Y_i = \frac{X_{i+1} - X_i}{\Lambda}$, $\sigma(\cdot) > 0$ is unknown nonparametrically, and $\epsilon_i \sim N(0, \Lambda^{-1})$, in which Λ is the time between successive observations. Since we consider a daily data set, this gives $\Lambda = \frac{1}{250}$.

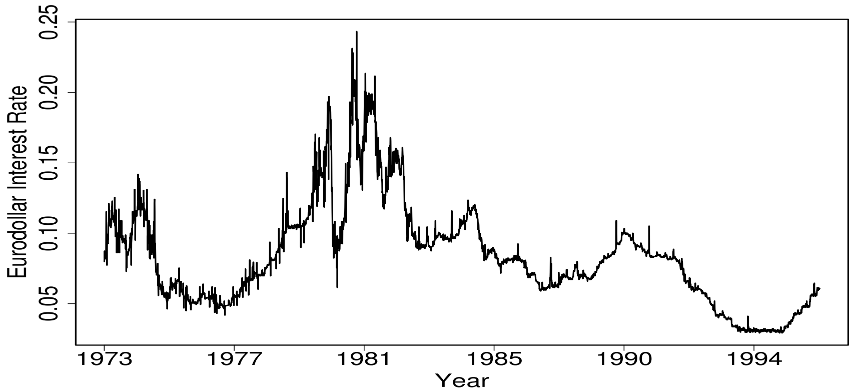

Figure 4.1: Seven–day Eurodollar deposit rate, June 1, 1973 to February 25, 1995.

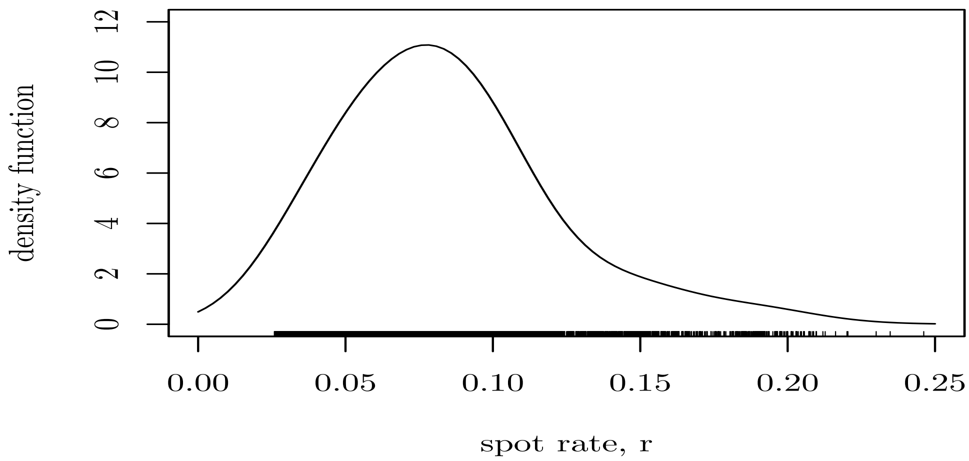

Figure 4.2: Nonparametric kernel density estimator of the Eurodollar rate.

On the question of whether there is any nonlinearity in the drift function $\mu(\cdot)$, existing studies have provided no definitive answer. For example, Aït–Sahalia (1996), and Arapis and Gao (2006) show that there is some evidence of supporting nonlinearity in the drift on the one hand. On the other hand, existing studies, such as

Chapman and Pearson (2000), and Fan and Zheng (2003), suggest that nonlinearity may just be caused by estimation biases when using nonparametric kernel estimation.

To further discuss whether the assumption on linearity in the drift is appropriate for the given set of data, we apply our test to propose testing

$$\mathcal{H}_{01}: \ \mu(x) = \mu(x;\theta_0) = \beta_0(\alpha_0 - x) \quad \text{versus} \quad \mathcal{H}_{11}: \ \mu(x) = \beta_1(\alpha_1 - x) + c_n \Delta(x) \tag{4.11}$$

for some $\theta_j = (\alpha_j, \beta_j) \in \Theta$ for $j = 0, 1$ and $c_n = \sqrt{n^{-1} \log\log(n)}$, where Θ is a parameter space in $I\!R^2$ and $\Delta(x)$ is a continuous function.

It can be shown that the proposed test in Section 2 has an asymptotically equivalent version of the form:

$$\widetilde{T}_n(h) = \frac{\sum_{j=1}^n \sum_{i=1, i\neq j}^n \widehat{e}_j \ K\left(\frac{X_i - X_j}{h}\right) \ \widehat{e}_i}{\sqrt{2 \sum_{j=1}^n \sum_{i=1}^n \widehat{e}_j^2 \ K^2\left(\frac{X_i - X_j}{h}\right) \ \widehat{e}_i^2}}, \tag{4.12}$$

where $\widehat{e}_i = Y_i - \widehat{\beta}(\widehat{\alpha} - X_i)$, in which $(\widehat{\alpha}, \widehat{\beta})$ is the pair of the conventional least squares estimators minimizing $\sum_{i=1}^n \left(Y_i - \widehat{\beta}(\widehat{\alpha} - X_i)\right)^2$.

As pointed out in the literature (Arapis and Gao 2006), $\widetilde{T}_n(h)$ is independent of the structure of the conditional variance $\sigma^2(\cdot)$. The kernel function used is the standard normal density function given by $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

Let $\widetilde{h}_{\text{test}}$ be the corresponding version of (B.43). It has been shown in Appendix B below that

$$\widetilde{h}_{\text{test}} = \widehat{a}_1^{-\frac{1}{2}} \widehat{t}_n^{-\frac{3}{2}}, \tag{4.13}$$

where $\widehat{t}_n$ and $\widehat{a}_1$ are the same as in (B.43), in which $\widehat{c}(\pi)$ becomes

$$\widehat{c}(\pi) = \frac{1}{n} \sum_{i=1}^n \widehat{\pi}^2(X_i) \widehat{\sigma}^6(X_i) \cdot \left(\frac{1}{n} \sum_{i=1}^n \widehat{\pi}(X_i) \widehat{\sigma}^4(X_i) \right)^{-\frac{3}{2}} \tag{4.14}$$

with

$$\widehat{\sigma}^2(X_i) = \frac{\sum_{u=1}^n K\left(\frac{X_i - X_u}{\widetilde{h}_{\text{cv}}}\right) \widehat{e}_u^2}{\sum_{v=1}^n K\left(\frac{X_i - X_v}{\widetilde{h}_{\text{cv}}}\right)}.$$

Let $L_1 = \widetilde{T}_n(\widetilde{h}_{\rm test})$ and $L_2 = \widetilde{T}_n(\widetilde{h}_{\rm cv})$. To apply the test L_j for each $j = 1, 2$ to test $\mathcal{H}_{01}$, we propose the following procedure for computing the p–value of L_j:

- Compute $\widehat{e}_i = Y_i - \widehat{\beta}(\widehat{\alpha} - X_i)$ and then generate a sequence of bootstrap resamples $\{\widehat{e}_i^*\}$ given by $\widehat{e}_i^* = \widehat{\sigma}(X_i)\ \epsilon_i^*$, where $\{\epsilon_i^*\}$ is a sequence of i.i.d. bootstrap resamples generated from $N(0, \Lambda^{-1})$ and $\widehat{\sigma}^2(\cdot)$ is defined as above.

- Generate $\widehat{Y}_i^* = \widehat{\beta}(\widehat{\alpha} - X_i) + \widehat{e}_i^*$. Compute the corresponding version L_j^* of L_j for each $j = 1, 2$ based on $\{\widehat{Y}_i^*\}$.

- Repeat the above steps $M = 1000$ times to find the bootstrap distribution of L_j^* and then compute the proportion that $L_j < L_j^*$ for each $j = 1, 2$. This proportion is a simulated p–value of L_j.

Our simulation results return the simulated p–values of $\widehat{p}_1 = 0.102$ for L_1 and $\widehat{p}_2 = 0.072$ for L_2. While both of the simulated p–values suggest that there is no enough evidence of rejecting the linearity in the drift at the 5% significance level, the evidence of accepting the linearity based on L_1 is stronger than that based on L_2.

As our test $\widetilde{T}_n(\widetilde{h}_{\rm test})$ involves no estimation biases, the process of computing the simulated p–values is quite robust. We therefore believe that this improved test further reinforces the findings of Chapman and Pearson (2000) and Fan and Zhang (2003) that there is no definitive answer to the question whether the short rate drift is actually nonlinear.

5. Conclusion

This paper has addressed the issue of how to appropriately choose the bandwidth parameter when using a nonparametric kernel–based test. Both the size and power properties of the proposed test have been studied systematically. The established theory and methodology has shown that a suitable bandwidth can be optimally chosen after appropriately balancing the size and power functions. Furthermore, the new methodology has resulted in a closed–form representation for the leading term of such an optimal bandwidth in the finite–sample case.

Existing results (see, for example, Li and Wang 1998; Li 1999; Fan and Linton 2003; Gao 2007) show that this kind of nonparametric kernel test associated with a large sample critical value may not have good size and power properties. Our small and medium–sample studies in both the simulated and real–data examples have shown that the performance of such a test can be significantly improved when it is coupled with a power–based optimal bandwidth as well as a bootstrap simulated critical value.

It is pointed out that the established theory and methodology has various applications in providing solutions to some other related testing problems, in which nonparametric methods are involved. Future extensions also include dealing with cases where both X_i and e_i may be strictly stationary time series.

Appendix A

This appendix lists the necessary assumptions for the establishment and the proofs of the main results given in Section 2.

A.1. *Assumptions*

Assumption A.1. (i) *Assume that* $\{e_i\}$ *is a sequence of i.i.d. continuous random errors with* $E[e_1] = 0$, $0 < \sigma^2 = E[e_1^2] = \sigma^2 < \infty$ *and* $E[e_1^6] < \infty$.

(ii) *We assume that* $\{X_i\}$ *is strictly stationary and* α*–mixing with mixing coefficient* $\alpha(t)$ *being defined by*

$$\alpha(t) = \sup\{|P(A \cap B) - P(A)P(B)| : A \in \Omega_1^s, B \in \Omega_{s+t}^{\infty}\} \leq C_\alpha \ \alpha^t \tag{A.1}$$

for all $s, t \geq 1$, *where* $0 < C_\alpha < \infty$ *and* $0 < \alpha < 1$ *are constants, and* Ω_i^j *denotes the* σ*–field generated by* $\{X_k : i \leq k \leq j\}$.

(iii) *We also assume that* $\{X_s\}$ *and* $\{e_t\}$ *are independent for all* $1 \leq s \leq t \leq n$. *Let* $\pi(\cdot)$ *be the marginal density such that* $0 < \int \pi^3(x)dx < \infty$, *and* $\pi_{\tau_1,\tau_2,\cdots,\tau_l}(\cdot)$ *be the joint probability density of* $(X_{1+\tau_1}, \ldots, X_{1+\tau_l})$ $(1 \leq l \leq 4)$. *Assume that* $\pi_{\tau_1,\tau_2,\cdots,\tau_l}(\cdot)$ *for all* $1 \leq l \leq 4$ *do exist and are continuous and bounded.*

(iv) *Assume that the univariate kernel function* $K(\cdot)$ *is a symmetric and bounded probability density function. In addition, we assume the existence of both* $K^{(3)}(\cdot)$*, the three–time convolution of* $K(\cdot)$ *with itself, and* $K_2^{(2)}(\cdot)$*, the two–time convolution of* $K^2(\cdot)$ *with itself.*

(v) *The bandwidth parameter* h *satisfies both* $\lim_{n\to\infty} h = 0$ *and* $\lim_{n\to\infty} nh^d = \infty$.

Assumption A.2. (i) *Let* $\mathcal{H}_0$ *be true. Then for any sufficiently small* $\varepsilon_1 > 0$ *and some* $B_{1L} > 0$

$$\lim_{n\to\infty} P\left(\sqrt{n}||\widehat{\theta} - \theta_0|| > B_{1L}\right) < \varepsilon_1,$$

where θ_0 *is the same as defined in (1.2).*

(ii) *Let* $\mathcal{H}_1$ *be true. Then for any sufficiently small* $\varepsilon_2 > 0$ *and some* $B_{2L} > 0$

$$\lim_{n\to\infty} P\left(\sqrt{n}||\widehat{\theta} - \theta_1|| > B_{2L}\right) < \varepsilon_2,$$

where θ_1 *is the same as defined in (1.2).*

(iii) *There exist some absolute constants* $\varepsilon_3 > 0$ *and* $0 < B_{3L} < \infty$ *such that the following*

$$\lim_{n\to\infty} P\left(\sqrt{n}||\widehat{\theta}^* - \widehat{\theta}|| > B_{3L}|\mathcal{W}_n\right) < \varepsilon_3$$

holds in probability, where $\widehat{\theta}^*$ *is as defined in the Simulation Procedure above Theorem 2.1.*

(iv) *Let* $m_\theta(x)$ *be differentiable with respect to* θ *and* $\frac{\partial m_\theta(x)}{\partial\theta}$ *be continuous in both* x *and* θ*. In addition,* $E\left[\left(\frac{m_{\theta_0}(X_1)}{\partial\theta}\right)\left(\frac{m_{\theta_0}(X_1)}{\partial\theta}\right)^\tau\right]$ *is a positive definite matrix, and*

$$0 < \int \left|\left|\frac{\partial m_\theta(x)}{\partial\theta}|_{\theta=\theta_0}\right|\right|^2 \pi^2(x)dx < \infty.$$

Assumption A.3. (i) *Let* $\{\Delta_n(x)\}$ *be a sequence of continuous functions such that* $0 < \int \Delta_n^2(x)\pi^2(x)dx < \infty$.

(ii) *Let* C_n^2 *satisfy* $\lim_{n\to\infty} n\ \sqrt{h^d}\ C_n^2 = \infty$ *and* $\lim_{n\to\infty} n\ C_n^6 = 0$*, where*

$$C_n^2 = \frac{\int \Delta_n^2(x)\pi^2(x)dx}{\sigma^2\sqrt{2\nu_2\ \int K^2(v)dv}},$$

in which $\nu_2 = E\left[\pi^2(X_1)\right] < \infty$.

Assumptions A.1–A.3 are standard and justifiable conditions. Some detailed justifications are given in Appendix C below.

A.2. *Technical lemmas*

Recall that using $\lim_{n\to\infty} nh^d = \infty$

$$\begin{aligned}
\kappa_n &= \frac{\sqrt{h^d}\left(\frac{\mu_3^2 K^2(0)}{nh^d} + \frac{4\mu_2^3\nu_3}{3}K^{(3)}(0)\right)}{\sigma_n^3} \equiv c_1\ \sqrt{h^d} + c_2\ \frac{1}{n\sqrt{h^d}} \\
&= c_1\ \sqrt{h^d}\left(1 + c_2 c_1^{-1}\frac{1}{nh^d}\right) \approx c_1\ \sqrt{h^d}. \qquad \text{(A.2)}
\end{aligned}$$

In order to establish some useful lemmas without including non–essential technicality, we introduce the following simplified notation:

$$\begin{aligned}
a_{ij} &= \frac{1}{n\sqrt{h^d}\sigma_n}K\left(\frac{X_i - X_j}{h}\right), \quad L_n(h) = \sum_{i=1}^n \sum_{j=1,\neq i}^n a_{ij}e_i e_j, \\
\rho(h) &= \frac{\sqrt{2}K^{(3)}(0)\ \int \pi^3(u)du}{3}\left(\sqrt{\int \pi^2(u)du\ \int K^2(v)dv}\right)^{-3}\ \sqrt{h^d}. \qquad \text{(A.3)}
\end{aligned}$$

We need the following lemmas; their proofs are given in Appendix C below.

LEMMA A.1. *Suppose that the conditions of Theorem 2.2(i) hold. Then for any h*

$$\sup_{x\in I\!R^1}\left|P\left(L_n(h) \leq x\right) - \Phi(x) + \rho(h)\ (x^2-1)\ \phi(x)\right| = O\left(h^d\right). \qquad \text{(A.4)}$$

Recall $L_n(h) = \sum_{i=1}^n \sum_{j=1,\neq i}^n e_i\ a_{ij}e_j$ as defined in (A.3) and let

$$\begin{aligned}
\overline{T}_n(h) &= \frac{h^{\frac{d}{2}}}{n\sigma_n}\sum_{i=1}^n \sum_{j=1,\neq i}^n \widehat{e}_i\ K_h(X_i - X_j)\ \widehat{e}_j = \frac{h^{\frac{d}{2}}}{n\sigma_n}\sum_{i=1}^n \sum_{j=1,\neq i}^n e_i\ K_h(X_i - X_j)\ e_j \\
&+ \frac{h^{\frac{d}{2}}}{n\sigma_n}\sum_{i=1}^n \sum_{j=1,\neq i}^n\ K_h(X_i - X_j)\left[m(X_i) - m_{\widehat{\theta}}(X_i)\right]\left[m(X_j) - m_{\widehat{\theta}}(X_j)\right] \\
&+ \frac{2h^{\frac{d}{2}}}{n\sigma_n}\sum_{i=1}^n \sum_{j=1,\neq i}^n e_i\ K_h(X_i - X_j)\left[m(X_j) - m_{\widehat{\theta}}(X_j)\right] \\
&\equiv L_n(h) + S_n(h) + D_n(h), \qquad \text{(A.5)}
\end{aligned}$$

where $S_n(h) = \frac{h^{\frac{d}{2}}}{n\sigma_n}\sum_{i=1}^n \sum_{j=1,\neq i}^n\ K_h(X_i - X_j)\left[m(X_i) - m_{\widehat{\theta}}(X_i)\right]\left[m(X_j) - m_{\widehat{\theta}}(X_j)\right]$ and

$$D_n(h) = \frac{2h^{\frac{d}{2}}}{n\sigma_n}\sum_{i=1}^n \sum_{j=1,\neq i}^n e_i\ K_h(X_i - X_j)\left[m(X_j) - m_{\widehat{\theta}}(X_j)\right]. \qquad \text{(A.6)}$$

Define $L_n^*(h)$, $S_n^*(h)$ and $D_n^*(h)$ as the corresponding versions of $L_n(h)$, $S_n(h)$ and $D_n(h)$ involved in (A.5) with (X_i, Y_i) and $\widehat{\theta}$ being replaced by (X_i, Y_i^*) and $\widehat{\theta}^*$ respectively.

Lemma A.2. *Suppose that the conditions of Theorem 2.2(i) hold. Then*

$$\sup_{x\in I\!R^1}\left|P^*\left(L_n^*(h)\le x\right)-\Phi(x)+\rho(h)\ (x^2-1)\ \phi(x)\right| = O_P\left(h^d\right). \tag{A.7}$$

Lemma A.3. *Suppose that the conditions of Theorem 2.2(i) hold. Then under* $\mathcal{H}_0$

$$E\left[S_n(h)\right] = O\left(\sqrt{h^d}\right) \quad \textit{and} \quad E\left[D_n(h)\right] = o\left(\sqrt{h^d}\right), \tag{A.8}$$

$$E^*\left[S_n^*(h)\right] = O_P\left(\sqrt{h^d}\right) \quad \textit{and} \quad E^*\left[D_n^*(h)\right] = o_P\left(\sqrt{h^d}\right), \tag{A.9}$$

$$E\left[S_n(h)\right]-E^*\left[S_n^*(h)\right] = O_P\left(\sqrt{h^d}\right) \ \textit{and}\ E\left[D_n(h)\right]-E^*\left[D_n^*(h)\right] = o_P\left(\sqrt{h^d}\right) \tag{A.10}$$

in probability with respect to the joint distribution of $\mathcal{W}_n$*, where* $E^*[\cdot]=E[\cdot|\mathcal{W}_n]$.

Lemma A.4. *Suppose that the conditions of Theorem 2.2(ii) hold. Then under* $\mathcal{H}_1$

$$\lim_{n\to\infty} E\left[S_n(h)\right]=\infty \quad \text{and} \quad \lim_{n\to\infty}\frac{E\left[D_n(h)\right]}{E\left[S_n(h)\right]}=0. \tag{A.11}$$

A.3. *Proof of Theorem 2.1*:

A.3.1. Proof of (2.10): Recall from (2.8) and (A.5)–(A.6) that

$$\widehat{T}_n(h) = \left(L_n(h)+S_n(h)+D_n(h)\right)\cdot\frac{\sigma_n}{\widehat{\sigma}_n}+o_P\left(\sqrt{h^d}\right), \tag{A.12}$$

$$\widehat{T}_n^*(h) = \left(L_n^*(h)+S_n^*(h)+D_n^*(h)\right)\cdot\frac{\sigma_n}{\widehat{\sigma}_n^*}+o_P\left(\sqrt{h^d}\right), \tag{A.13}$$

where σ_n^2, $\widehat{\sigma}_n^2$ and $\widehat{\sigma}_n^{*2}$ are as defined in (2.4), (2.7) and (2.9) respectively.

In view of Assumption A.2 and Lemmas A.1–A.3, we may ignore any terms with orders higher than $\sqrt{h^d}$ and then consider the following approximations:

$$\begin{aligned}\widehat{T}_n(h) &= L_n(h)+E\left[S_n(h)\right]+o_P(\sqrt{h^d}) \quad \text{and}\\ \widehat{T}_n^*(h) &= L_n^*(h)+E^*\left[S_n^*(h)\right]+o_P\left(\sqrt{h^d}\right).\end{aligned} \tag{A.14}$$

Let $s(h)=E[S_n(h)]$ and $s^*(h)=E^*\left[S_n^*(h)\right]$. We then apply Lemmas A.1 and A.2 to obtain that uniformly over $x\in I\!R^1$,

$$\begin{aligned}P\left(\widehat{T}_n(h)\le x\right) &= P\left(L_n(h)\le x-s(h)+o_P\left(\sqrt{h^d}\right)\right) \qquad \text{(A.15)}\\ &= \Phi(x-s(h))-\rho(h)((x-s(h))^2-1)\ \phi(x-s(h))+o\left(\sqrt{h^d}\right) \quad \text{and}\\ P^*\left(\widehat{T}_n^*(h)\le x\right) &= P^*\left(L_n^*(h)\le x-s^*(h)+o_P\left(\sqrt{h^d}\right)\right)\\ &= \Phi(x-s^*(h))-\rho(h)((x-s^*(h))^2-1)\ \phi(x-s^*(h))+o_P\left(\sqrt{h^d}\right).\end{aligned}$$

Theorem 2.2(i) follows consequently from (A.10) and (A.15).

A.3.2. Proof of (2.11): In view of the definition that $P^*\left(\widehat{T}_n^*(h) \geq l_\alpha^*\right) = \alpha$ and the conclusion from Theorem 2.1(i) that

$$P\left(\widehat{T}_n(h) \geq l_\alpha^*\right) - P^*\left(\widehat{T}_n^*(h) \geq l_\alpha^*\right) = O_P\left(\sqrt{h^d}\right), \tag{A.16}$$

the proof of $P\left(\widehat{T}_n(h) \geq l_\alpha^*\right) = \alpha + O\left(\sqrt{h^d}\right)$ follows unconditionally from the dominated convergence theorem.

A.4. *Proof of Theorem 2.2*

It follows from Lemmas A.1–A.4 that

$$\begin{aligned}
\alpha_n(h) &= P\left(\widehat{T}_n(h) \geq l_\alpha | \mathcal{H}_0\right) = P\left(L_n(h) \geq l_\alpha - S_n(h) + o_P(S_n(h)) | \mathcal{H}_0\right) \\
&= 1 - P\left(L_n(h) \leq l_\alpha - S_n(h) + o_P(S_n(h)) | \mathcal{H}_0\right), && \text{(A.17)} \\
\alpha_n^*(h) &= P\left(\widehat{T}_n(h) \geq l_\alpha^* | \mathcal{H}_0\right) = P\left(L_n(h) \geq l_\alpha^* - S_n(h) + o_P(S_n(h)) | \mathcal{H}_0\right) \\
&= 1 - P\left(L_n(h) \leq l_\alpha^* - S_n(h) + o_P(S_n(h)) | \mathcal{H}_0\right), && \text{(A.18)} \\
\beta_n(h) &= P\left(\widehat{T}_n(h) \geq l_\alpha | \mathcal{H}_1\right) = P\left(L_n(h) \geq l_\alpha - S_n(h) + o_P(S_n(h)) | \mathcal{H}_1\right) \\
&= 1 - P\left(L_n(h) \leq l_\alpha - S_n(h) + o_P(S_n(h)) | \mathcal{H}_1\right), && \text{(A.19)} \\
\beta_n^*(h) &= P\left(\widehat{T}_n(h) \geq l_\alpha^* | \mathcal{H}_1\right) = P\left(L_n(h) \geq l_\alpha^* - S_n(h) + o_P(S_n(h)) | \mathcal{H}_1\right) \\
&= 1 - P\left(L_n(h) \leq l_\alpha^* - S_n(h) + o_P(S_n(h)) | \mathcal{H}_1\right). && \text{(A.20)}
\end{aligned}$$

Using Assumptions A.2(iv) and A.3, a Taylor expansion of $m_\theta(\cdot)$ at θ_0 implies that for sufficiently large n

$$\begin{aligned}
S_n(h) &= C_0(m)\, \sqrt{h^d}\, (1 + o_P(1)) \quad \text{under} \quad \mathcal{H}_0 \quad \text{and} && \text{(A.21)} \\
S_n(h) &= n\, C_n^2\, \sqrt{h^d}\, (1 + o_P(1)) \;\; \text{under} \;\; \mathcal{H}_1 && \text{(A.22)}
\end{aligned}$$

hold in probability, where C_n^2 is as defined in Theorem 2.2(ii). The proof of Theorem 2.2 then follows from (A.15) and (A.17)–(A.22).

A.5. *Proof of Theorem 2.3*: The proof follows from that of Theorem 2.2. The details are given in Appendix C below.

Appendix B

B.1. Derivation of (3.9) in the submission

In the sequel we work with the approximate power–function $\beta(x)$, and discuss how to find a maximum for this function. Straightforward calculations imply that the first derivative may be written as

$$\beta'(x) = \phi(z_\alpha + d_3 x + d_4 x^{-1})\, x^{-6} \sum_{i=0}^{10} \vartheta_i x^i, \tag{B.1}$$

where the coefficients are given by

$$\begin{aligned}
\vartheta_0 &= c_2 n^{-1} d_4^4,\ \vartheta_1 = 3 z_\alpha d_4^3 c_2\ n^{-1},\ \ \vartheta_2 = 2c_2 n^{-1} d_3 d_4^3 - d_4^2 c_2 n^{-1}(4 - 3z_\alpha^2) + d_4^4 c_1,\\
\vartheta_3 &= 3 z_\alpha d_3 d_4^2 c_2 n^{-1} - c_2 n^{-1} z_\alpha d_4 (5 - z_\alpha^2) + 3 z_\alpha d_4^3 c_1,\\
\vartheta_4 &= 2 d_3 d_4(-c_2 n^{-1} + d_4^2 c_1) + d_4^2 c_1 (2 - 3z_\alpha^2),\\
\vartheta_5 &= -3 z_\alpha d_3^2 d_4 c_2 n^{-1} + d_3 \left(-z_\alpha d_4 + 3 z_\alpha d_4^2 c_1\right) - z_\alpha d_4 c_1,\\
\vartheta_6 &= -c_1(1 - z_\alpha^2) + d_3(-1 + 2 d_4 c_1) - d_3^2 c_2 n^{-1}(-2 + 3 z_\alpha^2) - 2 d_3^3 d_4 c_2 n^{-1},\\
\vartheta_7 &= -3 z_\alpha d_3^3 c_2 n^{-1} - 3 z_\alpha d_3^2 d_4 c_1 + z_\alpha d_3 c_1 (5 - z_\alpha^2),\\
\vartheta_8 &= -c_2 d_3^4 n^{-1} - 2 c_1 d_3^3 d_4 - c_1 d_3^2 (3 z_\alpha^2 - 4),\ \vartheta_9 = -3 c_1 z_\alpha d_3^3,\ \vartheta_{10} = -c_1 d_3^4.
\end{aligned} \tag{B.2}$$

Due to the complexity of the expressions of $\beta'(x)$, it is very difficult to find such an x_0 explicitly. Numerically, however, it may be possible to find $x_{\text{ew}} = \sqrt{h_{\text{ew}}^d}$ such that

$$x_{\text{ew}} = \arg \max_{x \in H_n(\alpha)} \beta(x) \quad \text{with} \quad H_n(\alpha) = \{x:\ \alpha - c_{\min} < \alpha(x) < \alpha + c_{\min}\} \tag{B.3}$$

when $c_{\min}$ is chosen as $c_{\min} = \frac{\alpha}{10}$ for example.

We now discuss how to get to an explicit expression of an optimal bandwidth, by maximizing the power function over a subset of $H_n(\alpha)$. Note that the minimal conditions that $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh^d = \infty$ imply that there is some large integer $N \geq 1$ such that for any arbitrarily large but finite $c_{\max} \geq 1$,

$$h \leq c_{\max}^{-1} \quad \text{and} \quad nh^d \geq c_{\max} \quad \text{for any } n \geq N. \tag{B.4}$$

In our finite–sample analysis, we then define a new interval of the form

$$H_n = \left[\left(\frac{c_{\max}}{n} \right)^{d^{-1}}, c_{\max}^{-1} \right]. \tag{B.5}$$

Since Theorem 2.2(i) shows that $\lim_{n\to\infty} \alpha_n(h) = \alpha$ holds in probability under the minimal conditions $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh^d = \infty$, we have $H_n \subseteq H_n(\alpha)$ at least for sufficiently large n.

In order to represent h_{ew} explicitly we consider, quite naturally, solving the optimization problem:

$$h_{\text{ew}} = \arg\max_{h \in H_n} \beta_n(h). \tag{B.6}$$

To keep the notation simple, we still use the notation h_{ew} even when H_n may not be identical to $H_n(\alpha)$. We impose the very natural condition that $\lim_{n\to\infty} nh^d = \infty$ and obtain the following approximations:

$$\begin{aligned} \sigma_n^2 &= 2\mu_2^2 \nu_2 \int K^2(u)du \equiv \sigma_0^2, \\ \kappa_n &= \sigma_n^{-3} \sqrt{h^d} \left(\frac{4\mu_2^3 \nu_3}{3} K^{(3)}(0) + \mu_3^2 \frac{K^2(0)}{nh^d} \right) \approx \sqrt{h^d}\, \sigma_0^{-3} \cdot \frac{4\mu_2^3 \nu_3}{3} K^{(3)}(0). \end{aligned} \tag{B.7}$$

Let

$$a_1 = \frac{4K^{(3)}(0)\mu_2^3 \nu^3}{3\sigma_0^3} = \frac{\sqrt{2}K^{(3)}(0)}{3} \left(\sqrt{\int K^2(u)du} \right)^{-3} c(\pi) \tag{B.8}$$

with $c(\pi) = \frac{\int \pi^3(x)dx}{\left(\sqrt{\int \pi^2(x)dx}\right)^3}$. We then have $\kappa_n \approx a_1 \sqrt{h^d}$. Let $b_1 = (z_\alpha^2 - 1)a_1$.

From the condition $\lim_{n\to\infty} nh^d = \infty$ and using (B.7) and (B.8), we then get the following simplified versions:

$$\begin{aligned} l_\alpha - r_n &= z_\alpha + \gamma_n - r_n \approx z_\alpha + \left(b_1 - nC_n^2 \right) x \\ &\equiv z_\alpha + a_2\, x, \end{aligned} \tag{B.9}$$

$$\begin{aligned} \beta_n(h) &\approx 1 - \Phi(l_\alpha - r_n) - \kappa_n\, (1 - (l_\alpha - r_n)^2)\, \phi(l_\alpha - r_n) \\ &\approx 1 - \Phi(z_\alpha + a_2 x) - a_1\, x\, (1 - (z_\alpha + a_2 x)^2)\, \phi(z_\alpha + a_2 x) \\ &\equiv \beta(x), \end{aligned} \tag{B.10}$$

where $a_2 = b_1 - nC_n^2$.

The natural condition $\lim_{n\to\infty} nh^d = \infty$ implies that we may take $c_2 = 0$, and subsequently $d_4 = 0$. This simplifies the expression of $\beta'(x)$ substantially. As a result, we have $\vartheta_i = 0$ for $0 \le i \le 5$. Thus, with the following new coefficients:

$$\begin{aligned}
\theta_0 &= \vartheta_6 = nC_n^2, \quad \theta_1 = \vartheta_7 = z_\alpha(5 - z_\alpha^2)a_1a_2, \\
\theta_2 &= \vartheta_8 = -(3z_\alpha^2 - 4)a_1a_2^2, \quad \theta_3 = \vartheta_9 = -3z_\alpha a_1a_2^3, \quad \theta_4 = \vartheta_{10} = -a_2^4a_1, \qquad \text{(B.11)}
\end{aligned}$$

equation (B.1) simplifies to

$$\beta'(x) = \phi(z_\alpha + a_2x)\left(\theta_0 + \theta_1x + \theta_2x^2 + \theta_3x^3 + \theta_4x^4\right). \qquad \text{(B.12)}$$

Since $\phi(\cdot)$ is nonnegative, the equation $\beta'(x) = 0$ is equivalent to

$$\theta_0 + \theta_1x + \theta_2x^2 + \theta_3x^3 + \theta_4x^4 = 0. \qquad \text{(B.13)}$$

We find the zeros of equation (B.13) by using existing results for a general quartic equation (see for example, http://mathworld.wolfram.com/QuarticEquation.html). Note that equation (B.13) can be written as

$$x^4 + r_3x^3 + r_2x^2 + r_1x + r_0 = 0 \quad \text{with} \quad r_i = \frac{\theta_i}{\theta_4}, \; i = 0, 1, 2, 3. \qquad \text{(B.14)}$$

Let $x = u - \frac{1}{4}r_3$,

$$p_2 = r_2 - \frac{3}{8}r_3^2, \quad p_1 = r_1 - \frac{1}{2}r_2r_3 + \frac{1}{8}r_3^3, \quad p_0 = r_0 - \frac{1}{4}r_1r_3 + \frac{1}{16}r_2r_3^2 - \frac{3}{256}r_3^4. \qquad \text{(B.15)}$$

We then may eliminate x^3 from (B.14) to obtain a standard equation of the form

$$u^4 + p_2u^2 + p_1u + p_0 = 0. \qquad \text{(B.16)}$$

Existing results immediately imply that the zeros can be represented by

$$\begin{aligned}
u_1 &= \frac{1}{2}(A_1 + A_2), \quad u_2 = \frac{1}{2}(A_1 - A_2), \qquad \text{(B.17)} \\
u_3 &= \frac{1}{2}(-A_1 + A_3), \quad u_4 = \frac{1}{2}(-A_1 - A_3), \qquad \text{(B.18)}
\end{aligned}$$

where $A_1 = \sqrt{-p_2 + y_1}$,

$$A_2 = \begin{cases} \sqrt{-A_1^2 - 2p_2 - 2p_1 A_1^{-1}} & \text{for} \quad A_1 \neq 0 \\ \sqrt{-2p_2 + 2\sqrt{y_1^2 - 4p_0}} & \text{for} \quad A_1 = 0 \end{cases}$$

$$A_3 = \begin{cases} \sqrt{-A_1^2 - 2p_2 + 2p_1 A_1^{-1}} & \text{for} \quad A_1 \neq 0, \\ \sqrt{-2p_2 - 2\sqrt{y_1^2 - 4p_0}} & \text{for} \quad A_1 = 0, \end{cases} \tag{B.19}$$

in which y_1 is a real root of the cubic equation

$$y^3 + q_2 y^2 + q_1 y + q_0 = 0, \tag{B.20}$$

where $q_2 = -p_2$, $q_1 = -4p_0$ and $q_0 = 4p_2p_0 - p_1^2$. A real root of equation (B.20) is

$$y_1 = -\frac{q_2}{3} + (B_1 + B_2), \tag{B.21}$$

in which

$$B_1 = \left(R_c + \sqrt{D_c}\right)^{1/3} \quad \text{and} \quad B_2 = \left(R_c - \sqrt{D_c}\right)^{1/3} \tag{B.22}$$

with

$$R_c = \frac{9q_2q_1 - 27q_0 - 2q_2^3}{54}, \quad Q_c = \frac{3q_1 - q_2^2}{9} \quad \text{and} \quad D_c = Q_c^3 + R_c^2. \tag{B.23}$$

In order to evaluate the four roots in (B.17) and (B.18), note first of all that the quantities R_c and D_c can be re–expressed as

$$R_c = \frac{-72p_0p_2 + 27p_1^2 + 2p_2^3}{54} \qquad \text{and} \qquad Q_c = \frac{-12p_0 - p_2^2}{9}. \tag{B.24}$$

For each of the quantities involved we now provide the dominant terms, in view of the fact that $t_n \equiv \theta_0 = nC_n^2$ can be viewed as sufficiently large when assuming that $t_n \to \infty$ as $n \to \infty$.

Note that for finding the dominant term in the quantity $B_1 + B_2$, some caution is needed since certain terms cancel out each other, and the dominant term comes from the second–order terms in the quantities B_1 and B_2. Indeed, we have

$$B_1 = \left(\sqrt{D_c}\right)^{1/3} \left(1 + \frac{R_c}{\sqrt{D_c}}\right)^{1/3} \approx \left(\sqrt{D_c}\right)^{1/3} \left\{1 + \frac{1}{3}\frac{R_c}{\sqrt{D_c}}\right\}, \tag{B.25}$$

where we used a Taylor expansion of the function $g(y) = y^{1/3}$ around the point 1. Similarly, using a Taylor expansion of the function $g(y) = y^{1/3}$ around the point -1 we have

$$B_2 = \left(\sqrt{D_c}\right)^{1/3} \left(-1 + \frac{R_c}{\sqrt{D_c}}\right)^{1/3} \approx \left(\sqrt{D_c}\right)^{1/3} \left\{-1 + \frac{1}{3}\frac{R_c}{\sqrt{D_c}}\right\}, \tag{B.26}$$

and then $B_1 + B_2$ is given by $B_1 + B_2 \approx \frac{2}{3} R_c \, D_c^{-1/3}$. Note that the dominant term for the latter term is

$$\frac{2}{3}\left(-\frac{4}{3}p_0 p_2\right)\left[-\left(\frac{4}{3}\right)^3 p_0^3\right]^{-1/3} = \frac{2}{3}p_2. \tag{B.27}$$

It may be shown that all four roots of the equation (B.16) are of order

$$u_{1,2,3,4} = (1 + o(1))\,(-p_0)^{\frac{1}{4}} = (1 + o(1))\,(-r_0)^{\frac{1}{4}} = (1 + o(1))\; a_1^{-\frac{1}{4}}\; t_n^{-\frac{3}{4}}, \tag{B.28}$$

but that only one root is real and non–negative, namely u_1 given by

$$u_1 = (1 + o(1))\; a_1^{-\frac{1}{4}}\; t_n^{-\frac{3}{4}}. \tag{B.29}$$

Letting $x_1 = u_1 - \frac{1}{4}r_3$, the optimal bandwidth is given approximately by

$$h_{\text{ew}} = x_1^{\frac{2}{d}} = (1 + o(1))\; u_1^{\frac{2}{d}} = (1 + o(1))\; a_1^{-\frac{1}{2d}} t_n^{-\frac{3}{2d}} \tag{B.30}$$

recalling the notation $x = \sqrt{h^d}$, where $a_1 = \frac{\sqrt{2}K^{(3)}(0)}{3}\left(\sqrt{\int K^2(u)du}\right)^{-3} c(\pi)$ is as defined before.

Thus, **equation (3.9)** of the submission has been derived.

B.2. Derivation of (3.10) in the submission

In order to show that h_{ew} is the maximizer of the power function $\beta_n(h)$ at $h = h_{\text{ew}}$ we need to verify that

$$\beta_n''(x)\big|_{x=\sqrt{h_{\text{ew}}^d}} < 0 \tag{B.31}$$

for at least sufficiently large n.

Let $x_0 = \sqrt{h_{\text{ew}}^d}$. Using $\beta'(x_0) = 0$ and equation (B.13), we have

$$\beta''(x_0) = \phi(z_\alpha + a_2 x_0)\left(4\theta_4 x_0^3 + 3\theta_3 x_0^2 + 2\theta_2 x_0 + \theta_1\right). \tag{B.32}$$

We thus have

$$\beta''(x)|_{x=\sqrt{h_{\mathrm{ew}}^d}} = \left[\phi(z_\alpha + a_2 x)\left(4\theta_4 x^3 + 3\theta_3 x^2 + 2\theta_2 x + \theta_1\right)\right]|_{x=\sqrt{h_{\mathrm{ew}}^d}}. \tag{B.33}$$

In order to verify (B.31), it suffices to show that

$$\left[4\theta_4 x^3 + 3\theta_3 x^2 + 2\theta_2 x + \theta_1\right]|_{x=\sqrt{h_{\mathrm{ew}}^d}} < 0, \tag{B.34}$$

which follows immediately from the fact that $\theta_4 x^3$ is the dominant term.

B.3. Comments on (3.11) in the submission

As given in (3.11), we have

$$\widehat{h}_{\mathrm{ew}} = \widehat{a}_1^{-\frac{1}{2d}} \widehat{t}_n^{-\frac{3}{2d}}, \tag{B.35}$$

where

$$\begin{aligned}
\widehat{t}_n &= n\widehat{C}_n^2 \quad \text{with} \quad \widehat{C}_n^2 = \frac{\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}_n^2(X_i)\widehat{\pi}(X_i)}{\widehat{\mu}_2\sqrt{2\widehat{\nu}_2 \int K^2(v)dv}} \quad \text{and} \\
\widehat{a}_1 &= \frac{\sqrt{2}K^{(3)}(0)}{3\left(\sqrt{\int K^2(u)du}\right)^3}\,\widehat{c}(\pi) \quad \text{with} \quad \widehat{c}(\pi) = \frac{\frac{1}{n}\sum_{i=1}^n \widehat{\pi}^2(X_i)}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^n \widehat{\pi}(X_i)}\right)^3},
\end{aligned}$$

in which $\widehat{\mu}_2$, $\widehat{\nu}_2$ and $\widehat{\pi}(\cdot)$ are as defined in (2.7) of the submission, and $\widehat{\Delta}_n(x)$ is given by

$$\widehat{\Delta}_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\mathrm{cv}}}\right)\left(Y_i - m_{\widehat{\theta}}(X_i)\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\mathrm{cv}}}\right)}$$

In real data application, $\widehat{\Delta}_n(x)$ can be computed directly from using the data. One may then imply whether there is any departure and how significant the departure is. In simulation study, computing $\widehat{h}_{\mathrm{ew}}$ requires the availability of the data $\{Y_i\}$ generated under either $\mathcal{H}_0$ or $\mathcal{H}_1$. As a result, the value of $\widehat{h}_{\mathrm{ew}}$ under $\mathcal{H}_0$ can be different from that of $\widehat{h}_{\mathrm{ew}}$ under $\mathcal{H}_1$ as they should be.

Note that under $\mathcal{H}_0$: $\Delta_n(\cdot) \equiv 0$

$$\widehat{\Delta}_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\mathrm{cv}}}\right)\left(Y_i - m_{\widehat{\theta}}(X_i)\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\mathrm{cv}}}\right)} = \frac{\frac{1}{n\widehat{b}_{\mathrm{cv}}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\mathrm{cv}}}\right)\left(m_{\theta_0}(X_i) - m_{\widehat{\theta}}(X_i)\right)}{\frac{1}{n\widehat{b}_{\mathrm{cv}}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\mathrm{cv}}}\right)}$$

$$
\begin{aligned}
&+\frac{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\ \Delta_n(X_i)}{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)}+\frac{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\ e_i}{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)}\\
&=\frac{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\left(m_{\theta_0}(X_i)-m_{\widehat{\theta}}(X_i)\right)}{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)}+\frac{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\ e_i}{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)}.
\end{aligned}
$$

Under $\mathcal{H}_0$: $\Delta_n(\cdot)\equiv 0$, it may be shown that

$$
\widehat{C}_n^2 = D_1\frac{1}{n\widehat{b}_{\rm cv}^d}\ (1+o_P(1)),
$$

where $D_1>0$ is some constant.

As a result, under $\mathcal{H}_0$,

$$
\widehat{h}_{\rm ew} = D_2\ \widehat{b}_{\rm cv}^{\frac{3}{2}}\ (1+o_P(1)),
$$

where $D_2>0$ is also some constant.

This shows that the optimal bandwidth used for computing sizes should be smaller (asymptotically) than the conventional bandwidth $\widehat{b}_{\rm cv}$ used for optimal estimation purposes.

Under $\mathcal{H}_1$: $\Delta_n(\cdot)\neq 0$, it may be shown that

$$
\begin{aligned}
\widehat{\Delta}_n(x) &= \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\left(Y_i-m_{\widehat{\theta}}(X_i)\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)}=\frac{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\left(m_{\theta_1}(X_i)-m_{\widehat{\theta}}(X_i)\right)}{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)}\\
&+\frac{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\ \Delta_n(X_i)}{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)}+\frac{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\ e_i}{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)} \qquad \text{(B.36)}\\
&= o_P(1)+\frac{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\ \Delta_n(X_i)}{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)}+o_P(1)\\
&= \Delta_n(x)+o_P(1).
\end{aligned}
$$

Thus, we have

$$
\widehat{h}_{\rm ew}\approx C_5\ \left(n\ C_n^2\right)^{-\frac{3}{2d}}\quad \text{with}\quad C_n^2=\frac{\int \Delta_n^2(x)\pi^2(x)dx}{\sigma^2\sqrt{2\nu_2\ \int K^2(v)dv}}.
$$

This shows that the order of $\widehat{h}_{\rm ew}$ is proportional to the rate of $\Delta_n(\cdot)$ converging to zero. As in Example 4.1 of the submission, when $\Delta_n(x)$ is specified for the simulation, one may clearly see the relationship between them. For example, when $\Delta_n(x) = c_n\ \Delta(x)$ as in the case in the simulation study in Example 4.1, the order of $\widehat{h}_{\rm ew}$ is proportional to $(n\ c_n^2)^{-\frac{3}{2d}}$.

B.4. Proof of Proposition 3.1

In view of (B.36), Assumption A.2 and the assumption that $\inf_{x\in I\!R^d} |\Delta_n(x)|\sqrt{n\widehat{b}_{\rm cv}^d} \to_P \infty$, we can obtain that uniformly in $x \in D_\pi = \{x \in I\!R^d : \pi(x) > 0\}$,

$$\begin{aligned}
\frac{\widehat{\Delta}_n(x)}{\Delta_n(x)} - 1 &= \frac{\widehat{\Delta}_n(x) - \Delta_n(x)}{\Delta_n(x)} = \frac{1}{\Delta_n(x)} \sum_{i=1}^{n} W_{ni}(x)\left(\Delta_n(X_i) - \Delta_n(x)\right) \\
&= \frac{1}{\Delta_n(x)} \sum_{i=1}^{n} W_{ni}(x)\left((X_i - x)^\tau\, \Delta_n'(\xi)\right) + o_P(1) \\
&= \frac{1}{\Delta_n(x)} \frac{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)\left((X_i - x)^\tau\, \Delta_n'(\xi)\right)}{\frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{i=1}^n K\left(\frac{x-X_i}{\widehat{b}_{\rm cv}}\right)} + o_P(1) \\
&= -\frac{\mathbf{1}^\tau \Delta_n'(x)}{\Delta_n(x)}(1 + o_P(1))\ \widehat{b}_{\rm cv}^d \int uK(u)du + o_P(1) = o_P(1) \qquad \text{(B.37)}
\end{aligned}$$

using Taylor expansions to $\Delta_n(x)$, where ξ is between x and X_i, and $\mathbf{1} = (1, 1, \cdots, 1)^\tau$ denotes the identity vector.

Let $\widehat{B}_n^2 = \widehat{\mu}_2\ \sqrt{2\widehat{\nu}_2\ \int K^2(v)dv}$, $B^2 = \sigma^2\sqrt{2\nu_2\ \int K^2(v)dv}$, $\widetilde{C}_n^2 = \frac{1}{n}\sum_{i=1}^n \widehat{\Delta}_n^2(X_i)\pi(X_i)$ and $\overline{C}_n^2 = \frac{1}{n}\sum_{i=1}^n \Delta_n^2(X_i)\pi(X_i)$.

Because of the conventional convergence, in order to show that $\frac{\widehat{C}_n^2}{C_n^2} \to_P 1$, it suffices to show that for n large enough

$$\begin{aligned}
\left|\frac{\widetilde{C}_n^2}{\overline{C}_n^2} - 1\right| &= \frac{\frac{1}{n}\left|\sum_{i=1}^n \left(\widehat{\Delta}_n^2(X_i) - \Delta_n^2(X_i)\right)\pi(X_i)\right|}{\frac{1}{n}\sum_{i=1}^n \Delta_n^2(X_i)\pi(X_i)} = \frac{\left|\int \left(\widehat{\Delta}_n^2(x) - \Delta_n^2(x)\right)\pi^2(x)dx + o_P(1)\right|}{\int \Delta_n^2(x)\pi^2(x)dx + o_P(1)} \\
&= \frac{\left|\int \left(\frac{\widehat{\Delta}_n^2(x)}{\Delta_n^2(x)} - 1\right)\Delta_n^2(x)\pi^2(x)dx + o_P(1)\right|}{\int \Delta_n^2(x)\pi^2(x)dx + o_P(1)} \\
&\le \sup_{x\in D_\pi}\left|\frac{\widehat{\Delta}_n^2(x)}{\Delta_n^2(x)} - 1\right| \frac{\int \Delta_n^2(x)\pi^2(x)dx + o_P(1)}{\int \Delta_n^2(x)\pi^2(x)dx + o_P(1)} \\
&\le \sup_{x\in D_\pi}\left|\frac{\widehat{\Delta}_n^2(x)}{\Delta_n^2(x)} - 1\right| = o_P(1) \qquad \text{(B.38)}
\end{aligned}$$

using (B.37).

In view of the Edgeworth expansion of $\beta_n(h)$ in Theorem 2.2(ii), in order to prove Proposition 3.1, it suffices to show that as $n \to \infty$

$$\frac{\widehat{h}_{\rm ew}}{h_{\rm ew}} - 1 \to_P 0, \tag{B.39}$$

which follows from using Taylor expansions as follows:

$$\begin{aligned}
\frac{\widehat{h}_{\rm ew}}{h_{\rm ew}} - 1 &= \left(\frac{t_n}{\widehat{t}_n}\right)^{\frac{3}{2d}} - 1 + o_P(1) = \left(\frac{C_n}{\widehat{C}_n}\right)^{\frac{3}{2d}} - 1 + o_P(1) \\
&= \left(1 + \frac{C_n}{\widehat{C}_n} - 1\right)^{\frac{3}{2d}} - 1 + o_P(1) \\
&= \frac{3}{2d}\left(\frac{C_n}{\widehat{C}_n} - 1\right)(1 + o_P(1)) = o_P(1) \qquad \text{(B.40)}
\end{aligned}$$

following (B.38). This completes the proof of Proposition 3.1.

B.5. Justification of the use of (4.12)

Similarly to existing proofs (Lemma A.2 of Li 1999 for example), it can be shown that $\widehat{T}_n(h)$ of (2.9) can be approximated by $\widetilde{T}_n(h)$ of (4.12) in the following form:

$$\widehat{T}_n(h) = \widetilde{T}_n(h) + o_P\left(\sqrt{h^d}\right). \tag{B.41}$$

As $\widetilde{T}_n(h)$ is invariant to the form of $\sigma(\cdot)$ and thus does not involve a consistent estimator of $\sigma(\cdot)$, it is used for the continuous–time model case.

B.6. Justification of the applicability of the proposed theory and methodology to model (4.10) and the equations (4.13) and (4.14)

By examining the derivation of (3.9) of the submission, it can be shown that

$$h_{\rm ew} = x_{\rm ew}^{\frac{2}{d}} = a_1^{-\frac{1}{2d}} t_n^{-\frac{3}{2d}}, \tag{B.42}$$

where $a_1 = \frac{\sqrt{2}K^{(3)}(0)}{3}\left(\sqrt{\int K^2(u)du}\right)^{-3} c(\pi)$ with

$$c(\pi) = \int \pi^3(x)\sigma^6(x)dx \cdot \left(\sqrt{\int \pi^2(x)\sigma^4(x)dx}\right)^{-3}.$$

Thus, equation (3.11) of the submission becomes

$$\widehat{h}_{\rm ew} = \widehat{a}_1^{-\frac{1}{2d}} \widehat{t}_n^{-\frac{3}{2d}}, \tag{B.43}$$

where $\widehat{t}_n = n\widehat{C}_n^2$, $\widehat{a}_1 = \frac{\sqrt{2}K^{(3)}(0)}{3}\left(\sqrt{\int K^2(u)du}\right)^{-3}\widehat{c}(\pi)$,

$$\widehat{c}(\pi) = \frac{1}{n}\sum_{i=1}^n \widehat{\pi}^2(X_i)\widehat{\sigma}^6(X_i) \cdot \left(\frac{1}{n}\sum_{i=1}^n \widehat{\pi}(X_i)\widehat{\sigma}^4(X_i)\right)^{-\frac{3}{2}},$$

in which $\widehat{\pi}(x) \cdot \widehat{\sigma}^2(x) = \frac{1}{n\widehat{b}_{\rm cv}^d}\sum_{s=1}^n K\left(\frac{x-X_s}{\widehat{b}_{\rm cv}}\right)\widehat{e}_s^2$ with $\widehat{e}_t = Y_t - m_{\widehat{\theta}}(X_t)$.

Before checking the proofs of the corresponding versions of Theorems 2.1–2.3 for model (4.10), we need to note that the leading term $Q_n(h)$ in (2.7) becomes

$$Q_n(h) = \frac{1}{n\sqrt{h^d}\sigma_n}\sum_{i=1}^n\sum_{j=1,\neq i}^n \epsilon_i\ \sigma(X_i)K\left(\frac{X_i - X_j}{h}\right)\sigma(X_j)\ \epsilon_j. \tag{B.44}$$

Since $a_{ij} = \sigma(X_i)K\left(\frac{X_i-X_j}{h}\right)\sigma(X_j)$ is still a symmetric function of (X_i, X_j) and $\{\epsilon_i\}$ is still a sequence of i.i.d. errors, it can be checked that the corresponding versions of Lemmas A.1–A.4 remain true.

Appendix C

C.1. Justification of Assumptions A.1–A.3

Assumption A.1 is quite standard in this kind of discussion. Similar conditions have been used in Li (1999) for example. Assumption A.1(i) can be extended to the case where $e_i = \sigma(X_i)\epsilon_i$ as discussed in (4.10).

Assumption A.2(i) is required to ensure that the parametric estimator $\widehat{\theta}$ is a $\sqrt{n}$–consistent estimator of θ_0. In addition, Assumption A.2(ii) requires that $\widehat{\theta}$ may also be considered as a $\sqrt{n}$–consistent estimator to θ_1. This is achievable in this kind of local alternative case as in (1.2). For example, in the univariate linear model case where $m_\theta(x) = \beta x$ and $m(x) = m_{\theta_0}(x) = \beta_0 x$ under $\mathcal{H}_0$ and $m(x) = m_{\theta_1}(x) + \Delta_n(x) = \beta_1 x + \Delta_n(x)$ under $\mathcal{H}_1$, the conventional least squares estimator

$$\widehat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

is a $\sqrt{n}$–consistent estimator of both β_0 and β_1 when the conditions $E[X_1\Delta_n(X_1)] = 0$ and $E[X_1 e_1] = 0$ are satisfied. This is because

$$\begin{aligned}
\widehat{\beta} - \beta_0 &= \frac{\sum_{i=1}^n X_i e_i}{\sum_{i=1}^n X_i^2} = \frac{E[X_1 e_1]}{E[X_1^2]} + O_P(n^{-1/2}) \\
&= O_P(n^{-1/2}) \ \text{ under } \mathcal{H}_0, \text{ and} \\
\widehat{\beta} - \beta_1 &= \frac{\sum_{i=1}^n X_i \Delta_n(X_i)}{\sum_{i=1}^n X_i^2} + \frac{\sum_{i=1}^n X_i e_i}{\sum_{i=1}^n X_i^2} \\
&= \frac{E[X_1\Delta_n(X_1)]}{E[X_1^2]} + \frac{E[X_1 e_1]}{E[X_1^2]} + O_P(n^{-1/2}) \\
&= O_P(n^{-1/2}) \quad \text{under} \quad \mathcal{H}_1
\end{aligned}$$

using both $E[X_1\Delta_n(X_1)] = 0$ and $E[X_1 e_1] = 0$.

Alternatively, one may use a semiparametric estimator $\overline{\theta}$ of the form

$$\overline{\beta} = \frac{\sum_{i=1}^n \widetilde{X}_i \widetilde{Y}_i}{\sum_{i=1}^n \widetilde{X}_i^2}, \tag{C.1}$$

where $\widetilde{X}_i = X_i - \sum_{j=1}^n W_{nj}(X_i) X_j$, $\widetilde{Y}_i = Y_i - \sum_{j=1}^n W_{nj}(X_i) Y_j$ and $W_{ni}(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)}$. In this case, unlike the case where there are two different regressors $\{X_i\}$ and $\{V_i\}$ and the conventional semiparametric estimation is applicable to a partially linear model of the form $Y_i = X_i\beta + g(V_i) + e_i$ discussed in Härdle *et al* (2000), $\overline{\beta}$ is not a $\sqrt{n}$–consistent estimator of β.

Assumption A.3 is needed to ensure that the power of the proposed test goes to one when $n \to \infty$. To ensure that the optimal bandwidth $\widehat{h}_{\text{ew}}$ satisfies Assumption A.1(v), Assumption A.3 is needed to impose certain rate of $C_n^2 \to 0$. In the case where $h = \widehat{h}_{\text{ew}}$, the assumptions that $\lim_{n\to\infty} nh^d = \infty$ and $\lim_{n\to\infty} n\sqrt{h^d}\, C_n^2 = \infty$ reduce to

$$\lim_{n\to\infty} n\, C_n^2 = \infty \ \text{ and } \ \lim_{n\to\infty} n\, C_n^6 = 0.$$

Assumption A.3 holds in most cases, including the case where the marginal density function $\pi(\cdot)$ either has compact support or satisfies $\lim_{x\to\infty} \Delta_n(x)\pi(x) = 0$.

C.2. Proofs of Lemmas A.1–A.4

In order to establish some useful lemmas without including non–essential technicality, we introduce the following simplified notation:

$$a_{ij} = \frac{1}{n\sqrt{h^d}\sigma_n} K\left(\frac{X_i - X_j}{h}\right), \quad L_n(h) = \sum_{i=1}^{n} \sum_{j=1,\neq i}^{n} a_{ij} e_i e_j,$$

$$\rho(h) = \frac{\sqrt{2}K^{(3)}(0)\ \int \pi^3(u)du}{3} \left(\sqrt{\int \pi^2(u)du\ \int K^2(v)dv}\right)^{-3} \sqrt{h^d}. \qquad \text{(C.2)}$$

To simplify the proof of Lemma A.1 below, we use the result that

$$\lim_{n\to\infty} \sigma_n^2 = 2\mu_2^2 \nu_2 \int K^2(u)du \equiv \sigma_0^2$$

and then assume without loss of generality that $\sigma_0 \equiv 1$.

We now have the following lemma.

LEMMA A.1. *Suppose that the conditions of Theorem 2.1(i) hold. Then for any h*

$$\sup_{x\in \mathbb{R}^1} \left| P\left(L_n(h) \le x\right) - \Phi(x) + \rho(h)\ (x^2 - 1)\ \phi(x) \right| = O\left(h^d\right). \qquad \text{(C.3)}$$

PROOF: The proof is based on a non–trivial application of Theorem 1.1 of Götze, Tikhomirov and Yurchenko (2004) listed in Appendix D below. As the proof itself is extremely technical, we provide only an outline below.

In view of the form of $L_n(h)$, we need to follow the proofs of Theorems 1.1 and 3.1 as well as Lemmas 3.2–3.5 of Götze, Tikhomirov and Yurchenko (2004) step by step to finish the proof of Lemma A.1. Note that their proofs of Theorems 1.1 and 3.1 remain true. The proofs of Lemmas 3.2–3.5 also remain true by successive conditioning arguments when needed.

We may also apply Lemma D.1 below to the conditional probability $P\left(L_n(h) \le x | \mathcal{X}_n\right)$ and then use the dominated convergence theorem to deduce (A.4) unconditionally.

To avoid repeating the conditioning argument (given $\mathcal{X}_n$) for each case in the following derivations, the corresponding conditioning arguments are all understood to be held in probability with respect to the joint distribution of $\mathcal{X}_n = (X_1, \cdots, X_n)$.

In any case, in order to apply Lemma D.1 listed in the Appendix D below, we need to verify certain conditions of Lemma D.1.

$$a_{jj} = n^{-1} h^{-d/2} K(0)$$

$$\mathbf{d}^{\tau} = n^{-1}h^{-d/2}K(0)\,(1,\cdots,1)^{\tau}$$

$$\mathbf{Tr}\mathbf{A} = h^{-d/2}K(0)$$

$$V^2 = (nh^d)^{-1}K^2(0)$$

$$\|\mathbf{A_0}\|^2 = n^{-2}h^{-d}\sum_{\substack{s,t=1\\ s\neq t}}^{n} K^2\left(\frac{X_s - X_t}{h}\right)$$

$$\mathbf{d}^{\tau}\mathbf{A_0}\mathbf{d} = n^{-3}h^{-3d/2}K^2(0)\sum_{\substack{s,t=1\\ s\neq t}}^{n} K\left(\frac{x_s - x_t}{h}\right). \tag{C.4}$$

Obviously

$$\begin{aligned}
\mathbf{Tr}(\mathbf{A_0^3}) &= \sum_{q=1}^{n}\sum_{\substack{k=1\\ k\neq q}}^{n}\sum_{\substack{j=1\\ j\neq k\\ j\neq q}}^{n} a_{qk}a_{kj}a_{jq} \\
&= \left(n^{-1}h^{-d/2}\right)^3 \sum_{q=1}^{n}\sum_{\substack{k=1\\ k\neq q}}^{n}\sum_{\substack{j=1\\ j\neq k\\ j\neq q}}^{n} K\left(\frac{X_q - X_k}{h}\right)K\left(\frac{X_k - X_j}{h}\right)K\left(\frac{X_j - X_q}{h}\right)
\end{aligned}$$

Using the stationary ergodic theorem and the α–mixing condition, the sums involving the kernel function K in (C.4) can be approximated as follows:

$$\begin{aligned}
\frac{1}{n^2}\sum_{\substack{j,k=1\\ j\neq k}}^{n} K^2\left(\frac{X_i - X_j}{h}\right) &\approx \int\int K^2\left(\frac{x-y}{h}\right)\pi(x)\pi(y)dxdy \\
&\approx h^d\int\int K^2(u)\pi(y+uh)\pi(y)dudy \\
&\approx h^d\int\int K^2(u)\pi^2(v)dudv,
\end{aligned} \tag{C.5}$$

where $\pi(x)$ is the marginal density function of X_1.

Similarly, for the second sum in expression (C.4)

$$\frac{1}{n^2}\sum_{\substack{s,t=1\\ s\neq t}}^{n} K\left(\frac{X_s - X_t}{h}\right) \approx h^d\int K(u)du\,\int \pi^2(v)dv. \tag{C.6}$$

For the triple sum in expression (C.5) we find

$$\frac{1}{n^3}\sum_{q=1}^{n}\sum_{\substack{k=1\\ k\neq q}}^{n}\sum_{\substack{j=1\\ j\neq k\\ j\neq q}}^{n} K\left(\frac{X_q - X_k}{h}\right)K\left(\frac{X_k - X_j}{h}\right)K\left(\frac{X_j - X_q}{h}\right)$$

$$
\begin{aligned}
&\approx \int\int\int K\left(\frac{x-y}{h}\right) K\left(\frac{y-z}{h}\right) K\left(\frac{z-x}{h}\right) \pi(x)\pi(y)\pi(z)dxdydz \\
&\approx h^{2d} \int\int\int K(-(u+v))K(v)K(u)\pi(z-uh)\pi(z+vh)\pi(z)dudvdz \\
&\approx h^{2d} \int\int\int K(-(u+v))K(v)K(u)\pi^3(z)dudvdz \\
&= h^{2d} \int\int K(u+v)K(v)K(u)dudv \int \pi^3(z)dz \\
&= h^{2d} \int \left(\int K(w)K(w-v)dw\right) K(v)dv \ \int \pi^3(z)dz \\
&= h^{2d} \int K * K(v)K(v)dv \ \int \pi^3(u)du \\
&= h^{2d} \, (K*K*K)(0) \ \int \pi^3(u)du. \qquad \text{(C.7)}
\end{aligned}
$$

Combining (C.4)—(C.7) we obtain the following approximate behaviours

$$
\begin{aligned}
&\mathbf{TrA} \approx h^{-d/2}K(0) \\
&V^2 \approx n^{-1}h^{-d}K^2(0) \\
&\|\mathbf{A_0}\|^2 \approx \int K^2(u)du \ \int \pi^2(v)dv \\
&\mathbf{d}^\tau \mathbf{A_0 d} \approx n^{-1}h^{-d/2}K^2(0) \int K(u)du \ \int \pi^2(v)dv \\
&\mathbf{Tr}(\mathbf{A_0^3}) \approx h^{d/2} \, K^{(3)}(0) \ \int \pi^3(u)du, \qquad \text{(C.8)}
\end{aligned}
$$

where we denoted $K^{(3)}(\cdot) = (K*K*K)(\cdot)$ the three times convolution of K with itself.

From this we get approximations for the quantities σ_*^2 and κ involved in Lemma B.1 below:

$$
\begin{aligned}
\sigma_*^2 &\approx n^{-1}h^{-d}(\mu_4-\mu_2^2)K^2(0) + 2\mu_2^2 \int K^2(u)du \ \int \pi^2(v)dv \\
\kappa &\approx \frac{\frac{\mu_3^2 K^2(0)}{nh^d} + \frac{4\mu_2^3 \ \int \pi^3(u)du}{3} K^{(3)}(0)}{\sigma_*^3} \ \sqrt{h^d} \\
&\approx \ \frac{\sqrt{2}K^{(3)}(0)}{3} \left(\sqrt{\int K^2(u)du}\right)^{-3} \ c(\pi) \ \sqrt{h^d} \equiv \rho(h),
\end{aligned}
$$

where $c(\pi) = \frac{\int \pi^3(x)dx}{\left(\sqrt{\int \pi^2(x)dx}\right)^3}$

In order to apply Lemma D.1 to finish the proof, we need to show that the upperbound of Lemma D.1 tends to 0 as $n \to \infty$. Observe that

$$\begin{aligned} \|\mathbf{A}\|^2 &= \|\mathbf{A_0}\|^2 + \sum_{j=1}^{n} a_{jj}^2 \\ &\approx \int K^2(u)du \int \pi^2(v)dv + (nh^d)^{-1}K^2(0). \end{aligned} \tag{C.9}$$

Similar to (C.7), we may show that

$$\begin{aligned} \sum_{t=1}^{n} \mathcal{L}_t^4 &= \sum_{s=1}^{n}\left(\sum_{t=1}^{n} a_{st}^2\right)^2 = \sum_{s=1}^{n}\sum_{t=1}^{n} a_{st}^4 + \sum_{s=1}^{n}\sum_{t_1=1}^{n}\sum_{t_2=1,\neq t_1}^{n} a_{st_1}^2 a_{st_2}^2 \\ &= \frac{1}{n^2h^d}\int K^4(u)du \int \pi^2(v)dv \\ &+ \frac{1}{nh^d}\int\int K^2(w)K^2\left(w + \frac{u-v}{h}\right)\pi(u)\pi(v)dwdudv \\ &= \frac{1}{n^2h^d}\int K^4(u)du \int \pi^2(v)dv + \frac{1}{n}K_2^{(2)}(0)\int \pi^2(v)dv, \end{aligned} \tag{C.10}$$

where $K_2^{(2)}(0)$ is the two–time convolution of $K^2(\cdot)$ with itself.

Similarly, we may show that as $n \to \infty$

$$\lambda_1 \leq \max_{1\leq j\leq n}\sum_{i=1}^{n}|a_{ij}| \leq \sqrt{h^d}\int K(u)du \int \pi^2(v)dv. \tag{C.11}$$

Consequently, using that $h \to 0$ and $nh^d \to \infty$, we find that

$$\frac{|\lambda_1|^2}{\|\mathbf{A}\|^2} \approx \frac{h^d \int \pi^2(v)dv}{\int K^2(u)du}. \tag{C.12}$$

From (C.7)–(C.10) we then find that

$$\frac{\left(\sum_{t=1}^{n}\mathcal{L}_t^4\right)^{1/2}}{\|\mathbf{A}\|^2} \approx \frac{\sqrt{K_2^{(2)}(0)\int \pi^2(v)dv}}{\sqrt{n}}. \tag{C.13}$$

Thus, (C.12) and (C.13) imply that there is some constant C_∞ such that

$$M \approx C_\infty h^d, \tag{C.14}$$

which shows that the upperbound in Lemma B.1 tends to 0 at a rate proportional to h^d. This completes the proof of Lemma A.1.

Recall $L_n(h) = \sum_{t=1}^n \sum_{s=1,\neq t}^n e_s \ a_{st} e_t$ as defined before and let

$$\begin{aligned} \overline{T}_n(h) &= \frac{h^{\frac{d}{2}}}{n\sigma_n} \sum_{i=1}^n \sum_{j=1,\neq i}^n \widehat{e}_i \ K_h(X_i - X_j) \ \widehat{e}_j = \frac{h^{\frac{d}{2}}}{n\sigma_n} \sum_{i=1}^n \sum_{j=1,\neq i}^n e_i \ K_h(X_i - X_j) \ e_j \\ &+ \frac{h^{\frac{d}{2}}}{n\sigma_n} \sum_{i=1}^n \sum_{j=1,\neq i}^n K_h(X_i - X_j) \left[m(X_i) - m_{\widehat{\theta}}(X_i)\right] \left[m(X_j) - m_{\widehat{\theta}}(X_j)\right] \\ &+ \frac{2h^{\frac{d}{2}}}{n\sigma_n} \sum_{i=1}^n \sum_{j=1,\neq i}^n e_i \ K_h(X_i - X_j) \left[m(X_j) - m_{\widehat{\theta}}(X_j)\right] \\ &\equiv L_n(h) + S_n(h) + D_n(h), \end{aligned} \tag{C.15}$$

where

$$S_n(h) = \frac{h^{\frac{d}{2}}}{n\sigma_n} \sum_{i=1}^n \sum_{j=1,\neq i}^n K_h(X_i - X_j) \left[m(X_i) - m_{\widehat{\theta}}(X_i)\right] \left[m(X_j) - m_{\widehat{\theta}}(X_j)\right],$$

$$D_n(h) = \frac{2h^{\frac{d}{2}}}{n\sigma_n} \sum_{i=1}^n \sum_{j=1,\neq i}^n e_i \ K_h(X_i - X_j) \left[m(X_j) - m_{\widehat{\theta}}(X_j)\right]. \tag{C.16}$$

We then define $L_n^*(h)$, $S_n^*(h)$ and $D_n^*(h)$ as the corresponding versions of $L_n(h)$, $S_n(h)$ and $D_n(h)$ involved in (A.5) with (X_t, Y_t) and $\widehat{\theta}$ being replaced by (X_t, Y_t^*) and $\widehat{\theta}^*$ respectively.

LEMMA A.2. *Suppose that the conditions of Theorem 2.2(i) hold. Then the following*

$$\sup_{x \in I\!R^1} \left| P^* \left(L_n^*(h) \leq x\right) - \Phi(x) + \rho(h) \ (x^2 - 1) \ \phi(x) \right| = O_P\left(h^d\right) \tag{C.17}$$

holds in probability.

PROOF: Since the proof follows similarly from that of Lemma A.1 using conditioning arguments given $\mathcal{W}_n = \{(X_i, Y_i) : 1 \leq i \leq n\}$, we do not wish to repeat the details.

LEMMA A.3. (i) *Suppose that the conditions of Theorem 2.2(ii) hold. Then under $\mathcal{H}_0$*

$$E\left[S_n(h)\right] = O\left(\sqrt{h^d}\right) \quad \text{and} \quad E\left[D_n(h)\right] = o\left(\sqrt{h^d}\right). \tag{C.18}$$

(ii) *Suppose that the conditions of Theorem 2.2(ii) hold. Then under $\mathcal{H}_0$*

$$E^*\left[S_n^*(h)\right] = O_P\left(\sqrt{h^d}\right) \quad \textit{and} \quad E^*\left[D_n^*(h)\right] = o_P\left(\sqrt{h^d}\right) \tag{C.19}$$

in probability with respect to the joint distribution of $\mathcal{W}_n$*, where* $E^*[\cdot] = E[\cdot|\mathcal{W}_n]$.

(iii) *Suppose that the conditions of Theorem 2.2(i) hold. Then under* $\mathcal{H}_0$

$$E\left[S_n(h)\right] - E^*\left[S_n^*(h)\right] = O_P\left(\sqrt{h^d}\right) \quad \text{and} \quad E\left[D_n(h)\right] - E^*\left[D_n^*(h)\right] = o_P\left(\sqrt{h^d}\right) \quad \text{(C.20)}$$

in probability with respect to the joint distribution of $\mathcal{W}_n$.

PROOF: As the proofs of (i)–(iii) are quite similar, we need only to prove the first part of (iii). In view of the definition of $\{a_{st}\}$ and (A.6), we have

$$\begin{aligned} S_n(h) &= \sum_{t=1}^n \sum_{s=1,\neq t}^n \left(m(X_s) - m_{\widehat{\theta}}(X_s)\right) a_{st} \left(m(X_t) - m_{\widehat{\theta}}(X_t)\right), \\ S_n^*(h) &= \sum_{t=1}^n \sum_{s=1,\neq t}^n \left(m(X_s) - m_{\widehat{\theta}^*}(X_s)\right) a_{st} \left(m(X_t) - m_{\widehat{\theta}^*}(X_t)\right). \end{aligned} \quad \text{(C.21)}$$

Ignoring the higher–order terms, it can be shown that the leading term of $S_n^*(h) - S_n(h)$ is represented approximately by

$$S_n^*(h) - S_n(h) = (1 + o_P(1)) \sum_{t=1}^n \sum_{s=1,\neq t}^n \left(m_{\widehat{\theta}}(X_s) - m_{\widehat{\theta}^*}(X_s)\right) a_{st} \left(m_{\widehat{\theta}}(X_t) - m_{\widehat{\theta}^*}(X_t)\right). \quad \text{(C.22)}$$

Using (C.22), Assumption A.2 and the fact that

$$\begin{aligned} E[a_{st}] &= \frac{1}{n\sqrt{h^d}\sigma_n} E\left[K\left(\frac{X_s - X_t}{h}\right)\right] \\ &= \frac{\sqrt{h^d}}{n\sigma_n} \int K(u)du \int \pi^2(v)dv = \frac{\sqrt{h^d}}{n\sigma_n} \int \pi^2(v)dv, \end{aligned} \quad \text{(C.23)}$$

we can deduce that

$$E[S_n(h)] - E^*[S_n^*(h)] = O_P\left(\sqrt{h^d}\right), \quad \text{(C.24)}$$

which completes an outline of the proof.

LEMMA A.4. *Suppose that the conditions of Theorem 2.2(iii) hold. Then under* $\mathcal{H}_1$

$$\lim_{n\to\infty} E\left[S_n(h)\right] = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{E\left[D_n(h)\right]}{E\left[S_n(h)\right]} = 0. \quad \text{(C.25)}$$

Proof: In view of the definitions of and $S_n(h)$ and $D_n(h)$, we need only to show the first part of (A.11). Observe that for θ_1 defined in the second part of Assumption A.2(ii),

$$\begin{aligned} S_n(h) &= \sum_{t=1}^{n}\sum_{s=1,\neq t}^{n}\left(m(X_s) - m_{\widehat{\theta}}(X_s)\right) a_{st}\left(m(X_t) - m_{\widehat{\theta}}(X_t)\right) \\ &= \sum_{t=1}^{n}\sum_{s=1,\neq t}^{n}\left(m(X_s) - m_{\theta_1}(X_s)\right) a_{st}\left(m(X_t) - m_{\theta_1}(X_t)\right) \\ &+ \sum_{t=1}^{T}\sum_{s=1,\neq t}^{T}\left(m_{\theta_1}(X_s) - m_{\widehat{\theta}}(X_s)\right) a_{st}\left(m_{\theta_1}(X_t) - m_{\widehat{\theta}}(X_t)\right) \\ &+ o_P\left(S_n(h)\right). \end{aligned} \tag{C.26}$$

In view of (C.26), using the second part of Assumption A.2(ii), in order to prove (A.11) it suffices to show that for $n \to \infty$ and $h \to 0$,

$$E\left[\sum_{t=1}^{n}\sum_{s=1,\neq t}^{n}\left(m(X_s) - m_{\theta_1}(X_s)\right) a_{st}\left(m(X_t) - m_{\theta_1}(X_t)\right)\right] \to \infty. \tag{C.27}$$

Simple calculations imply that as $n \to \infty$ and $h \to 0$

$$\begin{aligned} &E\left[\sum_{t=1}^{n}\sum_{s=1,\neq t}^{n}\left(m(X_s) - m_{\theta_1}(X_s)\right) a_{st}\left(m(X_t) - m_{\theta_1}(X_t)\right)\right] \\ &= c_n^2 E\left[\sum_{t=1}^{n}\sum_{s=1,\neq t}^{n}\Delta_n(X_s)a_{st}\Delta_n(X_t)\right] \\ &= \sigma_n^{-1}(1+o(1))\ \sqrt{h^d}\ n\ \int K(u)du\ \int \Delta_n^2(v)\pi^2(v)dv \\ &= \sigma_n^{-1}(1+o(1))\ n\ \sqrt{h^d}\ \int \Delta_n^2(v)\pi^2(v)dv \to \infty \end{aligned} \tag{C.28}$$

using Assumption A.3, where σ_n is as defined before.

C.3. Proof of Theorem 2.3

Define $F_{n,h}(x)$ and $F_{n,h}^*(x)$ as the exact finite–sample distributions of $\widehat{T}_n(h)$ and $\widehat{T}_n^*(h)$, respectively. Using existing results (Serfling 1980; Hall 1992) and Theorem 2.2(i) imply

$$l_\alpha - z_\alpha = \frac{\Phi(z_\alpha) - F_{n,h}(l_\alpha)}{\phi(z_\alpha)} + o_P\left(|l_\alpha - z_\alpha|\right) = \frac{1}{\phi(z_\alpha)}\ \left((z_\alpha^2 - 1)\ \phi(z_\alpha)\ a_1\sqrt{h^d}\right)$$

$$
\begin{aligned}
& + \quad o_P\left(|l_\alpha - z_\alpha|\right) = b_1 \sqrt{h^d} + o_P\left(|l_\alpha - z_\alpha|\right), \\
l_\alpha^* - z_\alpha \quad & = \quad \frac{\Phi(z_\alpha) - F_{n,h}^*(l_\alpha^*)}{\phi(z_\alpha)} + o_P\left(|l_\alpha^* - z_\alpha|\right) = \frac{1}{\phi(z_\alpha)} \left((z_\alpha^2 - 1)\ \phi(z_\alpha)\ a_1 \sqrt{h^d}\right) \\
& + \quad o_P\left(|l_\alpha^* - z_\alpha|\right) = b_1 \sqrt{h^d} + o_P\left(|l_\alpha^* - z_\alpha|\right), \qquad \text{(C.29)}
\end{aligned}
$$

where a_1 and b_1 are as defined above Theorem 2.3. The proof is now finished.

Appendix D

Götze, Tikhomirov and Yurchenko (2004) consider a quadratic sum of the form

$$
Q_n = \sum_{j=1}^{n} a_{jj}(X_j^2 - E[X_j^2]) + \sum_{1 \le j \ne k \le n} a_{jk} X_j X_k \ , \qquad \text{(D.1)}
$$

where $X_1, \cdots, X_n$ are independent and identically distributed random variables with $E[X_1] = 0$. $E[X_1^2] < \infty$ and $E[X_1^6] < \infty$. Here $\{a_{ij}\}$ is a sequence of real numbers possibly depending on n. They actually consider a more general setup than this, but for simplicity we briefly present this simplified form.

The following notations are needed.

$\mathbf{A} = (a_{jk})_{j,k=1}^{n}$: $n \times n$ matrix containing all coefficients a_{jk}

$\|\mathbf{A}\| = \sum_{j,k=1}^{n} a_{jk}^2$;

$\mathbf{TrA} = \sum_{j=1}^{n} a_{jj}$: the trace of the matrix $\mathbf{A}$;

$V^2 = \sum_{j=1}^{n} a_{jj}^2 \qquad \mathcal{L}_j^2 = \sum_{k=1}^{n} a_{jk}^2,\ j = 1, \cdots, n;$

$\mathbf{d}^\tau = (a_{11}, a_{22}, \cdots, a_{nn}) \in I\!R^n$: n-dimensional column vector containing all diagonal elements of the matrix $\mathbf{A}$;

$\mathbf{A_0}$ denotes the $n \times n$ matrix with elements a_{jk} if $j \ne k$ and equal to 0, if $j = k$;

$\mu_k = E(X_1^k)$, $\beta_k = E|X_1|^k \qquad k = 1, \cdots, 6;$

λ_1 the maximal in absolute value, eigenvalue of the matrix $\mathbf{A}$;

$$M = \max \left(\frac{|\lambda_1|^2}{\|\mathbf{A}\|^2}, \frac{(\sum_{j=1}^{n} \mathcal{L}_j^4)^{\frac{1}{2}}}{\|\mathbf{A}\|^2} \right);$$

$$\sigma_*^2 = (\mu_4 - \mu_2^2)V^2 + 2\mu_2^2 \|\mathbf{A_0}\|^2 \qquad \kappa = \sigma_*^{-3} \left(\mu_3^2 \mathbf{d}^\tau \mathbf{A_0} \mathbf{d} + \tfrac{4}{3}\mu_3^2 \mathbf{Tr} \mathbf{A_0}^3 \right).$$

In the above notations the superscript τ denotes the transposition of a vector or matrix. Note that the matrix $\mathbf{A_0}$ is obtained from the matrix $\mathbf{A}$ by replacing all diagonal elements by 0.

The coefficients in the quadratic form (D.1) should satisfy some conditions:

$\mathbf{Q}$(i): $\|\mathbf{A}\| < \infty$; $\mathbf{Q}$(ii): there exists some absolute positive constant $b_1^2 > 0$ such that

$$1 - \frac{V^2}{\|\mathbf{A}\|^2} \geq b_1^2 \,. \tag{D.2}$$

Note that (D.2) means that the 'non-diagonal' part of $\|\mathbf{A}\|$ is bounded away from zero and has a non-negligible influence on the distribution of Q_n.

Lemma D.1. *Under conditions* $\mathbf{Q}$*(i) and* $\mathbf{Q}$*(ii), we have*

$$\sup_{x \in I\!R^1} \left| P\{Q_n/\sigma_* \leq x\} - \Phi(x) + \kappa \Phi'''(x) \right| \leq C b_1^{-4} \left(\beta_3^2 + V \|\mathbf{A}\|^{-1} \beta_6 \right) \mu_2^{-3} M \,, \tag{D.3}$$

with C being an absolute positive constant and where Φ denotes the cumulative distribution function of the standard normal.

Proof: Its proof is given in the proof of Theorem 1.1 in Götze, Tikhomirov and Yurchenko (2004).

REFERENCES

Aït-Sahalia, Y. (1996), "Nonparametric Pricing of Interest Rate Derivative Securities," *Econometrica*, 64, 527–560.

Andrews, D. W. (1997), "A Conditional Kolmogorov Test," *Econometrica*, 65, 1097–1028.

Arapis, M., and Gao, J. (2006), "Empirical Comparisons in Short–Term Interest Rate Models Using Nonparametric Methods," *Journal of Financial Econometrics*, 4, 310–345.

Chapman, D., and Pearson, N. (2000), "Is the Short Rate Drift Actually Nonlinear ?" *Journal of Finance*, 54, 355–388.

Chen, S. X., Härdle, W. and Li, M. (2003), "An Empirical Likelihood Goodness–of–Fit Test for Time Series," *Journal of the Royal Statistical Society Series B*, 65, 663–678.

Dette, H. (1999), "A consistent text for the functional form of a regression based on a difference of variance estimators, " *The Annals of Statistics*, 27, 1012–1040.

Dette, H., and Spreckelsen, I. (2004), "Some Comments on Specification Tests in Nonparametric Absolutely Regular Processes," *Journal of Time Series Analysis*, 25, 159–172.

Fan, J. (1996), "Test of Significance Based on Wavelet Thresholding and Neyman's Truncation," *Journal of the American Statistical Association*, 434, 674–688.

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

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

Fan, J., and Zhang, C. M. (2003), "A Re-examination of Stanton's Diffusion Estimation with Applications to Financial Model Validation," *Journal of the American Statistical Association*, 461, 118–134.

Fan, J., Zhang, C. M., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *Annals of Statistics*, 29, 153–193.

Fan, Y., and Linton, O. (2003), "Some Higher–Order Theory for a Consistent Nonparametric Model Specification Test," *Journal of Statistical Planning and Inference*, 109, 125–154.

Franke, J., Kreiss, J. P., and Mammen, E. (2002), "Bootstrap of Kernel Smoothing in Nonlinear Time Series," *Bernoulli*, 8, 1–38.

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

González-Manteiga, W. and Cao, R. (1993), "Testing the hypothesis of a general linear model using nonparametric regression estimation," *Test*, 2, 161–188.

Götze, F., Tikhomirov, A. and Yurchenko, V. (2004). Asymptotic expansion in the central limit theorem for quadratic forms. Preprint 04–004, Probability and Statistics, The University of Bielefeld, Germany.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansions.* Springer, New York.

Härdle, W., Liang, H. and Gao, J. (2000). *Partially Linear Models.* Springer Series: Contributions to Statistics. Physica-Verlag, New York.

Härdle, W., and Mammen, E. (1993), "Comparing Nonparametric Versus Parametric Regression Fits," *Annals of Statistics*, 21, 1926–1947.

Hart, J. (1997), *Nonparametric Smoothing and Lack-of-Fit Tests*, New York: Springer.

Hjellvik, V., and Tjøstheim, D. (1995), "Nonparametric Tests of Linearity for Time Series," *Biometrika*, 82, 351–368.

Hjellvik, V., Yao, Q., and Tjøstheim, D. (1998), "Linearity Testing using Local Polynomial Approximation," *Journal of Statistical Planning and Inference*, 68, 295–321.

Horowitz, J. (2003), "Bootstrap Methods for Markov Processes," *Econometrica*, 71, 1049–1082.

Horowitz, J., and Spokoiny, V. (2001). An Adaptive Rate–Optimal Test of a Parametric Mean–Regression Model against a Nonparametric Alternative," *Econometrica*, 69, 599–631.

Hunsberger, S., and Follmann, D. A. (2001), "Testing for Treatment and Interaction Effects in Semiparametric Analysis of Covariance," *Statistics in Medicine*, 20, 1–19.

Juhl, T., and Xiao, Z. (2005), "A Nonparametric Test for Changing Trends," *Journal of Econometrics*, 127, 179–199.

Kulasekera, K. B., and Wang, J. (1997), "Smoothing Parameter Selection for Power Optimality in Testing of Regression Curves," *Journal of the American Statistical Association*, 438, 500–511.

Li, Q. (1999), "Consistent Model Specification Tests for Time Series Econometric Models," *Journal of Econometrics*, 92, 101–147.

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

Li, Q., and Wang, S. (1998), "A Simple Consistent Bootstrap Test for a Parametric Regression Functional Form," *Journal of Econometrics*, 87, 145–165.

Nishiyama, Y., and Robinson, P. (2000), "Edgeworth Expansions for Semiparametric Averaged Derivatives," *Econometrica*, 68, 931–980.

Nishiyama, Y., and Robinson, P. (2005), "The Bootstrap and the Edgeworth Correction for Semiparametric Averaged Derivatives," *Econometrica*, 73, 903–948.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Weihrather, G. (1993), "Testing linear regression model against nonparametric alternatives, " *Metrika*, 40, 367–379.

Zhang, C. M., and Dette, H. (2004), "A Power Comparison Between Nonparametric Regression tests," *Statistics & Probability Letters*, 66, 289–301.

Zheng, J. X. (1996), "A Consistent Test of Functional Form via Nonparametric Estimation Techniques," *Journal of Econometrics*, 75, 263–289.

Zhu, L. (2005), Nonparametric Monte Carlo Tests and Their Applications, *Lecture Notes in Statistics.* 182. New York: Springer.