Universidade de Vigo

Studying the Bandwidth in *k*-Sample Smooth Tests

Pablo Martínez-Clambor and Jacobo de Uña Álvarez

Report 09/02

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://eioweb.uvigo.es/
E-mail: depc05@uvigo.es

Universidade de Vigo

Studying the Bandwidth in *k*-Sample Smooth Tests

Pablo Martínez-Clambor and Jacobo de Uña Álvarez

Report 09/02

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita: Universidade de Vigo
Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

Studying the *Bandwidth* in k–Sample Smooth Tests

Pablo Martínez-Camblor[1] and Jacobo de Uña-Álvarez[2]

[1] *CIBER Epidemiología y SP, Subdirección de SP de Gipuzkoa, Donostia.*
E-mail: `pmcamblor@hotmail.com`
[2] *Departamento de Estadística e IO, Universidad de Vigo, Vigo*
E-mail: `jacobo@uvigo.es`

Summary In this paper, the problem of bandwidth choice in smooth k-sample tests is considered. Three different bootstrap methods are discussed and implemented. All the methods persecute the bandwidth leading to the maximum power, while preserving the level of the test. The relative performance of the methods is investigated in a simulation study. Illustration through real medical data is provided. The main conclusion is that the bootstrap minimum (BM) method provides a good compromise between statistical power and conservativeness. Robustness of the methods with respect to the number of bootstrap resamples and practical limitations are discussed.

Keywords: *k–Sample tests; Kernel estimator; Bandwidth selection; Double Bootstrap; Double Minimum; BM algorithm.*

1. Introduction

Smoothing methods have become a very popular tool when exploring data, because of their capability to estimate the population structure without any a priori type of parametric (e.g. Gaussian) assumption. On the other hand, smoothing-based statistics are a very natural way of testing the goodness-of-fit of the data to a given model specification. However, it is recognized that the choice of the smoothing degree or bandwidth may greatly influence the final shape of a smooth estimator, while having a big impact in testing for significance too.

Many methods for the selection of the smoothing degree have been provided. Focusing on the kernel density estimator (KDE), we only mention Park and Marron (1990), Devroye (1997), Bowman and Azzalini (2001) or, more recently, Ahmad and Amezziane (2007) for the access to a huge literature. The proposed methods look for a small error when approximating the underlying population curve by the smooth estimate. The goal in testing problems is different to that in nature, since one will be interested in (rather than a good estimator) the construction of a powerful test statistic. Optimal data-driven smoothing selectors in the sense of integrated deviations may not be appropriated to that end. Still, methods for bandwidth choice in testing problems have received relatively little attention in the related literature. Exceptions to this have been recently reviewed in Gao and Gijbels (2008).

In this paper we consider smooth tests for the k-sample problem. Specifically, it is assumed that k (continuous) populations are sampled independently, and that the test statistic $\mathcal{T}_h$ measures the discrepancy among the kernel density estimators pertaining to each one of the samples. Here, h stands for the bandwidth or smoothing degree when computing the kernel density estimators. Anderson et al. (1994) investigated the test statistic based on the L_2-norm for the two-sample problem, when using a fixed common bandwidth h for the two kernel density estimators. Louani (1998, 2000) studied large deviations for the L_1 and L_∞ distances between a kernel density estimator and the true density, and he proved that the Bahadur efficiency of the L_1-based (resp. L_∞-based) smooth test is greater (resp. smaller) than that of the Kolmogorov-Smirnov one-sample test. Cao and Lugosi (2005) analyzed minimum L_1-distance automatic bandwidth choice for the L_1-based one-sample smooth test. Cao and Van Keilegom (2006) introduced a new smooth test for the two-sample problem via empirical likelihood; the test statistic involves a comparison of two kernel density estimators based on the same bandwidth h. Cao and Van Keilegom (2006) showed that the choice of h influences the power of the test to a great extent, and in practice they suggested to use the bandwidth leading to an optimal (estimated) power over a given grid of smoothing levels.

Smooth k-sample tests were considered in Martínez-Camblor et al. (2008); these authors introduced the common area measure as a generalization of the L_1-norm to the k-sample case. Later, Martínez-Camblor and de Uña-Álvarez (2009) compared the common area test statistic to other type of discrepancy measures, including a different generalization of the L_1-distance and generalized L_2 and L_∞ distances too. These papers indicate that (for the k-sample problem):

(a) Smooth tests may be more powerful than tests based on the comparison of empirical distribution functions;
(b) The chosen distance among kernel density estimators has a big impact in the power of the test;
(c) The bandwidth may influence a lot the power.

However, no optimal solution to the problem of bandwidth choice in smooth test is currently available.

To be more specific, let f_1,...,f_k k probability densities which are independently sampled, and let $X_i = \{x_{i,1}, \dots, x_{i,n_i}\}$ be the random sample taken from the density f_i $(1 \leq i \leq k)$. The kernel density estimator of f_i is defined through

$$\hat{f}_{h_i}(X_i, t) = \frac{1}{n_i h_i} \sum_{j=1}^{n_i} K\left(\frac{x_{i,j} - t}{h_i}\right)$$

where K is a kernel function and h_i is a smoothing parameter or bandwidth. Then, any discrepancy measure among the $\hat{f}_{h_i}(X_i, \cdot)$'s can be the basis of a test statistic for the null hypothesis $H_0 : f_1 = ... = f_k$. Martínez-Camblor and de Uña-Álvarez (2009) found that the most powerful smooth test (among four) was that based on the generalized L_1-distance given by

$$L_{k,1}(h) = \frac{1}{N} \sum_{i=1}^{k} n_i \int |\hat{f}_{h_i}(X_i, t) - \hat{f}_{\bar{h}}(X, t)| dt$$

where $N = \sum_{i=1}^{k} n_i$, and where $\hat{f}_{\bar{h}}(X, t)$ stands for the kernel density estimator based on the pooled sample X with bandwidth $\bar{h}$. The single bandwidth h controls the amount of smoothing for each sample via $h_i = h\hat{\sigma}_i n_i^{-1/5}$ and $\bar{h} = h\hat{\bar{\sigma}} N^{-1/5}$, where $\hat{\sigma}_i$ and $\hat{\bar{\sigma}}$ are the standard deviation of the i-th and the pooled samples respectively. Hence, the asymptotically optimal rate $n^{-1/5}$ (where n is the sample size) in estimation is used for each bandwidth, but our aim is to choose the factor h to improve the power, while preserving the level of the test. This is the test we consider in this paper.

This paper is organized as follows. In Section 2 we discuss bandwidth choice principles for a general smooth test, say $\mathcal{T}_h$. Specific algorithms to compute the bandwidth in practice are introduced in Section 3. In Section 4 we investigate via simulations the performance of the *best* k-sample smooth test $L_{k,1}(h)$ in Martínez-Camblor and de Uña-Álvarez (2009) when based on different bandwidth selectors. Section 5 reports an illustration with real medical data. Finally, a discussion of our main findings is given in Section 6.

2. Bandwidth selection criterions

In this Section we discuss two different approaches to bandwidth selection for a smooth test statistic $\mathcal{T}_h$. The first one concentrates on the maximization of the power of the test, while the second considers the idea of minimizing the P-value. It will be seen that both approaches have a strong relationship. Practical implementation of these ideas will be discussed in Section 3. Without loss of generality, we will assume that the null hypothesis (whatever it is) is rejected for large values of $\mathcal{T}_h$. Besides, we will denote by $\mathcal{T}_{0,h}$ and $\mathcal{T}_{1,h}$ independent random variables with the null and the alternative distribution of $\mathcal{T}_h$ respectively.

Let $\{\mathcal{T}_h > c_{\alpha,h}\}$ be the rejection region at level α. One appealing idea is choosing the smoothing level h to maximize the power of the test (e.g. Cao and Van Keilegom, 2006). This leads to the maximum power bandwidth

$$h_{M,\alpha} = argmax_{\{h>0\}}\mathcal{P}\left\{\mathcal{T}_{1,h} > c_{\alpha,h}\right\}.$$

Note that this bandwidth depends on α.

Given the actual value of the test statistic, $\mathcal{T}_h = T_h$ say, more evidence against the null is obtained for smaller P-values of T_h. Hence, to construct a powerful test it makes sense to minimize $\pi_v(T_h) = \mathcal{P}\{\mathcal{T}_{0,h} > T_h\}$ along h. This idea is related to looking for a minimum in the so-called significance trace of the test (Hart, 1997, p. 160). The minimum P-value bandwidth is defined as

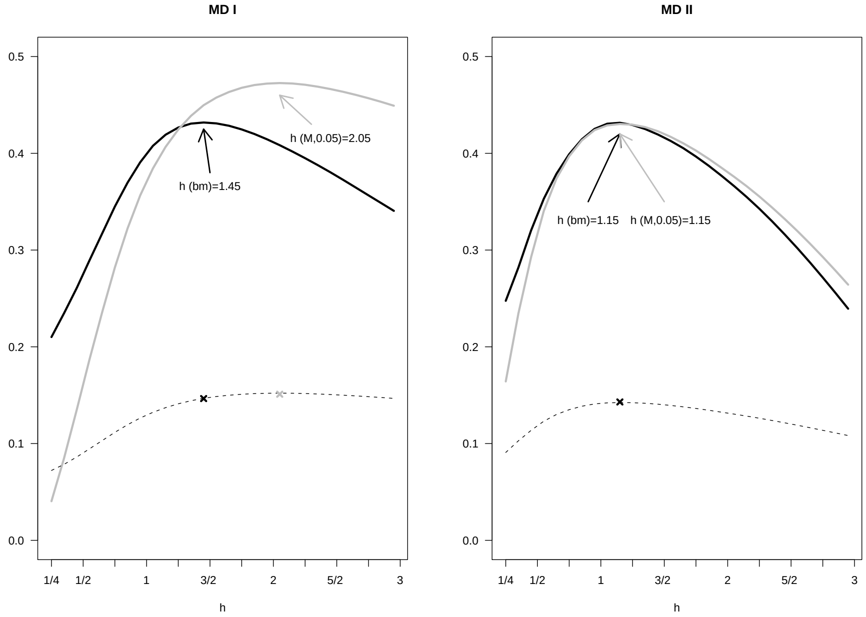
$$h_m = argmin_{\{h>0\}}\pi_v(T_h)$$

This idea was used in de Uña-Álvarez and Martínez-Camblor (2009). An interesting question is if P-values should be computed (and then minimized) for the actual values of the test $\{T_h\}_{h>0}$ *even* when the null hypothesis is true (a fact that is unknown in practice). Since our objective is the construction of a powerful test statistic, one may be tempted to minimize the P-values of the T_h that *would be obtained* under the alternative, and then averaging w.r.t the distribution of $\mathcal{T}_{1,h}$; this criterion gives the alternative bandwidth

$$h_{\bar{m}} = \mathbb{E}\left[argmin_{\{h>0\}}\pi_v(\mathcal{T}_{1,h})\right].$$

For going deeply into the comparison of these ideas, let us assume that the null and the alternative distributions of the test are Gaussian, that is, $\mathcal{T}_{0,h} \sim \mathcal{N}(\mu_{0,h}, \sigma_{0,h})$ and $\mathcal{T}_{1,h} \sim \mathcal{N}(\mu_{1,h}, \sigma_{1,h})$. In practice we will have this situation asymptotically for many smooth tests. Specifically, Martínez-Camblor and de Uña-Álvarez (2009) established asymptotic normality for the k-sample smooth test based on the L_p-distance. Under this assumption

Figure 1. Estimating $\mu_{0,h}$, $\sigma_{0,h}$, $\mu_{1,h}$ and $\sigma_{1,h}$ for $L_{k,1}(h)$ statistic from 1000 random samples for the models MD 1 (left) and MD 2 (rigth) with $a = 3/4$ and with $n = (25, 25, 25)$. Graphical representation for the functions $\mathbb{E}[\mathcal{Z}_m(h)]$ (black lines) and $\mathcal{Z}_{M,\alpha}(h) + 3/2$ for $\alpha = 0.05$ (grey lines). With $h_i = h\hat{\sigma}_i n_i^{-1/5}$ (n_i and $\hat{\sigma}_i$ are the sample size and the standard deviation for the i-th sample, respectively). The dashed line represents the statistical power at level $\alpha = 0.05$.



we have (putting Z for the standard normal):

$$h_m = argmin_{\{h>0\}}\pi_v(T_h) = argmax_{\{h>0\}}\left\{\frac{T_h - \mu_{0,h}}{\sigma_{0,h}}\right\},$$

$$\begin{aligned} h_{bm} =& argmin_{\{h>0\}}\pi_v(\mathcal{T}_{1,h}) \\ =& argmax_{\{h>0\}}\left\{\frac{\sigma_{1,h}Z + \mu_{1,h} - \mu_{0,h}}{\sigma_{0,h}}\right\} \equiv argmax_{\{h>0\}}\mathcal{Z}_m(h), \end{aligned}$$

(note that $h_{\bar{m}} = \mathbb{E}[h_{bm}]$) and (noting that $c_{\alpha,h} = \sigma_{0,h}z_\alpha + \mu_{0,h}$ where z_α is the $100(1-\alpha)\%$ percentile of Z)

$$\begin{aligned} h_{M,\alpha} =& argmax_{\{h>0\}}\mathcal{P}\left\{\mathcal{T}_{1,h} > c_{\alpha,h}\right\} \\ =& argmax_{\{h>0\}}\left\{\frac{\mu_{1,h} - \mu_{0,h} - \sigma_{0,h}z_\alpha}{\sigma_{1,h}}\right\} \equiv argmax_{\{h>0\}}\mathcal{Z}_{M,\alpha}(h). \end{aligned}$$

In Figure 1 we depict the functions $\mathcal{Z}_m(h)$ (averaged) and $\mathcal{Z}_{M,\alpha}(h)$ (in the case $\alpha = 0.05$), together with their corresponding maximizers, for two of the specific models simulated in Section 4. We see that both criterions seem to be closely related.

3. Bandwidth selection algorithms

In this Section we present the algorithms for performing the bandwidth choice in practice. They are related to the general ideas discussed in Section 2, and all of them make use of some preliminary bootstrap estimators of the involved objective functions. We will refer to null and alternative bootstrap resamples, indicating the fact that the bootstrap incorporates the null hypothesis (former case) or the alternative hypothesis (latter). In the k-sample smooth tests investigated in Section 4, null bootstrap resamples are drawn from a (pilot) smoother of the empirical distribution of the pooled sample; while the alternative bootstrap resamples are drawn independently from each of the kernel density estimators based on some pilot bandwidth. See Section 4 for the details. Maximization in the algorithms below is performed on a fixed grid of bandwidths $\mathcal{H} = \{h_1, ..., h_t\}$. Practical choice of this grid is important as it will be discussed in Section 4. For the moment we only mention that the finer the grid, the better the approximation of the bandwidth.

First, we introduce the double minimum (DM) method, which is oriented to the estimation of h_m (i.e. the minimum P-value bandwidth). This algorithm was introduced by Martínez-Camblor et al. (2008) in the scope of smooth k-sample tests, see also Martínez-Camblor and de Uña-Álvarez (2009). The method's name is a consequence of the two minimization steps performed in the plan below. The steps of the algorithm are the following.

DM_0. Let be $\mathcal{H} = \{h_1, \ldots, h_t\}$ a grid of h-values among which the optimal one is to be selected.

DM_1. Draw B_0 bootstrap resamples under the null. Let $\mathcal{T}_{0,h}^{b}$ be the statistic $\mathcal{T}_h$ when based on the b-th null bootstrap resample, $b = 1, ..., B_0$, and the bandwidth h.

DM_2. Compute

$$P_B = min_{\{h\in\mathcal{H}\}} \left\{ \frac{1}{B_0} \sum_{b=1}^{B_0} I\left\{\mathcal{T}_{0,h}^{b} > T_h\right\} \right\} \equiv min_{\{h\in\mathcal{H}\}} \pi_v^B(T_h) \equiv \pi_v^B(T_{h_{DM}})$$

where $h_{DM} = argmin_{\{h\in\mathcal{H}\}} \pi_v^B(T_h)$.

DM_3. Draw independently B_0' bootstrap resamples under the null. Let $\mathcal{T}_{0,h}^{b,*}$ be the statistic $\mathcal{T}_h$ when based on the b-th null bootstrap resample, $b = 1, ..., B_0'$, and the bandwidth h. Then, reject the null hypothesis if and only if

$$P_{B^*} \equiv \frac{1}{B_0'} \sum_{b'=1}^{B_0'} I\left\{P_{B^*}^{b'} < P_B\right\} < \alpha$$

where for $1 \leq b' \leq B_0'$,

$$P_{B^*}^{b'} = min_{\{h\in\mathcal{H}\}} \left\{ \frac{1}{B_0} \sum_{b=1}^{B_0} I\left\{\mathcal{T}_{0,h}^{b} > \mathcal{T}_{0,h}^{b',*}\right\} \right\} \equiv min_{\{h\in\mathcal{H}\}} \pi_v(\mathcal{T}_{0,h}^{b',*})$$

and α is the significance level of the test.

Note that, although calculation of h_{DM} is involved in the above procedure (Step DM_2), the goal of the algorithm is to provide the P-value of the test, for which keeping the specific value of the bandwidth is unimportant. In other words: only the minimum P-values reported in P_B (original sample), $P^b_{B^*}$ (b-th second bootstrap resample, $b = 1, ..., B'_0$), and P_{B^*} (final P-value of the test) are needed to make a decision. The test statistic is only computed $t(B_0 + B'_0)$ times along Steps DM_1-DM_3. In order to improve the computational cost of the DM algorithm, a stopping rule for the obvious case in which $\pi^B_v(T_h) > \alpha$ (resp. $\pi^B_v(T_h) < \alpha$) along the grid can be introduced; in such a case, the null hypothesis is accepted (resp. rejected) after Step DM_2.

Let us now introduce a modification of the DM algorithm which is oriented to the approximation of $h_{\bar{m}}$ rather than of h_m. The algorithm is called BM (taken from bootstrap minimum). The steps of the BM method are:

BM_0. Let be $\mathcal{H} = \{h_1, \ldots, h_t\}$ a grid of h-values among which the optimal one is to be selected.

BM_1. Draw B_0 bootstrap resamples under the null. Let $\mathcal{T}^b_{0,h}$ be the statistic $\mathcal{T}_h$ when based on the b-th null bootstrap resample, $b = 1, ..., B_0$, and the bandwidth h.

BM_2. Draw independently B_1 bootstrap resamples under the alternative. Let $\mathcal{T}^b_{1,h}$ be the statistic $\mathcal{T}_h$ when based on the b-th alternative bootstrap resample, $b = 1, ..., B_1$, and the bandwidth h.

BM_3. For each $b = 1, ..., B_1$ compute

$$h^b_{BM} = argmin_{\{h\in\mathcal{H}\}} \left\{ \frac{1}{B_0} \sum_{b'=1}^{B_0} I\left\{ \mathcal{T}^{b'}_{0,h} > \mathcal{T}^b_{1,h} \right\} \right\}$$

BM_4. Finally, compute

$$h_{BM} = \frac{1}{B_1} \sum_{b=1}^{B_1} h^b_{BM}$$

Note that the computation of the BM bandwidth as described along Steps BM_1-BM_4 implies $B_1 + B_0$ evaluations of the test statistic $\mathcal{T}_h$, multiplied by the number of bandwidths in the grid, t. Besides, in order to approximate the P-value of the data-driven test $T_{h_{BM}}$, one needs to perform B'_0 extra evaluations of the statistic under the null (total number of evaluations: $t(B_1+B_0)+B'_0$). A decision on the null and alternative hypotheses is reached after the following Step:

BM_5. Draw independently B'_0 bootstrap resamples under the null. Let $\mathcal{T}^b_{0,h_{BM}}$ the statistic $\mathcal{T}_h$ when based on the b-th null bootstrap resample, $b = 1, ..., B'_0$, and the bandwidth h_{BM}. Then, reject the null hypothesis if and only if

$$\frac{1}{B'_0} \sum_{b=1}^{B'_0} I\left\{ \mathcal{T}^b_{0,h_{BM}} > T_{h_{BM}} \right\} < \alpha$$

where $T_{h_{BM}}$ is the actual value of the test statistic (when based on h_{BM}) and α is the significance level of the test.

The third algorithm is called double bootstrap (DB), and it follows the original idea described in Cao and Van Keilegom (2006) for their empirical likelihood two-sample smooth test. Given the level of the test (α), the DB algorithm is a practical implementation of the bandwidth $h_{M,\alpha}$ which maximizes the power of the test statistic $\mathcal{T}_h$. The steps of the DB algorithm are as follows.

DB_0. Let be $\mathcal{H} = \{h_1, \ldots, h_t\}$ a grid of h-values among which the optimal one is to be selected.

DB_1. Draw B_1 bootstrap resamples under the alternative. Let $\mathcal{T}_{1,h}^b$ be the statistic $\mathcal{T}_h$ when based on the b-th alternative bootstrap resample, $b = 1, ..., B_1$, and the bandwidth h.

DB_2. Draw independently B_0 bootstrap resamples under the null. Let $\mathcal{T}_{0,h}^b$ be the statistic $\mathcal{T}_h$ when based on the b-th null bootstrap resample, $b = 1, ..., B_0$, and the bandwidth h.

DB_3. For each h compute

$$c_{\alpha,h}^B = inf \left\{ t : F_{0,h}^{B_0}(t) \geq 1 - \alpha \right\}$$

where $F_{0,h}^{B_0}$ is the empirical distribution function of the $\mathcal{T}_{0,h}^b$'s.

DB_4. Finally, compute

$$h_{DB} = argmax_{\{h \in \mathcal{H}\}} \left\{ \frac{1}{B_1} \sum_{b=1}^{B_1} I \left\{ \mathcal{T}_{1,h}^b > c_{\alpha,h}^B \right\} \right\}.$$

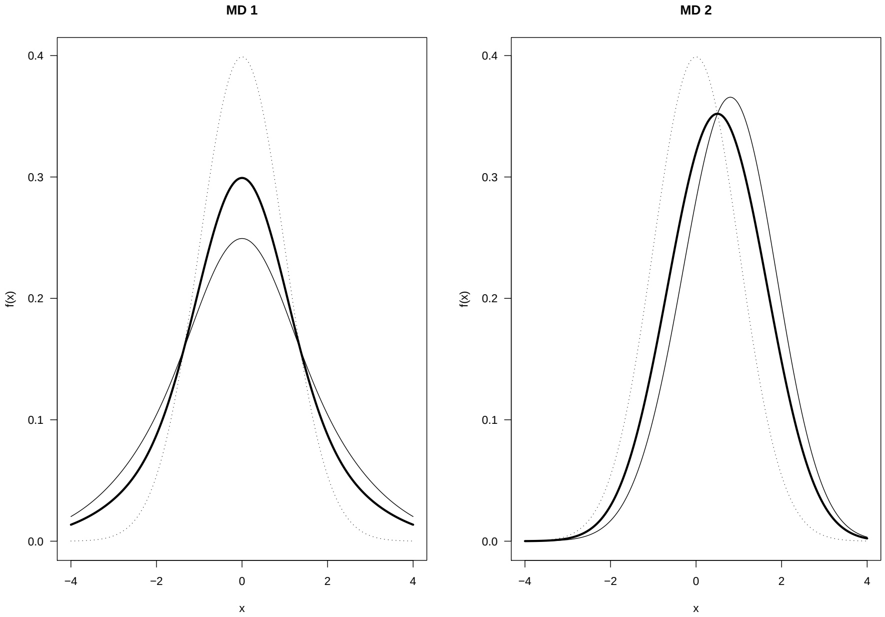
This algorithm involves $t(B_1 + B_0)$ evaluations of the test statistic. As for the DM method, the total number of evaluations is $t(B_1 + B_0) + B_0'$, when B_0' new null bootstrap resamples are drawn for the approximation (in an obvious way) of the p-value of the test based on h_{DB}. For the computation of the DB bandwidth, Cao and Van Keilegom (2006) have proposed to draw specific null bootstrap resamples in Step DB_2 for each resampled alternative in Step DB_1. This implies the calculation of a $c_{\alpha h}^B(b)$ value in Step DB_3, where b refers to the b-th alternative bootstrap resample fixed in Step DB_1. The number of evaluations of $\mathcal{T}_h$ over the grid of t bandwidths grows up to $tB_1B_0 + B_0'$ in this case, and this significantly increases the computational times in intensive Monte Carlo simulations. However, Cao and Van Keilegom (2006) did not correct the P-value of the test, thus resulting in an anticonservative method (see the simulations below, in which the DB method is implemented as described in that paper).

4. Simulation study

In order to investigate the practical behaviour of the described procedures for bandwidth choice, a small simulation study for the k-sample smooth test $L_{k,1}(h)$ was carried out. We consdered the case $k = 3$, where the first two densities are standard normal and the third density can be different. Given the computational cost involved in the calculations, the statistical powers (the significance level is $\alpha = 0.05$) were estimated from 500 Monte Carlo simulations for two different alternative models:

MD 1: $Z \equiv (1 - a)\mathcal{N}(0, 1) + a\mathcal{N}(0, 2)$.
MD 2: $Z \equiv (1 - a)\mathcal{N}(0, 1) + a\mathcal{N}(1, 1)$.

Figure 2. Simulated densities under MD 1 and MD 2 for the cases a=1/2 (thick line), a=3/4 (thin line) and a=0 (dotted line).



Two levels of contamination were used: $a = 1/2$ and $a = 3/4$. The case $a = 0$ corresponds to the null hypothesis (MD 0). In Figure 2 the two simulated alternative densities MD 1 and MD 2 are depicted. For MD 1, a change in shape (but not in location) occurs for a non-zero a. MD 2 represents the opposite situation of changing location while mantaining a similar shape under the alternative.

We considered both the balanced and unbalanced designs by taking triplets of sample sizes $(n_1, n_2, n_3) = (25, 25, 25)$ and $(n_1, n_2, n_3) = (25, 50, 75)$. Note that these situations include low ($N = 75$) to moderate ($N = 150$) total sample sizes. As mentioned in the Introduction, we used bandwidths of the form $h_i = h\hat{\sigma}_i n_i^{-1/5}$ and $\bar{h} = h\hat{\bar{\sigma}}m^{-1/5}$, where $\hat{\sigma}_i$ and $\hat{\bar{\sigma}}$ are the standard deviation of the i-th and the pooled samples respectively. Four different grids of bandwidths $\mathcal{H} = \{h_1, ..., h_t\}$ were considered for the selection of h: $\mathcal{H}_1 = \{1/4, 1/2, 1, 2, 3, 4\}$ (the largest), $\mathcal{H}_2 = \{1/2, 1, 2\}$ (shortest with small bandwidths), $\mathcal{H}_3 = \{1, 2, 3\}$ (shortest with large bandwdiths), and $\mathcal{H}_4 = \mathcal{H}_2 \cup \mathcal{H}_3 = \{1/2, 1, 2, 3\}$ (intermediate). These grids include the optimal smoothing levels for the simulated models, as explained below.

As mentioned in Section 3, we used the smoothed bootstrap (Hall et al., 1989) to obtain the bootstrap resamples under the null and under the alternative. The pilot bandwidth used in the smoothed bootstrap was always $g = Cn^{-1/3}$ (where n is the corresponding sample size), which corresponds with the bandwidth minimizing asymptotically the mean integrated squarred error of the smoothed empirical distribution function. For simplicity, we took $C = 1$. The number of null bootstrap replicates was chosen as $B_0 = 199$ for the DB algorithm, and as $B_0 = 499$ for the DM and BM algorithms. Note that the

Table 1. Observed rejection probabilities for the $L_{k,1}(h)$ statistic in the proposed models for bandwidth $h\hat{\sigma}_i n_i^{-1/5}$ where $\hat{\sigma}_i$ is the sample standard deviation, n_i $(1 \leq i \leq 3)$ is the sample size and $h \in \{1/4, 1/2, 1, 2, 3, 4\}$.

		$n = (25, 25, 25)$						$n = (25, 50, 75)$					
		h						h					
MD	a	**1/4**	**1/2**	**1**	**2**	**3**	**4**	**1/4**	**1/2**	**1**	**2**	**3**	**4**
0		0.062	0.054	0.044	0.050	0.060	0.056	0.052	0.054	0.052	0.038	0.052	0.050
1	1/2	0.146	0.174	0.268	0.292	0.250	0.222	0.328	0.394	0.574	0.632	0.594	0.560
	3/4	0.328	0.414	0.526	0.558	0.498	0.442	0.668	0.816	0.908	0.938	0.932	0.898
2	1/2	0.194	0.242	0.276	0.210	0.130	0.096	0.426	0.538	0.630	0.590	0.502	0.364
	3/4	0.380	0.476	0.556	0.442	0.258	0.130	0.786	0.894	0.946	0.946	0.898	0.744

computational savings of DM and BM methods allow for this extra effort in the bootstrap approximation of the P-values. The number of alternative bootstrap resamples was $B_1 = 100$ for DB and BM (note that DM does not make use of alternative bootstrap resamples). $B_0' = 499$ new null bootstrap resamples were used to compute the P-value correction of the test statistic when based on h_{DM} or h_{BM}, while no correction was performed for h_{DB} according to the original conception of Cao and Van Keilegom (2006).

Table 1 summarizes the observed statistical powers for different h values along the largest grid $\mathcal{H}_1$. In this case, no method for bandwidth selection was used, since first we were concerned with the influence of the bandwidth on the power of the smooth test for the simulated scenarios. The nominal levels were well respected in all the considered cases. The optimal h for MD 1 was $h = 2$ (which is included within the four considered grids) and for model MD 2 was $h = 1$, although for $a = 3/4$ and unbalanced samples the observed statistical power for $h = 2$ is the same (both values are included within the four grids).

The rejection levels of the test statistic when using the automatic bandwidth selectors h_{DM}, h_{BM} and h_{DB} are displayed in Table 2. We see that the DB method resulted in a clearly anticonservative test (specially for the largest grids). This is because the method, as originally proposed in Cao and Van Keilegom (2006), does not correct the P-value of the test for the multiplicity of bandwidths. Indeed, the DB obtained a rejected percentage larger than the optimal whithin the reference grid in 32 of the 40 considered situations. This point, which is positive in the sense of the power, makes that the nominal level was not well respected by this method. The results obtained by the BM procedure are always close to the optimum within the reference grid, while the nominal level was well respected. DM method was too conservative, showing a power uniformly below that of BM, which in summary could be considered as the winner in the comparison of the three bandwidth selectors. Means and standard deviations of the bandwidths h_{BM} and h_{DB} along the 500 trials are reported in Table 3. It is seen that the BM gave bandwidths more concentrated around its mean than the DB.

Table 2. Rejection probabilities for the $L_{k,1}(h)$ statistic in the proposed models for the *Double Bootstrap* (*DB*; $B_0 = 199 = B_0'$ and $B_1 = 100$), *Double Minimum* (*DM*; $B_0 = 499$) and *BM* ($B_0 = 499 = B_0'$ and $B_1 = 100$) algorithms on the grids $\mathcal{H}_1 = \{1/4, 1/2, 1, 2, 3, 4\}$, $\mathcal{H}_2 = \{1/2, 1, 2\}$, $\mathcal{H}_3 = \{1, 2, 3\}$ and $\mathcal{H}_4 = \{1/2, 1, 2, 3\}$.

						$n = (25, 25, 25)$							
		$\mathcal{H}_1$			$\mathcal{H}_2$			$\mathcal{H}_3$			$\mathcal{H}_4$		
MD	a	***DB***	***DM***	***BM***	***DB***	***DM***	***BM***	***DB***	***DM***	***BM***	***DB***	***DM***	***BM***
0		0.092	0.052	0.056	0.072	0.054	0.058	0.060	0.046	0.056	0.080	0.048	0.056
1	1/2	0.316	0.230	0.276	0.294	0.246	0.276	0.294	0.254	0.296	0.298	0.230	0.282
	3/4	0.582	0.460	0.532	0.576	0.488	0.524	0.562	0.514	0.572	0.576	0.470	0.562
2	1/2	0.300	0.226	0.280	0.306	0.236	0.278	0.272	0.220	0.264	0.294	0.222	0.276
	3/4	0.582	0.436	0.540	0.560	0.488	0.540	0.548	0.454	0.530	0.558	0.460	0.550
						$n = (25, 50, 75)$							
		$\mathcal{H}_1$			$\mathcal{H}_2$			$\mathcal{H}_3$			$\mathcal{H}_4$		
MD	a	***DB***	***DM***	***BM***	***DB***	***DM***	***BM***	***DB***	***DM***	***BM***	***DB***	***DM***	***BM***
0		0.080	0.054	0.052	0.058	0.042	0.052	0.054	0.040	0.054	0.070	0.046	0.054
1	1/2	0.644	0.526	0.614	0.649	0.554	0.602	0.660	0.586	0.644	0.652	0.548	0.626
	3/4	0.926	0.894	0.926	0.938	0.900	0.916	0.940	0.910	0.934	0.932	0.898	0.926
2	1/2	0.656	0.542	0.634	0.628	0.592	0.630	0.644	0.584	0.634	0.650	0.584	0.642
	3/4	0.946	0.922	0.948	0.952	0.928	0.944	0.948	0.936	0.950	0.946	0.928	0.954

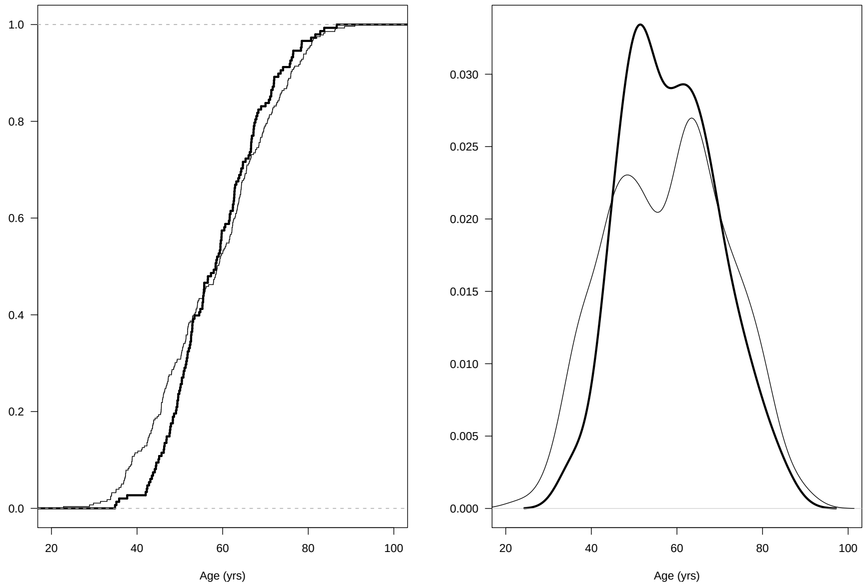
Table 3. Mean and standard deviation (between brackets) for the final used bandwidths for the *DB* and *BM* algortithms on the grids $\mathcal{H}_1 = \{1/4, 1/2, 1, 2, 3, 4\}$, $\mathcal{H}_2 = \{1/2, 1, 2\}$, $\mathcal{H}_3 = \{1, 2, 3\}$ and $\mathcal{H}_4 = \{1/2, 1, 2, 3\}$.

					$n = (25, 25, 25)$				
		$\mathcal{H}_1$		$\mathcal{H}_2$		$\mathcal{H}_3$		$\mathcal{H}_4$	
MD	a	***DB***	***BM***	***DB***	***BM***	***DB***	***BM***	***DB***	***BM***
0		1.98 (1.58)	1.97 (1.58)	1.19 (0.69)	1.19 (0.67)	2.05 (0.92)	2.07 (0.91)	1.68 (1.10)	1.70 (1.07)
1	1/2	1.58 (1.14)	1.77 (1.12)	1.40 (0.65)	1.34 (0.61)	1.79 (0.79)	1.82 (0.75)	1.62 (0.89)	1.57 (0.87)
	3/4	1.34 (0.90)	1.26 (1.10)	1.29 (0.63)	1.26 (0.57)	1.53 (0.66)	1.61 (0.63)	1.37 (0.76)	1.38 (0.73)
2	1/2	1.36 (1.15)	1.34 (1.03)	1.62 (0.77)	1.13 (0.56)	1.55 (0.78)	1.62 (0.77)	1.32 (0.87)	1.33 (0.82)
	3/4	1.34 (0.90)	1.03 (0.75)	1.62 (0.77)	1.03 (0.49)	1.30 (0.66)	1.37 (0.56)	1.07 (0.67)	1.13 (0.65)
					$n = (25, 50, 75)$				
		$\mathcal{H}_1$		$\mathcal{H}_2$		$\mathcal{H}_3$		$\mathcal{H}_4$	
MD	a	***DB***	***BM***	***DB***	***BM***	***DB***	***BM***	***DB***	***BM***
0		2.03 (1.56)	2.06 (1.55)	1.14 (0.68)	1.15 (0.68)	2.09 (0.91)	2.10 (0.89)	1.62 (1.10)	1.66 (1.05)
1	1/2	1.63 (1.08)	1.38 (0.92)	1.39 (0.65)	1.30 (0.56)	1.74 (0.74)	1.73 (0.66)	1.58 (0.88)	1.45 (0.75)
	3/4	1.03 (0.85)	0.93 (0.65)	0.99 (0.60)	1.05 (0.56)	1.31 (0.55)	1.37 (0.54)	1.06 (0.75)	1.07 (0.62)
2	1/2	1.28 (1.05)	1.26 (0.92)	1.11 (0.60)	1.16 (0.55)	1.48 (0.72)	1.51 (0.67)	1.25 (0.83)	1.26 (0.76)
	3/4	0.70 (0.63)	0.68 (0.51)	0.89 (0.55)	0.82 (0.47)	1.10 (0.31)	1.14 (0.36)	0.77 (0.48)	0.85 (0.52)

5. Real data analysis

In order to illustrate the proposed methods in a practical setup, in this Section we consider 427 cases of breast cancer registered in Gipuzkoa region, North of Spain. These data were previously analized in Martínez-Camblor et al. (2009). Two groups of women are considered, corresponding to stages I (148 women) and II (279 women) of cancer at diagnosis. The null hypothesis is that the distribution of the age at diagnosis is the same for the two groups. Empirical distribution functions and kernel density estimates for both samples are depicted in Figure 3.

Figure 3. Empirical distribution functions (left) and kernel density estimators with bandwidth 3.5 (right) for the stage I (thick line) and stage II (thin line) groups, breast cancer data.

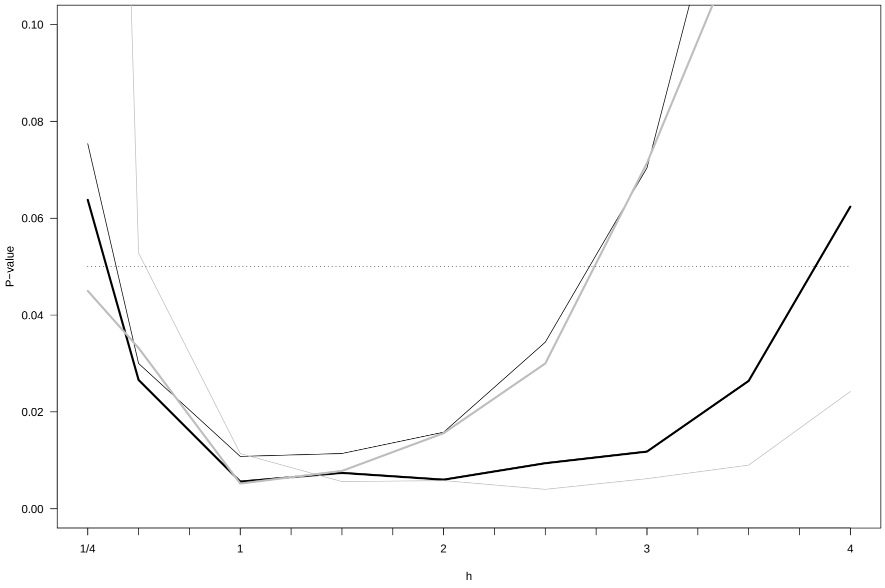


The mean and standard deviation of the age at diagnosis (in years) were 58.5 and 10.99 for stage I, and 57.9 and 13.7 for stage II. The P-values of the Levene test for equal variances and of the Welch test for equal means were 0.001 and 0.630 respectively, while the nonparametric Mann-Whitney-Wilcoxon test yielded a P-value of 0.7216. So, in principle, one can say that both populations share the same location but they are different in their shape (as it can be seen in Figure 3). We also applied the three likelihood-ratio tests based on the empirical distribution functions introduced by Zhang and Wu (2007), obtaining P-values between 0.01 and 0.05, and thus rejecting the null hypothesis of equal distributions.

Figure 4 provides the P-values of the smooth test based on $L_{k,1}(h)$ along the grid $\mathcal{H}_1 = \{1/4, 1/2, 1, 3/2, 2, 5/2, 3, 7/2, 4\}$ estimated through 5000 bootstrap resamples. We also considered the other three smooth tests investigated in Martínez-Camblor and de Uña-Álvarez (2009), namely:

$$\begin{aligned}
\mathcal{AC}(h) &= \int min\{f_{h_1}(t), \ldots, f_{h_k}(t)\}dt, \\
L_{k,2}(h) &= \frac{1}{N}\sum_{i=1}^{k} n_i \int (f_{h_i}(t) - f_{\bar{h}}(t))^2 dt, \\
\mathcal{S}_k(h) &= \frac{1}{N}\sum_{i=1}^{k} n_i sup_{\{t\in\mathbb{R}\}} |f_{h_i}(t) - f_{\bar{h}}(t)|.
\end{aligned}$$

Figure 4. P-values (estimated from 5000 bootstrap samples) for the breast cancer data, obtained by the statistics $L_{k,1}(h)$ (black thick line) $\mathcal{AC}(h)$ (grey thick line), $L_{k,2}(h)$ (black thin line) and $\mathcal{S}_k(h)$ (grey thin line) given by Martínez-Camblor and de Uña-Álvarez (2009) and the significance level (dotted line). $h_i = hn_i^{-1/5}\hat{\sigma}_i$.

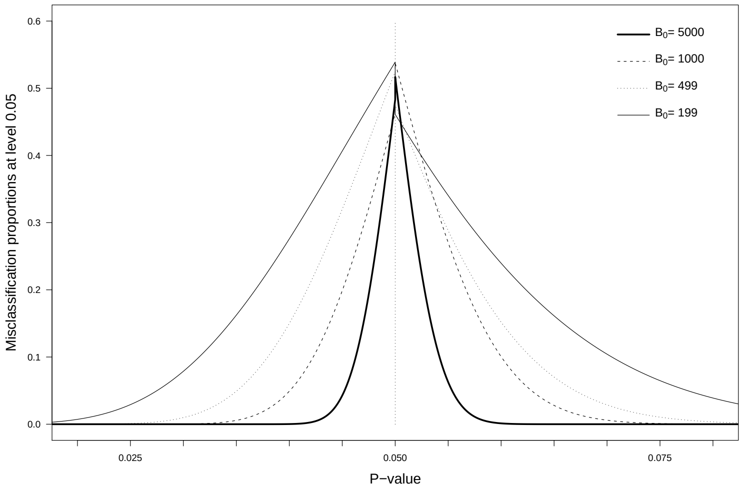


All of them reported highly significant P-values on a large range of values of h, although when oversmoothing and undersmoothing in an extreme way, P-values above 0.05 could be obtained. In order to make a decision about the suitable level of smoothness in the test statistic, we applied the DM, the BM, and the DB methods, as described in Section 3, for two different grids: $\mathcal{H}_1$ (see above) and $\mathcal{H}_2 = \{1/2, 1, 2\}$ (a reduction of 67% on the number of possible bandwidths with respect to $\mathcal{H}_1$). The numbers of bootstrap replicates were $B_0 = B_0' = 1000$ and $B_1 = 500$ when needed. The P-values obtained by the DM method for $L_{k,1}(h)$ were 0.01 ($\mathcal{H}_1$) and 0.002 ($\mathcal{H}_2$); BM reported a P-value of 0.007 for both grids (bandwidths $h = 0.891$ and $h = 0.885$ respectively), while DB gave a P-value of 0.002 and a bandwidth $h = 1$ for both grids. Hence, all bandwidth selectors allowed to reject the equality of the two groups at the standard 0.05 level. We mention that the tests $\mathcal{AC}(h)$, $L_{k,2}(h)$ and $\mathcal{S}_k(h)$ reported final P-values below 0.025 when based on the automatic bandwidths, thus reporting similar conclusions.

6. Discussion

In this paper we have addressed the problem of automatic bandwidth selection in smooth tests, where the goal is choosing the optimal bandwidth on a given grid $\mathcal{H}$ according to some suitable criterion. Different methods have been proposed in a general setup, and they have been compared through simulations and real data analysis in the special case of k-sample smooth tests. Since both the amount of data dispersion and the sample size determine the reasonable level of smoothing, we suggest to adapt the grid

Figure 5. Misclassification probability for different B_0 values against the real P-value.



taking these parameters into account. The basic optimality criterion used in this work is to reach the highest statistical power while preserving the given significance level of the test.

The issue of the influence of the bandwidth in smooth tests has received some attention in the recent literature. As reviewed in Gao and Gijbels (2008), roughly speaking there exist two different approaches to solve the problem of bandwidth choice. A first approach is to use an estimation-based optimal smoothing parameter (e.g. cross-validation) to construct the test. That approach can not be justified in theory and practice since estimation-based optimal values may not be optimal in testing problems. The second approach starts with an initial set of suitable values for the bandwidth and proceeds further from there. Our proposal follows this second philosophy.

Up to three different methods for bandwidth choice have been formalized in this paper. The double minimum (DM) method looks for the bandwidth minimizing the P-value of the test statistic along the grid. That is, the best bandwidth is the one reporting the most significant value for the smooth test. This idea connects to that of the max-statistic, which has been found useful in different scenarios (e.g. González et al., 2008). As usual, P-values and P-value correction for the multiplicity of tests is conducted through the (smoothed) bootstrap. A modification of the DM method which results in a less conservative test, the bootstrap minimum (BM) method, has been introduced. The BM bandwidth proceeds by minimizing the P-value of the test statistic when averaged along its alternative distribution. This implies bootstrapping not only the null but also the alternative hypothesis, and it resembles in some way the idea of the double bootstrap (DB) as described in Cao and Van Keilegom (2006). The DB bandwidth persecutes the maximum power; comparison with the DM and BM methods provided in this paper shows that the DB method rejects too many times the null, while being very computationally demmanding. Hence, further refinements of the method are in order.

Table 4. Observed rejection probabilities for the DB and BM bandwidths in the proposed models for different B_1 values and $n = (25, 25, 25)$. The reference grid is $\mathcal{H}_1$.

		Double Bootstrap				*BM*			
		B_1				B_1			
MD	a	**10**	**25**	**50**	**100**	**10**	**25**	**50**	**100**
0		0.074	0.090	0.086	0.092	0.056	0.064	0.062	0.056
1	1/2	0.262	0.300	0.320	0.316	0.274	0.274	0.284	0.276
	3/4	0.484	0.544	0.572	0.582	0.528	0.528	0.536	0.532
2	1/2	0.264	0.282	0.294	0.300	0.262	0.278	0.282	0.280
	3/4	0.484	0.544	0.572	0.582	0.538	0.544	0.546	0.540

An interesting issue is that of the influence of the number of (null and alternative) bootstrap resamples in the performance of the methods. Too small values of B_0 can be associated to incorrect decisions, specially when the true P-values are close to the nominal level. Missclassification errors for different values of B_0 versus the real P-value at level $\alpha = 0.05$ are shown in Figure 5. On the other hand, the BM and DB methods estimate the alternative hypothesis through the using of B_1 bootstrap resamples. We have found that both methods are quite robust with respect to B_1. Table 4 reports the statistical power of the $L_{k,1}(h)$ smooth test for for the bootstrap minimum and the double bootstrap methods and the models simulated in Section 4 (with the largest grid of bandwidths), along different values of B_1. From this Table we see that a very small number of alternative bootstrap resamples may lead to a decreasement in the percentages of rejections for the DB method, and that the BM method seems to be more robust with respect to the choice of B_1. Recall also that DB is anticonservative because no P-value correction was used (see Section 3). Table 5 shows the figures corresponding to the real medical data in Section 5, when using the $L_{k,1}(h)$ test statistic, the BM and DB methods, and the grid $\mathcal{H}_1 = \{1/4, 1/2, 1, 3/2, 2, 5/2, 3, 7/2, 4\}$. We let the value of B_1 vary between B_1 =10 and B_1 =500, with almost no influence on the reported P-value. From our whole experience, we thus recommend to dedicate more computational time to the estimation of the null hypothesis (i. e. a large B_0) than to the intensive resampling of the alternative.

In summary, we can say that the BM method showed the best performance in the simulations, reaching a good compromise between respecting the nominal level of the test and maximizing the power. BM bandwidth could be considered as a suitable modification of the DM method as originally conceived in Martínez-Camblor et al. (2008), which is too conservative. On the other hand, DB method (as implemented, see Section 3) is too anticonservative (although its final P-value could be corrected in the spirit of the other two methods), while being very computationally intensive. Besides, BM enjoys of the attractiveness of reporting a P-value independent of the specific level of the test, while DB is oriented to get the maximum power for a given α. Finally, it should be noted that, in principle, the bandwidth selectors discussed in our paper can be used in any type of

Table 5. Breast cancer data. Final BM and DB bandwidths used and respective estimated P-values from differents B_1 values.

	B_1					
	10	**25**	**50**	**100**	**250**	**500**
$\hat{h}_{BM}$	1.000	0.830	0.925	0.922	0.920	0.891
P-**value**(BM)	0.006	0.007	0.008	0.007	0.007	0.007
$\hat{h}_{DB}$	1.000	0.500	1.000	1.000	1.000	1.000
P-**value**(DB)	0.002	0.002	0.002	0.002	0.002	0.002

smooth test; to this end, the test statistic should be properly parametrized first so it depends on a single smoothing level in a suitable way.

Acknowledgements

The authors are grateful to Nerea Larrañaga (Subdirección de Salud Pública de Gipuzkoa) and her colleagues for permission to use the breast cancer data. Second author was supported by the research Grant MTM2008-03129 of the Spanish Ministerio de Ciencia e Innovación, and by the Grant PGIDIT07PXIB300191PR of the Xunta de Galicia.

REFERENCES

Ahmad AI & Amezziane M, (2007) A general and fast convergent bandwidth selection method of kernel estimator, *Journal of Nonparametric Statistics*, **19**, 165–187.

Anderson NH, Hall P & Titterington DM, (1994) Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariante Analysis*, **50**, 41–54.

Bowman A & Azzalini A, (2001) *Applied smoothing techniques for data analysis*, Oxford University Press, Oxford.

Cao R & Lugosi I, (2005) Goodness-of-fit based on the kernel density estimator, *Scandinavian Journal of Statistics. Theory and Applications*, **32**, 599–615.

Cao R & Van Keilegom I, (2006) Empirical likelihood tests for two-sample problems via nonparametric density estimation, *Canad. J. Statist.*, **34**, 61-77.

Devroye L, (1997) Universal smoothing factor selection in density estimation: Theory and practice, *Test*, **6**, 2, 223–320.

Gao J & Gijbels I, (2008) Bandwidth selection in nonparametric kernel testing. *Journal of the American Statistical Association*, **484**, 1584–1594.

González JR, Carrasco JL, Dudbridge F, Armengol Ll, Estivill X & Moreno V (2008) Maximizing association statistics over genetic models, *Genetic Epidemiology*, **32**, 246–254.

Hall P, DiCiccio JT & Romano JP, (1989) On smoothing and the bootstrap, *Annals of Statistics*, **17**, 2, 692–704.

Louani D, (1998) Large deviations limit theorems for the kernel density estimator. *Scandinavian Journal of Statistic*, **25**, 243–253.

Louani D, (2000) Large deviation for L_1-distance in kernel density estimation. *Journal of Statistical Planning & Inference*, **90**, 177–182.

Martínez-Camblor P, de Uña-Álvarez J & Corral N, (2008) k-Sample test based on the common area of kernel density estimator, *Journal of Statistical Planning & Inference*, **138**, 12, 4006-4020.

Martínez-Camblor P & de Uña-Álvarez J, (2009) Nonparametric k-sample tests: Density function vs. Distribution function, *Computational Statistics & Data Analysis*, **53**, 9, 3344–3357.

Martínez-Camblor P, Larrañaga N, Sarasqueta C & Basterretxea M, (2009) "Esa corporeidad mortal y rosa": Análisis del tiempo libre de enferemedad del cáncer de mama en Gipuzkoa en presencia de riesgos competitivos, To appear in *Gaceta Sanitaria.*

Park BU & Marron JS, (1990) Comparison of data-dirven bandwidth selectors, *Journal of American Statistics Association*, **85**, 409, 66–72.

Zhang J & Wu Y, (2007) k-Sample test based on the likelihood ratio, *Computational Statistics & Data Analysis*, **51**, 9, 4682–4691.