Detecting the Presence of Mixing with Multiscale Maximum Likelihood

Guenther WALTHER

A test of homogeneity tries to decide whether observations come from a single distribution or from a mixture of several distributions. A powerful theory has been developed for the case where the component distributions are members of an exponential family. When no parametric assumptions are appropriate, the standard approach is to test for bimodality, which is known not to be very sensitive for detecting heterogeneity. To develop a more sensitive procedure, this article builds on an approach employed in sampling literature and models the component distributions as logarithmically concave densities. It is shown how this leads to a special semiparametric model, in which the homogeneity problem is equivalent to testing whether a parameter c equals zero, versus the alternative that $c > 0$. This set-up leads naturally to a novel multiscale maximum likelihood procedure, where the multiscale character reflects the desirable property of adaptivity to the unknown value of the parameter c under the alternative, to ensure a test with high power. The test can also be extended, in principle, to a multivariate situation. In a univariate setting, the multiscale procedure is well suited to computation with the iterative convex minorant algorithm, a recent innovation in computational nonparametric statistics. The test is applied to simulated data and to the stamp data of Izenman and Sommer.

KEY WORDS: Log-concave; Mixture detection

1. INTRODUCTION

This article is concerned with the homogeneity problem: Given a sample of observations, one wishes to decide whether the observations come from one distribution (the homogeneity case) or from a mixture of distributions. The homogeneity problem arises in many different contexts. One prominent example is the Pickering/Platt debate on the nature of high blood pressure. The English internist Platt claimed that hypertension was a disease, which one either had or not. Thus the observed blood pressure measurements would follow a mixture of distributions. Pickering argued that there was no single dominating cause for hypertension, and that doctors arbitrarily labeled people in the right tail of the blood pressure distribution as having hypertension. See Swales (1985) for an account of this dispute. A more recent example from the physical sciences concerns the study of so-called moving groups of stars. Stars tend to form groups, which astrophysicists try to detect by examining the distribution of the radial velocity of a large number of stars. The presence of a moving group of stars will result in a mixture of velocity distributions. See Côté et al. (1993) for the discovery of a moving group via a statistical analysis of the velocity distribution.

The statistical theory on the homogeneity problem has been developed very successfully in the case where the component densities are assumed to be from a known exponential family. Important references are Lindsay and Roeder (1992, 1997) and Roeder (1994). More references to relevant work can be found in the authoritative monograph of Lindsay (1995). A key aspect of those approaches is the observation that mixing in exponential families manifests itself in the sign-change behavior of certain functions, which can then be examined statistically. Aspects of that approach go back to Shaked (1980).

The conclusions of such an analysis usually depend quite sensitively on the assumed parametric model for the component densities. In particular, skewed component distributions can cause problems (see Roeder 1994, p. 493, and Schork et al. 1990). Hence if no prior information is available concerning the component distributions, then the standard approach is a nonparametric test for the number of modes (see Titterington et al. 1985, p. 48, for a judicious discussion). It is well known, however, that such an approach is not very sensitive for detecting the presence of mixing (see, e.g., Roeder 1994, p. 493). Thus there exists a gap between the criteria and methods available in the nonparametric case on the one hand and the quite definitive results (see the articles cited in the previous paragraph) for the parametric case on the other hand. The aim of this article is to develop a counterpart for a more sensitive analysis in the nonparametric case. The article builds on a proposal from the sampling literature that models single-component distributions as logarithmically concave functions and then develops the necessary tools for statistical inference.

Section 2 motivates this set-up and gives a representation theorem that provides a useful framework for the statistical analysis. Section 3 shows that inference in this model leads naturally to a novel method of multiscale maximum likelihood analysis, which can also be motivated by certain optimality criteria. Section 4 addresses computational issues. It turns out that the multiscale maximum likelihood estimator (MLE) can be computed in a nice and fast way with the Iterative Convex Minorant Algorithm (see Jongbloed 1998), a recently developed tool in computational statistics. In Section 5 simulation results are reported, and the procedure is applied to the stamp data of Izenman and Sommer (1988). Proofs are deferred to the Appendix.

2. A REPRESENTATION THEOREM

Clearly, the homogeneity problem described in the Introduction becomes meaningful only if some assumptions are made about the component densities. We will adopt an approach commonly used in the sampling literature (see Gilks and Wild 1992; Dellaportas and Smith 1993; and Brooks 1998) and

Guenther Walther is Associate Professor, Department of Statistics, 390 Serra Mall, Stanford University, Stanford, CA 94305 (E-mail: *walther@stat.stanford.edu*). The author thanks the associate editor and two referees for their careful reading of the manuscript and their suggestions for improvement. Support from grants DMS 9704557 and DMS 9875598 is gratefully acknowledged.

model the class of single-component distributions with the class of logarithmically concave densities. That is, we consider component densities of the form $f(x) = e^{\phi(x)}$, where ϕ is a concave (but not necessarily smooth) function. As expounded in the above references, the most commonly used parametric densities are log-concave, the prime example of course being the normal density where ϕ is a quadratic. Hence log-concave densities are a quite natural choice, if one looks for a class that contains the commonly used distributions, but is also parsimonious. Note that the class of log-concave distributions is strictly contained in the class of unimodal distributions. Hence modeling single-component distributions as log-concave densities makes it possible to construct a procedure that is more sensitive for detecting mixing than would be possible under a unimodal model. This proposition will be illustrated by a normal mixture example in Section 5. As log-concave densities are precisely the densities that are totally positive of order 2, this model can also be regarded as a nonparametric counterpart to the remarkably successful approach developed by Lindsay and Roeder (1992, 1997) and Roeder (1994) in the parametric case.

The following representation theorem provides a framework for the analysis. It is proved without much extra effort for a multivariate set-up. In fact, the methodology described below can be readily transferred to a multivariate situation, but the computational algorithms will be different.

Theorem 1. Let the f_i be logarithmically concave densities on $\mathbf{R}^d$ and $p_i \geq 0, i = 1, \ldots, m$. Then on any compact set $G \subset \bigcap_{i=1}^m$ (support of f_i) the representation

$$f(x) := \sum_{i=1}^{m} p_i f_i(x) = \exp(\phi(x) + c \parallel x \parallel^2) \qquad (1)$$

holds for a concave function ϕ on $\mathbf{R}^d$ and a constant $c \geq 0$.

Figure 1 displays several examples of functions obeying (1), which were obtained in the fitting of a data set discussed in Section 5. For univariate functions f, (1) asserts an asymmetric smoothness condition for $\log f$: its (one-sided) derivative can jump down, but an increase of the derivative must satisfy a Lipschitz condition due to the derivative of the term $c \parallel x \parallel^2$, as the one-sided derivative of the concave function ϕ is nonincreasing. It will be seen that an approach based on maximum likelihood readily accounts for this property. The constant c in (1) is uniquely determined if we consider its smallest value for which (1) holds. We will use this convention in the following. The technical condition placed on G will turn out not to be essential for the methodology that will be developed below.

Theorem 1 shows that testing whether f can be represented as a single log-concave distribution is equivalent to testing whether $c = 0$ in (1). We will call the case $c = 0$ the *null model*. Assessing the evidence of mixing amounts to assessing the evidence that $c > 0$. It can happen that mixing two log-concave distributions will again result in a log-concave distribution, that is, in $c = 0$ in (1), in which case it is impossible to determine whether mixing is present. This lack of identifiability is the price one has to pay if no prior information about the component distributions is available and one chooses a nonparametric approach.

3. THE MULTISCALE MLE

As interest centers on the parameter c of the log density, it is natural to use the method of maximum likelihood for statistical inference. Note that the parameter c in (1) is unknown, and its MLE clearly does not exist. This motivates the use of maximum likelihood in a *multiscale* manner: given n i.i.d. observations $X_1, \ldots, X_n$, we will compute the nonparametric MLE $\hat{f}_n^c$ in (1) for various fixed positive values of c and then assess and combine the evidence for these various values of c versus the null model. The rationale for this procedure is the following. For fixed c, $\log \hat{f}_n^c$ is a piecewise parabolic function with parabolas of curvature c (see Proposition 1 below and the examples in Fig. 2). For large values of c, $\log \hat{f}_n^c$ will be a rough function with deep dips between the observations, whereas for c small, $\log \hat{f}_n^c$ will have only shallow dips between the observations and will be nearly concave under the null model. If the mixture f satisfies (1) with a small value c_{true}, then the deviation of $\log f$ from concavity will manifest

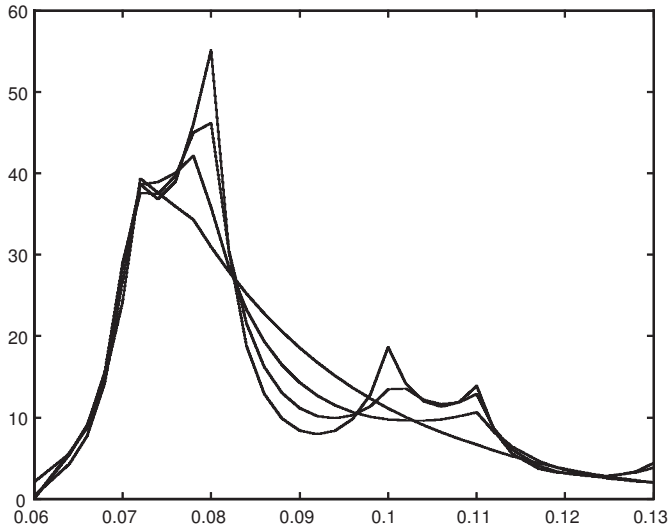

Figure 1. $\hat{f}_n^c$ for the Stamp Data for c = 0, 0.6, 1.8, 3.

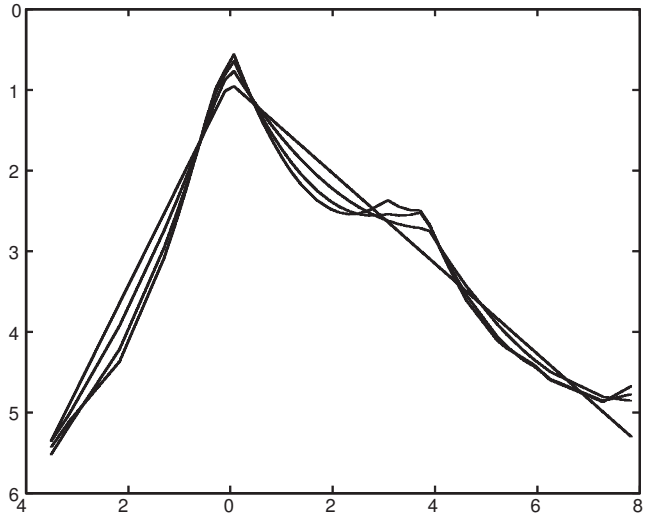

Figure 2. $\log \hat{f}_n^c$ for the Data in Figure 3 for c = 0, 0.4, 0.8, 1.2.

itself in a shallow dip, which may be impossible to detect in the rough function $\log \hat{f}_n^c$ when c is large. However, such a shallow but elongated dip in $\log \hat{f}_n^{c_{\text{true}}}$ may well represent a significant deviation from the expected nearly concave shape of this function under the null model. Conversely, a large value of c_{true} allows a deep but localized dip in $\log f$. If c is small, then $\log \hat{f}_n^c$ is not flexible enough to fit this localized dip and hence will not produce the significant result obtainable with $\log \hat{f}_n^{c_{\text{true}}}$. But c_{true} is unknown. Combining the evidence for various values of c is a way to make the procedure adaptive to the unknown c_{true}.

For the related model where ϕ is a decreasing function and the term cx^2 is replaced by cx in (1) (i.e., one tests whether the density is decreasing versus alternatives that are locally smoothly increasing), Walther (in press) showed that the multiscale MLE approach indeed results in a procedure that is adaptive and optimal in the asymptotic minimax framework. This motivates the use of the corresponding approach for the model (1), for which a theoretical analysis appears to be much more difficult. The next proposition summarizes some properties of the MLE:

Proposition 1. Fix $c \geq 0$. Then the MLE $\hat{f}_n^c$ for the model (1) exists and is of the form $\hat{f}_n^c(x) = \exp(\hat{\phi}_n(x) + cx^2)$, where $\hat{\phi}_n$ is the solution of the optimization problem

$$\max_{\phi \text{ concave}} \frac{1}{n} \sum_{i=1}^{n} \phi(X_i) - \int_{X_{(1)}}^{X_{(n)}} \exp(\phi(x) + cx^2)\, dx. \qquad (2)$$

The concave function $\hat{\phi}_n$ is piecewise linear with changes of slopes at the observations. The left-hand slopes $\{\hat{\phi}_n'(X_{(i)}-), 1 \leq i \leq n\}$ at the observations are contained in the cone

$$C_n := \{y \in \mathbf{R}^n : y_1 \geq y_2 \geq \cdots \geq y_n\}. \qquad (3)$$

We will assess the evidence against the null model by measuring the distance of $\log \hat{f}_n^c$ from the class of concave functions or, alternatively, the distance $T_n(c)$ of the (left-hand) derivative of $\log \hat{f}_n^c$ from *Mon*, the class of decreasing functions from $\mathbf{R}$ into $\mathbf{R}$, the extended real line. More precisely, the distance of a function g from *Mon* is defined as $d_w(g, Mon) = \inf_{m \in Mon} \| (g-m)w \|_\infty$, where w is a weight function which we will take to be $\hat{f}_n^0$. Down-weighting the tails of the distribution is a desirable feature for many applications. The choice of the supremum norm is motivated by its sensitivity to local violations of monotonicity (see Liero et al. 1998) and by the fact that it results in a relevant and interpretable measure of the deviation from the null model. It is shown in Section 4 that the evaluation of this distance $T_n(c)$ reduces to the computation of a simple expression and that computing the MLE $\hat{f}_n^c$ for various scales c goes hand in hand with the workings of the iterative algorithm employed. The implementation will of course use a finite set $\mathcal{C}$, usually a grid, for the range of possible values of c that are to be inspected.

Finally, the reference distribution for assessing the significance of the statistic is obtained by Monte Carlo sampling from the estimated null model $\hat{f}_n^0$. This approach follows that of Silverman (1981). The complete description of the test is then as follows.

Draw $X_1^*, \ldots, X_n^*$ i.i.d. from $\hat{f}_n^0$ and compute $T_n^*(c)$ for each $c \in \mathcal{C}$. Repeat this B times and compute the mean $m(c)$ and the standard deviation $s(c)$ of the B copies $T_n^{*1}(c), \ldots, T_n^{*B}(c)$. The p value of this test is now estimated by comparing the so standardized original values $T_n(c)$ with the standardized reference values:

$$p = \left(\#\left\{ \max_{c \in \mathcal{C}} \big((T_n(c) - m(c))/s(c) \big) \right.\right.$$
$$\left.\left. \leq \max_{c \in \mathcal{C}} \big((T_n^{*i}(c) - m(c))/s(c) \big), 1 \leq i \leq B \right\} + 1 \right) \Big/ (B+1).$$

Results of a small simulation study that compares observed relative frequencies of rejection with nominal levels are reported in Section 5.

While a precise theoretical analysis of the adaptivity properties of the test in the model (1) appears to be difficult, the following proposition shows that the test is consistent:

Proposition 2. Let the density f be of the form (1) with $c > 0$ and with compact support G. Then the above multiscale test is consistent against f; that is, its power converges to 1 as $n \to \infty, B \to \infty$.

4. COMPUTING THE MULTISCALE MLE

The optimization problem (2) can be put as the minimization of a convex function over the cone C_n given in (3), and as such is well suited for treatment with the Iterative Convex Minorant Algorithm (ICMA), a recent advance with potential applications in many nonparametric problems (see Groeneboom and Wellner 1992 and Jongbloed 1998). The key idea of that algorithm is to approximate the convex function locally around the current candidate solution by a quadratic form, which can then be minimized over the cone with standard tools from isotonic regression, such as the pool-adjacent-violators algorithm. This procedure is then iterated to the final solution.

We will now give some more details for the present context. It is helpful for the binning described below to scale the observations as well as the Monte Carlo samples to have variance unity, and then perform the test on this scale. Linearly binning the data, that is, assigning the empirical measure of each observation to each of its two neighboring grid points in proportion to the distance of the observation from the other grid point (see, e.g., Wand 1994), was not found to affect the solution much but provides two obvious advantages: First, quantized data, such as those involving rounding errors, are readily incorporated into this setting. Second, the running time of the algorithm will be shorter, as the optimization problem is reduced to a lower-dimensional space. Distances between the bin centers as large as one-tenth of the standard deviation of the data were found to give satisfactory results. Let $x_2 < \cdots < x_{n+1}$ denote the bin centers, so that n now refers to the number of bins, not the sample size. $x_1 := x_2 - 0.1$ is a dummy point whose purpose will become clear immediately. The binning procedure will associate weights w_i with the x_i such that $\sum_i w_i = 1$ (see Wand 1994 for the details of linear binning). For a given candidate solution f write $l_i := \log f(x_i)$. The associated concave function $\phi(x) = \log f(x) - cx^2$ can be taken to be piecewise linear by Proposition 1, and hence the

candidate solution is specified by the slopes s_i of ϕ between x_i and x_{i+1}, $1 \le i \le n$, together with the value l_1 at the dummy point x_1. Setting $l_1 := -10$ should be a small enough choice for almost any situation. In summary, we have

$$l_i = l_1 + \sum_{j=1}^{i-1} s_j(x_{j+1} - x_j) + c(x_i^2 - x_1^2), \quad 1 \le i \le n+1, \quad (4)$$

where the vector $\mathbf{s} = (s_1, \ldots, s_n)$ must be contained in C_n given in (3). The likelihood for the binned data is $\sum_{i=2}^{n+1} w_i l_i = \sum_{i=1}^{n} s_i(x_{i+1} - x_i) \sum_{j=i+1}^{n+1} w_j$. In all of the examples given in Section 5, the binned likelihood and the resulting MLE give an excellent approximation to the corresponding quantities for the unbinned data. Using a standard approximation to the integral in (2), the latter optimization problem becomes

$$\min_{\mathbf{s}\in C_n} \left\{ -\sum_{i=1}^{n} s_i(x_{i+1} - x_i) \sum_{j=i+1}^{n+1} w_j + \frac{1}{2} \left(\sum_{i=3}^{n} e^{l_i}(x_{i+1} - x_{i-1}) + e^{l_2}(x_3 - x_2) + e^{l_{n+1}}(x_{n+1} - x_n) \right) \right\}, \quad (5)$$

where the l_i are given in (4). Note that the criterion function (5) is convex in $\mathbf{s}$. Thus the problem is now in a form where the ICMA can be applied directly; the required gradient and the diagonal of the Hessian of the criterion function can readily be found by differentiating (5) and are given in the Appendix. Two modifications to the ICMA were found to be helpful: first, adding an additional stopping criterion that ends the iterations once the improvement in the criterion function (5) is less then 10^{-5}, and second, replacing each Hessian diagonal weight by 10^{-3} if it falls below that number, to avoid problems with inverting those elements. It can be proved that this change of weights will still make the algorithm converge to the solution. This choice of accuracy parameters follows the choice in Jongbloed (1998) for similar situations and was found to work well.

Alternatively to using the ICMA, (5) can be solved by an interior point method (see Terlaky and Vial 1998). Interior point methods require more programming effort and do not appear to be as fast and stable as the ICMA for the situation at hand. Note that the iterative nature of the ICMA goes hand in hand with the structure of the multiscale procedure: The solution of (5) serves as a starting value for that problem with the next larger value of c under consideration. With the distance of 0.1 between the bin centers as described above, the running time of the algorithm per scale is about 4 s on a present-day PC. In Section 5 a normal fit was used to provide the initial starting value for the ICMA in the null case $c = 0$.

Recall that $T_n(c)$ is defined as $d_w(g, Mon)$, with the weight function $w := \hat{f}_n^0$ and g equal to the left-hand derivative of $\log \hat{f}_n^c$. By Proposition 1, the latter equals $s_i + 2cx$ for $x \in (x_i, x_{i+1}]$, where $\mathbf{s}$ is the solution of (5). Furthermore, it can be shown that for general functions g and nonnegative w, $d_w(g, Mon) = \sup_{s<t}(g(t) - g(s))w(s)w(t)/(w(s) + w(t))$, with the usual convention $0/0 := 0$. Thus one obtains $T_n(c) = \max_{2\le i\le j\le n}(s_j - s_i + 2c(x_{j+1} - x_i))\hat{f}_n^0(x_i)\hat{f}_n^0(x_{j+1})/(\hat{f}_n^0(x_i) + \hat{f}_n^0(x_{j+1}))$. Hence $T_n(c)$ gives the largest weighted increase of the derivative of $\log \hat{f}_n^c$ and thus measures the size of the largest "dip" or local convexity in $\log \hat{f}_n^c$.

Sampling from $\hat{f}_n^0$ is straightforward because of the piecewise linear form of its logarithm: First sample an interval from the collection (x_i, x_{i+1}), $2 \le i \le n$, with pertaining probabilities $e^{l_i}(e^{s_i(x_{i+1}-x_i)} - 1)/s_i$, respectively $e^{l_i}(x_{i+1} - x_i)$, if $s_i = 0$. Let $U \sim U[0, 1]$. The desired random variable is then $x_i + X$, where $X = U(x_{i+1} - x_i)$ if $|s_i| < 10^{-5}$ (say), and otherwise $X = \log(U(e^{s_i(x_{i+1}-x_i)} - 1) + 1)/s_i$.

The Matlab code for the described procedure is available from the author upon request.

5. EXAMPLES

In all of the examples and simulations below, $B = 299$ resamples were used to compute the significance of the statistic. Including the null model $c = 0$, $\mathcal{C}$ consisted of 11 equally spaced values c, with the largest c taken to be 3. The results appear not to be sensitive to those choices.

A small simulation study was performed to assess the accuracy of the simulated p values. Three hundred samples of size 100 were generated from the standard uniform distribution, and the test was evaluated for each sample. Of the samples, 0.7% resulted in p values of less than 1%, and 5.3% of the samples resulted in p values of less than 5%. The uniform distribution is a least favorable distribution for this situation, because its log-density is constant, and, hence, everywhere on its support, it lies on the boundary to local violations of concavity. The above simulations show that the relevant observed significance levels coincide quite satisfactorily with the nominal levels. Simulating instead from the standard normal distribution resulted in significant outcomes at relative frequencies below the nominal levels, that is, in conservative results.

Next the methodology was applied to a sample of size 250 from the mixture $\frac{1}{2}N(0, \frac{1}{4}) + \frac{1}{2}N(2, 5)$. A histogram is displayed in Figure 3. This mixture distribution is unimodal, and hence one would expect it to be difficult to detect the two components in the distribution. The test clearly detected

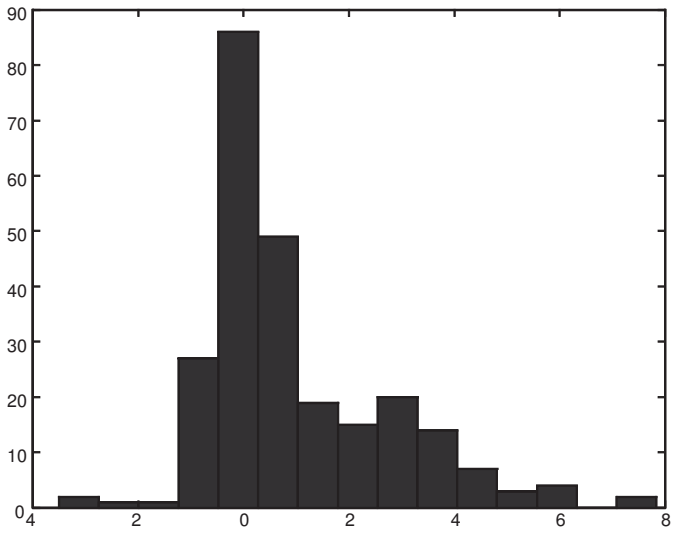

Figure 3. Histogram of 250 Observations From the Mixture $\frac{1}{2}N(0, \frac{1}{4}) + \frac{1}{2}N(2, 5)$.

the presence of mixing (p value <0.01). Figure 2 shows the graphs of $\log \hat{f}_n^c$ for various values of c; the deviation from concavity is quite evident.

Finally, the procedure was applied to the stamp data of Izenman and Sommer (1988). The data are measurements of the thicknesses of 485 stamps of the 1872 Hidalgo Issue of Mexico. A histogram is displayed in Figure 1 in the cited article. The data are quantized because of rounding. Figure 1 shows the graphs of $\hat{f}_n^c$ for various values of c. The test is again significant (p value <0.01). Izenman and Sommer (1988) pursue a further inference about the number of components in the mixture. One of their nonparametric approaches proceeds by counting the number of significant modes in the density. An analogous approach can be developed with the procedure introduced in this article. As was seen in Section 4, $T_n(c)$ measures the size of the largest "dip," that is, the largest local convexity in $\log \hat{f}_n^c$. As the null distribution is determined by the largest (standardized) value of $T_n(c)$ across scales $c \in \mathcal{C}$, that is, the largest "dip" found across locations and scales, it is possible to assign simultaneously valid p values to individual dips by simply evaluating the test statistic at the location and scale of the individual dip only and then referring its appropriately standardized value to the overall null distribution. Proceeding in this way, one finds the dips around 0.075 mm and 0.09 mm to be significant. The conclusion of three significant "bumps" coincides with the parametric analysis of Izenman and Sommer (1988, sect. 7.2).

Note, however, that any nonparametric procedure for inference about the number of components requires a careful justification of its correctness. For example, identifying the number of modes in the density with the number of components in the mixture is an approach that can go badly wrong: One can easily find examples showing that a mixture of (say) two unimodal densities can have any number of modes (see Walther 2001). The approach taken in this article is based on the concept of log-concavity, which is intimately related to the theory of total positivity. Using the latter theory, it may be possible to develop and justify a procedure for inference about the number of components, for example, along the lines delineated in the last paragraph. That would be another advantage over the modality approach besides the greater sensitivity to detect mixing. Developing such a nonparametric theory beyond the homogeneity problem treated in this article and thus providing a counterpart to the parametric theory developed in Lindsay and Roeder (1997) is an important open problem for future research.

APPENDIX: PROOFS

A.1 Proof of Theorem 1

Let $\phi = \log f_1$ and $M := \max_G e^{\phi}$. For $x_1, x_2 \in G$ and $\alpha \in (0,1)$ set $x_\alpha := \alpha x_1 + (1-\alpha)x_2$ and $\phi_\alpha := \alpha\phi(x_1) + (1-\alpha)\phi(x_2)$. The function $F(t) := e^t - Mt^2/2$ is concave on $(-\infty, \log M]$. Thus

$$\begin{aligned} f_1(x_\alpha) &\geq e^{\phi_\alpha} \qquad \text{as } \phi \text{ is concave} \\ &\geq \alpha F(\phi(x_1)) + (1-\alpha)F(\phi(x_2)) + M(\phi_\alpha)^2/2 \\ &= \alpha f_1(x_1) + (1-\alpha)f_1(x_2) - M\alpha(1-\alpha)(\phi(x_1) - \phi(x_2))^2/2. \end{aligned}$$

By theorem 41D in Roberts and Varberg (1973), ϕ is Lipschitz with some constant L. Hence the last expression in the above display is not smaller than

$$\begin{aligned} &\alpha f_1(x_1) + (1-\alpha)f_1(x_2) - L^2 M\alpha(1-\alpha) \parallel x_1 - x_2 \parallel^2 /2 \\ &\quad = \alpha(f_1(x_1) - L^2 M \parallel x_1 \parallel^2 /2) \\ &\qquad + (1-\alpha)(f_1(x_2) - L^2 M \parallel x_2 \parallel^2 /2) + L^2 M \parallel x_\alpha \parallel^2 /2. \end{aligned}$$

Thus we proved that $f_1(x) = \Psi_1(x) + b_1 \parallel x \parallel^2$ for a concave function Ψ_1 on G and $b_1 \geq 0$. Clearly also $\sum_{i=1}^m p_i f_i(x) = \Psi(x) + b \parallel x \parallel^2$ on G for a concave Ψ and some $b \geq 0$. The proof is complete once we show that $\log(\Psi(x) + b \parallel x \parallel^2) - c \parallel x \parallel^2$ is concave for some $c \geq 0$. The concavity of Ψ implies the existence of a nonnegative $D \leq b(\parallel x \parallel^2)_\alpha - b \parallel x_\alpha \parallel^2$ with $\Psi(x_\alpha) - b \parallel x_\alpha \parallel^2 \geq \Psi_\alpha + b(\parallel x \parallel^2)_\alpha - D \geq \min_G(\Psi + b \parallel x \parallel^2)$, where $(\parallel x \parallel^2)_\alpha := \alpha \parallel x_1 \parallel^2 + (1-\alpha) \parallel x_2 \parallel^2$. Thus

$$\begin{aligned} &\log(\Psi(x_\alpha) + b \parallel x_\alpha \parallel^2) \\ &\quad \geq \log(\Psi_\alpha + b(\parallel x \parallel^2)_\alpha - D) \\ &\quad \geq \log(\Psi_\alpha + b(\parallel x \parallel^2)_\alpha) - D/\min_G(\Psi(x) + b \parallel x \parallel^2) \\ &\quad \geq (\log(\Psi(x) + b \parallel x \parallel^2))_\alpha - c(\parallel x \parallel^2)_\alpha + c \parallel x_\alpha \parallel^2, \end{aligned}$$

where $c := b/\min_G(\Psi(x) + b \parallel x \parallel^2) < \infty$ and the second-to-last inequality follows from the fact that $\log(y - D) \geq \log(y) - D/m$ for $y \geq 0$ and $y - D \geq m > 0$. □

A.2 Proof of Proposition 2

First let f be a log-concave density. Then $\parallel \hat{f}_n^0 - f \parallel_\infty \to 0$ a.s. obtains by following the proof of theorem 3.2 in Groeneboom et al. (in press) (abbreviated GJW in the following). Their use of monotonicity and convexity to obtain various continuity and uniformity results is readily adapted to the use of log-concavity instead. The only additional argument required is to establish that $\hat{f}_n^0$ is uniformly bounded above. This can be deduced from (3.7) in GJW (in press) together with the log-concavity of f. The argument also applies to $\hat{f}_n^c$ (with a fixed c) in place of $\hat{f}_n^0$ by the fact that $\hat{\phi}_n$ in the representation $\hat{f}_n^c(x) = \exp(\hat{\phi}_n(x) + cx^2)$ is concave. Together with the uniform Glivenko–Cantelli theorem 2.8.1 in van der Vaart and Wellner (1996), said arguments yield $\parallel \hat{f}_n^{*c} - \hat{f}_n^0 \parallel \to 0$ in P_n^n-probability a.s., as well as uniform convergence of $\hat{f}_n^0$. Here P_n^n denotes the law of $(X_1^*, \ldots, X_n^*)$, and the X_i are iid from the f given in the statement of the proposition. It is then readily shown that this result in conjunction with the representation $\hat{f}_n^{*c} = \exp(\hat{\phi}_n^*(x) + cx^2)$, $\hat{\phi}_n^*$ concave, implies $T_n^*(c) \to 0$ in P_n^n-probability a.s. Using $T_n^*(c) \leq c|G| \parallel \hat{f}_n^{*c} \parallel_G$, one concludes that $T_n^*(c) \to^{L^2} 0$ a.s., and thus

$$m(c) \to 0, \quad s(c) \to 0 \text{ a.s.} \qquad \text{as } n \to \infty, B \to \infty, \tag{A.1}$$

for every fixed c. Next, let $\bar{c}$ be the smallest element in $\mathcal{C}$ that is greater than or equal to the c given in (1). (If no such element exists, set $\bar{c}$ equal to the largest element in $\mathcal{C}$; the following considerations will then apply, with some extra arguments.) Again, $\parallel \hat{f}_n^{\bar{c}} - f \parallel_\infty \to 0$ a.s. can be established by retracing the proof of theorem 3.2 in GJW (in press) and employing the shape restriction $f(x) = \exp(\phi(x) + cx^2)$, ϕ concave, in place of monotonicity and convexity to obtain the various continuity and uniformity results. Together with $c > 0$ this implies $\liminf_{n\to\infty} T_n(\bar{c}) > 0$ a.s. So with (A.1) we get

$$\begin{aligned} &\left| \left\{ \max_{c\in\mathcal{C}} \frac{T_n(c) - m(c)}{s(c)} \leq \max_{c\in\mathcal{C}} \frac{T_n^{*i}(c) - m(c)}{s(c)}, 1 \leq i \leq B \right\} \right| \\ &\quad \leq |\mathcal{C}| \max_{c\in\mathcal{C}} \left| \left\{ \frac{T_n(\bar{c}) - m(\bar{c})}{s(\bar{c})} \leq \frac{T_n^{*i}(c) - m(c)}{s(c)}, 1 \leq i \leq B \right\} \right| \end{aligned}$$

$$\le |\mathcal{C}| \left(\frac{s(\bar{c})}{T_n(\bar{c}) - m(\bar{c})} \right)^2 B \qquad \text{by Chebyshev's inequality}$$
$$= o(B) \text{ a.s.} \qquad \text{as } n \to \infty, B \to \infty.$$

The assertion follows. □

A.3 Proof of Proposition 1

We will first prove that if the MLE $\hat{\phi}_n$ exists, it is the solution of (2). Define the functionals $M_0(\phi) := \frac{1}{n}\sum_{i=1}^n \phi(X_i)$ and $M(\phi) := \frac{1}{n}\sum_{i=1}^n \phi(X_i) - \int \exp(\phi(x) + cx^2)\,dx$. For a given function ϕ with $0 < \int \exp(\phi(x) + cx^2)\,dx < \infty$, set $\phi^* = \phi - \log \int \exp(\phi(x) + cx^2)\,dx$. Then $M(\phi^*) = \frac{1}{n}\sum_{i=1}^n \phi(X_i) - \log \int \exp(\phi(x) + cx^2)\,dx - 1 = M(\phi) + \int \exp(\phi(x) + cx^2)\,dx - \log \int \exp(\phi(x) + cx^2)\,dx - 1$. Using the fact that $t - \log t - 1 \ge 0$ for $t > 0$, with equality only if $t = 1$, one sees that $M(\phi^*) \ge M(\phi)$, with equality only if $\int \exp(\phi(x) + cx^2)\,dx = 1$. This shows that the concave function ϕ maximizes M if and only if ϕ maximizes M under the constraint $\int \exp(\phi(x) + cx^2)\,dx = 1$. But under the latter constraint, we have $M(\phi) = M_0(\phi) + 1$, and the concave function that maximizes M_0 under said constraint is by definition the MLE $\hat{\phi}_n$. This argument follows that of Silverman (1982).

Proving existence and the stated properties of the MLE can be done in a way similar to the proof of Robertson et al. (1988, p. 326). The idea is that if the concave function $\hat{\phi}_n$ were not linear between the observations X_i, then the function $\hat{\hat{\phi}}_n$ obtained by linearly interpolating the $\hat{\phi}_n(X_i)$ would still be concave and satisfy $\hat{\hat{\phi}}_n \le \hat{\phi}_n$, with strict inequality on a set of positive Lebesgue measure. Thus $\sum_i \hat{\phi}_n(X_i) = \sum_i \hat{\hat{\phi}}_n(X_i)$, whereas the integral in (2) would be smaller for $\hat{\hat{\phi}}_n$ than for $\hat{\phi}_n$, contradicting the fact that $\hat{\phi}_n$ is a solution to (2). The assertion about the sequence of slopes being contained in the cone C_n is an immediate consequence of the concavity of $\hat{\phi}_n$. □

The gradient and the diagonal of the Hessian matrix of the criterion function $\psi(\mathbf{s})$ to be minimized in (5) are as follows:

$$\frac{\partial \psi(\mathbf{s})}{\partial \mathbf{s}_1} = -(x_2 - x_1)\sum_{j=2}^{n+1} w_j + \frac{1}{2}\left(\sum_{i=3}^{n} e^{l_i}(x_{i+1} - x_{i-1}) + e^{l_2}(x_3 - x_2) + e^{l_{n+1}}(x_{n+1} - x_n) \right)(x_2 - x_1)$$

$$\frac{\partial \psi(\mathbf{s})}{\partial \mathbf{s}_k} = -(x_{k+1} - x_k)\sum_{j=k+1}^{n+1} w_j + \frac{1}{2}\left(\sum_{i=k+1}^{n} e^{l_i}(x_{i+1} - x_{i-1}) + e^{l_{n+1}}(x_{n+1} - x_n) \right)(x_{k+1} - x_k), \qquad k \ge 2$$

$$\frac{\partial^2 \psi(\mathbf{s})}{\partial \mathbf{s}_1^2} = \frac{1}{2}\left(\sum_{i=3}^{n} e^{l_i}(x_{i+1} - x_{i-1}) + e^{l_2}(x_3 - x_2) + e^{l_{n+1}}(x_{n+1} - x_n) \right)(x_2 - x_1)^2$$

$$\frac{\partial^2 \psi(\mathbf{s})}{\partial \mathbf{s}_k^2} = \frac{1}{2}\left(\sum_{i=k+1}^{n} e^{l_i}(x_{i+1} - x_{i-1}) + e^{l_{n+1}}(x_{n+1} - x_n) \right) \times (x_{k+1} - x_k)^2, \qquad k \ge 2.$$

[Received June 2000. Revised January 2001.]

REFERENCES

Brooks, S. P. (1998), "MCMC Convergence Diagnosis via Multivariate Bounds on Log-concave Densities," *Annals of Statistics,* 26, 398–433.

Côté, P., Welch, D. L., and Fischer, P. (1993), "The Detection of an Extended Moving Group Near the Galactic Disk," *Astrophysical Journal Letters*, 406, L59–L62.

Dellaportas, P., and Smith, A. F. M. (1993), "Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling," *Journal of the Royal Statistical Society, C* 42, 443–460.

Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Journal of the Royal Statistical Society, C* 41, 337–348.

Groeneboom, P., and Wellner, J. A. (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Basel: Birkhäuser.

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (in press), "Estimation of a Convex Function: Characterizations and Asymptotic Theory," *Annals of Statistics.*

Izenman, A. J., and Sommer, C. J. (1988), "Philatelic Mixtures and Multimodal Densities," *Journal of the American Statistical Association*, 83, 941–953.

Jongbloed, G. (1998), "The Iterative Convex Minorant Algorithm for Nonparametric Estimation," *Journal of Computational and Graphical Statistics*, 7, 310–321.

Liero, H., Läuter, H., and Konakov, V. (1998), "Nonparametric versus Parametric Goodness of Fit," *Statistics*, 31, 115–149.

Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*, Hayward, CA: Institute of Mathematical Statistics.

Lindsay, B. G., and Roeder, K. (1992), "Residual Diagnostics for Mixture Models," *Journal of the American Statistical Association*, 87, 785–794.

——— (1997), "Moment-Based Oscillation Properties of Mixture Models," *Annals of Statistics*, 25, 378–386.

Roberts, A. W., and Varberg, D. E. (1973), *Convex Functions*, New York and London: Academic Press.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference*, New York: Wiley.

Roeder, K. (1994), "A Graphical Technique for Determining the Number of Components in a Mixture of Normals," *Journal of the American Statistical Association*, 89, 487–495.

Schork, N. J., Weder, A. B., and Schork, A. (1990), "On the Asymmetry of Biological Frequency Distributions," *Genetic Epidemiology*, 7, 427–446.

Shaked, M. (1980), "On Mixtures from Exponential Families," *Journal of the Royal Statistical Society*, 42, 192–198.

Silverman, B. W. (1981), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society*, 43, 97–99.

——— (1982), "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method," *Annals of Statistics*, 10, 795–810.

Swales, J. D. (ed.) (1985), *Platt vs. Pickering: An Episode in Recent Medical History*, Cambridge, UK: The Keynes Press.

Terlaky, T., and Vial, J. Ph. (1998), "Computing Maximum Likelihood Estimators of Convex Density Functions," *SIAM Journal on Scientific Computing*, 19, 675–694.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Berlin and New York: Springer-Verlag.

Walther, G. (in press), "Multiscale maximum likelihood analysis of a semiparametric model, with applications," *Annals of Statistics.*

——— (2001), "Kernel Oscillation Analysis for the Mixture Complexity," manuscript in preparation.

Wand, M. P. (1994), "Fast computation of multivariate kernel estimators," *Journal of Computational and Graphical Statistics*, 3, 433–445.