TEST
https://doi.org/10.1007/s11749-018-0611-5

ORIGINAL PAPER

Mode testing, critical bandwidth and excess mass

Jose Ameijeiras-Alonso[1] · **Rosa M. Crujeiras**[1] · **Alberto Rodríguez-Casal**[1]

Received: 31 July 2017 / Accepted: 18 August 2018

Abstract

The identification of peaks or maxima in probability densities, by mode testing or bump hunting, has become an important problem in applied fields. For real random variables, this task has been approached in the statistical literature from different perspectives, with the proposal of testing procedures which are based on kernel density estimators or on the quantification of excess mass. However, none of the existing proposals for testing the number of modes provides a satisfactory performance in practice. In this work, a new procedure which combines the previous approaches (smoothing and excess mass) is presented together with a revision on the previous proposals. The new method is compared with the existing ones in an extensive simulation study, showing a superior behaviour, with good calibration and power results. Theoretical justification for its performance is also obtained. A real data example on philatelic data is also included for illustration purposes, revising previous approaches and discussing the results with the new procedure.

Keywords Bootstrap calibration · Multimodality · Testing procedure · Philately

Mathematics Subject Classification 62G07 · 62G09 · 62G10

The authors gratefully acknowledge the support of Projects MTM2016-76969-P (Spanish State Research Agency, AEI) and MTM2013-41383-P (Spanish Ministry of Economy, Industry and Competitiveness), both co-funded by the European Regional Development Fund (ERDF), IAP network from Belgian Science Policy. Work of J. Ameijeiras-Alonso has been supported by the Ph.D. Grant BES-2014-071006 from the Spanish Ministry of Economy, Industry and Competitiveness.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11749-018-0611-5) contains supplementary material, which is available to authorized users.

✉ Jose Ameijeiras-Alonso
jose.ameijeiras@usc.es

Rosa M. Crujeiras
rosa.crujeiras@usc.es

Alberto Rodríguez-Casal
alberto.rodriguez.casal@usc.es

[1] Department of Statistics, Mathematical Analysis and Optimization, Faculty of Mathematics, Universidade de Santiago de Compostela, Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela, A Coruña, Spain

Published online: 31 August 2018

1 Introduction

Simple distribution models, such as the Gaussian density, may fail to capture the stochastic underlying structure driving certain mechanism in applied sciences. Complex measurements in geology, neurology, economics, ecology or astronomy exhibit some characteristics that cannot be reflected by unimodal densities. In addition, the identification of the (unknown) number of *peaks* or modes is quite common in these fields. Some examples include the study of the percentage of silica in chondrite meteors (Good and Gaskins 1980), the analysis of the macaques neurons when performing an attention-demanding task (Mitchell et al. 2007), the distribution of household incomes of the United Kingdom (Marron and Schmitz 1992), the study of the body-size in endangered fishes (Olden et al. 2007) or the analysis of the velocity at which galaxies are moving away from ours (Roeder 1990). In all these examples, identifying the number (and location) of local maxima of the density function (i.e. modes) is important per se, or as a previous step for applying other procedures.

An illustrative example which has been extensively considered in mode testing literature can be found in philately (the study of stamps and postal history and other related items). Research in this field has been motivated by the use of stamps for investment purposes. The value of stamps depends on its scarcity, and thickness is determinant in this sense. However, in some stamp issues, there is not a differentiation between groups available in stamps catalogs. The importance of establishing an objective criterion specially appears in stamp issues printed on a mixture of paper types, such as the 1872 Hidalgo issue. This particular example has been shown in several references in the literature as a paradigm of the problem of determining the number of modes/groups. In this work, this example will be revisited, recalling previous analysis and comparing results with the ones provided by the new testing procedure presented in this paper.

A formal hypothesis test for a null hypothesis of a certain number of modes can be stated as follows. Let f be the density function of a real random variable X and denote by j the number of modes. For $k \in \mathbb{Z}^+$, the testing problem on the number of modes can be formulated as:

$$H_0 : j = k \quad \text{versus} \quad H_a : j > k. \tag{1}$$

There have been quite a few proposals in the statistical literature for solving (1) and the different techniques can be classified in two groups: a first group of tests based on or using a critical bandwidth, introduced by Silverman (1981), further studied by Hall and York (2001) and also used by Fisher and Marron (2001); and a second group of tests based on the *excess mass*, such as those ones proposed by Hartigan and Hartigan (1985), Müller and Sawitzki (1991) and Cheng and Hall (1998). These methods are briefly revised and compared in this paper, where a new proposal gathering strength from both areas is also introduced, outperforming the existing procedures, in testing unimodality and more general hypotheses.

Apart from the formal testing procedures, and as a complementary tool for them, a first step when confronting the problem of identifying modes in a data distribution is the exploration of a nonparametric estimator of the underlying probability density,

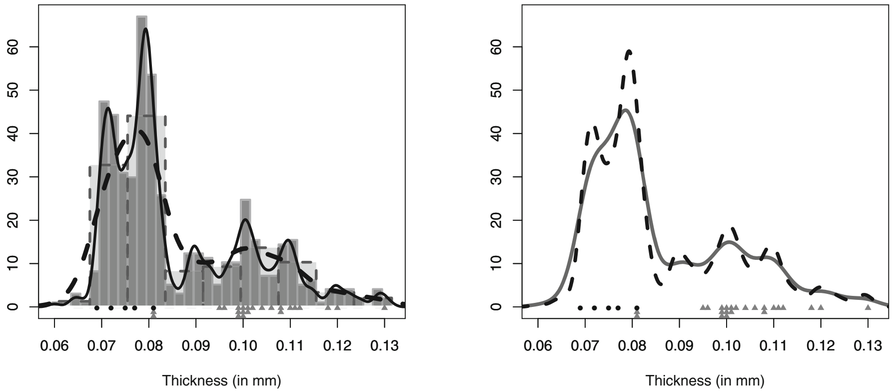


Fig. 1 Sample of 485 stamps from the 1872 Hidalgo Issue of Mexico. Points: stamps watermarked with *LA+-F* (circles) and *Papel sellado* (triangles). Kernel density estimators with Gaussian kernel and different bandwidths; left panel: $h = 0.003910$ (rule of thumb—solid line) and $h = 0.001205$ (Wand and Jones 1995,plug-in rule—dashed line, see, Ch. 3); right panel: critical bandwidths $h_4 = 0.002831$ (solid line) and $h_7 = 0.001487$ (dashed line). Left: histograms with different bin widths (0.002—continuous border and 0.008—dashed border)

which can be done by kernel methods. Classical kernel density estimation (Wand and Jones 1995, Ch. 2) allows for the reconstruction of the data density structure without imposing parametric restrictions (only subjected to mild regularity assumptions) but at the expense of choosing an appropriate bandwidth parameter, which controls the degree of smoothing. A direct observation of a kernel estimator may lead to inaccurate or even wrong conclusions about the mode density structure. This can be noticed from the plots shown in Fig. 1, where with the kernel density estimator for the stamp dataset, different conclusions can be drawn about the number of modes with different bandwidths. Based on this estimator, from an exploratory perspective, there are several alternatives for identifying modes such as the SiZer map (Chaudhuri and Marron 1999), the mode tree and the random forest (Minnotte and Scott 1993; Minnotte et al. 1998). Although these tools are helpful in supporting the results of formal testing procedures on the number and location of modes, apart from giving some insight on the global mode structure, the interpretation of the outputs from these procedures requires an *expert eye*.

This paper presents a new testing procedure combining the use of a critical bandwidth and an excess mass statistic, which can be applied to solve (1) in a general setting. In Sect. 2, a review on mode testing methods is presented, considering both tests based on critical bandwidth and on excess mass, jointly with the new proposal. A simulation study comparing all the procedures, in terms of empirical size and power, is included in Sect. 3. Section 4 is devoted to data analysis, revising the stamp dataset and presenting new results. Some final comments and discussion are given in Sect. 5. Details on the simulated models, the technical proofs, a modification of the proposal when the modes and antimodes lie in a known closed interval, a more flexible testing scenario, the computation of the proposed test and further details of the example analysed in Sect. 4 are provided as Supplementary Material available from the journal website.

2 A review on multimodality tests

Different proposals for multimodality tests will be briefly revised in this section. Section 2.1 includes a review on the methods using the critical bandwidth, and excess mass approaches are detailed in Sect. 2.2. A new proposal, borrowing strength from both alternatives, is presented in Sect. 2.3: an excess mass statistic will be calibrated from a modified nonparametric kernel density estimator using a critical bandwidth.

2.1 Tests based on the critical bandwidth

For a certain number of modes $k \in \mathbb{Z}^+$, the critical bandwidth (Silverman 1981) is the smallest bandwidth such that the kernel density estimator has at most k modes:

$$h_k = \inf\{h : \hat{f}_h \text{ has at most } k \text{ modes}\},$$

where $\hat{f}_h$ denotes the kernel density estimator, computed from a random sample $\mathcal{X} = (X_1, \ldots, X_n)$, with kernel K and bandwidth h:

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{2}$$

Silverman (1981) proposed to use the critical bandwidth with the Gaussian kernel as a statistic to test H_0: $j \leq k$ versus H_a: $j > k$, being its use justified by the fact that, with a Gaussian kernel, the number of modes of $\hat{f}_h$ is a nonincreasing function of h. Hence, H_0 is rejected for large values of h_k, whose distribution is approached using bootstrap procedures. Specifically, the proposed methodology consists in obtaining B samples $\mathcal{Z}^{*b} = (Z_1^{*b}, \ldots, Z_n^{*b})$ with $b = 1, \ldots, B$, where $Z_i^{*b} = (1 + h_k^2/\hat{\sigma}^2)^{-1/2} X_i^{*b}$, being $\hat{\sigma}^2$ the sample variance and X_i^{*b} generated from $\hat{f}_{h_k}$. By computing the critical bandwidth, h_k^{*b}, from each sample $\mathcal{Z}^{*b}$, given a significance level α, the null hypothesis is rejected if $\mathbb{P}(h_k^* \leq h_k | \mathcal{X}) \geq 1 - \alpha$. When $k = 1$, the testing problem tackled by Silverman (1981) coincides with (1). However, for a general k, the null hypothesis in (1) is more restrictive than the one considered by Silverman (1981), but asymptotic consistency of the test is only derived for $j = k$ (Hall and York 2001, for a deeper insight see , or Section SM4 in Supplementary Material).

Hall and York (2001) proved that the previous bootstrap algorithm does not provide a consistent approximation of the test statistic distribution under the null hypothesis and suggested a way for accurate calibration when $k = 1$. Given a closed interval I where the null hypothesis is tested (f has a single mode in I), if both the support of f an the interval I are unbounded then properties of h_1 (critical bandwidth when $k = 1$) are generally determined by extreme values in the sample, not by the modes of f. To avoid this issue, the testing problem (1) is reformulated as follows:

$$H_0 : j = 1 \text{ in the interior of a given closed interval } I \text{ and no local minimum in } I, \tag{3}$$

and the critical bandwidth is redefined accordingly as:

$$h_{\mathrm{HY}} = \inf\{h : \hat{f}_h \text{ has exactly one mode in } I\}. \tag{4}$$

An issue that should be kept in mind in the computation of this critical bandwidth is that even if K is the Gaussian kernel, the number of modes of $\hat{f}_h$ inside I is not necessarily a monotone function of h. But under relatively general conditions (see Hall and York 2001), the probability that the number of modes is monotone in h converges to 1 for such a kernel. Hall and York (2001) proposed using h_{HY} as a statistic to test (3). The null distribution of h_{HY} is approximated by bootstrap, generating bootstrap samples from $\hat{f}_{\mathrm{HY}}$.

Unfortunately, the critical bandwidths for the bootstrap samples h^{*b}_{HY}, are smaller than h_{HY}, so for an α-level test, a correction factor λ_α to compute the p value $\mathbb{P}(h^*_{\mathrm{HY}} \leq \lambda_\alpha h_{\mathrm{HY}} | \mathcal{X}) \geq 1 - \alpha$ must be considered. Two different methods were suggested for computing this factor λ_α, the first one based on a polynomial approximation and a second one using Monte Carlo techniques considering a simple unimodal distribution.

The previous proposal could be extended, as mentioned by Hall and York (2001), to test that f has exactly k modes in I, against the alternative that it has $(k+1)$ or more modes there, extending the critical bandwidth in (4) for k modes, namely $h_{\mathrm{HY},k}$. Nevertheless, in this scenario, the bootstrap test cannot be directly calibrated under the hypothesis that f has k modes and $(k-1)$ antimodes, since it depends on the $(2k-2)$ unknowns (c_i/c_1), where $c_i = f^{1/5}(t_i)/|f''(t_i)|^{2/5}$ (assuming $f''(t_i) \neq 0$ for all i), and t_i being the ordered turning points of f in I with $i = 1, \ldots, (2k-1)$, which notably complicates the computations.

Finally, it should be also commented that the use of the critical bandwidth for testing (1) is not limited to its use as a test statistic. Consider a Cramér–von Mises test statistic:

$$T = n\int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 dF_0(x) = \sum_{i=1}^{n} \left(F_0(X_{(i)}) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}, \tag{5}$$

where F_0 is a given continuous distribution function, $\{X_{(1)} \leq \cdots \leq X_{(n)}\}$ denotes the ordered sample and F_n is the empirical distribution function. Fisher and Marron (2001) proposed the use of (5) for solving the general problem of testing k modes (H_0: $j \leq k$) by taking $F_0(x) = \hat{F}_{h_k}(x) = \int_{-\infty}^{x} \hat{f}_{h_k}(t)\mathrm{d}t$ and derived the statistic:

$$T_k = \sum_{i=1}^{n} \left(\hat{F}_{h_k}(X_{(i)}) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}, \tag{6}$$

where the null hypothesis is rejected for large values of T_k. To approximate the distribution of the test statistic (6) under the null hypothesis, a bootstrap procedure is also proposed. It will be seen in Sect. 3 that the behaviour of the Fisher and Marron (2001) proposal is far from satisfactory.

2.2 Tests based on excess mass

Müller and Sawitzki (1991) confront the testing problem (1), employing a different perspective, under the following premise: a mode is present where an excess of probability mass is concentrated. Specifically, given a continuous real density function f and a constant λ, the excess mass is defined as:

$$E(\mathbb{P}_X, \lambda) = \mathbb{P}_X(C(\lambda)) - \lambda||C(\lambda)|| = \int_{C(\lambda)} f(x)\mathrm{d}x - \lambda||C(\lambda)||,$$

where $C(\lambda) = \{x : f(x) \geq \lambda\}$, and $||C(\lambda)||$ denotes the measure of $C(\lambda)$. If f has k modes, independently on λ, it can be divided in at most k disjoint connected sets over the support of f, called λ-clusters. If f has k λ-clusters, then the excess mass can be defined as:

$$E_k(\mathbb{P}_X, \lambda) = \sup_{C_1(\lambda),\ldots,C_k(\lambda)} \left\{ \sum_{m=1}^{k} (\mathbb{P}_X(C_m(\lambda)) - \lambda||C_m(\lambda)||) \right\}, \tag{7}$$

where the supremum is taken over all families $\{C_m(\lambda) : m = 1, \ldots, k\}$ of λ-clusters. Under the assumption that f has k λ-clusters, the excess mass defined in (7) can be empirically estimated with $E_{n,k}(\mathbb{P}_n, \lambda)$ in the following way

$$E_{n,k}(\mathbb{P}_n, \lambda) = \sup_{\hat{C}_1(\lambda),\ldots,\hat{C}_k(\lambda)} \left\{ \sum_{m=1}^{k} \mathbb{P}_n(\hat{C}_m(\lambda)) - \lambda||\hat{C}_m(\lambda)|| \right\},$$

where the empirical sets $\{\hat{C}_m(\lambda) : m = 1, \ldots, k\}$ are closed intervals with endpoints at data points, and $\mathbb{P}_n(\hat{C}_m(\lambda)) = (1/n)\sum_{i=1}^{n} \mathcal{I}(X_i \in \hat{C}_m(\lambda))$, being $\mathcal{I}$ the indicator function. The difference $D_{n,k+1}(\lambda) = E_{n,k+1}(\mathbb{P}_n, \lambda) - E_{n,k}(\mathbb{P}_n, \lambda)$ measures the plausibility of the null hypothesis, that is, large values of $D_{n,k+1}(\lambda)$ would indicate that H_0 is false. Using these differences, Müller and Sawitzki (1991) proposed the following test statistic:

$$\Delta_{n,k+1} = \max_{\lambda}\{D_{n,k+1}(\lambda)\}, \tag{8}$$

rejecting the null hypothesis that f has k modes for large values of $\Delta_{n,k+1}$. Note that just the sample is needed for computing the value of the excess mass test statistic. Müller and Sawitzki (1991) showed that this statistic is an extension of the *dip* test introduced by Hartigan and Hartigan (1985), just valid for the unimodal case, since both quantities (dip and excess mass) coincide up to a factor, for the unimodality case. In addition, the proposal of Müller and Sawitzki (1991) for testing unimodality is the same as that one of Hartigan and Hartigan (1985) and considers a Monte Carlo calibration, generating resamples from the uniform distribution.

In view of the extremely conservative behaviour of the calibration of the previous proposals (see Sect. 3 for results), Cheng and Hall (1998) designed a calibration procedure based on the following result: for large samples and under the hypothesis that f is unimodal, the distribution of $\Delta_{n,2}$ is independent of unknowns except for a factor $c = \left(f^3(x_0)/|f''(x_0)|\right)^{1/5}$, where x_0 denotes the unique mode of f. Using this fact, for the case $k = 1$, Cheng and Hall (1998) approximated the distribution of $\Delta_{n,2}$ employing the values of $\Delta^*_{n,2}$ obtained from the samples generated from a parametric calibration distribution $\Psi(\cdot, \beta)$, being β a certain parameter. Depending on the value of $d = c^{-5}$, different parametric distributions were suggested by the authors: a normal ($d = 2\pi$), a beta distribution ($d < 2\pi$) or a rescaled Student t ($d > 2\pi$). For estimating d, if $\widehat{x_0}$ denotes the largest mode of $\hat{f}_h$, then $\hat{d} = |\hat{f}''_{h'}(\widehat{x_0})|/\hat{f}^3_h(\widehat{x_0})$, is used, were $\hat{f}''$ and $\hat{f}$ are kernel estimators with a Gaussian kernel and h' and h are their respective asymptotically optimal global bandwidths, replacing the unknown quantities for the ones associated with a $N(0, \hat{\sigma}^2)$. The methodology proposed by Cheng and Hall (1998) consists in generating samples from $\Psi(\cdot, \hat{\beta})$, where $\hat{\beta}$ and the distribution family are chosen using $\hat{d}$. The excess mass statistic given in (8) when $k = 1$, that is $\Delta^*_{n,2}$, is computed from the resamples and, for a given significance level α, the null hypothesis is rejected if $\mathbb{P}(\Delta^*_{n,2} \leq \Delta_{n,2}|\mathcal{X}) \geq 1 - \alpha$.

2.3 A new proposal

The previous tests show some limitations for practical applications: first, just the proposals of Silverman (1981) and Fisher and Marron (2001) allow to test (1) for $k > 1$. Despite the efforts of Cheng and Hall (1998) and Hall and York (2001) for providing good calibration algorithms, it will be shown in Sect. 3 that the behaviour of all the proposals is far from satisfactory. Specifically, the test presented by Silverman (1981) is very conservative in general (although sometimes can show the opposite behaviour) and the proposal of Fisher and Marron (2001) does not have a good level accuracy. The new method proposed in this work overcomes these drawbacks by considering an excess mass statistic, as the one proposed by Müller and Sawitzki (1991) with bootstrap calibration. Unlike Cheng and Hall (1998), a completely data-driven procedure will be designed, using the critical bandwidth under H_0: $j = k$, $k \in \mathbb{Z}^+$.

The proposal, in a nutshell Consider the testing problem (1) and take the excess mass statistic given in (8), under the null hypothesis. Given $\mathcal{X}$, generate B resamples $\mathcal{X}^{*b}$ ($b = 1, \ldots, B$) of size n from a modified version of $\hat{f}_{h_k}$, namely the *calibration function* and subsequently denoted by g. For a significance level α, the null hypothesis will be rejected if $\mathbb{P}(\Delta^*_{n,k+1} \leq \Delta_{n,k+1}|\mathcal{X}) \geq 1 - \alpha$, where $\Delta^*_{n,k+1}$ is the excess mass statistic obtained from the generated samples. It should be also noted that the procedure can be easily adapted to handle Hall and York (2001) scenario: to test the null hypothesis that f has at most k modes in the interior of a given closed interval I, if I is known, use (a modified version of) $\hat{f}_{h_{\mathrm{HY},k}}$ to generate the samples. From this brief description, two questions arise: How is this *modified* version of $\hat{f}_{h_k}$ constructed? Does the procedure guarantee a correct calibration of the test? In fact, the construction of the calibration function as a modification of the kernel density estimator ensures the correct calibration, under some regularity conditions.

Regularity conditions (RC1) The density function f is bounded with continuous derivative. (RC2) There exist t_1 and t_2, such that f is monotone in $(-\infty, t_1)$ and in (t_2, ∞). (RC3) There are $(2j-1)$ points satisfying $\{x : f'(x) = 0 \text{ and } f(x) \neq 0\}$, which are the modes and antimodes of f, denoted as x_i, with $i = 1, \ldots, (2j-1)$; and $f''(x_i) \neq 0$. (RC4) f'' exists and is Hölder continuous in a neighbourhood of each x_i.

Define $d_i = |f''(x_i)|/f^3(x_i)$. To guarantee the asymptotic correct behaviour of the test, f must satisfy the regularity conditions (RC1)–(RC4) and the calibration function g is going to be build in order to preserve them, and to ensure the convergence, in probability, of the values $\widehat{d_i} = |g''(\widehat{x_i})|/g^3(\widehat{x_i})$ to d_i, in the modes and antimodes of g, namely $\widehat{x_i}$, as $n \to \infty$, for $i = 1, \ldots, (2j-1)$. As mentioned before, the calibration function g from which the bootstrap resamples are generated is obtained by modifying $\hat{f}_{h_j}$. Function g is constructed preserving the regularity conditions (RC1)-(RC4), by modifying $\widehat{f}_{h_j}$ in a neighbourhood of $\{x : \widehat{f}'_{h_j}(x) = 0\}$, being such values a finite collection see (Silverman 1981), having positive estimated density. This modification also ensures that the only points that satisfy $\{x : \widehat{g}'(x) = 0\}$ are the modes and antimodes of g. The estimator of d_i will be equal to the following ratio,

$$\widehat{d_i} = |\widehat{f}''_{h_{\text{PI}}}(\widehat{x_i})|/\widehat{f}^3_{h_j}(\widehat{x_i}), \text{ being } h_{\text{PI}} \text{ a plug-in bandwidth,} \tag{9}$$

where in this work, the plug-in rule for the second derivative will be obtained deriving the asymptotic mean integrated squared error and replacing f in its expression using a two-step procedure see, for example, (Wand and Jones 1995, Ch.3). Employing this calibration function g, which complete expression is given in (10), the assumptions over g of the Theorem 1 (the proofs of this result and Proposition 1 are provided as Supplementary Material) will be satisfied.

Theorem 1 *Let g be a modified version f_{h_j}, having j modes and satisfying that $|g''(\widehat{x_i})|/g^3(\widehat{x_i})$ converges in probability to $|f''(x_i)|/f^3(x_i)$, where x_i and $\widehat{x_i}$ are, respectively, the modes and antimodes of f and g, for $i = 1, \ldots, (2j-1)$. If both, f and g, satisfy conditions (RC1)–(RC4), then the limiting bootstrap distribution of $\Delta^*_{n,j+1}$ (calculated from the resamples associated to g) is identical to the asymptotic distribution of $\Delta_{n,j+1}$ (calculated from the sample associated to f), and so the test $\mathbb{P}(\Delta^*_{n,j+1} \leq \Delta_{n,j+1}|\mathcal{X}) \geq 1-\alpha$ has an asymptotic level α.*

Remark 1 Following Cheng and Hall (1998), a parametric family having the desired values of d_i, for $i = 1, \ldots, (2j-1)$, could be used as calibration function g when $j > 1$. Two issues appear related with their calibration procedure. First, it is not an easy task to construct this family. In addition, the second-order limiting properties of the test depend on the form of the density function. Then, a better behaviour is expected if the calibration function is "more similar" to the real density function. Our method deals with these two issues to get a test having a good performance in the finite-sample case and allowing to solve the general problem of testing k modes.

As mentioned before, our calibration function is constructed by modifying $\widehat{f}_{h_k}$ in a neighbourhood of the points $\{x : \widehat{f}'_{h_k}(x) = 0\}$. Depending on the nature of these points, two modifications in their neighbourhood will be done. If the point is a mode or an antimode of $\widehat{f}_{h_k}$, namely $\widehat{x_i}$, in its neighbourhood, $\widehat{f}_{h_k}$ will be replaced by the function J, described in (12). This modification will preserve the location

$\widehat{x_i}$, its estimated density value and it will satisfy that $g''(\widehat{x_i}) = \widehat{f}''_{h_{\text{PI}}}(\widehat{x_i})$.[1] In fact, this procedure guarantees the correct estimation of d_i and (RC4). The second modification, achieved by the function L, defined in (14), will remove the t saddle points of $\widehat{f_{h_k}}$, denoted as ζ_p, with $p = \{1, \ldots, t\}$. This modification is done in order to satisfy (RC3). Since all the modifications are made in bounded neighbourhoods, condition (RC2) will continue to be fulfilled and the modifications of the functions J and L will be carried out preserving condition (RC1). The calibration function g for k modes will be constructed as follows, to ensure that the assumptions of Theorem 1 are satisfied (see Proposition 1).

$$g(x; h_k, h_{\text{PI}}, \boldsymbol{\varsigma}) = \begin{cases} J(x; \widehat{x_i}, h_k, h_{\text{PI}}, \varsigma_i) & \text{if } x \in (\mathfrak{r}_i, \mathfrak{s}_i) \text{ for some } i \in \{1, \ldots, (2k-1)\}, \\ L(x; z_{(2p-1)}, z_{(2p)}, h_k) & \text{if } x \in (z_{(2p-1)}, z_{(2p)}) \text{ for some } p \in \{1, \ldots, t\}, \\ & \text{and } \zeta_p \notin (\mathfrak{r}_i, \mathfrak{s}_i) \text{ for any } i \in \{1, \ldots, (2k-1)\}, \\ \widehat{f_{h_k}}(x) & \text{otherwise.} \end{cases} \tag{10}$$

In (10), $\boldsymbol{\varsigma}$ has k elements $\varsigma_i \in (0, 1/2)$, with $i = 1, \ldots, k$, determining at which height of the kernel density estimation the modification is done. Values of ς_i close to 0 imply a modification in a "small" neighbourhood around the mode or antimode. Note that a little abuse of notation was made as g will depend on the function $\widehat{f_{h_k}}$ (not only on h_k) and on the values $\widehat{f}''_{h_{\text{PI}}}(\widehat{x_i})$, for $i = 1, \ldots, (2k-1)$. An example of the effect of g can be seen in Fig. 2. As shown in the Proposition 1, from this calibration function g, an asymptotic correct behaviour of our test can be obtained if the critical bandwidth satisfies the following condition.

Critical bandwidth condition (CBC) The critical bandwidth h_k satisfies that $a_n \leq h_k \leq b_n$, eventually with probability one, being a_n and b_n two sequences of positive numbers such as $b_n \to 0$ and $na_n/\log n \to \infty$.

Proposition 1 *Let g be defined as in (10), and where the functions J and L are defined as in (12) and (14). If h_k verifies (CBC), then g satisfies the conditions of Theorem 1.*

Remark 2 From the proof of Proposition 1, the reason for not using just a kernel density estimation with the critical bandwidth can be derived. Under some conditions (see Supplementary Material), the critical bandwidth is of order $n^{-1/5}$ and this order is not enough to guarantee that $\widehat{f}''_{h_k}(\widehat{x_i})$ will converge in probability to $f''(x_i)$.

The remaining part of this section will be devoted to further describe the construction of this calibration function g and two final remarks will be provided.

Before defining functions J and L, to ensure that g has continuous derivative, a link function l must be introduced:

[1] Note that, although asymptotically the sign of $\widehat{f}''_{h_{\text{PI}}}(\widehat{x_i})$ is always correct (under the assumptions of Theorem 1), in the finite-sample case, it may not be negative in the modes or positive in the antimodes. In that case, an abuse of notation will be done, denoting as h_{PI} to the critical or other plug-in bandwidth in order to guarantee that the sign of this second derivative remains correct.

$$l(x; u, v, a_0, a_1, b_0, b_1) = \frac{a_0 - a_1}{2}\left(1 + 2\left(\frac{x-u}{v-u}\right)^3 - 3\left(\frac{x-u}{v-u}\right)^2\right)$$
$$\times \exp\left(\frac{2(x-u)b_0}{a_0 - a_1}\right) + \frac{a_0 - a_1}{2}\left(2\left(\frac{x-u}{v-u}\right)^3 - 3\left(\frac{x-u}{v-u}\right)^2\right)$$
$$\times \exp\left(\frac{2(v-x)b_1}{a_0 - a_1}\right) + \frac{a_0 + a_1}{2}, \tag{11}$$

where $a_0 \neq a_1$ and $v > u$. Two issues must be noticed in this function. First, it allows a smooth connection between two functions, being u and v the starting and ending points where the link function is used, a_0 and a_1 the values of the connected functions on these points and b_0 and b_1 their first derivative values. Second, if the signs of b_0, b_1 and $(a_1 - a_2)$ are the same, then the first derivative of l will not be equal to 0 for any point inside $[u, v]$.

The form of J is given in Eq. (12) and its construction guarantees that $\widehat{x_i}$ is the unique point in which the derivative is equal to 0 in the neighbourhood where it is defined. The construction of J is achieved with the $\mathcal{K}$ function defined below and properly linked with the link function (11) to preserve (RC1). The $\mathcal{K}$ function is defined as follows:

$$\mathcal{K}(x; \widehat{x_i}, \mathfrak{p}_i, \mathfrak{q}_i, \eta_i) = \mathfrak{p}_i\left(1 + \delta_i\left(\frac{x - \widehat{x_i}}{\eta_i}\right)^2\right)^{\eta_i^2 \frac{\delta_i \cdot \mathfrak{q}_i}{2\mathfrak{p}_i}},$$

being δ_i a value indicating if $\widehat{x_i}$ is a mode ($\delta_i = -1$) or an antimode ($\delta_i = 1$). The value η_i will be defined later and it will depend on ς_i. The second derivative of this function exists and is Hölder continuous in $(\widehat{x_i} - \eta_i/2, \widehat{x_i} + \eta_i/2)$. The following equalities are also satisfied: $\mathcal{K}(\widehat{x_i}; \widehat{x_i}, \mathfrak{p}_i, \mathfrak{q}_i, \eta_i) = \mathfrak{p}_i$ and $\mathcal{K}''(\widehat{x_i}; \widehat{x_i}, \mathfrak{p}_i, \mathfrak{q}_i, \eta_i) = \mathfrak{q}_i$. Then, denoting as $\boldsymbol{\rho}_i = (\widehat{x_i}, \widehat{f_{h_k}}(\widehat{x_i}), \widehat{f}''_{h_{\mathrm{PI}}}(\widehat{x_i}))$, the J function can be defined as follows

$$J(x; \widehat{x_i}, h_k, h_{\mathrm{PI}}, \varsigma_i)$$
$$= \begin{cases} l\left(x; \mathfrak{r}_i, \mathfrak{v}_i, \widehat{f_{h_k}}(\mathfrak{r}_i), \mathcal{K}(\mathfrak{v}_i; \boldsymbol{\rho}_i, \eta_i), \widehat{f}'_{h_k}(\mathfrak{r}_i), \mathcal{K}'(\mathfrak{v}_i; \boldsymbol{\rho}_i, \eta_i)\right) & \text{if } x \in (\mathfrak{r}_i, \mathfrak{v}_i), \\ \mathcal{K}(x; \boldsymbol{\rho}_i, \eta_i) & \text{if } x \in [\mathfrak{v}_i, \mathfrak{w}_i], \\ l\left(x; \mathfrak{w}_i, \mathfrak{s}_i, \mathcal{K}(\mathfrak{w}_i; \boldsymbol{\rho}_i, \eta_i), \widehat{f_{h_k}}(\mathfrak{s}_i), \mathcal{K}'(\mathfrak{w}_i; \boldsymbol{\rho}_i, \eta_i), \widehat{f}'_{h_k}(\mathfrak{s}_i)\right) & \text{if } x \in (\mathfrak{w}_i, \mathfrak{s}_i), \end{cases} \tag{12}$$

being $\mathfrak{v}_i = \widehat{x_i} - \eta_i/2$ and $\mathfrak{w}_i = \widehat{x_i} + \eta_i/2$. As it was mentioned, the function J described in (12) (and hence also the calibration function g) depends on the constant $\varsigma_i \in (0, 1/2)$. Ordering the modes and denoting as $\widehat{x}_0 = -\infty$ and $\widehat{x}_{(2k)} = \infty$, that is $-\infty = \widehat{x}_0 < \widehat{x}_1 < \cdots < \widehat{x}_{2k-1} < \widehat{x}_{2k} = \infty$, the remaining unknowns values in (12) will be obtained as follows. First, it is necessary to decide at which height ϑ_i the modification in $\widehat{f_{h_k}}$ is done. For values of ς_i close to 0, ϑ_i will be close to $\widehat{f_{h_k}}(\widehat{x_i})$; while for values close to 0.5, ϑ_i will be in the middle point between $\widehat{f_{h_k}}(\widehat{x_i})$ and the highest (or lowest if $\widehat{x_i}$ is an antimode) value of $\widehat{f_{h_k}}$ in the two closest modes

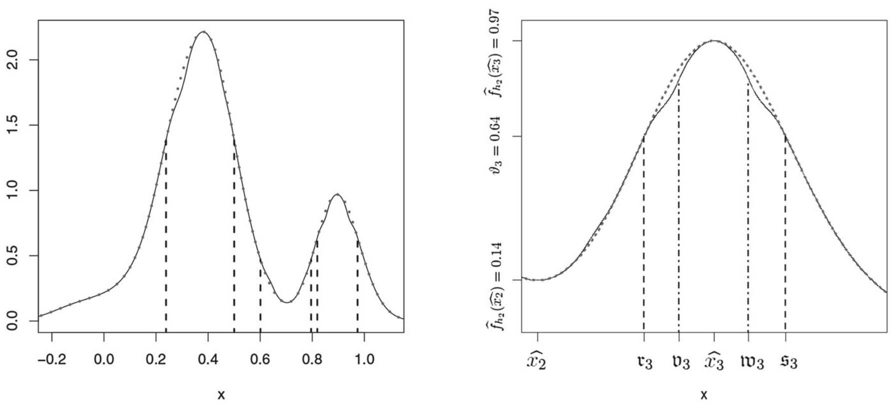


Fig. 2 Sample of $n = 1000$ observations from model M16. Dotted grey line: $\widehat{f_{h_2}}$. Solid line: $g(\cdot; h_k, h_{\mathrm{PI}}, (0.4, 0.4, 0.4))$. Dashed line: support of $J(\cdot; \widehat{x_i}, h_2, h_{\mathrm{PI}}, 0.4)$, with $i = 1, 2, 3$, Dot-dashed line: support of $\mathcal{K}(\cdot; \boldsymbol{\rho}_3, \eta_3)$. Left: in the support $(-0.2, 1.1)$. Right: in a neighbourhood of the mode $\widehat{x_3}$

or antimodes ($\widehat{x}_{i-1}$ and $\widehat{x}_{i+1}$). Second, once the height is decided, $\mathfrak{r}_i$ and $\mathfrak{s}_i$ will be the left and the right closest points to $\widehat{x_i}$ at which the density estimation is equal to ϑ_i. Third, in order to link correctly the $\mathcal{K}$ function, it is necessary to define η_i ensuring that $\mathcal{K}(\widehat{x_i} \pm \eta_i/2; \boldsymbol{\rho}_i, \eta_i)$ will be higher (lower in the antimodes) than ϑ_i. With this objective, η_i is chosen in such a way that $\mathcal{K}(\widehat{x_i} \pm \eta_i/2; \boldsymbol{\rho}_i, \eta_i)$ is near $\widehat{f_{h_k}}(\widehat{x_i})$ and as close as possible to the middle point between ϑ_i and $\widehat{f_{h_k}}(\widehat{x_i})$. Also, the value η_i will ensure that the neighbourhood $[\mathfrak{v}_i, \mathfrak{w}_i]$ in which $\mathcal{K}$ is defined is inside $(\mathfrak{r}_i, \mathfrak{s}_i)$. Finally, $\widehat{f'_{h_k}}$ must be different to 0 in the four points ($\mathfrak{r}_i$, $\mathfrak{v}_i$, $\mathfrak{w}_i$ and $\mathfrak{s}_i$) where the two link functions are employed. An example of the modifications achieved by the J function in the modes and antimodes of $\widehat{f_{h_k}}$ is shown in Fig. 2 and the complete characterization is provided below

$$
\begin{aligned}
\vartheta_i &= \widehat{f_{h_k}}(\widehat{x_i}) + \delta_i \cdot \varsigma_i \cdot \min\left(|\widehat{f_{h_k}}(\widehat{x_i}) - \widehat{f_{h_k}}(\widehat{x}_{i-1})|, |\widehat{f_{h_k}}(\widehat{x_i}) - \widehat{f_{h_k}}(\widehat{x}_{i+1})|\right), \\
\mathfrak{r}_i &= \inf\{x : x > \widehat{x}_{i-1}, \delta_i \cdot \widehat{f_{h_k}}(x) \le \delta_i \cdot \vartheta_i \text{ and } \widehat{f'_{h_k}}(x) \neq 0\}, \\
\mathfrak{s}_i &= \sup\{x : x < \widehat{x}_{i+1}, \delta_i \cdot \widehat{f_{h_k}}(x) \le \delta_i \cdot \vartheta_i \text{ and } \widehat{f'_{h_k}}(x) \neq 0\}, \\
\eta_i &= \sup\{\gamma : \gamma \in (0, \min(\widehat{x_i} - \mathfrak{r}_i, \mathfrak{s}_i - \widehat{x_i})), \delta_i \mathcal{K}(\widehat{x_i} + \gamma/2; \boldsymbol{\rho}_i, \gamma) \\
&\quad \le \delta_i(\widehat{f_{h_k}}(\widehat{x_i}) + \vartheta_i)/2 \text{ and } \widehat{f'_{h_k}}(\widehat{x_i} \pm \gamma/2) \neq 0\}.
\end{aligned}
\tag{13}
$$

In order to proceed with the modification achieved with the L function, assume that this estimator has t saddle points ζ_p, with $p = 1, \ldots, t$. Define as $\xi = \min\{|x - y| : x, y \in (\zeta_1, \ldots, \zeta_t) \cup (\mathfrak{r}_1, \mathfrak{s}_1, \ldots, \mathfrak{r}_{2k-1}, \mathfrak{s}_{2k-1})\}$. Then, if ζ_p is not inside the interval where the J functions are defined, the neighbourhood used to remove the stationary and turning points will be delimited by $z_{(2p-1)} = \zeta_p - \varpi\xi$ and $z_{(2p)} = \zeta_p + \varpi\xi$, with $\varpi \in (0, 1/4)$. In the simulation study, the value of ϖ will be taken close enough to 0 to avoid an impact in the value of the integral associated to g. Once these points are calculated, the saddle points can be removed from g with the link function by taking L equal to

$$L(x; z_{(2p-1)}, z_{(2p)}, h_k) = l\left(x; z_{(2p-1)}, z_{(2p)}, \widehat{f}_{h_k}(z_{(2p-1)}), \widehat{f}_{h_k}(z_{(2p)}), \widehat{f}'_{h_k}(z_{(2p-1)}), \widehat{f}'_{h_k}(z_{(2p)})\right). \tag{14}$$

To construct the calibration function, first, values of $\varsigma_i \in (0, 1/2)$, for $i \in \{1, \ldots, (2k-1)\}$ must be fixed. Then, using the J function (12) with the values given in (13) and the L function (14), the function g defined in (10) satisfies the specified regularity conditions and $|g''(\widehat{x_i}; h_k, h_{\text{PI}}, \boldsymbol{\varsigma})|/g^3(\widehat{x_i}; h_k, h_{\text{PI}}, \boldsymbol{\varsigma})$ converges in probability to d_i. With this modification the calibration function also preserves the structure of the data under the hypothesis that f has k modes. However, this calibration function g may not be a density function since

$$q(\boldsymbol{\varsigma}) = \int_{-\infty}^{\infty} g(x; h_k, h_{\text{PI}}, \boldsymbol{\varsigma})\text{d}x \tag{15}$$

may not be equal to 1. To ensure that g is indeed a density, a possible approach consists in proceeding with a search of values for $\boldsymbol{\varsigma}$ such that $q(\boldsymbol{\varsigma})$ is equal to 1. It can be seen that the convergence of this algorithm is guaranteed just considering "small enough" neighbourhoods (where the J function is applied), so that the calibration function is "almost equal" to the kernel density estimation which integral over the entire space is equal to one, that is,

$$\lim_{\varsigma_i \to 0^+; \forall i \in (1, \ldots, 2k-1)} q(\boldsymbol{\varsigma}) = \int_{-\infty}^{\infty} \widehat{f}_{h_k}(x)\text{d}x = 1.$$

For convenience, in the simulation study, the employed approach will be followed using ς_i close enough to 0 ($\forall i \in \{1, \ldots, 2k-1\}$) in order to avoid an impact on the integral value.

Remark 3 Under some regularity conditions, when f is twice continuously differentiable, a sufficient condition for the convergence in distribution of $n^{1/5}h_j$ is obtained when f has a bounded support or when employing SM1 critical bandwidth (see Remark SM1 in Supplementary Material). Then, "better" asymptotic results are expected when using their critical bandwidth. Although our proposal presents satisfactory results even if the support is unbounded (as it can be seen in Sect. 3), if the modes and antimodes lie in a known closed interval I, $h_{\text{HY},k}$ can be employed. An alternative approach for this case is given in Section SM3 in Supplementary Material. After obtaining a conclusion about the number of modes, when the objective is to estimate their location, it should noted that, under some regularity conditions, the modes and antimodes of $\hat{f}_{h_{\text{HY},k}}$ will provide a good estimation of their locations.

Remark 4 It should be also reminded that, unlike Silverman (1981) and Fisher and Marron (2001) proposals, the one presented in this paper considers H_0: $j = k$ instead of H_0: $j \leq k$. A deeper insight is presented in Section SM4 in Supplementary Material.

3 Simulation study

The aim of the following simulation study is to compare the different proposals presented in Sect. 2. Samples of size $n = 50$, $n = 200$ and $n = 1000$ ($n = 100$ instead of $n = 1000$ in power studies) were drawn from twenty five different distributions, ten of them unimodal (M1–M10), ten bimodal (M11–M20) and five trimodal (M21–M25) (see Section SM1 in Supplementary Material). For each choice of sampling distribution and sample size, 500 realizations of the sample were generated. Conditionally on each of those samples, for testing purposes, 500 resamples of size n were drawn from the population. Tables 1, 2, 3 and 4 report the percentages of rejections for significance levels $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.10$ under different scenarios: testing unimodality versus multimodality (Table 1); testing bimodality against more than two modes (Table 3) and power analysis (respectively Tables 2 and 4). The procedures considered include the proposals by Silverman (1981) (SI), Fisher and Marron (2001) (FM), Hall and York (2001) (HY), Hartigan and Hartigan (1985) (HH), Cheng and Hall (1998) (CH) and the new proposal (NP) in this paper. Note that for testing $H_0 : j = 2$, only SI, FM and NP can be compared. For the critical bandwidth test HY, the two proposed methods for computing λ_α have been tried, with very similar results. The ones reported in this section correspond to a polynomial approximation for λ_α. $I = [0, 1]$ is used both for HY and for NP, when the interval containing the modes is assumed to be known (Table 5). Further computational details are included in Section SM5 in Supplementary Material.

Testing unimodality versus multimodality From the results reported in Table 1, it can be concluded that SI is quite conservative: even for high sample sizes, the percentage of rejections is below the significance level, and quite close to 0 even for $\alpha = 0.10$. Regarding FM, a systematic behaviour cannot be concluded: the percentage of rejections is above the significance level for models M1, M5, M7, M9 or M10, but it can be also below the true level for M2, M4 or M8.

The behaviour of HY is quite good when using $I = [0, 1]$ for the different distributions and large sample sizes. For $n = 1000$, the percentage of rejections is quite close to α, except for model M5 (for $\alpha = 0.05$, below level) and for models M6 and M7 (for $\alpha = 0.10$, above level). However, the percentage of rejections is usually below the significance level for small sample sizes. Exceptions to this general pattern are found for model M1 ($n = 200$), M3 ($n = 50$) and M10 ($n = 200$), where percentage of rejections is close to α and models M3 ($n = 200$), M6 ($n = 200$) and M7 with percentages above α. Nevertheless, it should be kept in mind that the support where unimodality is tested must be known. Similarly to SI, the results obtained with HH are quite conservative. For instance, for $n = 1000$, even taking $\alpha = 0.10$, the percentage of rejections is always below 0.002.

Calibration seems correct in simple models for CH, although slightly conservative in some cases, such as for models M4 ($n = 1000$), M5 ($n = 200$) and M8 ($n = 200$). As expected, the parametric calibration distributions do

Table 1 Percentages of rejections for testing H_0: $j = 1$, with 500 simulations (1.96 times their estimated standard deviation in parenthesis) and $B = 500$ bootstrap samples

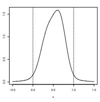
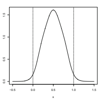
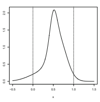
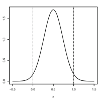
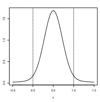
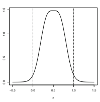
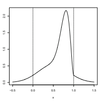
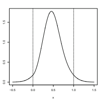
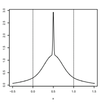
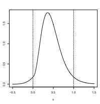
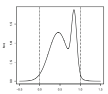
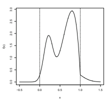
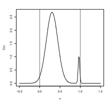
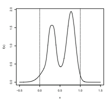
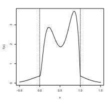
	α	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
M1		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.010(0.009)	0.076(0.023)	0.178(0.034)	0(0)	0.022(0.013)	0.050(0.019)
	$n = 200$	0(0)	0(0)	0(0)	0.056(0.020)	0.162(0.032)	0.262(0.039)	0.002(0.004)	0.046(0.018)	0.090(0.025)
	$n = 1000$	0(0)	0(0)	0(0)	0.036(0.016)	0.126(0.029)	0.210(0.036)	0.002(0.004)	0.052(0.019)	0.096(0.026)
		HH			CH			NP		
	$n = 50$	0(0)	0.006(0.007)	0.022(0.013)	0.022(0.013)	0.072(0.023)	0.140(0.030)	0.010(0.009)	0.064(0.021)	0.120(0.028)
	$n = 200$	0(0)	0.002(0.004)	0.002(0.004)	0.014(0.010)	0.058(0.020)	0.122(0.029)	0.010(0.009)	0.044(0.018)	0.120(0.028)
	$n = 1000$	0(0)	0(0)	0(0)	0.006(0.007)	0.048(0.019)	0.104(0.027)	0.008(0.008)	0.052(0.019)	0.108(0.027)
M2		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.004(0.006)	0.040(0.017)	0.076(0.023)	0(0)	0.024(0.013)	0.068(0.022)
	$n = 200$	0(0)	0(0)	0(0)	0(0)	0.006(0.007)	0.056(0.020)	0(0)	0.030(0.015)	0.082(0.024)
	$n = 1000$	0(0)	0(0)	0.004(0.006)	0(0)	0.014(0.010)	0.040(0.017)	0.004(0.006)	0.038(0.017)	0.080(0.024)
		HH			CH			NP		
	$n = 50$	0(0)	0.012(0.010)	0.028(0.014)	0.046(0.018)	0.100(0.026)	0.140(0.030)	0.016(0.011)	0.070(0.022)	0.122(0.029)
	$n = 200$	0(0)	0.002(0.004)	0.004(0.006)	0.020(0.012)	0.074(0.023)	0.164(0.032)	0.004(0.006)	0.050(0.019)	0.114(0.028)
	$n = 1000$	0(0)	0(0)	0(0)	0.008(0.008)	0.032(0.015)	0.092(0.025)	0.006(0.007)	0.030(0.015)	0.082(0.024)
M3		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.026(0.014)	0.112(0.028)	0.222(0.036)	0.008(0.008)	0.066(0.022)	0.108(0.027)
	$n = 200$	0(0)	0(0)	0(0)	0.014(0.010)	0.072(0.023)	0.146(0.031)	0.030(0.015)	0.088(0.025)	0.146(0.031)
	$n = 1000$	0(0)	0(0)	0(0)	0.002(0.004)	0.050(0.019)	0.128(0.029)	0.018(0.012)	0.070(0.022)	0.120(0.028)
		HH			CH			NP		
	$n = 50$	0(0)	0(0)	0.004(0.006)	0.002(0.004)	0.032(0.015)	0.056(0.020)	0.004(0.006)	0.042(0.018)	0.078(0.024)
	$n = 200$	0(0)	0(0)	0.002(0.004)	0.002(0.004)	0.004(0.006)	0.030(0.015)	0.002(0.004)	0.022(0.013)	0.054(0.020)
	$n = 1000$	0(0)	0(0)	0(0)	0.002(0.004)	0.012(0.010)	0.032(0.015)	0.006(0.007)	0.032(0.015)	0.082(0.024)
M4		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.002(0.004)	0.018(0.012)	0.060(0.021)	0(0)	0.020(0.012)	0.050(0.019)
	$n = 200$	0(0)	0(0)	0(0)	0(0)	0.012(0.010)	0.044(0.018)	0.004(0.006)	0.026(0.014)	0.074(0.023)
	$n = 1000$	0(0)	0.002(0.004)	0.002(0.004)	0(0)	0.010(0.009)	0.046(0.018)	0.008(0.008)	0.052(0.019)	0.090(0.025)
		HH			CH			NP		
	$n = 50$	0(0)	0.004(0.006)	0.018(0.012)	0.016(0.011)	0.064(0.021)	0.118(0.028)	0.014(0.010)	0.050(0.019)	0.102(0.027)
	$n = 200$	0(0)	0(0)	0(0)	0.008(0.008)	0.032(0.015)	0.082(0.024)	0.004(0.006)	0.030(0.015)	0.080(0.024)
	$n = 1000$	0(0)	0(0)	0(0)	0.004(0.006)	0.034(0.016)	0.066(0.022)	0(0)	0.028(0.014)	0.066(0.022)
M5		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.186(0.034)	0.366(0.042)	0.494(0.044)	0(0)	0.006(0.007)	0.038(0.017)
	$n = 200$	0(0)	0(0)	0(0)	0.268(0.039)	0.500(0.044)	0.612(0.043)	0.002(0.004)	0.030(0.015)	0.074(0.023)
	$n = 1000$	0(0)	0(0)	0(0)	0.210(0.036)	0.380(0.043)	0.504(0.044)	0.006(0.007)	0.028(0.014)	0.080(0.024)
		HH			CH			NP		
	$n = 50$	0(0)	0(0)	0.006(0.007)	0.004(0.006)	0.052(0.019)	0.084(0.024)	0.006(0.007)	0.062(0.021)	0.106(0.027)
	$n = 200$	0(0)	0(0)	0(0)	0.010(0.009)	0.034(0.016)	0.064(0.021)	0.012(0.010)	0.050(0.019)	0.092(0.025)
	$n = 1000$	0(0)	0(0)	0(0)	0.006(0.007)	0.022(0.013)	0.082(0.024)	0.006(0.007)	0.052(0.019)	0.106(0.027)
	α	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
M6		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.004(0.006)	0.040(0.017)	0.082(0.024)	0.002(0.004)	0.022(0.013)	0.074(0.023)
	$n = 200$	0(0)	0.004(0.006)	0.008(0.008)	0.010(0.009)	0.064(0.021)	0.122(0.029)	0.012(0.010)	0.110(0.027)	0.196(0.035)
	$n = 1000$	0(0)	0.008(0.008)	0.028(0.014)	0.008(0.008)	0.042(0.018)	0.100(0.026)	0.048(0.019)	0.118(0.028)	0.216(0.036)
		HH			CH			NP		
	$n = 50$	0(0)	0.006(0.007)	0.012(0.010)	0.028(0.014)	0.092(0.025)	0.168(0.033)	0.008(0.008)	0.050(0.019)	0.112(0.028)
	$n = 200$	0(0)	0.008(0.008)	0.012(0.010)	0.050(0.019)	0.136(0.030)	0.236(0.037)	0.018(0.012)	0.088(0.025)	0.160(0.032)
	$n = 1000$	0(0)	0.002(0.004)	0.002(0.004)	0.038(0.017)	0.112(0.028)	0.202(0.035)	0.016(0.011)	0.046(0.018)	0.116(0.028)
M7		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.072(0.023)	0.246(0.038)	0.378(0.043)	0.012(0.010)	0.072(0.023)	0.146(0.031)
	$n = 200$	0(0)	0(0)	0(0)	0.064(0.021)	0.210(0.036)	0.368(0.042)	0.016(0.011)	0.078(0.024)	0.144(0.031)
	$n = 1000$	0(0)	0(0)	0(0)	0.060(0.021)	0.214(0.036)	0.346(0.042)	0.004(0.006)	0.072(0.023)	0.134(0.030)
		HH			CH			NP		
	$n = 50$	0(0)	0(0)	0.010(0.009)	0.008(0.008)	0.026(0.014)	0.082(0.024)	0.006(0.007)	0.032(0.015)	0.084(0.024)
	$n = 200$	0(0)	0(0)	0(0)	0(0)	0.014(0.010)	0.042(0.018)	0.002(0.004)	0.028(0.014)	0.070(0.022)
	$n = 1000$	0(0)	0(0)	0(0)	0(0)	0.012(0.010)	0.036(0.016)	0.004(0.006)	0.042(0.018)	0.094(0.026)
M8		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.006(0.007)	0.024(0.013)	0.062(0.021)	0(0)	0.012(0.010)	0.046(0.018)
	$n = 200$	0(0)	0(0)	0(0)	0.002(0.004)	0.022(0.013)	0.064(0.021)	0(0)	0.024(0.013)	0.054(0.020)
	$n = 1000$	0(0)	0(0)	0(0)	0.002(0.004)	0.018(0.012)	0.048(0.019)	0.010(0.009)	0.054(0.020)	0.092(0.025)
		HH			CH			NP		
	$n = 50$	0(0)	0(0)	0.006(0.007)	0.006(0.007)	0.034(0.016)	0.078(0.024)	0.006(0.007)	0.032(0.015)	0.076(0.023)
	$n = 200$	0(0)	0(0)	0(0)	0.004(0.006)	0.026(0.014)	0.066(0.022)	0.006(0.007)	0.028(0.014)	0.088(0.025)
	$n = 1000$	0(0)	0(0)	0.002(0.004)	0.016(0.011)	0.038(0.017)	0.082(0.024)	0.014(0.010)	0.044(0.018)	0.088(0.025)
M9		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.018(0.012)	0.084(0.024)	0.198(0.035)	0(0)	0.006(0.007)	0.014(0.010)
	$n = 200$	0(0)	0(0)	0(0)	0.048(0.019)	0.182(0.034)	0.328(0.041)	0.002(0.004)	0.022(0.013)	0.060(0.021)
	$n = 1000$	0(0)	0(0)	0(0)	0.014(0.010)	0.160(0.032)	0.318(0.041)	0.012(0.010)	0.048(0.019)	0.086(0.025)
		HH			CH			NP		
	$n = 50$	0(0)	0(0)	0(0)	0.002(0.004)	0.016(0.011)	0.032(0.015)	0.004(0.006)	0.026(0.014)	0.068(0.022)
	$n = 200$	0(0)	0(0)	0(0)	0.002(0.004)	0.018(0.012)	0.042(0.018)	0.010(0.009)	0.046(0.018)	0.084(0.024)
	$n = 1000$	0(0)	0(0)	0(0)	0.002(0.004)	0.010(0.009)	0.014(0.010)	0.004(0.006)	0.020(0.012)	0.062(0.021)
M10		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.016(0.011)	0.054(0.020)	0.116(0.028)	0(0)	0.014(0.010)	0.050(0.019)
	$n = 200$	0(0)	0(0)	0(0)	0.018(0.012)	0.092(0.025)	0.182(0.034)	0.006(0.007)	0.038(0.017)	0.086(0.025)
	$n = 1000$	0(0)	0(0)	0(0)	0.022(0.013)	0.094(0.026)	0.168(0.033)	0.010(0.009)	0.050(0.019)	0.096(0.026)
		HH			CH			NP		
	$n = 50$	0(0)	0.002(0.004)	0.008(0.008)	0.004(0.006)	0.046(0.018)	0.086(0.025)	0.012(0.010)	0.044(0.018)	0.094(0.026)
	$n = 200$	0(0)	0(0)	0(0)	0.014(0.010)	0.042(0.018)	0.078(0.024)	0.010(0.009)	0.062(0.021)	0.094(0.026)
	$n = 1000$	0(0)	0(0)	0(0)	0.008(0.008)	0.028(0.014)	0.074(0.023)	0.008(0.008)	0.040(0.017)	0.104(0.027)

Table 2 Percentages of rejections for testing H_0: $j = 1$, with 500 simulations (1.96 times their estimated standard deviation in parenthesis) and $B = 500$ bootstrap samples

	α	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
M11		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0.002(0.004)	0.084(0.024)	0.250(0.038)	0.394(0.043)	0.008(0.008)	0.168(0.033)	0.330(0.041)
	$n = 100$	0(0)	0.002(0.004)	0.066(0.022)	0.374(0.042)	0.640(0.042)	0.738(0.039)	0.168(0.033)	0.502(0.044)	0.630(0.042)
	$n = 200$	0(0)	0.066(0.022)	0.260(0.038)	0.600(0.043)	0.788(0.036)	0.860(0.030)	0.378(0.043)	0.604(0.043)	0.714(0.040)
		HH			CH			NP		
	$n = 50$	0.012(0.010)	0.088(0.025)	0.156(0.032)	0.196(0.035)	0.370(0.042)	0.494(0.044)	0.090(0.025)	0.238(0.037)	0.376(0.042)
	$n = 100$	0.060(0.021)	0.182(0.034)	0.274(0.039)	0.442(0.044)	0.630(0.042)	0.722(0.039)	0.228(0.037)	0.418(0.043)	0.542(0.044)
	$n = 200$	0.106(0.027)	0.238(0.037)	0.356(0.042)	0.584(0.043)	0.768(0.037)	0.824(0.033)	0.328(0.041)	0.506(0.044)	0.600(0.043)
M12		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0.006(0.007)	0.332(0.041)	0.584(0.043)	0.716(0.040)	0(0)	0.174(0.033)	0.422(0.043)
	$n = 100$	0(0)	0(0)	0.042(0.018)	0.662(0.041)	0.838(0.032)	0.896(0.027)	0.224(0.037)	0.642(0.042)	0.802(0.035)
	$n = 200$	0(0)	0.026(0.014)	0.312(0.041)	0.914(0.025)	0.966(0.016)	0.986(0.010)	0.584(0.043)	0.912(0.025)	0.950(0.019)
		HH			CH			NP		
	$n = 50$	0.042(0.018)	0.118(0.028)	0.202(0.035)	0.268(0.039)	0.480(0.044)	0.594(0.043)	0.100(0.026)	0.266(0.039)	0.400(0.043)
	$n = 100$	0.158(0.032)	0.346(0.042)	0.460(0.044)	0.636(0.042)	0.788(0.036)	0.844(0.032)	0.346(0.042)	0.578(0.043)	0.680(0.041)
	$n = 200$	0.262(0.039)	0.490(0.044)	0.622(0.043)	0.818(0.034)	0.926(0.023)	0.956(0.018)	0.524(0.044)	0.752(0.038)	0.844(0.032)
M13		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0(0)	0.064(0.021)	0.492(0.044)	0.776(0.037)	0.634(0.042)	0.858(0.031)	0.892(0.027)
	$n = 100$	0(0)	0.004(0.006)	0.066(0.022)	0.914(0.025)	0.998(0.004)	1(0)	0.994(0.007)	1(0)	1(0)
	$n = 200$	0.010(0.009)	0.372(0.042)	0.784(0.036)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)
		HH			CH			NP		
	$n = 50$	0(0)	0.004(0.006)	0.014(0.010)	0.014(0.010)	0.052(0.019)	0.096(0.026)	0.016(0.011)	0.058(0.020)	0.112(0.028)
	$n = 100$	0(0)	0(0)	0(0)	0(0)	0.032(0.015)	0.062(0.021)	0.008(0.008)	0.050(0.019)	0.102(0.027)
	$n = 200$	0(0)	0(0)	0.002(0.004)	0.006(0.007)	0.048(0.019)	0.366(0.042)	0.016(0.011)	0.276(0.039)	0.758(0.038)
M14		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0.022(0.013)	0.662(0.041)	0.864(0.030)	0.922(0.024)	0.050(0.019)	0.576(0.043)	0.830(0.033)
	$n = 100$	0(0)	0.056(0.020)	0.314(0.041)	0.990(0.009)	1(0)	1(0)	0.920(0.024)	1(0)	1(0)
	$n = 200$	0.018(0.012)	0.398(0.043)	0.902(0.026)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)
		HH			CH			NP		
	$n = 50$	0.226(0.037)	0.514(0.044)	0.646(0.042)	0.718(0.039)	0.860(0.030)	0.924(0.023)	0.460(0.044)	0.716(0.040)	0.806(0.035)
	$n = 100$	0.784(0.036)	0.946(0.020)	0.978(0.013)	0.994(0.007)	0.996(0.006)	1(0)	0.950(0.019)	0.992(0.008)	0.994(0.007)
	$n = 200$	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)
M15		SI			FM			HY		
	$n = 50$	0(0)	0(0)	0.002(0.004)	0.042(0.018)	0.118(0.028)	0.228(0.037)	0.014(0.010)	0.128(0.029)	0.268(0.039)
	$n = 100$	0(0)	0(0)	0.012(0.010)	0.118(0.028)	0.334(0.041)	0.472(0.044)	0.066(0.022)	0.342(0.042)	0.500(0.044)
	$n = 200$	0(0)	0.020(0.012)	0.094(0.026)	0.266(0.039)	0.524(0.044)	0.636(0.042)	0.298(0.040)	0.576(0.043)	0.736(0.039)
		HH			CH			NP		
	$n = 50$	0.010(0.009)	0.046(0.018)	0.072(0.023)	0.098(0.026)	0.242(0.038)	0.374(0.042)	0.054(0.020)	0.156(0.032)	0.274(0.039)
	$n = 100$	0.014(0.010)	0.070(0.022)	0.150(0.031)	0.232(0.037)	0.424(0.043)	0.542(0.044)	0.124(0.029)	0.288(0.040)	0.400(0.043)
	$n = 200$	0.026(0.014)	0.104(0.027)	0.194(0.035)	0.364(0.042)	0.582(0.043)	0.690(0.041)	0.192(0.035)	0.400(0.043)	0.548(0.044)

not capture, for example, the skewness and this affects the second-order properties in more complex models. This effect is reflected in the asymmetric M3, M7 and M10 ($n = 1000$), or model M9, where the percentage of rejections is below α, and for M6 where is considerably higher than the significance level.

Finally, regarding the new proposal NP, it can be concluded that the calibration is quite satisfactory, even for complicated models, with a slightly conservative performance for M3 ($n = 200$), M4 ($n = 1000$) or M7 ($n = 200$), being this effect more clear for model M9. The only scenario where the percentage of rejections is above α is for M6 with $n = 200$, but this behaviour is corrected when increasing the sample size. Although the performance is better for higher sample sizes, in some cases, such as M9 or M16 (in the bimodal case), it can be seen that even for $n = 1000$, a percentage of rejections close to α is hard to get. In this difficult cases, the knowledge of the support can be used for obtaining better results as it was reported in Table 5, where the percentage of rejections is close to α for the sample sizes $n = 200$ and $n = 1000$.

Regarding power behaviour (and just commenting on the three methods which exhibit a correct calibration), results are reported in Table 2: none of the proposals is clearly more powerful. For instance, for M13, HY clearly detects the appearance of the second small mode, whereas the other approaches do not succeed in doing so. For

Table 3 Percentages of rejections for testing H_0: $j = 2$, with 500 simulations (1.96 times their estimated standard deviation in parenthesis) and $B = 500$ bootstrap samples

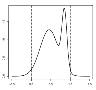
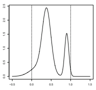
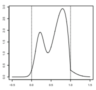
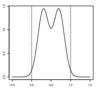
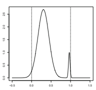
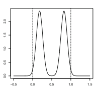
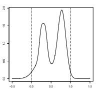
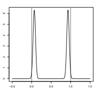
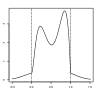
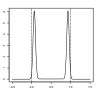
		α	0.01	0.05	0.10			α	0.01	0.05	0.10
M11	SI	$n = 50$	0(0)	0(0)	0.018(0.012)	M16	SI	$n = 50$	0(0)	0.006(0.007)	0.038(0.017)
		$n = 200$	0(0)	0.018(0.012)	0.042(0.018)			$n = 200$	0.004(0.006)	0.020(0.012)	0.040(0.017)
		$n = 1000$	0.008(0.008)	0.020(0.012)	0.044(0.018)			$n = 1000$	0(0)	0.004(0.006)	0.010(0.009)
	FM	$n = 50$	0.016(0.011)	0.076(0.023)	0.152(0.031)		FM	$n = 50$	0.092(0.025)	0.210(0.036)	0.284(0.040)
		$n = 200$	0.092(0.025)	0.276(0.039)	0.392(0.043)			$n = 200$	0.058(0.020)	0.128(0.029)	0.186(0.034)
		$n = 1000$	0.080(0.024)	0.218(0.036)	0.316(0.041)			$n = 1000$	0.038(0.017)	0.082(0.024)	0.156(0.032)
	NP	$n = 50$	0.028(0.014)	0.088(0.025)	0.178(0.034)		NP	$n = 50$	0.006(0.007)	0.064(0.021)	0.112(0.028)
		$n = 200$	0.014(0.010)	0.056(0.020)	0.108(0.027)			$n = 200$	0.016(0.011)	0.098(0.026)	0.200(0.035)
		$n = 1000$	0.006(0.007)	0.046(0.018)	0.084(0.024)			$n = 1000$	0.010(0.009)	0.074(0.023)	0.150(0.031)
M12	SI	$n = 50$	0(0)	0.004(0.006)	0.012(0.010)	M17	SI	$n = 50$	0(0)	0.002(0.004)	0.006(0.007)
		$n = 200$	0(0)	0.012(0.010)	0.018(0.012)			$n = 200$	0.002(0.004)	0.008(0.008)	0.024(0.013)
		$n = 1000$	0.002(0.004)	0.006(0.007)	0.008(0.008)			$n = 1000$	0.002(0.004)	0.002(0.004)	0.012(0.010)
	FM	$n = 50$	0.030(0.015)	0.100(0.026)	0.178(0.034)		FM	$n = 50$	0.002(0.004)	0.028(0.014)	0.060(0.021)
		$n = 200$	0.038(0.017)	0.134(0.030)	0.184(0.034)			$n = 200$	0.008(0.008)	0.046(0.018)	0.096(0.026)
		$n = 1000$	0.052(0.019)	0.094(0.026)	0.168(0.033)			$n = 1000$	0.002(0.004)	0.026(0.014)	0.048(0.019)
	NP	$n = 50$	0.004(0.006)	0.034(0.016)	0.074(0.023)		NP	$n = 50$	0.012(0.010)	0.060(0.021)	0.136(0.030)
		$n = 200$	0.002(0.004)	0.030(0.015)	0.076(0.023)			$n = 200$	0.008(0.008)	0.070(0.022)	0.106(0.027)
		$n = 1000$	0.008(0.008)	0.046(0.018)	0.082(0.024)			$n = 1000$	0.008(0.008)	0.038(0.017)	0.074(0.023)
M13	SI	$n = 50$	0(0)	0(0)	0.006(0.007)	M18	SI	$n = 50$	0(0)	0(0)	0(0)
		$n = 200$	0(0)	0.002(0.004)	0.002(0.004)			$n = 200$	0(0)	0(0)	0.020(0.012)
		$n = 1000$	0.002(0.004)	0.014(0.010)	0.034(0.016)			$n = 1000$	0(0)	0.010(0.009)	0.022(0.013)
	FM	$n = 50$	0.006(0.007)	0.044(0.018)	0.102(0.027)		FM	$n = 50$	0(0)	0.008(0.008)	0.038(0.017)
		$n = 200$	0(0)	0.024(0.013)	0.056(0.020)			$n = 200$	0.004(0.006)	0.032(0.015)	0.050(0.019)
		$n = 1000$	0.004(0.006)	0.036(0.016)	0.072(0.023)			$n = 1000$	0.004(0.006)	0.028(0.014)	0.066(0.022)
	NP	$n = 50$	0.006(0.007)	0.052(0.019)	0.118(0.028)		NP	$n = 50$	0.004(0.006)	0.054(0.020)	0.108(0.027)
		$n = 200$	0.006(0.007)	0.028(0.014)	0.070(0.022)			$n = 200$	0.006(0.007)	0.048(0.019)	0.108(0.027)
		$n = 1000$	0.010(0.009)	0.044(0.018)	0.088(0.025)			$n = 1000$	0.002(0.004)	0.034(0.016)	0.080(0.024)
M14	SI	$n = 50$	0(0)	0(0)	0.004(0.006)	M19	SI	$n = 50$	0(0)	0(0)	0(0)
		$n = 200$	0(0)	0(0)	0.002(0.004)			$n = 200$	0(0)	0(0)	0(0)
		$n = 1000$	0(0)	0(0)	0.004(0.006)			$n = 1000$	0(0)	0(0)	0(0)
	FM	$n = 50$	0.020(0.012)	0.072(0.023)	0.132(0.030)		FM	$n = 50$	0(0)	0.004(0.006)	0.034(0.016)
		$n = 200$	0.018(0.012)	0.088(0.025)	0.152(0.031)			$n = 200$	0(0)	0.008(0.008)	0.030(0.015)
		$n = 1000$	0.024(0.013)	0.058(0.020)	0.114(0.028)			$n = 1000$	0(0)	0.022(0.013)	0.044(0.018)
	NP	$n = 50$	0.008(0.008)	0.034(0.016)	0.074(0.023)		NP	$n = 50$	0.024(0.013)	0.070(0.022)	0.132(0.030)
		$n = 200$	0.004(0.006)	0.034(0.016)	0.088(0.025)			$n = 200$	0.012(0.010)	0.066(0.022)	0.118(0.028)
		$n = 1000$	0.008(0.008)	0.056(0.020)	0.092(0.025)			$n = 1000$	0.008(0.008)	0.040(0.017)	0.100(0.026)
M15	SI	$n = 50$	0(0)	0.002(0.004)	0.020(0.012)	M20	SI	$n = 50$	0.108(0.027)	0.384(0.043)	0.506(0.044)
		$n = 200$	0(0)	0.004(0.006)	0.020(0.012)			$n = 200$	0.290(0.040)	0.412(0.043)	0.564(0.043)
		$n = 1000$	0(0)	0.004(0.006)	0.032(0.015)			$n = 1000$	0.358(0.042)	0.498(0.044)	0.610(0.043)
	FM	$n = 50$	0.008(0.008)	0.072(0.023)	0.136(0.030)		FM	$n = 50$	0.002(0.004)	0.004(0.006)	0.022(0.013)
		$n = 200$	0.032(0.015)	0.128(0.029)	0.214(0.036)			$n = 200$	0.958(0.018)	0.974(0.014)	0.982(0.012)
		$n = 1000$	0.062(0.021)	0.154(0.032)	0.224(0.037)			$n = 1000$	0.976(0.013)	0.990(0.009)	0.998(0.004)
	NP	$n = 50$	0.012(0.010)	0.078(0.024)	0.158(0.032)		NP	$n = 50$	0.010(0.009)	0.068(0.022)	0.112(0.028)
		$n = 200$	0.024(0.013)	0.106(0.027)	0.200(0.035)			$n = 200$	0.016(0.011)	0.060(0.021)	0.128(0.029)
		$n = 1000$	0.014(0.010)	0.048(0.019)	0.104(0.027)			$n = 1000$	0.004(0.006)	0.038(0.017)	0.096(0.026)

Table 4 Percentages of rejections for testing H_0: $j = 2$, with 500 simulations (1.96 times their estimated standard deviation in parenthesis) and $B = 500$ bootstrap samples.

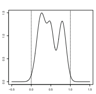
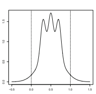
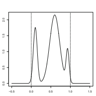
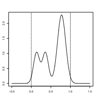
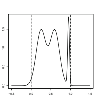
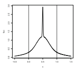
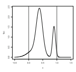
		α	0.01	0.05	0.10			α	0.01	0.05	0.10
M21	SI	$n = 50$	0(0)	0.008(0.008)	0.028(0.014)	M24	SI	$n = 50$	0(0)	0(0)	0.004(0.006)
		$n = 100$	0(0)	0.026(0.014)	0.062(0.021)			$n = 100$	0(0)	0.004(0.006)	0.008(0.008)
		$n = 200$	0.004(0.006)	0.048(0.019)	0.108(0.027)			$n = 200$	0(0)	0.004(0.006)	0.028(0.014)
	FM	$n = 50$	0.010(0.009)	0.052(0.019)	0.112(0.028)		FM	$n = 50$	0.004(0.006)	0.030(0.015)	0.082(0.024)
		$n = 100$	0.044(0.018)	0.148(0.031)	0.228(0.037)			$n = 100$	0.012(0.010)	0.052(0.019)	0.108(0.027)
		$n = 200$	0.076(0.023)	0.202(0.035)	0.278(0.039)			$n = 200$	0.050(0.019)	0.160(0.032)	0.288(0.040)
	NP	$n = 50$	0.014(0.010)	0.054(0.020)	0.112(0.028)		NP	$n = 50$	0.008(0.008)	0.060(0.021)	0.134(0.030)
		$n = 100$	0.020(0.012)	0.092(0.025)	0.168(0.033)			$n = 100$	0.034(0.016)	0.110(0.027)	0.176(0.033)
		$n = 200$	0.050(0.019)	0.134(0.030)	0.194(0.035)			$n = 200$	0.096(0.026)	0.232(0.037)	0.334(0.041)
M22	SI	$n = 50$	0(0)	0(0)	0(0)	M25	SI	$n = 50$	0(0)	0.012(0.010)	0.042(0.018)
		$n = 100$	0(0)	0.002(0.004)	0.036(0.016)			$n = 100$	0(0)	0.022(0.013)	0.096(0.026)
		$n = 200$	0.002(0.004)	0.144(0.031)	0.428(0.043)			$n = 200$	0.008(0.008)	0.048(0.019)	0.138(0.030)
	FM	$n = 50$	0.036(0.016)	0.142(0.031)	0.300(0.040)		FM	$n = 50$	0.068(0.022)	0.200(0.035)	0.312(0.041)
		$n = 100$	0.248(0.038)	0.610(0.043)	0.804(0.035)			$n = 100$	0.170(0.033)	0.344(0.042)	0.462(0.044)
		$n = 200$	0.740(0.038)	0.960(0.017)	0.982(0.012)			$n = 200$	0.190(0.034)	0.404(0.043)	0.552(0.044)
	NP	$n = 50$	0.080(0.024)	0.256(0.038)	0.402(0.043)		NP	$n = 50$	0.018(0.012)	0.098(0.026)	0.148(0.031)
		$n = 100$	0.266(0.039)	0.542(0.044)	0.706(0.040)			$n = 100$	0.098(0.026)	0.232(0.037)	0.320(0.041)
		$n = 200$	0.890(0.027)	0.956(0.018)	0.980(0.012)			$n = 200$	0.106(0.027)	0.248(0.038)	0.356(0.042)
M23	SI	$n = 50$	0(0)	0.012(0.010)	0.078(0.024)						
		$n = 100$	0(0)	0.130(0.029)	0.330(0.041)						
		$n = 200$	0.108(0.027)	0.570(0.043)	0.752(0.038)						
	FM	$n = 50$	0.054(0.020)	0.196(0.035)	0.338(0.041)						
		$n = 100$	0.334(0.041)	0.658(0.042)	0.780(0.036)						
		$n = 200$	0.832(0.033)	0.906(0.026)	0.938(0.021)						
	NP	$n = 50$	0.050(0.019)	0.204(0.035)	0.336(0.041)						
		$n = 100$	0.326(0.041)	0.624(0.042)	0.746(0.038)						
		$n = 200$	0.722(0.039)	0.878(0.029)	0.934(0.022)						

Table 5 Percentages of rejections for testing H_0: $j = 1$ (in model M9) and H_0: $j = 2$ (in model M16), with 500 simulations (1.96 times their estimated standard deviation in parenthesis) and $B = 500$ bootstrap samples

		α	0.01	0.05	0.10
			NP (known support)		
$n = 50$	M9		0 (0)	0.014 (0.010)	0.034 (0.016)
$n = 200$			0.012 (0.010)	0.052 (0.019)	0.090 (0.025)
$n = 1000$			0.010 (0.009)	0.040 (0.017)	0.092 (0.025)
$n = 50$	M16		0.002 (0.004)	0.024 (0.013)	0.076 (0.023)
$n = 200$			0.008 (0.008)	0.058 (0.020)	0.126 (0.029)
$n = 1000$			0.016 (0.011)	0.038 (0.017)	0.080 (0.024)

M11, M12, M14 and M15 ($n = 50$), CH presents the highest empirical power and HY shows the lowest one.

Assessing bimodality For testing $H_0 : j = 2$, Hall and York (2001) prove that, even knowing the density support, SI cannot be consistently calibrated by a bootstrap procedure, similar to the one used for the unimodality test. The conservative behaviour, observed in the unimodality test, is also perceived in most cases. But also, when testing bimodality, there is a model where the percentage of rejections is considerably higher than the significance level, M20, being this bimodal scenario similar to the conservative M19, just generating some outliers. FM presents again an erratic behaviour: for M17 (except for $n = 200$), M18 or M19, the percentage of rejections is below α, whereas the opposite happens for M11, M12, M15, M16 or M20 (except for $n = 50$).

For testing bimodality, NP presents good results. The percentage of rejections is close to the significance level, except for M12 ($n = 200$) and M13 ($n = 200$), slightly below α, and M11 ($n = 50$), M15 ($n = 50, n = 200$), M16 and M19 ($n = 50$), slightly above α. For $n = 1000$, all the results are good except for M16, but the calibration problem is corrected (as seen in Table 5) applying NP with known support, taking for that purpose $I = [0, 1]$. So, just the new proposal presents a correct calibration. Hence, power results reported in Table 4 are only judged for the new proposal: power increases with sample size, detecting that all the alternative distributions do not satisfy the null hypothesis, except for M21 ($n = 50$) and M24 ($n = 50$).

4 Real data analysis

Before the 1940 decade, stamps images were printed on a variety of paper types and, in general, with a lack of quality control in manufactured paper, which led to important fluctuations in paper thickness, being thin stamps more likely to be produced than thick ones. Given that the price of any stamp depends on its scarcity, the thickness of the paper is crucial for determining its value. However, there is not a standard rule for classifying stamps according to their thickness (not being available such a classification in catalogues), becoming this problem even harder in stamp issues printed on a mixture

Table 6 p value obtained using different proposals for testing k-modality, with k between 1 and 9

k	1	2	3	4	5	6	7	8	9
SI ($B = 100$)	0	0.04	0.06	0.01	0	0	0.44	0.31	0.82
FM ($B = 200$)	0	0.04	0	0	0	0	0.06	0.01	0.06
NP ($B = 500$)	0	0.022	0.004	0.506	0.574	0.566	0.376	0.886	0.808

Methods: SI, FM and NP

of paper types with possible differences in their thickness. For the 1872 Hidalgo stamp issue, it is known that the scarcity of ordinary white wove paper led the utilization of other types of paper (some of them watermarked), such us the white wove paper *Papel Sellado* or the *La Croix–Freres* (*LA+-F*). Some references exploring the number of groups in stamp thickness, and further comments, on this example are given in Section SM6 in Supplementary Material.

Taking a sample of 485 stamps, Izenman and Sommer (1988) revisited this problem previously studied by Wilson (1983) who concluded that there were only two kinds of paper (*Papel Sellado* and *La Croix–Freres*) by observing a histogram similar to the one represented in Fig. 1 (left panel, with dashed border). The same conclusions can be obtained using a kernel density estimator with a rule of thumb bandwidth (left panel, dashed curve). However, both the histogram and the kernel density estimator, depend heavily on the bin width and bandwidth, as it can be seen in Fig. 1 and different values of these tuning parameters may lead to different conclusions about the number of modes. Given that the exploratory tools did not provide a formal way of determining the number groups, Izenman and Sommer (1988) employed the multimodality test of Silverman (1981). Results from Izenman and Sommer (1988), applying SI with $B = 100$, are shown in Table 6. With a *flexible* rule (due the "conservative" nature of this test), these authors concluded that the number of groups in the 1872 Hidalgo Issue is seven. Fisher and Marron (2001) also analysed this example and the p values obtained in their studio ($B = 200$) are shown in Table 6: it is not clear which conclusion has to be made. They mentioned that their results are consistent with the previous studies, detecting 7 modes. As shown in Sect. 3, just NP has a good calibration behaviour, even with "small" sample sizes, while results of SI and FM are not accurate. Then, NP can be used to figure out how many groups are there in this stamp issue. The p values obtained with NP are also shown in Table 6, with $B = 500$. Similar results can be obtained employing the interval $I = [0.04, 0.15]$ in NP with known support, as Izenman and Sommer (1988) noticed that the thickness of the stamps is always in this interval. For a significance level $\alpha = 0.05$, it can be observed that the null hypothesis is rejected until $k = 4$, and then, there is no evidences to reject H_0 for larger values of k. Then, applying our new procedure, the conclusion is that the number of groups in the 1872 Hidalgo Issue is four.

In order to compare the results obtained by Izenman and Sommer (1988) and the ones derived applying the new proposal, two kernel density estimators, with critical bandwidths h_4 and h_7, are depicted in Fig. 1 (right panel). Izenman and Sommer (1988) argued that the stamps could be divided first in three groups (pelure paper with mode at 0.072 mm, related with the forged stamps; the medium paper in the point

0.080 mm; and the thick paper at 0.090 mm). Given the efforts made in the new issue in 1872 to avoid forged stamps, it seems quite reasonable to assume that the group associated with the pelure paper had disappeared in this new issue. In that case, the asymmetry in the first group can be attributed to the modifications in the paper made by the manufacturers. Also, this first and asymmetric group, justifies the application of nonparametric techniques to determine the number of groups. It can be seen, in the Section 7 of Izenman and Sommer (1988), that the parametric techniques (such as the mixture of Gaussian densities) have problems capturing this asymmetry, and they always determine that there are two modes in this first part of the density, one near the point 0.07 mm and another one near 0.08 mm. The other groups would correspond with stamps produced in 1872, on there it seems that the stamps of 1872 were printed on two different paper types, one with the same characteristics as the unwatermarked white wove paper used in the 1868 issue, and a second much thicker paper that disappeared completely by the end of 1872. Using this explanation, it seems quite reasonable to think that the two final modes using h_4, near the points 0.10 and 0.11 mm, correspond to the medium paper and the thick paper in this second block of stamps produced in 1872. Finally, for the two minor modes appearing near 0.12 and 0.13 mm, when h_7 is used, Izenman and Sommer (1988) do not find an explanation and they mention that probably they could be artefacts of the estimation procedure. This seems to confirm the conclusions obtained with our new procedure. The reason of determining more groups than the four obtained with our proposal seems to be quite similar to that of the model M20 in our simulation study. This possible explanation is that the spurious data in the right tail of the last mode are causing the rejection of H_0.

5 Discussion

Determining the number of modes in a distribution is a relevant practical problem in many applied sciences. The proposal presented in this paper provides a good performance for the testing problem (1), being in the case of a general number of modes k the only alternative with a reasonable behaviour. The totally nonparametric testing procedure can be extended to other contexts where a natural nonparametric estimator under the null hypothesis is available. For instance, the method can be adapted for dealing with periodic data, as it happens with the proposal by Fisher and Marron (2001).

In practical problems, where a large number of tests must be computed, obtaining a set of p values is a crucial task. In this setting, such a computation should be accompanied by the application of FDR correction techniques. The proposal in this work, based on the use of critical bandwidth and excess mass ideas, and its combination with FDR, is computationally feasible. With the aim of making this procedure accessible for the scientific community, and therefore, enabling its use in large size practical problems, an R package has been developed.

References

Chaudhuri P, Marron JS (1999) SiZer for exploration of structures in curves. J Am Stat Assoc 94:807–823
Cheng MY, Hall P (1998) Calibrating the excess mass and dip tests of modality. J R Stat Soc Ser B 60:579–589
Fisher NI, Marron JS (2001) Mode testing via the excess mass estimate. Biometrika 88:419–517
Good IJ, Gaskins RA (1980) Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. J Am Stat Assoc 75:42–56
Hall P, York M (2001) On the calibration of Silverman's test for multimodality. Stat Sin 11:515–536
Hartigan JA, Hartigan PM (1985) The dip test of unimodality. Ann Stat 13:70–84
Izenman AJ, Sommer CJ (1988) Philatelic mixtures and multimodal densities. J Am Stat Assoc 83:941–953
Marron JS, Schmitz HP (1992) Simultaneous density estimation of several income distributions. Econom Theory 8:476–488
Minnotte MC, Scott DW (1993) The mode tree: a tool for visualization of nonparametric density features. J Comput Graph Stat 2:51–68
Minnotte MC, Marchette DJ, Wegman EJ (1998) The bumpy road to the mode forest. J Comput Graph Stat 7:239–251
Mitchell JF, Sundberg KA, Reynolds JH (2007) Differential attention-dependent response modulation across cell classes in macaque visual area V4. Neuron 55:131–141
Müller DW, Sawitzki G (1991) Excess mass estimates and tests for multimodality. J Am Stat Assoc 86:738–746
Olden JD, Hogan ZS, Zanden M (2007) Small fish, big fish, red fish, blue fish: size-biased extinction risk of the world's freshwater and marine fishes. Global Ecol Biogeogr 16:694–701
Roeder K (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. J Am Stat Assoc 85:617–624
Silverman BW (1981) Using kernel density estimates to investigate multimodality. J R Stat Soc Ser B 43:97–99
Wand MP, Jones MC (1995) Kernel smoothing. Chapman and Hall, Great Britain
Wilson IG (1983) Add a new dimension to your philately. Am Philatel 97:342–349