Econometric Theory, **29**, 2013, 629–641.
doi:10.1017/S0266466612000382

A SMOOTH NONPARAMETRIC CONDITIONAL DENSITY TEST FOR CATEGORICAL RESPONSES

CONG LI
Shanghai University of Finance and Economics

JEFFREY S. RACINE
McMaster University

We propose a consistent kernel-based specification test for conditional density models when the dependent variable is categorical/discrete. The method is applicable to popular parametric binary choice models such as the logit and probit specification and their multinomial and ordered counterparts, along with parametric count models, among others. The test is valid when the conditional density function contains both categorical and real-valued covariates. Theoretical support for the test and for a bootstrap-based version of the test is provided. Monte Carlo simulations are conducted to assess the finite-sample performance of the proposed method.

1. INTRODUCTION

Model specification tests can be found in every practitioner's toolkit. In the discrete outcome context, existing tests include Hausman and McFadden's (1984) test of correct specification of the multinomial logit model, Mora and Moro-Egido's (2008) test of correct specification of ordered discrete models, or Rodriguez-Poo, Sperlich, and Vieu's (2004) test for correct specification of semiparametric models. However, existing tests are applicable to a particular subset of categorical dependent variable models or are not consistent tests in that there exist alternative models for which the tests lack power or are not applicable to models containing both numeric and categorical covariates. An omnibus procedure capable of handling a rich array of categorical dependent variable models that is consistent and can admit both numeric and categorical data would have obvious appeal.

Fan, Li, and Min (2006) proposed a nonsmoothing test for correct specification of parametric conditional distributions based on previous work by Zheng (2000), where nonsmoothing here refers to nonsmoothing of the response (they do smooth the categorical covariates). In the current paper we extend the approach of Fan et al. to obtain a smooth version of their nonsmoothing test that displays

Racine would like to acknowledge the generous support of SHARCNET (www.sharcnet.ca). Address correspondence to Jeffrey S. Racine, Department of Economics, McMaster University, Hamilton, ON, Canada L8S 4M4; e-mail: racinej@mcmaster.ca.

© Cambridge University Press 2012

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

finite-sample power improvements over its nonsmoothing counterpart (i.e., we smooth the categorical/discrete response in addition to the categorical/discrete covariates).

Zheng (2000) proposed a kernel-based nonparametric test for a parametric density function. Suppose that the data consist of $\{y_i, x_i\}_{i=1}^n$, an i.i.d. sample drawn from the distribution of (y, x) with the joint density function $p(y, x)$. Let $p(y|x)$ denote the conditional density function of y given x. We are interested in testing whether $p(y|x)$ belongs to a particular parametric family. Let $f(y|x, \theta)$ denote a parametric conditional density function with θ being an unknown $k \times 1$-dimensional parameter vector. The null hypothesis is given by

$$H_0\colon \Pr[p(y_i|x_i) = f(y_i|x_i, \theta_0)] = 1 \quad \text{for some } \theta_0 \in \Theta,$$

where Θ is a compact set in $\mathbb{R}^k$. The alternative hypothesis is the negation of the null

$$H_1\colon \Pr[p(y_i|x_i) = f(y_i|x_i, \theta)] < 1 \quad \text{for all } \theta \in \Theta.$$

Zheng (2000) proposed using the following quantity as a basis for a consistent test for H_0, which is simply a linearized version of the Kullback-Leibler information function $I(p, f)$ between two conditional distribution functions $p(y|x)$ and $f(y|x, \theta_0)$ and is given by

$$I_1(p, f) = \mathbb{E}\left[\frac{p(y_i, x_i) - f(y_i|x_i, \theta_0)p_1(x_i)}{f(y_i|x_i, \theta_0)}\right]. \tag{1}$$

Zheng has shown that $I_1(p, f) \geq 0$, and the equality holds if and only if H_0 is true (i.e., $p(y|x)$ equals $f(y|x, \theta_0)$ almost everywhere). Therefore, $I_1(p, f)$ also serves as a proper measure to test for H_0.

Zheng (2000) only considered the case for which all variables (i.e., covariates and outcome) are continuous. Fan et al. (2006) extended Zheng's test to include both continuous and discrete explanatory variables (x is a mixed variable), while they also allow the dependent variable y to be either continuous or discrete.

Let $x = (x^c, x^d)$, where x^c is a $q \times 1$ continuous variable, and x^d is an $r \times 1$ discrete variable. Fan et al. (2006) use x_{is}^c (x_{is}^d) to denote the sth component of x_i^c (x_i^d). They further assume that x_{is}^d takes the values $0, 1, \ldots, c_s - 1$ (i.e., it can assume c_s different values).

For each discrete component of x^d, Fan et al. (2006) use the Aitchison and Aitken (1976) kernel given by $l(x_{is}^d, x_{js}^d, \lambda_s) = 1 - \lambda_s$ if $x_{is}^d = x_{js}^d$, and $l(x_{is}^d, x_{js}^d, \lambda_s) = \lambda_s/(c_s - 1)$ if $x_{is}^d \neq x_{js}^d$; for the $r \times 1$ variable x^d they use the product kernel given by $L(x_i^d, x_j^d, \lambda) = \prod_{s=1}^r l(x_{is}^d, x_{js}^d, \lambda_s) = \prod_{s=1}^r \{\lambda_s/(c_s - 1)\}^{N_{is}(x)}(1 - \lambda_s)^{1-N_{is}(x)}$, where $N_{is}(x) = I(x_{is}^d \neq x_{js}^d)$, in which $I(.)$ is the usual indicator function and $\lambda = (\lambda_1, \ldots, \lambda_r)$ are the smoothing parameters for the discrete components that satisfy $0 \leq \lambda_s \leq (c_s - 1)/c_s$.

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

For the $q \times 1$ continuous component x^c, a standard (second-order) product kernel function can be used and is given by $W(x_i^c, x_j^c, h) = \prod_{s=1}^{q} w((x_{is}^c - x_{js}^c)/h_s)/h_s$, where $w((x_{is}^c - x_{js}^c)/h_s)/h_s$ is a standard univariate kernel such as the Epanechnikov kernel. Therefore, for the mix of continuous and categorical covariates $x = (x^c, x^d)$, the kernel function is defined by $K_{\gamma,ij} = K_\gamma(x_i, x_j) \stackrel{def}{=} W(x_i^c, x_j^c, h) \times L(x_i^d, x_j^d, \lambda)$, where $\gamma = (h, \lambda) \equiv (h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$.

Fan et al. (2006) propose estimating $p(y_i, x_i)$ and $p_1(x_i)$ using the following approach. Assuming that y_i is a discrete variable, then Fan et al. estimate $p(y_i, x_i)$ and $p_1(x_i)$ by the leave-one-out kernel estimators,

$$\hat{p}_{-i}(y_i, x_i) = \frac{1}{n} \sum_{j \neq i}^{n} I(y_i = y_j) K_\gamma(x_i, x_j) \quad \text{and} \quad \hat{p}_{1,-i}(x_i) = \frac{1}{n} \sum_{j \neq i}^{n} K_\gamma(x_i, x_j). \tag{2}$$

A smoothed leave-one-out estimator of $f(y_i|x_i, \theta_0) p_1(x_i)$ is given by

$$\begin{aligned} \tilde{p}_{-i}(y_i, x_i) &= \frac{1}{n} \sum_{j \neq i}^{n} \sum_{y} I(y_i = y) K_\gamma(x_i, x_j) f(y|x_j, \hat{\theta}) \\ &= \frac{1}{n} \sum_{j \neq i}^{n} K_\gamma(x_i, x_j) f(y_i|x_j, \hat{\theta}), \end{aligned} \tag{3}$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ_0 under the parametric null model.

Replacing $p(y_i, x_i)$, $f(y_i|x_i, \theta_0) p_1(x_i)$ and $f(y_i|x_i, \theta_0)$ by $\hat{p}_{-i}(y_i, x_i)$, $\tilde{p}_{-i}(y_i, x_i)$ and $f(y_i|x_i, \hat{\theta})$ in (1) leads to the feasible test statistic

$$J_{n,\gamma}^0 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left\{ \frac{K_\gamma(x_i, x_j)}{f(y_i|x_i, \hat{\theta})} \left[I(y_i = y_j) - f(y_i|x_j, \hat{\theta}) \right] \right\}. \tag{4}$$

2. A SMOOTH NONPARAMETRIC CONDITIONAL DENSITY TEST

When x contains discrete components, Fan et al. (2006) suggest also smoothing the discrete components of x. It is well known that one can choose to not smooth the discrete variables in nonparametric estimation or testing problems by resorting to sample splitting. The rationale for smoothing the discrete components of x is that this avoids splitting the sample into discrete subsets ("cells") and, by borrowing information from nearby cells, one can realize efficiency gains in finite-sample settings. However, when y is a discrete variable, Fan et al. rely on sample-splitting methods to deal with the dependent variable y. As was the case for the discrete components of x, it is reasonable to conjecture that by smoothing the discrete outcome y in an appropriate manner, one may obtain further finite-sample (i.e., power) gains for testing a conditional parametric density's functional form. We

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

will observe via Monte Carlo simulation that the improvement in power can be nontrivial.

In this paper we propose smoothing y by replacing the indicator function $I(y_i = y_j)$ and $I(y_i = y)$ in Fan et al.'s (2006) test by the smooth kernel functions $l(y_i, y_j, \lambda_0)$ and $l(y_i, y, \lambda_0)$, where λ_0 is the smoothing parameter, and

$$l(y_i, y_j, \lambda_0) = (1-\lambda_0) I(y_i = y_j) + \frac{\lambda_0}{c_0 - 1} I(y_i \neq y_j) \quad \text{if } y_i \text{ is an unordered variable,} \quad \textbf{(5)}$$

where $\lambda_0 \in [0, (c_0 - 1)/c_0]$, and

$$l(y_i, y_j, \lambda_0) = (1-\lambda_0) I(y_i = y_j) + \frac{\lambda_0^{|y_i - y_j|}}{2} I(y_i \neq y_j) \quad \text{if } y_i \text{ is an ordered variable,} \quad \textbf{(6)}$$

where $\lambda_0 \in [0, 1]$.

The kernel defined in equation (5) is referred to as a "uniform weight kernel function" because it gives equal (uniform) weight $\lambda_0/(c_0 - 1)$ for all $y_j \neq y_i$ when estimating the conditional density $f(y_i|x_i)$. The kernel defined in equation (6) is referred to as "geometric weight kernel function." When $y_j \neq y_i$, the weight attached to y_j is proportional to $\lambda_0^{|y_j - y_i|}$. Since $\lambda_0 < 1$ in general, its weight decays geometrically as $|y_j - y_i|$ increases.

We estimate $p(y_i, x_i)$ and $p_1(x_i)$ by the leave-one-out kernel estimators

$$\hat{p}_{-i}(y_i, x_i) = \frac{1}{n} \sum_{j \neq i}^{n} l(y_i, y_j, \lambda_0) K_\gamma(x_i, x_j),$$

$$\hat{p}_{1,-i}(x_i) = \frac{1}{n} \sum_{j \neq i}^{n} K_\gamma(x_i, x_j),$$

$$\tilde{p}_{-i}(y_i, x_i) = \frac{1}{n} \sum_{j \neq i}^{n} \sum_{y} l(y_i, y, \lambda_0) K_\gamma(x_i, x_j) f(y|x_j, \hat{\theta}).$$

Using $\hat{p}_{-i}(y_i, x_i)$, $\hat{p}_{1,-i}(x_i)$ and $\tilde{p}_{-i}(y_i, x_i)$ introduced above, we define our test statistic as

$$J_{n,\gamma} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left\{ \frac{K_\gamma(x_i, x_j)}{f(y_i|x_i, \hat{\theta})} \left[l(y_i, y_j, \lambda_0) - \sum_{y \in D_y} l(y_i, y, \lambda_0) f(y|x_j, \hat{\theta}) \right] \right\}, \quad \textbf{(7)}$$

where D_y is the support of y.

We suggest using cross-validation methods for choosing the smoothing parameters, and denote by $\hat{\gamma} = (\hat{h}_1, \ldots \hat{h}_q, \hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_r)$ the cross-validated smoothing parameter vector. The asymptotic behavior of $\hat{\gamma}$ is summarized in the following lemma.

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

LEMMA 2.1. *Under the regularity conditions given in Hall, Recine, and Li (2004),*

$$n^{1/(q+4)}\hat{h}_s \xrightarrow{p} a_s^0 \text{ for } 1 \le s \le q;\ n^{2/(q+4)}\hat{\lambda}_s \xrightarrow{p} b_s^0 \quad \text{for } 0 \le s \le r,$$

where $a_s^0\ (>0)$ *and* $b_s^0\ (\ge 0)$ *are some finite constants.*

THEOREM 2.1. *Suppose that* y *is an unordered discrete variable. Under the regularity conditions given in Fan et al. (2006) and defining* $T_n = n\sqrt{\hat{h}_1 \dots \hat{h}_q}\, J_{n,\hat{\gamma}} \Big/ \sqrt{\hat{V}_a}$ *(*$\hat{V}_a$ *is defined in the Appendix), then we have*

(i) under H_0, $T_n \xrightarrow{d} N(0,1)$;
(ii) under H_1, $\Pr[T_n > B_n] \to 1$ *for any nonstochastic sequence* $B_n = o\left(n(h_1 \dots h_q)^{1/2}\right)$.

THEOREM 2.2. *Suppose that* y *is an ordered discrete variable. Under the same conditions as given in Fan et al. (2006) and defining* $T_n = n\sqrt{\hat{h}_1 \dots \hat{h}_q}\, J_{n,\hat{\gamma}} \Big/ \sqrt{\hat{V}_b}$ *(*$\hat{V}_b$ *is defined in the Appendix), we have*

(i) under H_0, $T_n \xrightarrow{d} N(0,1)$;
(ii) under H_1, $\Pr[T_n > B_n] \to 1$, *for any nonstochastic sequence* $B_n = o\left(n(h_1 \dots h_q)^{1/2}\right)$.

2.1. A Bootstrap Test Procedure

It is well known that consistent nonparametric tests often suffer from substantial finite-sample size distortions. To overcome this limitation, Fan et al. (2006) propose a bootstrap procedure to accurately approximate the finite-sample null distribution of T_n. It involves the following steps.

Step (i) Generate the ith bootstrap value of the dependent variable y from the parametric conditional distribution $f(\cdot|x_i, \hat{\theta})$. Denote this value by y_i^* $(i = 1, \dots, n)$. This delivers the bootstrap sample $\{x_i, y_i^*\}_{i=1}^n$.

Step (ii) Based on the parametric null model, estimate θ using the bootstrap sample. Let $\hat{\theta}^*$ denote the resulting estimator. Compute the bootstrap statistic T_n^* in the same way as T_n except that $\{y_i\}_{i=1}^n$ and $\hat{\theta}$ are replaced by $\{y_i^*\}_{i=1}^n$ and $\hat{\theta}^*$, respectively. Note that Fan et al. (2006) use the same cross-validation selected smoothing parameter $\hat{\gamma}$ in computing the bootstrap statistics. There is no re-cross-validation needed when computing $J_{n,\hat{\gamma}}^*$.

Step (iii) Repeat steps (i) and (ii) a large number of times, say B times, and use the empirical distribution of the B bootstrap statistics $\{T_n^*\}_{j=1}^B$ to approximate the null distribution of T_n.

Step (iv) The bootstrap test rejects H_0 at significance level α if T_n exceeds the empirical αth percentile of $\{T_n^*\}_{j=1}^B$.

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

2.2. Asymptotic Validity of the Bootstrap Procedure

We now establish the asymptotic validity of the proposed bootstrap procedure, thereby justifying the use of the bootstrap for our testing procedure.

THEOREM 2.3. *Suppose y is an unordered discrete variable. Assume the same conditions as in Theorem 2.2, except the null hypothesis, and define* $T_n^* = n\sqrt{\hat{h}_1 \dots \hat{h}_q} J_{n,\hat{\gamma}}^* \Big/ \sqrt{\hat{V}_a^*}$, *where* $\hat{V}_a^*$ *is defined the same way as* $\hat{V}_a$ *except that* y_i *is replaced by* y_i^*. *Then we have*

$$\sup_{z\in\mathbb{R}} \left| P\left(T_n^* \le z \middle| \{x_i, y_i\}_{i=1}^n\right) - \Phi(z) \right| = o_p(1),$$

where $\Phi(.)$ *is the cumulative distribution function of a standard normal random variable.*

THEOREM 2.4. *Suppose y is an ordered discrete variable. Assume the same conditions as in Theorem 2.2, except the null hypothesis, and define* $T_n^* = n\sqrt{\hat{h}_1 \dots \hat{h}_q} J_{n,\hat{\gamma}}^* \Big/ \sqrt{\hat{V}_b^*}$, *where* $\hat{V}_b^*$ *is defined the same way as* $\hat{V}_a$ *except that* y_i *is replaced by* y_i^*. *Then we have*

$$\sup_{z\in\mathbb{R}} \left| P\left(T_n^* \le z \middle| \{x_i, y_i\}_{i=1}^n\right) - \Phi(z) \right| = o_p(1),$$

where $\Phi(.)$ *is the cumulative distribution function of a standard normal random variable.*

Proofs appear in the Appendix.

3. MONTE CARLO RESULTS

We consider a smoothed version of Fan et al.'s (2006) conditional density test in an ordered outcome setting. We compare the smooth and frequency (nonsmooth) versions of the test, incorporating a modification of the binary outcome data generating process (DGP) found in Fan et al. (2006). Note that Fan et al. consider $\beta_2 \in \{0, 1\}$, while we consider $\beta_2 \in \{0, -1\}$, and the alternative DGP they consider involves $\beta_3 x^2$. Frequently their DGP lies to the right of the cutoff point; i.e., $\beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x^2 + u_i > 0$ for almost all n leading to $y_i = 1$ for almost all $i = 1, \dots, n$. By using $\beta_2 = \{0, -1\}$ to offset the positive intercept ($\beta_0 = 1$) and by using an alternative DGP involving $\sin(0.5\pi x)$ (which oscillates around zero, taking positive and negative values) we obtain a DGP that is more "balanced," which is desirable in the ordered setting we consider below.

For the results reported below we consider an ordered probit specification in which the null DGP is given by $\zeta_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i$ with $x_i \sim N(0, 1)$, $u_i \sim N(0, 0.25)$, and $z_i \in \{0, 1\}$ with $\Pr(z_i = 1) = 0.5$. We generate $y_i \in \{0, 1, 2, \dots, c-1\}$ with $c = c(2, 3, 5, 7, 9)$, where y_i is determined by the

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

value of ζ_i lying in equal probability intervals defined by the value of c and cutoff values from the normal distribution. For instance, when $c = 3$ $y_i \in \{0, 1, 2\}$ when $\zeta_i \in [[-\infty, -0.43), [-0.43, 0.43), [0.43, \infty]]$ representing three equally probable intervals. By changing the value of c we can assess the impact of smoothing y as the number of unique outcomes increases.

The alternative DGP is given by $\zeta_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 \sin(\pi x_i/2) + u_i$. We let $\beta_3 \in \{0, 1\}$. When $\beta_3 = 0$ we can investigate the test's size, while when $\beta_3 = 1$ we investigate the test's power.

We conduct $B = 399$ bootstrap replications and $M = 1,000$ Monte Carlo replications and report results for sample sizes $n = (200, 300, 400)$. We report empirical rejection frequencies in Tables 1 and 2 for levels $\alpha = (0.01, 0.05, 0.10)$.

Tables 1 and 2 reveal that the smoothed y test has reasonable size and also size comparable to the frequency y comparator. Furthermore, power for the smoothed and unsmoothed versions of the test increase with both n and c. Moreover, it is evident that power for the smoothed version of the test proposed herein is substantially higher than that for the unsmoothed version, thereby

TABLE 1. Size: Entries represent empirical rejection frequencies for the proposed bootstrap test of correct specification under the null. Columns corresponding to "Smoothed y" are those for the proposed test, while columns corresponding to "Frequency y" are those corresponding to the unsmoothed test of Fan et al. (2006).

		Smoothed y			Frequency y		
n	c	0.010	0.050	0.100	0.010	0.050	0.100
200	2	0.016	0.073	0.135	0.016	0.073	0.135
300	2	0.020	0.082	0.153	0.020	0.082	0.153
400	2	0.016	0.064	0.119	0.016	0.064	0.119
200	3	0.015	0.074	0.132	0.014	0.072	0.129
300	3	0.012	0.060	0.116	0.013	0.058	0.115
400	3	0.014	0.063	0.120	0.014	0.063	0.120
200	5	0.010	0.071	0.130	0.010	0.072	0.129
300	5	0.016	0.065	0.122	0.015	0.064	0.123
400	5	0.020	0.055	0.115	0.019	0.054	0.115
200	7	0.008	0.062	0.121	0.007	0.060	0.124
300	7	0.013	0.058	0.108	0.014	0.058	0.108
400	7	0.007	0.045	0.100	0.007	0.045	0.099
200	9	0.014	0.059	0.120	0.015	0.059	0.118
300	9	0.008	0.055	0.117	0.008	0.051	0.114
400	9	0.011	0.063	0.116	0.011	0.061	0.115

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

TABLE 2. Power: Entries represent empirical rejection frequencies for the proposed bootstrap test of correct specification under the alternative. Columns corresponding to "Smoothed y" are those for the proposed test, while columns corresponding to "Frequency y" are those corresponding to the unsmoothed test of Fan et al. (2006).

		Smoothed y			Frequency y		
n	c	0.010	0.050	0.100	0.010	0.050	0.100
200	2	0.130	0.220	0.303	0.100	0.204	0.293
300	2	0.217	0.308	0.377	0.140	0.277	0.357
400	2	0.291	0.361	0.421	0.197	0.315	0.386
200	3	0.263	0.353	0.412	0.141	0.291	0.363
300	3	0.408	0.485	0.537	0.205	0.368	0.450
400	3	0.507	0.575	0.618	0.270	0.427	0.504
200	5	0.370	0.502	0.581	0.191	0.357	0.463
300	5	0.572	0.650	0.692	0.324	0.448	0.537
400	5	0.696	0.768	0.798	0.419	0.572	0.646
200	7	0.422	0.546	0.612	0.213	0.367	0.458
300	7	0.604	0.687	0.720	0.354	0.495	0.554
400	7	0.739	0.801	0.822	0.473	0.593	0.666
200	9	0.438	0.579	0.647	0.245	0.398	0.474
300	9	0.622	0.713	0.765	0.381	0.524	0.609
400	9	0.768	0.827	0.851	0.537	0.652	0.706

underscoring the utility of smoothing the outcome in conditional probability models in finite-sample settings.

REFERENCES

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

Fan, Y., Q. Li, & I. Min (2006) A nonparametric bootstrap test of conditional distributions. *Econometric Theory* 22, 587–613.

Hall, P., J. Racine, & Q. Li (2004) Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* 99, 1015–1026.

Hausman, J. & D. McFadden (1984) Specification tests for the multinomial logit model. *Econometrica* 52(5), 1219–1240.

Mora, J. & A.I. Moro-Egido (2008) On specification testing of ordered discrete choice models. *Journal of Econometrics* 143, 191–205.

Rodriguez-Poo, J.M., S. Sperlich, & P. Vieu (2004) An adaptive specification test for semiparametric models. *SSRN eLibrary*.

Wang, M.C. & J. van Ryzin (1981) A class of smooth estimators for discrete distributions. *Biometrika* 68, 301–309.

Zheng, J.X. (2000) A consistent test of conditional parametric distributions. *Econometric Theory* 16, 667–691.

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

APPENDIX A: Proof of Main Results

In the proof of Theorems 2.1 and 2.2, we will replace $\hat{h}_1,\ldots,\hat{h}_q$, $\hat{\lambda}_0,\ldots,\hat{\lambda}_r$ by their non-stochastic leading terms: $(h_1,\ldots,h_q)=(a_1^0 n^{-1/(q+4)},\ldots,a_q^0 n^{-1/(q+4)})$ and $(\lambda_0,\ldots,\lambda_r)=(b_0^0 n^{-2/(q+4)},\ldots,b_r^0 n^{-2/(q+4)})$. This will greatly simplify the arguments in the proof. By Lemma A.4 of Fan et al. (2006), we know that the conclusion holds provided $\hat{h}_s - h_s = o_p(h_s)$ $(s=1,\ldots,q)$ and $\hat{\lambda}_s - \lambda_s = o_p(\lambda_s)$ $(s=0,\ldots,r)$, which are true by Lemma 2.1.

Proof of Theorem 2.1. The test statistic is given by equation (7). We write $\hat{f}_{ij} = \hat{f}(y_i|x_j) \equiv f(y_i|x_j,\hat{\theta})$ and $\hat{f}_i = \hat{f}(y_i|x_i) \equiv f(y_i|x_i,\hat{\theta})$. Then substituting $l(y_i,y_j,\lambda_0)$ defined in (5) into equation (7), we obtain

$$
\begin{aligned}
J_{n,\gamma} &= \frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\frac{K_{\gamma,ij}}{\hat{f}_i}\left[(1-\lambda_0)I_{ij}+\frac{\lambda_0}{c_0-1}\left(1-I_{ij}\right)\right.\\
&\qquad\left.-\sum_{y\in D_y}(1-\lambda_0)\hat{f}\left(y|x_j\right)I\left(y_i=y\right)-\sum_{y\in D_y}\frac{\lambda_0}{c_0-1}\hat{f}\left(y|x_j\right)\left(1-I\left(y_i=y\right)\right)\right]\\
&= \frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\frac{K_{\gamma,ij}}{\hat{f}_i}\left[(1-\lambda_0)\left[I_{ij}-\hat{f}_{ij}\right]+\frac{\lambda_0}{c_0-1}\left(1-I_{ij}\right)\right.\\
&\qquad\left.-\frac{\lambda_0}{c_0-1}\sum_{y\in D_y}\hat{f}\left(y|x_j\right)+\frac{\lambda_0}{c_0-1}\sum_{y\in D_y}\hat{f}\left(y|x_j\right)I\left(y_i=y\right)\right]\\
&= \left[1-\lambda_0-\frac{\lambda_0}{c_0-1}\right]\left[\frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\frac{K_{\gamma,ij}}{\hat{f}_i}\left(I_{ij}-\hat{f}_{ij}\right)\right]\\
&= \frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\frac{K_{\gamma,ij}}{\hat{f}_i}\left[I_{ij}-\hat{f}_{ij}\right]+(s.o.)\\
&= \frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\frac{K_{\gamma,ij}}{f_i}\left[I_{ij}-f_{ij}\right]+(s.o.), \qquad \textbf{(A.1)}
\end{aligned}
$$

where $(s.o.)$ denotes smaller order terms, $f_{ij}=f(y_i|x_j,\theta_0)$ and $f_i=f(y_i|x_i,\theta_0)$. The fourth equality follows because $\lambda_0 = O(n^{-2/(4+q)}) = o(1)$, while the last equality follows from a Taylor expansion argument (see Fan et al., 2006, for detailed derivations).

We see that the leading term of our statistic given in equation (A.1) is the same as the leading term of Fan et al.'s (2006) test statistic. Hence, Theorem 2.1 follows from Theorem 2.1 of Fan et al. with $\hat{V}_a = \frac{2(h_1\ldots h_q)}{n(n-1)}\sum_i\sum_{j\neq i}\{K_{\hat{\gamma},ij}[I(y_i=y_j)-f(y_i|x_j,\hat{\theta})]/\hat{f}(y_i|x_i,\hat{\theta})\}^2$, which is a consistent estimator of $V_a = 2[\int W^2(v)dv]\mathbb{E}[(1-f(y_i|x_i,\theta_0))f^{-1}(y_i|x_i,\theta_0)p_1(x_i)]$ (the asymptotic variance of $n(\hat{h}_1\ldots\hat{h}_q)^{1/2}J_{n,\hat{\gamma}}$). This completes the proof of Theorem 2.1. ■

Proof of Theorem 2.2. When y is an ordered discrete variable, we use the Wang and van Ryzin (1981) kernel defined in equation (6). Substituting the ordered kernel function into equation (7), we have

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

$$J_{n,\gamma} = \frac{1-\lambda_0}{n(n-1)} \sum_i \sum_{j\neq i} \frac{K_{\gamma,ij}}{\hat{f}_i} \left\{ I_{ij} + \frac{1}{2} I\left(y_i \neq y_j\right) \lambda_0^{|y_i - y_j|} \right.$$
$$\left. - \sum_{y\in D_y} \hat{f}\left(y|x_j\right) \left[I\left(y_i = y\right) + \frac{1}{2} I\left(y_i \neq y\right) \lambda_0^{|y_i - y|} \right] \right\}$$
$$= J_{n1} - \lambda_0 J_{n1},$$

where $J_{n1} = \frac{1}{n(n-1)} \sum_i \sum_{j\neq i} \frac{K_{\gamma,ij}}{\hat{f}_i} \{ I_{ij} + \frac{1}{2} I(y_i \neq y_j) \lambda_0^{|y_i - y_j|} - \sum_{y\in D_y} \hat{f}(y|x_j) [I(y_i = y) + \frac{1}{2} I(y \neq y_i) \lambda_0^{|y_i - y|}] \}$. The leading term of $J_{n,\gamma}$ is J_{n1}, because $\lambda_0 = o(1)$. Therefore, we have

$$J_{n1} = \frac{1}{n(n-1)} \sum_i \sum_{j\neq i} \frac{K_{\gamma,ij}}{f_i} \left[I_{ij} - f_{ij} \right]$$
$$+ \frac{1}{2n(n-1)} \sum_i \sum_{j\neq i} \frac{K_{\gamma,ij}}{f_i} \left\{ I\left(y_i \neq y_j\right) \lambda_0^{|y_i - y_j|} - \sum_{y\in D_y} f\left(y|x_j\right) I\left(y \neq y_i\right) \lambda_0^{|y_i - y|} \right\}$$
$$+ (s.o.)$$
$$= J_{n1,1} + J_{n1,2} + (s.o.),$$

where the definition of $J_{n1,j}$ $(j = 1, 2)$ should be apparent, and $(s.o.)$ denotes the smaller order terms of the expansion of $1/\hat{f}_i$ and the Taylor expansion of $\hat{f}(y|x_j)$. We see that $J_{n1,1}$ is identical to the leading term in equation (A.1). Hence, $n\sqrt{h_1 \ldots h_q} J_{n1,1} \xrightarrow{d} N(0, V_a)$ by Theorem 2.1.

Now $J_{n1,2}$ can be written as a second-order U-statistic,

$$J_{n1,2} = \frac{1}{2n(n-1)} \sum_i \sum_{j\neq i} J_{n,ij} = \frac{2}{n(n-1)} \sum_i \sum_{j>i} H_{n,ij}, \tag{A.2}$$

where $J_{n,ij} = \frac{K_{\gamma,ij}}{f_i} [I(y_i \neq y_j) \lambda_0^{|y_i - y_j|} - \sum_{y\in D_y} \lambda_0^{|y_i - y|} f(y|x_j) I(y \neq y_i)]$ and $H_{n,ij} = [J_{n,ij} + J_{n,ji}]/4$.

We will use the U-statistic H-decomposition to analyze $J_{n1,2}$. We first evaluate $\mathbb{E}[H_{n,ij}|z_i] = (1/4)\mathbb{E}[J_{n,ij} + J_{n,ji}|z_i]$, where below we use the shorthand notation $\int [\ldots] dx_j = \sum_{x_j^d} \int [\ldots] dx_j^c$.

$$\mathbb{E}[J_{n,ij}|z_i] = \frac{1}{f_i} \sum_{y\in D_y} \int p(y, x_j) K_{\gamma,ij} \lambda_0^{|y_i - y|} I(y_i \neq y) dx_j$$
$$- \frac{1}{f_i} \sum_{y\in D_y} \int p_1(x_j) f(y|x_j) K_{\gamma,ij} \lambda_0^{|y_i - y|} I(y \neq y_i) dx_j$$
$$= 0,$$

because $p_1(x_j) f(y|x_j) = p(y, x_j)$.

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

Next, for $\mathbb{E}[J_{n,ji}|z_i]$, we have

$$\begin{aligned}\mathbb{E}\left[J_{n,ji}|z_i\right] &= \sum_{y\in D_y}\int p\left(y,x_j\right)\frac{K_{\gamma,ij}}{f\left(y|x_j,\theta_0\right)}\lambda_0^{|y_i-y|}I\left(y_i\neq y\right)dx_j \\ &\quad - \sum_{u\in D_y}\int p\left(u,x_j\right)\frac{K_{\gamma,ij}}{f\left(u|x_j,\theta_0\right)}\sum_{y\in D_y}\lambda_0^{|u-y|}f\left(u|x_i,\theta_0\right)I\left(y\neq u\right)dx_j \\ &= A_{1n}-A_{2n}.\end{aligned}$$

Define $m(y_i)=\sum_{y\in D_y}\lambda_0^{|y_i-y|}I(y_i\neq y)$. Recalling that $f_j=f(y_j|x_j)$ and noticing that $p(u,x_j)/f(u|x_j,\theta_0)=p_1(x_j)$, we have $A_{1n}=m(y_i)\int p_1(x_j)K_{\gamma,ij}dx_j=m(y_i)\mathbb{E}[K_{\gamma,ij}|x_i]$ and $A_{2n}=\sum_{y\in D_y}[f(y|x_i)m(y)]\int p_1(x_j)K_{\gamma,ij}dx_j=\mathbb{E}[m(y_i)|x_i]\mathbb{E}[K_{\gamma,ij}|x_i]$.

Therefore, we have shown that $\mathbb{E}[J_{n,ji}|z_i]=\epsilon_i\mathbb{E}[K_{\gamma,ij}|x_i]$, where $\epsilon_i=m(y_i)-\mathbb{E}[m(y_i)|x_i]$. Note that $m(y_i)=\sum_{y\in D_y}\lambda_0^{|y_i-y|}I(y_i\neq y)=\lambda_0\sum_{y\in D_y}\lambda_0^{|y-y_i|-1}I(|y-y_i|\geq 1)=\lambda_0\xi(y_i)$, where $\xi(y_i)=\sum_{y\in D_y}\lambda_0^{|y-y_i|-1}I(|y-y_i|\geq 1)$. Hence, $\epsilon_i=m(y_i)-\mathbb{E}[m(y_i)|x_i]=\lambda_0 v_i$, where $v_i=\xi(y_i)-\mathbb{E}[\xi(y_i)|x_i]$. Hence, we have shown that

$$\mathbb{E}[J_{n,ji}|z_i]=\lambda_0 v_i\mathbb{E}[K_{\gamma,ij}|x_i].$$

Now $\mathbb{E}[v_i|x_i]=0$ implies that $\mathbb{E}[J_{n,ji}]=0$. Also, $\mathbb{E}[J_{n,ij}|z_i]=0$ implies that $\mathbb{E}[J_{n,ij}]=0$. Hence, $\mathbb{E}[H_{n,ij}]=0$. Thus, by H-decomposition we have

$$J_{n1,2}=\frac{2}{n(n-1)}\sum_i\sum_{j>i}H_n\left(z_i,z_j\right)=\frac{\lambda_0}{2n}\sum_{i=1}^n v_i\mathbb{E}\left[K_{\gamma,ij}|x_i\right]+(s.o.), \tag{A.3}$$

where the partial sum term has an order $O_p\left(n^{-8-q/8+2q}\right)$, $(s.o.)=O_p\left(n^{-12-q/8+2q}\right)$.

By the Liapunov central limit theorem (CLT) we have that

$$\frac{\sqrt{n}}{\lambda_0}J_{n1,2}\xrightarrow{d}N(0,V_2), \tag{A.4}$$

where $V_2=(1/4)\mathbb{E}[v_i^2p_1^2(x_i)]+o(1)$ because $Var(v_i\mathbb{E}[K_{\gamma,ij}|x_i])=\mathbb{E}[v_i^2\mathbb{E}^2(K_{\gamma,ij}|x_i)]=\mathbb{E}[v_i^2p_1^2(x_i)]+o(1)$.

To show that $\mathrm{Cov}(J_{n1,1},J_{n1,2}^0)=0$, where $J_{n1,2}^0$ is the leading term of $J_{n1,2}$, we use the fact that $J_{n1,1}$ can be written as a degenerate U-statistic, i.e., $J_{n1,1}=\frac{2}{n(n-1)}\sum_i\sum_{j>i}H_{n,ij}^a$, where $H_{n,ij}^a=\frac{K_{\gamma,ij}}{2}\left[\frac{I_{ij}-f_{ij}}{f_i}+\frac{I_{ji}-f_{ji}}{f_j}\right]$. It is easy to verify that $\mathbb{E}[H_{n,ij}^a|z_i]=0$, $z_i=(x_i,y_i)$. Hence, $\mathrm{Cov}[H_{n,ij}^a,v_ip_1(x_i)]=\mathbb{E}\{v_ip_1(x_i)\mathbb{E}[H_{n,ij}^a|z_i]\}=0$. Therefore, $\mathrm{Cov}(J_{n1,1},J_{n1,2}^0)=0$. Now $J_{n1,1}$ and $J_{n1,2}^0$ are asymptotically independent.

Thus we have that $n\sqrt{h_1\ldots h_q}J_{n,\gamma}\xrightarrow{d}N(0,V_b)$, where $V_b=V_a+d_0^2V_2$ with $d_0=(a_1^0\ldots a_q^0)^{1/2}b_0^0=\lim_{n\to\infty}n(h_1\ldots h_q)^{1/2}\lambda_0/\sqrt{n}$ (a_s^0 and b_0^0 are defined in Lemma 2.1). A consistent estimator of V_b is $\hat{V}_b=\hat{V}_a+\hat{d}_0^2\hat{V}_2$, where $\hat{d}_0=(n\hat{h}_1\ldots\hat{h}_q)^{1/2}\hat{\lambda}_0$, $\hat{V}_2=\frac{1}{4n}\sum_{i=1}^n\hat{v}_i^2\hat{p}_1^2(x_i)$, $\hat{v}_i=[\hat{\xi}(y_i)-\sum_{y\in D_y}\hat{\xi}(y)f(y|x_i,\hat{\theta})]$, and $\hat{\xi}(y_i)=\sum_{y\in D_y}\hat{\lambda}_0^{|y_i-y|-1}I(|y-y_i|\geq 1)$. ■

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

Proof of Theorem 2.3. When y is an unordered discrete variable, we have

$$J^*_{n,\hat{\gamma}} = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} \frac{K_{\hat{\gamma},ij}}{f^*_i} \left[I^*_{ij} - f^*_{ij} \right] + (s.o.). \tag{A.5}$$

We see that the leading term of the test statistic is exactly the same as the leading term of Fan et al.'s (2006) test statistic. Hence, the proof of this case follows from the Proof of Theorem 2.3 of Fan et al., with $\hat{V}^*_a$ defined the same way as $\hat{V}_a$ except that y_i is replaced by y^*_i. ■

Proof of Theorem 2.4. When y is an ordered discrete variable, we have $J^*_{n,\hat{\gamma}} = J^*_{n1} - \hat{\lambda}_0 J^*_{n1}$, where $J^*_{n1} = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} \frac{K_{\hat{\gamma},ij}}{f^*_i} \{ I^*_{ij} + \frac{1}{2} I(y^*_i \neq y^*_j) \hat{\lambda}_0^{|y^*_i - y^*_j|} - \sum_{y \in D_y} f(y|x_j) \left[I(y^*_i = y) + \frac{1}{2} I(y \neq y^*_i) \hat{\lambda}_0^{|y^*_i - y|} \right] \}$. The leading term of $J^*_{n,\hat{\gamma}}$ is J^*_{n1}, because $\hat{\lambda}_0 = o(1)$. Therefore, we have

$$\begin{aligned} J^*_{n1} &= \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} \frac{K_{\hat{\gamma},ij}}{f^*_i} \left[I^*_{ij} - f^*_{ij} \right] \\ &\quad + \frac{1}{2n(n-1)} \sum_i \sum_{j \neq i} \frac{K_{\hat{\gamma},ij}}{f^*_i} \left\{ I\left(y^*_i \neq y^*_j\right) \hat{\lambda}_0^{|y^*_i - y^*_j|} - \sum_{y \in D_y} f\left(y|x_j\right) I\left(y \neq y^*_i\right) \hat{\lambda}_0^{|y^*_i - y|} \right\} \\ &\quad + (s.o.) \\ &= J^*_{n1,1} + J^*_{n1,2} + (s.o.), \end{aligned}$$

where the definition of $J^*_{n1,j}$ $(j = 1, 2)$ should be apparent, and $(s.o.)$ denotes the smaller order terms. We see $J^*_{n1,1}$ is identical to the leading term in (A.5). Hence, $n\sqrt{\hat{h}_1 \ldots \hat{h}_q} J^*_{n1,1} / \hat{V}^*_a \xrightarrow{d} N(0,1)$ by Theorem 2.3.

$J^*_{n1,2}$ can be written as a second-order U-statistic

$$J^*_{n1,2} = \frac{1}{2n(n-1)} \sum_i \sum_{j \neq i} J^*_{n,ij} = \frac{2}{n(n-1)} \sum_i \sum_{j > i} H^*_{n,ij},$$

where $J^*_{n,ij} = \frac{K_{\hat{\gamma},ij}}{f^*_i} [I(y^*_i \neq y^*_j) \hat{\lambda}_0^{|y^*_i - y^*_j|} - \sum_{y \in D_y} \hat{\lambda}_0^{|y^*_i - y|} f(y|x_j) I(y \neq y^*_i)]$ and $H^*_{n,ij} = [J^*_{n,ij} + J^*_{n,ji}]/4$.

We will use the U-statistic H-decomposition to analyze $J^*_{n1,2}$. We first evaluate $\mathbb{E}^*[H^*_{n,ij}|z^*_i] = (1/4)\mathbb{E}^*[J^*_{n,ij} + J^*_{n,ji}|z^*_i]$.

$$\begin{aligned} \mathbb{E}^*\left[J^*_{n,ij} \left| z^*_i \right. \right] &= \frac{K_{\hat{\gamma},ij}}{f^*_i} \sum_{y \in D_y} I\left(y^*_i \neq y\right) \hat{\lambda}_0^{|y^*_i - y|} f\left(y|x_j\right) \\ &\quad - \frac{K_{\hat{\gamma},ij}}{f^*_i} \sum_{y \in D_y} I\left(y \neq y^*_i\right) \hat{\lambda}_0^{|y^*_i - y|} f\left(y|x_j\right) \\ &= 0 \end{aligned}$$

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press

Next, for $\mathbb{E}[J^*_{n,ji}|z^*_i]$, we have

$$\begin{aligned}\mathbb{E}^*\left[J^*_{n,ji}\,|z^*_i\right] &= K_{\hat{\gamma},ij}\mathbb{E}^*\left[\frac{1}{f^*_j}I\left(y^*_i \neq y^*_j\right)\hat{\lambda}_0^{|y^*_i-y^*_j|}|z^*_i\right] \\ &\quad - K_{\hat{\gamma},ij}\mathbb{E}^*\left[\frac{1}{f^*_j}\sum_{y\in D_y}\hat{\lambda}_0^{|y^*_j-y|}f\left(y|x_j\right)I\left(y \neq y^*_j\right)|z^*_i\right] \\ &= A^*_{1n} - A^*_{2n}.\end{aligned}$$

We have $A^*_{1n} = K_{\hat{\gamma},ij}\sum_{y\in D_y} I(y^*_i \neq y)\hat{\lambda}_0^{|y^*_i-y|} = m(y^*_i)K_{\hat{\gamma},ij}$ and $A^*_{2n} = \mathbb{E}^*[m(y^*_i)]$ $K_{\hat{\gamma},ij}$. Therefore, we have shown that $\mathbb{E}^*[J^*_n(z^*_j,z^*_i)|z^*_i] = A^*_{1n} - A^*_{2n} = \epsilon^*_i K_{\hat{\gamma},ij}$, where $\epsilon^*_i = m(y^*_i) - \mathbb{E}^*[m(y^*_i)]$. Note that $m(y^*_i) = \sum_{y^*}\hat{\lambda}_0^{|y^*_i-y^*|}I(y^*_i \neq y^*) = \hat{\lambda}_0\sum_{y^*}$ $\hat{\lambda}_0^{|y^*-y^*_i|-1}I(|y^*-y^*_i| \geq 1) = \hat{\lambda}_0\xi(y^*_i)$, where $\xi(y^*_i) = \sum_{y^*}\hat{\lambda}_0^{|y^*-y^*_i|-1}I(|y^*-y^*_i| \geq 1)$. Hence, $\epsilon^*_i = m(y^*_i) - \mathbb{E}^*[m(y^*_i)] = \hat{\lambda}_0 v^*_i$, where $v^*_i = \xi(y^*_i) - \mathbb{E}^*[\xi(y^*_i)]$. Hence, we have shown that

$$\mathbb{E}^*\left[J^*_{n,ji}|z^*_i\right] = \hat{\lambda}_0 v^*_i K_{\hat{\gamma},ij}.$$

Here, $\mathbb{E}^*[v^*_i] = 0$ implies that $\mathbb{E}^*[J^*_{n,ji}] = 0$. Also, $\mathbb{E}^*[J^*_{n,ij}|z^*_i] = 0$ implies that $\mathbb{E}^*[J^*_{n,ij}] = 0$. Hence, $\mathbb{E}^*[H^*_{n,ij}] = 0$. Thus, by H-decomposition we have

$$J^*_{n1,2} = \frac{2}{n(n-1)}\sum_i\sum_{j>i}H^*_n\left(z^*_i,z^*_j\right) = \frac{\hat{\lambda}_0}{2n}\sum_{i=1}^n v^*_i K_{\hat{\gamma},ij} + (s.o.).$$

By the Liapunov CLT we have that

$$\frac{\sqrt{n}}{\hat{\lambda}_0}J^*_{n1,2}\Big/\sqrt{\hat{V}^*_2} \xrightarrow{d} N(0,1),$$

where $\hat{V}^*_2$ is defined the same way as $\hat{V}_2$ except that y_i is replaced by y^*_i. Following a procedure similar to that we used in the proof of Theorem 2.2, we have that $n\sqrt{\hat{h}_1\ldots\hat{h}_q}J^*_{n,\gamma}\Big/\sqrt{\hat{V}^*_b} \xrightarrow{d} N(0,1)$, where $\hat{V}^*_b = \hat{V}^*_a + \hat{d}_0^2\hat{V}^*_2$, $\hat{V}^*_a$ is defined in Theorem 2.3, $\hat{d}_0 = (n\hat{h}_1\ldots\hat{h}_q)^{1/2}\hat{\lambda}_0$ (which is defined in the proof of Theorem 2.2). ■

https://doi.org/10.1017/S0266466612000382 Published online by Cambridge University Press