Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 140 (2007) 802–826
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

A consistent model specification test with mixed discrete and continuous data

Cheng Hsiao[a,b,*], Qi Li[c], Jeffrey S. Racine[d]

[a]*Department of Economics, University of Southern California, Los Angeles, CA 90089, USA*
[b]*Wang Yanan Institute for Studies in Economics, Xiamen University, China*
[c]*Department of Economics, Texas A&M University, College Station, TX 77843-4228, USA*
[d]*Department of Economics, McMaster University, Hamilton, Ont., Canada L8S 4M4*

Available online 28 August 2006

Abstract

In this paper we propose a nonparametric kernel-based model specification test that can be used when the regression model contains both discrete and continuous regressors. We employ discrete variable kernel functions and we smooth both the discrete and continuous regressors using least squares cross-validation (CV) methods. The test statistic is shown to have an asymptotic normal null distribution. We also prove the validity of using the wild bootstrap method to approximate the null distribution of the test statistic, the bootstrap being our preferred method for obtaining the null distribution in practice. Simulations show that the proposed test has significant power advantages over conventional kernel tests which rely upon frequency-based nonparametric estimators that require sample splitting to handle the presence of discrete regressors.

JEL classification: C12; C14

Keywords: Consistent test; Parametric functional form; Nonparametric estimation

1. Introduction

The demand for consistent model specification tests has given rise to a rich literature on the subject. Bierens (1982, 1990) and Eubank and Spiegelman (1990) consider the problem

*Corresponding author. Department of Economics, University of Southern California, Los Angeles, CA 90089, USA. Tel.: +1 213 740 2103; fax: +1 213 740 8543.
E-mail address: chsiao@rcf.usc.edu (C. Hsiao).

0304-4076/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2006.07.015

of testing for correct specification of parametric regression models, Robinson (1989, 1991) considers testing parametric and semiparametric regression models with time series data, Eubank and Hart (1992), Hong and White (1995), and de Jong (1996) propose series-based tests for parametric regression models, while Härdle and Mammen (1993) suggest using the wild bootstrap to better approximate the finite-sample distribution of a kernel test for parametric regression models. More recently, Fan and Li (2000) discuss the relationship between smoothing and nonsmoothing tests, and show that smoothing tests are more powerful than nonsmoothing tests for high frequency alternatives, while Horowitz and Spokoiny (2001) propose an adaptive rate-optimal test for regression models and suggest using several different smoothing parameters to compute a kernel-based test in order to ensure that the test has good power against both the low and high frequency alternatives.[1] Most of this literature relies upon the application of semiparametric and nonparametric methods for detecting relevant structure that has inadvertently been omitted from a parametric model.

Kernel methods constitute one of the most popular approaches towards both nonparametric estimation and the construction of consistent model specification tests. However, to the best of our knowledge, existing kernel-based tests are limited to situations involving continuous regressors only. This is unfortunate because data in the social sciences frequently contains discrete regressors such as family size, gender, choices made by economic agents, and so forth. Consistent model specification tests that are directly applicable in the presence of discrete regressors would clearly be of value to applied researchers.

It is widely known that one can readily generalize existing kernel-based tests to admit discrete regressors by using a conventional frequency estimation method that splits the sample into subsets ('cells'). However, when using this frequency approach in finite-sample applications, there may not be enough observations remaining in each cell to produce reliable nonparametric estimates which has the undesirable effect of a commensurate loss in *finite-sample* efficiency. It is therefore expected that frequency-based procedures would suffer from low power in the presence of discrete regressors due to such sample splitting. However, this need not be the case if one were also to smooth the discrete regressors as is suggested in this paper.

Bierens (1983, 1987) and Ahmad and Cerrito (1994) studied the problem of nonparametric estimation of regression functions in the presence of mixed discrete and continuous data, and both works feature the use of kernel smoothing for discrete regressors. However, Bierens did not consider data driven methods of smoothing parameter selection. Though Ahmad and Cerrito suggested the use of cross-validation (CV) methods, they did not derive theoretical results such as rates of convergence of the CV smoothing parameters for the mixed regressor case. Recently, Hall et al. (2004, 2005), Li and Racine (2004), and Racine and Li (2004) have analyzed the nonparametric estimation of conditional density and regression functions involving mixed discrete and continuous regressors where CV methods were used to choose smoothing parameters, and they derived the rates of convergence of the CV smoothing parameters to their optimal benchmark values along with the asymptotic distributions of the resulting estimators. Moreover, Hall et al. (2005) have shown that the CV method has the

[1] Wooldridge (1992) and Yatchew (1992) are two early papers on the construction of consistent model specification tests using nonparametric methods. See Andrews (1997), Bierens and Ploberger (1997), Donald (1997), and Ait-Sahalia et al. (2001) and the references therein for recent developments in consistent model specification testing.

rather amazing ability to *automatically* remove 'irrelevant regressors' by oversmoothing such regressors, a property not shared by other smoothing parameter selection rules (e.g., plug-in methods). We have found that, for empirical applications involving economic data, irrelevant regressors are surprisingly common. Li and Racine (2006) have shown in addition that, for a variety of empirical data sets (e.g., U.S. patent data, crop yield data, female labor force participation data, marketing data, medical treatment data, etc.), smoothing discrete regressors often leads to substantially better out-of-sample predictions than those generated from conventional sample splitting nonparametric methods or commonly used parametric models. The fact that nonparametric methods can outperform common parametric models simply underscores the fact that common parametric models are often misspecified which itself argues for the development of reliable consistent model specification tests that can handle the mix of discrete and continuous data often present in applied settings.

In this paper we show that the superior performance of the CV nonparametric estimator in mixed data settings carries over to the model specification test setting. We propose a kernel-based test that does not use sample splitting in finite-sample settings building on results found in Racine and Li (2004).[2] We use least squares CV methods for selecting the smoothing parameters for both the discrete and the continuous regressors in the proposed kernel-based test. We demonstrate that the proposed test is substantially more powerful than frequency-based kernel tests in *finite-sample* settings in part because our approach does not suffer from efficiency losses which arise from the use of sample splitting.

It should be mentioned that, though CV methods are optimal for regression function estimation, they are not necessarily optimal for testing procedures. Indeed, some recent literature (e.g., Bierens and Ploberger, 1997) has advocated basing tests on constant bandwidths (i.e., bandwidths that do not shrink to zero as the sample size increases) which can have better power for low frequency data generating processes (DGP) than tests based upon shrinking bandwidths. On the other hand, Fan and Li (2001) clearly demonstrate that tests based upon shrinking bandwidths may be more powerful for high frequency DGPs. Horowitz and Spokoiny (1999) propose a general testing procedure that has good power in the direction of both low and high frequency DGPs. However, when faced with a mix of discrete and continuous data types, if the number of discrete cells is not very small relative to the sample size, a non-smoothing (frequency method) testing approach may have low power when few observations remain in each cell. In such situations, a smoothing test (at least one that smooths over the discrete regressors) may be advantageous. Also, using the same CV bandwidths for both estimation and testing can help researchers judge findings arising from a visual inspection of the fitted data. Therefore, we view the CV-based tests as complementary to non-smoothing tests.

2. Consistent tests with mixed discrete and continuous regressors

Before we present the proposed test statistic, we introduce some notation and briefly discuss how one estimates a nonparametric regression function in the presence of mixed discrete and continuous data.

[2] Racine and Li's (2004) approach builds on the work of Aitchison and Aitken (1976) who proposed a novel method for kernel density estimation with multivariate discrete data (see also Bowman, 1980; Hall, 1981; Hall and Wand, 1988 for related work).

2.1. Kernel estimation with mixed discrete and continuous regressors

We consider the case in which a subset of regressors are discrete and the remainder are continuous. Although it is well known that one can use a nonparametric frequency method to handle the presence of discrete regressors (theoretically), such an approach cannot be used in practice if the number of discrete cells is large relative to the sample size, as this will result in discrete cells containing insufficient data to meaningfully apply nonparametric methods (as is often the case with economic data sets containing mixed data types). Following the approach of Aitchison and Aitken (1976), we smooth the discrete regressors to avoid this problem. We would like to alert readers who may be unfamiliar with this area that there is an extensive literature on the kernel smoothing of discrete variables in statistics (see Fahrmeir and Tutz, 1994; Grund and Hall, 1993; Hart, 1997; Scott, 1992; Simonoff, 1996 and the references therein for further discussion).

Let x_i^{d} denote a $k \times 1$ vector of discrete regressors, and let $x_i^{\mathrm{c}} \in \mathbb{R}^q$ denote the remaining continuous regressors. We assume that some of the discrete regressors have a natural ordering, examples of which would include preference orderings (like, indifference, dislike), health conditions (excellent, good, poor) and so forth. Let $\tilde{x}_i^{\mathrm{d}}$ denote a $k_1 \times 1$ vector (say, the first k_1 components of x_i^{d}, $0 \leqslant k_1 \leqslant k$) of discrete regressors that have a natural ordering ($0 \leqslant k_1 \leqslant k$), and let $\bar{x}_i^{\mathrm{d}}$ denote the remaining $k_2 = k - k_1$ discrete regressors that do not have a natural ordering. We use x_{is}^{d} to denote the sth component of x_i^{d} ($s = 1, \ldots, k$).

For an unordered regressor, we use a variation on Aitchison and Aitken's (1976) kernel function defined by

$$\bar{l}(\bar{x}_{is}^{\mathrm{d}}, \bar{x}_{js}^{\mathrm{d}}) = \begin{cases} 1 & \text{if } \bar{x}_{is}^{\mathrm{d}} = \bar{x}_{js}^{\mathrm{d}}, \\ \lambda_s & \text{otherwise}, \end{cases} \tag{2.1}$$

where $\lambda_s \in [0, 1]$ is the smoothing parameter. Note that $\lambda_s = 0$ leads to an indicator function, and $\lambda_s = 1$ gives a uniform weight function. In this latter case, the $\bar{x}_t^{\mathrm{d}}$ regressor will be completely smoothed out (automatically removed as it will not affect the nonparametric estimation result).

For an ordered regressor, we suggest using the following kernel:

$$\tilde{l}(\tilde{x}_{is}^{\mathrm{d}}, \tilde{x}_{js}^{\mathrm{d}}, \lambda_s) = \begin{cases} 1 & \text{if } \tilde{x}_{is}^{\mathrm{d}} = \tilde{x}_{js}^{\mathrm{d}}, \\ \lambda_s^{|\tilde{x}_{is}^{\mathrm{d}} - \tilde{x}_{js}^{\mathrm{d}}|} & \text{if } \tilde{x}_{is}^{\mathrm{d}} \neq \tilde{x}_{js}^{\mathrm{d}}. \end{cases} \tag{2.2}$$

Again, when $\lambda_s = 0$ ($\lambda_s \in [0, 1]$), $l(\tilde{x}_{it}^{\mathrm{d}}, \tilde{x}_{jt}^{\mathrm{d}}, \lambda_s = 0)$ becomes an indicator function, and when $\lambda_s = 1$, $l(\tilde{x}_{it}^{\mathrm{d}}, \tilde{x}_{jt}^{\mathrm{d}}, \lambda_s = 1) = 1$ is a uniform weight function.[3]

Let $1(A)$ denote an indicator function which assumes the value 1 if A occurs and 0 otherwise. Combining (2.2) and (2.1), we obtain the product kernel function given by

$$L(x_i^{\mathrm{d}}, x_j^{\mathrm{d}}, \lambda) = \left[\prod_{s=1}^{k_1} \lambda_s^{|\tilde{x}_{is}^{\mathrm{d}} - \tilde{x}_{js}^{\mathrm{d}}|}\right] \left[\prod_{s=k_1+1}^{k} \lambda_s^{1 - \mathbf{1}(\bar{x}_{is}^{\mathrm{d}} = \bar{x}_{js}^{\mathrm{d}})}\right]. \tag{2.3}$$

[3]Although we do not cover the infinite support discrete variable case theoretically, in practice, infinite support discrete variables are likely to be ordered discrete variables, so one can use the kernel function introduced in Eq. (2.2) to handle the presence of infinite support discrete variables.

We will require a leave-one-out kernel estimator of $g(x_i)$ given by

$$\hat{g}_{-i}(x_i) = n^{-1} \sum_{j=1, j\neq i}^{n} y_j L_{\lambda,ij} W_{h,ij} / \hat{f}_{-i}(x_i), \tag{2.4}$$

where $L_{\lambda,ij} = L(x_i^{\mathrm{d}}, x_j^{\mathrm{d}}, \lambda)$, $W_{h,ij} = \prod_{s=1}^{q} h_s^{-1} w(x_{is}^{\mathrm{c}} - x_{js}^{\mathrm{c}}/h_s)$ is a product kernel function for the continuous regressor x^{c} (e.g., Gaussian, Epanechnikov etc.), h_s is the smoothing parameter associated with x_{is}^{c}, and

$$\hat{f}_{-i}(x_i) = n^{-1} \sum_{j=1, j\neq i}^{n} L_{\lambda,ij} W_{h,ij} \tag{2.5}$$

is the leave-one-out kernel estimator of $f(x_i)$ ($f(x)$ is the density function of $x = (x^{\mathrm{c}}, x^{\mathrm{d}})$).

We choose $(h, \lambda) = (h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$ by minimizing the following least squares CV function

$$CV(h, \lambda) = \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{g}_{-i}(x_i)]^2 M(x_i^{\mathrm{c}}), \tag{2.6}$$

where $\hat{g}_{-i}(x_i)$ is the leave-one-out kernel estimator defined in (2.4) and $M(\cdot)$ is a weight function having compact support. The use of a compactly supported weight function ensures that $CV(h, \lambda)$ is finite asymptotically and also mitigates boundary bias problems (e.g., Hall et al., 2005). Let

$$H_n = \left\{ (h_1, \ldots, h_q) \in R_+^q \,|\, n^{-(1-\varepsilon)} \leqslant \prod_{s=1}^{q} h_s \leqslant b^{-\varepsilon}, n^{-c} < h_s < n^{c}, s = 1, \ldots, q \right\}. \tag{2.7}$$

We assume that $(h_1, \ldots, h_q) \in H_n$. This condition basically assume that each h_s does not converge to 0, or to ∞, too fast, and that $nh_1 \ldots h_q \to \infty$. The minimization of the CV objective function (2.6) is done over $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r) \in H_n \times [0, 1]^r$. Hall et al. (2005) show that when a continuous (discrete) regressor, say x_s^{c} (x_s^d), is irrelevant in the sense that x_{is}^{c} is independent of y_i, then the CV selected smoothing parameter, say $\hat{h}_s$, will diverge to $+\infty$ (i.e., $\mathrm{P}(\hat{h}_s > C) \to 1$ for all $C > 0$). Similarly, if x_s^{d} is an irrelevant regressor, then $\hat{\lambda}_s \to 1$ in probability. Therefore, irrelevant regressors can be automatically (asymptotically) smoothed out. For expositional simplicity, in the remaining part of this paper, we will assume that the regressors are relevant ones, or equivalently, one can presume that the irrelevant regressors are already detected by the CV method and removed from the regression model.

We use $(\hat{h}_1, \ldots, \hat{h}_q, \hat{\lambda}_1, \ldots, \hat{\lambda}_r)$ to denote the CV choices of (h, λ) that minimize (2.6). Racine and Li (2004) and Hall et al. (2005) have derived the rate of convergence of the CV smoothing parameters to their (non-stochastic) optimal benchmark values. More specifically, defining $h_s^0 = a_s^0 n^{-1/(4+q)}$ and $\lambda_s^0 = b_s^0 n^{-2/(4+q)}$, where a_s^0 is a positive constant and b_s^0 is a non-negative constant,[4] Racine and Li (2004) (assuming all regressors are relevant) show that

$$(\hat{h}_s - h_s^0)/h_s^0 = \mathrm{O_p}(n^{-\varepsilon/(4+q)}) \quad \text{for } s = 1, \ldots, q,$$

[4] h_s^0 ($s = 1, \ldots, q$) and λ_s^0 ($s = 1, \ldots, r$) minimize the leading term of the non-stochastic objective function $E[CV(h, \lambda)]$ (see Hall et al., 2005 for details).

and

$$\hat{\lambda}_s - \lambda_s^0 = \mathrm{O_p}(n^{-\delta}) \quad \text{for } s = 1, \ldots, r, \tag{2.8}$$

where $\varepsilon = \min\{q/2, 2\}$ and $\delta = \min\{1/2, 4/(4+q)\}$.

Note that (2.8) implies that $\hat{h}_s = \mathrm{O_p}(h_s^0) = \mathrm{O_p}(n^{-1/(4+q)})$ and $\hat{\lambda}_s = \mathrm{O_p}(\lambda_s^0) = \mathrm{O_p}(n^{-2/(4+q)})$. The bias of the nonparametric estimator is of order $\mathrm{O_p}(\sum_{s=1}^{q}(h_s^0)^2 + \sum_{s=1}^{r}\lambda_s^0)$, and the variance is of order $\mathrm{O_p}((nh_1^0 \ldots h_q^0)^{-1})$. The leading terms h_s^0 and λ_s^0 are obtained by balancing the variance and the bias (squared). Moreover, (2.8) gives the rates of convergence of the CV smoothing parameters to their optimal benchmark values, which can be used to derive the asymptotic distribution of the CV-based test proposed in this paper. However, if one considers the smoothing parameter to be an index and treats the test statistic as a stochastic process, one can readily establish the asymptotic distribution of the test statistic using tightness/stochastic equicontinuity arguments which hold under quite weak conditions (e.g., Mammen, 1992; Ichimura, 2000).[5]

From a statistical point of view, smoothing the discrete regressors may introduce some *finite-sample* bias, but at the same time it will reduce the finite-sample variance. The CV selection of λ can be interpreted as a way of minimizing the finite-sample mean square error (MSE). Therefore, the reduction in variance more than offsets the increase in (squared) bias, and, as a result, the finite-sample MSE may be reduced substantially. The simulation results presented in Racine and Li (2004) reveal that the nonparametric estimator of $g(x)$ based on CV bandwidth selection $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$ performs much better than a conventional frequency estimator (which corresponds to $\lambda_s = 0$) because the former does not rely on sample splitting in finite-sample applications. Based on this intuition, we expect that the new kernel test outlined in the next subsection, which uses CV for choosing $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$, will be more powerful than a conventional kernel test that uses $\lambda_s = 0$. This intuition is confirmed in simulations reported in Section 3.

2.2. A consistent kernel based test

We are interested in testing the null hypothesis that a parametric model is correctly specified, which we state as

$$\mathrm{H_0} : \mathrm{P}[\mathrm{E}(y_i|x_i) = m(x_i, \beta)] = 1 \quad \text{for some } \beta \in \mathscr{B}, \tag{2.9}$$

where $m(\cdot,\cdot)$ is a known function with β being a $p \times 1$ vector of unknown parameters, where $\mathscr{B}$ is a compact subset in $\mathbb{R}^p$. The alternative hypothesis is the negation of $\mathrm{H_0}$, i.e.,

$$\mathrm{H_1} : \mathrm{P}[\mathrm{E}(y_i|x_i) = m(x_i, \beta)] < 1 \quad \text{for all } \beta \in \mathscr{B}. \tag{2.10}$$

We consider a test statistic that was independently proposed by Fan and Li (1996) and Zheng (1996). The test statistic is based on $I \overset{\text{def}}{=} \mathrm{E}[u_i \mathrm{E}(u_i|x_i) f(x_i)]$, where $u_i = y_i - m(x_i, \beta)$. Note that $I = \mathrm{E}\{[\mathrm{E}(u_i|x_i)]^2 f(x_i)\} \geqslant 0$, and $I = 0$ if and only if $\mathrm{H_0}$ is true. Therefore, I serves as a valid candidate for testing $\mathrm{H_0}$. The sample analogue

[5] We are indebted to an anonymous referee who suggested this approach.

of I is given by

$$I_n = n^{-1}\sum_{i=1}^{n}\hat{u}_i\hat{\mathrm{E}}_{-i}(u_i|x_i)\hat{f}_{-i}(x_i) = n^{-1}\sum_{i=1}^{n}\hat{u}_i\left\{n^{-1}\sum_{j=1,j\neq i}^{n}\hat{u}_j W_{h,ij}L_{\lambda,ij}\right\}$$
$$= n^{-2}\sum_{i}\sum_{j\neq i}\hat{u}_i\hat{u}_j K_{\gamma,ij}, \tag{2.11}$$

where $K_{\gamma,ij} = W_{h,ij}L_{\lambda,ij}$ $(\gamma = (h,\lambda))$, $\hat{u}_i = y_i - m(x_i,\hat{\beta})$ is the parametric null model's residual, $\hat{\beta}$ is a $\sqrt{n}$-consistent estimator of β (under H_0), and $\hat{\mathrm{E}}_{-i}(u_i|x_i)\hat{f}_{-i}(x_i)$ is a leave-one-out kernel estimator of $\mathrm{E}(y_i|x_i)f(x_i)$. In the case where we have only continuous regressors x_i^{c} and use a non-stochastic value of h_s ($h_s = \mathrm{o}(1)$ and $(nh_1\ldots h_q)^{-1} = \mathrm{o}(1)$), the asymptotic null (normal) distribution of the I_n test is derived independently by Fan and Li (1996) and Zheng (1996).

We advocate the use of CV methods for selecting the smoothing parameter vectors h and λ. We use $\hat{I}_n$ to denote our CV-based test, i.e, $\hat{I}_n$ is defined the same way as I_n given in (2.11) but with $(h_1,\ldots,h_q,\lambda_1,\ldots,\lambda_r)$ replaced by the CV smoothing parameters $(\hat{h}_1,\ldots,\hat{h}_q,\hat{\lambda}_1,\ldots,\hat{\lambda}_r)$. The asymptotic distribution of our CV-based test is given in the next theorem.

Theorem 2.1. *Under Assumptions* (A1)–(A3) *given in the appendix, we have*

(i) $n(\hat{h}_1\ldots\hat{h}_q)^{1/2}\hat{I}_n \to \mathrm{N}(0,\Omega)$ *in distribution under* H_0, *where* $\Omega = 2\mathrm{E}[\sigma^4(x_i)f(x_i)]\times[\int W^2(v)\,\mathrm{d}v]$.
A consistent estimator of Ω *is given by*

$$\hat{\Omega} = \frac{2(\hat{h}_1\ldots\hat{h}_q)}{n^2}\sum_{i}\sum_{j\neq i}\hat{u}_i^2\hat{u}_j^2 W_{\hat{h},ij}^2 L_{\hat{\lambda},ij}^2. \tag{2.12}$$

Hence, we have
(ii) $\hat{J}_n \overset{\mathrm{def}}{=} n(\hat{h}_1\ldots\hat{h}^q)^{1/2}\hat{I}_n/\sqrt{\hat{\Omega}} \to \mathrm{N}(0,1)$ *in distribution under* H_0.

The proof of Theorem 2.1 is given in the appendix.

It can be easily shown that the $\hat{J}_n$ test diverges to $+\infty$ if H_0 is false; thus it is a consistent test. When there are only continuous regressors and with the use of non-stochastic smoothing parameters $h_1,\ldots,h_q$, Zheng (1996) shows that the $\hat{J}_n$ test can detect Pitman local alternatives that approach the null model at the rate $(n(h_1\ldots h_q)^{1/2})^{-1/2}$. A sequence of Pitman local alternatives is defined as:

$$\mathrm{H}_{1L}:\ \mathrm{P}[\mathrm{E}(y_i|x_i) = m(x_i,\beta) + \delta_n l(x_i)] = 1 \quad \text{for some } \beta\in\mathscr{B}.$$

Using similar derivations as in Zheng (1996), one can show that, if $\delta_n = (n(h_1^0\ldots h_q^0)^{1/2})^{-1/2}$, then

$$\hat{J}_n \to \mathrm{N}(\mu,1) \text{ in distribution,}$$

where $\mu = \mathrm{E}[l^2(x_i)f(x_i)]/\sqrt{\Omega}$, Ω is defined in Theorem 2.1. Therefore, our CV based $\hat{J}_n$ test can detect Pitman local alternatives that approach the null model at the rate $(n(h_1^0\ldots h_q^0)^{1/2})^{-1/2} = n^{-(4+q/2)/[2(4+q)]}$ (because $h_s^0 \sim n^{-1/(4+q)}$ for $s = 1,\ldots,q$). Hence, the local power property of our smoothing $\hat{J}_n$ test is the same as the case considered by Zheng (1996) with only continuous regressors, and it is also the same as the case with mixed discrete and continuous regressors but where one uses the conventional frequency method

to deal with the discrete regressors. Therefore, when all regressors are relevant, smoothing the discrete regressors does not lead to power gains asymptotically. This is not surprising because, asymptotically, $\lambda_s \to 0$ for all $s = 1, \ldots, q$, and our smoothing test statistic and a frequency based test statistic should be close to each other when n is sufficiently large. This being said, however, in finite-sample applications, we expect the smoothing (discrete regressors) test to be more powerful than a frequency-based test because, by smoothing discrete regressors, we obtain more accurate estimates of the unknown conditional mean function in the MSE sense. This conjecture is confirmed in simulations reported in Section 4.

It can be shown that the above local power property is the same for tests based upon other nonparametric methods such as series-based tests (e.g., Hong and White, 1995) and k-nearest-neighbor-based (k-nn) tests (e.g., Li, 2005). However, when some of the regressors are in fact independent of the dependent variable (i.e., are irrelevant regressors), then the $\hat{J}_n$ test will (asymptotically) automatically remove these regressors (e.g., Hall et al., 2005). In this case, our smoothing test will be more powerful than a frequency based test, both in finite-samples and asymptotically. The finite-sample power gains can be substantial as evidenced by the simulations reported in Section 4. In the case of irrelevant regressors, we also expect that our kernel CV-based test is more powerful than series-based or k-nn-based tests because, even with data-driven methods for selecting the number of series terms for a series-based test and for selecting k in k-nn-based tests, it is unclear how to construct series based and k-nn based tests that can automatically remove irrelevant regressors (asymptotically). We would like to emphasize here that the ability to remove irrelevant continuous regressors x^c is a special feature associated with the local constant (Nadaraya–Watson) kernel method when coupled with the least squares CV method. If one were to use local linear methods to estimate $E(u_i|x_i)$ (e.g., Liu et al., 2001), the irrelevant continuous regressors cannot be smoothed out.[6] That is, the local linear estimation method coupled with the least squares CV method can detect linearity (in the continuous regressor x^c), but it cannot detect irrelevant regressors in x^c (e.g., Li and Racine, 2004).

Theorem 2.1 is valid asymptotically. Numerous simulations have revealed that the asymptotic normal approximation performs poorly in finite-sample settings for the J_n test in the case with only continuous regressors (e.g., Li and Wang, 1998, also see Härdle and Mammen, 1993 who employ a slightly different statistic). In fact, Li and Wang (1998) show that the $\hat{J}_n$ test with only continuous regressors (and with non-stochastic h) approaches the asymptotic standard normal distribution at the rate $O_p((h_1 \ldots h_q)^{1/2})$. When $q = 1$ with $h \sim n^{-1/5}$, this yields the rate of $O_p(n^{-1/10})$ which is an extremely slow rate of convergence. Simulations reported in Li and Wang show substantial size distortions for the $\hat{J}_n$ test. Our simulations for the CV test ($\hat{J}_n$) also show that the asymptotic normal approximation does not work well in finite-sample settings. Therefore, we suggest using bootstrap methods as a viable alternative for approximating the finite-sample null distribution of the CV-based test statistic $\hat{J}_n$.

We advocate using the residual-based wild bootstrap method to approximate the null distribution of $\hat{J}_n$. The wild bootstrap error u_i^* is generated via a two point distribution $u_i^* = [(1-\sqrt{5})/2]\hat{u}_i$ with probability $(1+\sqrt{5})/[2\sqrt{5}]$, and $u_i^* = [(1+\sqrt{5})/2]\hat{u}_i$ with probability $(\sqrt{5}-1)/[2\sqrt{5}]$. From $\{u_i^*\}_{i=1}^n$, we generate $y_i^* = m(x_i, \hat{\beta}) + u_i^*$ for $i = 1, \ldots, n$. $\{x_i, y_i^*\}_{i=1}^n$ is called the 'bootstrap sample', and we use this bootstrap sample to obtain a nonlinear least squares estimator of β (a least squares estimator if $m(x_i, \beta) = x_i'\beta$, the prime

[6]The local linear CV method can detect and remove discrete irrelevant regressors (see Li and Racine, 2004).

denoting transpose), while we let $\hat{\beta}^*$ denote the resulting estimator. The bootstrap residual is given by $\hat{u}_i^* = y_i^* - m(x_i, \hat{\beta}^*)$. The bootstrap test statistic $\hat{J}_n^*$ is obtained from $\hat{J}_n$ with $\hat{u}_i$ being replaced by $\hat{u}_i^*$. Note that we use the same CV selected smoothing parameters $\hat{h}$ and $\hat{\lambda}$ when computing the bootstrap statistics. That is, there is no need to rerun CV with the bootstrap sample. Therefore, our bootstrap test is computationally quite simple. In practice, we repeat the above steps a large number of times, say $B = 399$ times, the original test statistic $\hat{J}_n$ plus the B bootstrap test statistics give us the empirical distribution of the bootstrap statistics, which is then used to approximate the finite-sample null distribution of $\hat{J}_n$.

We will use the concept of 'convergence in distribution in probability' (e.g., Li et al., 2003) to study the asymptotic distribution of the bootstrap statistic $\hat{J}_n^*$.[7] The next theorem shows that the wild bootstrap works for the CV-based $\hat{J}_n$ test.

Theorem 2.2. *Under Assumptions* (A1)–(A3) *given in the appendix, we have*

$$\sup_{z\in\mathbb{R}} |\mathrm{P}(\hat{J}_n^* \leqslant z | \{x_i, y_i\}_{i=1}^n) - \Phi(z)| = \mathrm{o_p}(1), \tag{2.13}$$

where $\hat{J}_n^* = n(\hat{h}_1 \dots \hat{h}_q)^{1/2} \hat{I}_n^* / \sqrt{\hat{\Omega}^*}$, $\Phi(\cdot)$ *is the cumulative distribution function of a standard normal random variable.*

The proof of Theorem 2.2 is given in the appendix.

3. Monte Carlo results

In this section we report some Monte Carlo simulations which were designed to examine the finite-sample performance of the proposed bootstrap test. We adopt slightly different notation in this section, and for clarity we will use x_i to denote discrete regressors and z_i to denote continuous ones.

The null model that we consider is

$$DGP0\text{:}\ y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + z_i\beta_3 + z_i^2\beta_4 + u_i, \tag{3.1}$$

where, for $t = 1, 2$, x_{it} takes values in $\{0, 1\}$ with $\mathrm{P}(x_{ti} = l) = 0.5$ for $l = 0, 1$, $z_i \sim \mathrm{N}(0, 1)$ and $u_i \sim \mathrm{N}(0, 1)$. We choose $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (1, 1, 1, 1, 1)$.

We consider three alternative models,

$$DGP1\text{:}\ y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + z_i\beta_3 + z_i^2\beta_4 + z_i^3\beta_5 + u_i \tag{3.2}$$

with $\beta_5 = 0.25$,

$$DGP2\text{:}\ y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + z_i\beta_3 + z_i^2\beta_4 + x_{i1}z_i\beta_5 + x_{2i}z_i^2\beta_6 + u_i \tag{3.3}$$

with $\beta_5 = \beta_6 = 0.5$, and

$$DGP3\text{:}\ y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + z_i\beta_3 + z_i^2\beta_4 + (x_{i1} + x_{2i})\sin(4\pi z_i)\beta_6 + u_i \tag{3.4}$$

with $\beta_6 = 1.0$.

[7] In the literature, the concept 'convergence in distribution with probability one' is used to describe the asymptotic behavior of bootstrap tests. 'Convergence in distribution in probability' is much easier to establish than 'convergence in distribution with probability one', and runs parallel to that of 'convergence in probability' and 'convergence with probability one'.

Compared with the null model DGP0, DGP1 is a low frequency alternative that has an extra cubic term, DGP2 is a low frequency alternative that has additional interaction terms between the continuous and discrete regressors, and DGP3 is a high frequency alternative.

For all simulations we use a Gaussian kernel for $W(\cdot)$, while the discrete regressor kernel $L(\cdot,\cdot,\cdot)$ is defined in (2.3). We consider samples of size $n = 100$, 200, and 400. The number of Monte Carlo replications is 1,000, and the empirical distribution of the test statistic is computed from the 399 wild bootstrap test statistics generated under the null model. We compare bootstrap size and power performance for four test statistics: (i) the proposed CV based test $\hat{J}_n$; (ii) a test using CV h and $\lambda = 0$; (iii) an ad hoc plug-in method for selecting h, $h = z_{\mathrm{sd}} n^{-1/5}$, and $\lambda = 0$; and (iv) an ad hoc plug-in method for selecting both h and λ with $h = z_{\mathrm{sd}} n^{-1/5}$ and $\lambda = x_{j,\mathrm{sd}} n^{-2/5}$, where z_{sd}, $x_{j,\mathrm{sd}}$ are the sample standard deviation of $\{z_i\}_{i=1}^n$ and $\{x_{j,i}\}_{i=1}^n$ $(j = 1, 2)$, respectively. Both (ii) and (iii) are frequency-based tests and are expected to be less powerful than our proposed test $\hat{J}_n$. The ad hoc selection of h in (iii) is suggested in Li and Wang (1998), while the choice of (iv) was suggested by an anonymous referee. We compute empirical rejection frequencies for conventional significance levels $\alpha = 0.01$, 0.05 and 0.10. The results are reported in Table 1. For brevity we will refer to: (i) as the proposed CV test; (ii) as the CV h/frequency (with $\lambda = 0$) test; (iii) as the ad hoc h/frequency test; and (iv) as the ad hoc h/λ test. Tables 2 and 3 summarize the behavior of the CV bandwidths for the discrete regressors.

Examining Table 1, we first note that all the (CV and the frequency) tests have empirical sizes that do not differ significantly from their nominal sizes (entries associated with DGP0). Next, for the low frequency alternative DGP1, all three tests are quite powerful, with our proposed CV test and the ad hoc h/λ test being the most powerful ones. To see the power improvement due to smoothing the discrete regressors, we compare our CV test

Table 1
Empirical rejection frequencies for DGP0, DGP1, DGP2, and DGP3 (CV indicates cross-validation, AH means ad hoc)

α	CV λ, h			CV h, $\lambda = 0$			AH h, $\lambda = 0$			AH h, AH λ		
	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
$n = 100$												
DGP0	0.010	0.048	0.100	0.007	0.034	0.087	0.005	0.033	0.076	0.002	0.027	0.080
DGP1	0.346	0.631	0.737	0.103	0.317	0.447	0.117	0.308	0.433	0.277	0.532	0.635
DGP2	0.193	0.473	0.585	0.155	0.417	0.570	0.181	0.443	0.573	0.200	0.455	0.560
DGP3	0.257	0.432	0.475	0.051	0.212	0.303	0.009	0.038	0.104	0.009	0.034	0.090
$n = 200$												
DGP0	0.004	0.047	0.105	0.009	0.060	0.115	0.005	0.061	0.107	0.009	0.045	0.112
DGP1	0.669	0.920	0.963	0.436	0.741	0.831	0.459	0.770	0.856	0.668	0.893	0.942
DGP2	0.543	0.794	0.884	0.506	0.797	0.874	0.550	0.835	0.902	0.578	0.813	0.887
DGP3	0.802	0.820	0.831	0.653	0.781	0.810	0.006	0.043	0.087	0.006	0.036	0.092
$n = 400$												
DGP0	0.013	0.060	0.110	0.009	0.061	0.107	0.007	0.056	0.114	0.005	0.056	0.111
DGP1	0.867	0.988	0.998	0.804	0.965	0.989	0.850	0.978	0.989	0.912	0.991	0.997
DGP2	0.948	0.991	0.995	0.932	0.993	0.998	0.958	0.996	0.997	0.956	0.988	0.993
DGP3	0.982	0.985	0.985	0.979	0.984	0.984	0.003	0.040	0.087	0.004	0.046	0.086

Table 2
$\hat{\lambda}_1$: Lower quartile, median, and upper quartile values of the CV smoothing parameters for the discrete regressors

DGP	P_{25}	P_{50}	P_{75}
$n = 100$			
DGP0	0.081	0.141	0.210
DGP1	0.086	0.150	0.226
DGP2	0.095	0.175	0.291
$n = 200$			
DGP0	0.070	0.106	0.145
DGP1	0.083	0.118	0.164
DGP2	0.088	0.133	0.185
$n = 400$			
DGP0	0.053	0.077	0.100
DGP1	0.063	0.089	0.116
DGP2	0.066	0.094	0.125

Table 3
$\hat{\lambda}_2$: Lower quartile, median, and upper quartile values of the CV smoothing parameters for the discrete regressors

DGP	P_{25}	P_{50}	P_{75}
$n = 100$			
DGP0	0.084	0.143	0.210
DGP1	0.085	0.144	0.225
DGP2	0.091	0.170	0.282
$n = 200$			
DGP0	0.070	0.106	0.145
DGP1	0.083	0.118	0.164
DGP2	0.088	0.133	0.185
$n = 400$			
DGP0	0.054	0.076	0.099
DGP1	0.061	0.087	0.115
DGP2	0.066	0.094	0.121

with the CV h/frequency (with $\lambda = 0$) test. Considering the case of $n = 100$, at the 1% level the power of our CV test is triple that of the CV h/frequency test, while at the 5% level the power of the CV test is double that of the CV h/frequency test. This is consistent with our theoretical analysis, as the frequency test splits the sample into four parts (x_1 and x_2 each assume two values leading to four subsets or 'cells') when estimating the regression function nonparametrically, which results in a loss of efficiency thereby leading to a loss of power. Tables 2 and 3 reveal that $\hat{\lambda}_1$ and $\hat{\lambda}_2$ indeed converge to 0 as $n \to \infty$; however, there appears to be a fair bit of smoothing occurring across cells in finite-samples. The CV test significantly dominates the frequency-based test in terms of its ability to detect departures from the null in the presence of mixed discrete and continuous data.

We observe that for DGP3, the ad hoc h/frequency (with $\lambda = 0$) test and the ad hoc h/λ test have almost no power at all for the sample sizes considered. This may at first glance appear surprising, particularly in light of the good power demonstrated by the other two tests. The intuition underlying the lack of power of the ad hoc h (and ad hoc λ for test (iv)) test is in fact straightforward. The ad hoc bandwidth, h, is too large for the *high frequency* alternative DGP3. That is, it dramatically oversmooths the data, thereby completely obscuring the deviations from the null model present in the data (roughly 500% larger than that selected via CV). In contrast, the CV method automatically selects a much smaller value of h resulting in a more powerful test. Our proposed CV test is also significantly more powerful than the CV h/frequency (with $\lambda = 0$) test, again due to the fact that we do not split the sample into discrete cells in finite-sample applications.

The main purpose of this paper is to construct a consistent model specification test which can be applied in the presence of mixed discrete and continuous data. However, even for the case with only continuous regressors, our CV based test is new in the literature. With continuous data it is known that, in order to obtain a kernel-based test having high power, one should use a relatively large value of h for low frequency data and a relatively small value of h with high frequency data (e.g., Fan and Li, 2000). Recently, Horowitz and Spokoiny (2001) propose using a range of smoothing parameters in a kernel-based test to guard against low power in the direction of either low or high frequency alternatives. Horowitz and Spokoiny show that their proposed test is adaptive and rate optimal. The intuition behind the adaptive test is that, when the DGP deviates from the null model and is a high frequency type, one should use a relatively small h in order to have high power against a high frequency alternative. On the other hand, if the DGP is a low frequency type, one should use a relatively large value of h in the kernel test (see Fan and Li, 2000 for a more detailed discussion).

Our CV based test is not adaptive or rate optimal in the sense of Horowitz and Spokoiny. However, it does have the ability to select a relatively large smoothing parameter for low frequency data and a relatively small smoothing parameter for high frequency data in *finite-sample* applications. Thus, even in the simple case with only continuous regressors, our CV based test is expected to have much better *finite-sample* power than a number of existing tests based on ad hoc smoothing parameter selection (e.g., Zheng, 1996; Li and Wang, 1998). Below we consider a DGP similar to the one found in Horowitz and Spokoiny (2001) having only one continuous regressor. We show that our CV smoothing parameter h tends to assume relatively large values for low frequency DGPs and relatively small values for high frequency ones. Consequently, our test is more powerful than a test which uses ad hoc bandwidth selection.

We consider a DGP similar to that used in Horowitz and Spokoiny (2001). The null DGP is given by

$$DGP4\text{: } y_i = 1 + z_i + u_i,$$

where $z_i \sim \mathrm{N}(0, 25)$ is truncated at its 5th and 95th percentiles,[8] and where $u_i \sim \mathrm{N}(0, 4)$. The alternative models have the form

$$DGP5\text{: } y_i = 1 + z_i + (5/\tau)\phi(z_i/\tau) + u_i,$$

where z_i and u_i are the same as in DGP4, $\phi(\cdot)$ is a standard normal density function, and $\tau = 0.25$ or 1. $\tau = 0.25$ corresponds to a high frequency alternative, and $\tau = 1$ corresponds

[8] Horowitz and Spokoiny (2001) consider a fixed set of $\{z_i\}_{i=1}^n$. Here, we consider the case of random $\{z_i\}_{i=1}^n$.

Table 4
Empirical rejection frequencies for DGP4 and DGP5

	Horowitz–Spokoiny			$\tilde{J}_n$ (ad hoc h)			$\hat{J}_n$ CV-test		
	$\alpha = 0.01$	0.05	0.10	$\alpha = 0.01$	0.05	0.10	$\alpha = 0.01$	0.05	0.10
DGP4 (size)	0.013	0.055	0.108	0.013	0.055	0.109	0.015	0.060	0.112
DGP5 ($\tau = 0.25$)	0.551	0.821	0.891	0.324	0.585	0.710	0.504	0.801	0.882
DGP5 ($\tau = 1$)	0.460	0.685	0.779	0.406	0.638	0.760	0.498	0.728	0.818

to a low frequency one. The sample size is $n = 100$. For the adaptive test we first select $h = \{2.5, 3, 3.5, 4, 4.5\}$, the values used by Horowitz and Spokoiny in their simulation using a similar DGP. Surprisingly, we find that the resulting adaptive test has lower power than the CV test for both $\tau = 0.25$ and 1. We then examine the values of h selected by the CV test. For $\tau = 0.25$, the median value of h_{CV} (over 1,000 replications) is 0.551, and for $\tau = 1$, the median value of h_{CV} is 0.956, both of which are less than the smallest h used in the adaptive test. We therefore recompute the adaptive test using $h = \{0.5, 1, 1.5, 2, 2.5\}$. The resulting power of the adaptive test is now slightly better than the CV test for $\tau = 0.25$. This result suggests that it is important to choose the correct range for h in computing the adaptive test, and that the CV method can be used as a guide to help identify the range of h's to be used for the adaptive test.

For the ad hoc h test we use a popular ad hoc rule to select h, $h = z_{sd}n^{-1/5}$. We use $\tilde{J}_n$ to denote the ad hoc h test. The estimated size and power of the three tests for DGP4 and DGP5 are reported in Table 2 below (the $\hat{J}_n$ and the $\tilde{J}_n$ tests are based on a bootstrap approximation of the null distribution).

From Table 4 we observe that both the CV-based test and the adaptive test are more powerful than the $\tilde{J}_n$ test. The ad hoc bandwidth selection rule ($h = z_{sd}n^{-1/5}$) ignores high and low frequency features present in the data. For the low frequency alternative, our CV h is, on average, twice as large as the h for the high frequency alternative. The simulation results also show that our CV test, in the case with only continuous regressors, provides a complement to the adaptive test of Horowitz and Spokoiny in *finite-sample* applications, and it can help identify the ranges of h's to be used when computing the adaptive test. Moreover, in the case when there exist irrelevant regressors, our CV-based test can (asymptotically) automatically remove the irrelevant regressors, while Horowitz and Spokoiny's test does not possess this property. Finally, our approach can also be applied to the mixed discrete and continuous regressor case in a straightforward way. While one can generalize the adaptive test to the mixed data type case by using the frequency estimator to deal with the presence of discrete regressors, it will suffer finite-sample power loss due to sample splitting (especially when there exist irrelevant discrete regressors). It might be possible to generalize the adaptive test to cover the mixed regressor case by smoothing the discrete regressors, however this extension seems quite challenging and lies beyond the scope of the present paper.

3.1. Hypothesis testing in the presence of irrelevant regressors

We now examine the finite-sample performance of the test when there exist irrelevant regressors by adding an irrelevant binary and an irrelevant continuous regressor to

Table 5
Empirical rejection frequencies in the presence of irrelevant regressors ($n = 400$)

DGP	CV λ, h			CV h, $\lambda = 0$			ad hoc h, $\lambda = 0$		
	$\alpha = 0.01$	0.05	0.10	$\alpha = 0.01$	0.05	0.10	$\alpha = 0.01$	0.05	0.10
DGP0	0.016	0.071	0.121	0.017	0.070	0.123	0.012	0.063	0.121
DGP1	1.000	1.000	1.000	0.894	0.960	0.979	0.642	0.793	0.858
DGP2	0.999	0.999	0.999	0.987	0.994	0.997	0.016	0.082	0.138

Table 6
$\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\lambda}_3$: Lower quartile, median, and upper quartile values of the CV smoothing parameters for the discrete regressors (relevant, $n = 400$)

DGP	$\hat{\lambda}_1$			$\hat{\lambda}_2$			$\hat{\lambda}_3$		
	P_{25}	P_{50}	P_{75}	P_{25}	P_{50}	P_{75}	P_{25}	P_{50}	P_{75}
DGP0	0.062	0.073	0.075	0.060	0.071	0.073	0.646	0.986	1.000
DGP1	0.062	0.078	0.103	0.056	0.080	0.105	0.642	0.982	1.000
DGP2	0.058	0.079	0.103	0.057	0.080	0.103	0.731	1.000	1.000

DGP0–DGP2. That is, y_i is still generated by DGP0–DGP2, but we overspecify the null model and estimate a model for which y_i is linear in $(x_{1i}, x_{2i}, x_{3i}, z_i, z_i^2, z_{2i})$, where $x_{3i} \in \{0, 1\}$ and $z_{2i} \sim \mathrm{N}(0, 1)$ are the irrelevant regressors. The test statistic involves nonparametric kernel estimators having five regressors (e.g., $\hat{\mathrm{E}}_{-i}(u_i|x_{1i}, x_{2i}, x_{3i}, z_i, z_{2i})$). The results of Hall et al. (2005) imply that the smoothing parameters for the irrelevant regressors should converge to their upper extremities ($\lambda_s \to 1$ and $h_s \to \infty$ as $n \to \infty$ for irrelevant discrete and continuous regressors x_s^{d} and x_s^{c}) so that these irrelevant regressors are effectively removed.[9] Empirical rejection frequencies are given in Table 5. As this setting involves nonparametric regression with five regressors, we only conduct simulations with $n = 400$, a small sample size by nonparametric standards given the number of regressors involved. Table 6 summarizes the behavior of the CV bandwidths for the discrete regressors.

The key feature highlighted by this simulation is the relative performance of the various versions of the test. In particular, were one to use ad hoc plug-in rules or use the conventional frequency approach ($\lambda = 0$) with cross-validatory choices of h, the power of the test would be dramatically reduced relative to the case in which all regressors are relevant, while the proposed approach retains much of its power in either setting. This arises due to the ability of CV to remove irrelevant regressors by oversmoothing the irrelevant regressors through selecting large bandwidths for irrelevant regressors while delivering optimal bandwidths for the relevant regressors. Table 6 reveals that bandwidths for the relevant discrete regressors $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are small (and are converging to 0 as $n \to \infty$); however, for the irrelevant regressor, $\hat{\lambda}_3$ tends to assume values at or near its upper bound value of 1, underscoring the ability of CV to remove irrelevant regressors by oversmoothing.

[9] Hall and Wehrly (1992) observed a similar phenomenon with only continuous regressors.

4. Testing for correct specification for wage equations

In this section, we examine the behavior of the proposed test relative to its frequency-based counterpart, the latter being obtained by setting the smoothing parameters for the discrete regressors equal to zero in the proposed test ($\lambda = 0$). In addition, we compute the conventional frequency-based test when ad hoc plug-in methods are used to select the bandwidths for the continuous regressors. We also investigate the ability of the proposed test to help guide applied researchers in their search for data-consistent parametric models.

By way of example we offer a labor application and consider testing for correct specification of a wage equation, the most popular being quadratic in age (see, for example, Lanot and Walker, 1998). These models are often estimated at the industry level, and also are estimated on sub-groups of data such as, say, male unionized high school graduates. They are also frequently estimated using an additive dummy variable specification to model the discrete data types. Data was obtained from the 2000 CPS-ORG data files compiled by Hirsch and Macpherson which is a rich source of industry-level data.[10]

We consider a common specification incorporating additive dummy regressors for the discrete regressors given by

$$\log(W_i) = \beta_0 + \beta_1 F_i + \beta_2 E_i + \beta_3 U_i + \beta_4 A_i + \beta_5 A_i^2 + \varepsilon_i, \tag{4.1}$$

where W is hourly wage, F sex (Male = 0, Female = 1), E education (High School Graduate = 0, College Graduate = 1), U union status[11] (Non-union = 0, Union = 1), and A age ($16 \leqslant A \leqslant 64$). All workers were full time employees. However, model (4.1) is restrictive since it does not allow interaction terms among the discrete regressors and the age variable. It is therefore unlikely that model (4.1) will be an adequate description of the true relationship between the explanatory regressors and the dependent variable, log(wage). This conjecture is confirmed by our specification test.

The full sample size for this data is 64,138 which is computationally costly for using our CV-based test. Moreover, when we try a sample size larger than 2000, almost all tests reject the null of a simple linear regression model. We treat this large sample as a population and we know that the linear model is not true for this population. We then randomly selected a sample size of 500, 1000 and 1500 to compare the power performance of different tests. The purpose of this exercise is to show that the CV based smoothing test is more powerful than frequency-based tests with empirical data. To conduct the test, 399 bootstrap replications are used to obtain p-values. For the CV-based tests, the bandwidths are selected via leave-one-out CV, and the search algorithm is restarted ten times from different initial random values in an attempt to ensure that the search process does not become ensnared in local minima.

We consider sample sizes of $n = 500$ through 1500 in increments of 500 to examine the test's power properties. For a given model we conduct 21 tests, one for each of seven industries for each of the three sample sizes considered. For this specification we also

[10]We are indebted to Barry Hirsch for his generous help and for providing access to these data.

[11]We recognize the potential endogeneity of wages and union status. Given the controversy surrounding selectivity corrections in this setting, we ignore this issue and focus solely on functional form (see Freeman and Medoff, 1984 for well-known 'dismissals' of selection models for estimating union wage differentials). However, estimates of wage differentials generated from these 'unadjusted' wage equations are frequently reported and compared to 'adjusted' differentials (selectivity-adjusted), so we present these simply as tests for correct specification of unadjusted wage equations.

Table 7
p-Values for testing correct specification of the quadratic model

Industry	Sample size		
	500	1000	1500
Proposed CV $\hat{J}_n$ test			
Construction	0.028	0.006	0.000
Manufacturing	0.106	0.000	0.000
Transportation/communication/public utilities	0.000	0.000	0.000
Wholesale/retail trade	0.833	0.000	0.000
Finance/insurance/real estate	0.056	0.009	0.000
Service	0.039	0.018	0.000
Professional/related services	0.004	0.007	0.004
CV h with $\lambda = 0$			
Construction	0.879	0.074	0.004
Manufacturing	0.534	0.000	0.000
Transportation/communication/public utilities	0.008	0.005	0.000
Wholesale/retail trade	0.811	0.000	0.000
Finance/insurance/real estate	0.294	0.001	0.006
Service	0.099	0.302	0.005
Professional/related services	0.693	0.042	0.033
Ad hoc Plug-in h with $\lambda = 0$			
Construction	0.887	0.024	0.007
Manufacturing	0.302	0.000	0.000
Transportation/communication/public utilities	0.008	0.585	0.014
Wholesale/retail trade	0.529	0.000	0.000
Finance/insurance/real estate	0.138	0.084	0.007
Service	0.474	0.836	0.050
Professional/related services	0.498	0.074	0.015

generate results for the conventional frequency-based test and for the frequency-based test using ad hoc plug-in bandwidth selection methods. Results are summarized in Table 7 in the form of the empirical *p*-values for each test (using the wild bootstrap procedure to approximate the null distribution of the tests).

First, we note that our proposed test appears to be more powerful than the conventional frequency-based test ($\lambda = 0$). The *p*-values tabulated in Table 7 indicate that this quadratic additive specification is rejected at all conventional levels by the proposed $\hat{J}_n$ for $n = 1,000$, while it is also rejected by the other frequency-based tests when $n = 1500$.

A popular alternative specification, based in part on the seminal work of Murphy and Welch (1990), is quartic in age, given by

$$\log(W_i) = \beta_0 + \beta_1 F_i + \beta_2 E_i + \beta_3 U_i + \beta_4 A_i + \beta_5 A_i^2 + \beta_6 A_i^3 + \beta_7 A_i^4 + \varepsilon_i. \tag{4.2}$$

Results for the proposed $\hat{J}_n$ test are summarized[12] in Table 8. While the quartic specification is not rejected quite as frequently as the quadratic, it is evident that the

[12]We do not include results for the frequency-based tests listed in Table 7 for space considerations, but these results are available from the authors upon request.

Table 8
p-Values for testing correct specification of the quartic model

Industry	Sample size		
	500	1000	1500
Construction	0.842	0.882	0.000
Manufacturing	0.806	0.000	0.000
Transportation/communication/public utilities	0.000	0.000	0.000
Wholesale/retail trade	0.814	0.000	0.000
Finance/insurance/real estate	0.528	0.002	0.000
Service	0.359	0.084	0.084
Professional/related services	0.919	0.012	0.005

proposed test strongly suggests that neither (4.1) nor (4.2) appears to be appropriate for the data at hand.

In principle, it is always possible to get a good statistical fit by adding higher order polynomial terms in A_i. However, it is another matter altogether to render meaningful economic interpretations of the coefficients of higher order polynomials in A_i. Moreover, often a higher order polynomial equation will provide a good within-sample fit but will yield disastrous post-sample predictions. As an alternative, we consider the specification of a quadratic model with interaction among all regressors. Such a model not only provides more flexibility but also is consistent with human capital theory, which predicts that education, gender, union status, and experience all have an impact on one's wage rate, not just additively, but mutually reinforcing each other.

The results for testing the specification of a quadratic model with interaction among all regressors is given in Table 9. The p-values reported in Table 9 indicate that, at any conventional level for all industries but the manufacturing industry, we cannot reject the null hypothesis of correct specification for the model with all interaction terms between the dummy variables, A_i, and A_i^2. It appears that by considering the interaction terms, a log wage equation with a quadratic experience term not only produces a close approximation to the underlying DGP, but is also sufficiently simple to provide an economically meaningful interpretation of the data.

Parametric models involving the quadratic and quartic formulations often have linear additive dummy variables for various attributes, as in the studies of union wage differentials, sex-based discrimination and the like. Common parametric specifications focus mainly on nonlinearity present in age while often maintaining linear additive dummy variables for attributes. Using current industry-level data, our proposed specification soundly rejects these two common parametric specifications. Our specification test supports the alternative specification, which includes interaction terms between all regressors found in the standard quadratic specification.

The application considered here suggests that, in parametric settings where there may be an insufficient number of continuous regressors to capture the variation of a continuous dependent variable, it may be more fruitful to focus on interaction terms rather than focusing on potentially omitted higher-order polynomial terms in age. On the basis of findings reported above, we would recommend that applied researchers include such models in their list of candidate specifications.

Table 9
p-Values for testing the specification with all the interaction terms

Industry	Sample size		
	500	1000	1500
Construction	0.692	0.804	0.856
Manufacturing	0.029	0.074	0.181
Transportation/communication/public utilities	0.762	0.854	0.912
Wholesale/retail trade	0.829	0.825	0.848
Finance/insurance/real estate	0.747	0.829	0.852
Service	0.735	0.800	0.866
Professional/related services	0.826	0.801	0.887

5. Concluding remarks

In this paper we propose using cross-validation (CV) methods for choosing smoothing parameters when constructing a consistent kernel-based model specification test with both discrete and continuous data. Simulations demonstrate that the proposed test enjoys a substantial power advantage over frequency-based tests in the presence of discrete data. This methodology can be applied to a number of other model diagnostic situations such as nonparametric significance tests (omitted variable tests) or testing a semiparametric null regression model with mixed discrete and continuous regressors.

Acknowledgments

Insightful comments from two referees and a co-editor have led to a greatly improved version of this paper. In particular, the use of a stochastic-equicontinuity and tightness argument for proving the main result was suggested by a referee to whom we are extremely grateful. We would also like to thank Jushan Bai, Xiaohong Chen and Max Stinchcombe for their insightful discussions that led to much improved proofs of the main result. Li's research is partially supported by the Private Enterprises Research Center, Texas A&M University. Racine would like to thank the NSF, NSERC, and SSHRC for their generous support.

Appendix A. Proofs of Theorems 2.1 and 2.2

We first list the assumptions that will be used to prove Theorems 2.1 and 2.2.

Assumption A1. (i) (y_i, x_i), $i = 1, 2, \ldots, n$, are independent and identically distributed as (y_1, x_1). (ii) $\nabla m(x, .)$ and $\nabla^2 m(x, .)$ are continuous in x^c and dominated by functions with finite second moments, where $\nabla m(x, .)$ and $\nabla^2 m(x, .)$ are the $p \times 1$ vector of first order partial derivatives and the $p \times p$ matrix of second order partial derivatives of m with respect to β respectively. (iii) y has finite fourth moment. $g(x)$, $f(x)$, $\sigma(x) = \mathrm{E}(u_i^2 | x_i = x)$, and $\mu_4(x) = \mathrm{E}(u_i^4 | x_i = x)$ all satisfy some Lipschitz type conditions: $|H(x^c + v, x^d) - H(x^c, x^d)| \leqslant G(x^c, x^d)\|v\|$ with $\mathrm{E}[G^2(x_i)] < \infty$ for all $x^d \in \mathscr{D}$, where $\|.\|$ is the Euclidean norm and $\mathscr{D}$ is the domain of x^d.

Assumption A2. The univariate kernel function $w(\cdot)$: $\mathbb{R}^q \to \mathbb{R}$ is non-negative, bounded, symmetric and compactly supported with $\int w(v)\,dv = 1$, and $\int w(v)\|v\|^4\,dv < \infty$. Also, $w(\cdot)$ satisfies a Lipschitz condition: $|w(v') - w(v)| \leqslant G(v)\|v' - v\|$ with $G(v)$ is bounded and integrable. Further, $w(0) > w(\delta)$ for all $\delta > 0$.

Assumption A3. The smoothing parameters $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r) \in H_n \times [0,1]^r$, where H_n is defined in (2.7).

Assumption A1 is quite mild, requiring only some smoothness (for the continuous variables) and some moment conditions for the regression and density functions. Many commonly used kernels (say, kernels with bounded range) satisfy condition A2. Assumption A2 can be relaxed to allow for the Gaussian kernel, but we will not pursue this generality in this paper. Assumption A3 requires that each h_s does not converge to 0, or to ∞, too fast. Therefore, we allow for the existence of irrelevant regressors. However, by the result of Hall et al. (2005), we know that irrelevant regressors will be smoothed out by the CV method. Therefore, we only need to deal with relevant regressors in deriving the asymptotic distribution of our test statistic.

Because $h_s \sim n^{-1/(4+q)}$ for all $s = 1, \ldots, q$ and $\lambda_s \sim n^{-2/(4+q)}$ for all $s = 1, \ldots, r$, for expositional simplicity, in the appendix we will assume that $h_1 = \cdots = h_q = h$ and $\lambda_1 = \cdots = \lambda_r = \lambda$. Hence, $\hat{h}_1 = \cdots = \hat{h}_q = \hat{h}$ and $\hat{\lambda}_1 = \cdots = \hat{\lambda}_r = \hat{\lambda}$. We use h_0 and λ_0 denote the non-stochastic smoothing parameters such that $\hat{h}/h_0 - 1 = o_p(1)$ and $\hat{\lambda}/\lambda_0 - 1 = o_p(1)$ (e.g., Racine and Li, 2004).

Proof of Theorem 2.1. We first prove the case for a linear model, i.e., $m(x_i, \beta) = x_i'\beta$. Using $\hat{u}_i = u_i - x_i'(\hat{\beta} - \beta)$ we have (recall that $\gamma = (h, \lambda)$)

$$\begin{aligned}\hat{I}_n &= \frac{1}{n^2\hat{h}^q}\sum_i\sum_{j\neq i} u_i u_j K_{\hat{\gamma},ij} - 2\left[\frac{1}{n^2\hat{h}^q}\sum_i\sum_{j\neq i} u_i x_j' K_{\hat{\gamma},ij}\right](\hat{\beta}-\beta) \\ &\quad + (\hat{\beta}-\beta)'\left[\frac{1}{n^2\hat{h}^q}\sum_i\sum_{j\neq i} x_i x_j' K_{\hat{\gamma},ij}\right](\hat{\beta}-\beta) \\ &\equiv I_{1n} - 2I_{2n}(\hat{\beta}-\beta) + (\hat{\beta}-\beta)'I_{3n}(\hat{\beta}-\beta), \end{aligned} \tag{A.1}$$

where the definition of I_{1n} $(l = 1, 2, 3)$ should be apparent.

The proof will be done in two steps. In step one, we show that: (i) $nh_0^{q/2} I_{1n}(h_0, \lambda_0) \to N(0, \Omega)$ in distribution; (ii) $I_{2n}(h_0, \lambda_0) = O_p(n^{-1/2})$; (iii) $I_{3n}(h_0, \lambda_0) = O_p(1)$; and (iv) $\hat{\Omega}(h_0, \lambda_0) = \Omega + o_p(1)$. These results together with $\hat{\beta} - \beta = O_p(n^{-1/2})$ prove the result of Theorem 2.1 for the case with $(h, \lambda) = (h_0, \lambda_0)$.

In the second step we show that: (i) $n\hat{h}^{q/2} I_{1n}(\hat{h}, \hat{\lambda}) - nh_0^{1/2} I_{1n}(h_0, \lambda_0) = o_p(1)$; (ii) $I_{2n}(\hat{h}, \hat{\lambda}) - I_{jn}(h_0, \lambda_0) = o_p(n^{-1/2})$; (iii) $I_{3n}(\hat{h}, \hat{\lambda}) - I_{3n}(h_0, \lambda_0) = o_p(1)$; and (iv) $\hat{\Omega}(\hat{h}, \hat{\lambda}) - \hat{\Omega}(h_0, \lambda_0) = o_p(1)$. Steps one and two complete the proof of Theorem 2.1. □

Step one's results are proved in Lemma A.1. Below we prove step two.

Proof of step two (i). $n\hat{h}^{q/2} I_{1n}(\hat{h}, \hat{\lambda}) - nh_0^{1/2} I_{1n}(h_0, \lambda_0) = o_p(1)$.

Note that $h_0 = a_0 n^{-1/(q+4)}$, and $\lambda_0 = b_0 n^{-2/(q+4)}$. Write $\hat{h} = \hat{c}_1 n^{-1/(q+4)}$, and $\hat{\lambda} = \hat{c}_2 n^{-2/(q+4)}$, and denote $c_0 = (a_0, b_0)'$ and $\hat{c} = (\hat{c}_1, \hat{c}_2)'$. From $\hat{h}/h_0 \to 1$ and $\hat{\lambda}/\lambda_0 \to 1$ (in probability), we know that $\|\hat{c} - c_0\| \to 0$ in probability.

Let $h = c_1 n^{-1/(q+4)}$ and $\lambda = c_2 n^{-2/(q+4)}$. Define $A_n(c) = A_n(c_1, c_2) = nh^{q/2} I_{1n}(h, \lambda)$, and $B_n(c) = A_n(c) - A_n(c_0)$. Then $A_n(\cdot)$ and $B_n(\cdot)$ are both stochastic processes indexed by c. Then (i) becomes $B_n(\hat{c}) = o_p(1)$, i.e., we want to show that, for all $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr[|B_n(\hat{c})| < \varepsilon] = 1. \tag{A.2}$$

For $\delta > 0$, denote the δ-ball centered at c_0 by $C_\delta = \{c : \|c - c_0\| \leqslant \delta\}$. By Lemma A.2 we know that $A_n(\cdot)$ is tight. By the Arzela–Ascoli Theorem (see Theorem 8.2 of Billingsley, 1968, p. 55) we know that tightness implies the following stochastic equicontinuity condition: for any $\varepsilon > 0$, $\eta_1 > 0$, there exist a δ $(0 < \delta < 1)$ and an n_1, such that

$$\Pr\left[\sup_{\|c'-c\|<\delta} |A_n(c') - A_n(c)| > \varepsilon\right] < \eta_1 \tag{A.3}$$

for all $n \geqslant n_1$

Eq. (A.3) implies that

$$\Pr[|B_n(\hat{c})| > \varepsilon, \hat{c} \in C_\delta] \leqslant \Pr\left[\sup_{c\in C_\delta} |B_n(c)| > \varepsilon\right] < \eta_1 \tag{A.4}$$

for all $n \geqslant n_1$.

Also, from $\hat{c} \to c_0$ in probability we know that for any $\eta_2 > 0$ and for the δ given in (A.4), there exists an n_2 such that

$$\Pr[\hat{c} \notin C_\delta] \equiv \Pr[\|\hat{c} - c_0\| > \delta] < \eta_2 \tag{A.5}$$

for all $n \geqslant n_2$.

Therefore,

$$\begin{aligned}\Pr[|B_n(\hat{c})| > \varepsilon] &= \Pr[|B_n(\hat{c})| > \varepsilon, \hat{c} \in C_\delta] + \Pr[|B_n(\hat{c})| > \varepsilon, \hat{c} \notin C_\delta] \\ &< \eta_1 + \eta_2\end{aligned} \tag{A.6}$$

for all $n \geqslant \max\{n_1, n_2\}$ by (A.4) and (A.5), where we have also used the fact that $\{|B_n(\hat{c})| > \varepsilon, \hat{c} \notin C_\delta\}$ is a subset of $\{\hat{c} \notin C_\delta\}$ (if A is a subset of B, then $P(A) \leqslant P(B)$).

Eq. (A.6) is equivalent to (A.2) because ε, η_1 and η_2 can all be arbitrarily small. This completes the proof of (i). □

Proof of step two (ii)–(iv). *Proof of* (ii). $I_{2n}(\hat{h}, \hat{\lambda}) = O_p(n^{-1/2})$.

Write $h = c_1 n^{-1/(q+4)}$ and $\lambda = c_2 n^{-2/(q+4)}$. Then by using the same proof as Lemma A.2, one can show that $J_{2n}(c) \overset{\text{def}}{=} \sqrt{n}[n^{-2}\sum_i \sum_{j\neq i} u_i x_j' h^{-q} K_{\gamma,ij}] = O_p(1)$ for any c, and $J_{2n}(c)$, as a stochastic process (indexed by c), is tight under the sup-norm. The remaining arguments are the same as in the proof of (i) above; one can show that $J_{2n}(\hat{c}) - J_{2n}(c_0) = o_p(n^{-1/2})$. Therefore, $J_{2n}(\hat{c}) = O_p(J_{2n}(c_0)) = O_p(n^{-1/2})$ by Lemma A.1(ii).

Proof of (iii). $I_{3n}(\hat{c}) = O_p(1)$, and (iv): $\hat{\Omega}(\hat{c}) - \hat{\Omega}(c_0) = o_p(1)$.

Similarly, one can show that $I_{3n}(c)$ is tight, and that $I_{3n}(\hat{c}) = I_{3n}(c_0) + o_p(1) = O_p(n^{-1})$ because $I_{3n}(c_0) = I(h_0, \lambda_0) = O_p(n^{-1})$ by Lemma A.1. The proof of (iv) follows the same argument as (iii). □

Lemma A.1.

(i) $nh_0^{q/2} I_{1n}(h_0, \lambda_0) \to N(0, \Omega)$ *in distribution*,
(ii) $I_{2n}(h_0, \lambda_0) = O_p(n^{-1/2})$,

(iii) $I_{3n}(h_0, \lambda_0) = O_p(1)$, *and*
(iv) $\hat{\Omega}(h_0, \lambda_0) = \Omega + o_p(1)$.

Proof of (i). $I_{1n}(h_0, \lambda_0) = (n^2 h_0^q)^{-1} \sum_i \sum_{j \neq i} u_i u_j W(x_i^c - x_j^c / h_0) L(x_i^d, x_j^d, \lambda_0)$ is a second order degenerate U-statistic; its asymptotic variance is $(n^2 h_0^q)^{-1}\{E[\sigma^2(x_i) f(x_i)] \times [\int W^2(v)\, dv] + O(h_0^2 + \lambda_0)\} = (n^2 h_0^q)^{-1}\{\Omega + o(1)\}$. We also need to check the conditions for Hall's (1984) Theorem 1 (a central limit theorem (CLT) for a degenerate U-statistic). Define $H_n(z_1, z_2) = u_i u_j K_{\gamma_0, ij}$, where $K_{\gamma_0, ij} = W_{h_0, ij} L_{\lambda_0, ij}$, and $G(z_1, z_2) = E[H_n(z_1, z_3) H_n(z_2, z_3) | z_3]$. Define $W_{ij} = h^q W_{h_0, ij} \equiv W((x_i^c - x_j^c)/h_0)$. Then it is easy to verify that $E[H_n^2(z_i, z_j)] = E[\sigma^2(x_i)\sigma^2(x_j) W_{ij}^2 L_{\lambda_0, ij}^2] = O(h_0^{-q})$,
$E[H_n^4(z_i, z_j)] = E[\sigma^4(x_i)\sigma^4(x_j) W_{h_0, ij}^4 L_{\lambda_0, ij}^4] = O(h_0^{-3q})$, and $E[G_n^2(z_i, z_j)] = O(h_0^{-q})$. Hence,

$$\frac{E[G_n^2(z_1, z_2)] + n^{-1} E[H_n^4(z_1, z_2)]}{\{E[H_n^2(z_1, z_2)]\}^2} = O(h_0^q + (n h_0^q)^{-1}) = o(1). \tag{A.7}$$

Thus, by Hall's (1984) Theorem 1 we have

$$n h_0^{q/2} I_{1n}(h_0, \lambda_0) \to N(0, \Omega) \text{ in distribution.} \quad \square \tag{A.8}$$

Proof of (ii)–(iv). Straightforward calculations show that $E[\|I_{2n}\|^2] = O(n^{-1})$ and $E[\|I_{3n}\|] = O(1)$, which imply that $I_{2n} = O_p(n^{-1/2})$ and $I_{3n} = O_p(1)$. These together with $\hat{\beta} - \beta = O_p(n^{-1/2})$ imply that the last two terms at the right-hand side of Eq. (A.11) are smaller that I_{1n}, the first term, when $h = h_0$ and $\lambda = \lambda_0$.

Finally, it is easy to see that replacing $\hat{u}_i^2$ ($\hat{u}_j^2$) by u_i^2 (u_j^2) in $\hat{\Omega}(h_0, \lambda_0)$ yields the leading term of $\hat{\Omega}$. Let this leading term be denoted $\tilde{\Omega}$. Straightforward calculation shows that $E(\tilde{\Omega}) = \Omega + o(1)$, and $E\{[\tilde{\Omega}]^2\} = o(1)$. Therefore, $\hat{\Omega}(h_0, \lambda_0) = \tilde{\Omega} + o_p(1) = \Omega + o_p(1)$.

For the general nonlinear regression case, we can prove the results by using the Taylor expansion of $m(x_i, \hat{\beta}) = m(x_i, \beta) + \nabla m(x_i, \beta)(\hat{\beta} - \beta) + (1/2)(\hat{\beta} - \beta)' \nabla^2 m(x_i, \tilde{\beta})(\hat{\beta} - \beta)$, where $\tilde{\beta}$ is in the line segment between $\hat{\beta}$ and β. Using the fact that $\hat{\beta} - \beta = O_p(n^{-1/2})$, the proof carries though to the general case in a straightforward way. $\square$

Lemma A.2. *Let* $A_n(c) = n h^{q/2} I_{1n}(h, \lambda)$, *where* $h = c_1 n^{-1/(q+4)}$, $\lambda = c_2 n^{-2/(q+4)}$, $c = (c_1, c_2)$, $c_j \in [C_{j,1}, C_{j,2}]$ *with* $0 < C_{j,1} < C_{j,2} < \infty$ $(j = 1, 2)$.

Then the stochastic process $A_n(c)$ *indexed by* c *is tight under the sup-norm.*

Proof. Let $K_{c,ij}$ denote $K_{\gamma, ij}$ with $h = c_1 n^{1/(q+4)}$ and $\lambda = c_2 n^{-2/(q+4)}$. Then $K_{c,ij} = W(X_j - X_i / c_1 n^{-1/(q+4)}) L(X_j^d, X_i^d, c_2 n^{-2/(q+4)})$. Also, letting $\delta = q/(4+q)$, then $h^q = c^q n^{-q/(q+4)} = c^q n^{-\delta}$. Note that $h = c_1 n^{-\delta}$, and $\lambda = c_2 n^{-2\delta}$. Thus, $h^{-q/2} K_{c,ij} = c_1^{-q/2} n^{-\delta/2} W_{c_1, ij} L_{c_2, ij}$. Also note that $|L_{c_2', ij} - L_{c_2, ij}| \leqslant |(c_2')^{d_{x_i, x_j}} - c_2^{d_{x_i, x_j}}| \leqslant |c_2' - c_2|$ we have

$$\begin{aligned}
&(h')^{-q/2} K_{c', ij} - h^{-q/2} K_{c, ij} \\
&\quad = n^{\delta/2}\{(c_1')^{-q/2} W_{c_1', ij} L_{c_2', ij} - c_1^{-q/2} W_{c_1, ij} L_{c_2, ij}\} \\
&\quad = n^{\delta/2}\{(c_1')^{-q/2} W_{c_1', ij}[L_{c_2', ij} - L_{c_2, ij}] + [(c_1')^{-q/2} W_{c_1', ij} - c_1^{-q/2} W_{c_1, ij}] L_{c_2, ij}\} \\
&\quad \leqslant D_1 \left\{ (h')^{-q/2} W_{c_1', ij} |c_2' - c_2| + h^{-q/2} G\left(\frac{x_j - x_i}{h}\right) |c_1' - c_1| \right\},
\end{aligned} \tag{A.9}$$

where $D_1>0$ is a finite constant. In the last equality we used $|L_{c_2,ij}|\leqslant 1$ and Assumption (A2). We also replaced one of the $(c_1')^{-q/2}$ by $c_1^{-q/2}$, because $c_1, c_1' \in [C_{1,1}, C_{1,2}]$, are all bounded from above and below. The difference can be absorbed into D_1.

Using (A.9), we have

$$
\begin{aligned}
&\mathrm{E}\{[A_n(c')-A_n(c)]^2\}\\
&\quad=\frac{2(n-1)}{n}\,\mathrm{E}\left\{\sigma^2(x_i)\sigma^2(x_j)n^{\delta}\left[\frac{1}{(h')^{q/2}}K_{c',ij}-\frac{1}{h^{q/2}}K_{c,ij}\right]^2\right\}\\
&\quad\leqslant 4D_1\mathrm{E}\left\{\sigma^2(x_i)\sigma^2(x_j)\left[(h')^{-q}W^2\left(\frac{x_j-x_i}{h'}\right)|c_2'-c_2|^2+h^{-q}G\left(\frac{x_j-x_i}{h}\right)|c_1'-c_1|^2\right]\right\}\\
&\quad=4D_1\left\{\int f(x_i)^2\sigma^4(x_i)\left[\int W^2(v)\,\mathrm{d}v+\mathrm{O}(n^{-2\delta})\right](c_2'-c_2)^2\right.\\
&\qquad\left.+\left[\int G(v)^2\,\mathrm{d}v+\mathrm{O}(n^{-2\delta})\right](c_1'-c_1)^2\right\}\\
&\quad\leqslant D\|c'-c\|^2,
\end{aligned}
\tag{A.10}
$$

where D is a finite positive constant, and the $\mathrm{O}(n^{-2\delta})$ terms come from $\mathrm{O}(h^2)$ and $\mathrm{O}((h')^2)$. Therefore, $A_n(\cdot)$ and $B_n(\cdot)$ are tight by Theorem 15.6 of Billingsley (1968, p. 128), or Theorem 3.1 of Ossiander (1987). □

Proof of Theorem 2.2. The proof of Theorem 2.2 is almost identical to that of Theorem 2.1. We will only prove the linear null model case, as the general nonlinear regression model case follows in a straightforward manner. Using $\hat{u}_i^*=y_i^*-x_i'\hat{\beta}^*=u_i^*-x_i'(\hat{\beta}^*-\hat{\beta})$, $\hat{\beta}^*-\hat{\beta}=\mathrm{O_p}(n^{-1/2})$, and using arguments similar to the proof of Theorem 2.2, one can show that

$$
\begin{aligned}
\hat{I}_n^* &= \frac{1}{n(n-1)}\sum_i\sum_{j\neq i}u_i^*u_j^*K_{\hat{\gamma},ij}-2\left[\frac{1}{n(n-1)}\sum_i\sum_{j\neq i}u_i^*x_j'K_{\hat{\gamma},ij}\right](\hat{\beta}^*-\hat{\beta})\\
&\quad+(\hat{\beta}^*-\hat{\beta})'\left[\frac{1}{n(n-1)}\sum_i\sum_{j\neq i}x_ix_j'K_{\hat{\gamma},ij}\right](\hat{\beta}^*-\hat{\beta})\\
&= I_{1n}^*-2I_{2n}(\hat{\beta}^*-\hat{\beta})+(\hat{\beta}^*-\hat{\beta})'I_{3n}^*(\hat{\beta}^*-\hat{\beta}).
\end{aligned}
\tag{A.11}
$$

The proof parallels that of the proof of Theorem 2.1. We will show that

(i) $n\hat{h}^{q/2}I_{1n}^*/\sqrt{\hat{\Omega}^*}\,|\{x_i,y_i\}_{i=1}^n\to \mathrm{N}(0,1)$ in distribution in probability,
(ii) $I_{2n}^*=\mathrm{O_p}(n^{-1/2})$,
(iii) $I_{3n}^*=\mathrm{O_p}(1)$, and
(iv) $\hat{\Omega}^*=\hat{\Omega}+\mathrm{o_p^*}(1)$, where $A_n=\mathrm{O_p^*}(1)$ means that $E^*[|A_n|^2]\equiv E[|A_n|^2|\{x_i,y_i\}_{i=1}^n]=\mathrm{O_p}(1)$ and that $A_n=\mathrm{o_p^*}(\cdot)$ is defined as $\mathrm{E}^*[|A_n|^2]=\mathrm{o_p}(1)$.

We first prove (i). Using $\hat{u}_i^*=u_i^*+x_i'(\hat{\beta}-\hat{\beta}^*)$ and noting that $\hat{\beta}-\hat{\beta}^*=\mathrm{O_p^*}(n^{-1/2})$, it is easy to show that $I_{1n}^*=U_n^*+\mathrm{O_p^*}(n^{-1})$, where $U_n^*=2/n(n-1)\sum_i\sum_{j>i}u_i^*u_j^*W_{h_0}(x_i^{\mathrm{c}}-x_j^{\mathrm{c}}/h_0)L(x_i^{\mathrm{d}},x_j^{\mathrm{d}},\lambda_0)$. Obviously, conditional on the random sample $\{x_i,y_i\}_{i=1}^n$, u_i^* are mean

zero and mutually independent and have variance $\hat{u}_i^2$. Hence, $n\hat{h}^{q/2}U_n^*$ is a degenerate U-statistic with conditional variance $2\hat{h}^q/[n(n-1)]\sum_i\sum_{j\neq i}\hat{u}_i^2\hat{u}_j^2 W_{\hat{h}}^2(x_i^c - x_j^c/\hat{h})L(x_i^d, x_j^d, \hat{\lambda}) = \hat{\Omega}$. It is easy to check that $E^*(\hat{\Omega}^*) = \hat{\Omega} + o_p(1)$, and $var^*(\hat{\Omega}^*) = o_p(1)$. Thus, $n\hat{h}^{q/2}U_n^*/\sqrt{\hat{\Omega}^*}$ has mean zero and conditional variance $1 + o_p(1)$.

To show that U_n^* has an asymptotic normal distribution conditional on the random sample, we define $U_{n,ij}^* = 2/n(n-1)H_n^*(z_i^*, z_j^*)$, where $H_n^*(z_i^*, z_j^*) = u_i^* u_j^* K_{\hat{\gamma},ij}$. Then, $U_n^* = 2/n(n-1)\sum_i\sum_{j>i}H_n^*(z_i^*, z_j^*)$. We apply the CLT of de Jong (1987) for generalized quadratic forms to derive the asymptotic distribution of $U_n^*|\{x_i, y_i\}_{i=1}^n$. The reason for using de Jong's CLT instead of the one in Hall (1984) is that in the bootstrap world, the function $H_n^*(z_i^*, z_j^*)$ depends on i and j, because $z_i^* = (x_i, y_i^*)$. By de Jong (1987, Proposition 3.2) we know that $U_n^*/S_n^* \to N(0,1)$ in distribution in probability if G_I^*, G_{II}^* and G_{IV}^* are all $o_p(S_n^{*4})$, where $S_n^{*2} = E^*[U_n^{*2}]$, $G_I^* = \sum_i\sum_{j>i}E^*[U_{n,ij}^{*4}]$, $G_{II}^* = \sum_i\sum_{j>i}\sum_{l>j>i}[E^*(U_{n,ij}^{*2}U_{n,il}^{*2}) + E^*(U_{n,ji}^{*2}U_{n,jl}^{*2}) + E^*(U_{n,li}^{*2}U_{n,lj}^{*2})]$, and $G_{IV}^* = \binom{1}{2}\sum_i\sum_{j>i}\sum_s\sum_{t>s}E^*(U_{n,is}^{*2}U_{n,sj}^{*2}U_{n,ti}^{*2}U_{n,js}^{*2})$.

Now,

$$E^*[H_n^{*2}(z_i^*, z_j^*)] = E^*[u_i^{*2}u_j^{*2}K_{\hat{\gamma},ij}^2] = K_{\hat{\gamma},ij}^2\hat{u}_i^2\hat{u}_j^2.$$

Hence,

$$\begin{aligned} S_n^{*2} &= \frac{4}{n^2(n-1)^2}\sum_i\sum_{j>i}E^*[H_n(z_i^*, z_j^*)^2] = \frac{4}{n^2(n-1)^2}\sum_i\sum_{j>i}K_{\hat{\gamma},ij}^2\hat{u}_i^2\hat{u}_j^2 \\ &= \frac{1}{n(n-1)(\hat{h}_1\ldots\hat{h}_q)}\hat{\Omega} = O_p((\hat{h}_1\ldots\hat{h}_q)n^2). \end{aligned}$$

Therefore,

$$1/S_n^{*2} = O_p(n^2(\hat{h}_1\ldots\hat{h}_q)) \quad \text{and} \quad 1/S_n^{*4} = O_p(n^4(\hat{h}_1\ldots\hat{h}_q)^2).$$

Next, $E^*[H_n^{*4}(z_i^*, z_j^*)] = cK_{\hat{\gamma},ij}^4\hat{u}_i^4\hat{u}_j^4$, where c is a finite positive constant. Therefore,

$$\begin{aligned} G_I^* &= \frac{16}{n^4(n-1)^4}\sum_i\sum_{j>i}E^*[U_{n,ij}^{*4}] = \frac{c16}{n^4(n-1)^4}\sum_i\sum_{j>i}K_{\hat{\gamma},ij}^4\hat{u}_i^4\hat{u}_j^4 \\ &= O_p(n^{-6}(\hat{h}_1\ldots\hat{h}_q)^{-3}). \end{aligned}$$

From the above calculation it should be apparent that the probability orders of G_I^*, G_{II}^* and G_{IV}^* are solely determined by the factor of n's and $(\hat{h}_1\ldots\hat{h}_p)$'s through $K_{ij,\hat{\gamma}}$. Therefore, tedious but straightforward calculations show that

$$G_{II}^* \sim n^{-8}\sum_i\sum_{j>i}\sum_{s>j>i}[K_{ij,\hat{\gamma}}^2K_{is,\hat{\gamma}}^2 + K_{js,\hat{\gamma}}^2K_{ji,\hat{\gamma}}^2 + K_{si,\hat{\gamma}}^2K_{sj,\hat{\gamma}}^2] = O_p(n^{-5}(\hat{h}_1\ldots\hat{h}_q)^{-2}),$$

$$G_{IV}^* \sim n^{-8}\sum_i\sum_{j>i}\sum_s\sum_{t>s}[K_{si,\hat{\gamma}}K_{sj,\hat{\gamma}}K_{ti,\hat{\gamma}}K_{tj,\hat{\gamma}}] = O_p(n^{-4}(\hat{h}_1\ldots\hat{h}_q)^{-1}).$$

Therefore, $G_k^*/S_n^{*4} = o_p(1)$ for all $k = I, II, IV$, and by de Jong's (1987) CLT for generalized quadratic forms, we know that

$$U_n^*/S_n^* \to N(0,1) \text{ in distribution in probability.} \tag{A.12}$$

Next, it is easy to see that

$$\begin{aligned}\hat{\Omega}^*_{n,\hat{\gamma}} &= \frac{2(\hat{h}_1 \dots \hat{h}_q)}{n(n-1)} \sum_i \sum_{j \neq i} \hat{u}_i^{*2} \hat{u}_j^{*2} K^2_{\hat{\gamma},ij} \\ &= \frac{2(\hat{h}_1 \dots \hat{h}_q)}{n(n-1)} \sum_i \sum_{j \neq i} \hat{u}_i^2 \hat{u}_j^2 K^2_{\hat{\gamma},ij} + \mathrm{o}^*_{\mathrm{p}}(1) \\ &= \hat{\Omega} + \mathrm{o}^*_{\mathrm{p}}(1). \end{aligned} \tag{A.13}$$

Eqs. (A.12), (A.13) and $S^{*2}_n = 1/n(n-1)\hat{h}_1 \dots \hat{h}_q \hat{\Omega}$ lead to

$$n(\hat{h}_1 \dots \hat{h}_q)^{1/2} U^*_n / \sqrt{\hat{\Omega}^*} \to \mathrm{N}(0,1) \text{ in distribution in probability,}$$

completing the proof of (i). □

Similarly, one can show that $\mathrm{E}^*[\|I^*_{2n}\|^2] = \mathrm{O}_{\mathrm{p}}(n^{-1})$ and $\mathrm{E}^*[\|I_{3n}\|] = \mathrm{O}_{\mathrm{p}}(1)$. Hence, $I^*_{2n} = \mathrm{O}^*_{\mathrm{p}}(n^{-1/2})$ and $I^*_{3n} = \mathrm{O}^*_{\mathrm{p}}(1)$. Therefore, we conclude that $n(\hat{h}_1 \dots \hat{h}_q)^{1/2} \hat{I}^*_n / \sqrt{\hat{\Omega}^*_n}$ has the same asymptotic distribution as that of $n(\hat{h}_1 \dots \hat{h}_q)^{1/2} U^*_n / \sqrt{\hat{\Omega}^*_{n,\hat{\gamma}}}$. Hence,

$$n(\hat{h}_1 \dots \hat{h}_q)^{1/2} \hat{I}^*_n / \sqrt{\hat{V}^*_{n,\hat{\gamma}}} \to \mathrm{N}(0,1) \text{ in distribution in probability.} \tag{A.14}$$

Since N(0, 1) is a continuous distribution, by Polyā's Theorem (Bhattacharya and Rao, 1986), we know that (A.14) is equivalent to (2.13). This finishes the proof of Theorem 2.2.

References

Ahmad, I.A., Cerrito, P.B., 1994. Nonparametric estimation of joint discrete-continuous probability densities with applications. Journal of Statistical Planning and Inference 41, 349–364.

Ait-Sahalia, Y., Bickel, P., Stoker, T.M., 2001. Goodness-of-fit tests for kernel regression with an application to option implied volatilities. Journal of Econometrics 105, 363–412.

Aitchison, J., Aitken, C.G.G., 1976. Multivariate binary discrimination by the kernel method. Biometrika 63, 413–420.

Andrews, D.W.K., 1997. A conditional Kolmogorov test. Econometrica 65, 1097–1128.

Bhattacharya, R.N., Rao, R.R., 1986. Normal Approximations and Asymptotic Expansions. Krieger, Malabar, FL.

Bierens, H.J., 1983. Uniform consistency of kernel estimators of a regression function under generalized conditions. Journal of American Statistical Association 77, 699–707.

Bierens, H.J., 1987. Kernel estimators of regression functions. In: Truman, F.B. (Ed.), Advances in Econometrics: Fifth World Congress, vol. 1. Cambridge University Press, Cambridge, MA, pp. 99–144.

Bierens, H.J., Ploberger, W., 1997. Asymptotic theory of integrated conditional moment tests. Econometrica 65, 1129–1152.

Billingsley, P., 1968. Convergence of Probability Measure, second ed. Wiley, New York.

Bowman, A.W., 1980. A note on consistency of the kernel method for the analysis of categorical data. Biometrika 67, 682–684.

de Jong, P., 1987. A central limit theorem for generalized quadratic forms. Probability and Related Fields 75, 261–277.

de Jong, R.M., 1996. The Bierens test under data dependence. Journal of Econometrics 72, 1–32.

Eubank, R., Hart, J., 1992. Testing goodness-of-fit in regression via order selection criteria. The Annals of Statistics 20, 1412–1425.

Eubank, R., Spiegelman, S., 1990. Testing the goodness-of-fit of a linear model via nonparametric regression techniques. Journal of the American Statistical Association 85, 387–392.

Fan, Y., Li, Q., 1996. Consistent model specification tests: omitted variables and semiparametric functional forms. Econometrica 64, 865–890.
Fan, Y., Li, Q., 2000. Consistent model specification tests: kernel-based test versus Bierens' ICM tests. Econometric Theory 16, 1016–1041.
Fahrmeir, L., Tutz, G., 1994. Multivariate Statistical Modeling Based on Generalized Models. Springer, New York.
Freeman, R.B., Medoff, J.L., 1984. What Do Unions Do? Basic Books, New York.
Grund, B., Hall, P., 1993. On the performance of kernel estimators for high-dimensional sparse binary data. Journal of Multivariate Analysis 44, 321–344.
Hall, P., 1981. On nonparametric multivariate binary discrimination. Biometrika 68, 287–294.
Hall, P., 1984. Central limit theorem for integrated square error of multivariate nonparametric density estimators. Journal of Multivariate Analysis 14, 1–16.
Hall, P., Wand, M., 1988. On nonparametric discrimination using density differences. Biometrika 75, 541–547.
Hall, P., Racine, J., Li, Q., 2004. Cross-validation and the estimation of conditional probability densities. Journal of The American Statistical Association 99, 1015–1026.
Hall, P., Li, Q., Racine, J., 2005. Nonparametric estimation of regression functions in the presence of irrelevant regressors. Manuscript, Texas A&M University.
Härdle, W., Mammen, E., 1993. Comparing nonparametric versus parametric regression fits. The Annals of Statistics 21, 1926–1947.
Hart, J.D., 1997. Nonparametric Smoothing and Lack-of-fit Tests. Springer, New York.
Hong, Y., White, H., 1995. Consistent specification testing via nonparametric series regression. Econometrica 63, 1133–1159.
Horowitz, J.T., Spokoiny, V.G., 2001. An adaptive, rate-optimal test of a parametric model against a nonparametric alternative. Econometrica 69, 599–631.
Ichimura, H., 2000. Asymptotic distribution of non-parametric and semiparametric estimators with data dependent smoothing parameters. Manuscript.
Lanot, G., Walker, I., 1998. The union/non-union wage differential: an application of semi-parametric methods. Journal of Econometrics 84, 327–349.
Li, Q., Racine, J., 2004. Cross-validated local linear nonparametric regression. Statistica Sinica 14, 485–512.
Li, Q., Racine, J., 2006. Nonparametric Econometrics: Theory and Practice. Princeton University Press, Princeton, NJ.
Li, Q., Wang, S., 1998. A simple consistent bootstrap test for a parametric regression function. Journal of Econometrics 87, 145–165.
Li, Q., Hsiao, C., Zinn, J., 2003. Consistent specification tests for semiparametric/nonparametric models based on series estimation methods. Journal of Econometrics 112, 295–325.
Mammen, E., 1992. When Does Bootstrap Work? Asymptotic Results and Simulations. Springer, New York.
Murphy, K.M., Welch, F., 1990. Empirical age-earnings profiles. Journal of Labor Economics 8, 202–229.
Ossiander, M., 1987. A central limit theorem under metric entropy with L_2 bracketing. Annals of Probability 15, 897–919.
Racine, J., Li, Q., 2004. Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics 119, 99–130.
Robinson, P.M., 1989. Hypothesis testing in semiparametric and nonparametric models for econometric time series. Review of Economic Studies 56, 511–534.
Robinson, P.M., 1991. Consistent nonparametric entropy-based testing. Review of Economic Studies 58, 437–453.
Scott, D., 1992. Multivariate in Density Estimation: Theory, Practice, and Visualization. Wiley, New York.
Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.
Wooldridge, J., 1992. A test for functional form against nonparametric alternatives. Econometric Theory 8, 452–475.
Yatchew, A.J., 1992. Nonparametric regression tests based on least squares. Econometric Theory 8, 435–451.
Zheng, J.X., 1996. A consistent test of functional form via nonparametric estimation technique. Journal of Econometrics 75, 263–289.