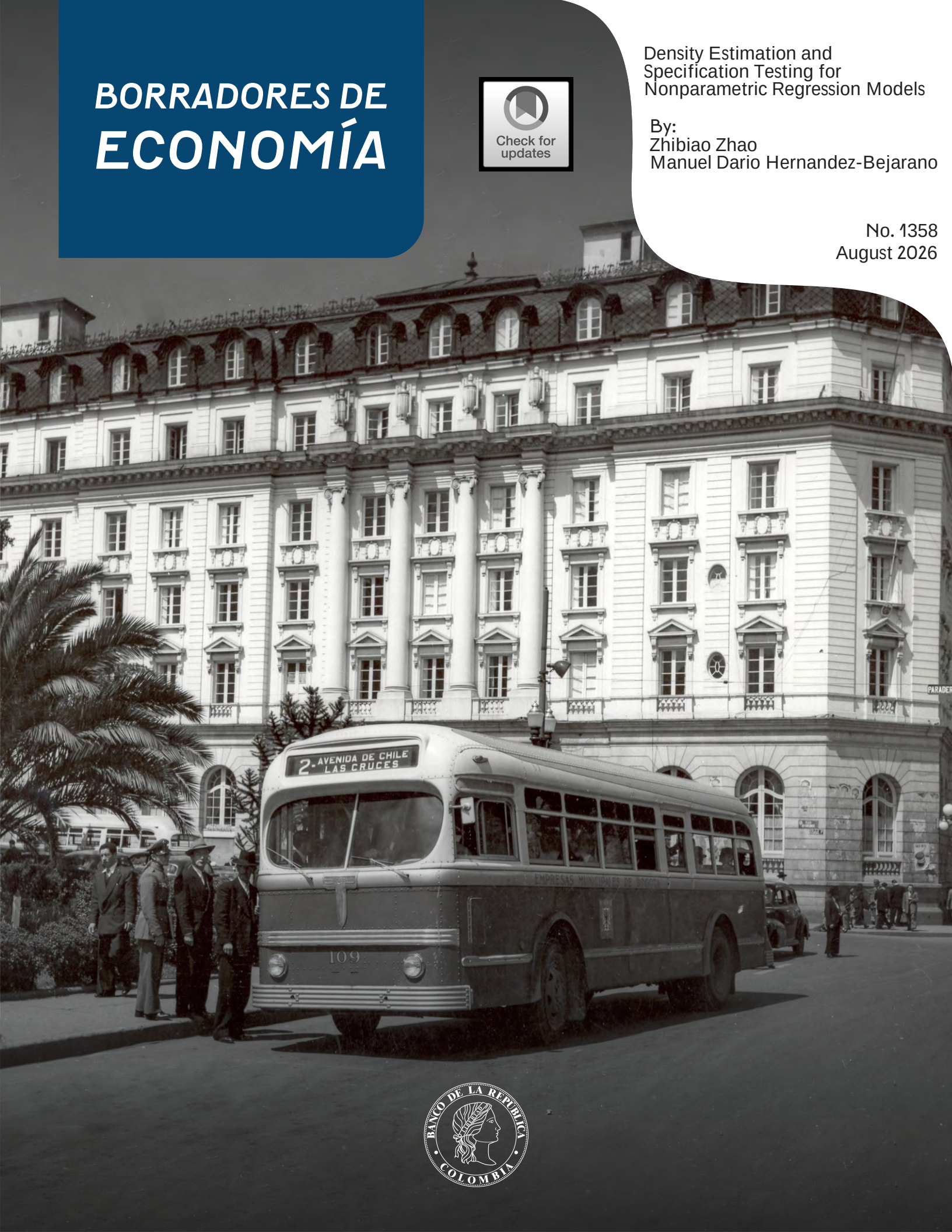
BORRADORES DE ECONOMÍA

Density Estimation and Specification Testing for Nonparametric Regression Models

By:
Zhibiao Zhao
Manuel Dario Hernandez-Bejarano

No. 1358
August 2026



BANCO DE LA REPÚBLICA
COLOMBIA

Density Estimation and Specification Testing for Nonparametric Regression Models

The opinions contained in this document are the sole responsibility of the authors and do not commit Banco de la República nor its Board of Directors.

Zhibiao Zhao*†
Manuel Dario Hernandez-Bejarano‡

Abstract

Motivated by the slow convergence rate of the classical nonparametric kernel density estimator, we study more efficient density and density derivative estimations for the marginal density of nonparametric regression models. In the presence of an unknown nonparametric regression function, the proposed density and density derivative estimators can achieve a parametric convergence rate, $\sqrt{n}$, and possess several appealing properties that the classical estimator lacks. In the absence of nonparametric regression function, in the normal case the proposed method performs as well as if we have known the model and estimated the density using maximum likelihood method. Based on the new density estimator, we further propose a more powerful density-based specification test for the nonparametric regression function. Extensive numerical studies show that the proposed density estimator, density derivative estimator, and specification test significantly outperform existing ones.

JEL Codes: C12; C14.

Keywords: Density estimation; Functional convergence; Nonparametric kernel density estimator; Nonparametric regression; Specification testing.

*Corresponding author. E-mail: zuz13@psu.edu
†Department of Statistics, The Pennsylvania State University, University Park, USA.
‡Research and Econometrics Unit, Banco de la República de Colombia. E-mail: mhernabe@banrep.gov.co

Estimación de densidad y prueba de especificación para modelos de regresión no paramétricos

Zhibiao Zhao*†,
Manuel Dario Hernandez-Bejarano‡

Resumen

Motivados por la lenta convergencia del estimador clásico de densidad de kernel no paramétrico, en este documento estudiamos estimaciones más eficientes de la densidad y de su derivada para la densidad marginal en modelos de regresión no paramétricos. En presencia de una función de regresión no paramétrica desconocida, los estimadores de densidad y su derivada propuestos alcanzan una tasa de convergencia paramétrica, $\sqrt{n}$, y poseen varias propiedades atractivas de las que carece el estimador clásico. En ausencia de una función de regresión no paramétrica, en el caso normal, el método propuesto se desempeña tan bien como si se conociera el modelo y se estimara la densidad mediante el método de máxima verosimilitud. Con base en el nuevo estimador de densidad, se propone adicionalmente una prueba de especificación más potente basada en la densidad para la función de regresión no paramétrica. Estudios numéricos exhaustivos demuestran que el estimador de densidad propuesto, el estimador de la derivada de la densidad y la prueba de especificación superan significativamente a los métodos existentes.

Códigos JEL: C12; C14.

Palabras Clave: Estimación de densidad; Convergencia funcional; Estimador de densidad kernel no paramétrico; Regresión no paramétrica; Prueba de especificación.

*Autor correspondiente. E-mail: zuz13@psu.edu
†Departamento de Estadística, Universidad Estatal de Pensilvania, University Park, EE. UU.
‡Unidad de Investigaciones y Econometría, Banco de la República de Colombia. E-mail: mhernabe@banrep.gov.co

1 Introduction

Consider observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the nonparametric regression model

$$Y = \mu(X) + e \quad \text{with} \quad \mathbb{E}(e) = 0, \tag{1}$$

where $\mu(\cdot)$ is an unknown function, and e is independent of X. While linear models, or more generally nonlinear parametric models, enjoy better interpretability, it is often difficult to know in advance such parametric forms, and there could always be a potential risk of mis-specification using, for example, a linear model to describe a highly nonlinear phenomenon. In contrast to parametric models, the nonparametric model (1) can let the data speak for themselves by imposing no specific structure on $\mu(\cdot)$. Thanks to this appealing feature, the nonparametric regression model (1) has extensive applications in economics, finance, and statistics [Fan and Gijbels (1996); Li and Racine (2006)].

The density function plays an important role in statistical analysis as it contains all distributional information about the data. Denote by $f_Y(y)$ the density of Y. Given $Y_1, \ldots, Y_n$, the classical nonparametric kernel density estimator [see Silverman (1986)] of $f_Y(y)$ is

$$\tilde{f}_Y(y) = \frac{1}{nb_n} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{b_n}\right), \tag{2}$$

for kernel function $K(\cdot)$ and bandwidth $b_n \to 0$. Under appropriate conditions, $\tilde{f}_Y(y)$ is $\sqrt{nb_n}$-consistent, and the slow convergence rate can be problematic for small sample sizes.

To overcome the aforementioned slow convergence issue, $\sqrt{n}$-consistent density estimations have been studied in the literature under two directions. One direction is $\sqrt{n}$-consistent marginal density estimation for some parametric models, including moving average process [Saavedra and Cao (1999, 2000); Schick and Wefelmeyer (2004)], invertible linear process [Schick and Wefelmeyer (2007)], nonlinear autoregressive conditional heteroscedastic model with parametric noise [Zhao (2010)], and nonlinear autoregressive model with nonparametric noise [Kim et al. (2015)]. The second direction is $\sqrt{n}$-consistent estimation for certain special integral functionals of the marginal density and/or its derivatives [Hall and Marron (1987); Bickel and Ritov (1988); Ritov and Bickel (1990); Frees (1994); Birgé and Massart (1995); Kerkyacharian and Picard (1996); Giné and Mason (2007)]. Both directions rely on either the parametric form of the model itself or an explicitly known form of the functional of the marginal density, such as integrated squared density and convolution of two or more densities, with the density being the only unknown quantity. However, such parametric assumptions can hardly be justified in many applications, and there could be a risk of mis-specification.

Our first goal is to construct $\sqrt{n}$-consistent density estimation for the nonparametric regression model (1). Instead of imposing parametric form on $\mu(\cdot)$ as in the existing literature, we impose a parametric distribution on noise e while leaving $\mu(\cdot)$ completely unspecified. Specifically, we assume

Assumption 1. In (1), $e \sim f_e(\cdot|\theta)$ for a density $f_e(\cdot|\theta)$ with unknown parameter θ.

Generally speaking, Assumption 1 is not as restrictive as imposing parametric model assumptions on $\mu(\cdot)$. In mathematical finance, for instance, for discretely sampled data from stochastic diffusion models driven by Brownian motion, the noise is normally distributed. Furthermore, many practical data or their proper transformations satisfy (approximate) normality or, in general, the Student t_d distribution with an unknown d degrees of freedom. By contrast, it can be more difficult or even unrealistic to know in advance the parametric form of $\mu(\cdot)$, especially when it is actually the goal of the study to examine the relation between variables Y and X and when such relation exhibits a highly nonlinear pattern; see Models 2–4 in Section 3. Therefore, our nonparametric-regression-parametric-noise approach provides a more realistic and flexible framework than those parametric-model-nonparametric-noise approaches in the literature.

In (1), under Assumption 1, conditioning on X, Y has density $f_e(y - \mu(X)|\theta)$. Thus,

$$f_Y(y) = \mathbb{E} f_e(y - \mu(X)|\theta). \tag{3}$$

Denote by $(\hat{\mu}(\cdot), \hat{\theta})$ some estimates of $(\mu(\cdot), \theta)$. By (3), we propose estimating $f_Y(y)$ by

$$\hat{f}_Y(y) = \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} f_e(y - \hat{\mu}(X_i)|\hat{\theta}), \quad |\mathcal{N}| \text{ is the cardinality of set } \mathcal{N}. \tag{4}$$

Here $\mathcal{N}$, to be specified later, is a subset of $\{1, 2, \ldots, n\}$ to avoid boundary issues in nonparametric regression. We establish the interesting result that $\hat{f}_Y(y)$ can achieve the parametric convergence rate $\sqrt{n}$, despite the nonparametric nature of the unknown function $\mu(\cdot)$ so that $\hat{\mu}(\cdot)$ is obtained through a nonparametric regression method. In addition to the faster convergence rate over the nonparametric kernel density estimator $\tilde{f}_Y(y)$ in (2), the new estimator $\hat{f}_Y(y)$ has other appealing properties, such as being uniformly bounded, having uniformly bounded derivatives, and possessing smoothness properties which the classical estimator $\tilde{f}_Y(y)$ lacks.

The second goal is to construct $\sqrt{n}$-consistent estimator for density derivatives $f_Y^{(k)}(y) = \partial^k f_Y(y)/\partial y^k, k = 1, 2, \ldots$. The traditional nonparametric kernel density derivative estimator is to take k-th order derivative in (2), yielding

$$\tilde{f}_Y^{(k)}(y) = \frac{1}{n b_n^{k+1}} \sum_{i=1}^{n} K^{(k)}\left(\frac{y - Y_i}{b_n}\right), \quad \text{where} \quad K^{(k)}(u) = \frac{\partial^k K(u)}{du^k}. \tag{5}$$

Then $\tilde{f}_Y^{(k)}(y)$ has convergence rate $b_n^k \sqrt{n b_n}$ which deteriorates quickly as k increases, causing even more serious issue than $\tilde{f}_Y(y)$ in small to medium sample size settings. To address this, we propose estimating $f_Y^{(k)}(y)$ by taking k-th order derivative in (4) instead, that is,

$$\hat{f}_Y^{(k)}(y) = \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} f_e^{(k)}(y - \hat{\mu}(X_i)|\hat{\theta}) \quad \text{with} \quad f_e^{(k)}(z|\theta) = \frac{\partial f_e(z|\theta)}{\partial z^k}. \tag{6}$$

Similar to the case of $\hat{f}_Y(y)$, we establish $\sqrt{n}$ consistency for $\hat{f}_Y^{(k)}(y)$. As demonstrated by extensive simulation studies in Section 3, while $\tilde{f}_Y^{(k)}(y)$ in (5) can be quite unstable across realizations, $\hat{f}_Y^{(k)}(y)$ in (6) performs well and can reduce the mean squared error by 95%.

In addition to providing distributional information about the data, the density function has also been used in specification testing in the literature. For model (1), consider testing

$$H_0: \text{ in (1), } \mu(x) = \mu_\beta(x), \quad \text{for all } x. \tag{7}$$

Here, $\mu_\beta(x)$ is a given parametric model with unknown parameter β. In the literature [e.g., Aït-Sahalia (1996); Gao and King (2004); Hong and Li (2005); Aït-Sahalia et al. (2009); Zhao (2011); Kim et al. (2015)], density-based specification test measures the distance between two estimates of $f_Y(y)$: the nonparametric version $\tilde{f}_Y$ in (2), which is $\sqrt{nb_n}$-consistent regardless whether or not H_0 holds, and some parametric version (under H_0) which is usually $\sqrt{n}$-consistent under H_0. For example, using integrated-squared-distance (maximal deviation being another popular distance), the test is

$$D_n = \int_{\mathcal{Y}} [\tilde{f}_Y(y) - \check{f}_Y(y)]^2 dy, \quad \text{with } \tilde{f}_Y \text{ in (2) and a parametric estimate } \check{f}_Y. \tag{8}$$

Here, the construction of parametric estimate $\check{f}_Y$ is given in (23) below and $\mathcal{Y}$ is some bounded interval. Since $\check{f}_Y$ usually has a faster convergence rate $\sqrt{n}$ than the nonparametric estimate $\tilde{f}_Y$, the random variation in D_n is determined by the relatively poorer estimator $\tilde{f}_Y$, leading to poor performance of D_n. Pritsker (1998) pointed out that such tests [e.g., Aït-Sahalia (1996)] often require an extremely large sample size.

Our third goal is to propose more efficient density-based specification test for (7). In (8), we propose replacing the $\sqrt{nb_n}$-consistent estimator $\tilde{f}_Y$ by our newly constructed $\sqrt{n}$-consistent density estimate $\hat{f}_Y(y)$ in (4) to form the new test

$$T_n = \int_{\mathcal{Y}} [\hat{f}_Y(y) - \check{f}_Y(y)]^2 dy, \quad \text{with } \hat{f}_Y \text{ in (4) and a parametric estimate } \check{f}_Y. \tag{9}$$

Due to the same $\sqrt{n}$-consistency of $\hat{f}_Y(y)$ (for nonparametric $\mu(\cdot)$) and $\check{f}_Y(y)$ (under H_0), it is expected that T_n converges at rate n under H_0. Simulation studies in Section 3 show that $\hat{f}_Y(y)$-based test T_n performs well whereas $\tilde{f}_Y(y)$-based test D_n in (8) has little power.

The rest of the document is organized as follows. Section 2 presents large sample theory for the proposed density and density derivative estimators, the construction of simultaneous confidence bands, and the specification test. Section 3 contains numerical studies. Finally, proofs are included in the appendix A.

2 Main results

The notation $\xrightarrow{p}$ means convergence in probability. For a matrix (or vector) $A = \{a_{ij}\}$, write $|A| = \sum_{i,j} |a_{ij}|$. For two random variables U and V, $\mathrm{Cov}(U, V)$ is their covariance.

2.1 $\sqrt{n}$-rate functional CLT for $\hat{f}_Y(y)$ and $\hat{f}_Y^{(k)}(y)$

For $\hat{\mu}(\cdot)$ in (4), we use the widely used nonparametric kernel smoothing estimate

$$\hat{\mu}(x) = \frac{\sum_{i=1}^n Y_i K\{(x - X_i)/h_n\}}{\sum_{i=1}^n K\{(x - X_i)/h_n\}}, \tag{10}$$

for bandwidth $h_n > 0$ and kernel function $K(\cdot)$.

Assumption 2. (i) $K(\cdot)$ satisfies $\int_{\mathbb{R}} K(u)du = 1$, $\int_{\mathbb{R}} uK(u)du = 0$, and has bounded support $[-L, L]$ and bounded derivative. (ii) h_n satisfies $nh_n^4 \to 0$, $\sqrt{n}h_n / \log n \to \infty$. (iii) $(X_1, Y_1, e_1), \ldots, (X_n, Y_n, e_n)$ are iid from (1), $\mathbb{E}(e_i) = 0$, $\mathbb{E}(e_i^2) < \infty$, and e_i is independent of X_i. (iv) X has support on a bounded interval $\mathcal{X} := [a, A]$ for some $a < A$, and the density $f_X(x)$ of X satisfies $f_X(x) > c$ for all $x \in \mathcal{X}$ for some constant $c > 0$. (v) $\mu(\cdot)$ and $f_X(\cdot)$ are twice differentiable on $\mathcal{X}$.

Under Assumption 2, it is well known [Fan and Yao (2003); Hansen (2008)] that

$$\sup_{x \in [a + Lh_n, A - Lh_n]} |\hat{\mu}(x) - \mu(x)| = O_p\left(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\right) = o_p(n^{-1/4}). \tag{11}$$

Throughout, we take

$$\mathcal{N} = \{i : \ X_i \in [a + Lh_n, A - Lh_n]\}, \quad \text{see Remark 1 below.} \tag{12}$$

Remark 1. The introduction of $\mathcal{N}$ in (4) is due to pure technical consideration and, in our practical implementation, we simply take $\mathcal{N} = \{1, 2, \ldots, n\}$. It is also possible to extend the bounded support of X in Assumption 2 to unbounded support case by imposing proper decaying rate on $f_X(x) \to 0$ as $x \to \infty$.

For $\hat{\theta}$, since the log likelihood function for $(e_1, \ldots, e_n)$ is $\sum_{i=1}^n \log f_e(e_i|\theta)$, replacing e_i by the estimate $\hat{e}_i = Y_i - \hat{\mu}(X_i)$, we use the maximum likelihood estimate (MLE):

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \sum_{i \in \mathcal{N}} \log f_e(\hat{e}_i|\theta) \quad \text{with} \quad \hat{e}_i = Y_i - \hat{\mu}(X_i). \tag{13}$$

Remark 2. Instead of using the Nadaraya-Watson type kernel smoothing estimator $\hat{\mu}(x)$ in (10), an alternative approach is the local likelihood method. Note that, for $X_i \approx x$, $Y_i \approx \mu(x) + e_i$ has density $f_e(y - \mu(x)|\theta)$. In the first step, the local MLE of $\mu(x)$ is

$$(\tilde{\mu}(x), \tilde{\theta}) = \operatorname*{argmax}_{(\mu, \theta)} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \log f_e(Y_i - \mu|\theta). \tag{14}$$

In the second step, θ can be estimated by similar MLE in (13) with $\hat{e}_i$ therein replaced by $\tilde{e}_i = Y_i - \tilde{\mu}(X_i)$. Computationally, $\hat{\mu}(x)$ has a closed-form expression while $\tilde{\mu}(x)$ has to be solved numerically, subject to numerical stability issues. Furthermore, the consistency of $\hat{\mu}(x)$ only requires $\mathbb{E}(e_i) = 0$ and $\mathbb{E}(e_i^2) < \infty$, which are much less stringent than the exact distributional assumption for $\tilde{\mu}(x)$. Also, in the most important normality case, $\hat{\mu}(x)$ and $\tilde{\mu}(x)$ are equivalent. Due to these considerations, we use $\hat{\mu}(x)$.

Assumption 3. For the density $f_e(z|\theta)$ in Assumption 1, (i) $\theta \in \Theta$ for a compact set Θ. (ii) All first and second order partial derivatives of $f_e(z|\theta)$ with respect to (z, θ^T) are uniformly bounded in $z \in \mathbb{R}, \theta \in \Theta$. (iii) There exists $\epsilon > 0$ such that $\sup_{|z-e|\le\epsilon, \theta\in\Theta} |\partial \log f_e(z|\theta)/\partial z| \le \ell_1(e)$ for some function $\ell_1(\cdot)$ satisfying $\mathbb{E}\ell_1(e) < \infty$. (iv) There exists function $\ell_2(\cdot)$ such that $\sup_{\theta\in\Theta} f_e(z|\theta) \le \ell_2(z)$ and $\int_{\mathbb{R}} |z|\ell_2(z)dz < \infty$. (v) $f_e(z|\theta)$ satisfies the regularity conditions [see, e.g., Casella and Berger (2002)] for the asymptotic normality of MLE.

The regularity conditions in Assumption 3 are similar to those in the study of MLE. Some conditions may be relaxed, for example, the uniform boundedness in Assumption 3(ii) can be replaced by certain moment conditions. However, we shall not pursue this direction, since these weaker conditions entail lengthier technical proofs without adding meaningful insights, and, furthermore, Assumption 3 is already flexible enough to include most commonly used distributions, such as normal and Student t distributions.

Theorem 1. *For $\hat{f}_Y(y)$ in (4) with $(\hat{\mu}(\cdot), \hat{\theta})$ defined in (10) and (13), under Assumptions 1–3, for any bounded interval $\mathcal{Y}$, the functional Central Limit Theorem (CLT) holds:*

$$\Big\{\sqrt{n}[\hat{f}_Y(y) - f_Y(y)], \quad y \in \mathcal{Y}\Big\} \Rightarrow \Big\{G(y), \quad y \in \mathcal{Y}\Big\}, \tag{15}$$

for a centered Gaussian process $\{G(y)\}$ with autocovariance $\Sigma(y, y') = \mathrm{Cov}\{G(y), G(y')\}$,

$$\begin{aligned} \Sigma(y, y') &= \Sigma_0(y, y') + \Sigma_\mu(y, y') + \Sigma_\theta(y, y'), \\ \textit{where}\ \ \Sigma_0(y, y') &= \mathrm{Cov}\{f_e(y - \mu(X)|\theta), f_e(y' - \mu(X)|\theta)\}, \\ \Sigma_\mu(y, y') &= \mathbb{E}(e^2) \times \mathbb{E}[f_e'(y - \mu(X)|\theta) f_e'(y' - \mu(X)|\theta)], \\ \Sigma_\theta(y, y') &= \mathbb{E}[\dot{f}_e(y - \mu(X)|\theta)]^T \times \mathcal{I}_\theta^{-1} \times \mathbb{E}[\dot{f}_e(y' - \mu(X)|\theta)]. \end{aligned} \tag{16}$$

Here, $\dot{f}_e(z|\theta) = \partial f_e(z|\theta)/\partial\theta$, $f_e'(z|\theta) = \partial f_e(z|\theta)/\partial z$, $\mathcal{I}_\theta = \mathbb{E}[\psi(e|\theta)\psi(e|\theta)^T]$ is the Fisher information matrix with the score function $\psi(e|\theta) = \partial \log f_e(e|\theta)/\partial\theta$.

By Theorem 1, for each fixed y,

$$\sqrt{n}[\hat{f}_Y(y) - f_Y(y)] \Rightarrow N\Big(0, \Sigma_0(y, y) + \Sigma_\mu(y, y) + \Sigma_\theta(y, y)\Big). \tag{17}$$

On the other hand, if $(\mu(\cdot), \theta)$ were known, by (3), we would have estimated $f_Y(y)$ by

$$f_Y^*(y) = \frac{1}{n}\sum_{i=1}^n f_e(y - \mu(X_i)|\theta).$$

By CLT for iid data, for each fixed y,

$$\sqrt{n}[f_Y^*(y) - f_Y(y)] \Rightarrow N\Big(0, \Sigma_0(y,y)\Big). \tag{18}$$

Compared to (18), the limiting variance in (17) consists of three components: the common term $\Sigma_0(y,y)$ accounts for the unknown distribution of X, and the two extra terms $\Sigma_\mu(y,y)$ and $\Sigma_\theta(y,y)$ reflect the variations introduced by the estimation of $\mu(\cdot)$ and θ, respectively. Furthermore, by the same argument in the proof of Theorem 1, if $\mu(\cdot)$ is known, i.e., we replace $\hat{\mu}(X_i)$ by $\mu(X_i)$ in (4), then the limiting variance becomes $\Sigma_0(y,y)+\Sigma_\theta(y,y)$; if θ is known, i.e., we replace $\hat{\theta}$ by θ in (4), then the limiting variance becomes $\Sigma_0(y,y) + \Sigma_\mu(y,y)$.

In comparison to $\sqrt{nb_n}$-consistent $\tilde{f}_Y(y)$ in (2), the new estimate $\hat{f}_Y(y)$ in (4) has a faster convergence rate $\sqrt{n}$. Intuitively, $\hat{f}_Y(y)$ uses information from the model structure which, while being nonparametric in nature, can still provide information to the density through the identity (3), thus leading to more efficient density estimation. Below, we present additional appealing properties of this new density estimate.

(A) Uniform boundedness.

In statistical inference, we often require the density function be bounded. From (4), if $f_e(z|\theta)$ is uniformly bounded, then $\hat{f}_Y(y)$ has the same uniform bound. However, for $\tilde{f}_Y(y)$ in (2), even though $K(\cdot)$ is bounded, there always exist sample paths (with asymptotically vanishing probability though) such that all Y_i's are close to y so that $\tilde{f}_Y(y) = O(1/b_n)$ can be arbitrarily large as $b_n \to 0$.

(B) Smoothness of the limiting process $\{G(y)\}$ in Theorem 1.

Since $\zeta_i(y)$ in (40) (see Section A) is differentiable, $\{G(y)\}$ is differentiable. In general, if $f_e(z|\theta)$ has k-th order derivative (with respect to z), then the limiting process $\{G(y)\}$ has $(k-1)$-th order derivative. However, the nonparametric kernel density estimator $\tilde{f}_Y(y)$ in (2) exhibits completely different behavior. To see this, let

$$U_n(y) = \frac{\sqrt{nb_n}[\tilde{f}_Y(y) - f_Y(y) - b_n^2 f_Y''(y)\int_{\mathbb{R}} u^2K(u)du/2]}{\sqrt{f_Y(y)\int_{\mathbb{R}} K^2(u)du}}.$$

Under Assumption 2, it is well known that $U_n(y) \Rightarrow N(0,1)$ for each fixed y and that $U_n(y)$ and $U_n(y')$ are asymptotically independent $N(0,1)$ for $|y-y'| > 2Lb_n$. If we take $|y-y'| = 3Lb_n \to 0$, $U_n(y) - U_n(y') \Rightarrow N(0,2)$. That is, even when y and y' are arbitrarily close, the distance $U_n(y) - U_n(y')$ is a non-degenerate random variable $N(0,2)$, indicating discontinuity. Roughly speaking, this means that, for large enough n, $U_n(y)$ consists of pure jumps without any continuity, and consequently these large random variations in $\tilde{f}_Y$ causes the test D_n in (8) to be very insensitive to $\check{f}_Y$, or in other words, the test has no power distinguishing parametric specifications.

(C) Functional CLT.

The unpleasant non-smoothness of $\tilde{f}_Y(y)$ discussed in (B) above is caused by the lack of tightness, which is why $\tilde{f}_Y(y)$ only has pointwise asymptotic normality and does not admit the type of functional CLT as in Theorem 1. As will be discussed in Section 2.2, the functional CLT is useful in studying the asymptotic distribution of the specification test.

To apply Theorem 1 for statistical inference purpose (for example, confidence interval construction), we estimate the autocovariance $\Sigma(y, y')$ in Theorem 1 by

$$\begin{aligned}
\hat{\Sigma}(y, y') &= \hat{\Sigma}_0(y, y') + \hat{\Sigma}_\mu(y, y') + \hat{\Sigma}_\theta(y, y'), \quad \text{where} \qquad (19)\\
\hat{\Sigma}_0(y, y') &= \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} f_e(y - \hat{\mu}(X_i)|\hat{\theta}) f_e(y' - \hat{\mu}(X_i)|\hat{\theta}) - \hat{f}_Y(y)\hat{f}_Y(y'),\\
\hat{\Sigma}_\mu(y, y') &= \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} \hat{e}_i^2 \times \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} f'_e(y - \hat{\mu}(X_i)|\hat{\theta}) f'_e(y' - \hat{\mu}(X_i)|\hat{\theta}),\\
\hat{\Sigma}_\theta(y, y') &= \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} \dot{f}_e(y - \hat{\mu}(X_i)|\hat{\theta})^T \times \left[\sum_{i\in\mathcal{N}} \psi(\hat{e}_i|\hat{\theta})\psi(\hat{e}_i|\hat{\theta})^T\right]^{-1} \times \sum_{i\in\mathcal{N}} \dot{f}_e(y' - \hat{\mu}(X_i)|\hat{\theta}).
\end{aligned}$$

Theorem 2. *Under the conditions in Theorem 1, we have uniform consistency*

$$\sup_{y,y'\in\mathcal{Y}} |\hat{\Sigma}(y, y') - \Sigma(y, y')| \xrightarrow{p} 0.$$

For the density derivative estimator $\hat{f}_Y^{(k)}(y)$ in (6), we have

Theorem 3. *Suppose Assumptions 1–3 hold, where Assumption 3(ii) is replaced by: All first and second order partial derivatives of $f_e(z|\theta), f'_e(z|\theta), f''_e(z|\theta), \ldots, f_e^{(k)}(z|\theta)$ with respect to (z, θ^T) are uniformly bounded in $z \in \mathbb{R}, \theta \in \Theta$. For $\hat{f}_Y^{(k)}(y)$ in (6), we have*

$$\Big\{\sqrt{n}[\hat{f}_Y^{(k)}(y) - f_Y^{(k)}(y)], \quad y \in \mathcal{Y}\Big\} \Rightarrow \Big\{G^{(k)}(y), \quad y \in \mathcal{Y}\Big\}, \qquad (20)$$

where $\{G^{(k)}(y)\}$ is the k-th order derivative of the centered Gaussian process $\{G(y)\}$ in Theorem 1. Furthermore, $\{G^{(k)}(y)\}$ is again a centered Gaussian process with autocovariance $\mathrm{Cov}\{G^{(k)}(y), G^{(k)}(y')\} = \partial^{2k}\Sigma(y, y')/\partial y^k \partial y'^k$.

By Theorem 3, all the estimated derivatives $\hat{f}_Y^{(k)}(y)$ have $\sqrt{n}$ convergence rate, whereas the nonparametric kernel density derivative estimator $\tilde{f}_Y^{(k)}(y)$ in (5) has a much slower convergence rate $b_n^k\sqrt{nb_n}$. Furthermore, if the derivative $f_e^{(k)}(z|\theta)$ is bounded, then $\hat{f}_Y^{(k)}(y)$ has the same bound. However, for $\tilde{f}_Y^{(k)}(y)$ in (5), as argued in (A) above, there exist sample paths such that $\tilde{f}_Y^{(k)}(y)$ grows at rate $O(1/b_n^{k+1})$, which causes the estimated curve quite unstable. The simulation studies in Section 3 confirm this phenomenon, and $\hat{f}'_Y(y)$ can reduce the mean squared error by 95% compared to $\tilde{f}'_Y(y)$.

Bandwidth selection: For the bandwidth h_n in (10), if we let $h_n \propto n^{-\kappa}$, then Assumption 2(ii) holds for $\kappa \in (1/4, 1/2)$. On the other hand, for nonparametric mean regression function estimation, it is well known that Ruppert et al. (1995)'s automatic plug-in bandwidth selector (implemented via the `dpill` function in `R`) selects the optimal bandwidth $h_n^* \propto n^{-1/5}$. Thus, a reasonable choice of h_n is

$$h_n = n^{-\gamma} h_n^* \text{ for some } \gamma, \quad \text{where } h_n^* \text{ is automatic plug-in bandwidth (dpill in R).} \tag{21}$$

From $h_n^* \propto n^{-1/5}$, $h_n \propto n^{-(1/5+\gamma)}$. From $1/5+\gamma \in (1/4, 1/2)$, we have $\gamma \in (1/20, 3/10)$. By the simulation studies in Section 3, the proposed method is quite robust against the choice of γ, and in practice we can take $\gamma = 7/40$, the middle point of $(1/20, 3/10)$.

2.2 Specification test

The $\sqrt{n}$-consistent density estimate $\hat{f}_Y(y)$ can be used to address specification testing problem (7) through the test T_n in (9). We need to construct the parametric estimate $\check{f}_Y$ of $f_Y(y)$. Under H_0, we estimate β by the nonlinear least-squares method:

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} [Y_i - \mu_\beta(X_i)]^2. \tag{22}$$

For θ, we use the same estimate $\hat{\theta}$ in (13) where the nonparametric estimate $\hat{\mu}(\cdot)$ is used to obtain $\hat{e}_i$, due to the consideration that $\hat{\theta}$ is always $\sqrt{n}$-consistent, regardless of whether or not H_0 holds. Therefore, similar to (4), under H_0, the parametric estimate of $f_Y(y)$ is

$$\check{f}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} f_e(y - \mu_{\hat{\beta}}(X_i)|\hat{\theta}), \quad \text{with } \hat{\theta} \text{ in (13).} \tag{23}$$

The proposed test T_n is then obtained by plugging $\hat{f}_Y$ and $\check{f}_Y$ into (9).

Assumption 4. (i) $\beta \in \Omega$ for a compact set Ω. (ii) $\mathbb{P}\{\mu_{\beta_1}(X) = \mu_{\beta_2}(X)\} = 1$ implies $\beta_1 = \beta_2$. (iii) For each $\beta \in \Omega$, $\mu_\beta(x)$ is twice differentiable in $x \in \mathcal{X}$ ($\mathcal{X}$ is defined in Assumption 2). (iv) For each $x \in \mathcal{X}$, the partial derivatives $\dot{\mu}_\beta(x) = \partial \mu_\beta(x)/\partial \beta$ and $\ddot{\mu}_\beta(x) = \partial^2 \mu_\beta(x)/\partial \beta \partial \beta^T$ are continuous in $\beta \in \Omega$. (v) $\mathbb{E}[\dot{\mu}_\beta(X)\dot{\mu}_\beta(X)^T]$ is non-singular.

Theorem 4. *Suppose Assumptions 1–4 hold. Write* $\dot{\mu}_\beta(X) = \partial \mu_\beta(X)/\partial \beta, M(y) = \mathbb{E}[f_e'(y - \mu_\beta(X)|\theta)\dot{\mu}_\beta(X)^T]$, $\mathcal{J}_\beta = \mathbb{E}[\dot{\mu}_\beta(X)\dot{\mu}_\beta(X)^T]$. *For* T_n *in (9) with* $\hat{f}_Y$ *in (4) and* $\check{f}_Y$ *in (23),*

$$nT_n \Rightarrow \int_{\mathcal{Y}} Z(y)^2 dy, \quad \textit{under } H_0,$$

where $\{Z(y)\}$ *is a centered Gaussian process with autocovariance*

$$\mathrm{Cov}\{Z(y), Z(y')\} = \mathbb{E}(e^2)\Big\{\mathbb{E}[f_e'(y - \mu_\beta(X)|\theta) f_e'(y' - \mu_\beta(X)|\theta)] - M(y)\mathcal{J}_\beta^{-1} M(y')^T\Big\}.$$

Since the asymptotic distribution in Theorem 4 involves complicated unknown functions, we adopt bootstrap approach to obtain the cutoff value at significance level α:

(1) Based on the original data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and given interval $\mathcal{Y}$, obtain h_n in (21) with $\gamma = 7/40$ [see the discussion below (21)], calculate $\hat{\mu}(\cdot)$ in (10), $\hat{\theta}$ in (13), $\hat{\beta}$ in (22), and further calculate T_n in (9) with $\hat{f}_Y$ in (4) and $\check{f}_Y$ in (23).

(2) Using $\hat{\mu}(\cdot)$ and $\hat{\beta}$ in step (1) to obtain $\hat{e}_i = Y_i - \hat{\mu}(X_i)$ and $\mu_{\hat{\beta}}(X_i), i = 1, \ldots, n$.

(3) Draw n samples, denoted by $\{\hat{e}_1^b, \ldots, \hat{e}_n^b\}$, from $\{\hat{e}_1, \ldots, \hat{e}_n\}$ with replacement, and calculate the test statistic based on the bootstrap data $(X_1, Y_1^b), \ldots, (X_n, Y_n^b)$, where $Y_i^b = \mu_{\hat{\beta}}(X_i) + \hat{e}_i^b$, using the same interval $\mathcal{Y}$ and bandwidth h_n in step (1).

(4) Repeat step (3) a large number (say, M) of times to obtain M realizations of the test statistic, and reject H_0 if T_n exceeds the $(1-\alpha)$ quantile of these realizations.

2.3 The special normality case

The assumption of normality is one of the most common assumptions in statistical analysis, often satisfied in many practical data with appropriate transformations. Below, we consider the special normality case of (1):

$$Y = \mu(X) + e, \quad e \sim N(0, \theta) \text{ with unknown } \theta > 0. \tag{24}$$

Denote by $\phi(y; \mu, \theta)$ the normal density of $N(\mu, \theta)$, and denote the partial derivatives by $\dot{\phi}_\mu(y; \mu, \theta) = \partial\phi(y; \mu, \theta)/\partial\mu$ and $\dot{\phi}_\theta(y; \mu, \theta) = \partial\phi(y; \mu, \theta)/\partial\theta$. Then (4) becomes

$$\hat{f}_Y(y) = \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \phi(y; \hat{\mu}(X_i), \hat{\theta}) \quad \text{with} \quad \hat{\theta} = \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} [Y_i - \hat{\mu}(X_i)]^2, \quad \hat{\mu}(\cdot) \text{ in (10)}. \tag{25}$$

From Theorem 1, after some calculations we can obtain

Corollary 1. *For model (24), under Assumption 2, the functional CLT in Theorem 1 holds with the autocavariance $\Sigma(y, y')$ therein reducing to*

$$\begin{aligned} \Sigma(y, y') &= \mathrm{Cov}\{\phi(y; \mu(X), \theta), \phi(y'; \mu(X), \theta)\} + \theta \times \mathbb{E}[\dot{\phi}_\mu(y; \mu(X), \theta)\dot{\phi}_\mu(y'; \mu(X), \theta)] \\ &\quad + 2\theta^2 \times \mathbb{E}[\dot{\phi}_\theta(y; \mu(X), \theta)] \times \mathbb{E}[\dot{\phi}_\theta(y'; \mu(X), \theta)]. \end{aligned}$$

As in Theorem 2, a uniformly consistent estimate of $\Sigma(y, y')$ is given by

$$\begin{aligned} \hat{\Sigma}(y, y') &= \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \phi(y; \hat{\mu}(X_i), \hat{\theta})\phi(y'; \hat{\mu}(X_i), \hat{\theta}) - \hat{f}_Y(y)\hat{f}_Y(y') \\ &\quad + \hat{\theta} \times \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \dot{\phi}_\mu(y; \hat{\mu}(X_i), \hat{\theta})\phi_\mu(y'; \hat{\mu}(X_i), \hat{\theta}) \\ &\quad + 2\hat{\theta}^2 \times \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \dot{\phi}_\theta(y; \hat{\mu}(X_i), \hat{\theta}) \times \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \dot{\phi}_\theta(y'; \hat{\mu}(X_i), \hat{\theta}). \end{aligned} \tag{26}$$

According to Theorem 1 and the discussions in Section 2.1, in the presence of an unknown nonparametric mean regression function $\mu(\cdot)$ in (1), the proposed density estimator can achieve a parametric convergence rate of $\sqrt{n}$ and possesses several appealing properties. A related question is how the method performs in the absence of such nonparametric mean function. Below, we address this question specifically for the case of normality, as outlined in equation (24).

Consider the special case of (1) where $\mu(\cdot) \equiv \mu$ is actually an unknown constant μ:

$$Y = \mu + e, \quad \mu \text{ is unknown constant, } e \sim N(0, \theta) \text{ with unknown } \theta > 0. \tag{27}$$

However, even without knowing that $\mu(\cdot) \equiv \mu$ is indeed a constant, we still use $\hat{f}_Y(y)$ in (25) to estimate $f_Y(y)$. Theorem 5 below establishes that $\hat{f}_Y(y)$ is asymptotically efficient in the sense that it is asymptotically equivalent to the MLE that would be obtained if we knew that $\mu(\cdot) \equiv \mu$ is a constant. This means that in the absence of a nonparametric mean function, the proposed method will not lose any efficiency by fitting nonparametric regression.

Theorem 5. *For the constant mean model (27), denote by* $\hat{f}_Y^{\text{MLE}}(y) = \phi(y; \overline{\mu}, \overline{\theta})$ *the MLE of* $f_Y(y)$*, where* $\overline{\mu} = n^{-1}\sum_{i=1}^n Y_i$ *and* $\overline{\theta} = n^{-1}\sum_{i=1}^n (Y_i - \overline{\mu})^2$ *are the MLE of* (μ, θ)*. For* $\hat{f}_Y(y)$ *in (25), under the conditions in Theorem 1, for any bounded interval* $\mathcal{Y}$*, we have*

$$\sup_{y \in \mathcal{Y}} |\hat{f}_Y(y) - \hat{f}_Y^{\text{MLE}}(y)| = o_p(n^{-1/2}).$$

3 Numerical studies

In this Section, we will investigate the performance of our proposed density estimate. We have the following four main objectives:

(i) Comparing the new density estimate $\hat{f}_Y(y)$ in (4) to the nonparametric kernel density estimator $\tilde{f}_Y(y)$ in (2) which is implemented using `R` function `density` with the default rule-of-thumb bandwidth choice for b_n [Silverman (1986)].

(ii) Comparing the new density derivative estimate $\hat{f}_Y^{(k)}(y)$ in (6) to the nonparametric kernel density derivative estimator $\tilde{f}_Y^{(k)}(y)$ in (5) which is implemented by `dkde` function in `R` package `kedd` with the default unbiased cross-validation choice for b_n.

(iii) Examining the effect of sample size n, bandwidth h_n, and different models on the performance of the proposed density and density derivative estimates.

(iv) Examining the performance of the specification test T_n in (9).

Consider $(X_1, Y_1), \ldots, (X_n, Y_n)$ from four increasingly more complicated models

$$\begin{aligned}
&\text{Model 1:} \quad Y = 0.3 + 0.4X + e;\\
&\text{Model 2:} \quad Y = 0.3 + 0.4\sin(2\pi X) + e;\\
&\text{Model 3:} \quad Y = 0.3 + 0.4X + 0.4\sin(2\pi X) + e;\\
&\text{Model 4:} \quad Y = 0.3 + 0.4X + 0.4\sin(2\pi X) + 0.5\sqrt{1+X^2} + e.
\end{aligned}$$

For X, we use $X \sim \text{Unif}(0,1)$. For noise e, we consider two most widely used distributions:

(i) $e \sim N(0, 0.6^2)$;
(ii) $e \sim 0.6t_4$, student t with 4 degrees of freedom and scale parameter 0.6.

The second distribution is a heavy-tailed distribution with infinite fourth moment and infinite kurtosis. The true density $f_Y(y)$ of Y is computed by numerically evaluating (3).

Comparison of density estimations:

To investigate the impact of the bandwidth h_n in equation (10), we consider four values of $\gamma \in (1/20, 3/10)$: $2/20, 3/20, 4/20, 5/20$ in (21), as discussed further below. Additionally, we use four different sample sizes: $n = 50, 100, 200, 400$, to represent a range from small to medium sample sizes. For $\tilde{f}_Y(y)$, we calculate its mean squared error (MSE) as follows:

$$\text{MSE}(\tilde{f}_Y) = \text{average of 1000 realizations of } \left\{ \frac{q_Y(0.975) - q_Y(0.025)}{100} \sum_{j=1}^{100} [\tilde{f}_Y(y_j) - f_Y(y_j)]^2 \right\},$$

where $q_Y(0.025)$ and $q_Y(0.975)$ represent the 2.5% and 97.5% quantile of Y, respectively, and $y_1, \ldots, y_{100}$ are equally spaced points on $[q_Y(0.025), q_Y(0.975)]$. The MSE for $\hat{f}_Y(y)$ is defined similarly. For ease of comparison, we also compute the relative MSE (RMSE) of $\hat{f}_Y$ as $\text{MSE}(\hat{f}_Y)/\text{MSE}(\tilde{f}_Y)$, with RMSE ≤ 1 indicating better performance of $\hat{f}_Y(y)$.

MSE and RMSE of $\hat{f}_Y$ are presented in Table 1. We make the following observations:

(1^0) As sample size increases, MSE decreases quickly, indicating increasingly better performance.

(2^0) The proposed density estimate $\hat{f}_Y(y)$ achieves substantial reductions in MSE with RMSE below 70% for most cases. For example, with $n = 100$, the MSE reduction is approximately 50% for a normal distribution and 35% for a student t distribution.

(3^0) The larger the sample size, the smaller the RMSE, and the greater the relative advantage of $\hat{f}_Y(y)$ over $\tilde{f}_Y(y)$. Both $\hat{f}_Y(y)$ and $\tilde{f}_Y(y)$ have bias terms which, while being theoretically negligible, may have some effect in finite samples, and these bias terms decrease as sample size increases. Since MSE = Bias^2 + Variance, as bias diminishes, the variance plays a dominating role, leading to

increasingly better relative performance of $\sqrt{n}$-consistent $\hat{f}_Y(y)$ than $\sqrt{nb_n}$-consistent $\tilde{f}_Y(y)$, where $b_n \propto n^{-1/5}$ decreases with sample size.

(4^0) The performance of $\hat{f}_Y(y)$ is consistent across the choices of bandwidth h_n through the choices of γ in (21). Thus, in practice, a reasonable and safe choice of γ is 7/40, the middle point of $(1/20, 3/10)$.

(5^0) The relative advantage of $\hat{f}_Y(y)$ compared to $\tilde{f}_Y(y)$ is consistent across all four models, which span from a simple linear model to a highly nonlinear model. Generally, this relative advantage is less pronounced in models that incorporate student t_4 noise, which is to be expected given that this distribution features a very heavy tail and possesses an infinite fourth moment.

Comparison of density derivative estimations:

Table 2 displays the MSE of the new density derivative estimate $\hat{f}'_Y(y)$ in (6) and its RMSE to the nonparametric kernel density derivative estimator $\tilde{f}'_Y(y)$ in (5). Compared to the density estimation case, the reduction in MSE of density derivative estimation is even more striking and exceeds 90% (i.e., RMSE $\leq 10\%$) in all cases considered. Since $\hat{f}'_Y(y)$ is $\sqrt{n}$-consistent while $\tilde{f}'_Y(y)$ is only $b_n\sqrt{nb_n}$-consistent, the larger convergence rate ratio $b_n^{-1.5}$ explains the significantly larger reductions in MSE compared to the density estimation case where the convergence rate ratio is $b_n^{-0.5}$.

From the discussion below Theorem 3, the nonparametric kernel density derivative estimator $\tilde{f}'_Y(y)$ in (5) can be quite unstable. To illustrate this, Figure 1 plots two realizations of density estimations (left panel) and density derivative estimations (right panel) from Model 2 with $e \sim 0.6t_4$ and $n = 400$. Using $\gamma = 7/40$ in (21), $\hat{f}_Y(y)$ (dashed curve in left panel) and $\hat{f}'_Y(y)$ (dashed curve in right panel) match the true curves (solid curve in all plots) very well for both realizations. By contrast, while $\tilde{f}_Y(y)$ (dotted curve in left panel) works reasonably well for density estimate (not as good as $\hat{f}_Y(y)$ though), $\tilde{f}'_Y(y)$ (dotted curve in right panel) has very poor performance in the second realization. In our simulations, realizations with such poor performance of $\tilde{f}'_Y(y)$ occur quite often, resulting in MSE($\tilde{f}'_Y$) being about 20 times larger than MSE($\hat{f}'_Y$).

Comparison of specification tests T_n in (9) and D_n in (8):

Consider testing $H_0 : \mu(x) = \beta_0 + \beta_1 x$ when the data are actually generated from Model 1 (null model) and Models 2–4 (alternative models) with noise $e \sim N(0, 0.6^2)$. As discussed in Section 2.2, the cutoff value of T_n is obtained by bootstrap method with 1000 bootstrap realizations and, for fair comparison, the cutoff value of D_n is obtained by similar bootstrap method. Using $\gamma = 7/40$ in (21), Table 3 summarizes the empirical size (Model 1) and empirical power (Models 2–4) based on 100 realizations, for sample size $n = 50, 100, 200, 400$, and at significance level $\alpha = 5\%$. Clearly, test T_n has empirical size close to the nominal level and its power increases drastically with sample size. At $n = 200$, T_n has power about 85%; at $n = 400$, T_n has power 100%. By contrast, test D_n has little power detecting deviations from null hypothesis, and this poor performance is consistent with the findings in Pritsker (1998) and Kim et al. (2015). In our simulations we also tried significance levels 10% and 1% and the conclusions are similar.

Table 1: MSE of $\hat{f}_Y$ and relative MSE defined as RMSE $= \frac{\text{MSE}(\hat{f}_Y)}{\text{MSE}(\tilde{f}_Y)}$, using bandwidth h_n in (21) with $\gamma = 2/20, 3/20, 4/20, 5/20$. RMSE ≤ 1 indicates better performance of $\hat{f}_Y(y)$.

			$\gamma = 2/20$		$\gamma = 3/20$		$\gamma = 4/20$		$\gamma = 5/20$	
		n	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE
Model 1	$e \sim N(0, 0.6^2)$	50	0.0106	0.614	0.0104	0.609	0.0106	0.636	0.0113	0.653
		100	0.0047	0.479	0.0045	0.468	0.0047	0.492	0.0050	0.514
		200	0.0021	0.380	0.0021	0.373	0.0023	0.408	0.0025	0.419
		400	0.0010	0.277	0.0010	0.282	0.0010	0.306	0.0011	0.316
	$e \sim 0.6t_4$	50	0.0122	0.751	0.0124	0.738	0.0121	0.744	0.0130	0.754
		100	0.0060	0.641	0.0063	0.673	0.0071	0.714	0.0061	0.635
		200	0.0033	0.573	0.0032	0.558	0.0032	0.564	0.0034	0.602
		400	0.0016	0.465	0.0017	0.505	0.0017	0.502	0.0017	0.544
Model 2	$e \sim N(0, 0.6^2)$	50	0.0098	0.642	0.0103	0.670	0.0102	0.648	0.0114	0.738
		100	0.0044	0.491	0.0045	0.503	0.0046	0.539	0.0049	0.546
		200	0.0020	0.377	0.0021	0.397	0.0022	0.410	0.0023	0.428
		400	0.0010	0.326	0.0010	0.307	0.0011	0.342	0.0010	0.338
	$e \sim 0.6t_4$	50	0.0114	0.776	0.0115	0.794	0.0115	0.758	0.0126	0.811
		100	0.0059	0.683	0.0056	0.655	0.0059	0.694	0.0055	0.656
		200	0.0027	0.558	0.0028	0.559	0.0031	0.612	0.0029	0.575
		400	0.0015	0.483	0.0014	0.469	0.0015	0.507	0.0015	0.525
Model 3	$e \sim N(0, 0.6^2)$	50	0.0098	0.585	0.0106	0.652	0.0105	0.626	0.0115	0.698
		100	0.0045	0.480	0.0045	0.479	0.0046	0.478	0.0054	0.554
		200	0.0021	0.387	0.0022	0.393	0.0022	0.405	0.0042	0.419
		400	0.0009	0.296	0.0010	0.305	0.0010	0.304	0.0011	0.319
	$e \sim 0.6t_4$	50	0.0114	0.704	0.0116	0.733	0.0116	0.774	0.0121	0.776
		100	0.0058	0.631	0.0062	0.703	0.0057	0.636	0.0061	0.663
		200	0.0031	0.591	0.0031	0.569	0.0032	0.593	0.0032	0.592
		400	0.0016	0.479	0.0016	0.489	0.0015	0.470	0.0016	0.509
Model 4	$e \sim N(0, 0.6^2)$	50	0.0095	0.575	0.0114	0.659	0.0106	0.653	0.0114	0.684
		100	0.0046	0.478	0.0045	0.468	0.0047	0.496	0.0050	0.534
		200	0.0021	0.378	0.0023	0.395	0.0023	0.413	0.0024	0.429
		400	0.0010	0.316	0.0010	0.317	0.0011	0.327	0.0011	0.340
	$e \sim 0.6t_3$	50	0.0120	0.753	0.0115	0.738	0.0116	0.733	0.0126	0.776
		100	0.0062	0.655	0.0060	0.641	0.0065	0.679	0.0061	0.675
		200	0.0031	0.576	0.0030	0.571	0.0031	0.582	0.0033	0.597
		400	0.0016	0.504	0.0016	0.492	0.0016	0.498	0.0015	0.470

Table 2: MSE of the derivative estimate $\hat{f}'_Y$ and relative MSE defined as RMSE $= \frac{\text{MSE}(\hat{f}'_Y)}{\text{MSE}(\tilde{f}'_Y)}$, using bandwidth h_n in (21) with $\gamma = 2/20, 3/20, 4/20, 5/20$. RMSE ≤ 1 indicates better performance of $\hat{f}'_Y(y)$.

			$\gamma = 2/20$		$\gamma = 3/20$		$\gamma = 4/20$		$\gamma = 5/20$	
		n	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE
Model 1	$e \sim N(0, 0.6^2)$	50	0.0633	0.038	0.0626	0.046	0.0689	0.047	0.0784	0.053
		100	0.0249	0.037	0.0256	0.029	0.0299	0.046	0.0312	0.037
		200	0.0111	0.016	0.0109	0.020	0.0123	0.023	0.0127	0.022
		400	0.0051	0.017	0.0051	0.013	0.0055	0.014	0.0060	0.020
	$e \sim 0.6t_4$	50	0.0694	0.078	0.0642	0.093	0.0672	0.099	0.0695	0.084
		100	0.0356	0.073	0.0365	0.058	0.0360	0.054	0.0371	0.056
		200	0.0187	0.047	0.0187	0.068	0.0196	0.059	0.0201	0.053
		400	0.0100	0.048	0.0101	0.046	0.0103	0.044	0.0101	0.049
Model 2	$e \sim N(0, 0.6^2)$	50	0.0493	0.040	0.0569	0.062	0.0593	0.062	0.0666	0.062
		100	0.0212	0.030	0.0236	0.035	0.0242	0.030	0.0251	0.033
		200	0.0098	0.020	0.0099	0.025	0.0106	0.029	0.0113	0.022
		400	0.0044	0.015	0.0048	0.015	0.0049	0.019	0.0050	0.019
	$e \sim 0.6t_4$	50	0.0546	0.067	0.0568	0.061	0.0573	0.094	0.0635	0.098
		100	0.0269	0.074	0.0291	0.064	0.0282	0.066	0.0307	0.065
		200	0.0142	0.043	0.0153	0.044	0.0141	0.051	0.0150	0.060
		400	0.0069	0.034	0.0076	0.041	0.0727	0.042	0.0747	0.039
Model 3	$e \sim N(0, 0.6^2)$	50	0.0528	0.033	0.0621	0.049	0.0656	0.052	0.0678	0.047
		100	0.0232	0.030	0.0239	0.042	0.0268	0.037	0.0278	0.038
		200	0.0108	0.023	0.0118	0.025	0.0170	0.026	0.0123	0.025
		400	0.0051	0.018	0.0051	0.016	0.0053	0.013	0.0056	0.019
	$e \sim 0.6t_4$	50	0.0632	0.0692	0.0612	0.085	0.0637	0.082	0.0665	0.072
		100	0.0310	0.059	0.0371	0.071	0.0322	0.057	0.0320	0.061
		200	0.0160	0.038	0.0171	0.052	0.0170	0.066	0.0172	0.050
		400	0.0089	0.037	0.0087	0.039	0.0088	0.051	0.0907	0.043
Model 4	$e \sim N(0, 0.6^2)$	50	0.0553	0.030	0.0623	0.063	0.0669	0.060	0.0773	0.057
		100	0.0259	0.032	0.0269	0.027	0.0252	0.036	0.0294	0.030
		200	0.0109	0.018	0.0114	0.023	0.0125	0.023	0.0132	0.028
		400	0.0053	0.013	0.0055	0.014	0.0054	0.017	0.0060	0.014
	$e \sim 0.6t_3$	50	0.0668	0.082	0.0709	0.083	0.0697	0.089	0.0677	0.089
		100	0.0324	0.061	0.0331	0.057	0.0346	0.068	0.0352	0.068
		200	0.0174	0.044	0.0161	0.059	0.0184	0.051	0.0184	0.046
		400	0.0886	0.041	0.0856	0.030	0.0927	0.043	0.0096	0.044

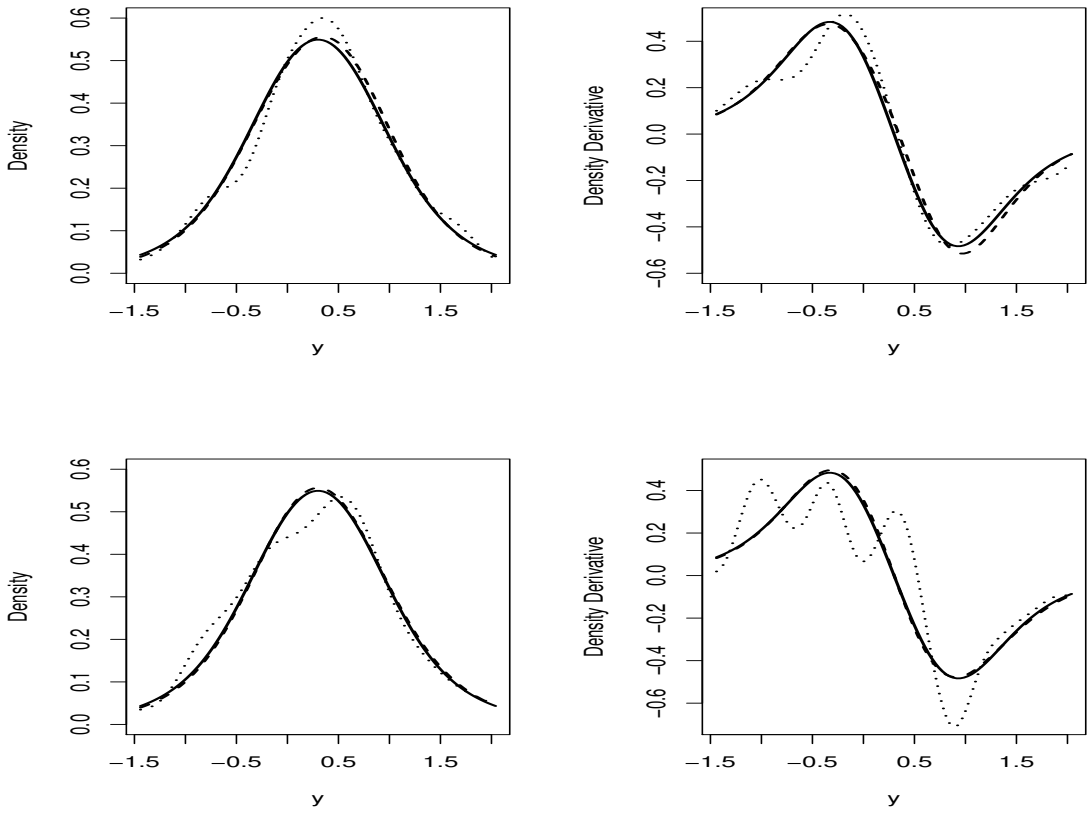


Figure 1: Two realizations (one realization on each row) of density estimates (left plots) and density derivative estimates (right plots) from Model 2 with $e \sim 0.6t_4$ and $n = 400$. In each plot, solid curve (——) is the true function, dotted curve ($\cdots$) is the nonparametric kernel estimates, and dashed curve (−−) is the proposed method.

Table 3: Comparison of size (under null Model 1) and power (under alternative Models 2–4) for specification tests T_n (the first number in each cell) and D_n (the bracketed number), significance level 5%.

		$n = 50$	$n = 100$	$n = 200$	$n = 400$
Null model:	Model 1	0.04(0.01)	0.03(0.01)	0.05(0.00)	0.04(0.00)
Alternative models:	Model 2	0.18(0.03)	0.45(0.06)	0.88(0.09)	1.00(0.09)
	Model 3	0.25(0.04)	0.51(0.03)	0.85(0.04)	1.00(0.03)
	Model 4	0.19(0.02)	0.49(0.03)	0.84(0.02)	1.00(0.04)

Acknowledgements

This paper is based on one chapter of the author's doctoral dissertation (Hernandez-Bejarano (2024)) submitted to the Department of Statistics at The Pennsylvania State University, University Park, USA, 2024.

References

Aït-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate. *Review of Financial Studies*, 9:385–426.

Aït-Sahalia, Y., Fan, F., and Peng, H. (2009). Nonparametric transition-based tests for jump-diffusions. *Journal of the American Statistical Association*, 104:1102–1116.

Bickel, P. and Ritov, Y. (1988). Estimating integrated squares density derivatives: sharp best order of convergence estimates. *Sankhya, Series A*, 50:381–393.

Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Annals of Statistics*, 23:11–29.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. 2nd Edition. Belmont, CA, Duxbury.

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

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

Frees, E. W. (1994). Estimating densities of functions of observations. *Journal of the American Statistical Association*, 89:517–525.

Gao, J. and King, M. (2004). Adaptive testing in continuous-time diffusion models. *Econometric Theory*, 20:844–882.

Giné, E. and Mason, D. M. (2007). On local U-statistic processes and the estimation of densities of functions of several sample variables. *Annals of Statistics*, 35:1105–1145.

Hall, P. and Marron, J. (1987). Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, 6:109–115.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748.

Hernandez-Bejarano, M. D. (2024). *Density Estimation for Some Semiparametric Models*. Doctoral dissertation, Pennsylvania State University, University Park, PA. https://www.proquest.com/docview/3110363675?sourcetype=Dissertations

Hong, Y. and Li, H. (2005). Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *Review of Financial Studies*, 18:37–84.

Jennrich, R. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, 40:633–643.

Kerkyacharian, G. and Picard, D. (1996). Estimating nonquadratic functionals of a density using Haar wavelets. *Annals of Statistics*, 24:485–507.

Kim, K. H., Zhang, T., and Wu, W. B. (2015). Parametric specification test for nonlinear autoregressive models. *Econometric Theory*, 31(5):1078–1101.

Li, Q. and Racine, J. (2006). *Nonparametric Econometrics: Theory and Practice, 1st.* Princeton University Press, 1st.

Pritsker, M. (1998). Nonparametric density estimation and tests of continuous time interest rate models. *Review of Financial Studies*, 11:449–487.

Ritov, Y. and Bickel, P. (1990). Achieving information bounds in non and semiparametric models. *Annals of Statistics*, 18:925–938.

Rudin, W. (1976). *Principles of Mathematical Analysis.* 3rd Edition. McGraw-Hill.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90:1257–1270.

Saavedra, A. and Cao, R. (1999). Rate of convergence of a convolution-type estimator of the marginal density of an MA(1) process. *Stochastic Processes and their Applications*, 80:129–155.

Saavedra, A. and Cao, R. (2000). On the estimation of the marginal density of a moving average process. *Canadian Journal of Statistics*, 28:799–815.

Schick, A. and Wefelmeyer, W. (2004). Functional convergence and optimality of plug-in estimators for stationary densities of moving average processes. *Bernoulli*, 10:889–917.

Schick, A. and Wefelmeyer, W. (2007). Uniformly root-n consistent density estimators for weakly dependent invertible linear processes. *Annals of Statistics*, 35:815–843.

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

Zhao, Z. (2010). Density estimation for nonlinear parametric models with conditional heteroscedasticity. *Journal of Econometrics*, 155:71–82.

Zhao, Z. (2011). Nonparametric model validations for hidden Markov models with applications in financial econometrics. *Journal of Econometrics*, 162.:225–239.

Appendices

A Proofs

Write $\|Z\|^2 = \mathbb{E}(Z^2)$ for a random variable Z. Throughout, $c_1, c_2, \ldots$, are generic constants that may vary from places to places. Also, $\mathcal{Y}$ is any bounded interval with length denoted by $|\mathcal{Y}|$. First, we present two useful lemmas.

Lemma 1. *(i) For any differentiable random function $H(y)$,*

$$\sup_{y\in\mathcal{Y}} |H(y)| = O_p\left\{\inf_{y\in\mathcal{Y}} \|H(y)\| + \sup_{y\in\mathcal{Y}} \|H'(y)\|\right\}. \tag{28}$$

(ii) If $H(y) = \sum_{i=1}^n H_i(y)$, where $H_1(y), \ldots, H_n(y)$ are iid and differentiable, then

$$\sup_{y\in\mathcal{Y}} |H(y) - n\mathbb{E}H_1(y)| = O_p\left\{\sqrt{n}\left[\inf_{y\in\mathcal{Y}} \|H_1(y)\| + \sup_{y\in\mathcal{Y}} \|H_1'(y)\|\right]\right\}. \tag{29}$$

Proof. (i) For any $y, y_0 \in \mathcal{Y}$, $|H(y) - H(y_0)| = |\int_{y_0}^y H'(z)dz| \le \int_{\mathcal{Y}} |H'(z)|dz$. Thus,

$$\sup_{y\in\mathcal{Y}}[H(y) - H(y_0)]^2 \le \left[\int_{\mathcal{Y}} |H'(z)|dz\right]^2 \le |\mathcal{Y}| \int_{\mathcal{Y}} H'(z)^2 dz,$$

where the second "$\le$" follows from Cauchy-Schwarz inequality. Taking expectation yields

$$\mathbb{E}\Big\{\sup_{y\in\mathcal{Y}}[H(y) - H(y_0)]^2\Big\} \le |\mathcal{Y}| \int_{\mathcal{Y}} \mathbb{E}[H'(z)^2]dz \le |\mathcal{Y}|^2 \sup_{y\in\mathcal{Y}} \mathbb{E}[H'(y)^2] = O(1) \sup_{y\in\mathcal{Y}} \mathbb{E}[H'(y)^2],$$

which, after taking square root and using triangle inequality, implies

$$\left\|\sup_{y\in\mathcal{Y}} |H(y)|\right\| = O(1)\left[\|H(y_0)\| + \sup_{y\in\mathcal{Y}} \|H'(y)\|\right].$$

Since y_0 is arbitrary, (28) follows. (ii) By writing $H(y) - n\mathbb{E}H_1(y) = \sum_{i=1}^n [H_i(y) - \mathbb{E}H_i(y)]$ with the summands being iid with zero mean, the desired result follows from (28). ♢

The following lemma can greatly simplify the presentation of the main proofs.

Lemma 2. *Let $g(\cdot)$ and $h(\cdot,\cdot)$ be functions such that $\mathbb{E}[g(e_i)] = 0$, $\mathbb{E}[g^2(e_i)] < \infty$, $\sup_{y\in\mathcal{Y}} \mathbb{E}\{h^2(y, X_i) + [\partial h(y, X_i)/\partial y]^2\} < \infty$. For $\mathcal{N}$ in (12), define*

$$\Lambda_n(y) = \frac{1}{|\mathcal{N}|} \sum_{i\in\mathcal{N}} g(e_i)h(y, X_i) - \frac{1}{n}\sum_{i=1}^n g(e_i)h(y, X_i).$$

Then $\sup_{y\in\mathcal{Y}} |\Lambda_n(y)| = o_p(n^{-1/2})$.

Proof. Let $\mathcal{N}^c$ be the complement of $\mathcal{N}$ and $\mathbf{1}$ the indicator function. Introducing

$$\Lambda_{n,1}(y) = \sum_{i=1}^{n} g(e_i)h(y, X_i) \quad \text{and} \quad \Lambda_{n,2}(y) = \sum_{i=1}^{n} g(e_i)h(y, X_i)\mathbf{1}_{i\in\mathcal{N}^c},$$

we can rewrite

$$\Lambda_n(y) = \left(\frac{1}{|\mathcal{N}|} - \frac{1}{n}\right)\Lambda_{n,1}(y) - \frac{1}{|\mathcal{N}|}\Lambda_{n,2}(y). \tag{30}$$

By Assumption 2(iv), $|\mathcal{N}^c| = O_p(nh_n) = o_p(n)$, which implies $1/|\mathcal{N}| - 1/n = o_p(1/n)$. By (30), to prove $\sup_{y\in\mathcal{Y}} |\Lambda_n(y)| = o_p(n^{-1/2})$, it suffices to prove

$$\sup_{y\in\mathcal{Y}} |\Lambda_{n,1}(y)| = O_p(\sqrt{n}) \quad \text{and} \quad \sup_{y\in\mathcal{Y}} |\Lambda_{n,2}(y)| = o_p(\sqrt{n}). \tag{31}$$

Since $\{(e_i, X_i)\}$ are iid and e_i is independent of X_i, the assumption $\mathbb{E}[g(e_i)] = 0$ implies that the summands in $\Lambda_{n,1}(y)$ are iid and have zero mean. By the same argument, the summands in the derivative $\Lambda'_{n,1}(y)$ are also iid and have zero mean. Therefore, an application of Lemma 1(ii) gives $\sup_{y\in\mathcal{Y}} |\Lambda_{n,1}(y)| = O_p(\sqrt{n})$. For $\Lambda_{n,2}(y)$, under the given conditions, by the Dominated Convergence Theorem,

$$\|g(e_i)h(y, X_i)\mathbf{1}_{i\in\mathcal{N}^c}\|^2 \to 0 \quad \text{and} \quad \left\|g(e_i)\frac{\partial h(y, X_i)}{\partial y}\mathbf{1}_{i\in\mathcal{N}^c}\right\|^2 \to 0.$$

Then an application of Lemma 1(ii) gives $\sup_{y\in\mathcal{Y}} |\Lambda_{n,2}(y)| = o_p(\sqrt{n})$. ♢

Remark 3. Thanks to the introduction of $\mathcal{N}$ in (12), by (11), we have

$$\max_{i\in\mathcal{N}} |\hat{\mu}(X_i) - \mu(X_i)| = o_p(n^{-1/4}). \tag{32}$$

A careful examination of the proofs below reveals that the only reason we need to impose $i \in \mathcal{N}$ is to ensure the uniform approximation (32). In subsequent proofs, for ease of presentation, we shall pretend that the uniform approximation (32) holds for all i, that is,

$$\max_{1\le i\le n} |\hat{\mu}(X_i) - \mu(X_i)| = o_p(n^{-1/4}). \tag{33}$$

Otherwise, we can always restrict our attention on $\mathcal{N}$ to derive various approximations and then apply Lemma 2 to further approximate such results established on $\mathcal{N}$ to the counterparts on $\{1, 2, \ldots, n\}$ with negligible error term $o_p(n^{-1/2})$. Thus, without loss of generality, from now on we shall simply use $\mathcal{N} = \{1, 2, \ldots, n\}$.

Proof of Theorem 1. From (33), $\max_{1\le i\le n} |\hat{\mu}(X_i) - \mu(X_i)| = o_p(n^{-1/4})$; see Remark 3. Also, by Lemma 3 below, $\hat{\theta} = \theta + O_p(n^{-1/2})$. Under Assumption 3(ii), applying Taylor's expansion, we can obtain (again,

as discussed in Remark 3, we take $\mathcal{N} = \{1, 2, \ldots, n\}$):

$$\hat{f}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} f_e(y - \mu(X_i)|\theta) - I_n(y) + J_n(y)^T(\hat{\theta} - \theta) + O(\delta_n), \tag{34}$$

where

$$\begin{aligned} I_n(y) &= \frac{1}{n}\sum_{i=1}^{n} f_e'(y - \mu(X_i)|\theta)[\hat{\mu}(X_i) - \mu(X_i)], & (35)\\ J_n(y) &= \frac{1}{n}\sum_{i=1}^{n} \dot{f}_e(y - \mu(X_i)|\theta) \quad \text{with} \quad \dot{f}_e(z|\theta) = \frac{\partial f_e(z|\theta)}{\partial \theta}, & (36)\\ \delta_n &= \sup_{1\le i\le n} [\hat{\mu}(X_i) - \mu(X_i)]^2 + |\hat{\theta} - \theta|^2 = o_p(n^{-1/2}). & (37) \end{aligned}$$

Under Assumption 3(ii), an application of Lemma 1(ii) gives

$$\sup_{y\in\mathcal{Y}} \left| J_n(y) - \mathbb{E}[\dot{f}_e(y - \mu(X)|\theta)] \right| = o_p(1). \tag{38}$$

From (34), (37), (38) and Lemmas 3–4 below, we can obtain the uniform approximation

$$\sqrt{n}[\hat{f}_Y(y) - f_Y(y)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\zeta_i(y) - f_Y(y)] + o_p(1), \tag{39}$$

uniformly over $y \in \mathcal{Y}$, where

$$\zeta_i(y) = f_e(y - \mu(X_i)|\theta) - e_i f_e'(y - \mu(X_i)|\theta) + \mathbb{E}[\dot{f}_e(y - \mu(X)|\theta)]^T \times \mathcal{I}_\theta^{-1} \times \psi(e_i|\theta). \tag{40}$$

From the identity (3), $\mathbb{E}(e_i) = 0$, and $\mathbb{E}\psi(e_i|\theta) = 0$ since $\psi(e_i|\theta)$ is the score function, the summands $\{\zeta_i(y) - f_Y(y)\}_{i=1}^n$ are iid and have zero mean. The finite-dimensional convergence easily follows from the Cramér-Wold device. Also, tightness follows since all the functions involving y has bounded derivative with respect to y (Assumption 3(ii)). Thus, $\{\sqrt{n}[\hat{f}_Y(y) - f_Y(y)], y \in \mathcal{Y}\} \Rightarrow \{G(y), y \in \mathcal{Y}\}$ for a centered Gaussian process $\{G(y)\}$ with $\mathrm{Cov}\{G(y), G(y')\} = \mathrm{Cov}\{\zeta_i(y), \zeta_i(y')\}$.

Note that, we have

$$\mathbb{E}[e_i\psi(e_i|\theta)] = \mathbb{E}\left[e_i \frac{\dot{f}_e(e_i|\theta)}{f_e(e_i|\theta)}\right] = \int_{\mathbb{R}} z\dot{f}_e(z|\theta)dz = \frac{\partial \int_{\mathbb{R}} z f_e(z|\theta)dz}{\partial\theta} = \frac{\partial \mathbb{E}(e_i)}{\partial\theta} = 0, \tag{41}$$

where the last "=" follows from $\mathbb{E}(e_i) = 0$ and the third "=" follows from exchanging the order of integral and derivative, which is justified via the Dominated Convergence Theorem under Assumption 3(iv). Using (41), $\mathbb{E}(e_i) = 0$, $\mathbb{E}\psi(e_i|\theta) = 0$, and the independence between e_i and X_i, we can verify that $\mathrm{Cov}\{\zeta_i(y), \zeta_i(y')\}$ is given by (16). ♢

Proof of Theorem 2. We only consider the uniform consistency for $\hat{\Sigma}_0(y, y')$ since the other cases $\hat{\Sigma}_\mu(y, y')$ and $\hat{\Sigma}_\theta(y, y')$ can be treated similarly. By $\hat{\theta} = \theta + O_p(n^{-1/2})$, (33), and Assumption 3(ii), $\sup_{1\le i\le n, y\in\mathcal{Y}} |f_e(y - \hat{\mu}(X_i)|\hat{\theta}) - f_e(y - \mu(X_i)|\theta)| = o_p(1)$. Thus,

$$\sup_{y,y'\in\mathcal{Y}} \left| \frac{1}{n}\sum_{i=1}^{n} f_e(y - \hat{\mu}(X_i)|\hat{\theta}) f_e(y' - \hat{\mu}(X_i)|\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} f_e(y - \mu(X_i)|\theta) f_e(y' - \mu(X_i)|\theta) \right| \xrightarrow{p} 0.$$

Using the same argument in the proof of (38), we can prove

$$\sup_{y,y'\in\mathcal{Y}} \left| \frac{1}{n}\sum_{i=1}^{n} f_e(y - \mu(X_i)|\theta) f_e(y' - \mu(X_i)|\theta) - \mathbb{E}[f_e(y - \mu(X)|\theta) f_e(y' - \mu(X)|\theta)] \right| \xrightarrow{p} 0.$$

Combining the above two results, we have $n^{-1}\sum_{i=1}^{n} f_e(y - \hat{\mu}(X_i)|\hat{\theta}) f_e(y' - \hat{\mu}(X_i)|\hat{\theta}) \xrightarrow{p} \mathbb{E}[f_e(y - \mu(X)|\theta) f_e(y' - \mu(X)|\theta)]$, uniformly in $y, y' \in \mathcal{Y}$. Also, from Theorem 1, $\hat{f}_Y(y) \xrightarrow{p} f_Y(y)$, uniformly in $y \in \mathcal{Y}$. Thus, $\hat{\Sigma}_0(y, y') \xrightarrow{p} \Sigma_0(y, y')$, uniformly in $y, y' \in \mathcal{Y}$. ♢

Proof of Theorem 3. By Theorem 1, $\sqrt{n}[\hat{f}_Y(y) - f_Y(y)]$ converges to a Gaussian process $G(y)$ on $\mathcal{Y}$. Note that, under the newly imposed strengthened version of Assumption 3(ii), $\zeta_i(y)$ in (40) is k times differentiable, so the limiting process $G(y)$ is also k times differentiable. By the same argument [Equation (34)] in the proof of Theorem 1, we can show that $\sqrt{n}[\hat{f}_Y^{(k)}(y) - f_Y^{(k)}(y)]$ converges uniformly to a Gaussian process on $\mathcal{Y}$. Under this setting, convergence of function implies convergence of derivatives [see Rudin (1976)], and thus (20) holds. Since the Gaussian process $G(y)$ is k times differentiable, the derivative $G^{(k)}(y)$ is again a Gaussian process with the desired autocovariance. ♢

Proof of Theorem 4. Under the regularity conditions in Assumption 4, for $\hat{\beta}$ in the nonlinear least-squares estimation (22), we have $|\hat{\beta} - \beta| = O_p(n^{-1/2})$ and

$$\hat{\beta} - \beta = \frac{\mathcal{J}_\beta^{-1}}{n}\sum_{i=1}^{n} e_i \dot{\mu}_\beta(X_i) + o_p(n^{-1/2}). \tag{42}$$

See, e.g., Jennrich (1969). For $\check{f}_Y(y)$ in (23), by a similar Taylor expansion in (34),

$$\check{f}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} f_e(y - \mu_\beta(X_i)|\theta) - N_n(y)^T(\hat{\beta} - \beta) + J_n(y)^T(\hat{\theta} - \theta) + o_p(n^{-1/2}), \tag{43}$$

where J_n is defined in (36) with $\mu(X_i)$ therein replaced by $\mu_\beta(X_i)$, and

$$N_n(y) = \frac{1}{n}\sum_{i=1}^{n} f_e'(y - \mu_\beta(X_i)|\theta)\dot{\mu}_\beta(X_i).$$

An application of Lemma 1(ii) gives

$$\sup_{y\in\mathcal{Y}} |N_n(y) - \mathbb{E}[f'_e(y-\mu_\beta(X)|\theta)\dot{\mu}(X;\beta)]| = o_p(1). \tag{44}$$

Comparing (34) and (43) and using (42), (44), and (50) below, we can obtain

$$\sqrt{n}[\hat{f}_Y(y) - \check{f}_Y(y)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_i(y) + o_p(1),$$

uniformly over $y \in \mathcal{Y}$, where

$$\eta_i(y) = \Big\{\mathbb{E}[f'_e(y-\mu_\beta(X)|\theta)\dot{\mu}_\beta(X)^T]\mathcal{J}_\beta^{-1}\dot{\mu}_\beta(X_i) - f'_e(y-\mu_\beta(X_i)|\theta)\Big\}\, e_i.$$

As in the proof of Theorem 1, tightness and finite-dimensional convergence can be established and thus

$$\Big\{\sqrt{n}[\hat{f}_Y(y) - \check{f}_Y(y)], \quad y\in\mathcal{Y}\Big\} \Rightarrow \Big\{Z(y), \quad y\in\mathcal{Y}\Big\},$$

where $\{Z(y)\}$ is centered Gaussian process and $\mathrm{Cov}\{Z(y), Z(y')\} = \mathrm{Cov}\{\eta_i(y), \eta_i(y')\}$ which, after some calculations, reduces to the autocovariance function given in Theorem 4. The convergence of nT_n then follows from the continuous mapping theorem. ♢

Proof of Theorem 5. For model (27), $f_Y(y) = \phi(y;\mu,\theta)$ and $\zeta_i(y)$ in (40) becomes $\zeta_i(y) = \phi(y;\mu,\theta) + e_i\dot{\phi}_\mu(y;\mu,\theta) + e_i^2\dot{\phi}_\theta(y;\mu,\theta)$. Therefore, from (43), we have

$$\hat{f}_Y(y) = \phi(y;\mu,\theta) + \frac{1}{n}\sum_{i=1}^n \Big\{e_i\dot{\phi}_\mu(y;\mu,\theta) + (e_i^2-\theta)\dot{\phi}_\theta(y;\mu,\theta)\Big\} + o_p(n^{-1/2}), \tag{45}$$

uniformly over $y \in \mathcal{Y}$. On the other hand, by Taylor's expansion,

$$\hat{f}_Y^{\mathrm{MLE}}(y) = \phi(y;\mu,\theta) + (\overline{\mu}-\mu)\dot{\phi}_\mu(y;\mu,\theta) + (\overline{\theta}-\theta)\dot{\phi}_\theta(y;\mu,\theta) + O_p(n^{-1}), \tag{46}$$

uniformly over $y \in \mathcal{Y}$. Using $\overline{\mu}-\mu = n^{-1}\sum_{i=1}^n e_i$ and $\overline{\theta} = n^{-1}\sum_{i=1}^n e_i^2 + O_p(1/n)$ and from (45)–(46), we conclude that $\sup_{y\in\mathcal{Y}}|\hat{f}_Y(y) - \hat{f}_Y^{\mathrm{MLE}}(y)| = o_p(n^{-1/2})$. ♢

Lemma 3. *For $\hat{\theta}$ in (13), under the conditions in Theorem 1, we have*

$$\hat{\theta} = \theta + \frac{\mathcal{I}_\theta^{-1}}{n}\sum_{i=1}^n \psi(e_i|\theta) + o_p(n^{-1/2}), \tag{47}$$

where $\psi(e|\theta)$ and $\mathcal{I}_\theta$ are defined in Theorem 1. In particular, $\hat{\theta} = \theta + O_p(n^{-1/2})$.

Proof. The maximization problem (13) is equivalent to $\hat{\theta} = \text{argmin}_\theta \frac{1}{n}\sum_{i=1}^n \log f_e(\hat{e}_i|\theta)$. Define $\hat{\theta}^* = \text{argmin}_\theta \frac{1}{n}\sum_{i=1}^n \log f_e(e_i|\theta)$, which is the MLE of θ based on $e_1, \ldots, e_n$. Under regularity conditions in Assumption 3(iv), the classical result in MLE states that $\hat{\theta}^*$ satisfies the asymptotic representation (47). Thus, from the classical proof of consistency and asymptotic normality of MLE, to show that $\hat{\theta}$ also satisfies (47), it suffices to establish the uniform convergence

$$\Delta_n := \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \log f_e(\hat{e}_i|\theta) - \frac{1}{n}\sum_{i=1}^n \log f_e(e_i|\theta)\right| \xrightarrow{p} 0. \tag{48}$$

From (33),

$$\max_{1\le i\le n}|\hat{e}_i - e_i| = \max_{1\le i\le n}|\hat{\mu}(X_i) - \mu(X_i)| \xrightarrow{p} 0. \tag{49}$$

Thus, for $\epsilon > 0$ in Assumption 3(iii), for large enough n, $\max_{1\le i\le n}|\hat{e}_i - e_i| < \epsilon$ with probability tending to one. By mean value theorem and Assumption 3(iii),

$$\begin{aligned}
\Delta_n &\le \frac{1}{n}\sum_{i=1}^n \sup_{\theta\in\Theta}|\log f_e(\hat{e}_i|\theta) - \log f_e(e_i|\theta)| \\
&\le \frac{1}{n}\sum_{i=1}^n |\hat{e}_i - e_i|\ell_1(e_i) \qquad \Big(\ell_1(e_i) \text{ is defined in Assumption 3(iii)}\Big) \\
&\le \max_{1\le i\le n}|\hat{e}_i - e_i| \times \frac{1}{n}\sum_{i=1}^n \ell_1(e_i) \xrightarrow{p} 0,
\end{aligned}$$

in view of (49) and $n^{-1}\sum_{i=1}^n \ell_1(e_i) \xrightarrow{p} \mathbb{E}\ell_1(e) < \infty$. This completes the proof. ♢

Lemma 4. *For $I_n(y)$ in (35), under conditions in Theorem 1, we have*

$$I_n(y) = \frac{1}{n}\sum_{j=1}^n e_j f_e'(y - \mu(X_j)|\theta) + o_p(n^{-1/2}), \quad \textit{uniformly over } y \in \mathcal{Y}. \tag{50}$$

Proof. For convenience, we write $K_{h_n}(u) = K(u/h_n)/h_n$. Rewrite

$$\hat{\mu}(x) = \sum_{i=1}^n w_i(x)Y_i \quad \text{with} \quad w_i(x) = \frac{K_{h_n}(x - X_i)}{n\hat{f}_X(x)}, \quad \hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^n K_{h_n}(x - X_i). \tag{51}$$

Using $\sum_{i=1}^n w_i(x) = 1$ and $Y_i = \mu(X_i) + e_i$, we can obtain

$$\hat{\mu}(x) - \mu(x) = B_n(x) + \sum_{i=1}^n w_i(x)e_i \quad \text{with} \quad B_n(x) = \sum_{i=1}^n w_i(x)[\mu(X_i) - \mu(x)]. \tag{52}$$

The bias term $B_n(x)$ satisfies $\sup_x |B_n(x)| = O_p(h_n^2)$. Thus, by (52) and the uniform boundedness of

$f'_e(z|\theta)$ (Assumption 3(ii)), we have

$$I_n(y) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_j(X_i)e_j f'_e(y-\mu(X_i)|\theta) + O_p(h_n^2). \tag{53}$$

For $\hat{f}_X(x)$ in (51), by property of nonparametric kernel density estimator,

$$\hat{f}_X(X_i) = f_X(X_i) + O_p\Big(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\Big), \quad \text{uniformly in } i = 1, 2, \ldots. \tag{54}$$

Again, as discussed in Remark 3, here we have used all $i \in \{1, 2, \ldots, n\}$ instead of $i \in \mathcal{N}$. Substituting this into $w_i(x)$ in (51), by (53) we have

$$I_n(y) = \left[1 + O_p\Big(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\Big)\right]\tilde{I}_n(y) + O_p(h_n^2), \tag{55}$$

where

$$\tilde{I}_n(y) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\xi_{ij}(y)e_j \quad \text{with} \quad \xi_{ij}(y) = \frac{K_{h_n}(X_i - X_j)f'_e(y-\mu(X_i)|\theta)}{f_X(X_i)}.$$

Further write

$$\begin{aligned}\tilde{I}_n(y) &= \frac{1}{n^2}\sum_{i=1}^{n}\xi_{ii}(y)e_i + \frac{1}{n^2}\sum_{i\neq j}\{\xi_{ij}(y) - \mathbb{E}[\xi_{ij}(y)|X_j]\}e_j + \frac{1}{n^2}\sum_{i\neq j} e_j\mathbb{E}[\xi_{ij}(y)|X_j] \\ &:= \tilde{I}_{n,0}(y) + \tilde{I}_{n,1}(y) + \tilde{I}_{n,2}(y).\end{aligned} \tag{56}$$

For $\tilde{I}_{n,0}(y)$, by Lemma 1(ii), we have

$$\tilde{I}_{n,0}(y) = \frac{K(0)}{n^2h_n}\sum_{i=1}^{n}\frac{f'_e(y-\mu(X_i)|\theta)}{f_X(X_i)} = O_p\Big(\frac{1}{nh_n}\Big) = o_p(n^{-1/2}),$$

uniformly over $y \in \mathcal{Y}$. Therefore, from (56), in order to prove (50), it suffices to establish:

$$\sup_{y\in\mathcal{Y}}|\tilde{I}_{n,1}(y)| = o_p(n^{-1/2}) \text{ and } \sup_{y\in\mathcal{Y}}\left|\tilde{I}_{n,2}(y) - \frac{1}{n}\sum_{j=1}^{n} e_j f'_e(y-\mu(X_j)|\theta)\right| = o_p(n^{-1/2}). \tag{57}$$

[proof of the first assertion in (57)]: Essentially, (56) is a Hoeffding-type decomposition, and similar to the property of Hoeffding decomposition, it can be verified that $\{\xi_{ij}(y) - \mathbb{E}[\xi_{ij}(y)|X_j]\}e_j, 1 \le i \neq j \le n$, are uncorrelated so

$$\|\tilde{I}_{n,1}(y)\|^2 = \frac{1}{n^4}\sum_{i\neq j}\|\{\xi_{ij}(y) - \mathbb{E}[\xi_{ij}(y)|X_j]\}e_j\|^2. \tag{58}$$

Note that we have the following facts:

(i) $\xi_{ij}(y) - \mathbb{E}[\xi_{ij}(y)|X_j]$ and e_j are independent.

(ii) $\|\xi_{ij}(y) - \mathbb{E}[\xi_{ij}(y)|X_j]\| \leq \|\xi_{ij}(y)\| + \|\mathbb{E}[\xi_{ij}(y)|X_j]\| \leq 2\|\xi_{ij}(y)\|$ via the triangle inequality and Jensen's inequality $\|\mathbb{E}[\xi_{ij}(y)|X_j]\| \leq \|\xi_{ij}(y)\|$.

(iii) $|\xi_{ij}(y)| \leq c_1 K_{h_n}(X_i - X_j)$ for some constant c_2, in view of the boundedness of $f'_e(z|\theta)$ and $f_X(x) > c$ in Assumption 2(iv).

(iv) $\mathbb{E}K^2_{h_n}(X_i - X_j) = O(1/h_n)$ for $i \neq j$.

Combining the above facts with (58), we see that there exists constant c_3 such that

$$\|\tilde{I}_{n,1}(y)\|^2 \leq \frac{c_3}{n^4}\frac{n^2 - n}{h_n} = o(n^{-1}), \quad \text{for all } y \in \mathcal{Y}. \tag{59}$$

For the derivative $\tilde{I}'_{n,1}(y) = n^{-2}\sum_{j=1}^n \sum_{i=1}^n \{\xi'_{ij}(y) - \mathbb{E}[\xi'_{ij}(y)|X_j]\}e_j$, the same argument in (58)–(59) yields $\|\tilde{I}'_{n,1}(y)\|^2 = o(1/n)$. By Lemma 1(ii), the first assertion in (57) follows.

[proof of the second assertion in (57)]: For $\tilde{I}_{n,2}(y)$, we calculate

$$\begin{aligned} \mathbb{E}[\xi_{ij}(y)|X_j] &= \int_{-\infty}^{\infty} \frac{1}{h_n} K\left(\frac{x - X_j}{h_n}\right) f'_e(y - \mu(X_i)|\theta)dx \\ \left(\text{by transformation } u = \frac{x - X_j}{h_n}\right) &= \int_{-\infty}^{\infty} K(u) f'_e(y - \mu(X_j + uh_n)|\theta)du. \end{aligned} \tag{60}$$

In (60), applying Taylor's expansion $f'_e(y - \mu(X_j + uh_n)|\theta) = f'_e(y - \mu(X_j)|\theta) + f''_e(y - \mu(X_j)|\theta)\mu'(X_j)uh_n + O(u^2h_n^2)$, where all the derivatives here are uniformly bounded, and using the conditions in Assumption 2(i), we can obtain

$$\sup_{y \in \mathcal{Y}} |\mathbb{E}[\xi_{ij}(y)|X_j] - f'_e(y - \mu(X_j)|\theta)| = O(h_n^2), \quad \text{for all } i, j. \tag{61}$$

The second assertion in (57) follows from (61) under conditions $nh_n^4 \to 0$, $\mathbb{E}|e_j| < \infty$. ♢