Goodness-of-fit Tests Based on the Kernel Density Estimator

RICARDO CAO

Department of Mathematics, Universidade da Coruña

GÁBOR LUGOSI

Department of Economics, Pompeu Fabra University

ABSTRACT. Given an i.i.d. sample drawn from a density f on the real line, the problem of testing whether f is in a given class of densities is considered. Testing procedures constructed on the basis of minimizing the L_1-distance between a kernel density estimate and any density in the hypothesized class are investigated. General non-asymptotic bounds are derived for the power of the test. It is shown that the concentration of the data-dependent smoothing factor and the 'size' of the hypothesized class of densities play a key role in the performance of the test. Consistency and non-asymptotic performance bounds are established in several special cases, including testing simple hypotheses, translation/scale classes and symmetry. Simulations are also carried out to compare the behaviour of the method with the Kolmogorov-Smirnov test and an L_2 density-based approach due to Fan [*Econ. Theory* 10 (1994) 316].

Key words: bandwidth, goodness-of-fit test, kernel density estimator, smoothing factor selection

1. Introduction

Given a class of densities $\mathcal{F}$ on $\mathbb{R}$ and a sample of n i.i.d. random variables $X_1, \ldots, X_n$ drawn from an unknown density f, the problem is to decide whether the null hypothesis $\mathcal{H}_0 : f \in \mathcal{F}$ is true or not. For all goodness-of-fit tests we require that, under the null hypothesis, the probability of rejecting $\mathcal{H}_0$ be at most α where $\alpha \in (0, 1)$ is a pre-specified value.

The tests we propose are based on the kernel density estimator. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function with $\int K = 1$. For convenience we assume that K is non-negative. For a smoothing factor $h > 0$, the kernel density estimator $f_{n,h}$ is defined as $f_{n,h}(x) = 1/n \sum_{i=1}^{n} K_h(x - X_i)$, where $K_h(\cdot) = (1/h)K(\cdot/h)$ (see Akaike, 1954; Parzen, 1962; Rosenblatt, 1956). In most of the cases along this paper (but not always) the smoothing parameter will tend to zero when the sample size goes to infinity. We will use h_n instead of h to make explicit the dependence on the sample size.

The composite goodness-of-fit tests we investigate in this paper all have the form: accept $\mathcal{H}_0$ if and only if $T_n \leq c_\alpha$, where c_α is an appropriate constant and the test statistic T_n has the form

$$T_n = \inf_{g \in \mathcal{F}} \int |f_{n,h_n} - g|.$$

Different versions of these tests differ in their choices of the smoothing factor h_n and the constant c_α. In general, we allow the smoothing factor $h_n = h_n(X_1, \ldots, X_n)$ to depend on the data. The main results of the paper point out that the choice of h_n plays a crucial role in the performance of the test. In particular, we require that the L_1 error of the estimator f_{n,h_n} be sharply concentrated around its mean. We show that this property is satisfied in several natural choices of the smoothing factor.

Some killing-the-bias version of our test statistic may be defined as

$$T_n' = \inf_{g \in \mathcal{F}} \int |f_{n,h_n} - K_{h_n} * g|.$$

However, this version will not be considered in this paper as it is incompatible with the minimum-distance bandwidth choice studied in section 3. This smoothing-the-model approach has been widely used in the regression context for smooth alternatives (see Härdle & Kneip, 1999, and references therein for details).

To keep the discussion simple we only consider univariate data in this paper. Extensions to the multivariate case are straightforward.

Classical approaches to solve this goodness-of-fit problem use empirical process theory. For instance, the popular Kolmogorov–Smirnov test considers the maximal difference between the empirical measure and the hypothesized probability measure over all sets of the form $(-\infty, x]$. The L_1 approach is very much related to this, as it coincides with the total variation distance, i.e. the maximal difference between the measure induced by the density estimator and the hypothesized measure over all Borel sets.

The idea of using non-parametric density estimators for goodness-of-fit tests goes back to Bickel & Rosenblatt (1973) and Rosenblatt (1975).

More recent work includes Ahmad & Cerrito (1993) and Fan (1994, 1998). All these papers base their tests on the L_2 error of the kernel density estimator. A distinguishing feature of the approach of this paper is the use of the L_1 error instead. This allows to drop unnecessary assumptions as well as to obtain non-asymptotic performance bounds. In fact, our results are strongly based on some specific desirable properties of the L_1 error.

Using the L_1 distance in hypothesis testing is not new: Györfi & van der Meulen (1991) have proven the universal consistency of a similar test, in the case of a simple hypothesis, based on the histogram estimator. In fact, the study of testing based on the L_1 (or total variation) distance goes back to Hoeffding & Wolfowitz (1958) and LeCam (1973). For related results in a somewhat different setup we refer to Devroye & Lugosi (2002). Finally, we mention that the results of the paper may be used to derive upper bounds for the minimax rates of testing in certain situations. (For the definition of minimax rates see e.g. Lepski & Tsybakov, 2000.) However, these upper bounds are very general and it remains to see in what concrete problems the minimal rates are actually achieved by the L_1 kernel-based test.

The rest of the paper proceeds as follows. In section 2, non-asymptotic bounds are derived for the performance of the test which are valid for all densities f and for all classes $\mathcal{F}$. The main assumption in these results is that the test statistic T_n is sharply concentrated around its mean, a property which is proven in most natural examples. These general results show that the test has an excellent behaviour whenever the class $\mathcal{F}$ is not 'too large' in a certain sense and/or the kernel estimator has certain stability properties.

In section 3, a general smoothing factor is defined which fits naturally in the framework of the minimum-distance test studied in this paper. The corresponding test statistic is shown to satisfy the concentration property required for the general results of section 2. As an example, the test is shown to be universally consistent under the only requirement that $\mathcal{F}$ is totally bounded, a surprisingly strong property.

In section 4, the general results of section 2 are applied to the simplest special case when $\mathcal{F}$ contains a single density f_0, to the composite hypothesis case, where $\mathcal{F}$ is a translation/scale class, and also for testing symmetry.

To make the main ideas more transparent, in the majority of the paper we ignore computational issues and assume that all quantities which depend on the data and the class $\mathcal{F}$ can be computed. Nevertheless, in section 5, we describe some simulation results.

2. General results

In this section, we investigate some basic properties of tests of the proposed form. We establish several results under general assumptions on the class $\mathcal{F}$ and the smoothing factor h_n. In

subsequent sections, we illustrate in several applications how these results can be used in concrete examples.

In the rest of the paper, $\mathbb{E}_f$ and $\mathbb{P}_f$ denote expectation and probability under the assumption that the data are drawn according to the density f.

Recall that we investigate tests based on the statistic $T_n = \inf_{g \in \mathcal{F}} \int |f_{n,h_n} - g|$ where $h_n = h_n(X_1, \ldots, X_n)$ is a possibly data-dependent smoothing factor. Then, given a class of densities $\mathcal{F}$ and a constant $\alpha \in (0, 1)$, we may compute

$$c_\alpha = \inf\left\{ c : \sup_{f \in \mathcal{F}} \mathbb{P}_f[T_n > c] \leq \alpha \right\}.$$

Computability of this constant is a simplifying assumption we use here to focus on essentials. In any case, c_α may be approximated with arbitrary precision using Monte-Carlo simulations. As we accept $\mathcal{H}_0$ if and only if $T_n \leq c_\alpha$, the definition of c_α immediately guarantees that if the data are indeed drawn from a density in $\mathcal{F}$, then the rejection probability is bounded by α, as desired. It remains to see how the test behaves when $f \notin \mathcal{F}$. The starting point for bounding the probability of acceptance is the following simple result. The key assumption is that, regardless of whether $f \in \mathcal{F}$ or not, the test statistic T_n is assumed to be concentrated around its expected value. Later we will see that this assumption is satisfied for several natural choices of the smoothing factor.

Theorem 1
Let $b_n \stackrel{\text{def}}{=} \sup_{f \in \mathcal{F}} \mathbb{E}_f T_n$ *and assume that there exist positive constants* κ_1, κ_2 *such that for every* $\epsilon > 0$ *and for all densities* f,

$$\mathbb{P}_f[|T_n - \mathbb{E}_f T_n| > \epsilon] \leq \kappa_1 \, \mathrm{e}^{-\kappa_2 n \epsilon^2}.$$

Then $c_\alpha \leq b_n + \sqrt{(1/\kappa_2 n) \log(\kappa_1/\alpha)}$ *and for any density* f *and* $\delta > 0$, *if*

$$\mathbb{E}_f T_n > b_n + \delta + \sqrt{\frac{1}{\kappa_2 n} \log \frac{\kappa_1}{\alpha}},$$

then

$$\mathbb{P}_f[\mathcal{H}_0 \textit{ is accepted}] \leq \kappa_1 \, \mathrm{e}^{-\kappa_2 n \delta^2}.$$

Note that $b_n \leq \sup_{f \in \mathcal{F}} \mathbb{E}_f \int |f_{n,h_n} - f|$ which is the largest L_1 error of the kernel density estimator used in the procedure *within the class* $\mathcal{F}$ and therefore b_n is typically small (unless the class $\mathcal{F}$ is very large and/or contains densities which are hard to estimate with a kernel estimator). On the other hand, if $f \notin \mathcal{F}$ then it is expected that $\mathbb{E}_f T_n$ is large. For example, in typical situations we will have that $b_n \to 0$ and for $f \notin \mathcal{F}$, $\mathbb{E}_f T_n$ does not converge to zero. In such cases the probability of making a mistake is exponentially small.

Proof. Fix any $\delta > 0$. Using the definition of the test and the assumptions,

$$\begin{aligned} \mathbb{P}_f[\mathcal{H}_0 \text{ is accepted}] &\leq \mathbb{P}_f[\mathbb{E}_f T_n - T_n \geq \delta] + \mathbb{I}\{\mathbb{E}_f T_n \leq c_\alpha + \delta\} \\ &\leq \kappa_1 - \mathrm{e}^{-\kappa_2 n \delta^2} + \mathbb{I}\{\mathbb{E}_f T_n \leq c_\alpha + \delta\}, \end{aligned}$$

where $\mathbb{I}$ denotes the indicator function. To finish the proof, we need to show that $c_\alpha \leq b_n + \sqrt{(1/\kappa_2 n) \log(\kappa_1/\alpha)}$. Recall the definition of c_α and fix any $c > 0$ and $\gamma < c$, then

$$\begin{aligned}\sup_{f\in\mathcal{F}}\mathbb{P}_f[T_n > c] &= \sup_{f\in\mathcal{F}}\mathbb{P}_f[(T_n - \mathbb{E}_f T_n) + \mathbb{E}_f T_n > c] \\ &\le \kappa_1\, \mathrm{e}^{-\kappa_2 n\gamma^2} + \sup_{f\in\mathcal{F}} \mathbb{I}\{\mathbb{E}_f T_n > c-\gamma\} \\ &= \kappa_1\, \mathrm{e}^{-\kappa_2 n\gamma^2} + \mathbb{I}\left\{\sup_{f\in\mathcal{F}} \mathbb{E}_f T_n > c-\gamma\right\} \\ &= \kappa_1\, \mathrm{e}^{-\kappa_2 n\gamma^2} + \mathbb{I}\{b_n > c-\gamma\}.\end{aligned}$$

The obtained upper bound equals α if $\gamma = \sqrt{(1/\kappa_2 n)\log(\kappa_1/\alpha)}$ and $c \ge b_n + \gamma$, which concludes the proof.

As mentioned above, the main message of theorem 1 is that the performance of the test for $f \notin \mathcal{F}$ depends on the size of b_n and $\mathbb{E}_f T_n$. The sequence b_n simply depends on the performance of the kernel estimator for densities *within* the class $\mathcal{F}$. Bounding b_n is relatively straightforward, one may use standard techniques for analysing the L_1 error of the kernel density estimator (see the books of Devroye & Györfi, 1985; Devroye, 1987; Eggermont & LaRiccia, 2001; for exhaustive studies). To make the theorem useful, we also need to assure that for $f \notin \mathcal{F}$, $\mathbb{E}_f T_n$ is not too small. Ideally, $\mathbb{E}_f T_n$ should be something of the order of $\inf_{g\in\mathcal{F}}\int|f-g|$, a strictly positive quantity. In particular, under very general conditions, theorem 1 implies consistency and exponentially vanishing probability of acceptance. We formulate this in the following immediate corollary.

Corollary 1
Assume the conditions of theorem 1 *and suppose* $\lim_{n\to\infty} b_n \to 0$ *and that for* $f \notin \mathcal{F}$, $\liminf_{n\to\infty}\mathbb{E}_f T_n = \inf_{g\in\mathcal{F}}\int|g-f|$. *Then the test is consistent* (*that is*, *for any* $f \notin \mathcal{F}$ *the test rejects the null hypothesis eventually almost surely*), *and moreover*, *if n is so large that*

$$b_n + \sqrt{\frac{1}{\kappa_2 n}\log\frac{\kappa_1}{\alpha}} < \frac{1}{2}\inf_{g\in\mathcal{F}}\int|g-f| - \epsilon$$

for some $\epsilon > 0$ *and* $\mathbb{E}_f T_n \ge \inf_{g\in\mathcal{F}}\int|g-f| - \epsilon$, *then*

$$\mathbb{P}_f[\mathcal{H}_0 \textit{ is accepted}] \le \kappa_1\, \mathrm{e}^{-\kappa_2 n(\inf_{g\in\mathcal{F}}\int|g-f|)^2/4}.$$

The asymptotic condition on $\mathbb{E}_f T_n$ is satisfied under minimal assumptions on the smoothing factor as can be seen in the following lemma. Its proof (not detailed here) can be done using standard arguments and the L_1 universal consistency of the kernel density estimator (see Devroye & Györfi, 1985).

Lemma 1
Assume that $f \notin \mathcal{F}$ *and* (h_n) *is such that* $h_n \to 0$ *and* $nh_n \to \infty$ *almost surely as* $n \to \infty$. *Then* $\liminf_{n\to\infty}\mathbb{E}_f T_n = \inf_{g\in\mathcal{F}}\int|g-f| > 0$.

In the sequel, we offer two basic tools for proving non-asymptotic lower bounds for $\mathbb{E}_f T_n$. In both cases we relate $\mathbb{E}_f T_n$ to $\inf_{g\in\mathcal{F}}\mathbb{E}_f\int|f_{n,h}-g|$. More precisely, our aim is to understand under what conditions on h_n and $\mathcal{F}$ (and possibly f) we have

$$\mathbb{E}_f T_n \approx \inf_{g\in\mathcal{F}}\mathbb{E}_f\int|f_{n,h_n}-g|.$$

Lemma 1 shows that a desirable lower bound can be proven whenever h_n is such that the kernel estimator is stable. The second lemma offers a very different argument to show that

such lower bounding is possible if the class $\mathcal{F}$ is totally bounded in L_1. The notation $J_g \stackrel{\text{def}}{=} \int |f_{n,h} - g|$ is introduced for shortening.

Lemma 2
Let the auxiliary random variables $X_1', \ldots, X_n'$ be identically distributed with the X_is and independent of them. Let the kernel estimate defined by the auxiliary sample be denoted by f'_{n,h'_n} where $h'_n = h_n(X_1', \ldots, X_n')$. Then for any density f, and any class $\mathcal{F}$,

$$\mathbb{E}_f T_n \geq \inf_{g \in \mathcal{F}} \mathbb{E}_f \int |f_{n,h_n} - g| - \mathbb{E}_f \int |f'_{n,h'_n} - f_{n,h_n}|.$$

Proof. The proof is based on a symmetrization argument. Denote $J'_g \stackrel{\text{def}}{=} \int |f'_{n,h'} - g|$. Then,

$$\begin{aligned}
\inf_{g \in \mathcal{F}} \mathbb{E}_f \int |f_{n,h_n} - g| - \mathbb{E}_f T_n &= \mathbb{E}_f \left(\inf_{g \in \mathcal{F}} \mathbb{E}_f J_g - \inf_{g \in \mathcal{F}} J_g \right) \\
&\leq \mathbb{E}_f \sup_{g \in \mathcal{F}} (\mathbb{E}_f J_g - J_g) \\
&= \mathbb{E}_f \sup_{g \in \mathcal{F}} (\mathbb{E}_f (J'_g - J_g | X_1, \ldots, X_n)) \\
&\leq \mathbb{E}_f \sup_{g \in \mathcal{F}} (J'_g - J_g) \\
&\leq \mathbb{E}_f \int |f'_{n,h'} - f_{n,h}|,
\end{aligned}$$

which concludes the proof.

Remark. The term $\mathbb{E}_f \int |f'_{n,h'_n} - f_{n,h_n}|$ appearing in the lemma is independent of the size of the class $\mathcal{F}$, and it measures the stability of the kernel estimator. It may be bounded further by $2\mathbb{E}_f \int |f_{n,h_n} - \mathbb{E}_f f_{n,h_n}|$, twice the 'variation term' of the L_1 error. If h_n is independent of the data, the behaviour of this term is well understood. For example, if f is of bounded support of length $s(f)$, then the variation term is bounded by

$$\frac{\sqrt{s(f) + 2h_n} \sqrt{\int K^2}}{\sqrt{nh_n}}$$

(see e.g. Devroye & Lugosi, 2000). On the other hand, it is easy to see that no non-trivial density-free upper bound exists for $\mathbb{E}_f \int |f'_{n,h'_n} - f_{n,h_n}|$, even if h_n is independent of the data. In contrast to this, the next lemma provides a density-free bound which is only meaningful if the class $\mathcal{F}$ is not too 'large'.

The bound we offer next involves the covering numbers of the class $\mathcal{F}$. The *ϵ-covering number* of a class of densities $\mathcal{F}$ is the smallest integer $N_{\mathcal{F}}(\epsilon) = N$ such that there exist N densities $f_1, \ldots, f_N$ such that for all $f \in \mathcal{F}$, $\min_{j \leq N} \int |f - f_j| \leq \epsilon$. If no such integer exists, then we say that $N_{\mathcal{F}}(\epsilon) = \infty$.

Lemma 3
Assume that the smoothing factor h_n is such that, for all densities f and g, and any $\delta > 0$, $\mathbb{P}_f [\mathbb{E}_f J_g - J_g > \delta] \leq \kappa_1 \, \mathrm{e}^{-\kappa_2 n \delta^2}$, where κ_1, κ_2 are constants. Then for any density f,

$$\mathbb{E}_f T_n \geq \inf_{g \in \mathcal{F}} \mathbb{E}_f \int |f_{n,h_n} - g| - \inf_{\epsilon > 0} \left(2\epsilon + \sqrt{\frac{\log(\kappa_1 e N_{\mathcal{F}}(\epsilon))}{\kappa_2 n}} \right).$$

Proof. Fix $\epsilon > 0$ and let the densities $g_1, \ldots, g_N$ form an ϵ-covering of $\mathcal{F}$, where $N = N_{\mathcal{F}}(\epsilon)$. Then, just like in lemma 2,

$$\inf_{g \in \mathcal{F}} \mathbb{E}_f \int |f_{n,h_n} - g| - \mathbb{E}_f T_n \leq \mathbb{E}_f \sup_{g \in \mathcal{F}} (\mathbb{E}_f J_g - J_g).$$

The right-hand side, in turn, may be bounded, using the triangle inequality twice, as

$$\mathbb{E}_f \sup_{g \in \mathcal{F}} (\mathbb{E}_f J_g - J_g) \leq 2\epsilon + \mathbb{E}_f \max_{j=1,\ldots,N} (\mathbb{E}_f J_{g_j} - J_{g_j}).$$

By assumption, and using the union-of-events bound, for any $\delta > 0$,

$$\mathbb{P}_f \left[\max_{j=1,\ldots,N} (\mathbb{E}_f J_{g_j} - J_{g_j}) > \delta \right] \leq N \kappa_1 \, \mathrm{e}^{-\kappa_2 n \delta^2}.$$

As for any non-negative random variable Z and positive number u,

$$\mathbb{E}Z \leq \sqrt{\mathbb{E}(Z^2)} = \left(\int_0^\infty \mathbb{P}[Z^2 > \delta] \, \mathrm{d}\delta \right)^{1/2} \leq \left(u + \int_u^\infty \mathbb{P}[Z^2 > \delta] \, \mathrm{d}\delta \right)^{1/2}$$

we may integrate the above inequality and optimize the bound in u to obtain the desired inequality.

We end this section by investigating the assumptions of concentration in theorem 1 and also in lemma 3. We begin by pointing out that if h_n does not depend on the data then these conditions are indeed satisfied.

Lemma 4

Assume that the smoothing factor h_n may depend on the sample size n and on the class $\mathcal{F}$ but not on the data $X_1, \ldots, X_n$. Then for all f and $\epsilon > 0$,

$$\mathbb{P}_f[|T_n - \mathbb{E}_f T_n| > \epsilon] \leq 2 \, \mathrm{e}^{-n\epsilon^2/2}.$$

Moreover, for any density g,

$$\mathbb{P}_f[\mathbb{E}_f J_g - J_g > \delta] \leq \mathrm{e}^{-n\delta^2/2}.$$

Proof. The proof of both inequalities is based on a well-known concentration inequality due to McDiarmid (1989), and is similar to arguments of Devroye (1991).

We only prove the first inequality, the second is similar. Write

$$\phi(x_1, \ldots, x_n) = T_n = \inf_{g \in \mathcal{F}} \int \left| \frac{1}{n} \sum_{i=1}^n K_{h_n}(x - x_i) - g \right|.$$

To apply McDiarmid's inequality, we only need to show that the difference of the function ϕ evaluated at two vectors that only differ in their ith component can be bounded by $2/n$ for all $i = 1, \ldots, n$. Indeed, writing $f^{(i)}_{n,h_n}$ for the kernel density estimate obtained by replacing x_i by x_i' (but leaving the other data points unchanged), and g' for the density minimizing $\int |f^{(i)}_{n,h_n} - g|$ over $g \in \mathcal{F}$, we obtain

$$\begin{aligned} &\phi(x_1, \ldots, x_n) - \phi(x_1, \ldots, x_i', \ldots, x_n) \\ &\leq \int |f_{n,h_n} - g'| - \int \left| f^{(i)}_{n,h_n} - g' \right| \end{aligned}$$

$$\leq \int |f_{n,h_n} - f_{n,h_n}^{(i)}|$$
$$= \frac{1}{n}\int |K_{h_n}(x - x_i) - K_{h_n}(x - x_i')|\, \mathrm{d}x$$
$$\leq \frac{2}{n}$$

and the proof is complete.

Very often, the smoothing factor depends on the data but it is concentrated, with high probability, around its expected value. We will face such situations in several specific examples below. Here we show that concentration of h_n implies the concentration of T_n (and also that of J_g) as required by the conditions of theorem 1 and lemma 3. For simplicity, we assume that the kernel K is of bounded support. This assumption may be weakened easily.

Lemma 5

Let $h_n = h_n(X_1, \ldots, X_n)$ be an arbitrary data-dependent smoothing factor and let $\bar{h}_n$ be a deterministic sequence of positive numbers (e.g. one may take $\bar{h}_n = \mathbb{E}_f h_n$). Assume that K is supported in the interval $[-1/2, 1/2]$ and that K is Lipschitz with constant L. Writing

$$Z_n = \max\left(\frac{|h_n - \bar{h}_n|}{\bar{h}_n}, \frac{|h_n - \bar{h}_n|}{h_n}\right),$$

we have, for every $\epsilon \geq 3(L/4 + 1)\mathbb{E}Z_n$,

$$\mathbb{P}_f[|T_n - \mathbb{E}_f T_n| > \epsilon] \leq 2\, \mathrm{e}^{-n\epsilon^2/18} + \mathbb{P}_f\left[Z_n > \frac{\epsilon}{3(L/4 + 1)}\right].$$

Proof. Define the uncomputable version of the test statistic based on the deterministic smoothing factor $\bar{h}_n$ by $\bar{T}_n = \inf_{g \in \mathcal{F}} \int |f_{n,\bar{h}_n} - g|$. Then

$$\mathbb{P}_f[|T_n - \mathbb{E}_f T_n| > \epsilon] \leq \mathbb{P}_f\left[|T_n - \bar{T}_n| > \frac{\epsilon}{3}\right] + \mathbb{P}_f\left[|\bar{T}_n - \mathbb{E}_f \bar{T}_n| > \frac{\epsilon}{3}\right] + \mathbb{I}\left\{|\mathbb{E}_f \bar{T}_n - \mathbb{E}_f T_n| > \frac{\epsilon}{3}\right\}.$$

By lemma 4 the middle term is bounded by $2\,\mathrm{e}^{-n\epsilon^2/18}$. Define the sequence $a_n = \min(h_n/\bar{h}_n, \bar{h}_n/h_n)$, then,

$$|T_n - \bar{T}_n| \leq \frac{1}{n}\sum_{i=1}^{n} \int \left|\left(K_{h_n}(x - X_i) - K_{\bar{h}_n}(x - X_i)\right)\right| \mathrm{d}x$$
$$= \int |K(x) - a_n K(x a_n)|\, \mathrm{d}x$$
$$\leq \int |K(x) - K(x a_n)|\, \mathrm{d}x + \int |K(x a_n) - a_n K(x a_n)|\, \mathrm{d}x$$
$$\leq \frac{L}{4}|1 - a_n| + \left|1 - \frac{1}{a_n}\right|$$
$$\leq \left(\frac{L}{4} + 1\right) \max\left(|1 - a_n|, \left|1 - \frac{1}{a_n}\right|\right).$$

The statement now follows.

3. Minimum distance smoothing factor

In this section, we investigate a general principle of selecting the smoothing factor h_n. For any class $\mathcal{F}$, we may define the smoothing factor by

$$h_n = \arg\min_{h \in A} \inf_{g \in \mathcal{F}} \int |f_{n,h} - g|$$

where A is some fixed subset of $(0, \infty)$. Very often $A = (0, \infty)$ but sometimes it may be convenient to restrict the choice either for theoretical or for computational reasons. For simplicity we assume that the minimum exists. Otherwise straightforward modifications yield essentially identical results. This choice of h_n leads to the test statistic

$$T_n = \inf_{h \in A} \inf_{g \in \mathcal{F}} \int |f_{n,h} - g|.$$

An important feature of this choice is that the test statistic is sharply concentrated around its mean, regardless of the class $\mathcal{F}$. This property will allow us to use theorem 1 in a convenient way. The final conclusion is presented in the following lemma. Its proof is omitted as it is similar to that of lemma 4.

Lemma 6
With the minimum-distance choice of the smoothing factor we have, for all f,

$$\mathbb{P}_f[|T_n - \mathbb{E}T_n| > \epsilon] \leq 2\, \mathrm{e}^{-n\epsilon^2/2}.$$

Thus, for all choices of $\mathcal{F}$ and A, theorem 1 may be applied easily. According to the theorem (and, in particular, corollary 1), the test performs well if b_n is small (ideally converging to zero at a fast rate) and $\mathbb{E}_f T_n$ is large whenever $f \notin \mathcal{F}$.

To bound b_n, simply recall that by the definition of h_n,

$$\begin{aligned} b_n &= \sup_{f \in \mathcal{F}} \mathbb{E}_f \inf_{g \in \mathcal{F}} \int |f_{n,h_n} - g| \\ &= \sup_{f \in \mathcal{F}} \mathbb{E}_f \inf_{g \in \mathcal{F}} \inf_{h \in A} \int |f_{n,h} - g| \\ &\leq \sup_{f \in \mathcal{F}} \inf_{h \in A} \mathbb{E}_f \int |f_{n,h} - f|. \end{aligned}$$

Thus, for this choice of the smoothing factor, b_n is always bounded by the largest expected error of the kernel estimator in the class, with optimally chosen smoothing factor. This quantity may typically be bounded easily if the class $\mathcal{F}$ is not too large. For example, if the class $\mathcal{F}$ is such that every $f \in \mathcal{F}$ has a bounded support with length $s(f)$ and is absolutely continuous with an absolutely continuous first derivative such that $c(f) = \int |f''| < \infty$ such that $\sup_{f \in \mathcal{F}} s^2(f) c(f) < \infty$ then, assuming that K is bounded, symmetric, and of bounded support, $b_n \leq C n^{-2/5}$ for some constant C (see Devroye & Lugosi, 2000).

To complete the analysis of the test based on the minimum-distance smoothing factor, one needs to understand the behaviour of $\mathbb{E}_f T_n$ when $f \notin \mathcal{F}$. To this end, we offer the following version of lemma 3.

Lemma 7
For any density f,

$$\mathbb{E}_f \inf_{g\in\mathcal{F}} \inf_{h\in A} \int |f_{n,h} - g| \geq \inf_{g\in\mathcal{F}} \mathbb{E}_f \inf_{h\in A} \int |f_{n,h} - g| - \inf_{\epsilon>0} \left(2\epsilon + \sqrt{\frac{2\log N_{\mathcal{F}}(\epsilon)}{n}} \right).$$

Proof. Let the densities $g_1, \ldots, g_N$ form an ϵ covering of $\mathcal{F}$, where $N = N_{\mathcal{F}}(\epsilon)$. Then, as in the proof of lemma 3,

$$\begin{aligned} &\inf_{g\in\mathcal{F}} \mathbb{E}_f \inf_{h\in A} \int |f_{n,h} - g| - \mathbb{E}_f \inf_{g\in\mathcal{F}} \inf_{h\in A} \int |f_{n,h} - g| \\ &\quad \leq 2\epsilon + \mathbb{E}_f \max_{j\leq N} \left(\mathbb{E}_f \inf_{h\in A} \int |f_{n,h} - g_j| - \inf_{h\in A} \int |f_{n,h} - g_j| \right). \end{aligned}$$

Denote $Z_j = \mathbb{E}_f \inf_{h\in A} \int |f_{n,h} - g_j| - \inf_{h\in A} \int |f_{n,h} - g_j|$. Then again, it follows by McDiarmid's inequality that for all $j \leq N$ and $s > 0$, $\mathbb{E}_f \, e^{sZ_j} \leq e^{s^2/2n}$, and therefore, using a Jensen's inequality,

$$\begin{aligned} \mathbb{E}_f \max_{j\leq N} Z_j &= \frac{1}{s} \log e^{s\mathbb{E}_f \max_{j\leq N} Z_j} \leq \frac{1}{s} \log \mathbb{E}_f \max_{j\leq N} e^{sZ_j} \\ &\leq \frac{1}{s} \log \mathbb{E}_f \sum_{j=1}^{N} e^{sZ_j} \leq \frac{\log N}{s} + \frac{s}{2n} \\ &= \sqrt{2\frac{\log N}{n}} \qquad \text{(by choosing } s = \sqrt{2n\log N}\text{)}. \end{aligned}$$

This concludes the proof.

The term

$$C_n(\mathcal{F}) = \inf_{\epsilon>0} \left(2\epsilon + \sqrt{\frac{2\log N_{\mathcal{F}}(\epsilon)}{n}} \right)$$

may be considered as a measure of *complexity* of the class $\mathcal{F}$. Clearly, for any totally bounded class, $C_n(\mathcal{F}) \to 0$ as $n \to \infty$. Also, for many classes, it is easy to obtain explicit estimates for the covering numbers. Several examples may be found in Kolmogorov & Tikhomirov (1961) (see also Devroye, 1987; Devroye & Lugosi, 2000). For example, if $\mathcal{F}$ is any class of Lipschitz densities supported in [0,1] with Lipschitz constant C, then, using an estimate from Devroye (1987), we have

$$C_n(\mathcal{F}) \leq (nC)^{-1/3} (\log 3)^{1/3} (2^{-2/3} + 2^{-4/3}).$$

Summarizing the arguments above, we obtain the following corollary of Theorem 1.

Corollary 2
Let $\mathcal{F}$ be a totally bounded class of densities with complexity $C_n(\mathcal{F})$ and consider the test based on the minimum-distance smoothing factor. Then

$$c_\alpha \leq \sup_{f\in\mathcal{F}} \inf_{h\in A} \mathbb{E}_f \int |f_{n,h} - f| + \sqrt{\frac{2}{n} \log \frac{2}{\alpha}}$$

and for any density $f \notin \mathcal{F}$ and $\delta > 0$, if

$$\inf_{g\in\mathcal{F}} \mathbb{E}_f \inf_{h\in A} \int |f_{n,h} - g| > \sup_{f'\in\mathcal{F}} \inf_{h\in A} \mathbb{E}_{f'} \int |f_{n,h} - f'| + C_n(\mathcal{F}) + \sqrt{\frac{2}{n}\log\frac{2}{\alpha}} + \delta,$$

the probability of acceptance is at most 2 $\mathrm{e}^{-n\delta^2/2}$.

In the rest of this section we show that the test based on the minimum-distance smoothing factor is consistent for all f under the only assumption that $\mathcal{F}$ is totally bounded, that is, $C_n(\mathcal{F}) = o(1)$. To achieve this, we need to restrict the minimum-distance smoothing factor such that $A \subset [a_n, b_n]$ where $\{a_n\}$ and $\{b_n\}$ are arbitrary sequences of positive numbers such that $a_n \leq b_n$, $na_n \to \infty$, and $b_n \to 0$.

Corollary 3

Let $\mathcal{F}$ be a totally bounded class of densities and consider the test based on the minimum-distance smoothing factor with $A \subset [a_n, b_n]$. Then for any density $f \notin \mathcal{F}$, the hypotheses $f \in \mathcal{F}$ is rejected, almost surely, as $n \to \infty$.

Proof. In order to make sure that the test is consistent almost surely, one may choose δ to be the order of $n^{-1/2+\epsilon}$ for some small positive ϵ. Then by corollary 2, consistency can be proven if (i) $\sup_{f\in\mathcal{F}} \inf_{h\in A} \mathbb{E}_f \int |f_{n,h} - f|$ converges to zero, and (ii) $\inf_{g\in\mathcal{F}} \mathbb{E}_f \inf_{h\in A} \int |f_{n,h} - g|$ stays bounded away from zero.

To prove (i), fix any $\epsilon > 0$, and let $f_1, \ldots, f_{N_{\mathcal{F}}(\epsilon)}$ be any ϵ-covering of $\mathcal{F}$. Let $f \in \mathcal{F}$, and assume that $i \leq N_{\mathcal{F}}(\epsilon)$ is such that $\int |f - f_i| \leq \epsilon$. Then, if h_i denotes the smoothing factor in A minimizing $\mathbb{E}_{f_i} \int |f_{n,h_i} - f_i|$, then

$$\begin{aligned}
\inf_{h\in A} \mathbb{E}_f \int |f_{n,h} - f| &\leq \mathbb{E}_f \int |f_{n,h_i} - f| \\
&\leq \mathbb{E}_{f_i} \int |f_{n,h_i} - f_i| + 2\epsilon \\
&= \inf_{h\in A} \mathbb{E}_{f_i} \int |f_{n,h} - f_i| + 2\epsilon
\end{aligned}$$

where the second inequality follows by the *embedding device* of Devroye (1987, pp. 46–47). Thus,

$$\begin{aligned}
\limsup_{n\to\infty} \sup_{f\in\mathcal{F}} \inf_{h\in A} \mathbb{E}_f \int |f_{n,h} - f| &\leq 2\epsilon + \limsup_{n\to\infty} \max_{i=1,\ldots,N_{\mathcal{F}}(\epsilon)} \inf_{h\in A} \mathbb{E}_{f_i} \int |f_{n,h} - f_i| \\
&= 2\epsilon
\end{aligned}$$

by the L_1 consistency of the kernel estimator. As ϵ is arbitrary, (i) is proven.

To show (ii), note that by an application of the triangle inequality,

$$\inf_{g\in\mathcal{F}} \mathbb{E}_f \inf_{h\in A} \int |f_{n,h} - g| \geq \inf_{g\in\mathcal{F}} \int |f - g| - \mathbb{E}_f \sup_{h\in A} \int |f_{n,h} - f|.$$

As the choice of the minimum-distance smoothing factor is restricted such that $A \subset [a_n, b_n]$ where $na_n \to \infty$ and $b_n \to 0$, Devroye & Györfi (1985, p. 148), implies that $\mathbb{E}_f \sup_{h\in A} \int |f_{n,h} - f| \to 0$ and, indeed,

$$\liminf_{n\to\infty} \inf_{g\in\mathcal{F}} \mathbb{E}_f \inf_{h\in A} \int |f_{n,h} - g| \geq \inf_{g\in\mathcal{F}} \int |f - g| > 0$$

which implies the universal consistency of the test.

4. Applications

In this section, three particular cases are considered: testing a simple null hypothesis, testing for translation/scale families and testing for symmetry.

4.1. Simple hypotheses

As a first application, we consider the simplest case when $\mathcal{F}$ contains a single nominal density f_0, that is, $\mathcal{F} = \{f_0\}$. In other words, one has to test the simple null hypothesis whether the data $X_1, \ldots, X_n$ are distributed according to f_0 or not.

In this case, it is natural to define the smoothing factor h_n to minimize the expected L_1 error, that is,

$$h_n = \arg\min_{h>0} \mathbb{E}_{f_0} \int |f_{n,h} - f_0|.$$

Note that h_n is not data-dependent, it may be computed by Monte-Carlo approximation even before seeing the data. Similarly, a Monte-Carlo approximation of the c_α is straightforward since $c_\alpha = \inf\{c : \mathbb{P}_{f_0}[T_n > c] \leq \alpha\}$ in this case.

For example, once h_n has been calculated, the following simple Monte-Carlo (or bootstrap) approach can be used to calculate c_α : (i) draw an artificial sample of size n from f_0, say $\overrightarrow{X}^*$; (ii) compute the kernel estimator with this new sample, say f^*_{n,h_n}, and approximate numerically the 'bootstrap' L_1-distance: $T_n^* = \int |f^*_{n,h_n} - f_0|$; (iii) repeat steps (i)–(ii) a large number of times (call it B) to obtain the 'bootstrap' replications T_n^{*1}, $T_n^{*2}, \ldots, T_n^{*B}$; (iv) sort these values and define $c_\alpha^* = T_n^{*(\lceil B(1-\alpha)\rceil)}$, i.e. the $\lceil B(1-\alpha)\rceil$th order statistic of the bootstrap replications.

This c_α^* is the approximation that will be used for the true value c_α. It is clear that in the case of simple null hypothesis, c_α^* is a consistent estimator (as $B \to \infty$) of c_α and hence the practical level of the test approaches to the nominal α provided that B is large (and we can choose B with the 'only' price of computing time!)

Now theorem 1 can be applied directly, together with lemma 4. We obtain the following corollary.

Corollary 4
Let $\bar{b}_n$ denote the expected L_1 error of the kernel density estimator of f_0, based on the optimal bandwidth h_n:

$$\bar{b}_n \stackrel{\text{def}}{=} \inf_{h>0} \mathbb{E}_{f_0} \int |f_{n,h} - f_0|.$$

Then for all $\alpha > 0$, $c_\alpha \leq \bar{b}_n + \sqrt{(2/n)\log(2/\alpha)}$ and for any density f and $\delta > 0$, if

$$\mathbb{E}_f \int |f_{n,h_n} - f_0| > \bar{b}_n + \delta + \sqrt{\frac{2}{n}\log\frac{2}{\alpha}},$$

then

$$\mathbb{P}_f[\mathcal{H}_0 \text{ is accepted}] \leq 2\, \mathrm{e}^{-n\delta^2/2}.$$

Also, for any f_0, $\lim_{n\to\infty} \bar{b}_n = 0$ and the test is consistent for all f.

Obviously, a sufficient condition for the lower bound of the expected L_1 error in the previous corollary is $\int |\mathbb{E} f_{n,h} - f_0| > \bar{b}_n + \delta + \sqrt{(2/n)\log(2/\alpha)}$. We remark here that in the case of a simple null hypothesis, consistency of a similar test based on the histogram density estimator was shown by Györfi & van der Meulen (1991).

The corollary shows that the performance of the test depends on how well the density f_0 can be estimated by the kernel estimate. The smaller the $\bar{b}_n$, the more efficient the test is. In most hypothesis testing problems f_0 is sufficiently regular so that $\bar{b}_n$ is of the order of $n^{-2/5}$. However, if f_0 is either unsmooth or heavy-tailed, the kernel estimator is poor and for a finite sample size, only densities far from f_0 (in the L_1 distance) will be rejected with a high probability. In such cases, by a simple modification of the test, the performance may be improved significantly. The idea is to first transform the data $X_1, \ldots, X_n$ by a transformation T so that if the density of the X_i was f_0 then the transformed data have a density which is easy to estimate with the kernel estimator, and the proposed test is performed on the transformed data. Some densities are well-known to be very easy to estimate with the kernel density estimator. One such example is the triangular density $(1 - |x|)_+$ if the Epanechnikov kernel $K(x) = (3/4)(1 - x^2)_+$ is used. In this case, $\bar{b}_n \leq 1.177 \cdot n^{-2/5}$ (see Devroye & Györfi, 1985) and the fact that the L_1 error is invariant under monotone transformations implies the following.

Corollary 5

Consider the simple null hypothesis that the data are distributed according to the density f_0 where f_0 is an arbitrary density on $\mathbb{R}$. Let T be the unique monotone increasing function with the property that if the random variable X has density f_0 then $T(X)$ has density $f^*(x) = (1 - |x|)_+$. Consider the test which first transforms the data to $Y_1, \ldots, Y_n$, where $Y_i = T(X_i)$ for all $i = 1, \ldots, n$ and then accepts the null hypotheses if and only if $T_n^* \leq c_\alpha^*$ where T_n^* and c_α^* are defined as above (using the Epanechnikov kernel) but for the null hypothesis that the Y_i have density f^*. Then for all $\alpha > 0$, $c_\alpha \leq 1.177 \cdot n^{-2/5} + \sqrt{(2/n)\log(2/\alpha)}$ and for any density f and $\delta > 0$, if

$$\mathbb{E}_f \int |f_{n,h} - f_0| > 1.177 \cdot n^{-2/5} + \delta + \sqrt{\frac{2}{n}\log\frac{2}{\alpha}},$$

then

$$\mathbb{P}_f[\mathcal{H}_0 \text{ is accepted}] \leq 2\, \mathrm{e}^{-n\delta^2/2}.$$

4.2. Translation/scale classes

In this section, we consider the problem of testing whether the underlying density of the data may be obtained by translation and/or scaling of a fixed density f_0.

To begin, observe that if $\mathcal{F}$ is a simple translation class, that is,

$$\mathcal{F} = \{f(x) = f_0(x - a) : a \in \mathbb{R},\},$$

for some nominal density f_0, then the translation-invariance of the kernel estimator immediately implies that the value of c_α remains the same as in the case of the simple null hypothesis $\mathcal{F} = \{f_0\}$ and moreover corollary 4 remains true in this case as well.

The case of translation/scale class is somewhat more interesting. In other words, we consider classes of the form

$$\mathcal{F} = \left\{ f(x) = \frac{1}{c} f_0\left(\frac{x - a}{c}\right) : a \in \mathbb{R}, c > 0 \right\}$$

where f_0 is a fixed density. For example, if the goal is to test normality, one may take f_0 to be the standard normal density.

Clearly, as the scale of the density is arbitrary, a data-dependent choice of the bandwidth h_n is necessary. To determine a good data-dependent choice, note that if $h_n^* = \arg\min_{h>0} \mathbb{E}_{f_0} \int |f_{n,h} - f_0|$ is the optimal smoothing factor for the density f_0, then ch_n^* is the optimal bandwidth for the scaled and translated density $(1/c)f_0((\cdot - a)/c)$. As h_n^* can be

computed before seeing the data (just like in the case of simple hypothesis), all one needs to estimate is the scaling factor c. Recall that by lemma 5 the estimated bandwidth should be concentrated around its mean value. One simple way to achieve this is to estimate the scale based on order statistics. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the ordered sample $X_1, \ldots, X_n$, and introduce the estimator

$$c_n = \frac{X_{(\lfloor 3n/4 \rfloor)} - X_{(\lfloor n/4 \rfloor)}}{q^{(0)}_{3/4} - q^{(0)}_{1/4}}$$

where $q^{(0)}_{1/4}$ and $q^{(0)}_{3/4}$ denote the 25th and 75th percentiles of the density f_0. The values 1/4 and 3/4 do not have any special role, they may be replaced by any pair of different numbers in (0,1). The only property we need for the estimator to work is that the density f_0 is bounded away from zero in a neighbourhood of $q^{(0)}_{3/4}$ and $q^{(0)}_{1/4}$. If this property is not satisfied, the values 1/4 and 3/4 should be modified appropriately.

Based on c_n, we define the data-based smoothing factor by $h_n = c_n h^*_n$. Thus, if $\bar{h}_n = c h^*_n$ denotes the optimal smoothing factor for the scaled density $(1/c) f_0((\cdot - a)/c)$, then, under mild conditions on the true underlying density f, we have the following stability property for the estimated smoothing factor.

Lemma 8
Let $0 < \epsilon \leq 1$. Assume that the density f is such that there exists a positive number τ such that $f(x) \geq \tau$ for all $x \in (q_{1/4} - \delta, q_{1/4} + \delta) \cup (q_{3/4} - \delta, q_{3/4} + \delta)$ where q_p denotes the $100p$ percentile of f and $\delta = (q_{3/4} - q_{1/4})\epsilon/2$. Define

$$\bar{h}_n = h^*_n \frac{q_{3/4} - q_{1/4}}{q^{(0)}_{3/4} - q^{(0)}_{1/4}}.$$

Then

$$\mathbb{P}_f[Z_n > \epsilon] \leq 12\, \mathrm{e}^{-n\epsilon^2\tau^2(q_{3/4}-q_{1/4})^2/8},$$

where Z_n is the random variable defined in lemma 5.

Proof. First observe that

$$\begin{aligned}
\mathbb{P}_f\left[\left|\frac{h_n - \bar{h}_n}{\bar{h}_n}\right| > \epsilon\right] &= \mathbb{P}_f\left[\left|\frac{X_{(\lfloor 3n/4 \rfloor)} - X_{(\lfloor n/4 \rfloor)}}{q_{3/4} - q_{1/4}} - 1\right| > \epsilon\right] \\
&\leq \mathbb{P}_f\left[|X_{(\lfloor 3n/4 \rfloor)} - q_{3/4}| > \frac{(q_{3/4} - q_{1/4})\epsilon}{2}\right] \\
&\quad + \mathbb{P}_f\left[|X_{(\lfloor n/4 \rfloor)} - q_{1/4}| > \frac{(q_{3/4} - q_{1/4})\epsilon}{2}\right].
\end{aligned}$$

Both terms on the right-hand side can be bounded similarly. For example, the first term may be written as a sum of two terms (one for the lower tail, one for the upper tail). Now using the assumption of the positivity of f in the neighbourhood of its 75th percentile, for the upper tail of the first term we obtain

$$\begin{aligned}
\mathbb{P}_f\left[X_{(\lfloor 3n/4 \rfloor)} - q_{3/4} > \frac{(q_{3/4} - q_{1/4})\epsilon}{2}\right] &\leq \mathbb{P}_f[F(q_{3/4} + \delta) - F_n(q_{3/4} + \delta) > \delta\tau] \\
&\leq \mathrm{e}^{-2n\delta^2\tau^2}
\end{aligned}$$

where F and F_n denote the cumulative distribution function of f and its empirical counterpart, respectively. The second inequality follows by Hoeffding's (1963) inequality for the tail of the binomial distribution. Thus, we have

$$\mathbb{P}_f\left[\left|\frac{h_n-\bar{h}_n}{\bar{h}_n}\right|>\epsilon\right]\le 4\,\mathrm{e}^{-n\epsilon^2\tau^2(q_{3/4}-q_{1/4})^2/2}.$$

Using the fact that $\epsilon\le 1$, we may similarly bound the tail probabilities of $(h_n-\bar{h}_n)/h_n$ by observing that

$$\begin{aligned}
&\mathbb{P}_f\left[\left|\frac{h_n-\bar{h}_n}{h_n}\right|>\epsilon\right]\\
&\quad\le \mathbb{P}_f\left[\left|\left(q_{3/4}-q_{1/4}\right)-\left(X_{(\lfloor 3n/4\rfloor)}-X_{(\lfloor n/4\rfloor)}\right)\right|>\frac{\left(q_{3/4}-q_{1/4}\right)\epsilon}{2}\right]\\
&\qquad+\mathbb{P}_f\left[\left(X_{(\lfloor 3n/4\rfloor)}-X_{(\lfloor n/4\rfloor)}\right)<\frac{q_{3/4}-q_{1/4}}{2}\right]\\
&\quad\le 2\mathbb{P}_f\left[\left|\left(q_{3/4}-q_{1/4}\right)-\left(X_{(\lfloor 3n/4\rfloor)}-X_{(\lfloor n/4\rfloor)}\right)\right|>\frac{\left(q_{3/4}-q_{1/4}\right)\epsilon}{2}\right].
\end{aligned}$$

Now the probability on the right-hand side may be bounded the same way as above. Collecting terms, we obtain the lemma.

In the corollary below we summarize the results of this lemma together with theorem 1, and lemmas 1 and 5 for the case when $\mathcal{F}$ is the class of all scaled translations of a fixed nominal density f_0. The corollary states that the test is consistent for *all* f_0 and f and moreover provides exponential estimates for the probability of acceptance under the only assumption of f that it is strictly positive in a neighbourhood of its 25th and 75th percentiles. We remark here that this assumption may easily be dropped at a price of a slightly more complicated bound but we do not pursue this issue further.

Corollary 6

Let f_0 be a fixed density and consider the test described in this section for deciding whether the density of the data is of the form $f(x)=(1/c)f_0((x-a)/c)$ for some $a\in\mathbb{R}$ and $c>0$.

Recall the definition of $\bar{b}_n$ in corollary 4. Then for any f_0, $\lim_{n\to\infty}\bar{b}_n=0$ and the test is consistent for all f.

Assume further that the kernel K is supported in the interval $[-1/2,\,1/2]$ and it is Lipschitz with constant L and let f satisfy the assumption of lemma 8. Define

$$\kappa_2=\min\left(\frac{1}{18},\frac{\left(q_{3/4}-q_{1/4}\right)^2\tau^2}{72(L/4+1)^2}\right).$$

Then for all $\alpha>0$, $c_\alpha\le\bar{b}_n+\sqrt{(1/\kappa_2 n)\log(14/\alpha)}$ and for any $\delta>0$, if

$$\mathbb{E}_f\int|f_{n,h}-f_0|>\bar{b}_n+\delta+\sqrt{\frac{1}{\kappa_2 n}\log\frac{14}{\alpha}},$$

then

$$\mathbb{P}_f[\mathcal{H}_0\ \textit{is accepted}]\le 14\,\mathrm{e}^{-\kappa_2 n\delta^2}.$$

Remark. Note that, as expected, the factor $(q_{3/4}-q_{1/4})^2\,\tau^2$ is translation and scale-free in the sense that its value does not change if the density f is changed to $(1/c)f((\cdot-a)/c)$ for any $c>0$ and $a\in\mathbb{R}$.

The corollary reveals that even though the problem of testing the composite hypothesis of a translation/scale class is considerably more difficult than testing a simple hypothesis as in the previous section, the main result we obtained above is comparable with corollary 4. In other words, we do not pay a high price for having to use a data-dependent smoothing factor. Note

that the bound for c_α is almost identical in both cases and it is determined by the error of the kernel estimator for f_0. In the case of a simple hypothesis we could boost the performance of the test by transforming the data such that the error of the kernel estimator was small. Here we cannot do this directly. However, a similar trick may be adopted if one estimates the scale first (similarly as in the selection of the smoothing factor) and use a data-dependent transformation chosen from a class. We omit the straightforward but tedious analysis here.

4.3. Testing symmetry

In this section we investigate the testing procedure with the 'minimum-distance' choice of the smoothing factor proposed in the previous section, in the special case of testing symmetry of a density.

Thus, we define $\mathcal{F}$ as the class of all symmetric densities on the real line. Unfortunately, in this case the test based on the minimum-distance smoothing factor cannot work. The reason is that the class of all symmetric densities is 'too large' in the sense that in this case c_α is bounded from below by a positive constant independently of the sample size. However, by assuming some additional regularity conditions on the density f, meaningful results may be obtained.

As a simple example, we may consider a situation in which the statistician has a reason to assume that (say) the unknown density is Lipschitz, supported on $[-a, a]$, with Lipschitz constant C. Denote the class of such densities by $\mathcal{L}(-a, a, C)$. Then we let $\mathcal{F}$ be the class of all densities in $\mathcal{L}(-a, a, C)$ which are symmetric around some point. In this case, as it is pointed out in section 3, the complexity $C_n(\mathcal{F})$ is bounded by a constant times $(nC)^{-1/3}$, and corollary 2 is applicable.

Let us denote by $\mathcal{S}$ the class of all symmetric density functions and by $\mathcal{S}_m$ the subclass of symmetric densities around $m \in \mathbb{R}$. The computational issue of calculating T_n is simple as soon as one finds some device for finding the closest symmetric density to an arbitrary given density f. The next lemma provides such a device.

Lemma 9

For any density f, the infimum $\inf_{g\in\mathcal{S}_0}\int|f-g|$ is attained at any symmetric density f^s satisfying $f^\ell \le f^s \le f^u$, with $f^\ell(x) = \min\{f(x), f(-x)\}$, $f^u(x) = \max\{f(x), f(-x)\}$. In particular, $f^s = \frac{1}{2}(f^\ell + f^u)$ is the closest (possibly among others) symmetric density in the sense of the L_1 distance. Furthermore,

$$\inf_{g\in\mathcal{S}_0}\int|f-g| = \int|f^s - f| = \frac{1}{2}\int(f^u - f^\ell)$$
$$= \frac{1}{2}\int|f(x) - f(-x)|\,\mathrm{d}x = \int_0^\infty |f(x) - f(-x)|\,\mathrm{d}x.$$

Proof. First of all, given a density f, consider any symmetric density function, f^s, satisfying $f^\ell \le f^s \le f^u$. Then standard algebra gives

$$\int|f^s - f| = \int_{\mathbb{R}^+}(f^u - f^\ell) = \frac{1}{2}\int(f^u - f^\ell).$$

Now, consider any $g \in \mathcal{S}_0$ and define $A = \{x : g(x) < f^\ell(x)\}$, $B = \{x : f^\ell(x) \le g(x) \le f^u(x)\}$ and $C = \{x : g(x) > f^u(x)\}$, which are symmetric subsets of $\mathbb{R}$. Then, tedious but straightforward calculations lead to

$$\int|g - f| = \int_A (f^\ell - g) + \int_C (g - f^u) + \int_{\mathbb{R}^+}(f^u - f^\ell)$$
$$\ge \int_{\mathbb{R}^+}(f^u - f^\ell) = \frac{1}{2}\int(f^u - f^\ell) = \int|f^s - f|$$

and the proof is concluded.

As a consequence of the previous lemma, it is straightforward to compute the closest symmetric density, in the L_1 sense, to a given non-parametric kernel density estimator, $f_{n,h}$. A numerical minimization in h solves the problem then.

In the general case, as $\mathcal{S} = \cup_{m\in\mathbb{R}}\mathcal{S}_m$, standard manipulations can be used to show that

$$T_n = \inf_{m\in\mathbb{R},h>0} \int_m^\infty |f_{n,h}(x) - f_{n,h}(2m - x)| \, \mathrm{d}x$$

and the computation of the test statistic only needs the numerical minimization of a function of two real variables.

5. Simulations

A small simulation study has been carried out for the standard goodness-of-fit problem of testing normality. Thus, $\mathcal{F}$ is the class of all normal densities. As $\mathcal{F}$ is a translation/scale family we may compute a Monte-Carlo approximation of c_α by just drawing B samples from a standard normal distribution as described in section 4.1. The smoothing factor was selected by minimum L_1-distance as detailed in section 3. To do this, we restricted $\mathcal{F}$ to normal densities with mean and variance within compact intervals in order for the class to be totally bounded, as required by corollary 3. Of course this is also the case from the practical implementation viewpoint. In order for the kernel estimator to be sufficiently far from the class of densities (and to avoid degenerate minimum distance bandwidths) we used the Epanechnikov kernel for the test statistic T_n. In our simulation study we have chosen $B = 10{,}000$.

To examine the size and the power of the test, we selected either $f \in \mathcal{F}$ or $f \notin \mathcal{F}$. The final results offer the rejection percentages along the 10,000 trials used in the simulation. The nominal significance level selected was $\alpha = 0.05$.

For the null hypothesis scenario we used a standard normal distribution (model L_0), while we considered models L_1–L_6, six distributions of the lambda family (cases 1–5 and 7 already used in the simulation study by Fan, 1994) for the alternative. This family provides a wide range of distributions that are easily generated as the quantile function is given by

$$F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1 - u)^{\lambda_3}]/\lambda_2.$$

The particular choices for this four parameters used in models L_1–L_6 are listed in Table 1.

This study was carried out in order to compare our minimum L_1-distance approach with Fan's L_2-distance test based on the kernel method too. Fan's procedure is based on the test statistic

Table 1. *Parameter choices for models L_1–L_6*

Model	λ_1	λ_2	λ_3	λ_4
L_1	0	2	1	1
L_2	0	−0.397012	−0.16	−0.16
L_3	0	−1	−0.24	−0.24
L_4	0	1	1.4	0.25
L_5	3.586508	0.04306	0.025213	0.094029
L_6	−0.116734	−0.351663	−0.13	−0.16

$$\int \left(f_{n,h_n} - \hat{f}\right)^2,$$

where $\hat{f}$ is the maximum likelihood estimator of f under the hypothesized parametric model. As Fan did not give any automatic method for selecting the smoothing factor, she used several values in a range to explore the behaviour of her test statistic. Table 2 contains the range of rejection percentages obtained by Fan for different bandwidths in a reasonable interval (see Fan, 1994 for details). The Kolmogorov–Smirnov test was also included in the simulation as a standard competitor. It is based on the test statistic $\sup_{x\in\mathbb{R}} |F_n(x) - \hat{F}(x)|$, where $\hat{F}$ is the maximum likelihood estimator of the underlying distribution function F under the parametric model. The results for the three tests are summarized in Tables 2–4.

It is clearly seen that, in general, the minimum L_1-distance method performs better than Fan's test. The new method slightly outperforms Kolmogorov–Smirnov test in terms of power. However, it seems that the Kolmogorov–Smirnov test is a little closer to the nominal size of the test.

Similar conclusions have been drawn from other simulation results (not reported here) for testing normality under some normal mixture models already used by Marron & Wand (1992).

Table 2. *Rejection percentages of Fan's test*

Model	$n = 100$	$n = 200$	$n = 400$	$n = 800$
L_0	0.3–3.4	1.7–5.8	3.2–6.1	2.6–4.6
L_1	95.2–99.6			
L_2	0.1–0.9	3.1–8.6	48.8–62.1	87.4–93.4
L_3	1.6–5.7	20.8–38.9	90.3–95.6	
L_4	94.2–98.6			
L_5	9.9–16.3	63.1–66-7	99.3-99.4	
L_6	0.0–1.0	4.5–11.8	52.7-66.3	89.4–93.8

Table 3. *Rejection percentages of the minimum L_1-distance test*

Model	$n = 100$	$n = 200$
L_0	5.08	6.72
L_1	75.72	100
L_2	100	
L_3	100	
L_4	91.80	99.78
L_5	100	
L_6	100	

Table 4. *Rejection percentages of the Kolmogorov–Smirnov test*

Model	$n = 100$	$n = 200$
L_0	4.65	4.94
L_1	59.07	95.08
L_2	100	
L_3	100	
L_4	40.17	78.91
L_5	99.97	100
L_6	100	

Acknowledgements

The work of the first author was supported by the Spanish Ministry of Science and Technology (European Regional Development Funds included) and by the Xunta de Galicia. The work of the second author was supported by the Spanish Ministry of Science and Technology and the PASCAL Network of Excellence (European Commission).

References

Ahmad, I. A. & Cerrito, P. B. (1993). Goodness of fit tests based on the L_2-norm of multivariate probability density functions. *J. Nonparametr. Statist.* **2**, 169–181.

Akaike, H. (1954). An approximation to the density function. *Ann. Inst. Statist. Math.* **6**, 127–132.

Bickel, P. J. & Rosenblatt, M. (1973). On some global measures of the deviation of density function estimates. *Ann. Statist.* **1**, 1071–1095.

Devroye, L. (1987). *A course in density estimation*. Birkhäuser, Boston.

Devroye, L. (1991). Exponential inequalities in nonparametric estimation. In *Nonparametric functional estimation and related topics* (ed. G. Roussas), 31–44. NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

Devroye, L. & Györfi, L. (1985). *Nonparametric density estimation: the L_1 view*. John Wiley, New York.

Devroye, L. & Lugosi, G. (2000). *Combinatorial methods in density estimation*. Springer-Verlag, New York.

Devroye, L. & Lugosi, G. (2002). Almost sure classification of densities. *J. Nonparametr. Statist.* **14**, 675–698.

Eggermont, P. P. B. & LaRiccia, V. N. (2001). *Maximum penalized likelihood estimation, volume I: Density estimation*. Springer-Verlag, New York.

Fan, Y. (1994). Testing the goodness of fit of a parametric density function by kernel method. *Econ. Theory* **10**, 316–356.

Fan, Y. (1998). Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. *Econ. Theory* **14**, 604–621.

Györfi, L. & van der Meulen, E. C. (1991). A consistent goodness of fit test based on the total variation distance. In *Nonparametric functional estimation and related topics* (ed. G. Roussas), 631–646. NATO ASI Series, Kluwer Academic Publishers. Dordrecht.

Härdle, W. & Kneip, A. (1999). Testing a regression model when we have smooth alternatives in mind. *Scand. J. Statist.* **26**, 221–238.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

Hoeffding, W. & Wolfowitz, J. (1958). Distinguishability of sets of distributions. *Ann. Math. Statist.* **29**, 700–718.

Kolmogorov, A. N. & Tikhomirov, V. M. (1961). ϵ-entropy and ϵ-capacity of sets in function spaces. *Translat. Am. Math. Soc.* **17**, 277–364.

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1**, 38–53.

Lepski, O. V. & Tsybakov, A. B. (2000). Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probab. Theory Related Fields* **117**, 17–48.

Marron, J. S. & Wand, M. P. (1992). Exact mean integrated square error. *Ann. Statist.* **20**, 712–736.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in combinatorics 1989* (ed. J. Siemons), 148–188. Cambridge University Press, Cambridge.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Statist.* **33**, 1065–1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832–837.

Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3**, 1–14.

Received April 2004, in final form June 2005

Ricardo Cao, Department of Mathematics, Universidade da Coruña, Campus de Elviña, s/n, 15071 A Coruña, Spain.
E-mail: rcao@udc.es