An Omnibus Test of Goodness-of-fit for Conditional Distribution Functions with Applications to Regression Models

Gilles R. Ducharme, Sandie Ferrigno

An Omnibus Test of Goodness-of-fit for Conditional Distribution Functions with Applications to Regression Models

by

Gilles R. Ducharme* and Sandie Ferrigno**

* Équipe de Probabilités et Statistique, Institut de Mathématiques et de Modélisation de Montpellier (I3M) Université Montpellier II, Place Eugène Bataillon, 34095, Montpellier, Cedex 5, France (gilles.ducharme@univ-montp2.fr)

** Équipe de Probabilités et Statistique, Institut de Mathématiques Elie Cartan, Université Henri Poincaré Nancy 1, B.P. 70239, F-54506 Vandoeuvre-lès-Nancy, Cedex, France (sandie.ferrigno@iecn-unancy.fr)

Key-words: Conditional distribution function, Cramer-von Mises Statistic Goodness-of-fit test, Local polynomial estimation, Regression Model.

Abstract

We introduce a goodness-of-fit test for statistical models about the conditional distribution function of a random variable. This test is useful for assessing whether a regression model fits a data set regarding all assumptions made in its elaboration. The test is based on a generalization of the Cramer-von Mises statistic and involves a local polynomial estimator of the conditional distribution function. First, the uniform almost sure consistency of this estimator is established. Then, the asymptotic distribution of the test statistic is derived under the null hypothesis and local alternatives. The extension to the case where unknown parameters appear in the model is developed. Finally, a simulation study is performed to see how the test behaves with moderate samples. It emerges that, although the test can detect any departure from the null model, its power is comparable to that of other nonparametric tests designed to examine only specific departures.

1. Introduction

It is often the case in statistical applications that a model is entertained for the conditional distribution of a random variable Y on some random vector $\boldsymbol{X}$ (in the sequel, vectors are represented in italic boldface, matrices in bold and real quantities in plain italic). This may happen when the distribution of $\boldsymbol{X}$ is known or of secondary importance and the conditional cumulative distribution function (*cdf*)

$$F(y \mid \boldsymbol{x}) = \mathbb{P}\Big[Y \leq y \;\Big|\boldsymbol{X} = \boldsymbol{x}\Big],$$

which embodies all information about the relationship between $\boldsymbol{X}$ and Y, is the element that commands attention. It is then of importance to determine if the entertained model for $F(y \mid \boldsymbol{x})$ is consonant with data.

An important case where the above problem occurs is when a regression model is set up to predict the value of Y from that of $\boldsymbol{X}$. To this end a common *modus operandi* consists of selecting a parametric model relating $\boldsymbol{X}$ to Y from a catalogue, the choice being guided by more or less solid knowledge about the mechanism that relates $\boldsymbol{X}$ to Y. The parameters of this model are estimated and prediction proceeds from there.

Such parametric models usually involve many assumptions and together they induce a parametric form for $F(y \mid \boldsymbol{x})$ that must be validated in each application. As

an example, consider the problem of predicting the height Y of a human subject based on his age X. One model for this is:

$$Y = g(X; \boldsymbol{\theta}_1) + \sigma(X; \boldsymbol{\theta}_2)\varepsilon. \tag{1.1}$$

In addition to assumptions about the functional form of g and σ, this model requires that the error term ε be additive and, usually, independent of X. Often the distribution of ε is also assumed normal. These assumptions induce the parametric conditional *cdf* $F(y \mid x) = N(g(x; \boldsymbol{\theta}_1), \sigma^2(x; \boldsymbol{\theta}_2))$. When all assumptions hold, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ can be estimated in some optimal fashion and this confers to the statistical inference some principled merit in predicting Y for given values x of X.

Typically, in models like (1.1) the task of selecting the structural part g receives most of the attention because it contains visual information on the relationship between $\boldsymbol{X}$ and Y. In the human height example, the catalogue of possible g has grown very large (see Ducharme and Fontez, 2004 for some references). The selection of σ is, however, often done more casually: in human height a constant σ is usually taken, based on an argument that, to the best of our knowledge, has not been fully substantiated (Ducharme and Fontez, 2004). The situation is similar for the other assumptions; some do get attention (Q-Q plots are routinely produced to check the normality of ε) others, little or none at all. This may be tolerable when the ensuing inference is robust to some misspecifications or when an asymptotic argument renders these inconsequential. Otherwise, unsubstantiated assumptions or incorrect selections may have a bearing on the validity of the ensuing inference.

Methods that can assess whether a model is concordant with some data fall under the banner of goodness-of-fit (GoFIT) tests. For models as (1.1), many tests have been proposed to assess if the assumption about a functional form for g is consonant with data; for some examples, see Alcala *et al.* (1999), Ducharme and Fontez (2004), Van Keilegom *et al.* (2008). Some authors (Liero, 2003) have proposed tests for σ. GoFIT procedures for the normality of ε are discussed in D'Agostino and Stephens (1986). Note that many of these tests require, for their validity and for good power properties, that the other assumptions of the model be correct. Note also that these GoFIT tests are "directional" in the sense that they primarily detect departures from one assumption about the model.

Thus the need for an "omnibus" test that can detect any departures from a model for $F(y \mid \boldsymbol{x})$. If this model is $F_0(y \mid \boldsymbol{x})$, the GoFIT problem is that of testing

$$H_0\colon\ F(y \mid \boldsymbol{x}) = F_0(y \mid \boldsymbol{x}). \tag{1.2}$$

In this paper, such an omnibus test is developed. This test compares a nonparametric estimator of $F(y \mid \boldsymbol{x})$ with $F_0(y \mid \boldsymbol{x})$. Here, the nonparametric estimator $\hat{F}(y \mid \boldsymbol{x})$ is based on the local polynomial approach, which has been recognized as superior to classical estimates. The comparison of $\hat{F}(y \mid \boldsymbol{x})$ with $F_0(y \mid \boldsymbol{x})$ is done through a version of the generalized Cramer-von Mises statistic.

The paper is organized as follows. In Section 2, we present the local polynomial estimator $\hat{F}(y \mid \boldsymbol{x})$ and show its almost sure uniform convergence. In Section 3, we introduce the test statistic and obtain its asymptotic normality under (1.2) in the case where $F_0(y \mid \boldsymbol{x})$ is fully specified. Section 4 gives the local power properties of the test. Section 5 treats the case where $F_0(y \mid \boldsymbol{x})$ depends on unknown parameters and states that the asymptotic normality is not affected by the introduction of estimators in the test statistic. The results of a simulation study are presented in Section 6 and show that, surprisingly, the test yields powers comparable to other, more directional, tests. Section 7 contains all technical details.

2. The local polynomial estimator of $F(y \mid x)$

Let $(\boldsymbol{X}_i,\ Y_i)$, $i = 1,\ldots, n$, be independent and identically distributed (*iid*) random vectors with support $\mathcal{S} \times \mathbb{R}$ ($\mathcal{S} \subseteq \mathbb{R}^d$), joint density $f(\boldsymbol{x}, y) = f(y|\boldsymbol{x})f(\boldsymbol{x})$ and *cdf* $F(\boldsymbol{x}, y)$. The problem of estimating nonparametrically $f(\boldsymbol{x}, y)$, $f(\boldsymbol{x})$ and $F(\boldsymbol{x}, y)$ has been much investigated and the reader is referred to, among others, Prakasa Rao (1983). Here, we focus on the problem of estimating the conditional *cdf* $F(y|\boldsymbol{x})$. Work on this using the kernel method was initiated by Collomb (1980). However his method has recently lost momentum, at least for estimating functionals of $F(y \mid \boldsymbol{x})$ (as $\mathbb{E}(Y \mid \boldsymbol{x})$), and been progressively replaced by the design adaptive local polynomial approach. Following this trend, our task here is to produce a local polynomial estimator of $F(y \mid \boldsymbol{x})$ and investigate some of its properties in view of GoFIT testing.

This paragraph collects the notation for a concise expression of the Taylor expansion of a suitably regular function η in the neighborhood of some point $\boldsymbol{x}$. For $k \in \{0, 1,\ldots, p\}$, define the sets $\mathcal{J}_k = \{\boldsymbol{j} = (j_1,\ldots, j_d) \mid j_m \in \{0, 1,\ldots, k\}$ with $|\boldsymbol{j}| = j_1 + \ldots + j_d = k\}$. Write $\boldsymbol{j}! = j_1! \times\ldots\times j_d!$ and, for $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{x}^{\boldsymbol{j}} = x_1^{j_1} x_2^{j_2} \cdots x_d^{j_d}$ with $(\partial \boldsymbol{x})^{\boldsymbol{j}} = \partial x_1^{j_1} \partial x_2^{j_2} \cdots \partial x_d^{j_d}$. When x is real, interpret $x^{\boldsymbol{j}}$ as $x^{|\boldsymbol{j}|}$. The cardinal of $\mathcal{J}_k$ is $\mathcal{C}_k = (k + d - 1)!/((d - 1)!k!)$. Order the elements of $\mathcal{J}_k$ from "smallest" to "largest" according to the following rule: with $\boldsymbol{j}, \boldsymbol{j}' \in \mathcal{J}_k$, $\boldsymbol{j} < \boldsymbol{j}'$ if the concatenated string of characters $j_1\ldots j_d$ forms a number smaller than $j_1'\ldots j_d'$. Set $\bar{\mathcal{J}} = \bigcup_{k=0}^{p} \mathcal{J}_k$ and $\mathcal{P} = \sum_{k=0}^{p} \mathcal{C}_k$. The elements of $\bar{\mathcal{J}}$ are also ordered, first the sole element of $\mathcal{J}_0$, which is $\mathbf{0} \in \mathbb{R}^d$, then the

ordered elements of $\mathcal{J}_1$ and so on. When encountering expressions where it is implied that $\boldsymbol{j}$ runs through all elements of $\bar{\mathcal{J}}$, such as $Diag\{\boldsymbol{x}^{\boldsymbol{j}},\ \boldsymbol{j} \in \bar{\mathcal{J}}\}$, assume that $\boldsymbol{j}$ takes successively the values in $\bar{\mathcal{J}}$ ordered in this manner. The multivariate $(p+1)$-order Taylor expansion of $\eta(\boldsymbol{x}')$ in the neighborhood of $\boldsymbol{x}$ can be expressed as

$$\eta(\boldsymbol{x}') = \sum_{\boldsymbol{j}\in\bar{\mathcal{J}}} \frac{1}{\boldsymbol{j}!}(\boldsymbol{x}'-\boldsymbol{x})^{\boldsymbol{j}} \frac{\partial^{|\boldsymbol{j}|}}{(\partial\boldsymbol{x})^{\boldsymbol{j}}}\eta(\boldsymbol{x}) \ + \sum_{\boldsymbol{j}\in\mathcal{J}_{p+1}} \frac{1}{\boldsymbol{j}!}(\boldsymbol{x}'-\boldsymbol{x})^{\boldsymbol{j}} \frac{\partial^{p+1}}{(\partial\boldsymbol{x})^{\boldsymbol{j}}}\eta(\boldsymbol{x}^*), \tag{2.1}$$

for some $\boldsymbol{x}^*$ on the line segment joining $\boldsymbol{x}$ and $\boldsymbol{x}'$.

Next, we collect the notation needed to express the p-order local polynomial estimator of $F(y \mid \boldsymbol{x})$. Let $\mathbb{I}\{A\}$ be the indicator of event A pertaining to Y. Choose a *kernel* function K: $[-1,\ 1]^d \to \mathbb{R}_+$ and a *bandwidth* $h > 0$. Set $\eta(\boldsymbol{x}) = F(y \mid \boldsymbol{x})$ and replace each occurrence of $\partial^{|\boldsymbol{j}|}\big/(\partial\boldsymbol{x})^{\boldsymbol{j}}\, F(y \mid \boldsymbol{x})$ in the first part of the right-hand side of (2.1) by the parameter $\beta_{|\boldsymbol{j}|,\boldsymbol{j}}$. Collect these into $\boldsymbol{\beta} = \{\beta_{|\boldsymbol{j}|,\boldsymbol{j}},\ \boldsymbol{j} \in \bar{\mathcal{J}}\} \in \mathbb{R}^{\mathcal{P}}$. Consider

$$\hat{\boldsymbol{\beta}} \ = \ \underset{\boldsymbol{\beta}\in\mathbb{R}^{\mathcal{P}}}{\arg\min} \sum_{i=1}^{n} \left\{ \mathbb{I}\{Y_i \le y\} - \sum_{\boldsymbol{j}\in\bar{\mathcal{J}}} (\boldsymbol{X}_i - \boldsymbol{x})^{\boldsymbol{j}} \beta_{|\boldsymbol{j}|,\boldsymbol{j}} \right\}^2 K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right).$$

It is explained in Fan and Gijbels (1996) (see also Masry, 1996, for the case $d > 1$) that component $\hat{\beta}_{|\boldsymbol{j}|,\boldsymbol{j}}$ of $\hat{\boldsymbol{\beta}}$ is an estimate of $(\boldsymbol{j}!)^{-1}\, \partial^{|\boldsymbol{j}|}\big/(\partial\boldsymbol{x})^{\boldsymbol{j}}\, \mathbb{E}(\mathbb{I}\{Y \le y\} \mid \boldsymbol{x})$ and in particular $\hat{F}(y \mid \boldsymbol{x}) = \hat{\beta}_{0,\boldsymbol{0}}$ is an estimate of $F(y \mid \boldsymbol{x})$. Note that setting $p = 0$ in the above gives back Collomb's (1980) estimator.

An explicit expression for $\hat{\beta}_{|\boldsymbol{j}|,\boldsymbol{j}}$ can be obtained from a standard weighted least squares argument. Let $\mathbf{X}$ be the $n \times \mathcal{P}$ matrix with i-th line $\{(\boldsymbol{X}_i - \boldsymbol{x})^{\boldsymbol{j}},\ \boldsymbol{j} \in \bar{\mathcal{J}}\}$ and $\mathbf{W} = Diag\{K((\boldsymbol{X}_i - \boldsymbol{x})/h),\ i = 1,\ldots,\ n\}$. Set $\boldsymbol{e}_{|\boldsymbol{j}|,\boldsymbol{j}}$ as the $\mathcal{P}$-dimensional vector having a "1" at the position where $\beta_{|\boldsymbol{j}|,\boldsymbol{j}}$ appears in $\boldsymbol{\beta}$ and "0" elsewhere. Finally, put $\mathbb{I}_y = (\mathbb{I}\{Y_1 \le y\},\ \ldots,\ \mathbb{I}\{Y_n \le y\})^t$ (here ^{t} means transposition). Then, $\hat{\beta}_{|\boldsymbol{j}|,\boldsymbol{j}} = \boldsymbol{e}^t_{|\boldsymbol{j}|,\boldsymbol{j}}(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\mathbb{I}_y$ and the local polynomial estimator of $F(y \mid \boldsymbol{x})$ takes the form:

$$\hat{F}(y \mid \boldsymbol{x}) \ = \ \sum_{i=1}^{n} \mathcal{W}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right) \mathbb{I}\{Y_i \le y\}, \tag{2.2}$$

where, upon setting $\mathbf{1} = \sum_{\boldsymbol{j}\in\bar{\mathcal{J}}} \boldsymbol{e}_{|\boldsymbol{j}|,\boldsymbol{j}}$, $\mathbf{H} = Diag\{h^{\boldsymbol{j}},\ \boldsymbol{j} \in \bar{\mathcal{J}}\}$ and for $\boldsymbol{t} \in \mathbb{R}^d$, $\mathbf{T} = Diag\{\boldsymbol{t}^{\boldsymbol{j}},\ \boldsymbol{j} \in \bar{\mathcal{J}}\}$, we get $\mathcal{W}(\boldsymbol{t}) = \boldsymbol{e}^t_{0,\boldsymbol{0}}(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{H}\mathbf{T}\mathbf{1}K(\boldsymbol{t}) = \boldsymbol{e}^t_{0,\boldsymbol{0}}\mathbf{S}_n^{-1}(\boldsymbol{x})\,\mathbf{T}\mathbf{1}K(\boldsymbol{t})\,/\,nh^d$ with

$$\mathbf{S}_n(\boldsymbol{x}) = \mathbf{H}^{-1}\Big((nh^d)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{X}\Big)\mathbf{H}^{-1}. \tag{2.3}$$

Note for further use that for all $\boldsymbol{x} \in \mathcal{S}$

$$\sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{x})^{j} \mathcal{W}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right) = \begin{cases} 1 & \text{if } \left|\boldsymbol{j}\right| = 0 \\ 0 & \text{if } 1 \leq \left|\boldsymbol{j}\right| \leq p \end{cases}, \tag{2.4}$$

so that $\hat{F}(-\infty \mid \boldsymbol{x}) = 0$ and $\hat{F}(\infty \mid \boldsymbol{x}) = 1$. However $\hat{F}(y \mid \boldsymbol{x})$ may not be a proper *cdf* because it could decrease or take negative values, the weight function $\mathcal{W}$ in (2.2) not being necessarily positive. This mainly happens when there are no data in some sizeable parts of $\mathcal{S}$, an event whose probability decreases as n increases. The approaches in Hall *et al.* (1999) could be used to correct this, but in view of the added computational cost involved, we will not pursue such a refinement further.

Now $\mathcal{W}$ in (2.2) is a random function and this causes difficulties in studying the statistical properties of $\hat{F}(y \mid \boldsymbol{x})$. Ducharme and Mint el Mouvid (2001) have attacked the problem directly using the *U*-statistic structure of $\hat{F}(y \mid \boldsymbol{x})$ when $d = 1$, but their approach seems difficult to extend. However, Fan and Gijbels (1996, p. 64) explain that $\mathcal{W}$ can be approximated by an *equivalent kernel* (Huang and Fan, 1999). As shown in Masry (1996), this indirect approach appears easier when $d > 1$. Let

$$\mathcal{K}(\boldsymbol{t}) = e_{0,\boldsymbol{0}}^{t} \mathbf{S}^{-1} \mathbf{T1} K(\boldsymbol{t}) \tag{2.5}$$

where $\mathbf{S}$ is the $\mathcal{P} \times \mathcal{P}$ matrix with element $s_{\boldsymbol{j},\boldsymbol{j}'} = \int \boldsymbol{u}^{\boldsymbol{j}+\boldsymbol{j}'} K(\boldsymbol{u}) d\boldsymbol{u}$, $\boldsymbol{j}, \boldsymbol{j}' \in \bar{\mathcal{J}}$ that is assumed invertible throughout. Note that $\mathcal{K}(\boldsymbol{t})$ satisfies a theoretical version of (2.4):

$$\int \boldsymbol{u}^{j} \mathcal{K}(\boldsymbol{u}) d\boldsymbol{u} = \begin{cases} 1 & \text{if } \left|\boldsymbol{j}\right| = 0 \\ 0 & \text{if } 1 \leq \left|\boldsymbol{j}\right| \leq p \end{cases}.$$

Write

$$\mathcal{W}_{eq}(\boldsymbol{t}) = (nh^{d} f(\boldsymbol{x}))^{-1} \mathcal{K}(\boldsymbol{t}), \tag{2.6}$$

and define

$$\tilde{F}(y \mid \boldsymbol{x}) = \sum_{i=1}^{n} \mathcal{W}_{eq}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right) \mathbb{I}\left\{Y_i \leq y\right\}.$$

We now collect the "global" assumptions needed throughout, to which more specific assumptions will be added as needed. In the sequel, $\boldsymbol{x}$ is confined to a compact subset $\mathcal{C}$ of $\mathcal{S}$.

Assumption K: K is a continuous, bounded function on $[-1, 1]^d$ whose first order derivatives are also bounded.

Assumption X: The boundary of $\mathcal{C}$ is disjoint from that of $\mathcal{S}$ which is compact. Furthermore, at all points $\boldsymbol{x}$ in $\mathcal{C}$, the density f of $\boldsymbol{X}$ exists with $0 < \delta \leq f(\boldsymbol{x}) \leq \Delta < \infty$ and has a bounded continuous derivative.

Assumption H: *a*) As $n \to \infty$, $h \to 0$ in such a way that $\log n \big/ nh^d \to 0$.
b) $nh^{2p+2+d/2} \to 0$.

Note that the both parts of *Assumption H* impose an unwanted relationship between d and p that seems unavoidable with the tools used here. But, the ensuing condition on p is mild in practice because $p = 1$ can be taken as long as $d \leq 7$.

Asumption F: *a*) For all $\boldsymbol{x} \in \mathcal{C}$, $F(y \mid \boldsymbol{x})$ is differentiable in y with a bounded continuous conditional density $f(y \mid \boldsymbol{x})$. For each y, $F(y \mid \boldsymbol{x})$ possesses a $(p+1)$-order Taylor expansion as in (2.1) in a neighborhood $\mathcal{V}(\boldsymbol{x})$.

$$b) \text{ For all } \boldsymbol{j} \in \mathcal{J}_{p+1}, \quad \sup_{\boldsymbol{x} \in \mathcal{C}} \sup_{y \in \mathbb{R}} \left| \frac{\partial^{p+1} F(y \mid \boldsymbol{x})}{(\partial \boldsymbol{x})^{\boldsymbol{j}}} \right| \leq \mathcal{M} < \infty.$$

We can now relate $\tilde{F}(y \mid \boldsymbol{x})$ to $\hat{F}(y \mid \boldsymbol{x})$ through the following result, proved in Subsection 7.1. The first statement is a slight, but crucial, refinement of Lemma 2.1 in Huang and Fan (1999). Here and throughout the paper, the generic symbol $\mathcal{O}_{as}(1)$ represents an almost surely bounded random variable.

Theorem 2.1: *Under Assumptions K, X, H–a) and F–a), as* $n \to \infty$,

$$\sup_{\boldsymbol{x} \in \mathcal{C}} \sup_{y \in \mathbb{R}} \left| \hat{F}(y \mid \boldsymbol{x}) - \tilde{F}(y \mid \boldsymbol{x}) \right| = h \mathcal{O}_{as}(1). \tag{2.7}$$

As a consequence,

$$\sup_{\boldsymbol{x} \in \mathcal{C}} \sup_{y \in \mathbb{R}} \left| \hat{F}(y \mid \boldsymbol{x}) - F(y \mid \boldsymbol{x}) \right| \xrightarrow{a.s.} 0. \tag{2.8}$$

Ferrigno and Ducharme (2008) give the optimal bandwidth for (2.2), in the mean integrated squared error sense, when $d = 1$ and their results can be extended to $d > 1$. In the present GoFIT context however, one should not be too uncompromising about this optimal h because there appears to be no guarantee that it will work well in test settings, where optimality refers to power properties. We will get back to this in Section 6.

3. The test statistic and its behavior under H_0

Consider the weighted L_2 distance between $\hat{F}(y \mid \boldsymbol{x})$ of (2.2) and $F_0(y \mid \boldsymbol{x})$:

$$T_n(\boldsymbol{x}) = n\int_{\mathbb{R}} \left(\hat{F}(y \mid \boldsymbol{x}) - F_0(y \mid \boldsymbol{x})\right)^2 f_0(y \mid \boldsymbol{x}) w_{\boldsymbol{x}}(y) dy\,. \tag{3.1}$$

For the problem of testing (1.2), a test statistic is

$$T_n = \sqrt{h^d}\int_{\mathcal{C}} T_n(\boldsymbol{x}) w(\boldsymbol{x}) d\boldsymbol{x}\,. \tag{3.2}$$

In (3.1), (3.2), $w(\boldsymbol{x})$ and $w_{\boldsymbol{x}}(y)$ are user-supplied positive weight functions whose role is to direct power toward those alternatives that are believed more plausible. Test statistic (3.1) is the generalized Cramer-von Mises statistic (D'Agostino and Stephens, 1986, p. 100) applied to conditional *cdfs* with $\boldsymbol{x}$ fixed while (3.2) is an integrated (over the domain of scrutiny $\mathcal{C}$) version.

To build a test from (3.2), the behavior of T_n under H_0 is required. Set for conciseness $w^*(\boldsymbol{x}, y) = f_0(y \mid \boldsymbol{x})\, w(\boldsymbol{x}) w_{\boldsymbol{x}}(y)$ and add the following:

Assumption W: $w^*(\boldsymbol{x}, y)$ is bounded over $\mathcal{C} \times \mathbb{R}$ with $\int_{\mathbb{R}} w^*(\boldsymbol{x}, y) dy$ bounded over $\mathcal{C}$.

Define $a_{\mathcal{K}}(w^*)$ and $\sigma^2_{\mathcal{K}}(w^*)$ according to (7.5), (7.6) respectively. We state the main result of this section, whose proof is given in Subsection 7.2.

Theorem 3.1: *Suppose Assumptions K, X, H, W and suppose F_0 satisfies Assumption F. Then when H_0 is true,*

$$\frac{T_n - h^{-d/2} a_{\mathcal{K}}(w^*)(1 + h\mathcal{O}_{as}(1))}{\sigma_{\mathcal{K}}(w^*)} \xrightarrow{L} N(0, 1). \tag{3.3}$$

Note that $a_{\mathcal{K}}(w^*)$ and $\sigma^2_{\mathcal{K}}(w^*)$ involve f. When unknown, this quantity must be estimated, typically by a nonparametric density estimator involving a different bandwidth $\tilde{h}$. This leads to the estimators $\hat{a}_{\mathcal{K}}(w^*)$, $\hat{\sigma}^2_{\mathcal{K}}(w^*)$. When $d = 1$, (3.3) can be replaced by $(T_n - h^{-1/2}\hat{a}_{\mathcal{K}}(w^*)) \big/ \hat{\sigma}_{\mathcal{K}}(w^*)$ and it can be shown that the optimal rate $\tilde{h} = O(n^{-1/5})$ still leads to the $N(0, 1)$. When $d > 1$ however, $h^{-d/2} a_{\mathcal{K}}(w^*)(1 + h\mathcal{O}_{as}(1))$ is not approximated by $h^{-d/2}\hat{a}_{\mathcal{K}}(w^*)$ and it appears better to use another estimator of $\mathbb{E}(T_n)$ (or $\mathbb{V}(T_n)$) for example by bootstrapping as explained in Remark 7.6. This requires a moderate number of bootstrap replications. Another possibility is to bootstrap the whole distribution of T_n but this requires more calculations.

The above test calls for rejecting H_0 for large values of T_n. This obviously yields a consistent test strategy against fixed alternatives. The next section investigates the behavior of T_n under contiguous alternatives.

4. Asymptotic power

Consider the sequence of local alternatives

$$H_{1,n}\colon\ F_{1,n}(y \mid x) \ =\ F_0(y \mid x) \ +\ c_n R(y \mid x),$$

where $R(y|x) = \int_{-\infty}^{y} r_x(t)\,dt$ and $r_x(y)$ is such that $f_{1,n}(y|x) = f_0(y|x) + c_n r_x(y)$ is a positive bounded density. Because the present investigation is mainly for theoretical insight, we consider the simplest context and assume that $\mathcal{S} \subseteq \mathbb{R}$ with $w(x) \equiv w_x(y) \equiv 1$. Assume further that on $\mathcal{S}$, $F_{1,n}(y \mid x)$ is a polynomial in x of degree $\leq p$ satisfying *Assumption F*. Let $(X_{i,n},\ Y_{i,n})$, $i = 1,..,\ n$, $n \geq 1$, be a triangular array of row-wise *iid* random vector with joint *cdf* $F_{1,n}(x,\ y)$. We may as well consider that the marginal density f of the $X_{i,n}$ does not vary with n. All these simplifying assumptions can be lifted at the expense of a more tedious argument. In view of (2.4), (7.4), this entails

$$\hat{F}(y \mid x) - F_{1,n}(y \mid x) \ =\ (1 + h\mathcal{O}_{as}(1))\sum_{i=1}^{n} \mathcal{W}_{eq}\left(\frac{X_{i,n} - x}{h}\right)\Big(\mathbb{I}\{Y_{i,n} \leq y\} - F_{1,n}(y \mid X_{i,n})\Big). \qquad (4.1)$$

Finally put $c_n = (n\sqrt{h})^{-1/2}$. Then, we have the following.

Theorem 4.1: *Under $H_{1,n}$ and Assumptions K, F, X, H, and W, to which we add $nh^{3/2} \to \infty$, we have*

$$\frac{T_n - h^{-1/2} a_{\mathcal{K}}(f_0)}{\sigma_{\mathcal{K}}(f_0)} \xrightarrow{\ L\ } N\left(\frac{b_R(f_0)}{\sigma_{\mathcal{K}}(f_0)},\, 1\right), \qquad (4.2)$$

where $b_R(f_0) = \iint \Big(R(y \mid x)\Big)^2 f_0(y \mid x)dydx \ \geq 0.$

5. The case of unknown parameters

The methods of the previous sections apply to the case where $F_0(y\,|\boldsymbol{x})$ is entirely specified. This is rare in applications where usually, as in (1.1), the conditional *cdf* involves unknown parameters $\boldsymbol{\theta}$ that must be estimated. We now study how this added layer of complexity affects the behavior of our test.

First note that in this context, (1.2) evolves into the composite null hypothesis

$$H_0\text{: } F(y\,|\boldsymbol{x}) = F_{\boldsymbol{\theta}_0}(y\,|\,\boldsymbol{x}) \text{ for some } \boldsymbol{\theta}_0 \in \Theta \tag{5.1}$$

We suppose that f does not depend on $\boldsymbol{\theta}$ and change *Assumption F* into:

Asumption F_{θ}: For all $\boldsymbol{\theta}$ in a neighborhood $\mathcal{V}(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ and all $\boldsymbol{x} \in \mathcal{C}$, $F_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$ is continuously differentiable in y with conditional density $f_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$. In turn, $f_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$ is continuously differentiable with respect to $\boldsymbol{\theta}$ in $\mathcal{V}(\boldsymbol{\theta}_0)$ and there exist a function $h_{\boldsymbol{x}}(y)$ such that $\int_{\mathbb{R}} h_{\boldsymbol{x}}(y)dy < \mathcal{M}$ for all $\boldsymbol{x} \in \mathcal{S}$ and such that for each $\boldsymbol{\theta}$ in $\mathcal{V}(\boldsymbol{\theta}_0)$ and each $y \in \mathbb{R}$, $\left|\partial f_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})/\partial\boldsymbol{\theta}\right| \le h_{\boldsymbol{x}}(y)$. Also, for each $\boldsymbol{\theta} \in \mathcal{V}(\boldsymbol{\theta}_0)$, $y \in \mathbb{R}$, $\boldsymbol{x} \in \mathcal{C}$, $F_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$ has a $(p+1)$-order Taylor expansion in a neighborhood of $\boldsymbol{x}$. Moreover, for all $\boldsymbol{j} \in \mathcal{J}_{p+1}$,

$$\sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \sup_{\boldsymbol{x}\in\mathcal{C}} \sup_{y\in\mathbb{R}} \left|\frac{\partial^p F_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})}{(\partial\boldsymbol{x})^{\boldsymbol{j}}}\right| \le \mathcal{M}.$$

Finally for each $\boldsymbol{x} \in \mathcal{C}$, $F_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ in $\mathcal{V}(\boldsymbol{\theta}_0)$ and these derivatives are bounded on $\mathcal{V}(\boldsymbol{\theta}_0)$.

To which we add:

Assumption E: The estimator $\hat{\boldsymbol{\theta}}$ is $\sqrt{n}$ -consistent for $\boldsymbol{\theta}_0$.

With these, test statistic (3.2) becomes

$$T_n(\hat{\boldsymbol{\theta}}) = nh^{d/2}\int\int\left(\hat{F}(y \mid \boldsymbol{x}) - F_{\hat{\boldsymbol{\theta}}}(y \mid \boldsymbol{x})\right)^2 w(\boldsymbol{x})w_{\boldsymbol{x}}(y)f_{\hat{\boldsymbol{\theta}}}(y \mid \boldsymbol{x})\,dyd\boldsymbol{x}\,. \tag{5.2}$$

Theorem 5.1: *Under Assumptions K, X, H,* F_{θ}, *W and E, and under* (5.1), *we have*

$$\frac{T_n(\hat{\boldsymbol{\theta}}) - h^{-d/2}a_{\hat{\boldsymbol{\theta}},\mathcal{K}}(f_{\hat{\boldsymbol{\theta}}})(1 + h\mathcal{O}_{as}(1))}{\sigma_{\hat{\boldsymbol{\theta}},\mathcal{K}}(f_{\hat{\boldsymbol{\theta}}})} \xrightarrow{\;L\;} N\big(0,\,1\big). \tag{5.3}$$

where $a_{\theta,\mathcal{K}}(f_{\theta})$, $\sigma^2_{\theta,\mathcal{K}}(f_{\theta})$ *have the form* (7.5), (7.6) *respectively, with* F_{θ} *replacing* F_0.

The proof of this is given in Subsection 7.4. The resulting test strategy for (5.1) rejects H_0 when the left hand side of (5.3) is greater that $z_{1-\alpha}$, the $(1 - \alpha)$-th quantile of the $N(0, 1)$ distribution.

Note that it is possible to show, with added work, that a similar results holds when the weight function $w_{\boldsymbol{x}}(y)$ also depends on $\boldsymbol{\theta}$. Finally, as explained at the end of Section 3, the procedure can be adapted to the case where f is unknown ($d = 1$).

Otherwise bootstrapping (first from the empirical *cdf* of the $\boldsymbol{X}_i$ and then from $F_{\hat{\theta}}(y \mid \boldsymbol{x})$) can be used to get better estimates of $\mathbb{E}(T_n(\hat{\boldsymbol{\theta}})$, $\mathbb{V}(T_n(\hat{\boldsymbol{\theta}})$. Thus the test procedure can be adapted to contexts encountered in practical situations.

6. Simulations

The test of the previous section is based on an asymptotic approximation. It is of interest to check if this asymptotic behavior holds for moderate sample sizes and to study power against various departures from the null hypothesis.

As a first experiment, we have replicated the first simulation in Alcala *et al.* (1999) who develop a GoFIT test, based on a local polynomial estimator, for the structural part g in model (1.1). Their null and alternative hypothesis are H_0: $Y = \theta X + \varepsilon$, H_1: $Y = \theta X + aX^2 + \varepsilon$, where $\theta = 5$, $a = 1, 5$, $\varepsilon \sim N(0, \sigma^2)$ with $\sigma^2 = 1, 2, 3$ and independent from X. As for the X, Alcala *et al.* (1999) do not need the constraint in *Assumption* X and use $\mathcal{C} = \mathcal{S} = [0, 1]$ with $X \sim U(0, 1)$. In our context, $\mathcal{C}$ is constrained and to allow comparisons with their results, we have taken $\mathcal{C} = [0, 1]$ with $X \sim U(-0.1, 1.1)$. Sample sizes were $n = 50$, 100 and $p = 1$, a local linear smoother. Alcala *et al.* (1999) do not specify the kernel used in their test. We have taken the Epanechnikov kernel, after observing similar results with other kernels. As for the bandwidth, Alcala *et al.* (1999) consider $h = 0.1$, 0.25 and see little effects on their test. Thus we have taken the "reasonable" values $h = 0.20$ for $n = 50$ and $h = 0.12$ for $n = 100$. The parameters θ, a and σ^2 were estimated by least squares.

Alcala *et al.* (1999) do not specify their weight function but, for ease of computation, we have taken $w = w_x \equiv 1$. One advantage of this choice is that test statistic $T_n(x)$ has the more explicit form:

$$T_n(x)/n = 1/3 + \sum_{i=1}^{n-1} (\hat{F}(y_{(i)} \mid x))^2 \left[F_{\hat{\theta}}(y_{(i+1)} \mid x) - F_{\hat{\theta}}(y_{(i)} \mid x) \right]$$

$$- \sum_{i=1}^{n-1} (\hat{F}(y_{(i)} \mid x)) \left[F^2_{\hat{\theta}}(y_{(i+1)} \mid x) - F^2_{\hat{\theta}}(y_{(i)} \mid x) \right] + \left[F^2_{\hat{\theta}}(y_{(n)} \mid x) - F_{\hat{\theta}}(y_{(n)} \mid x) \right].$$

The simulation was conducted as follows. For each combinations of (n, a, σ^2), 1000 samples were generated from the above H_0, H_1 and the percentage of rejections at level $\alpha = 5\%$ was recorded. Results are shown in Table 6.1 along with the powers reported in Alcala *et al.* (1999) in parenthesis. The shaded values correspond to the actual levels of our tests.

Regarding levels, it is seen that the asymptotic normal is relatively accurate. More interesting is that the power of our omnibus test is in most cases comparable and often better than the "directional" test of Alcala *et al.* (1999). Thus even if the misspecification concerns only the structural part g, there is some advantage to use our test over this competitor.

Table 6.1: Percentage of rejection of H_0 at level $\alpha = 5\%$, based on 1000 replications. In parenthesis, the powers of Alcala *et al.* (1999). Shaded, the actual levels of our tests.

σ^2	a	n	% of rejection
1	0	50	4.9
		100	4.4
1	1	50	9.1 (8.7)
		100	9.0 (18.1)
1	5	50	90.7 (76.1)
		100	99.1 (96.8)
2	0	50	4.7
		100	4.3
2	1	50	6.9 (7.2)
		100	6.9 (12.5)
2	5	50	58.6 (53.6)
		100	86.1 (84.8)
3	0	50	4.6
		100	4.5
3	1	50	6.4 (6.9)
		100	5.9 (8.4)
3	5	50	42.1 (35.2)
		100	69.9 (68.1)

To investigate some other types of misspecification, a second experiment was conducted, inspired from the first simulation in Van Keilegom *et al.* (2008). Here the data is generated from either one of the models: H_0: $Y = \theta X + \varepsilon$; H_0^*: $Y = \theta X + \sigma(X)\varepsilon$; H_1^*: $Y = \theta X + a(X) + \sigma(X)\varepsilon$, independent X, ε and H_1^{**}: $Y = \theta X + a(X) + \sigma(X)\varepsilon$, correlated X, ε ($\rho = 0.3$). In these, $\theta = 1$ and $\varepsilon \sim N(0, 1)$. As for the functions $a(\bullet)$, $\sigma(\bullet)$, we followed Van Keilegom *et al.* (2008) and took $a(x) = x\exp(x)/2$, $0.3\sin(4\pi x)$ with $\sigma(x) = (1+X)/5$. The parameters θ and $\mathbb{V}(\varepsilon)$ were estimated by standard (for H_0) and weighted (for H_0^*) least squares. The rest of the settings were as in our first experiment. Table 6.2 gives the results along with the powers of T_{CM}, the best test (among four) of H_0^* against H_1^* investigated by Van Keilegom *et al.* (2008).

The shaded areas in Table 6.2 correspond to pairs of identical null and alternative hypotheses and serve as a check on the accuracy of the levels of the test, which are reasonably close to nominal. For a more precise control of the type 1 error, bootstrapping could be applied.

Regarding power, our omnibus test is again generally better than their directional test for H_0^* against H_1^*. Note that the departure from independence between X and ε (e.g. H_0, H_0^* against H_1^{**} with $a(x) = 0$) gets detected with some accuracy despite the correlation between X and ε being rather low.

Table 6.2: Percentage of rejection of H_0, H_0^* at level $\alpha = 5\%$, based on 1000 replications. In parenthesis, the results of Van Keilegom *et al.* (2008).

Null Hypothesis	$a(x)$	n	% of rejection H_1^*	% of rejection H_1^{**}
H_0	0	50	5.3	23.1
		100	6.4	38.2
	$x\exp(x)/2$	50	63.4	93.5
		100	88.1	99.8
	$0.3\mathrm{Sin}(4\pi x)$	50	16.1	22.7
		100	99.3	99.6
H_0^*	0	50	4.8	21.7
		100	4.1	32.5
	$x\exp(x)/2$	50	54.2 (35.0)	89.9
		100	76.4 (54.7)	99.4
	$0.3\mathrm{Sin}(4\pi x)$	50	18.3 (30.6)	21.4
		100	99. 3(77.1)	99.5

Thus as a general conclusion, the omnibus test of the paper can detect departures from any assumptions made in setting up a model and this, while paying a lower premium in power than the above directional competitors when the model is misspecified on its structural part. Thus, in testing the GoFIT of models like (1.1), there is generally much to gain by using the present approach.

7. Technical details

Throughout this section, the symbol $\mathcal{M}$ denotes a positive majoring constant that may change from one occurrence to another.

7.1 Proof of Theorem 2.1

The components of matrix $\mathbf{S}_n(\boldsymbol{x})$ in (2.3) have the form:

$$s_{n,\boldsymbol{j}}(\boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right)^{\boldsymbol{j}} K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right),$$

for $\boldsymbol{j} \in \bigcup_{k=0}^{2p}\mathcal{J}_k$, whereas those of matrix $\mathbf{S}$ in (2.5) are $\boldsymbol{\mu}_{\boldsymbol{j}} = \int \boldsymbol{u}^{\boldsymbol{j}} K(\boldsymbol{u})d\boldsymbol{u}$.

The following lemma is a version of Corollary 1 *ii*) of Masry (1996), adapted to the case of *iid* $(\boldsymbol{X}_i,\ Y_i)$.

Lemma 7.1: *Under Assumptions K, X and H–a), as* $n \to \infty$

$$\sup_{\boldsymbol{x}\in\mathcal{C}}\left|s_{n,\boldsymbol{j}}(\boldsymbol{x}) - \boldsymbol{\mu}_{\boldsymbol{j}} f(\boldsymbol{x})\right| = \left(\left(\frac{\log n}{nh^d}\right)^{1/2} + h\right)\mathcal{O}_{as}(1)\,,$$

and the optimal rate of convergence is obtained when $h \sim (\log n/n)^{1/(d+2)}$.

Lemma 7.2: *Under the assumptions of Theorem* 2.1,

$$\sup_{\boldsymbol{x}\in\mathcal{C}}\sup_{y\in\mathbb{R}}\left|\tilde{F}(y \mid \boldsymbol{x}) - F(y \mid \boldsymbol{x})\right| \xrightarrow{\ as\ } 0.$$

Proof. Obviously,

$$\left|\tilde{F}(y \mid \boldsymbol{x}) - F(y \mid \boldsymbol{x})\right| \le \left|\tilde{F}(y \mid \boldsymbol{x}) - \mathbb{E}(\tilde{F}(y \mid \boldsymbol{x}))\right| + \left|\mathbb{E}(\tilde{F}(y \mid \boldsymbol{x})) - F(y \mid \boldsymbol{x})\right|. \qquad (7.1)$$

An application of Theorem 2.1.1 of Prakasa Rao (1983, p. 35) to the uniformly continuous (in $\boldsymbol{x} \in \mathcal{C},\ y \in \mathbb{R}$) function $F(y \mid \boldsymbol{x})f(\boldsymbol{x})$ shows that the last term on the right-hand side of (7.1) is $o(1)$ as $h \to 0$. As for the first term, we have

$$\sup_{\boldsymbol{x}\in\mathcal{C}}\sup_{y\in\mathbb{R}}\left|\tilde{F}(y \mid \boldsymbol{x}) - \mathbb{E}(\tilde{F}(y \mid \boldsymbol{x}))\right| \le \delta^{-1}\sup_{\boldsymbol{x}\in\mathcal{C}}\sup_{y\in\mathbb{R}}\left|\tilde{F}_{num}(y \mid \boldsymbol{x}) - \mathbb{E}(\tilde{F}_{num}(y \mid \boldsymbol{x}))\right|,$$

where $\tilde{F}_{num}(y \mid \boldsymbol{x})$ is the numerator of $\tilde{F}(y \mid \boldsymbol{x})$:

$$\tilde{F}_{num}(y \mid \boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n}\mathcal{K}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right)\mathbb{I}\left\{Y_i \le y\right\}.$$

Now, we follow closely the first part of the proof of Theorem 3.1.7 p. 185 of Prakasa Rao (1983). For $\varepsilon > 0$, choose m points $\boldsymbol{z}_1,\ldots,\ \boldsymbol{z}_m$ such that for any $\boldsymbol{x} \in \mathcal{C}$, $\inf_k \left\|\boldsymbol{x} - \boldsymbol{z}_k\right\| < b_n = \varepsilon h^{d+1} / 8C$, where C is the Lipschitz constant of $\mathcal{K}$. It is easy to see that $m = O(h^{-d(d+1)})$. Then,

$$\mathbb{P}\left[\sup_{\boldsymbol{x}\in\mathcal{C}}\sup_{y\in\mathbb{R}}\left|\tilde{F}_{num}(y \mid \boldsymbol{x}) - \mathbb{E}(\tilde{F}_{num}(y \mid \boldsymbol{x}))\right| \ge \varepsilon\right] \le$$

$$m\max_{1\le k\le m}\left\{\mathbb{P}\left[\sup_{\left\|\boldsymbol{x}-\boldsymbol{z}_k\right\|\le b_n}\sup_{y\in\mathbb{R}}\left|\tilde{F}_{num}(y \mid \boldsymbol{x}) - \tilde{F}_{num}(y \mid \boldsymbol{z}_k)\right| \ge \frac{\varepsilon}{3}\right]\right.$$

$$+\ \mathbb{P}\left[\sup_{\|\boldsymbol{x}-\boldsymbol{z}_k\|\le b_n}\ \sup_{y\in\mathbb{R}}\left|\mathbb{E}(\tilde{F}_{num}(y\mid\boldsymbol{x}))-\mathbb{E}(\tilde{F}_{num}(y\mid\boldsymbol{z}_k))\right|\ge\frac{\varepsilon}{3}\right]$$

$$+\ \mathbb{P}\left[\sup_{y\in\mathbb{R}}\left|(\tilde{F}_{num}(y\mid\boldsymbol{z}_k))-\mathbb{E}(\tilde{F}_{num}(y\mid\boldsymbol{z}_k))\right|\ge\frac{\varepsilon}{3}\right]\Bigg]. \qquad (7.2)$$

But $\sup_{\|\boldsymbol{x}-\boldsymbol{z}_k\|\le b_n}\ \sup_{y\in\mathbb{R}}\left|\tilde{F}_{num}(y\mid\boldsymbol{x})-\tilde{F}_{num}(y\mid\boldsymbol{z}_k)\right|\ \le\ \frac{1}{h^d}\ \sup_{\substack{\|\boldsymbol{x}-\boldsymbol{z}_k\|\le b_n\\ \boldsymbol{w}\in S_X}}\left|\mathcal{K}\left(\frac{\boldsymbol{w}-\boldsymbol{x}}{h}\right)-\mathcal{K}\left(\frac{\boldsymbol{w}-\boldsymbol{z}_k}{h}\right)\right|\ \le \varepsilon/8,$

so that the first probability in the right hand side of (7.2) is zero. Similarly, Jensen's inequality shows that the second probability also vanishes. Now, we follow the argument in Pollard (1984), p. 14. Let $(\boldsymbol{X}_{n+i},Y_{n+i})$, $i = 1,\ldots, n$, be another *iid* sample from $F(\boldsymbol{x}, y)$, independent from the first and let E_1 ,..., E_n by *iid* Rademacher random variables independent from the "double sample" $\boldsymbol{\varphi} = \{(\boldsymbol{X}_i,\ Y_i),\ i = 1,\ldots, 2n\}$. By Chebyshev's inequality, for any $y \in \mathbb{R}$ and for n large enough,

$$\mathbb{P}\left[\left|\tilde{F}_{num}(y\mid\boldsymbol{z}_k))-\mathbb{E}(\tilde{F}_{num}(y\mid\boldsymbol{z}_k))\right|\le\frac{\varepsilon}{2}\right]\ge\ \frac{1}{2},$$

so that by a first and second symmetrization as in Pollard (1984), p. 15, we get

$$\mathbb{P}\left[\sup_{y\in\mathbb{R}}\left|(\tilde{F}_{num}(y\mid\boldsymbol{z}_k))-\mathbb{E}(\tilde{F}_{num}(y\mid\boldsymbol{z}_k))\right|\ge\frac{\varepsilon}{3}\right]\ \le\ 4\,\mathbb{P}\left[\sup_{y\in\mathbb{R}}\left|\sum_{i=1}^{n}E_iR_{ik}(y)\right|\ge\frac{nh^d\varepsilon}{12}\right]\Bigg],$$

where $R_{ik}(y) = \mathcal{K}\Big(\Big(\boldsymbol{X}_i-\boldsymbol{z}_k\Big)\Big/h\Big)\mathbb{I}\{Y_i\le y\}$. If we condition on $\boldsymbol{\varphi}$, the $R_{ik}(y)$ can take at most $n + 1$ different values, say $t_0,\ldots, t_n$, because the indicator function is constant between each pair of adjacent Y_i. Thus

$$\mathbb{P}\left[\sup_{y\in\mathbb{R}}\left|\sum_{i=1}^{n}E_iR_{ik}(y)\right|\ge\frac{nh^d\varepsilon}{12}\ \Big|\ \boldsymbol{\varphi}\right]\ =\ \mathbb{P}\left[\max_{0\le i'\le n}\left|\sum_{i=1}^{n}E_iR_{ik}(t_{i'})\right|\ge\frac{nh^d\varepsilon}{12}\ \Big|\ \boldsymbol{\varphi}\right],$$

$$\le\ (n+1)\max_{0\le i'\le n}\mathbb{P}\left[\left|\sum_{i=1}^{n}E_iR_{ik}(t_{i'})\right|\ge\frac{nh^d\varepsilon}{12}\ \Big|\ \boldsymbol{\varphi}\right].$$

At this stage, a standard idea is to apply some variant of Bernstein's inequality. However, this leads to a bound of the unwanted order $\exp\{-nh^{2d}O(1)\}$. Thus a finer argument must is required. To this end, note that $\mathbb{E}\Big(E_iR_{ik}(t_{i'})\Big)\ =\ 0$, $\Big(E_iR_{ik}(t_{i'})\Big)^2\ \le$ $\mathcal{K}^2\Big(\Big(\boldsymbol{X}_i-\boldsymbol{z}_k\Big)\Big/h\Big)$ which is independent of i' while, for $l \ge 2$,

$$\Big(\mathcal{K}\Big(\Big(\boldsymbol{X}_{i}-\boldsymbol{z}_{k}\Big)\Big/h\Big)\Big)^{l} \;\le\; C_{1}^{l-2}R_{ik}^{2}(\infty) \text{ where } C_{1} \ge \sup_{\boldsymbol{x}}\Big|\mathcal{K}(\boldsymbol{x})\Big|.$$

On one hand, for any $0 < t < C_{1}^{-1}$ we have,

$$\mathbb{E}\Big(\exp\{tE_{i}R_{ik}(t_{i'})\Big) \;\le\; 1+\frac{1}{2}t^{2}R_{ik}^{2}(\infty)\times\sum_{l=2}^{\infty}\frac{2}{l!}t^{l-2}C_{1}^{l-2} \;\le\; 1+\frac{3}{2}t^{2}R_{ik}^{2}(\infty).$$

On the other hand, in view of the inequality $\exp\{|x|\} \le \exp\{x\} + \exp\{-x\}$,

$$\mathbb{E}\left(\exp\left|t\sum_{i=1}^{n}E_{i}R_{ik}(t_{i'})\right| \mid \boldsymbol{\varphi}\right) \;\le\; 2\prod_{i=1}^{n}(1+\frac{3}{2}t^{2}R_{ik}^{2}(\infty)).$$

Hence, by Markov's inequality,

$$\mathbb{P}\left[\left|\sum_{i=1}^{n}E_{i}R_{ik}(t_{i'})\right| \ge \frac{nh^{d}\varepsilon}{12} \;\Big|\; \boldsymbol{\varphi}\right] \le 2\exp\left\{-t\frac{nh^{d}\varepsilon}{12}\right\}\prod_{i=1}^{n}\left(1+\frac{3}{2}t^{2}R_{ik}^{2}(\infty)\right). \tag{7.3}$$

We get rid of the conditioning by taking expectation on both side of (7.3). Now $\mathbb{E}(R_{ik}^{2}(\infty)) \le h^{d}C_{1}$. Upon using the inequality $1 + x \le \exp\{x\}$, we get

$$\mathbb{E}\left(\prod_{i=1}^{n}\left(1+\frac{3}{2}t^{2}R_{ik}^{2}(\infty)\right)\right) \;\le\; \exp\left\{nh^{d}t^{2}\,\frac{3}{2}C_{1}\right\}.$$

Hence

$$\mathbb{P}\left[\left|\sum_{i=1}^{n}E_{i}R_{ik}(t_{i'})\right| \ge \frac{nh^{d}\varepsilon}{12}\;\right] \;\le\; 8(n+1)\exp\{-tnh^{d}\varepsilon\,/\,12+t^{2}nh^{d}3C_{1}\,/\,2\}.$$

The term in the exponential is minimized at $t^{*} = \varepsilon/36C_{1}$. This shows that

$$\mathbb{P}\left[\sup_{\boldsymbol{x}\in\mathcal{C}}\sup_{y\in\mathbb{R}}\Big|\tilde{F}_{num}(y\mid\boldsymbol{x})-\mathbb{E}(\tilde{F}_{num}(y\mid\boldsymbol{x}))\Big| \ge \varepsilon\right] \;\le\; 8(n+1)h^{-d(d+1)}\exp\{-nh^{d}\varepsilon^{2}C_{3}\}.$$

where $C_{3} = 1/(648\,C_{1})$. The Borel-Cantelli lemma finishes the proof. ■

Proof of Theorem 2.1. From Lemma 7.1 and upon using standard results from matrix theory, we get $\mathbf{S}_{n}^{-1}(\boldsymbol{x}) = \mathbf{S}^{-1}\Big(1+h\,\mathcal{O}_{as}(1)\Big)\,/\,f(\boldsymbol{x})$. Thus, from (2.6)

$$\mathcal{W}(\boldsymbol{t}) = \mathcal{W}_{eq}(\boldsymbol{t})(1+h\mathcal{O}_{as}(1)). \tag{7.4}$$

Hence $\Big|\hat{F}(y\mid\boldsymbol{x})-\tilde{F}(y\mid\boldsymbol{x})\Big| = h\mathcal{O}_{as}(1)\Big|\tilde{F}(y\mid x)\Big|$ and Lemma 7.2 finishes the proof of the first assertion. Now, in view of this,

$$\sup_{\boldsymbol{x}\in\mathcal{C}}\sup_{y\in\mathbb{R}}\left|\hat{F}(y\mid\boldsymbol{x})-F(y\mid\boldsymbol{x})\right| \;\le\; h\mathcal{O}_{as}(1) \;+\; \sup_{\boldsymbol{x}\in\mathcal{C}}\sup_{y\in\mathbb{R}}\left|\tilde{F}(y\mid\boldsymbol{x})-F(y\mid\boldsymbol{x})\right|.$$

Applying Lemma 7.2 again gives the second assertion. ■

7.2 Proof of Theorem 3.1

Write $\mathcal{I}_i(y) = \mathbb{I}\{Y_i \le y\} - F_0(y \mid \boldsymbol{X}_i)$. In the following proofs, integration with respect to y, y_i or $y^{(j)}$ is over $\mathbb{R}$ while integration with respect to $\boldsymbol{x}_i$ (resp. $\boldsymbol{x}$, $\boldsymbol{x}^{(j)}$) is over $\mathcal{S}$ (resp. $\mathcal{C}$). Note that the results in this subsection are derived under H_0.

Lemma 7.3: *Suppose Assumption X and let $\mathcal{H}$ be a kernel satisfying Assumption K with h according to Assumption H–a). Let $g(\boldsymbol{x},y)$ satisfy Assumption W. Consider*

$$V = nh^{d/2}\sum_{i=1}^{n}\iint \frac{1}{n^2h^{2d}f^2(\boldsymbol{x})}\mathcal{H}^2\left(\frac{\boldsymbol{X}_i-\boldsymbol{x}}{h}\right)\mathcal{I}_i^2(y)g(\boldsymbol{x},y)d\boldsymbol{x}dy\,.$$

Then, setting $\mathcal{H}^{(2)} = \int_{[-1,1]^d}\mathcal{H}^2(\boldsymbol{u})d\boldsymbol{u}$ and

$$a_{\mathcal{H}}(g) = \mathcal{H}^{(2)}\iint \frac{F_0(y\mid\boldsymbol{x})(1-F_0(y\mid\boldsymbol{x}))}{f(\boldsymbol{x})}g(\boldsymbol{x},y)dyd\boldsymbol{x}\,, \tag{7.5}$$

we have $\mathbb{E}(V) = h^{-d/2}a_{\mathcal{H}}(g)\big(1+O(h)\big)$, $\mathbb{V}(V) = O((nh^d)^{-1})$, so that $V-\mathbb{E}(V) = o_p(1)$.

Proof: Because $\int \mathcal{I}_1^2(y)f_0(y_1\mid\boldsymbol{x}_1)dy_1 = F_0(y\mid\boldsymbol{x}_1)\big(1-F_0(y\mid\boldsymbol{x}_1)\big)$,

$$\mathbb{E}(V) = \frac{1}{nh^{3d/2}}\iint\left[\int\mathcal{H}^2\left(\frac{\boldsymbol{x}_1-\boldsymbol{x}}{h}\right)F_0(y\mid\boldsymbol{x}_1)\big(1-F_0(y\mid\boldsymbol{x}_1)\big)f(\boldsymbol{x}_1)\,d\boldsymbol{x}_1\right]\frac{g(\boldsymbol{x},y)}{f^2(\boldsymbol{x})}\,dy\,d\boldsymbol{x}\,.$$

By a change of variable, the integral in brackets is $h^d\mathcal{H}^{(2)}F_0(y\mid\boldsymbol{x})(1-F_0(y\mid\boldsymbol{x}))f(\boldsymbol{x}) + h^{d+1}O(1)$, uniformly in $(\boldsymbol{x}, y)$. Hence $\mathbb{E}(V)$ has the stated form. Now because $\mathbb{E}\Big(\mathcal{I}_1(y)^2\mathcal{I}_1(y')^2 \mid \boldsymbol{X}_1 = \boldsymbol{x}_1\Big) \le 1$ and in view of *Assumption X*

$$\mathbb{V}(V) \le \frac{\mathcal{M}}{nh^{3d}}\int\int\int\mathcal{H}^2\left(\frac{\boldsymbol{x}_1-\boldsymbol{x}}{h}\right)\mathcal{H}^2\left(\frac{\boldsymbol{x}_1-\boldsymbol{x}'}{h}\right)d\boldsymbol{x}\,d\boldsymbol{x}'\,d\boldsymbol{x}_1 \;+\; o((nh^d)^{-1}) \le \frac{\mathcal{M}}{nh^d}\Big(\mathcal{H}^{(2)}\Big)^2.$$

Thus $\mathbb{V}(V)$ is of the stated order. Chebyshev's inequality finished the proof. ■

Remark 7.3.1: Lemma 7.3 holds for a triangular array $\{(\boldsymbol{X}_{i,n}, Y_{i,n}),\ i = 1,\ldots,\ n,\ n \ge 1\}$ of row-wise *iid* random vectors with the joint structure described in Section 4.

Lemma 7.4: *Under the assumptions of Lemma* 7.3, *we have*

$$\sum_{i=1}^{n}\iint \frac{1}{nh^{d}}\frac{1}{f(\boldsymbol{x})}\mathcal{H}\left(\frac{\boldsymbol{X}_{i}-\boldsymbol{x}}{h}\right)\mathcal{I}_{i}(y)g(\boldsymbol{x},y)dyd\boldsymbol{x} \;=\sum\nolimits_{i=1}^{n} v_{i} \;=\; o_{p}((n^{1/2}h^{d/4})^{-1}).$$

When H–b) is added, $\sum_{i=1}^{n} v_{i} \;=\; o_{p}((nh^{p+1+d/4})^{-1})$.

Proof: Because $\mathbb{E}(\mathcal{I}_{i}(y) \mid \boldsymbol{X}_{1},...,\boldsymbol{X}_{n}) = 0$, we have $\mathbb{E}(v_{i}) = 0$. Moreover

$$\mathbb{E}(v_{i}^{2} \mid \boldsymbol{X}_{1},...,\boldsymbol{X}_{n}) \;\leq\; \frac{\mathcal{M}}{n^{2}h^{2d}}\iint\left|\mathcal{H}\left(\frac{\boldsymbol{X}_{1}-\boldsymbol{x}}{h}\right)\right|\left|\mathcal{H}\left(\frac{\boldsymbol{X}_{1}-\boldsymbol{x}'}{h}\right)\right|d\boldsymbol{x}\,d\boldsymbol{x}' \;\leq\; \frac{\mathcal{M}}{n^{2}},$$

hence $\mathbb{E}(v_{i}^{2}) \;=\; O(n^{-2})$. Applying Chebyshev's inequality concludes the proof. ■

Lemma 7.5: *Under the assumptions of Lemma* 7.3, *to which we add H–b), consider*

$$W = \frac{1}{nh^{3d/2}}\sum_{1\leq i\neq j\leq n}\iint\frac{1}{f^{2}(\boldsymbol{x})}\mathcal{H}\left(\frac{\boldsymbol{X}_{i}-\boldsymbol{x}}{h}\right)\mathcal{H}\left(\frac{\boldsymbol{X}_{j}-\boldsymbol{x}}{h}\right)\mathcal{I}_{i}(y)\mathcal{I}_{j}(y)g(\boldsymbol{x},y)d\boldsymbol{x}dy \;=\; \sum_{1\leq i\neq j\leq n} w_{ij,n}\,.$$

Then $W \xrightarrow{\;L\;} N(0,\ \sigma_{\mathcal{H}}^{2}(g))$ *where* $\sigma_{\mathcal{H}}^{2}(g) \;=\; \lim_{n\to\infty}\mathbb{V}(W)$, *e.g.*

$$\sigma_{\mathcal{H}}^{2}(g) \;=\; 2(\mathcal{H}^{(*2)})^{(2)}\int\int\left[\frac{\tau_{y,y'}(\boldsymbol{x})}{f(\boldsymbol{x})}\right]^{2}\; g(\boldsymbol{x},y)g(\boldsymbol{x},y')\,dy\,dy'd\boldsymbol{x}\,, \tag{7.6}$$

with $(\mathcal{H}^{(*2)})^{(2)} = \int\left(\int\mathcal{H}(\boldsymbol{t}+\boldsymbol{u})\mathcal{H}(\boldsymbol{u})d\boldsymbol{u}\right)^{2}d\boldsymbol{t}$ $,\tau_{y,y'}(\boldsymbol{x}) \;=\; F_{0}(y\wedge y' \mid \boldsymbol{x}) - F_{0}(y \mid \boldsymbol{x})F_{0}(y' \mid \boldsymbol{x})$.

Proof: Because W is a U-statistic, we use de Jong's (1987) Theorem 2.2 which state that asymptotic normality holds under the following conditions:

$$1)\ \mathbb{E}(\,w_{ij,n}\,|\,(\boldsymbol{X}_{i},\ Y_{i})) = 0,$$

$$2)\ \frac{\underset{1\leq i\leq n}{Max}\ \sum_{\substack{j=1\\ j\neq i}}^{n}\mathbb{V}(w_{ij,n})}{\mathbb{V}(W)}\ \to\ 0 \quad \text{while} \quad 3)\ \frac{\mathbb{E}(W^{4})}{\mathbb{V}^{2}(W)}\ \to\ 3.$$

Condition 1) follows immediately from $\mathbb{E}\left(\mathcal{I}_{1}(y) \mid \boldsymbol{X}_{1}\right) = 0$, as noted above. As for Condition 2), the use of de Jong's (1987) formula on p. 266 shows that the term on the left side is $1/2n$ because $\mathbb{V}(W) \;=\; 2n(n-1)\mathbb{E}(w_{12,n}^{2})$. But

$$\mathbb{E}(w_{12,n}^{2}) \;=\; \frac{1}{n^{2}h^{3d}}\iiint\iiint\mathcal{H}\left(\frac{\boldsymbol{x}_{1}-\boldsymbol{x}^{(1)}}{h}\right)\mathcal{H}\left(\frac{\boldsymbol{x}_{2}-\boldsymbol{x}^{(1)}}{h}\right)\mathcal{H}\left(\frac{\boldsymbol{x}_{1}-\boldsymbol{x}^{(2)}}{h}\right)\mathcal{H}\left(\frac{\boldsymbol{x}_{2}-\boldsymbol{x}^{(2)}}{h}\right)$$

$$\times \tau_{y_1,y_2}(\boldsymbol{x}_1)\tau_{y_1,y_2}(\boldsymbol{x}_2)f^{-2}(\boldsymbol{x}^{(1)})f^{-2}(\boldsymbol{x}^{(2)})g(\boldsymbol{x}^{(1)},y_1)g(\boldsymbol{x}^{(2)},y_2)d\boldsymbol{x}^{(1)}\,d\boldsymbol{x}^{(2)}\,d\boldsymbol{x}_1 d\boldsymbol{x}_2 dy_1 dy_2\,.$$

It follows directly that $2n^2\,\mathbb{E}(w_{ij,n}^2)$, and thus $\mathbb{V}(W)$, $\to \sigma_{\mathcal{H}}^2(g)$ of (7.6). Condition 3) is more difficult because it involves the term $\mathbb{E}(W^4)$. We use the decomposition given in Table 1 of de Jong (1987):

$$\mathbb{E}\left(\left(\sum_{1\le i\ne j\le n} w_{ij,n}\right)^4\right) = 16\,G_1 + 96\,G_2 + 192\,G_3 + 384\,G_4 + 96\,G_5,$$

where, because $(\boldsymbol{X}_i,\ Y_i)$ are *iid*, $16\,G_1 = 8n(n-1)\,\mathbb{E}\left(w_{12,n}^4\right)$, $96\,G_2 = 48n(n-1)(n-2)\,\mathbb{E}\left(w_{12,n}^2 w_{13,n}^2\right)$, $\left|G_3\right| \le G_2$, $384\,G_4 = 48n(n-1)(n-2)(n-3)\,\mathbb{E}\left(w_{12,n}w_{13,n}w_{42,n}w_{43,n}\right)$, $96\,G_5 = 12n(n-1)(n-2)(n-3)\,\mathbb{E}^2\left(w_{12,n}^2\right)$. The first four of the above expectations are respectively $O(n^{-4}h^{-d})$, $O(n^{-4})$, $O(n^{-4})$ and $O(n^{-4}h^{d})$, so that the terms G_1, G_2, G_3 and G_4 are $o(1)$. To see this requires lengthy calculations. We present only the case of G_1 that is typical. No new difficulty arises in the other cases.

$$\begin{aligned}
\mathbb{E}\left(w_{12,n}^4\right) &= \frac{1}{n^4h^{6d}}\int\cdots\int\prod_{j=1}^{4}\prod_{i=1}^{2}\mathcal{H}\left(\frac{\boldsymbol{x}_i-\boldsymbol{x}^{(j)}}{h}\right)\\
&\quad\times\prod_{i=1}^{2}\int\prod_{j=1}^{4}(\mathbb{I}\{y_i\le y^{(j)}\}-F_0(y^{(j)}\mid\boldsymbol{x}_i))f_0(y_i\mid\boldsymbol{x}_i)dy_i\\
&\quad\times\prod_{j=1}^{4}g(\boldsymbol{x}^{(j)},y^{(j)})\,dy^{(j)}\prod_{j=1}^{4}f^{-2}(\boldsymbol{x}^{(j)})\,d\boldsymbol{x}^{(j)}\prod_{i=1}^{2}f(\boldsymbol{x}_i)\,d\boldsymbol{x}_i\,,\\
&\le \mathcal{M}n^{-4}h^{-6d}\int\cdots\int\left(\prod_{i=1}^{2}\prod_{j=1}^{4}\left|\mathcal{H}\left(\frac{\boldsymbol{x}_i-\boldsymbol{x}^{(j)}}{h}\right)\right|\right)d\boldsymbol{x}^{(1)}\cdots d\boldsymbol{x}^{(4)}d\boldsymbol{x}_1 d\boldsymbol{x}_2\,.
\end{aligned}$$

But,

$$\frac{1}{h^{2d}}\int\prod_{i=1}^{2}\left|\mathcal{H}\left(\frac{\boldsymbol{x}_i-\boldsymbol{x}^{(j)}}{h}\right)\right|d\boldsymbol{x}^{(j)} \le \frac{1}{h^d}\mathcal{H}_+^{(2)}\left(\frac{\boldsymbol{x}_2-\boldsymbol{x}_1}{h}\right),$$

where $\mathcal{H}_+^{(2)}$ denotes the convolution of $\left|\mathcal{H}\right|$. Hence,

$$\mathbb{E}\left(w_{12,n}^4\right) \le \frac{\mathcal{M}}{n^4h^{2d}}\iint\left(\mathcal{H}_+^{(2)}\left(\frac{\boldsymbol{x}_2-\boldsymbol{x}_1}{h}\right)\right)^4 d\boldsymbol{x}_1 d\boldsymbol{x}_2\,,$$

$$\leq \frac{\mathcal{M}}{n^4 h^d} \iint \left(\mathcal{H}_+^{(2)}\left(\boldsymbol{t}\right)\right)^4 d\boldsymbol{t}\, d\boldsymbol{x}_1 \;=\; O(n^{-4}h^{-d}).$$

As for G_5, it was noted above that $\mathbb{E}\left(w_{12,n}^2\right) = \mathbb{V}(W)/(2n(n-1))$. Hence

$$96\mathrm{G}_5 = \; 3\mathbb{V}^2(W)\frac{(n-2)(n-3)}{n(n-1)}.$$

Condition 3 follows. ■

Proof of Theorem 3.1: First note that $\mathcal{K}$ and w^* satisfy the conditions of Lemmas 7.3, 7.4 and 7.5. Now write

$$\begin{aligned}\hat{F}(y \mid \boldsymbol{x}) - F_0(y \mid \boldsymbol{x}) \;&=\; (1 + h\mathcal{O}_{as}(1))\sum_{i=1}^{n} \mathcal{W}_{eq}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right)\mathcal{I}_i(y) \\ &\quad + \;\sum_{i=1}^{n} \mathcal{W}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right)\Big(F_0(y \mid \boldsymbol{X}_i) - F_0(y \mid \boldsymbol{x})\Big), \qquad (7.7) \\ &= \; (1 + h\mathcal{O}_{as}(h))\, A_n(\boldsymbol{x},\, y) \;+\; B_n(\boldsymbol{x},\, y).\end{aligned}$$

Making use of *Assumption F*, expand $F_0(y \mid \boldsymbol{X}_i)$ about $\boldsymbol{x}$. From (2.4), we get for $\boldsymbol{x}_i^*$ on the line segment between $\boldsymbol{x}$ and $\boldsymbol{X}_i$,

$$\begin{aligned}\left|B_n(\boldsymbol{x}, y)\right| \;&=\; \left|\sum_{\boldsymbol{j}\in\mathcal{J}_{p+1}} \frac{1}{\boldsymbol{j}!}\sum_{i=1}^{n} \mathcal{W}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right)(\boldsymbol{X}_i - \boldsymbol{x})^{\boldsymbol{j}}\, \frac{\partial^{p+1}}{(\partial \boldsymbol{x})^{\boldsymbol{j}}} F_0(y \mid \boldsymbol{x}_i^*)\right|, \\ &\leq \; (1 + h\mathcal{O}_{as}(1))\frac{\mathcal{M}h^{p+1}}{f(\boldsymbol{x})}\left|\sum_{\boldsymbol{j}\in\mathcal{J}_{p+1}} \frac{1}{\boldsymbol{j}!}\frac{1}{nh^d}\sum_{i=1}^{n}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right)^{\boldsymbol{j}} \mathcal{K}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\right)\right|,\end{aligned}$$

in view of (7.4). Applying Lemma 7.1 with $\mathcal{K}$ yields, uniformly in $\boldsymbol{x}$, y:

$$\left|B_n(\boldsymbol{x}, y)\right| \;\leq\; h^{p+1}\mathcal{O}_{as}(1).$$

Now, (7.7) and (3.2) give

$$\begin{aligned}T_n = \; &(1 + h\mathcal{O}_{as}(h))\; nh^{d/2}\iint A_n^2\left(\boldsymbol{x}, y\right) w^*(\boldsymbol{x}, y) dy d\boldsymbol{x} \\ &+ \; 2\, nh^{d/2+p+1}\mathcal{O}_{as}(1)\iint A_n\left(\boldsymbol{x}, y\right) w^*(\boldsymbol{x}, y) dy d\boldsymbol{x} \;+\; nh^{d/2\,+\,2p\,+2}\;\mathcal{O}_{as}(1), \\ = \; &(1 + h\mathcal{O}_{as}(h))\; U_1 + \; U_2 \;+\; o_p(1),\end{aligned}$$

in view of *Assumption H–b*). Using (7.4) and applying Lemma 7.4 shows that $U_2 = o_p(1)$, so that $T_n = (1 + h\mathcal{O}_{as}(h))\, U_1 + o_p(1)$. Now, again from (7.4),

$$U_1 = nh^{d/2} \iint \left(\sum_{i=1}^{n} \mathcal{W}_{eq}\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h} \right) \mathcal{I}_i(y) \right)^2 w^*(\boldsymbol{x}, y) dy \; d\boldsymbol{x} = \sum_{i=1}^{n} \zeta_{ii,n} + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \zeta_{ij,n} .$$

Lemmas 7.3, 7.5 show that $U_1 - h^{-d/2} a_{\mathcal{K}}(g)(1 + hO(1)) \xrightarrow{L} N(0,\ \sigma^2_{\mathcal{K}}(g))$. ■

Remark 7.6: The argument above shows that $\mathbb{E}(T_n) \approx n^2 h^{d/2} \mathbb{E}\Big(\psi(\boldsymbol{X}_i; \mathcal{W}_{eq}) \Big)$, where

$$\psi(\boldsymbol{X}_i; \mathcal{W}) = \iint \mathcal{W}^2 \left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h} \right) F_0(y \mid \boldsymbol{x})(1 - F_0(y \mid \boldsymbol{x})) w^*(\boldsymbol{x}, y) d\boldsymbol{x} dy .$$

When n increases, $\mathcal{W} \to \mathcal{W}_{\text{eq}}$ by (7.4) and by continuity, $\psi(\boldsymbol{X}_i; \mathcal{W}) \to \psi(\boldsymbol{X}_i; \mathcal{W}_{eq})$. The expectation of this quantity can be consistently estimated by taking a bootstrap sample $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*$ from the empirical distribution of the $\boldsymbol{X}_i$ computing $\mathcal{W}^*$ and the sum of the $\psi(\boldsymbol{X}_i^*; \mathcal{W}^*)$ and repeating this over B bootstrap replication. The average of these sums is a bootstrap estimator of $n\, \mathbb{E}\Big(\psi(\boldsymbol{X}_i; \mathcal{W}) \Big)$. More generally the whole distribution of T_n can be bootstrapped by first sampling $\boldsymbol{X}_i^*$ first from the empirical distribution of the $\boldsymbol{X}_i$ and then Y_i^* from $F_0(y \mid \boldsymbol{X}_i^*)$.

7.3 Proof of Theorem 4.1

Substituting $F_0(y|x) = F_{1,n}(y|x) - c_n R(y|x)$, one gets from (3.2)

$$\begin{aligned} T_n &= nh^{1/2} \iint \Big(\hat{F}(y \mid x) - F_{1,n}(y \mid x) \Big)^2 f_0(y \mid x) dy dx + b_R(f_0) \\ &\qquad + 2\, c_n^{-1} \iint (\hat{F}(y \mid x) - F_{1,n}(y \mid x)) R(y \mid x) f_0(y \mid x) dy dx , \\ &= C_{1,n} + b_R(f_0) + C_{2,n}. \end{aligned}$$

Writing $\mathcal{I}_{i,n}(y) = \mathbb{I}\{ Y_{i,n} \leq y \} - F_{i,n}(y \mid X_{i,n})$ it is easy to see, in view of (2.4), (7.4) and (4.1) that $C_{2,n} = 2(1 + h\mathcal{O}_{as}(1)) c_n^{-1} \sum_{i=1}^{n} v_{i,n}^*$, where

$$v_{i,n}^* = \iint \mathcal{W}_{eq}\left(\frac{X_{i,n} - x}{h} \right) \mathcal{I}_{i,n}(y) R(y \mid x) f_0(y \mid x) dy\, dx .$$

This is a function of $(X_{i,n}\ Y_{i,n})$ where the density of $X_{i,n}$ does not vary with n. Thus we can apply Lemma 7.4 with $\mathcal{H} = \mathcal{K}$, $g(x, y) = R(y \mid x)f_0(y \mid x)$ to get under $H_{1,n}$, $C_{2,n} \xrightarrow{P} 0$. Now, write

$$C_{1,n} = (1 + h\mathcal{O}_{as}(1))nh^{1/2}\left(\sum_{i=1}^{n} w^*_{ii,n} + \sum_{1\le i\ne j\le n} w^*_{ij,n}\right),$$

where $w^*_{ij,n} = \iint \mathcal{W}_{eq}\left(\frac{X_{i,n} - x}{h}\right)\mathcal{W}_{eq}\left(\frac{X_{j,n} - x}{h}\right)\mathcal{I}_{i,n}(y)\mathcal{I}_{j,n}(y)f_0(y \mid x)dydx\,.$

Applying the triangular version of Lemma 7.3 yields

$$nh^{1/2}\mathbb{E}\left(\sum_{i=1}^{n} w^*_{ii,n}\right) = h^{-1/2}a_{1,\mathcal{K}}(f_0)(1 + O(h))\,,$$

where $a_{1,\mathcal{K}}(f_0)$ has the same structure as (7.5) with $F_{1,n}$ replacing F_0, and is such that $h^{-1/2}\left|a_{1,\mathcal{K}}(f_0) - a_{\mathcal{K}}(f_0)\right| = O((nh^{3/2})^{-1/2})\,.$

The treatment of the remaining (U statistic) part of $C_{1,n}$ is a bit more involved because de Jong's (1987) work does not deal with triangular arrays. But it does allow for a kernel varying with n, so his Theorem 2.2 can be adapted to the present context via the following argument. Starting from a sequence of *iid* $Z_i = (X_i,\ Y_i) \sim F_0(x,\ y)$, one can build the *iid* sequence $Z_{i,n} = (X_i,\ F^{-1}_{1,n}(F_0(Y_i \mid X_i) \mid X_i)) = (X_i,\ H_n(Z_i)) \sim F_{1,n}(x,\ y)$ (here $F_{1,n}(x,\ y)$ is the joint cumulative *cdf* associated with $F_{1,n}(y \mid x)$ and $f(x)$). The moments of a U statistic of the form $\sum_{1\le i<j\le n} h_n(Z_{i,n}, Z_{j,n})$ are the same as those of $\sum_{1\le i<j\le n} h^*_n(Z_i, Z_j)$, where $h^*_n(Z_i, Z_j) = h_n((X_i, H_n(Y_i)), (X_j, H_n(Y_j)))$. This is the context of de Jong's theorem and, applying Lemma 7.5, we get

$$c_n^{-2}\sum_{1\le i\ne j\le n} w^*_{ij,n} \,/\, \sigma_{1,\mathcal{K}}(f_0) \xrightarrow{L} N(0,\ 1),$$

where $\sigma^2_{1,\mathcal{K}}(f_0)$ has the same expression as $\sigma^2_{\mathcal{K}}(f_0)$ in (7.6) with $F_{1,n}$ replacing F_0. Notice also that $\sigma^2_{1,\mathcal{K}}(f_0) \to \sigma^2_{\mathcal{K}}(f_0)$. Combining these results concludes the proof. ■

7.4 Proof of Theorem 5.1

We write the proof for $\theta \in \mathbb{R}$. Expand in Taylor series both $F_{\hat\theta}(y \mid \boldsymbol{x})$ and $f_{\hat\theta}(y \mid \boldsymbol{x})$ about the true value θ_0. Write $g_{\theta^*}(\boldsymbol{x}, y) = w(\boldsymbol{x})w_{\boldsymbol{x}}(y)(f_{\theta_0}(y \mid \boldsymbol{x}) + c_n\,\dot f_{\theta^*}(y \mid \boldsymbol{x}))\,,$ where θ^* lies on the line segment joining θ_0 and $\hat\theta$, $\dot f$ is the derivative of f with

respect to θ and $c_n = (\hat{\theta} - \theta_0) = O_p(n^{-1/2})$ by Assumption *E*. Also write $\dot{F}_\theta$, $\ddot{F}_\theta$ for the first and second derivative of F_θ with respect to θ. Injecting these into (5.2) yields,

$$\begin{aligned} T_n(\hat{\theta}) = \; & nh^{d/2} \int\int \Big(\hat{F}(y \mid \boldsymbol{x}) - F_{\theta_0}(y \mid \boldsymbol{x})\Big)^2 g_{\theta^*}(\boldsymbol{x}, y) dy d\boldsymbol{x} \\ & + 2nh^{d/2} c_n \int\int \Big(\hat{F}(y \mid \boldsymbol{x}) - F_{\theta_0}(y \mid \boldsymbol{x})\Big) F_{\theta_0}(y \mid \boldsymbol{x}) g_{\theta^*}(\boldsymbol{x}, y) dy d\boldsymbol{x} \\ & + nh^{d/2} c_n^2 \int\int \Big(\dot{F}_{\theta^*}(y \mid \boldsymbol{x})\Big)^2 g_{\theta^*}(\boldsymbol{x}, y) dy d\boldsymbol{x} \qquad (7.8) \\ & + 2nh^{d/2} c_n^2 \int\int \Big(\hat{F}(y \mid \boldsymbol{x}) - F_{\theta_0}(y \mid \boldsymbol{x})\Big) \ddot{F}_{\theta^*}(y \mid \boldsymbol{x}) g_{\theta^*}(\boldsymbol{x}, y) dy d\boldsymbol{x} \,, \\ = \; & T_n^*(\theta_0) + n^{1/2} h^{d/2} O_p(1) R_{n,1} + h^{d/2} O_p(1) R_{n,2} + h^{d/2} O_p(1) R_{n,3}. \end{aligned}$$

Obviously, the last two terms are $o_p(1)$. Note that both $g_{\theta^*}(\boldsymbol{x}, y)$ and $\dot{F}_{\theta^*}(y \mid \boldsymbol{x}) g_{\theta^*}(\boldsymbol{x}, y)$ satisfies *Assumption W* so that applying Lemmas 7.3, 7.4 and 7.5 as in the proof of Theorem 3.1 shows that $R_{n,1} = O_p((nh^{d/2})^{-1})$ and the second term is $o_p(1)$. Finally, $T_n^*(\theta_0) = T_n(\theta_0) + n^{1/2} h^{d/2} O_p(1) R_{n,4}$, where

$$\Big|R_{n,4}\Big| \leq nh^{d/2} \int\int \Big(\hat{F}(y \mid \boldsymbol{x}) - F_{\theta_0}(y \mid \boldsymbol{x})\Big)^2 w(\boldsymbol{x}) w_{\boldsymbol{x}}(y) h_{\boldsymbol{x}}(y) dy d\boldsymbol{x} \,.$$

Applying the reasoning of Theorem 3.1 shows that $R_{n,4} = O_p(1)$. Therefore, $T_n(\hat{\theta}) = T_n(\theta_0) + o_p(1)$. Because $T_n(\theta_0)$ is exactly the same statistic as T_n, we get

$$\frac{T_n(\hat{\theta}) - h^{-d/2} a_{\theta_0, \mathcal{K}}(f_{\theta_0})(1 + h\mathcal{O}_{as}(1))}{\sigma_{\theta_0, \mathcal{K}}(f_{\theta_0})} \xrightarrow{L} N\Big(0, 1\Big). \qquad (7.9)$$

Noting that $a_{\hat{\theta}, \mathcal{K}}(f_{\hat{\theta}}) = a_{\theta_0, \mathcal{K}}(f_{\theta_0}) + O_p(n^{-1/2})$ and similarly for $\sigma^2_{\hat{\theta}, \mathcal{K}}(f_{\hat{\theta}})$, concludes the proof. ■

References

Alcala, J.T., Cristobal, J.A., Conzalez-Manteiga, W. (1999): Goodness-of-fit test for linear models based on local polynomials. *Statistics and Probability Letters*, **42**, 39-46.

Collomb, G. (1980): Estimation non paramétrique de probabilités conditionnelles. *C. R. de Acad. Sci. Paris*, **291**, *Serie A*, 427-430.

D'Agostino, R.B., Stephens, M.A. (1986): *Goodness-of-fit Techniques*. Marcel Dekker, New-York.

Ducharme, G.R., Mint el Mouvid, M. (2001): Convergence presque sure de l'estimateur linéaire local de la fonction de repartition conditionnelle. *C.R. Acad. Sci. Paris*, **333,** *Série I*, 873-876.

Ducharme, G.R., Fontez, B. (2004): A smooth test of goodness-of-fit for growth curves and monotonic nonlinear regression models. *Biometrics*, **60**, p. 977-986.

Fan, J., Gijbels, I. (1996): *Local polynomial modeling and its application*. Chapman and Hall, New York.

Ferrigno, S., Ducharme, G.R. (2008): Un choix de fenêtre optimal en estimation polynomiale locale de la fonction de répartition conditionnelle. *C.R. Acad. Sci. Paris,* **346**, *Serie I*, 83-86.

Hall, P., Wolff, R.C.L., Yao, Q. (1999): Methods for estimating a conditional distribution function. *Journal of the American Statistical Association.*, **94**, p. 154-163.

Huang, L.S., Fan, J. (1999): Nonparametric estimation of quadratic regression functionals. *Bernoulli*, **5**, 927-949.

Liero, H. (2003): Testing homoscedasticity in nonparametric regression. *Journal of Nonparametric Statistics*, **15**, p.31-51.

Masry, E. (1996): Multivariate local polynomial regression for times series: uniform strong consistency and rates. *Journal of Time Series Analysis,* **17**, p. 571-599.

McCullagh, P., Nelder, J.A. (1983): *Generalized Linear Models*. Chapman and Hall, London.

Prakasa Rao, B.L.S. (1983): *Nonparametric Functional Estimation*. Academic Press NewYork.

Van Keilegom, I., Gonzalez-Manteiga, W., Sellero, C.S. (2008): Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *Test*, **17**, p.401-415.