J. R. Statist. Soc. B (1998)
60, *Part* 3, *pp.* 579–589

Calibrating the excess mass and dip tests of modality

Ming-Yen Cheng and Peter Hall†

Australian National University, Canberra, Australia

[Received October 1996. Final revision November 1997]

Summary. Nonparametric tests of modality are a distribution-free way of assessing evidence about inhomogeneity in a population, provided that the potential subpopulations are sufficiently well separated. They include the excess mass and dip tests, which are equivalent in univariate settings and are alternatives to the bandwidth test. Only very conservative forms of the excess mass and dip tests are available at present, however, and for that reason they are generally not competitive with the bandwidth test. In the present paper we develop a practical approach to calibrating the excess mass and dip tests to improve their level accuracy and power substantially. Our method exploits the fact that the limiting distribution of the excess mass statistic under the null hypothesis depends on unknowns only through a constant, which may be estimated. Our calibrated test exploits this fact and is shown to have greater power and level accuracy than the bandwidth test has. The latter tends to be quite conservative, even in an asymptotic sense. Moreover, the calibrated test avoids difficulties that the bandwidth test has with spurious modes in the tails, which often must be discounted through subjective intervention of the experimenter.

Keywords: Bandwidth test; Bimodal distribution; Bootstrap; Bump; Density estimation; Mode; Mode testing; Monte Carlo simulation; Unimodal distribution

1. Introduction

1.1. Assessing homogeneity

Testing for homogeneity of a population is traditionally conducted using parametric procedures, e.g. approximating the sampling distribution by a parametric mixture of unimodal distributions (Everitt and Hand, 1981) and assessing the goodness of fit (see for example Cox (1966), Aitkin and Rubin (1985) and Roeder (1994)). A disadvantage of this approach is that the assessment may be influenced almost as much by the validity of the particular parametric model — e.g. by the weight of the tails of the fitted distribution — as by the hypothesis of homogeneity. Testing the goodness of fit, even in the case of normal mixtures, can be particularly awkward (e.g. Hartigan (1985)) and can require sophisticated numerical methods (e.g. McLachlan (1987) and Finch *et al*. (1989)).

Provided that the components in a population have unimodal distributions and are sufficiently widely separated, they may be identified nonparametrically by 'bump hunting' (e.g. Good and Gaskins (1980)) or by testing for a multiplicity of modes in a nonparametric density estimator (e.g. the bandwidth test of Silverman (1981)). In principle these approaches are influenced relatively little by parametric issues such as the tail weight. However, methods that rely on fitting a nonparametric curve estimate can be affected by undesirable properties

†*Address for correspondence*: Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia.
E-mail: halpstat@pretty.anu.edu.au

 1369–7412/98/60579

of the procedure that is used, e.g. by the propensity of data values towards the edge of a sample to produce spurious bumps in the tail of a kernel density estimator that employs a single, global, bandwidth. Those problems reduce the attractiveness of the bandwidth test and lessen its power. Although they may sometimes be overcome by direct intervention of the experimenter, that adds a subjective element to the test and requires statistical skills that the experimenter may not have.

In this paper we suggest a way of calibrating the dip and excess mass tests, overcoming their well-known conservatism and rendering them superior to the bandwidth test in terms of level accuracy and power. The calibration retains the immunity of dip and excess mass tests to problems caused by spurious clumps of data in the tails of a distribution.

1.2. Dip and excess mass tests

Dip and excess mass approaches to testing hypotheses about modality were introduced by Hartigan and Hartigan (1985) and Müller and Sawitzki (1991a) respectively. They may be shown to be equivalent in the one-dimensional case, in that the excess mass statistic equals exactly twice the dip statistic. One of their attractive features is that they proceed without explicitly estimating a density.

Nevertheless, along with other tests for modality they share difficulties of implementation. Suggestions by Hartigan and Hartigan (1985) and Müller and Sawitzki (1991a) that the tests be based on comparisons with properties of samples of uniform random variables are attractive on the grounds of simplicity, but they lead to considerable conservatism. It may be shown that the asymptotic levels of such tests are zero, for each non-zero value of nominal level. By contrast, Silverman's (1981) bootstrap-based method for implementing the bandwidth test has a non-zero asymptotic level for each non-zero nominal level (Mammen *et al.*, 1992). Therefore it is no surprise to learn that, in moderate to large samples, Silverman's approach has greater power than the traditional dip and excess mass tests, if powers are compared at the same nominal level.

The bandwidth test is itself quite conservative, however, even in the asymptotic limit. This difficulty has been discussed by Mammen *et al.* (1992) and quantified by York (1998). It leads to reduced power. A calibrated form of the dip and excess mass tests, which we suggest in this paper, is competitive with the bandwidth test on grounds of power as well as immunity to spurious bump problems. It uses the bootstrap to emulate sampling under the null hypothesis of unimodality (i.e. homogeneity).

Since the dip and excess mass tests are equivalent, we may confine attention to the latter. Section 1.3 will detail the excess mass approach, Section 2 will describe our method, Section 3 will address its numerical properties and Section 4 will outline its main theoretical features. Technical arguments behind the results in Section 4 are given by Cheng and Hall (1997), on which this paper is based. Cheng and Hall (1997) also detail an alternative approach to calibrating the excess mass test, based on a version of Silverman's (1981) bootstrap argument. For the excess mass test, this approach produces asymptotically correct levels.

Work related to the dip and excess mass tests includes that of Hartigan (1987) on estimating a convex density contour, of Müller and Sawitzki (1987, 1991b) on the excess mass test and of Nolan (1991) on properties of the excess mass ellipsoid (a comparator set for the excess mass statistic) in the d-dimensional case. In this paper we treat only the case $d = 1$, where the selection of comparator sets avoids the ambiguities of higher dimensions. Polonik (1995a, b) has continued Müller and Sawitzki's development of the concept of excess mass, whereas Mammen and Tsybakov (1995) have used the concept as a tool for optimal estimation

of sets with smooth boundaries. Properties of the bandwidth test have been reported by Silverman (1983), Mammen *et al.* (1992) and Fisher *et al.* (1994). Work of Minnotte and Scott (1992) on modal trees should also be mentioned in this context.

1.3. Details of the excess mass test

Let F denote the distribution function corresponding to the sampling density f, and let $\hat{F}$ be the empirical distribution function of an n-sample drawn from F. The measure of empirical excess mass for m modes is defined to equal

$$E_{nm}(\lambda) = \sup_{C_1, \ldots, C_m} \left[\sum_{j=1}^{m} \{ \hat{F}(C_j) - \lambda \| C_j \| \} \right] \tag{1.1}$$

(Müller and Sawitzki, 1991a, b), where $\lambda > 0$, the supremum is taken over all sequences $\{C_1, \ldots, C_m\}$ of disjoint intervals (the comparator sets), $\hat{F}(C)$ is the $\hat{F}$-measure of C and $\|C\|$ is the length of C. Define

$$D_{nm}(\lambda) = E_{nm}(\lambda) - E_{n,m-1}(\lambda) \geqslant 0.$$

The excess mass statistic for testing the null hypothesis H_{m-1} that f has $m-1$ modes, against the alternative H_m that it has m modes, is

$$\Delta_{nm} = \sup_{\lambda > 0} \{ D_{nm}(\lambda) \}.$$

The hypothesis H_{m-1} is rejected in favour of H_m if the value of Δ_{nm} is too large. In the present paper we develop empirical methods for quantifying the notion of 'too large', with particular emphasis on the case $m = 1$.

2. Calibration method

The method presented here is based on a theoretical property that will be stated formally in Section 4: for large samples, and under the null hypothesis that f is unimodal, the distribution of Δ_{n2} is independent of unknowns except for a factor,

$$c = \{ f(x_0)^3 / |f''(x_0)| \}^{1/5}, \tag{2.1}$$

where x_0 denotes the unique mode of f. The method involves estimating c. It is immune to the spurious bump problems of the bandwidth test and may be adapted to the general problem of testing for m modes. It requires resampling from a 'calibration distribution'—a known unimodal distribution F^0, for which the properties of Δ_{n2} are similar to those under F if the null hypothesis is correct.

The value of $d = c^{-5} = |f''(x_0)| / f(x_0)^3$ may lie anywhere in the interval $[0, \infty)$, and we cover that range by three classes of distribution:

(a) the beta distribution with density

$$g_\beta(x) = \frac{1}{B(\beta, \beta)} \{ x(1-x) \}^{\beta - 1} \tag{2.2a}$$

for $0 < x < 1$ and $\beta > 1$, where $d = \gamma(\beta)$ ranges over $[0, 2\pi)$ (with $\gamma(1) = 0$ and $\gamma(\infty) = 2\pi$);

(b) any normal distribution, where $d = 2\pi$;

(c) the rescaled Student t-distribution with density

$$g_\beta(x) = \frac{1}{B(\beta - \frac{1}{2}, \frac{1}{2})} \frac{1}{(1 + x^2)^\beta} \tag{2.2b}$$

for $-\infty < x < \infty$ and $\beta > \frac{1}{2}$, where $d = \gamma(\beta)$ ranges over $(2\pi, \infty)$ (with $\gamma(\frac{1}{2}) = \infty$ and $\gamma(\infty) = 2\pi$).

Values of d for the beta and Student t reference distributions are respectively

$$\gamma(\beta) = \begin{cases} |g_\beta''(\frac{1}{2})| / g_\beta(\frac{1}{2})^3 \\ |g_\beta''(0)| / g_\beta(0)^3 \end{cases} = \begin{cases} 2^{4\beta-1}(\beta - 1)\, B(\beta, \beta)^2 \\ 2\beta\, B(\beta - \frac{1}{2}, \frac{1}{2})^2. \end{cases} \tag{2.3}$$

These reference distributions were chosen because they are straightforward to simulate from. However, any class of reference distributions which enjoyed a one-to-one relationship with the set $(0, \infty)$ of potential values of d would be suitable in principle. The symmetry and ranges of the reference distributions are not important to first-order characteristics of our method, such as first-order limiting properties, and so do not need to reflect those aspects of the data. Nevertheless, if the data were markedly skew then that characteristic would be reflected in the second-order properties of the test (which are not dealt with here) and might best be captured by using a similarly skewed reference distribution. Since c and d are invariant under changes to the scale of the sampling distribution, then changes to the scale of the densities at equation (2.2) do not affect equations (2.3).

Consider estimating f and f'', for example by using kernel methods. (In this setting, $\widehat{f''}$ would typically not be taken as the second derivative of $\hat{f}$, since different smoothing parameters would be employed to calculate $\hat{f}$ and $\widehat{f''}$. See Section 3.) Let $\hat{x}_0 = \arg\max(\hat{f})$ denote the 'largest mode' of $\hat{f}$, put

$$\hat{d} = |\widehat{f''}(\hat{x}_0)| / \hat{f}(\hat{x}_0)^3$$

and let $\hat{\beta} = \gamma^{-1}(\hat{d})$. Conditionally on $\mathcal{X} = \{X_1, \ldots, X_n\}$, draw a sample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ from the distribution with density $g_{\hat{\beta}}$, and compute the version Δ_{n2}^* of Δ_{n2} for the data $\mathcal{X}^*$. To construct a test at level α, use Monte Carlo methods to compute the critical point $\hat{z}_\alpha$ defined by

$$P_{g_{\hat{\beta}}}(\Delta_{n2}^* > \hat{z}_\alpha | \mathcal{X}) = \alpha,$$

and reject the null hypothesis that f is unimodal if $\Delta_{n2} > \hat{z}_\alpha$. Under mild regularity conditions, including consistency of $\hat{d}$ for the value of d under the null hypothesis, the test has asymptotically correct level. Details will be given in Section 4.

Our method may be applied to the more general problem of testing H_{m-1} against H_m. Specifically, compute estimators $\hat{d}_j$ of the respective values of $d_j = |f''(x_j)| / f(x_j)^3$, for $1 \leqslant j \leqslant 2m - 3$, where x_j represents the jth turning point. (For example, we might take $\hat{d}_j = |\widehat{f''}(\hat{x}_j)| / \hat{f}(\hat{x}_j)^3$, where $\hat{x}_{2j-1}$ is the jth largest mode of $\hat{f}$ and $\hat{x}_{2j}$ is at the position of the deepest local minimum between $\hat{x}_{2j-1}$ and $\hat{x}_{2j+1}$.) Resample from a calibration distribution with exactly $2m - 3$ turning points, for the jth of which the corresponding value of d_j is $\hat{d}_j$. (We may obtain such a distribution by taking mixtures of the distributions at equation (2.2). In this respect the invariance of d_j under scale changes is important.) Compute the bootstrap version Δ_{nm}^* of Δ_{nm}, with the latter defined as in Section 1.2. Then the conditional distribution of Δ_{nm}^*, given the data, is a consistent estimator of the distribution of Δ_{nm} under the null hypothesis; see Section 4.

3. Numerical performance

16 different distributions were included in the simulation study, six of them unimodal (illustrated in Fig. 1), eight bimodal and two trimodal. Most were mixtures of normal or Student t-distributions. The sample size n was 50, 100 or 200. Throughout, $\hat{f}$ and $\hat{f}^{(r)}$ were taken as kernel estimators based on the normal kernel. Note that $\hat{f}$ and $\hat{f}^{(r)}$ require different amounts of smoothing. We used the respective asymptotically optimal global bandwidths, with unknown quantities depending on the density replaced by those for a normal $N(0, S^2)$ distribution, where S^2 was the sample variance. Whenever values of the kernel estimates over a range were needed, they were computed on an equally spaced grid of 512 points. For brevity, not all the simulation results are given here. Further information may be obtained from the authors on request.

We compared our method with Silverman's (1981) bandwidth test, which is based on the density estimator $\hat{f}_{\text{crit}} = \hat{f}(\cdot|\hat{h}_{\text{crit}})$ where $\hat{h}_{\text{crit}}$ is the smallest bandwidth such that $\hat{f}(\cdot|h)$ has a unique mode. The reader is referred to Silverman (1981) for details of the test. For each

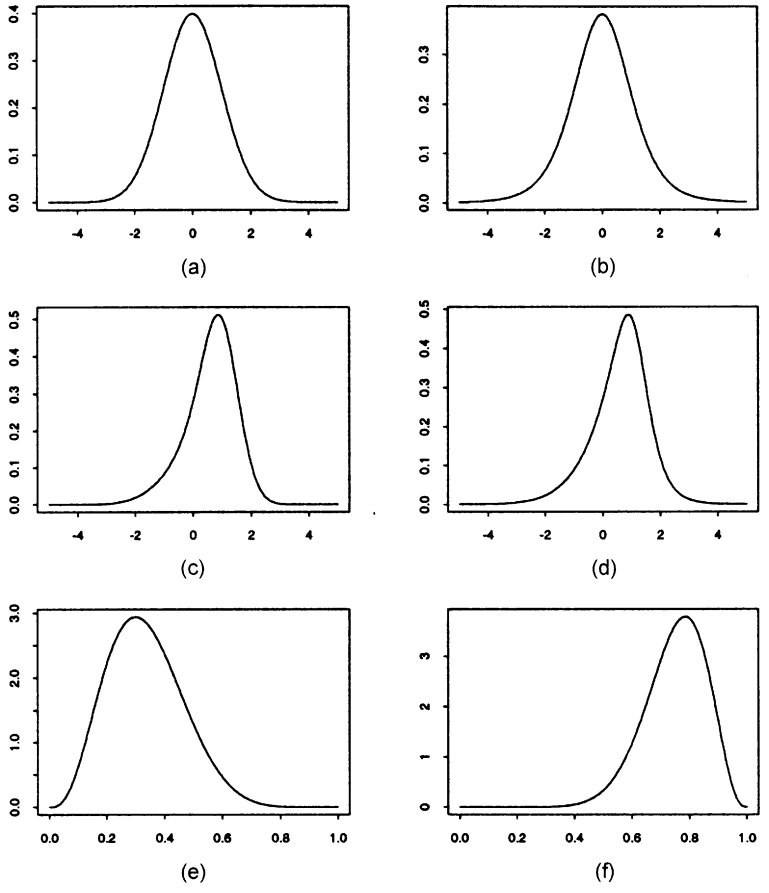

Fig. 1. Density functions of six unimodal distributions corresponding to those whose level accuracies are explored in Table 1: (a) standard normal; (b) Student's t with 6 degrees of freedom; (c) a skewed normal mixture; (d) a skewed Student t-mixture with 6 degrees of freedom; (e) beta(4, 8); (f) beta(12, 4)

choice of sampling distribution and sample size, 500 realizations of the sample $\mathcal{X}$ were generated. Conditionally on each of those samples, 500 resamples of size n were drawn from the population with density g_β (for our calibration approach to the excess mass test) or $\hat{f}_{\text{crit}}$ (for the bandwidth test), and the test statistic was evaluated using the resamples to approximate the conditional distribution. The actual probabilities of rejecting the null hypothesis were approximated by the proportion of times that the null hypothesis was rejected.

The bandwidth test is affected by data clusters in the tails of a distribution. For example, if the sampling distribution is a finite mixture of Student t-distributions, and if the bandwidth test is applied to the full data set, then $\hat{h}_{\text{crit}}$ will diverge to ∞ as the sample size increases, to 'smooth out' spurious modes caused by outlying data points. As a result, both the coverage accuracy and the power suffer. The problems are less dramatic, but nevertheless serious, in small to moderate samples. In order not to penalize the bandwidth test for this behaviour we conducted a separate simulation study in which we confined attention to the bandwidth test applied to data that lay within l standard deviations of at least one mode. (Of course, here we were using knowledge of the true positions of the modes.) We examined $l = 1.0, 1.5, 2.5, 3.5, 4.5$ and found that choosing l as small as 1.0 often excluded information that was helpful, even for good level accuracy under the null hypothesis, let alone good power under bimodal alternatives. In contrast, choosing $l \geqslant 2.5$ often led to excessively large values of $\hat{h}_{\text{crit}}$, and consequent poor level accuracy and power. We found that $l = 1.5$ gave the best results overall, and we shall report our results for the bandwidth test in that case. This special consideration gives the bandwidth test a greater advantage than it would be likely to find in practice.

Another approach to correcting the bandwidth test is to count only those modes of height greater than $1.5\,K(0)/nh$, say. This method was included in the simulation study, and in the sense of level accuracy for heavy-tailed distributions it was found to be somewhat superior to the approach of excluding modes in the tails. For example, for the Student t-distribution with 6 degrees of freedom when the nominal level was 0.01, the true levels were 0.008 (for both $n = 50$ and $n = 200$). When the nominal level was 0.05 the true levels were 0.032 ($n = 50$) and 0.040 ($n = 200$). However, for light-tailed distributions it performed very conservatively. For example, for the normal distribution the true levels were 0.002 ($n = 50$) and 0.004 ($n = 200$) when the nominal level was 0.01, and 0.012 (for both $n = 50$ and $n = 200$) when the nominal level was 0.05. Overall, the level accuracy and power were inferior to those of our calibrated version of the excess mass test and, for the light-tailed densities, inferior to those of the bandwidth test restricted to central parts of the sampling distribution.

Table 1 reports estimates of the true levels, for a variety of nominal levels and for the six unimodal test distributions depicted in Fig. 1, of our calibration approach to the excess mass test and of the bandwidth test (in parentheses). The standard errors of these estimates are approximately $0.04\,\alpha^{1/2}$, where α denotes the nominal level. In all cases the bandwidth test is conservative relative to our calibration of the excess mass test. Generally, our calibration approach has good level accuracy and is slightly conservative. Its level accuracy is slightly, but not to any severe extent, impaired by heavy tailedness (see our comparison of normal with Student t-mixtures below) or skewness (compare the standard normal with the skewed normal mixture, or the corresponding Student t and Student t-mixture). However, for both of the two skewed beta distributions it enjoys good performance. The slight problem that it has with heavy-tailed distributions is substantially less than that of the unrestricted bandwidth test and would vanish if, as in the case of the bandwidth test, we were to restrict attention to data within a few standard deviations of the mode.

Although we have demonstrated these advantages of our approach only for small to

Table 1. Level accuracy for six unimodal distributions†

n	*Estimated true levels for the following nominal levels:*			
	0.01	*0.05*	*0.10*	*0.20*
Standard normal				
50	0.014 (0.004)	0.036 (0.012)	0.076 (0.036)	0.176 (0.108)
200	0.010 (0.006)	0.048 (0.012)	0.086 (0.028)	0.198 (0.126)
Student t, 6 degrees of freedom				
50	0.012 (0.002)	0.032 (0.010)	0.068 (0.026)	0.152 (0.080)
200	0.012 (0.002)	0.024 (0.008)	0.066 (0.022)	0.138 (0.086)
Skewed normal mixture				
50	0.008 (0.006)	0.042 (0.020)	0.088 (0.068)	0.166 (0.164)
200	0.006 (0.002)	0.026 (0.010)	0.056 (0.040)	0.138 (0.110)
Skewed Student t-mixture				
50	0.008 (0.006)	0.028 (0.026)	0.068 (0.052)	0.158 (0.170)
200	0.008 (0.002)	0.020 (0.006)	0.054 (0.036)	0.122 (0.110)
beta(4, 8)				
50	0.018 (0.004)	0.056 (0.024)	0.092 (0.052)	0.190 (0.126)
200	0.010 (0.004)	0.042 (0.018)	0.090 (0.048)	0.186 (0.140)
beta(12, 4)				
50	0.018 (0.006)	0.062 (0.034)	0.108 (0.086)	0.226 (0.212)
200	0.018 (0.006)	0.056 (0.028)	0.118 (0.060)	0.186 (0.148)

†Estimated true levels, for the calibrated excess mass test and for the bandwidth test (values in parentheses), approximated by 500 simulations.

moderate sample sizes, they also appear in the limit as $n \to \infty$. Indeed, we shall show in Section 4 that our calibrated excess mass test has the asymptotically correct level. It is known that the asymptotic conservatism of the bandwidth test persists in the limit. This has recently been quantified by York (1998), who has shown that for nominal levels 0.01, 0.05, 0.1 and 0.2 the asymptotic levels of the bandwidth test are 0.000, 0.010, 0.032 and 0.102. The asymptotics alluded to here, for the bandwidth test, are for a distribution (such as the beta$(2, 2)$ distribution) which decreases steeply to 0 at the ends of its support, so that the outlier problems discussed earlier do not arise.

Figs 2–4 display the powers against most of the bimodal distributions considered in the simulations. Figs 2(a), 3(a) and 4(a) depict the sampling density (which is a mixture of normal distributions), and Figs 2(b), 2(c), 3(b), 3(c), 4(b) and 4(c) plot the approximate probability of rejecting the null hypothesis of unimodality, at the nominal level, for sample sizes $n = 50$ and $n = 200$, and for both tests.

In general, calibration of the excess mass test produces a test of greater power than the bandwidth test, particularly in challenging cases where the sampling distribution is almost unimodal (in the sense that the two modes are close to each other and the second mode is barely distinguishable from the major mode). This is due largely to conservatism of the bandwidth approach. However, if one considers carefully the definition of excess mass, it is possible to construct examples of bimodal densities where the true excess mass difference of the density is very small. Clearly, this will cause the excess mass test trouble in detecting bimodality. The distribution whose density is plotted in Fig. 4 was deliberately chosen to be of this type. It has a clearly visible second mode, but low true excess mass. As a result the

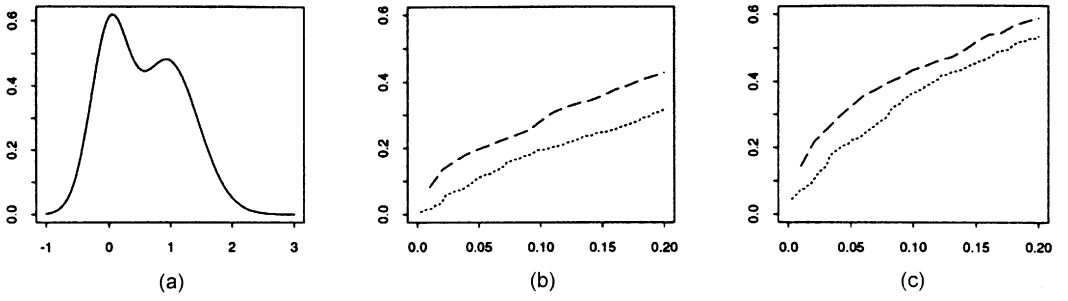

Fig. 2. Power against light-tailed bimodal distributions: (a) sampling density; (b) power against nominal level for our calibrated excess mass test (- - - -) and for the bandwidth test (·····), when the sample size is 50; (c) corresponding powers of the two tests when the sample size is 200

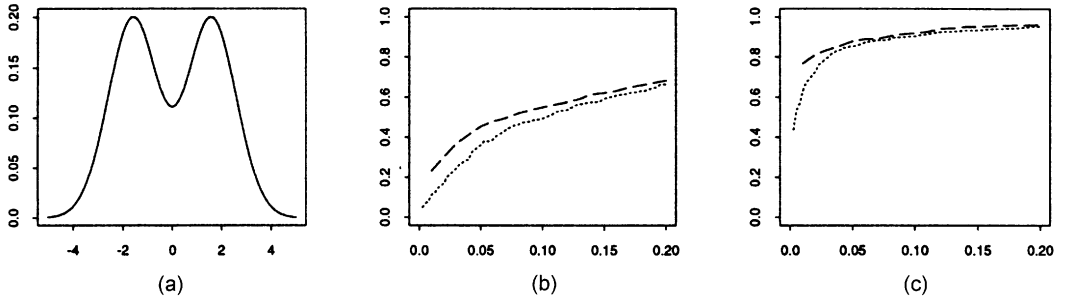

Fig. 3. Power against light-tailed bimodal distributions: (a) sampling density; (b) power against nominal level for our calibrated excess mass test (- - - -) and for the bandwidth test (·····), when the sample size is 50; (c) corresponding powers of the two tests when the sample size is 200

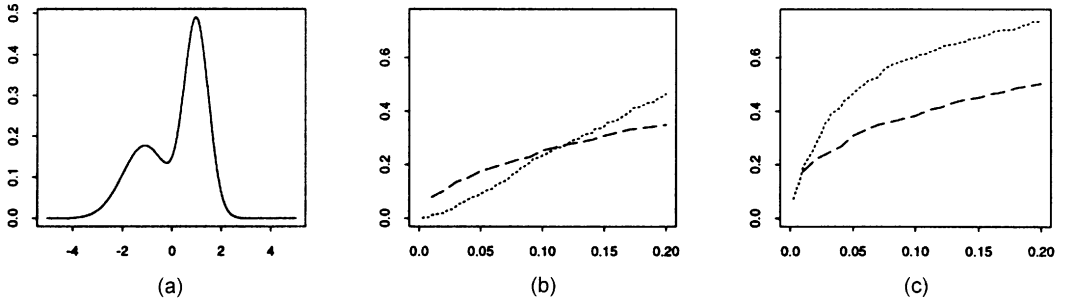

Fig. 4. Power against light-tailed bimodal distributions: (a) sampling density; (b) power against nominal level for our calibrated excess mass test (- - - -) and for the bandwidth test (·····), when the sample size is 50; (c) corresponding powers of the two tests when the sample size is 200

bandwidth test has, except for $n = 50$ or for $n = 100$ and small nominal levels, greater power than calibration of the excess mass test.

Both tests were applied to two trimodal (normal mixture) distributions, one symmetric and the other skewed. In each case the bandwidth test had less power than the calibrated excess mass test, and the power was a little less for the symmetric density than for the skewed density.

Fig. 5 shows the powers of the calibrated excess mass test and of the bandwidth test against heavy-tailed versions of the distributions addressed in Figs 2–4. Three mixtures of Student t-densities were designed to resemble the three normal density mixtures (in the earlier figures) in all respects except their tails, so that power differences arising from heavy tailedness would not be confounded with other effects. Comparing Figs 2–4 with the corresponding rows of Fig. 5 we see that heavy tailedness consistently reduces the power, for both types of test, and that in general the power decreases more for the bandwidth test. This is despite the fact that our construction of the bandwidth test was carefully designed to minimize effects of heavy tailedness. Without this care, the effects of heavy tailedness on the power of the bandwidth test can be very severe indeed.

The numerical results reported in Table 1 and Figs 2–5 convey in three ways the fact that our calibration of the excess mass test is generally more favourable than the bandwidth test.

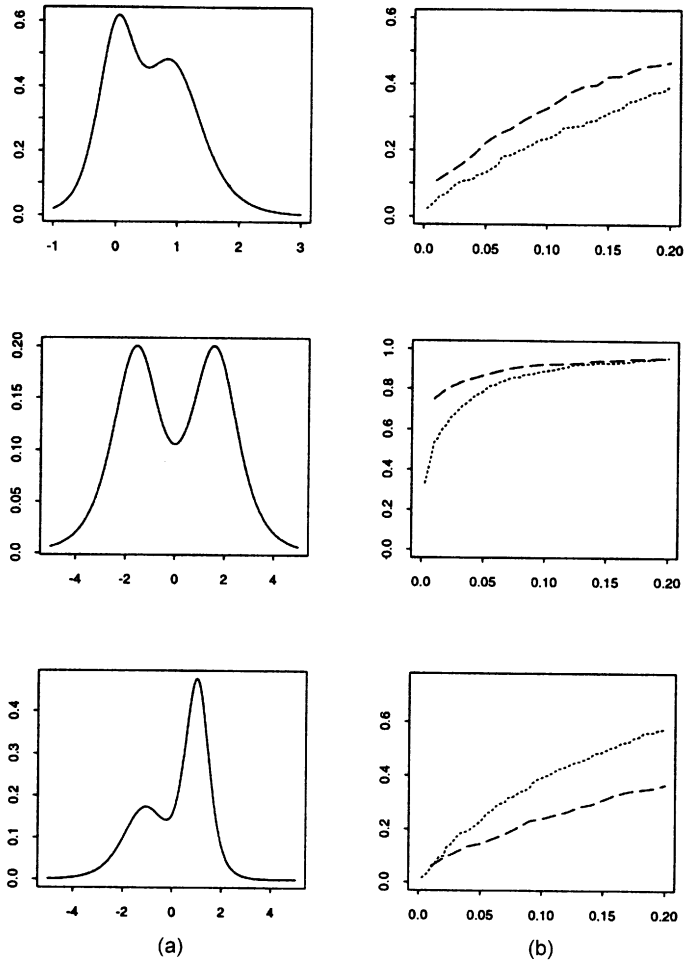

Fig. 5. Power against heavy-tailed bimodal distributions: (a) sampling density; (b) powers of the calibrated excess mass test (- - - -) and the bandwidth test (·····), against nominal level when the sample size n is 200 (each density is a mixture of two Student t-distributions with the same number of degrees of freedom, different locations and (for the top and bottom rows) different scales; the Student t-densities used for the middle row have 4 degrees of freedom, whereas those for the top and bottom rows have 6 degrees of freedom)

First, it has very good level accuracy in a wide variety of situations, including skewed or heavy-tailed unimodal distributions. Secondly, it has greater power in many bimodal cases, most significantly when the two modes are not well separated from each other. Thirdly, the effect of heavy tailedness on its power is comparatively minor.

Silverman (1981) applied the bandwidth test to the bench-mark chondrite data set, of size $n = 22$, describing the percentage of silica in 22 meteors. For comparison we applied our calibration of the excess mass test to the same data, obtained from Simonoff (1996). The calibration method rejects the null hypothesis of unimodality at level 0.03, whereas the bandwidth test rejects the null hypothesis only at the larger level 0.08. This reflects the fact that the bandwidth test tends to be conservative and therefore to have lower power.

4. Theoretical properties

We begin by describing the limiting distribution of the excess mass statistic Δ_{n2}. Let W denote a standard Wiener process on the real line. Given real numbers $t_1 < t_2$ and u, define

$$\Delta(t_1, t_2, u) = W(t_2) - W(t_1) - (t_2^3 - t_1^3) + u(t_2 - t_1).$$

Put

$$Z = 6^{1/5} \sup_{-\infty<u<\infty} \left[\sup_{-\infty<t_1<\ldots<t_4<\infty} \{\Delta(t_1, t_2, u) + \Delta(t_3, t_4, u)\} - \sup_{-\infty<t_1<t_2<\infty} \{\Delta(t_1, t_2, u)\} \right]. \quad (4.1)$$

Theorem 1. With probability 1, Z is finite and positive. Its distribution does not have any atoms.

Next we give regularity conditions, similar to those of Müller and Sawitzki (1991a), under which Δ_{n2} is asymptotically distributed as a multiple of Z:

> the sampling density f has a continuous derivative, ultimately monotone in each tail; the constraints $f'(x_0) = 0$ and $f(x_0) \neq 0$ are jointly satisfied at just one point x_0, and f'' exists and is Hölder continuous within a neighbourhood of x_0, with $f''(x_0) < 0$. (4.2)

Define c as at equation (2.1).

Theorem 2. Under condition (4.2), $n^{3/5}\,\Delta_{n2}$ converges in distribution to cZ as $n \to \infty$.

Theorem 2 justifies the assumption in Section 2 that in large samples the distribution of Δ_{n2} under the null hypothesis is virtually independent of unknowns, except for a factor. Our next result shows that the calibration method suggested in Section 2 consistently estimates the distribution of Δ_{n2}.

Let $\hat{d}$ be a positive function of the data $\mathcal{X}$, let $\mathcal{X}^*$ denote a resample drawn by sampling randomly, with replacement, from either of the two calibration distributions determined by equations (2.2) (for which β is replaced by $\hat{\beta} = \gamma^{-1}(\hat{d})$, with γ defined by equations (2.3)) and let Δ_{n2}^* be the corresponding version of Δ_{n2}.

Theorem 3. If $\hat{d}$ converges in probability to a constant $d > 0$, then

$$\sup_{-\infty<x<\infty} |P_{G_{\hat{\beta}}}(n^{3/5}\Delta_{n2}^* \leqslant x|\mathcal{X}) - P(cZ \leqslant x)| \to 0$$

in probability as $n \to \infty$, where $c = d^{-1/5}$.

Theorems 1–3 imply that, in the case of testing the null hypothesis of unimodality, our calibration method produces tests with asymptotically correct level. More generally, each of theorems 1–3 has an analogue in the setting of testing H_{m-1} against H_m. In particular, under suitable regularity conditions (which include the assumption that H_{m-1} holds), $n^{3/5}\,\Delta_{nm}$ converges in distribution to a random variable $Z(c_1, \ldots, c_{2m-3})$, depending only on the versions $c_1, \ldots, c_{2m-3}$ of c at the $2m-3$ turning points. Assuming that we have consistent estimators of these quantities, the limiting bootstrap distribution of $n^{3/5}\Delta^*_{nm}$ is identical with the distribution of $Z(c_1, \ldots, c_{2m-3})$, and so the test has asymptotically correct level.

References

Aitkin, M. and Rubin, D. B. (1985) Estimation and hypothesis testing in finite mixture models. *J. R. Statist. Soc.* B, **47**, 67–75.

Cheng, M.-Y. and Hall, P. (1997) Calibrating the excess mass test of modality. *Research Report SRR-002-97*. Centre for Mathematics and Its Applications, Australian National University, Canberra.

Cox, D. R. (1966) Notes on the analysis of mixed frequency distributions. *Br. J. Math. Statist. Psychol.*, **19**, 39–47.

Everitt, B. S. and Hand, D. J. (1981) *Finite Mixture Distributions*. London: Chapman and Hall.

Finch, S. J., Mendell, N. R. and Thode, H. C. (1989) Probabilistic measures of adequacy of a numerical search for a global maximum. *J. Am. Statist. Ass.*, **84**, 1020–1023.

Fisher, N. I., Mammen, E. and Marron, J. S. (1994) Testing for multimodality. *Comput. Statist. Data Anal.*, **18**, 499–512.

Good, I. J. and Gaskins, R. A. (1980) Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Am. Statist. Ass.*, **75**, 42–73.

Hartigan, J. A. (1985) A failure of likelihood asymptotics for normal mixtures. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* (eds L. LeCam and R. A. Olshen), vol. II, pp. 807–810. Pacific Grove: Wadsworth and Brooks.

———(1987) Estimation of a convex density contour in two dimensions. *J. Am. Statist. Ass.*, **82**, 267–270.

Hartigan, J. A. and Hartigan, P. M. (1985) The DIP test of unimodality. *Ann. Statist.*, **13**, 70–84.

Mammen, E., Marron, J. S. and Fisher, N. I. (1992) Some asymptotics for multimodality tests based on kernel density estimates. *Probab. Theory Reltd Flds*, **91**, 115–132.

Mammen, E. and Tsybakov, A. B. (1995) Asymptotic minimax recovery of sets with smooth boundaries. *Ann. Statist.*, **23**, 502–524.

McLachlan, G. J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.*, **36**, 318–324.

Minnotte, M. C. and Scott, D. W. (1992) The mode tree: a tool for visualization of nonparametric density features. *J. Comput. Graph. Statist.*, **2**, 51–68.

Müller, D. W and Sawitzki, G. (1987) Using excess mass estimates to investigate the modality of a distribution. *Sonderforschungsbereich 123*. Universität Heidelberg, Heidelberg.

———(1991a) Excess mass estimates and tests for multimodality. *J. Am. Statist. Ass.*, **86**, 738–746.

———(1991b) Using excess mass estimates to investigate the modality of a distribution. In *The Frontiers of Statistical Scientific Theory and Industrial Applications* (eds A. Öztürk, E. C. van der Meulen, E. J. Dudewicz and P. R. Nelsen), vol. II, pp. 355–382. Syracuse: American Sciences Press.

Nolan, D. (1991) The excess-mass ellipsoid. *J. Multiv. Anal.*, **39**, 348–371.

Polonik, W. (1995a) Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.*, **23**, 855–881.

———(1995b) Density estimation under qualitative assumptions in higher dimensions. *J. Multiv. Anal.*, **55**, 61–81.

Roeder, K. (1994) A graphical technique for determining the number of components in a mixture of normals. *J. Am. Statist. Ass.*, **89**, 487–495.

Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality. *J. R. Statist. Soc.* B, **43**, 97–99.

———(1983) Some properties of a test for multimodality based on kernel density estimates. In *Probability, Statistics and Analysis* (eds J. F. C. Kingman and G. E. H. Reuter), pp. 248–259. Cambridge: Cambridge University Press.

Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. New York: Springer.

York, M. G. (1998) Some problems in testing for modality. *PhD Thesis*. Australian National University, Canberra.