Industrial Mathematics and Statistics
Edited by J.C. Misra

18. Bikernel Mixture Analysis

Guenther Walther

Stanford University

Abstract

A challenging problem in the analysis of mixtures is the determination of the number of components in the mixture, in the case where one does not want to make parametric assumptions on the component distributions. It is shown how convolutions with the Gauss kernel and its derivative kernel allow to set lower confidence bounds on the number of components in a location mixture. The resulting procedure has the advantage over mode-hunting approaches that it is sensitive to detect mixing in more general unimodal situations, and at the same time it cannot be improved upon in the more restricted situation where mixing manifests itself in multimodality, even if one is allowed to use that knowledge a priori. This is explained heuristically and made precise in the asymptotic minimax framework.

Keywords: mixture complexity, upcrossings, modality, minimax.

1 Introduction

One important issue in the analysis of mixture data is the inference on the number of components in the mixture. Some well-known examples in the statistical and scientific literature are the number of kinds of chondrite in meteorites ([18, 12]), the number of different paper types used in the production of certain stamps ([15]), the number of groups of stars in certain locations in space ([6]), or the number of genetic components determining blood pressure ([10, 25]). The statistical framework for such a cluster analysis is a mixture model. This article focuses on the important case of a location mixture

$$f(x) = \sum_{i=1}^{k} p_i\, g(x - t_i), \quad g \in \mathcal{S}, \tag{1}$$

where the nonnegative weights sum to unity and $\mathcal{S}$ denotes a set of univariate single-component distributions. In this model the weights p_i, the location parameters t_i, and the single-component density g are unknown. The object of the inference is the *mixture complexity* $\inf\{k : f(x) = \sum_{i=1}^{k} p_i g(x - t_i), g \in \mathcal{S}\}$. Scientific and statistical interest focuses on lower confidence bounds for the mixture complexity; it is well known that no nontrivial upper confidence bounds exist, see [9].

Powerful theoretical results are available in the case where the component class $\mathcal{S}$ is parametric. [28, 20, 21, 25] exploit structural properties of exponential families, [19] and [8] derive procedures based on properties of moment matrices. There is a large literature on an approach where $\mathcal{S}$ is a nonparametric class of distributions, motivated by the fact that the parametric procedures are quite sensitive to the structure imposed on $\mathcal{S}$. For example, if $\mathcal{S}$ is taken to be the normal family in the analysis but the real g is skewed, then many normal components are required in the mixture to pick up the skewness, which can result in a considerable overestimate of the mixture complexity. See [25, p.493] and the references given there for a further discussion of this issue. The nonparametric approaches usually proceed by mode- or bump-hunting, i.e. by establishing a lower confidence bound on the number of modes of f or of 'bumps' (maxima of the density derivative, see [7, 12]). Various techniques have been developed to assess the modality of f, see e.g. [29, 14, 24, 23]; [4, 5] give further refinements for several of these procedures. One disadvantage of such an approach is that it is not very sensitive to detect mixing. For example, the means of two homoscedastic normal distributions need to be separated by at least two standard deviations before any mixture becomes bimodal. Another problem, which is apparently not widely recognized, is that counting modes and 'bumps' does not always lead to a valid inference: Figure 1 shows a unimodal density f (solid line) and the mixture $1/2(f(x) + f(x-4))$, which is trimodal.

In fact, it is not hard to modify f in figure 1 such that the two-component mixture possesses any prescribed number m of modes. Thus if the sample size is large enough, then mode-hunting will give a lower bound of 3 (or even m) for the mixture complexity with arbitrary high confidence level, while for unimodal components the mixture complexity

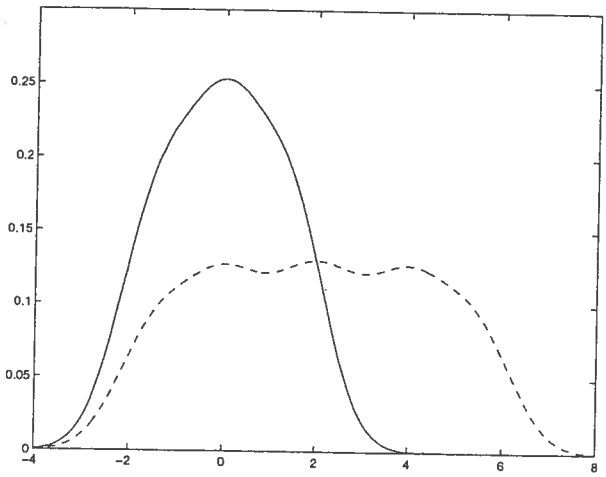

Figure 1: A mixture of two unimodal distributions which is trimodal.

is only two. This is a most serious scientific and statistical mistake. See [33] for an investigation of mode- and bump-hunting in this context.

In section 2 it is shown how convolutions with the Gaussian kernel and its derivative kernel allow valid inference on the mixture complexity, and the accompanying statistical theory is presented. Section 3 investigates how the new procedure compares vis-a-vis the more conservative mode-hunting approach in a setting where multimodality is in fact present. The main result is that the new nonparametric procedure dominates any mode-hunting technique: The new procedure is sensitive to detect mixing in a more general setting where the mixture is unimodal, yet it cannot be improved upon in the more restrictive setting where the mixture is multimodal, even if one is allowed to use that knowledge. This is first made plausible heuristically, and then made rigorous in the modern asymptotic minimax framework.

Examples are given intermittently throughout the exposition to illustrate various points. Some proofs are deferred to section 4.

2 Upcrossings for mixtures

A univariate density g is said to be of type PF_n, $n \geq 1$, if for all $x_1 < \ldots < x_n$,

$y_1 < \ldots < y_n$,

$$\det \| g(x_i - y_j) \|_{i,j=1}^{n} \geq 0. \tag{2}$$

g is called a Polya frequency density ($g \in PF_\infty$) if (2) holds for every positive integer n, see [16]. The nonparametric class PF_∞ contains many standard unimodal parametric densities, such as the normal family, but also skewed distributions such as the gamma family with integer shape parameter. PF_∞ is thus a flexible nonparametric model for single-component distributions. Another reason why it is appealing to use PF_∞ is that this is precisely the model under which mode- and bump-hunting will work, see [33]. Thus the literature on mode- and bump-hunting has implicitly used and approved PF_∞ as an appropriate nonparametric model for single-component distributions, and it is thus fitting to pursue the new developments below for the same model.

It is informative to note that PF_∞ is a strict subset of the class of unimodal distributions; for example, the unimodal f in figure 1 is not in PF_∞. The PF_1 condition stipulates that g is nonnegative, which is a vacuous condition for probability densities. The PF_2 condition is equivalent to g being logarithmically concave, see 18.A in [22]. Using this latter condition, procedures for testing the important homogeneity case $k = 1$ were developed in [32, 34] using a maximum likelihood technique. It was also shown there how this model extends to a multivariate situation. The following approach is restricted to a univariate set-up, but in turn allows to set lower confidence bounds for any mixture complexity $k \geq 1$.

The theory of Polya frequency functions was instigated by the qualitative study of the number of sign changes of certain transformations, such as convolutions, see [27, 16], and has turned out to be of fundamental importance in many areas of mathematics and the sciences. The key idea of this paper is that a simple but powerful approach to the mixture complexity problem can be developed by studying such a qualitative behavior for the *sum of two appropriate convolutions*. Let f be a univariate density, and let φ denote the standard normal density. Recall the usual notation $k_h(\cdot) := k(\cdot/h)/h$, in particular we

will write $\varphi_h'(\cdot) := \varphi'(\cdot/h)/h = -(\cdot/h^2)\varphi(\cdot/h)$. Consider the function

$$\bar{s}_{h,\lambda}(\cdot) := \Big((1-|\lambda|)\,(\varphi_h' \star f)\,(\cdot) + \lambda\,(\varphi_h \star f)\,(\cdot)\Big)/\sigma_\lambda, \tag{3}$$

where $h > 0, \lambda \in (-1,1)$, and $\sigma_\lambda := \|(1-|\lambda|)\varphi_h' + \lambda\varphi\|_2 = \sqrt{(1-|\lambda|)^2 + 2\lambda^2}/(2\pi^{1/4})$ is a standardizing constant. To understand the behavior of this function, set for a moment $\lambda = 0$ and $f = \varphi$. Then up to a scale factor, $\bar{s}_{h,\lambda}$ equals the derivative of a Gaussian density, which is depicted in figure 2. A key point is that if $f \in PF_2$, then no matter what $h > 0$ or $\lambda \in (-1,1)$ are, $\bar{s}_{h,\lambda}$ will have the same *qualitative* behavior as the derivative of the Gaussian, in that there will be no upcrossing through zero. See figure 2 for the case where f =gamma(2,1). Conversely, if this qualitative behavior holds for all $h > 0, \lambda \in (-1,1)$, then $f \in PF_2$; see theorem 4 in section 4.

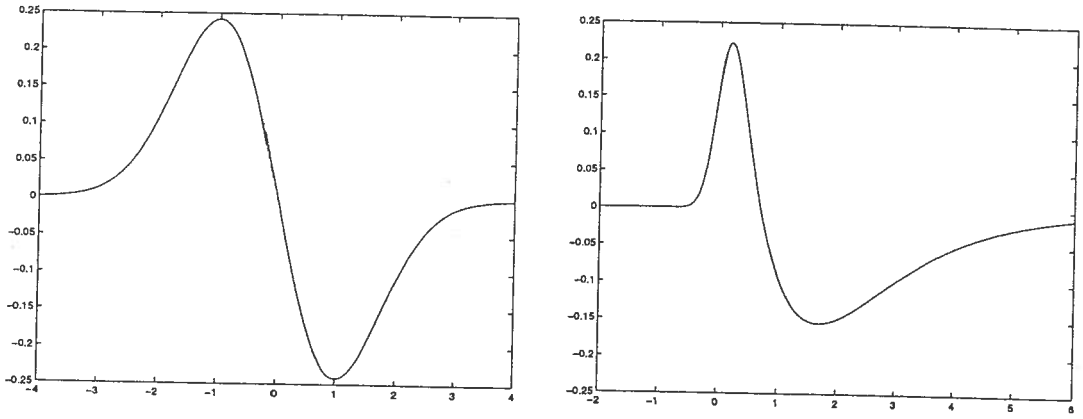

Figure 2: Left: The derivative of the standard normal density. Right: $\bar{s}_{0.2,-0.1}$ for the gamma(2,1) density.

Thus an upcrossing through zero implies mixing. Most importantly, counting upcrossings and adding 1 yields a quantity that never exceeds the mixture complexity and thus can be used to construct valid lower confidence bounds for the mixture complexity:

Theorem 1 *If f is given by the mixture (1) with $\mathcal{S} = PF_\infty$, then $\bar{s}_{h,\lambda}(\cdot)$ has at most $k-1$ upcrossings through zero.*

Proof of theorem 1: Assume to the contrary that $\bar{s}_{h,\lambda}(\cdot)$ has k upcrossings for some

$\lambda \in (-1,1), h > 0$. Then for $b > 0$ small enough, the same is true for the function

$$\begin{aligned} s(x) &:= (1-|\lambda|)\int \frac{\varphi_h(x+b-t)-\varphi_h(x-t)}{b/h} f(t)dt + \lambda \int \varphi_h(x-t)f(t)dt \\ &\overset{\psi:=\varphi_h\star g}{=} (1-|\lambda|)h/b \sum_{j=1}^{k} p_j\Big(\psi(x+b-t_j)-\psi(x-t_j)\Big) + \lambda \sum_{j=1}^{k} p_j \psi(x-t_j) \\ &= \sum_{j=1}^{2k} q_j \psi(x-y_j), \end{aligned} \tag{4}$$

where $y_{2j-1} = t_j - b$, $y_{2j} = t_j$, $q_{2j-1} = (1-|\lambda|)p_j h/b$, $q_{2j} = \lambda p_j - (1-|\lambda|)p_j h/b$, $j = 1,\ldots,k$. Hence there exist $x_1,\ldots,x_{2k+1}$ such that the $s(x_i)$ have alternating sign. W.l.o.g. we may assume that the sequence $\{t_i\}$ is strictly increasing, and by taking b small enough the same holds for $\{y_i\}$. (4) shows that the vector $\{s(x_i), 1 \le i \le 2k+1\}$ is a linear combination of the vectors $\{\psi(x_i - y_j), 1 \le i \le 2k+1\}$, $1 \le j \le 2k$, whence

$$\begin{aligned} 0 &= \begin{vmatrix} \psi(x_1-y_1) & \ldots & \psi(x_1-y_{2k}) & s(x_1) \\ \vdots & & & \vdots \\ \psi(x_{2k+1}-y_1) & \ldots & \psi(x_{2k+1}-y_{2k}) & s(x_{2k+1}) \end{vmatrix} \\ &= \sum_{i=1}^{2k+1} s(x_i)(-1)^{2k+1+i} \det \| \psi(x_l - y_m) \|_{l\neq i, m\neq 2k+1}. \end{aligned}$$

ψ is STP_∞ by theorem 3.3 in [2], i.e. satisfies (2) with strict inequality. Thus all the determinants in the last sum are positive, and the $s(x_i)(-1)^{2k+1+i}$are either all positive or all negative. Hence the sum cannot be zero and the desired contradiction obtains. □

The statistical approach to the problem is now evident. Given $X_1,\ldots,X_n$ iid from f, one convolves the kernels with the empirical measure instead of f to obtain the usual kernel estimates:

$$\hat{s}_{h,\lambda}(\cdot) := \Big((1-|\lambda|)\, h\hat{f}_h{}'(\cdot) + \lambda\, \hat{f}_h(\cdot)\Big)/\sigma_\lambda,$$

where $\hat{f}_h(x) := \frac{1}{nh}\sum_{i=1}^n \varphi\Big(\frac{x-X_i}{h}\Big)$ and $\hat{f}_h{}'(x) := \frac{1}{nh^2}\sum_{i=1}^n \varphi'\Big(\frac{x-X_i}{h}\Big)$. The recipe provided by theorem 1 is to look for upcrossings while varying h and λ. This corresponds to the paradigm introduced from computer vision into the statistics literature in [3]. It is argued

there that statistical interest often focuses on certain features of a function f having an unknown scale h, and that these features are hence best extracted by considering (say) kernel estimates $\hat{f}_h$ over a range of h simultaneously. Note that this approach has a natural and rigorous justification for the problem at hand here. In particular, the kernel estimate is not just any more a device to look at f at various scales of resolution, but the problem is characterized in terms of convolutions with the Gauss kernel and its derivative in the first place.

[11] give results in the Gaussian white noise model for the simultaneous behavior of kernel estimates over a range of bandwidths. By approximating $\hat{s}_{h,\lambda}(\cdot)$ by an appropriate Gaussian process, these results can be used to derive the desired confidence bounds for the number of upcrossings. The next theorem takes the range of bandwidths H_n to be a subset of $[n^{-1+\epsilon}, const]$, $\epsilon > 0$. Set $U_{n,h}(x) := \left(q_\alpha + \sqrt{2(\log \frac{1}{h\hat{f}_h(x)})^{\div}}\right)\sqrt{\frac{\hat{f}_h(x)}{nh}}$, and $L_{n,h}(x) := -U_{n,h}(x)$, where q_α is given below.

Theorem 2 *If f is given by (1), then*

$$\overline{\lim_{n\to\infty}} P_f\Big(\hat{s}_{h,\lambda}(\cdot) \text{ has more than } k-1 \text{ upcrossings of } [L_{n,h}(\cdot), U_{n,h}(\cdot)] \\ \text{for some } h \in H_n, \lambda \in (-1,1)\Big) \leq \alpha.$$

Thus a lower $1-\alpha$ confidence bound for the mixture complexity obtains by counting upcrossings of $\hat{s}_{h,\lambda}(\cdot)$ through the pair of lower and upper bounds $L_{n,h}(\cdot)$ and $U_{n,h}(\cdot)$, taking the largest number found this way while varying h and λ, and adding 1. Note that the $1-\alpha$ confidence level is simultaneous for all the components found. This aspect is known to be a perennial problem for mode-hunting procedures, or even parametric procedures such likelihood ratio tests, where the presence of components is tested sequentially one at a time with an unknown overall confidence level as a result. The quantile q_α in the definition of $U_{n,h}$ is given in the proof of theorem 2, but alternatively one can use $q_{n,\alpha}$ obtained by simulating the procedure with Monte Carlo samples of size n from the uniform distribution in place of f. Note that by employing a null distribution that is based on a range of bandwidths h, one avoids working with the extreme value distribution that

arises by using one bandwidth sequence under the sup-norm, see [1]. The convergence to the latter extreme value distribution is known to be extremely slow, see [13]. Using the analysis of this latter paper, one can show that the procedure introduced here is much more accurate even for small sample sizes.

If the kernel estimates are computed via the FFT as usual, then it is simple and fast to compute these estimates over a grid of bandwidths. In all of the following examples a grid H_n was used that consists of 50 equally spaced values from $SD(\underline{X})/\sqrt{n}$ to a quarter of the range of the data, and likewise a grid of 25 values was used for the λ.

An example is given in figure . 300 observations were sampled from the mixture $f(x) = 0.2\text{gamma}(x-5,2,1) + 0.4\text{norm}(x,2,1) + 0.4\text{unif}(x,-5,3)$. The purpose of the example is also to demonstrate the method when model 1 does not hold. The mixture has only two modes, but the procedure detects the three components: Figure shows two upcrossings of $\hat{s}_{0.3,-0.07}$ through the 10% band.

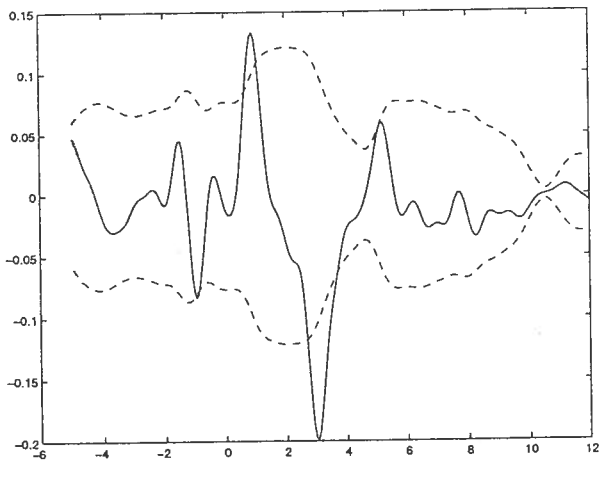

Figure 3: Two upcrossing of $\hat{s}_{0.3,-0.07}$ indicate the presence of three components at the 10% level.

Remarks:

1. It is pleasing to note the similarities to the approach in the exponential family case which relies on counting certain sign changes, see [25, 20, 21], even though the nonparametric method here is completely different from the parametric one. On the other hand,

the special case $\lambda = 0$ reduces to a procedure that counts modes, a well-known approach in the nonparametric set-up.

2. It is possible to refine the analysis and check the PF_n condition for $n > 2$. However, it remains to be seen how much would be gained. The PF_2 condition is equivalent to f being logarithmically concave, i.e. $f(x) = e^{\psi(x)}$ for a concave function ψ. The prime example is of course the normal family, where ψ is a quadratic. Thus this condition captures the appearance of a bell-shaped curve that is commonly associated with a single-component distribution. Further one deduces from this equivalence that the above upcrossing procedure looks for 'bumps' in the *log-density*. This fact establishes yet another pleasing connection to other nonparametric theory: It is a widely accepted principle that the inference should focus on the log-density, not the density itself, see e.g. [30] or [31].

3. It is essential that the Gauss kernel be used for the convolution, or at least a kernel that is in PF_∞, if the interest lies in the *number* of features in the data. Otherwise this number can be spuriously inflated by the convolution. This follows from a well-known result from the theory of total positivity, see [16].

3 Comparison with the modality approach

The above procedure allows to detect mixing even in situations where the mixture is unimodal. An important question is how this procedure compares to other modality approaches in a situation where the mixture *is* indeed multimodal. It would seem that in the more restricted case where multimodality is present, a procedure that is specially tailored to detect multimodality is more powerful to detect mixing. The reason is that such a procedure is allowed to concentrate its power solely on finding multimodality, rather than having to be sensitive to detect a larger class of alternatives.

The following heuristic suggests that this is in fact not so. Consider for simplicity distributions on the unit interval. The least favorable distribution for detecting multimodality is the uniform distribution: At each point in $[0, 1]$ the derivative is zero and hence on the boundary to bimodality. So a procedure must guard against erroneously

declaring bimodality by guarding against a sign change of the derivative from - to + simultaneously on $[0,1]$. On the other hand, consider the function $\bar{s}_{h,\lambda}$ in (3). $\varphi_h' \star f$ gives in essence the derivative of f, and $\varphi_h \star f$ is positive. Hence varying λ corresponds to increasing or decreasing the derivative term $\varphi_h' \star f$. Thus the least favorable situation is obtained by adjusting λ such that the resulting sum equals zero on $[0,1]$, because then the procedure has to guard against erroneous sign changes everywhere on $[0,1]$. But this worst-case situation is essentially the same as the one previously encountered for the modality approach. Hence detecting mixing in the more general context is not a harder problem than detecting mixing in the narrower context of multimodality.

This can be made precise in the asymptotic minimax framework. We measure the difficulty of detecting bimodality of f by how far f drops down between the modes, where the decrease and increase each have to occur in an interval of length l. More precisely, define the $l-drop\ of\ f := \inf\{d : \exists\, x_1 < x_2 < y_1 < y_2$ with $|x_2 - x_1| \le l$, $|y_2 - y_1| \le l$ and $f(x_1) \ge f(x_2) + d$, $f(y_2) \ge f(y_1) + d\}$. The difficulty of detecting such a drop depends on the smoothness of f. But we do not wish to make any smoothness assumptions on f, and hence consider the l-drop of $\bar{f}_h = \dot{\varphi}_h \star f$ instead, where $h = h(n) \downarrow 0$ with $h(n) \ge n^{-1+\epsilon}$. Consider within the class of densities supported on $[0,1]$ and uniformly bounded above and below away from the zero, the set $\mathcal{F}_n(d) := \{f :\ l\text{-drop of } \sqrt{\bar{f}_h} \ge dl\sqrt{\frac{(\log(1/h))^+}{8\sqrt{\pi}nh^3}}$ for some $l\}$. Part 1 of the following theorem shows that if the drop is below a certain threshold, then no procedure can detect it asymptotically. Part 2 shows that if the drop is above this threshold, then the above upcrossing procedure will detect it with power 1 asymptotically:

Theorem 3 *(1) Let $\Psi_n(\underline{X}_n)$ be any sequence of procedures for detecting bimodality at level $\alpha_n \to \alpha \in (0,1)$ that are based on an iid sample $\underline{X}_n = (X_1, \ldots, X_n)$ from f. If $d < 1$, then*

$$\overline{\lim}_{n\to\infty} \inf_{f \in \mathcal{F}_n(d)} P_f\Big(\Psi_n(\underline{X}_n) \text{ detects bimodality } \Big) \le \alpha$$

(2) If $d > 1$, then

$$\underline{\lim}_{n\to\infty} \inf_{f\in\mathcal{F}_n(d)} P_f\Big(\textit{above upcrossing procedure detects bimodality}\Big) = 1.$$

Hence there is a *sharp cut-off*: Below the cut-off, no procedure can detect bimodality. Above the cut-off, even the *more general* upcrossing procedure detects it with power one. Thus this sharp cut-off leaves no room for any specially tailored modality procedure to outperform the more general upcrossing procedure.

The conclusion of the theorem will now be demonstrated with a data example. There are two points in the result that need to be evaluated against a pertinent example: First, the minimax results are asymptotic. Second, while the upcrossing procedure attempts to be sensitive to small-scale features as well as large-scale features by considering various bandwidths simultaneously, the statement of theorem is most informative for small l, i.e. localized modes. The faculty quality data considered in [14] give assessments of 63 statistics departments. This sample size is already quite small for a nonparametric approach. Furthermore, the histogram in figure 4 shows that the purported bimodality is almost a global feature, in that it extends almost over the whole range of the data. Thus this example appears well suited for a procedure that is most sensitive to detecting global features, such as the dip-test considered in [14], which uses a distance between distribution functions.

The upcrossing procedure is significant at the 9%level for ascertaining that there are at least two components present. This level is in fact obtained for $\lambda = 0$, i.e. the special case where the procedure looks for bimodality. The corresponding plot of $\hat{s}_{h,\lambda}$ is given in figure 4. [14] quote a significance level of 10% for the dip test.

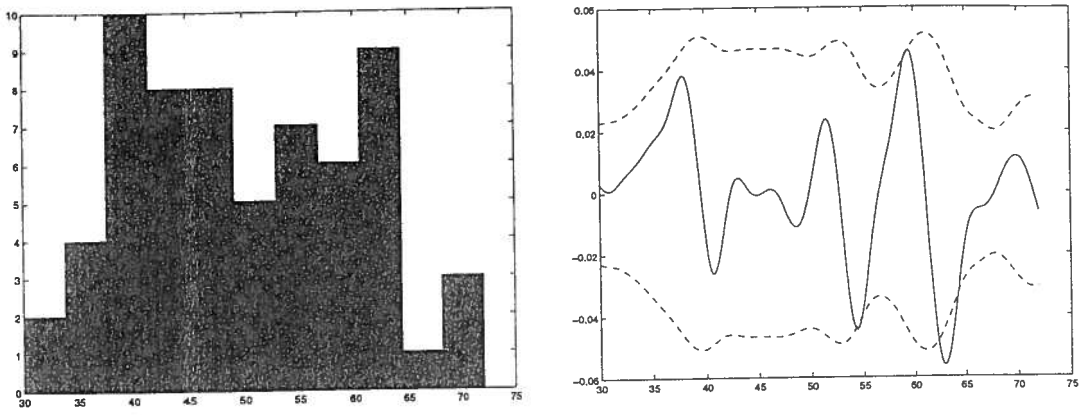

Figure 4: Left: The histogram of the faculty quality data. Right: $\hat{s}_{1.463,0}$ and the 10% bands.

4 Appendix

Theorem 4 *If $f \in PF_2$ and $\lambda \in (-1,1), h > 0$, then $\bar{s}_{h,\lambda}(\cdot)$ has at most one zero, which must be simple. Moreover, the only possible sign-change can only be a downcrossing through 0. Conversely, if $\bar{s}_{h,\lambda}(\cdot)$ has no upcrossing through 0 for all $\lambda \in (-1,1)$ and $h > 0$, then (a version of) f is in PF_2.*

Proof of theorem 4: Let $f \in PF_2$. As φ is STP_∞, i.e. satisfies (2) with strict inequality, it follows from theorem 3.3 in [2] that $\bar{f}_h$ is STP_2 and hence strictly log-concave, cf. 18.A.10 in [22]. As $\bar{f}_h$ is in C^∞, this implies that $(\log \bar{f}_h)'$ is strictly decreasing, and hence $(1-|\lambda|)h(\log \bar{f}_h)' + \lambda$ has at most one zero, which must be simple, and no upcrossing through 0. Writing $(\log \bar{f}_h)' = \bar{f}_h'/\bar{f}_h$ one sees that this is then also true for $\bar{f}_h[(1-|\lambda|)h(\log \bar{f}_h)' + \lambda] = \sigma_\lambda \bar{s}_{h,\lambda}$, as $\bar{f}_h$ is positive on $\mathbf{R}$.

Conversely, the last equation shows that if $\bar{s}_{h,\lambda}$ has no upcrossing through 0, then $(\log \bar{f}_h)'$ has no upcrossing through $\frac{-\lambda}{h(1-|\lambda|)}$. As this holds for all $\lambda \in (-1,1)$, $(\log \bar{f}_h)'$ is nonincreasing and hence $\bar{f}_h$ is log-concave. As the latter property is true for all $h > 0$, it follows from standard arguments that a version of f is log-concave and hence in PF_2. □

Proof of theorem (2): First consider the case $k = 1$. Let $[a,b]$ be an interval

contained in the interior of the support of f. Employing the strong approximation of [17], integrating by parts as in [26, p.31] and rearranging terms shows

$$\| \sqrt{\frac{nh}{\hat{f}_h}}\Big(\hat{s}_{h,\lambda} - \bar{s}_{h,\lambda}(1 - \frac{Z(1)}{\sqrt{n}})\Big) - \frac{1}{\sigma_\lambda \sqrt{h\bar{f}_h}} \int \Big((1-|\lambda|)\varphi' + \lambda\varphi\Big)\Big(\frac{x-t}{h}\Big) dZ(F(t)) \|_{[a,b]} = O\Big(\frac{\log n}{\sqrt{nh}}\Big) \ a.s., \quad (5)$$

uniformly in λ, where Z is a two-sided Brownian motion. Employing the LIL for Brownian motion as in [26, p.32] (see also [1, p.1075]) and using the exponential decay of $(1-|\lambda|)\varphi' + \lambda\varphi$ yields

$$\begin{aligned} &\frac{1}{\sigma_\lambda \sqrt{h\bar{f}_h}} \int \Big((1-|\lambda|)\varphi' + \lambda\varphi\Big)\Big(\frac{x-t}{h}\Big) dZ(F(t)) \\ &\quad - \frac{1}{\sigma_\lambda \sqrt{h}} \int \Big((1-|\lambda|)\varphi' + \lambda\varphi\Big)\Big(\frac{x-t}{h}\Big) dZ(t) \ = \ \sqrt{h} r(x), \end{aligned} \quad (6)$$

where the process $r(x)$ is a.s. bounded. It will be seen that the strong variation diminishing property of the Gauss kernel (cf. the result on the simple zero in theorem 4) allows to restrict attention to small bandwidths h, where the heteroscedasticity effect $\sqrt{h} r(x)$ is negligible. Let q_α be a positive number to be determined later. Abbreviate $\underline{h}_n := \min H_n, \bar{h}_n := \max H_n$ and $S^n_{h,\lambda} := \bar{s}_{h,\lambda}(1 - \frac{Z(1)}{\sqrt{n}})\sqrt{\frac{nh}{\hat{f}_h}}$. Fix $h_0 > 0$. Then by Fatou's lemma and (5,6):

$$\begin{aligned} &\overline{\lim_{n\to\infty}} P_f\Big(\hat{s}_{h,\lambda}(\cdot) \text{ has an upcrossing of } [L_{n,h}(\cdot), U_{n,h}(\cdot)] \text{ for some } h \in H_n, \lambda \in (-1,1)\Big) \\ \leq \ & P_f\Big(\text{for infinitely many } n \ \exists h \in [\underline{h}_n, h_0], |\lambda| < 1 : \\ & \quad \frac{1}{\sigma_\lambda \sqrt{h}} \int \Big((1-|\lambda|)\varphi' + +\lambda\varphi\Big)\Big(\frac{x-t}{h}\Big) dZ(t) + \sqrt{h} r(x) + S^n_{h,\lambda} \\ & \quad \text{has an upcrossing of } \Big[\mp\Big(q_\alpha + \sqrt{2\Big(\log \frac{1}{h\hat{f}_h}\Big)^+}\Big)\Big]\Big) \end{aligned} \quad (7)$$

$$\begin{aligned} + \ & P_f\Big(\text{for infinitely many } n \ \exists h \in [h_0, \bar{h}_n], |\lambda| < 1 : \\ & \quad \frac{1}{\sigma_\lambda \sqrt{h\bar{f}_h}} \int \Big((1-|\lambda|)\varphi' + \lambda\varphi\Big)\Big(\frac{x-t}{h}\Big) dZ(F(t)) + S^n_{h,\lambda} \end{aligned} \quad (8)$$

has an upcrossing of $\big[\mp\big(q_\alpha + \sqrt{2\big(\log\frac{1}{h\hat{f}_h}\big)^+}\big)\big]\big)$.

We first consider (8). If $h > 0$, then by theorem 4, $\bar{s}_{h,\lambda}$ has at most one zero $x_{h,\lambda}$, which must be simple. Standard arguments show that $\bar{s}'_{h,\lambda}(x_{h,\lambda})$ is bounded away from zero for (h,λ) in the compact set $[h_0,\bar{h}_n]\times[-1,1]$. Hence for some $C_1, C_2 > 0$ we have $|S^n_{h,\lambda}(x)| \geq C_1\sqrt{nh}(|x - x_{h,\lambda}| \wedge C_2)$ for n large enough. Abbreviating $I(h,\lambda,x) := \frac{1}{\sigma_\lambda\sqrt{h\bar{f}_h}}\int\big((1 - |\lambda|)\varphi' + \lambda\varphi\big)\big(\frac{x-t}{h}\big)dZ(F(t))$ and using the fact that $S^n_{h,\lambda}$ has no upcrossing by theorem 4, it follows that (8) is not larger than

$$P\Big(\sup_{h\in[h_0,\bar{h}_n],|\lambda|\leq 1}\ \sup_{|x-y|\leq\log n/\sqrt{nh}}|I(h,\lambda,x) - I(h,\lambda,y)| > q_\alpha\ i.o.\Big)$$
$$+\ P\Big(\sup_{h\in[h_0,\bar{h}_n],|\lambda|\leq 1}\ \sup_{x\in[a,b]}|I(h,\lambda,x)| > \frac{C_1}{2}\log n\ i.o.\Big),$$

where the first probability corresponds to the case where start x and endpoint y of the posited upcrossing fall in a $\log n/2\sqrt{nh}$- neighborhood of $x_{h,\lambda}$, and the second probability covers the case where x or y fall outside this neighborhood. But $I(\cdot,\cdot,\cdot)$ is a.s. continuous on the compact set $[h_0,\bar{h}_n]\times[-1,1]\times[a,b]$ by dominated convergence, and so both probabilities are zero, and hence so is (8).

Again by the fact that $S^n_{h,\lambda}$ has no upcrossing through 0 and by (5), the probability in (7) is not larger than

$$P\Big(\sup_{h\in(0,h_0),|\lambda|<1}\ \sup_{x\in[a,b]}\Big\{\Big|\frac{1}{\sigma_\lambda\sqrt{h}}\int\Big((1-|\lambda|)\varphi' + \lambda\varphi\Big)\Big(\frac{x-t}{h}\Big)dZ(t)\Big| - \sqrt{2\Big(\log\frac{1}{h\bar{f}_h}\Big)^+}\Big\}$$
$$> q_\alpha - 2\sqrt{h_0}\parallel r\parallel_\infty\Big) \qquad (9)$$

It follows from theorem 6.1 in [11] that $\sup_{h\in(0,h_0),|\lambda|<1}\sup_{x\in[a,b]}\Big\{\Big|\frac{1}{\sigma_\lambda\sqrt{h}}\int\Big((1-|\lambda|)\varphi' + \lambda\varphi\Big)\Big(\frac{x-t}{h}\Big)dZ(t)\Big| - \sqrt{2\Big(\log\frac{1}{h\bar{f}_h}\Big)^+}\Big\}^+ \downarrow 0$ as $h_0 \to 0$. This shows firstly that $q_\alpha := \inf\Big\{q : P\Big(\text{for some } h\in(0,\bar{h}_n], |\lambda|<1 : \frac{1}{\sigma_\lambda\sqrt{h}}\int\Big((1-|\lambda|)\varphi' + \lambda\varphi\Big)\Big(\frac{x-t}{h}\Big)dZ(t) \text{ has an upcrossing of } \mp(q + \sqrt{2(\log 1/(h\bar{f}_h))^+}) \text{ on } [a,b]\Big) \leq \alpha\Big\}$ is finite and thus well-defined. Secondly, it follows that for h_0 small enough the sup in (9) is smaller than the positive number

$q_\alpha - 2\sqrt{h_0} \parallel r \parallel_\infty$, so that the probability in (9) goes to zero as $h_0 \to 0$ by Fatou's lemma. Thus the test is asymptotically conservative, a fact which can be traced to the strong variation reducing property of the Gauss kernel. The proof in the case where there are $k > 1$ components in the mixture is quite analogous. □

Proof of theorem 3: For part (1) set $f_0 \equiv 1_{[0,1]}$ and consider a uniform prior on the alternatives $f_{j,k} = f_0 + \phi_j + \psi_k$, $1 \le j, k \le m$, where these quantities are defined as follows: Let h be as given as in the assumptions of the theorem, and for $b > 0$ to be determined shortly, set $m := \lfloor \frac{1}{4hb} \rfloor$ and $\tilde{\varphi} := \varphi 1_{[-b,b]}$. Then set $\phi_j(t) := \sqrt{\frac{2h(\log(1/h)^+}{\sigma_0^2 n}} \tilde{\varphi}_h'(t - [2(j-1)+1]bh)$ and $\psi_k(t) := -\sqrt{\frac{2h(\log(1/h)^+}{\sigma_0^2 n}} \tilde{\varphi}_h'(t - \frac{1}{2} - [2(k-1)+1]bh)$. As $d < 1$ one can choose b large enough (depending only on d) so that for some $\epsilon > 0$

$$\begin{aligned}(\varphi_h' \star f_{j,k})([2(j-1)+1]bh) &< -(d+\epsilon)\sqrt{\frac{2h(\log(1/h)^+}{\sigma_0^2 n}}|(\varphi_h' \star \varphi_h')(0)| \\ &= -(d+\epsilon)\sqrt{\frac{(\log(1/h)^+}{2\sqrt{\pi} n h}}\end{aligned}$$

and $(\varphi_h' \star f_{j,k})(\frac{1}{2} + [2(k-1)+1]bh) > (d+\epsilon)\sqrt{\frac{(\log(1/h)^+}{2\sqrt{\pi} n h}}$. Using $h\sqrt{(f_{j,k})_h}' = \frac{1}{2}\varphi_h' \star f_{j,k}/\sqrt{(f_{j,k})_h}$ and $\parallel f_{j,k} - f_0 \parallel_\infty \to 0$ $(m(n) \to \infty)$ and choosing l small enough one finds that $f_{j,k} \in \mathcal{F}_n(d)$ for n large enough. As in the proof of theorem 3 in [32] it can be shown that

$$\frac{1}{m^2} \sum_{1 \le j,k \le m} e^{\Lambda_{j,k}^n} \to 1 \quad \text{in } P_{f_0}\text{-probability as } n \to \infty, \tag{10}$$

where $\Lambda_{j,k}^n = \sum_{i=1}^n \log\Big(1 + \phi_j(X_i) + \psi_k(X_i)\Big)$ denotes the log-likelihood ratio. Hence if $\{\Psi_n(\underline{X}_n)\}$ is any procedure for detecting bimodality at level $\alpha_n \to \alpha$, then for arbitrary $\epsilon > 0$

$$\begin{aligned}\sup_{f \in \mathcal{F}_n(d)} P_f\Big(\Psi_n \text{ does not detect bimod.}\Big) &\ge \frac{1}{m^2} \sum_{1 \le j,k \le m} P_{f_{j,k}}(\Psi_n \text{ d.n.d.b.}) \\ &\ge E_{f_0}\Big(1(\Psi_n \text{ d.b.}) + \frac{1}{m^2} e^{\Lambda_{j,k}^n} 1(\Psi_n \text{ d.n.d.b.})\Big) - \alpha_n \\ &\ge (1-\epsilon) P_{f_0}\Big(\frac{1}{m^2} \sum_{1 \le j,k \le m} e^{\Lambda_{j,k}^n} \ge 1 - \epsilon\Big) - \alpha_n.\end{aligned}$$

So the assertion follows from (10).

For part (2) let $f \in \mathcal{F}_n(d)$ with $d > 1$. By the mean value theorem there exist $x < y$ such that $-d\sqrt{\frac{(\log(1/h)+}{8\sqrt{\pi}nh^3}} \geq \sqrt{\bar{f}_h(x)}' = \frac{1}{2}\overline{f_h(x)}'/\sqrt{\bar{f}_h(x)} = \frac{\frac{\sigma_0}{h}\bar{s}_{h,0}(x)}{2\sqrt{\bar{f}_h(x)}}$ and $d\sqrt{\frac{(\log(1/h)+}{8\sqrt{\pi}nh^3}} \leq \sqrt{\bar{f}_h(y)}' = \frac{\frac{\sigma_0}{h}\bar{s}_{h,0}(y)}{2\sqrt{\bar{f}_h(y)}}$. As $d > 1$ and $h(n) \to 0$, this yields $\sqrt{nh/\hat{f}_h(y)}\bar{s}_{h,0}(y) - \sqrt{nh/\hat{f}_h(y)}$ $U_{n,h}(y) \to \infty$ a.s. and $\sqrt{nh/\hat{f}_h(x)}\bar{s}_{h,0}(x) - \sqrt{nh/\hat{f}_h(x)}L_{n,h}(x) \to -\infty$ a.s., uniformly in $f \in \mathcal{F}_n(d)$, as f is uniformly bounded above and below. By the latter reason and (5), the joint law of $\left(\sqrt{nh/\hat{f}_h(x)}(\hat{s}_{h,0}(x) - \bar{s}_{h,0}(x)), \sqrt{nh/\hat{f}_h(y)}(\hat{s}_{h,0}(y) - \bar{s}_{h,0}(y))\right)$ converges to a bivariate normal distribution whose marginal variances are bounded above, uniformly in $f \in \mathcal{F}_n(d)$, x and y, and hence $\inf_{f\in\mathcal{F}_n(d)} P_f\left(\hat{s}_{h,0} \text{ has an upcrossing of } [L_{n,h}, U_{n,h}]\right) \to 1$. □

References

[1] Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Stat.* **1**, 1071–1095

[2] Brown, L.D., Johnstone, I.M. and MacGibbon, K.B. (1981). Variation diminishing transformations: A direct approach to total positivity and its statistical applications. *J. Amer. Statist. Assoc.* **76**, 824–832

[3] Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves. *J. Amer. Statist. Assoc.* **94**, 807–823

[4] Cheng, M.-Y. and Hall, P. (1998). Calibrating the excess mass and dip tests of multimodality. *J.R. Statist. Soc.* **B 60**, 579–589

[5] Cheng, M.-Y. and Hall, P. (1999). Mode testing in difficult cases. *Ann. Stat.* **27**, 1294-1315

[6] Côté, P., Welch, D.L., and Fischer, P.: The detection of an extended moving group near the galactic disk. Astrophysical Journal Lett. **406** (1993), L59–L62

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

[8] Dacunha-Castelle, D. and Gassiat, E. (1997). The estimation of the order of a mixture model. *Bernoulli* **3**, 279–299

[9] Donoho, D.L. (1988). One-sided inference about functionals of a density. *Ann. Stat.* **16**, 1390–1420

[10] Dudley, C.R.K., Giuffra, L.A., Raine, A.E.G., and Reeders, S.T. (1991), Assessing the role of APNH, a gene encoding for a human amiloridesensitive Na^+/H^+ antiporter, on the interindividual variation in red cell Na^+/Li^+ countertransport. *J. of the American Society of Nephrology* **2**, 937–943

[11] Dümbgen, L. and Spokoiny, V.G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Stat.* **29**, 124-152

[12] 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. Amer. Statist. Assoc.* **75**, 42–56

[13] Hall, P. (1991). On convergence rates of suprema. *Prob. Th. Rel. Fields* **89**, 447–455

[14] Hartigan, J.A. and Hartigan, P.M. (1985). The dip test of unimodality. *Ann. Stat.* **13**, 70–84

[15] Izenman, A.J. and Sommer, C.J. (1988). Philatelic mixtures and multimodal densities. *J. Amer. Statist. Assoc.* **83**, 941–953

[16] Karlin, S. (1968). *Total Positivity, Vol. I.* Stanford University Press

[17] Komlos, J., Major, P. and Tusnady, G. (1975). An approximation of partial sums of independent r.v.'s and the sample d.f. *Z. Wahrsch. verw. Gebiete* **32**, 111–131

[18] Leonard, T. (1978). Density estimation, stochastic processes, and prior information (with discussion). *JRSS B* **40**, 113–146

[19] Lindsay, B.G. (1989). Moment matrices: applications in mixtures. *Ann. Stat.* **17**, 722–740

[20] Lindsay, B. G. and Roeder, K. (1992). Residual diagnostics for mixture models. *J. Amer. Statist. Assoc.* **87**, 785–794

[21] Lindsay, B. G. and Roeder, K. (1997). Moment-based oscillation properties of mixture models. *Ann. Stat.* **25**, 378–386

[22] Marshall, A.W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and its Application.* Academic Press

[23] Minnotte, M.C. (1997). Nonparametric testing of the existence of modes. *Ann. Stat.* **25**, 1646–1660

[24] Müller, D.W. and Sawitzki, G. (1981). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86**, 738–746

[25] Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.* **89**, 487–495

[26] Rosenblatt, M. (1991). *Stochastic Curve estimation.* NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 3

[27] Schoenberg, I.J. (1950). On Polya frequency functions, II: Variation-diminishing integral operators of the convolution type. *Acta Sci. Math. Szeged* **12**, 97–106

[28] Shaked, M. (1980). On mixtures from exponential families. *JRSS B*, **43**, 97–99

[29] Silverman, B.W. (1981). Using kernel density estimates to investigate multimodality. *JRSS B* **43**, 97–99

[30] Silverman, B.W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Stat.* **10**, 795–810

[31] Stone, C.J., Hansen, M.H., Kooperberg, C. and Truong, Y.K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Ann. Stat.* **25**, 1371–1424

[32] Walther, G. (2001a). Multiscale maximum likelihood analysis of a semiparametric model, with applications. *Ann. Stat.* To appear.

[33] Walther, G. (2001b). Kernel oscillation analysis of the mixture complexity. Manuscript.

[34] Walther, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *J. Amer. Statist. Assoc.* To appear.

Guenther Walther

Department of Statistics, 390 Serra Mall

Stanford University, Stanford, CA 94305

email walther@stat.stanford.edu