

NON-STANDARD BOUNDARY BEHAVIOUR IN TWO-COMPONENT MIXTURE MODELS

HEATHER BATTEY, PETER MCCULLAGH, AND DANIEL XIANG

ABSTRACT. The empirical Bayes approach to multiple testing, widely used in modern statistical contexts, is frequently formulated as a two-component mixture model of the form $F_\theta = (1 - \theta)F_0 + \theta F_1$, where F_0 is standard Gaussian and F_1 is a completely specified heavy-tailed distribution with the same support. For a sample of n independent and identically distributed values $X_i \sim F_\theta$, the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically normal provided that $0 < \theta < 1$ is an interior point. This paper investigates the large-sample behaviour for boundary points, which is entirely different and strikingly asymmetric for $\theta = 0$ and $\theta = 1$. The reason for the asymmetry has to do with typical choices such that F_0 is an extreme boundary point and F_1 is usually not extreme. On the right boundary, well known results on boundary parameter problems are recovered, giving $\lim \mathbb{P}_1(\hat{\theta}_n < 1) = 1/2$. On the left boundary, $\lim \mathbb{P}_0(\hat{\theta}_n > 0) = 1 - 1/\alpha$, where $1 \leq \alpha \leq 2$ indexes the domain of attraction of the density ratio $f_1(X)/f_0(X)$ when $X \sim F_0$. For $\alpha = 1$, which is the most important case in practice, we show how the tail behaviour of F_1 governs the rate at which $\mathbb{P}_0(\hat{\theta}_n > 0)$ tends to zero. A new limit theorem for the joint distribution of the sample maximum and sample mean conditional on positivity establishes multiple inferential anomalies. Most notably, given $\hat{\theta}_n > 0$, the likelihood ratio statistic has a conditional null limit distribution $G \neq \chi_1^2$ determined by the joint limit theorem. We show through this route that no advantage is gained by extending the single distribution F_1 to the non-parametric composite mixture generated by the same tail-equivalence class.

Some key words: α -stable limit law; likelihood-ratio statistic; local false discovery rate; multiple testing; non-standard likelihood theory; regular variation; tail behaviour.

1. INTRODUCTION

Let F_0, F_1 be distinct probability distributions on the same measurable space. For each $0 \leq \theta \leq 1$, the mixture

$$F_\theta = (1 - \theta)F_0 + \theta F_1 \tag{1.1}$$

is also a probability distribution on the same space. In this paper, the phrase “mixture model generated by F_0, F_1 ” is interpreted as the set of convex combinations

$$\text{conv}(F_0, F_1) = \{F_\theta : 0 \leq \theta \leq 1\}$$

in which the generators F_0, F_1 are the boundary points.

The model (1.1) arises particularly in connection with the empirical Bayes approach to multiple testing (Efron et al., 2001), where interest is either in evaluating the simultaneous correctness of a set of statements concerning the same null hypothesis, or in the assessment of different null hypotheses, at most a small number of which are false. The case in which one of the components is standard Gaussian reflects a reduction by sufficiency to a set of pivotal test statistics and is particularly relevant for the applications we have in mind. The other component is typically chosen to have tails that are heavier than Gaussian, for instance a standard Gaussian convolution corresponding to a signal plus noise model at

Department of Mathematics, Imperial College London.
Department of Statistics, University of Chicago.

the relevant site. The interpretation of *site* is context-specific and may refer to a genomic locus in a genetics context or an energy level in a particle physics context. The response at site i is treated as a draw from F_θ , which means F_1 with probability θ and F_0 with probability $1 - \theta$, on account of the label being unknown.

The goal of this paper is to study the large-sample behaviour of the maximum-likelihood estimator and related statistics when the generators have the same support and the data are generated independently according to the boundary distribution $X_1, \dots, X_n \sim F_0$. In that case, we write $h(x) = f_1(x)/f_0(x)$ for the density ratio at x . For $\theta = 0$, the error probability $\mathbb{P}_0(\hat{\theta}_n > 0)$ has a large-sample limit, which is strictly positive in certain cases, and zero in other cases. It is shown in §3 that $\lim \mathbb{P}_0(\hat{\theta}_n > 0)$ is zero if and only if the random variables $h(X_i)$ belong to the domain of attraction of the Cauchy class, i.e., the α -stable family with index $\alpha = 1$. Otherwise, if $h(X_i)$ belongs to the domain of attraction of any non-Cauchy class, which necessarily has index $1 < \alpha \leq 2$, the limit probability $1 - \alpha^{-1}$ is strictly positive but not more than one half.

For reasons that are explained in §2.5, the Cauchy class is virtually the default in practical work with mixtures where the null generator is Gaussian and F_1 is symmetric with heavier tails. In such situations, it is of interest to characterize the rate at which the error tends to zero. For example, if F_1 has regularly-varying tails with index -2δ , i.e., the upper tail probability $\bar{F}_1(x) = \mathbb{P}_1(X > x)$ satisfies

$$\lim_{x \rightarrow \infty} \frac{\bar{F}_1(tx)}{\bar{F}_1(x)} = t^{-2\delta}$$

for $t > 0$, then the rate is logarithmic:

$$\mathbb{P}_0(\hat{\theta}_n > 0) \sim \frac{\delta}{\log n}.$$

By contrast, if F_1 has exponential tails, e.g., $-\log \bar{F}_1(x) \sim |x|^{2\gamma}$ for large x and $0 < \gamma < 1$, then

$$\mathbb{P}_0(\hat{\theta}_n > 0) \sim \frac{\gamma 2^\gamma}{(\log n)^{1-\gamma}}.$$

Here and henceforth, $\mathbb{P}_0(\cdot)$ denotes the probability with respect to the null generator or its n -fold product $F_0^{\otimes n}$, and the regular-variation notation $a(n) \sim b(n)$ means

$$\lim_{n \rightarrow \infty} \frac{a(n)}{b(n)} = 1,$$

or equivalently $a(n) = b(n)(1 + o(1))$ for positive functions $a(n), b(n)$. If a_n and b_n are random variables, $a_n \sim b_n$ means that the above limit holds for any realization.

The limit distribution of the likelihood ratio statistic conditional on $\hat{\theta}_n > 0$ is obtained in §4.3 using a non-quadratic local approximation. The derivation is based on a new limit theorem establishing a large-sample joint distribution for the sample mean and sample maximum of a random variable in the Cauchy domain of attraction, conditional on positivity of the sample mean.

The theoretical results derived in §3.4 and §3.5 may be of interest in their own right but here we focus on their implications in the context of the two-component mixture problem.

2. MAXIMUM LIKELIHOOD

2.1. Regularity conditions. Under the model (1.1), X_1, \dots, X_n are independent random variables with distribution F_θ for some $0 \leq \theta \leq 1$. Standard theory for maximum-likelihood estimators tells us that, under suitable regularity conditions, $n^{1/2}(\hat{\theta}_n - \theta)$ has a zero-mean Gaussian limit. A principal regularity condition is that all distributions in the model have the same support. This condition need not be satisfied by F_0, F_1 , but it

is automatic for the sub-model in which $0 < \theta < 1$. Whether F_0 and F_1 have the same support or not, a second principal regularity condition requires θ to be an interior point. Otherwise, if $\theta = 0$, the event $\hat{\theta}_n = 0$ has positive probability, so the limit distribution cannot be Gaussian. The mixture model can therefore be regular only if the boundary points are excluded. As discussed in §2.4, from a list of further regularity conditions, there is the possibility that more than one is violated. In the empirical Bayes formulations that motivated the present work, infinite variance of the density ratio $f_1(X)/f_0(X)$ presents considerable challenges not covered by existing literature on boundary inference problems.

2.2. Unequal supports. For completeness, we consider briefly the case where the supports $\text{supp}(F_0)$ and $\text{supp}(F_1)$ are not equal. The qualitative behaviour of the maximum-likelihood estimator is conveniently illustrated by an example in which F_0 is uniform on $(0, 2)$ and F_1 is uniform on $(1, 3)$. The likelihood function is

$$L(\theta) = 2^{-n}(1 - \theta)^{n_0}\theta^{n_2},$$

where n_r is the number of sample points in the interval $(r, r + 1)$, a realization of N_r . In the null case $X_1, \dots, X_n \sim F_0$, we have $n_2 = 0$ and

$$\mathbb{P}_0(\hat{\theta}_n > 0) = \mathbb{P}_0(N_0 = 0) = 2^{-n}.$$

Thus, $\hat{\theta}_n \rightarrow 0$ at an exponential rate.

The generators are said to have disjoint supports if there exist disjoint events S_0, S_1 such that $F_0(S_0) = F_1(S_1) = 1$. By modifying the preceding example so that either the distributions have disjoint supports or $\text{supp}(f_0)$ is a subset of $\text{supp}(f_1)$, we find $\hat{\theta}_n = 0$ with \mathbb{P}_0 -probability one for every $n \geq 1$. By contrast, if $\text{supp}(f_1)$ is a proper subset of $\text{supp}(f_0)$, we find that $\lim \mathbb{P}_0(\hat{\theta}_n > 0) = 1/2$.

It appears from this analysis for uniform distributions that the type-one error rate $\mathbb{P}_0(\hat{\theta}_n > 0)$ has a large-sample limit, which is either zero or one half. This behaviour is mirrored qualitatively in the equal-supports case, although the range of possible limits is zero to one half, depending on the tail behaviour of the random variable $h(X) = f_1(X)/f_0(X)$. The qualitative analogue of $\text{supp}(f_0) \subseteq \text{supp}(f_1)$ and $\text{supp}(f_1) \subseteq \text{supp}(f_0)$ is that f_0 has, respectively, lighter and heavier tails than f_1 . For the statistical applications we have in mind, $f_0(x) = o(f_1(x))$ as $|x| \rightarrow \infty$, with f_0 the standard Gaussian density function, and our discussion from §2.5 onwards concerns almost exclusively that case.

2.3. Equal supports. In the standard case, the generators share a common support and have positive densities $f_0(x), f_1(x)$, so that F_θ has density $f_\theta(x) = (1 - \theta)f_0(x) + \theta f_1(x)$, which is linear in θ . The contribution $\log f_\theta(x)$ to the log likelihood from a single observation at x is strictly concave as a function of θ , and the sum $\sum_i \log f_\theta(x_i)$ is also strictly concave. Consequently, the maximum-likelihood estimate is either a stationary point or a boundary point.

The log likelihood derivatives are

$$\begin{aligned} l'(\theta) &= \sum_{i=1}^n \frac{h(X_i) - 1}{1 - \theta + \theta h(X_i)} \\ l''(\theta) &= - \sum_{i=1}^n \left(\frac{h(X_i) - 1}{1 - \theta + \theta h(X_i)} \right)^2 \leq 0, \end{aligned} \tag{2.1}$$

and $l'(0) = \sum (h(X_i) - 1)$, where $h(x) = f_1(x)/f_0(x)$ is the density ratio of the boundary points. It follows that the event $\hat{\theta}_n > 0$ is the same as the event $n^{-1} \sum h(X_i) > 1$, where X_1, \dots, X_n are independently distributed as F_0 . In other words, $\hat{\theta}_n > 0$ if and only if the

sample average of the transformed variables $h(X_i)$ exceeds its expected value

$$E(h(X)) = \int h(x)f_0(x)dx = \int f_1(x)dx = 1.$$

If $h(X)$ has finite variance σ^2 , then $n^{-1/2}l'(0)$ is zero-mean Gaussian for large n , and $-n^{-1}l''(0) \rightarrow \sigma^2$ by the law of large numbers. In this case $\hat{\theta}_n = 0$ with probability $1/2$ in the large-sample limit; conditionally on $\hat{\theta}_n > 0$, the random variable $n^{1/2}\hat{\theta}_n > 0$ is distributed half-Gaussian with scale parameter σ . This is the familiar boundary-parameter result established by other authors (e.g. Chernoff, 1954; Self and Liang, 1987; Geyer, 1994). The boundary probabilities exhibit more interesting behaviours in the cases for which $h(X)$ does not have a finite second moment under the null model. It is those situations that we explore here.

2.4. Position within the literature. The inferential problem for two-component mixtures belongs to a class of boundary problems for which an extensive literature was carefully surveyed by Brazzale and Mamei (2024). Appendix A.1 of that work outlines an argument, due to Self and Liang (1987), establishing the limit distribution of the log likelihood-ratio statistic when the true value of the parameter is on the boundary. Two aspects of the argument are problematic when $h(X)$ does not have finite variance under the null distribution F_0 : that the Fisher information at $\theta = 0$ is not finite, this being the expectation of $\sum(h(X_i) - 1)^2$; and that a suitably rescaled version of $l'(0)$ is not asymptotically normally distributed. The same issues afflict the argument of Ghosh and Sen (1985) presented in Appendix A.3 of Brazzale and Mamei (2024).

In the context of a mixture model with two unit-Gaussian components, one having unknown mean, Bickel and Chernoff (1993) and Liu and Shao (2004) established that the likelihood ratio statistic diverges to infinity at rate $O(\log(\log n))$ and that the asymptotic null distribution of a suitably standardized version is of extreme-value type. The conclusions of the present paper are strikingly different: that the likelihood ratio statistic converges to zero under the null model and that its conditional distribution given positivity has a limit distribution G (Theorem 4.1) that is visibly close to but not exactly χ_1^2 .

Gaussian-Gaussian mixtures of unequal means and equal variances lead to density ratios $h(X)$ that are log-linear in X and thus in the normal domain of attraction. The extreme-value limit distribution emerges from estimation of the unknown mean parameter in the non-null component. In recent work, Shi and Drton (2024) have studied a split-sample version of the likelihood ratio statistic for the same mixture problem, following the generic *universal inference* construction of Wasserman et al. (2020). In well-behaved parametric problems, universal inference violates sufficiency and is inferior to a standard likelihood-based analysis. Its advantage in the two-component mixture setting is that sample splitting breaks some dependencies in the likelihood ratio construction, allowing Shi and Drton (2024) to show that the asymptotic null distribution of the split-sample likelihood ratio statistic is standard Gaussian after centering and scaling. In our case, the source of the difficulties is different and is not evaded by sample splitting.

An important conceptual difference between the models is that the signal in Shi and Drton (2024) is an arbitrary fixed real number corresponding to the mean of the non-null distribution. By contrast, the signal in our formulation is a random variable whose realizations are centered at zero and appreciably large with non-negligible probability.

Li, Chen and Marriott (2009) and Chen and Li (2009) allow two-component mixture constructions of the form $(1 - \theta)f(x; \lambda_0) + \theta f(x; \lambda_1)$ such that the Fisher information at $\theta = 0$ is not finite. However they study a different problem based on a penalized likelihood

function that forces the estimate $\hat{\theta}$ away from 0 and 1. Their implicit assessment of homogeneity is thus based on $\lambda_0 = \lambda_1$ rather than $\theta = 0$.

Contrasted with Ghosh and Sen (1985), Self and Liang (1987) and Bickel and Chernoff (1993), non-existence of second moments necessitates a radically different approach based on the theory of α -stable limits (Gnedenko and Kolmogorov, 1954) and regular variation (Bingham, Goldie and Teugels, 1987). From this we are able to delineate the role of tail properties in determining the type-one error rate $\mathbb{P}_0(\hat{\theta}_n > 0)$ and the anomalous limiting behaviour of likelihood-based statistics.

2.5. Tail behaviour for Gaussian mixtures. The motivating example for a large part of this paper is a restricted class of two-component Gaussian mixtures in which the density ratio $h(x)$ is an even function that is continuous, unbounded and ultimately monotone. Ultimate monotonicity means that to each η sufficiently large there corresponds a number ξ such that

$$\{x : h(x) > \eta\} = \{x : |x| > \xi\}. \quad (2.2)$$

Mills's approximation to the Gaussian tail probability yields

$$\mathbb{P}_0(|X| > \xi) = \frac{2\phi(\xi)}{\xi} \left(1 + O(\xi^{-2})\right).$$

On using the implicit definition $\phi(\xi) = \eta^{-1}f_1(\xi)$,

$$\mathbb{P}_0(h(X) > \eta) = \mathbb{P}_0(|X| > \xi) \sim 2\xi^{-1}\phi(\xi) = \frac{2f_1(\xi)}{\eta\xi}. \quad (2.3)$$

It follows that the asymptotic inverse relationship $\xi = h^{-1}(\eta)$ determines the null tail behaviour via $f_1(\xi)/\xi$, and thereby the limit distribution of normalized sums.

Let $L : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a slowly-varying function, in the sense that, for all $k > 0$

$$\lim_{x \rightarrow \infty} \frac{L(kx)}{L(x)} = 1. \quad (2.4)$$

For a large class of non-null generators, including all whose density satisfies

$$-\log f_1(x) \sim |x|^{2\kappa} L(x) \quad (2.5)$$

for some $0 \leq \kappa < 1$, we find that $\xi(\eta) \sim \sqrt{2\log \eta}$ as $\eta \rightarrow \infty$. This asymptotic inverse implies

$$\mathbb{P}_0(h(X) > \eta) \sim \frac{2f_1(\sqrt{2\log \eta})}{\eta\sqrt{2\log \eta}} \sim \frac{1}{\eta L_1(\eta)},$$

where, for every f_1 in the class (2.5), L_1 is also slowly varying. In all such cases, the random variable $h(X)$, or more correctly, its distribution, belongs to the Cauchy domain of attraction with index $\alpha = 1$, meaning that suitably standardized sums converge in distribution to a random variable in the Cauchy class, discussed in §3.2.

3. LIMIT DISTRIBUTIONS FOR SUMS

3.1. Introduction. The results of this section are stated in terms of a generic random variable X . This is both for notational convenience and for ease of application of the results beyond the Gaussian mixture setting for which they were initially conceived. For the statistical questions we have in mind, addressed in §5.3, the results are applied with $h(X)$ or $h(X) - 1$ in place of X , using the argument of §2.5.

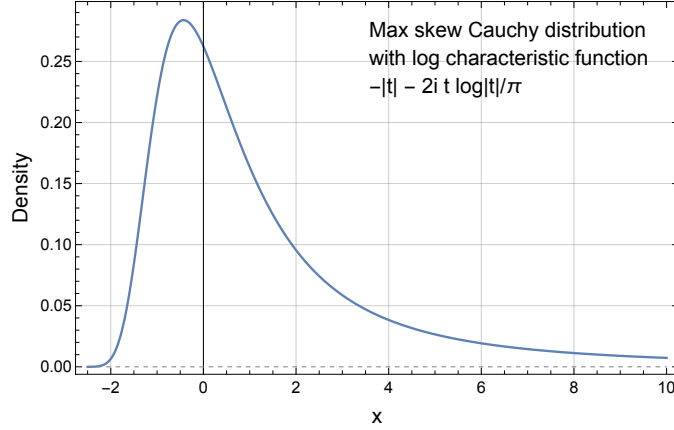


FIGURE 1. Density function of the maximally-skew Cauchy distribution

3.2. Stable limits. The theory of limit distributions for the sum of independent and identically distributed random variables is tied up with stability of convolutions. Modulo affine transformation, every distribution that has a convolution limit is associated with a pair (α, β) , and is said to be in the domain of attraction of the (α, β) stable law; in essence, the set of Borel distributions on the real line is partitioned into equivalence classes. There are also non-degenerate distributions that do not belong to the domain of attraction of any stable class, for example, $\bar{F}(x) \sim 1/\log x$.

Every stable distribution with $\alpha \geq 1$ has a density that is strictly positive on the real line; only in a few cases is it possible to express the density in terms of standard functions. However, the characteristic function of every limit distribution is necessarily of the form $\psi(a + bt)$ where

$$\log \psi(t) = \begin{cases} -t^2 & (\alpha = 2), \\ -|t|(1 + i\beta \operatorname{sign}(t) \frac{2}{\pi} \log |t|) & (\alpha = 1), \\ -|t|^\alpha (1 - i\beta \operatorname{sign}(t) \tan(\pi\alpha/2)) & (\alpha \neq 1). \end{cases}$$

See Bingham, Goldie and Teugels (1987, Theorem 8.3.2) for a statement of this result, or Gnedenko and Kolmogorov (1954, chapter 34) where the sign of β is mistakenly reversed for $\alpha \neq 1$ (see comment I.7 on page 253 of Zolotarev, 1986).

Note that not all (α, β) -combinations give rise to distinct distributions. In the Gaussian case ($\alpha = 2$) the last version of the log characteristic function reduces to $-t^2$, so β is immaterial and all limits are symmetric. However, limits in every other class are symmetric only if $\beta = 0$.

Zolotarev (1986, section 2.5) gives the tail behaviour of the (α, β) -limit distribution parameterized according to the characteristic function shown above:

$$P_{\alpha, \beta}(X > x) \sim (1 + \beta)C_\alpha x^{-\alpha}; \quad P_{\alpha, \beta}(X < -x) \sim (1 - \beta)C_\alpha x^{-\alpha},$$

where $C_\alpha = \pi^{-1}\Gamma(\alpha) \sin(\pi\alpha/2)$ and $C_1 = \pi^{-1}$. See also Feller (1966, section XVII.6). For $\beta = -1$, the first limit is interpreted as $x^\alpha P_{\alpha, \beta}(X > x) \rightarrow 0$ as $x \rightarrow \infty$; likewise for $\beta = 1$ in the second limit. Note that $C_2 = 0$, so this characterization $P_{2, \beta}(X > x) = o(x^{-2})$ is correct but not tight for Gaussian limits. We include a derivation of the above expression for the tail probability when $\alpha = 1$ in Appendix III.

For the two-component mixture problems considered in this paper, the class of limit distributions that can arise is a proper subset of those listed above. First, the fact that each summand $h(X_i)$ has finite mean implies $\alpha \geq 1$. Second, the fact that the random

variables are positive implies maximal skewness with $\beta = 1$.

$$P_{\alpha,\beta}(X < 0) = \frac{1}{2} \left(1 - \frac{b(\alpha - 2)}{2} \right), \quad (3.1)$$

where b relates to β as (Zolotarev, 1986, equation I.19)

$$\beta = \cot\left(\frac{\pi\alpha}{2}\right) \tan\left(\frac{\pi b(\alpha - 2)}{\alpha}\right),$$

with $\beta = b = 1$ at the boundary of the parameter space for $\beta \in [-1, 1]$. It follows that (3.1) reduces to $1/\alpha$ for $\beta = 1$.

In the asymmetric Cauchy class, the distribution with log characteristic function

$$\log \psi(t) = -|t| - 2it \log |t|/\pi$$

has the maximally skew density,

$$f(x) = \frac{1}{\pi} \int_0^\infty e^{-t} \cos(tx + 2t \log t/\pi) dt, \quad (3.2)$$

which is shown in Fig. 1. Although the summands in $l'(0)$ have finite mean, the characteristic function for $\alpha = 1$ does not have a first-order Taylor expansion so the limit distribution does not have a first moment. The left tail is sub-Cauchy and $P_{1,1}(X < 0) \simeq 0.3652$; the right tail behaviour is $f(x) \sim 2/(\pi x^2)$ or $\bar{F}(x) \sim 2/(\pi x)$, where $\bar{F}(x) = P_{1,1}(X > x)$.

3.3. Stabilizing sequences. A primary difficulty in understanding the boundary behavior of the MLE is finding an appropriate normalization of the sum $\sum_{i=1}^n h(X_i)$ that converges in distribution to a stable law. Since guidance on the construction of the stabilizing sequence is opaque in the literature on stable limits, we provide some background in this section.

Let F be the cumulative function of a distribution on the positive real line, and let $\bar{F}(x) = 1 - F(x)$ be the right-tail probability. In order that F belong to the domain of attraction of a stable law with index $0 < \alpha < 2$, it is necessary and sufficient that the tail be regularly varying, i.e.,

$$\bar{F}(x) \sim \frac{2C_\alpha}{x^\alpha L(x^\alpha)} \quad (3.3)$$

for large x and some slowly-varying function L . This is a re-statement of a special case of Theorem 2 from section 35 of Gnedenko and Kolmogorov (1954). The slow-variation factor in (3.3) is expressed in the form $L(x^\alpha)$ rather than $L(x)$ or $1/L(x)$ as this simplifies the scaling sequence (3.7). For the moment, the choice of constant is immaterial, but the particular choice

$$C_\alpha = \pi^{-1} \Gamma(\alpha) \sin(\pi\alpha/2), \quad (3.4)$$

matching the right tail of the limit distribution, will subsequently be convenient.

Let X_1, \dots, X_n be independent and identically distributed with distribution F in the domain of attraction of some stable law with index $0 < \alpha \leq 2$. Then there exist deterministic stabilizing sequences A_n, B_n such that

$$\frac{1}{B_n} \sum_{i=1}^n X_i - A_n$$

has the stable limit distribution $P_{\alpha,\beta}$ with index α . The skewness coefficient $|\beta| \leq 1$ is a balance between the two tails: if the support of F is bounded below, then $\beta = 1$. For $\alpha = 2$, the scaling coefficients are $B_n \propto n^{1/2}$. Otherwise, for $0 < \alpha < 2$, they are determined by the condition

$$\lim_{n \rightarrow \infty} n \bar{F}(B_n x) = 2C_\alpha x^{-\alpha}, \quad (3.5)$$

for each $x > 0$ (Gnedenko and Kolmogorov, 1954, section 35). Some partial intuition for why equation (3.5) determines the scaling sequence $\{B_n\}$ is that convergence to the stable limit, in terms of the characteristic function g of the original random variables, is equivalent to

$$e^{-iA_n t} g(t/B_n)^n \rightarrow \psi(t), \quad n \rightarrow \infty.$$

Since $t/B_n \rightarrow 0$ for fixed t , it follows that the behaviour of g near zero is key to solving for B_n . By the inversion formula for characteristic functions, the behavior of g near the origin manifests in the tail behaviour of \bar{F} . Conditions (3.3) and (3.5) imply

$$\lim_{n \rightarrow \infty} \frac{nB_n^{-\alpha}}{L(B_n^\alpha x^\alpha)} = \lim_{n \rightarrow \infty} \frac{nB_n^{-\alpha}}{L(B_n^\alpha)} = 1.$$

To calculate B_n , it suffices to invert the asymptotic relation $\tilde{B}_n L(n\tilde{B}_n) \sim 1$, where $\tilde{B}_n := B_n^\alpha/n$. This is achieved through the de Bruijn conjugate L^\dagger (de Bruijn, 1959), which is a slowly varying function defined (up to an asymptotic equivalence) by the condition,

$$L(xL^\dagger(x))L^\dagger(x) \sim L^\dagger(xL(x))L(x) \sim 1, \quad (3.6)$$

as $x \rightarrow \infty$. The condition $\tilde{B}_n L(n\tilde{B}_n) \sim 1$ implies that $\tilde{B}_n \sim L^\dagger(n)$, which defines the scaling sequence up to asymptotic equivalence as

$$B_n \sim (nL^\dagger(n))^{1/\alpha}. \quad (3.7)$$

For a proof of equation (3.6), see Theorem 1.5.13 of Bingham, Goldie and Teugels (1987). Conjugation is an involution $L^{\dagger\dagger} \sim L$, which is also a group inverse under the compositional operation discussed in Appendix I. For the moment, it suffices to remark that if $L(n) \sim K(\log n)^\gamma$ for any real γ and $K > 0$, then $L^\dagger(n) \sim 1/L(n)$. In general, however, the conjugate is not equivalent to the reciprocal, and the conjugate of a functional product is not the product of the conjugates.

For $1 < \alpha \leq 2$, the distribution F in (3.3) has a finite mean μ , and the footnote to Theorem 2 in section 35 of Gnedenko and Kolmogorov (1954) gives the centering sequence $A_n = n\mu/B_n$, which implies that

$$\frac{1}{B_n} \sum_{i=1}^n (X_i - \mu)$$

has the stable limit distribution $P_{\alpha,1}$ discussed in the preceding section. It follows from (3.1) that the exceedance event $\bar{X}_n > \mu$ has a large-sample limit probability

$$\lim_{n \rightarrow \infty} \mathbb{P}(\bar{X}_n > \mu) = P_{\alpha,1}(\varepsilon > 0) = 1 - 1/\alpha, \quad (3.8)$$

where ε is a random variable with distribution $P_{\alpha,1}$. This limit is strictly positive, but not more than one half. The case $0 < \alpha < 1$ does not arise in mixture models and is not discussed here. The single remaining case $\alpha = 1$ is a little more complicated because $A_n \neq n\mu/B_n$ even if the mean is finite. The development in §2.5 shows that it is also the most important case for Gaussian mixture models.

3.4. Stabilizing sequences for Cauchy limits. The main limit theorems for the boundary-parameter problem considered in this paper are developed here and in §3.5. They may be of standalone interest, and may have other applications.

In order for a distribution F with tail probability $\bar{F}(x) = 2C_1/(xL(x))$ to have a finite mean, it is necessary and sufficient that the tail contribution be finite. Integration by

parts gives

$$\begin{aligned} \int_T^\infty x dF(x) &= T\bar{F}(T) + \int_T^\infty \bar{F}(x) dx, \\ &= \frac{2C_1}{L(T)} + 2C_1 \int_T^\infty \frac{dx}{xL(x)}. \end{aligned}$$

For example, $L(x) = (\log x)^{1+\delta}$ suffices for finiteness only if $\delta > 0$. While both terms tend to zero for large T , the second term is dominant.

For $\alpha = 1$, the scaling sequence (3.7) is $B_n \sim nL^\dagger(n)$. Finiteness of the mean implies $L(n) \rightarrow \infty$, and hence $B_n/n \sim L^\dagger(n) \rightarrow 0$ as $n \rightarrow \infty$. The centering sequence given explicitly in the footnote to Theorem 2 of Gnedenko and Kolmogorov (1954) is

$$A_n \sim n\Im \log \psi(1/B_n) \sim n \int \sin(x/B_n) dF(x). \quad (3.9)$$

where $\Im c$ denotes the imaginary part of c . See also equation (6) in Chow and Teugels (1979). We now find an approximation for this sine-integral for a broad class of functions $L(x)$, conveniently parametrized in order to recover several important cases arising in Gaussian mixture models.

Theorem 3.1. Let F be a finite-mean distribution on the positive real line whose tail is $\bar{F}(x) \sim 2C_1/(xL(x))$, where $C_1 = 1/\pi$ and the slowly varying function $L(x)$ admits the parametrization

$$L(x) = (\beta_0 \log x)^{\delta+1} e^{(\beta_1 \log x)^\gamma} \quad (3.10)$$

for $\beta_0 > 0$. Finiteness of the mean implies either $\beta_1 > 0$ and $0 < \gamma < 1$, or $\beta_1 = 0$ and $\delta > 0$. With the convention that $\gamma = 0$ when $\beta_1 = 0$ and $0^0 = 0$, the sine-integral for large $T = 1/|t|$ is,

$$\int_0^\infty \sin(tx) dF(x) = \mu t - K_{\delta,\gamma,\beta_1} \frac{t(\log T)^{1-\gamma}}{L(T)} + o\left(\frac{t(\log T)^{1-\gamma}}{L(T)}\right), \quad |t| \rightarrow 0,$$

where μ is the mean, $K_{\delta,\gamma,\beta_1} = 2C_1/(\beta_1^\gamma \gamma)$ for $\gamma, \beta_1 > 0$, and $K_{\delta,0,0} = 2C_1/\delta$.

Remark 3.1. The form $\bar{F}(x) = 2C_1/(xL(x))$ implies by (3.3) that F belongs to the domain of attraction of the (maximally skew) Cauchy distribution. The parametrization (3.10) of the slowly varying function $L(x)$ is an encompassing form recovering many important examples as special cases; see §4.1.

Proof. The argument detailed in Appendix II shows that the dominant component of the sine-integral for small t is

$$\begin{aligned} \int_0^\infty \sin(tx) dF(x) &= t\mu - t \int_T^\infty x dF(x) + O\left(\frac{t}{L(T)}\right), \\ &= t\mu - 2C_1 t \int_T^\infty \frac{dx}{xL(x)} + O\left(\frac{t}{L(T)}\right), \end{aligned}$$

where the remainder is of smaller order than the second term.

For $\beta_1 = 0$, the second part is a straightforward integral

$$\int_T^\infty \frac{dx}{xL(x)} = \int_T^\infty \frac{dx}{x(\beta_0 \log x)^{\delta+1}} = \frac{1}{\beta_0 \delta (\beta_0 \log T)^\delta} = \frac{\log T}{\delta L(T)}.$$

For $\beta_1, \gamma > 0$, the transformation $u = (\beta_1 \log x)^\gamma$ gives rise to a gamma-tail integral

$$\int_T^\infty \frac{e^{-(\beta_1 \log x)^\gamma} dx}{x(\beta_0 \log x)^{\delta+1}} = \frac{1}{\beta_1 \gamma} \frac{\beta_1^{\delta+1}}{\beta_0^{\delta+1}} \int_{(\beta_1 \log T)^\gamma}^\infty u^{\delta/\gamma-1} e^{-u} \sim \frac{(\beta_1 \log T)^{1-\gamma}}{\beta_1 \gamma L(T)}.$$

□

Corollary 3.1. For a distribution satisfying the conditions of Theorem 3.1, the centering sequence (3.9) is

$$\begin{aligned} A_n &= \frac{n\mu}{B_n} - K_{\delta,\gamma,\beta_1} \frac{n(\log B_n)^{1-\gamma}}{B_n L(B_n)} (1 + o(1)), \\ &= \frac{\mu}{L^\dagger(n)} - K_{\delta,\gamma,\beta_1} \frac{(\log B_n)^{1-\gamma}}{L^\dagger(n) L(nL^\dagger(n))} (1 + o(1)), \\ &= \frac{\mu}{L^\dagger(n)} - K_{\delta,\gamma,\beta_1} (\log n)^{1-\gamma} (1 + o(1)). \end{aligned}$$

The third line follows from the definition of the de Bruijn conjugate in (3.6).

Corollary 3.2. For a distribution satisfying the conditions of Theorem 3.1, the distribution of the sample average for large n is

$$\frac{\bar{X}_n - \mu}{L^\dagger(n)} = -K_{\delta,\gamma,\beta_1} (\log n)^{1-\gamma} + \varepsilon + o_p(1),$$

where ε has the Cauchy limit with skewness $\beta = 1$. Given the right-tail behaviour of the limit distribution,

$$\mathbb{P}(\bar{X}_n > \mu) \sim 2K_{\delta,\gamma,\beta_1}^{-1} (\log n)^{\gamma-1} / \pi,$$

which tends to zero at rate $\beta_1^\gamma (\log n)^{\gamma-1}$ if $\beta_1, \gamma > 0$, or $\delta / \log n$ if $\beta_1 = 0$.

3.5. Joint distribution of $(\bar{X}_n, X_{(n)})$. In §4.3 we consider the conditional distribution of the likelihood ratio statistic, conditional on positivity of $\hat{\theta}_n$, for which we need the following theorem concerning the joint distribution of the sample average and the maximum order statistic $X_{(n)}$.

Theorem 3.2. Let F be a zero-mean distribution with support $(-1, \infty)$ and tail satisfying the conditions of Theorem 3.1, and let X_1, \dots, X_n be an iid sample from F . Let $T_n = B_n |A_n| = K_{\delta,\gamma,\beta_1} B_n (\log n)^{1-\gamma}$, where $B_n = n/L(n)$ and A_n is given in Corollary 3.1. Given $\bar{X}_n > 0$, the conditional limit distribution of $(\bar{X}_n, X_{(n)})$ is such that

$$\left(\frac{n\bar{X}_n}{T_n}, \frac{X_{(n)}}{T_n} \right) \sim \left(\frac{U}{1-U}, \frac{1}{1-U} \right),$$

where U is uniform on $(0, 1)$. It follows that $n\bar{X}_n/T_n \sim T_n/(n\bar{X}_n)$ is conditionally self-reciprocal, the ratio $n\bar{X}_n/X_{(n)} \sim U$ is conditionally uniform, and the support of the joint distribution degenerates to the line $X_{(n)}/T_n - n\bar{X}_n/T_n - 1 \rightarrow 0$.

Remark 3.2. The unconditional scaling factor for both $n\bar{X}_n$ and $X_{(n)}$ is B_n , so $(n\bar{X}_n/B_n - A_n, X_{(n)}/B_n)$ has a joint limit distribution, which has been studied by Chow and Teugels (1979) in this setting. However, the conditioning event $\bar{X}_n > 0$ has probability $O(B_n/T_n)$ tending to zero, so the joint limit distribution does not determine the limit of conditional distributions. Theorem 3.2 implies that $n\bar{X}_n > \epsilon T_n$ and $X_{(n)} > (1 + \epsilon)T_n$ are asymptotically equivalent events for each $\epsilon > 0$.

To prove the theorem, we first state a technical lemma that relates the sample mean to the sample maximum in the event where the latter is unusually large.

Lemma 3.1. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$, where $\bar{F}(x) \sim 2C_1/(xL(x))$ satisfies the conditions of Theorem 3.1, and let B_n, T_n be defined as in Theorem 3.2. For any fixed $y > 0$, conditional on $X_{(n)} > yT_n$, we have

$$\frac{n\bar{X}_n}{T_n} = \frac{X_{(n)} + \sum_{k \geq 1} X_{(n-k)}}{T_n} \sim \frac{X_{(n)}}{T_n} - 1 + o_p(1). \quad (3.11)$$

Consequently, for any fixed $y > 1$

$$\lim_{n \rightarrow \infty} \mathbb{P}(\bar{X}_n > 0 \mid X_{(n)} > yT_n) = 1,$$

so that $\mathbb{P}(\bar{X}_n > 0, X_{(n)} > yT_n) \sim \mathbb{P}(X_{(n)} > yT_n)$ as $n \rightarrow \infty$.

Proof. Let $X_{(n-k)}$ be the k th order statistic with $X_{(n+1)} = \infty$. Since the transformed variables $F(X_1), \dots, F(X_n)$ are independent uniform, the joint distribution of the top order statistics is such that successive differences of the probability-integral transformed values have exactly the same distribution as successive spacings of the top $k+1$ of n independent uniform order statistics, i.e.

$$n\{\bar{F}(X_{(n-r)}) - \bar{F}(X_{(n-r+1)})\}_{0 \leq r \leq k} \xrightarrow{(d)} (e_0, \dots, e_k),$$

where e_0, \dots, e_k are independent unit exponential variables, whose distribution approximates that of the spacings among the top $k+1$ uniform order statistics, for large n and $k \ll n$. Since $u := \bar{F}(x) \sim 2C_1/(xL(x))$ for large x implies $xL(x) \sim 2C_1/u$, and since the property $L^\dagger(xL(x))L(x) \sim 1$ implies $L(x) \sim 1/L^\dagger(2C_1/u)$, the inverse function inherits the asymptotic behaviour

$$\bar{F}^{-1}(u) \sim (2C_1/u) L^\dagger(2C_1/u), \quad (u \rightarrow 0).$$

Now since $n\bar{F}(X_{(n)}) \xrightarrow{(d)} e_0$ and

$$\bar{F}^{-1}(e_0/n) \sim (2C_1n/e_0) L^\dagger(2C_1n/e_0) \sim \frac{2C_1}{e_0} nL^\dagger(n) \sim 2C_1B_n/e_0,$$

it follows that $X_{(n)}/B_n \xrightarrow{(d)} 2C_1/e_0$. Similarly, since $n\bar{F}(X_{(n-k)}) \xrightarrow{(d)} e_0 + \dots + e_k$, we obtain

$$\bar{F}^{-1}((e_0 + \dots + e_k)/n) \sim \frac{2C_1B_n}{e_1 + \dots + e_k} \quad (3.12)$$

which implies $X_{(n-k)}/B_n \xrightarrow{(d)} 2C_1/(e_0 + \dots + e_k)$.

If $X_{(n)} > yT_n$ for some $y > 0$, then e_0 is unusually small compared with e_1, e_2, \dots , so that $e_0 + \dots + e_k \sim e_1 + \dots + e_k$. It follows that the conditional distribution of $X_{(n-1)}, X_{(n-2)}, \dots$ given $X_{(n)} > yT_n$ is approximately the same as the unconditional distribution of $X_{(n)}, X_{(n-1)}, \dots$. In that event, Corollary 3.2 implies

$$n\bar{X}_n = X_{(n)} + \sum_{k \geq 1} X_{(n-k)} \sim X_{(n)} - T_n + O_p(B_n), \quad (3.13)$$

which implies (3.11). For $y > 1$, $B_n = o(T_n)$ implies

$$\mathbb{P}(n\bar{X}_n > 0 \mid X_{(n)} > yT_n) \rightarrow 1.$$

□

Proof of Theorem 3.2. For $U \sim \text{Uniform}(0, 1)$,

$$\mathbb{P}\left(\frac{U}{1-U} > x, \frac{1}{1-U} > y\right) = \min\left\{\frac{1}{1+x}, \frac{1}{y}\right\}, \quad y > 1, x > 0.$$

By the last line of Lemma 3.1, we have

$$\mathbb{P}\left(\frac{n\bar{X}_n}{T_n} > x, \frac{X_{(n)}}{T_n} > y \mid \bar{X}_n > 0\right) \sim \mathbb{P}\left(\frac{n\bar{X}_n}{T_n} > x \mid \frac{X_{(n)}}{T_n} > y, \bar{X}_n > 0\right) \frac{\mathbb{P}(X_{(n)}/T_n > y)}{\mathbb{P}(\bar{X}_n > 0)}.$$

For any fixed y , the tail property of \bar{F} implies

$$\mathbb{P}(X_{(n)} > yT_n) \sim n\bar{F}(yT_n) \sim \frac{2C_1B_n}{yT_n},$$

and Corollary 3.2 implies $\mathbb{P}(\bar{X}_n > 0) \sim 2C_1 B_n / T_n$. It follows that for any fixed y ,

$$\lim_{n \rightarrow \infty} \frac{\mathbb{P}(X_{(n)}/T_n > y)}{\mathbb{P}(\bar{X}_n > 0)} \rightarrow \frac{1}{y}. \quad (3.14)$$

Since $X_{(n)}/T_n > y$ for $y > 1$ implies $\bar{X}_n > 0$ with probability 1 in the limit, we have

$$\mathbb{P}\left(\frac{n\bar{X}_n}{T_n} > x \mid \frac{X_{(n)}}{T_n} > y, \bar{X}_n > 0\right) \sim \mathbb{P}\left(\frac{n\bar{X}_n}{T_n} > x \mid \frac{X_{(n)}}{T_n} > y\right).$$

Lemma 3.1 implies that, conditional on $X_{(n)} > yT_n$,

$$\frac{n\bar{X}_n}{T_n} = \frac{X_{(n)} + \sum_{k \geq 2} X_{(n-k+1)}}{T_n} \sim \frac{X_{(n)}}{T_n} - 1 + o_p(1),$$

so the previous expression is, by the argument leading to (3.14),

$$\mathbb{P}\left(\frac{n\bar{X}_n}{T_n} > x \mid \frac{X_{(n)}}{T_n} > y\right) \sim \mathbb{P}\left(\frac{X_{(n)}}{T_n} > 1 + x \mid \frac{X_{(n)}}{T_n} > y\right) \rightarrow \begin{cases} \frac{y}{1+x} & \text{if } 1 + x > y \\ 1 & \text{otherwise.} \end{cases}$$

Together with (3.14), this proves the claim. \square

4. IMPLICATIONS FOR INFERENCE IN THE TWO-COMPONENT GAUSSIAN MIXTURE MODEL

4.1. Left-boundary behaviour induced by the tail of f_1 . We start by considering in Example 4.1 non-null generators that depart only slightly from Gaussianity in the relevant sense. Typical choices in practice are closer to Examples 4.2 and 4.3, and result in density ratios whose distributions belong to the domain of attraction of the Cauchy family. These require the more elaborate theory of §3.4.

Example 4.1. Let $\nu > -1$, and suppose that the non-null generator has density

$$f_1(x) \propto |x|^\nu \phi(x),$$

so that $h(x) \propto |x|^\nu$ is even. For $\nu \neq 0$, both boundary points are extreme.

Ultimate monotonicity requires $\nu > 0$, in which case $h(x) > \eta$ is equivalent to $|x| > \text{const } \eta^{1/\nu}$. In that case, the tail behaviour of the density ratio is

$$-\log \mathbb{P}_0(h(X) > \eta) \simeq \text{const } \eta^{2/\nu} + L(\eta),$$

where L is slowly varying. It follows that all moments exist and $h(X)$ belongs to the domain of attraction of the normal distribution. Thus, $\lim \mathbb{P}_0(\hat{\theta}_n > 0) = 1/2$ by the discussion of §2.3.

For $-1 < \nu < 0$, the ratio $h(x) \propto |x|^\nu$ is monotone decreasing in $|x|$, so condition (2.2) is not satisfied. However, the tail behaviour can be found by a simpler argument:

$$\mathbb{P}_0(h(X) > \eta) = \mathbb{P}_0(|X| < \text{const } \eta^{1/\nu}) \sim \text{const } \eta^{1/\nu}.$$

In other words, the tail is regularly varying with index $\alpha = -1/\nu > 1$. In this case, $h(X)$ belongs to the normal domain of attraction only if $\nu \geq -1/2$, in which case $\lim \mathbb{P}_0(\hat{\theta}_n > 0) = 1/2$. Otherwise, if $-1 < \nu < -1/2$, the behaviour is nonstandard and the limit is $\lim \mathbb{P}_0(\hat{\theta}_n > 0) = 1 - 1/\alpha = 1 + \nu$.

Example 4.2. In a continuation of Example 4.4, consider a Gaussian mixture with f_1 the standard Cauchy density function. The density ratio $h(x) \sim \text{const } x^{-2} e^{x^2/2}$ is not monotone in $|x|$, but it is monotone for $|x| > 1$, and thus ultimately monotone according to (2.2). The equation $h(\xi) = \eta$ has an asymptotic solution

$$\frac{1}{2}\xi^2 = \log \eta + o(\log \eta),$$

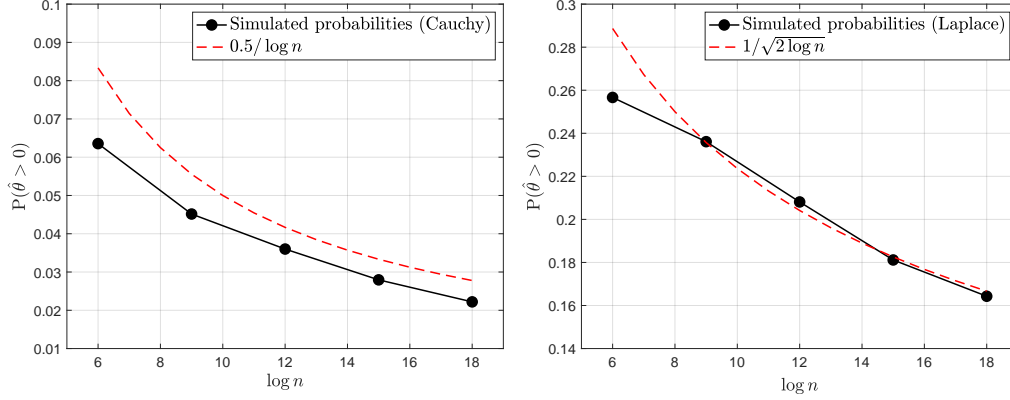


FIGURE 2. Simulated probabilities $\mathbb{P}(\hat{\theta}_n > 0)$ based on 2×10^4 replicates, where $\hat{\theta}_n$ estimates θ in the model $(1 - \theta)N(0, 1) + \theta F_1$, where F_1 is standard Cauchy (left) and standard Laplace (right).

so that $\xi^{-1}f_1(\xi) \sim (2 \log \eta)^{-3/2}/\pi$, and the tail approximation (2.3) is

$$\mathbb{P}_0(h(X) > \eta) \sim \frac{2}{\pi \eta (2 \log \eta)^{3/2}}.$$

This is of the form in Theorem 3.1 with $(\beta_0, \beta_1, \delta, \gamma) = (2, 0, 1/2, 0)$, leading by Corollary 3.2 to the conclusion that

$$\mathbb{P}_0(\hat{\theta}_n > 0) \sim \frac{1}{2 \log n}.$$

Example 4.3. Consider a Gaussian mixture in which the non-null generator has regularly varying tails according to (2.5) with tail index $0 \leq \kappa < 1$. Then $h(\cdot)$ is ultimately monotone, and the equation $h(\xi) = \eta$ has asymptotic solution

$$\xi^2 = 2 \log \eta + (\log \eta)^\kappa L((2 \log \eta)^{1/2}) + o((\log \eta)^\kappa).$$

For $\kappa < 1$, the composition $f_1(\xi(\eta))$ is slowly-varying as a function of η , and the tail approximation (2.3) is

$$\mathbb{P}_0(h(X) > \eta) \sim \frac{2f_1(\xi(\eta))}{\eta \xi(\eta)} \sim \frac{2c_\kappa e^{-(2 \log \eta)^\kappa}}{\eta \sqrt{2 \log \eta}},$$

where c_κ is a normalizing constant. This is of the form in Theorem 3.1 with $(\beta_0, \beta_1, \delta, \gamma) = (2/(c_\kappa \pi)^2, 2, -1/2, \kappa)$, leading by Corollary 3.2 to the conclusion that

$$\mathbb{P}_0(\hat{\theta}_n > 0) \sim \frac{2^\kappa \kappa}{(\log n)^{1-\kappa}}.$$

In particular, for the Laplace distribution $f_1(x) = \frac{1}{2}e^{-|x|}$, $\kappa = 1/2$ and the convergence rate is $(\sqrt{2 \log n})^{-1}$.

The conclusions of Examples 4.2 and 4.3 are compared with simulated probabilities in Figure 2.

4.2. Parametrization for two-component mixtures. When the data are generated by \mathbb{P}_θ for the right boundary point $\theta = 1$, the standard asymptotics apply. The purpose of this section is to provide intuition for why the left boundary behavior is different from the right, and why the latter follows from existing results.

There may exist alternative values of the mixture parameter $\theta < 0$ or $\theta > 1$ in (1.1) such that f_θ remains a valid probability density. Consider the set of all such θ , for which the extreme points are

$$\begin{aligned}\theta_{\min} &:= \min \left\{ \theta : f_\theta \geq 0 \text{ and } \int f_\theta(x) dx = 1 \right\} \\ \theta_{\max} &:= \max \left\{ \theta : f_\theta \geq 0 \text{ and } \int f_\theta(x) dx = 1 \right\}.\end{aligned}$$

When f_1 has heavier tails than f_0 , the extreme points typically satisfy $\theta_{\min} = 0$ and $\theta_{\max} > 1$, and the original model $\{f_\theta : 0 \leq \theta \leq 1\}$ is a strict subset of the extended model

$$\{(1 - \rho)f_0 + \rho f_{\theta_{\max}} : 0 \leq \rho \leq 1\}. \quad (4.1)$$

The extended model was considered by various authors (see, e.g. Genovese and Wasserman (2004) and Patra and Sen (2016)) in a multiple testing context. Patra and Sen (2016) noted that when $f_1 \geq cf_0$ for some constant $c > 0$, some of the probability mass in f_1 can be re-assigned to the null component without changing the density of the overall mixture. To address this non-identifiability issue, they define

$$\alpha_0 := \inf\{\gamma \in (0, 1] : (F_\theta - (1 - \gamma)F_0)/\gamma \text{ is a cdf}\},$$

which is the smallest mixture weight for the non-null component that is consistent with the marginal cdf F_θ . In the current setting, α_0 corresponds to ρ in model (4.1).

The extension (4.1) of the parameter space for θ is illustrated for two Gaussian mixture problems in Examples 4.4 and 4.5 below.

Example 4.4. Let $f_0(x) = \phi(x)$ be the standard Gaussian distribution, and let $f_1(x)$ be the standard Cauchy distribution. Then f_θ has a density

$$(1 - \theta)\phi(x) + \theta f_1(x) = \phi(x)(1 - \theta + \theta h(x)),$$

where $h(x) = f_1(x)/\phi(x)$ is the density ratio. The positivity condition $f_\theta \geq 0$ implies

$$\theta(h(x) - 1) \geq -1$$

for all x, θ . On the subset for which $h(x) \geq 1$, this implies the lower bound

$$\theta_{\min} = \sup_{x:h(x)>1} \frac{-1}{h(x) - 1} = 0, \quad (4.2)$$

which is zero since h is unbounded as $x \rightarrow \infty$. On the subset for which $h(x) \leq 1$, positivity implies the upper bound

$$\theta_{\max} = \inf_{x:h(x)<1} \frac{1}{1 - h(x)} = \frac{1}{1 - h(1)} \simeq 2.9218. \quad (4.3)$$

The density $f_{\theta_{\max}}$ is illustrated in Figure 3.

As depicted in Figure 3, the level set $\{x : h(x) \leq 1\}$ is a symmetric interval, approximately $|x| \leq 1.85$, while the minimum in (4.3) occurs at $x = \pm 1$, and a local maximum $h(0) \simeq 0.7979$ at $x = 0$. This is only partly typical for applied work in certain domains. The following example illustrates a broad family of symmetric distributions for which h is convex and the minimum occurs at the origin, implying $\theta_{\max} = 1/(1 - h(0))$.

Example 4.5. Signal-plus-noise model: Let $F_0 = N(0, 1)$, let G be a symmetric distribution on \mathbb{R} , and let $F_1 = G \star F_0$ be the convolution with density

$$\begin{aligned}f_1(y) &= \phi(y) \int_{\mathbb{R}} e^{yx - x^2/2} G(dx); \\ h(y) = f_1(y)/\phi(y) &= \int_{\mathbb{R}} e^{-x^2/2} \cosh(yx) G(dx).\end{aligned}$$

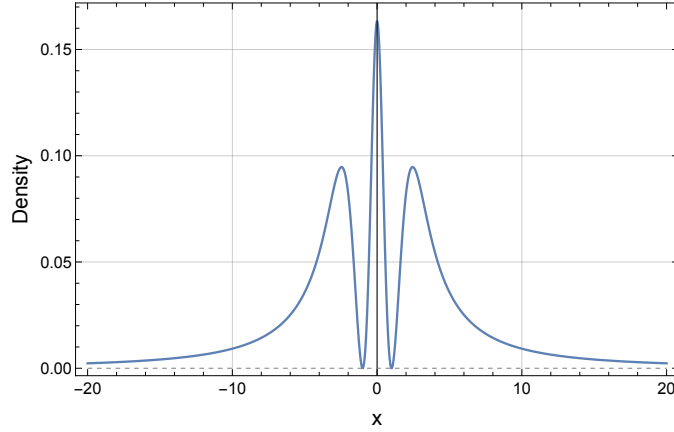


FIGURE 3. The density function $f_{\theta_{\max}}$ is symmetric and trimodal with zero density at ± 1 when f_1 is standard Cauchy and f_0 is standard normal. It is an extreme point relative to $N(0, 1)$ in the sense of (4.1).

Symmetry of G implies that $h(\cdot)$ is a positive combination of cosh-functions, and hence that $h(\cdot)$ is symmetric and convex with a minimum at the origin. Provided that $G \neq \delta_0$, the minimum is unique, and the value $h(0) = \int e^{-x^2/2} G(dx)$ is strictly less than one. The argument used in Example 4.4 implies $\theta_{\min} = 0$ and $\theta_{\max} = 1/(1 - h(0)) > 1$ for every symmetric convolution $F_1 = G \star N(0, 1)$. The upper extremity $F_{\theta_{\max}}$ has zero density at the origin.

That f_0 and f_1 are respectively extreme and non-extreme boundary points of the extended model (4.1) leads to the notably asymmetric boundary behaviour of the maximum likelihood estimator $\hat{\theta}_n$ when the data are independent draws from F_0 and F_1 respectively. Since the density ratio $h(X) = f_0(X)/f_1(X)$ has finite variance, it follows from classical results in the literature on boundary inference problems (see §2.4 for an overview) that the error probability $\mathbb{P}_1(\hat{\theta}_n < 1)$ tends to $1/2$ as $n \rightarrow \infty$. The behaviour under F_0 is non-standard because $\theta = 0$ is an extreme boundary point. As a result, the boundary behaviour $\mathbb{P}_0(\hat{\theta}_n > 0)$ is characterized by the tail probability of $h(X)$ induced by the upper tail of f_1 (see §2.3 and §2.5).

If the non-null component $F_1 = N(0, 1) \star G$ is a convolution with a symmetric signal distribution G , then F_1 has tails at least as heavy as those of G . We formalize this claim below.

Proposition 4.1. Suppose G is a symmetric distribution about zero with density $g(x)$. If $g(x)$ is regularly varying, then the standard normal convolution with density

$$m(y) = \int \phi(y - x)g(x)dx$$

is regularly varying with the same index. If $-\log g(x)$ is regularly varying with index 2κ , then

$$\log m(y) \geq -|y|^{2\kappa} L(y)$$

for some slowly varying L . Furthermore, if $\kappa < 1$, then $-\log m(y)$ is regularly varying with index 2κ .

Proof. First suppose $g(x)$ is regularly varying with index $\nu > 0$, i.e.

$$g(x) \sim x^\nu L(x)$$

for some slowly varying $L(x)$ as $x \rightarrow \infty$. Then for large $y > 0$ the convolution satisfies

$$\begin{aligned} m(y) &= \int g(y - \varepsilon) \phi(\varepsilon) d\varepsilon \\ &= \int_{|\varepsilon| < \sqrt{y}} g(y - \varepsilon) \phi(\varepsilon) d\varepsilon + O(e^{-y/2}) \sim g(y) \end{aligned}$$

since $y - \varepsilon = y(1 + o(1))$ over the range of the integral. Now suppose $-\log g(x)$ is regularly varying with index 2κ . By Jensen's inequality,

$$\log m(y) \geq \int \phi(\varepsilon) \log g(y - \varepsilon) d\varepsilon = - \int |y - \varepsilon|^{2\kappa} L(y - \varepsilon) \phi(\varepsilon) d\varepsilon.$$

Making the same restriction to $|\varepsilon| < \sqrt{y}$ as in the previous case and letting $y \rightarrow \infty$, we obtain the result. The proof that the inequality is tight when $\kappa < 1$ is recorded in Appendix IV. \square

It follows from the above result that if G has tails satisfying (2.5), then so does the convolution $G \star N(0, 1)$. Therefore, the distribution of the density ratio under the null remains in the Cauchy domain of attraction, implying that the type 1 error rates derived in Corollary 3.2 apply to the case where the non-null generator is a signal-plus-Gaussian noise convolution.

4.3. Likelihood-ratio statistic. The likelihood-ratio statistic for testing $\theta = 0$ is

$$\Lambda_n = 2 \sum_{i=1}^n \log(1 + \hat{\theta}_n(h(X_i) - 1)).$$

Our goal here is to establish the asymptotic null distribution, particularly the conditional distribution given $\hat{\theta}_n > 0$.

The conventional arguments based on Taylor expansion (see Appendices A.1 and A.3 of Brazzale and Mamei, 2024) do not apply because the r th log-likelihood derivative at zero is equal to $(-1)^{r-1}(r-1)! \sum_{i=1}^n (h(X_i) - 1)^r$, where $(h(X_i) - 1)^r$ belongs to the domain of attraction of the stable law with index $1/r$. The first derivative does not have a finite variance, so there is no concept of Fisher information. Higher-order derivatives do not have an expectation, and are strongly dependent. Nevertheless, one series of simulations shown in the right panel of Figure 4 suggests that the asymptotic null distribution given $\hat{\theta}_n > 0$ is close to $\kappa_n \chi_1^2$, where the Bartlett factor $\kappa_n \rightarrow 1$ tends to one. This section offers an explanation and a derivation of the correct limit distribution, which is not χ_1^2 in the setting of Theorem 3.1.

Theorem 4.1. Let X_1, \dots, X_n be independent standard Gaussian, let $h(X_i) \geq 0$ be in the domain of attraction of the maximally skew Cauchy law, and let $G(\cdot)$ be the cumulative distribution function whose u th quantile is

$$G^{-1}(u) = -2u - 2 \log(1 - u) = 2 \sum_{r \geq 2} u^r / r. \quad (4.4)$$

Then, the conditional limit distribution of the likelihood-ratio statistic is

$$\lim_{n \rightarrow \infty} \mathbb{P}_0(\Lambda_n \leq x \mid \Lambda_n > 0) = G(x).$$

Remark 4.1. The limit distribution is not dissimilar to χ_1^2 . Both densities behave like $x^{-1/2}$ near the origin. The first four cumulants of χ_1^2 are 1, 2, 8, and 96, while those of G are 1, 7/3, 32/3, and 3194/45. However, for $X \sim G$, the log density of $X^{1/2}$ has a Taylor expansion

$$\log(2xg(x^2)) = -\frac{2x}{3} - \frac{5x^2}{36} - \frac{23x^3}{810} - \frac{31x^4}{6480} + O(x^5),$$

which is essentially linear for $x < 1$, while the corresponding half-normal log density $\log(2/\pi)/2 - x^2/2$ is exactly quadratic and negative at zero. The difference between $2xg(x^2)$ and the degree-4 Taylor approximation is less than 1% for $x < 3$.

Proof. For notational simplicity we set $Z = h(X) - 1 \sim F$, so that F is in the Cauchy domain of attraction with zero mean, and Theorem 3.2 applies. Given $\bar{Z}_n > 0$, we construct an approximation for $l(\cdot)$ and its derivatives $l^{(r)}(0) = (-1)^{r-1}(r-1)!S_r$, where

$$S_r = \sum_{i=1}^n Z_i^r = \sum_{k \geq 0} Z_{(n-k)}^r = Z_{(n)}^r + \sum_{k \geq 1} Z_{(n-k)}^r.$$

Both $n\bar{Z}_n$ and $Z_{(n)}$ are conditionally $O_p(T_n)$, while the conditional distribution of $Z_{(n-1)}, Z_{(n-2)}, \dots$ is asymptotically the same as the unconditional distribution of $Z_{(n)}, Z_{(n-1)}, \dots$. Since Z_i^r is in the domain of attraction of the stable law with index $1/r$, the scaling constant is B_n^r , and we have

$$S_r - Z_{(n)}^r = \sum_{k \geq 1} Z_{(n-k)}^r = \begin{cases} n\bar{X}_n - Z_{(n)} & r = 1, \\ O_p(B_n^r) & r > 1. \end{cases}$$

It follows that $S_r = Z_{(n)}^r(1 + o_p(1))$ for $r \geq 2$.

Consider the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^n \log\{(1-\theta)f_0(X_i) + \theta f_1(X_i)\},$$

whose first two derivatives are given in (2.1). A first-order Taylor expansion around zero gives the local approximation

$$\tilde{l}(\theta) = \log(1 + \theta Z_{(n)}) + \theta(S_1 - Z_{(n)}), \quad (4.5)$$

where, in the Taylor expansion, we have substituted $Z_{(n)}$ for S_1 in the expression for $l'(0)$, and added and subtracted $\log(1 + \theta Z_{(n)}) = \theta Z_{(n)} + O(\theta^2)$. The approximation satisfies $l(\theta) - \tilde{l}(\theta) = o_p(1)$ for $\theta = O(T_n^{-1})$ owing to the form of the higher derivatives at zero.

Theorem 3.2 shows that $R = S_1/Z_{(n)}$ is less than one with high probability for large n , in which case \tilde{l} has a maximum at $\tilde{\theta}_n Z_{(n)} = R/(1-R)$. In that case $R = \tilde{\theta} Z_{(n)}/(1 + \tilde{\theta} Z_{(n)})$, and the approximate likelihood-ratio statistic

$$2\tilde{l}(\tilde{\theta}) = -2R - 2\log(1-R) = 2 \sum_{k \geq 2} R^k/k = \Lambda_n + o_p(1) \quad (4.6)$$

is a monotone function of R . Theorem 3.2 also shows that the limit distribution of R given $\bar{Z}_n > 0$ is uniform on $(0, 1)$, implying that $\Lambda_n \sim G$ in the limit. \square

Figure 4 shows a histogram of R and the Bartlett-adjusted likelihood ratio statistic $(\Lambda_n/\hat{\kappa}_n)^{1/2}$ using the exact likelihood and the exact maximum in the Gauss-Cauchy mixture model restricted to 2000 out of 74058 samples for which $\bar{Z}_n > 0$. For this model, $L(x) = (2 \log x)^{3/2}$ is the slow-variation function, $B_n \sim n/(2 \log n)^{3/2}$ and $T_n \sim 4\pi^{-1}B_n \log n$. For the simulation, $n = 10^7$, $\hat{\kappa}_n = 1.0170$ is the sample average, and $R > 1$ in 31 cases. The sample cumulant ratios $k_2/2$ and $k_3/8$ for $\Lambda_n/\hat{\kappa}_n$ compared with χ_1^2 are 1.160 and 1.346, which are surprisingly close to the theoretical limit values $7/6$ and $4/3$ respectively. The new limit density is shown on the same square-root scale, together with the half-normal density for comparison. While the difference between the two distributions is not large, the histogram clearly favours G . A standard 20-bin χ^2 -test gives $X^2 = 40.02$ for the χ_1^2 distribution, and $X^2 = 15.22$ for G .

Analogous simulations for the Gauss-Laplace mixture give virtually identical results for the likelihood-ratio statistic, though with $\hat{\kappa}_n = 0.9566$. However, the fraction of R -values greater than one was 249/2000, or 12%, consistent with a $(\log n)^{-1/2}$ convergence rate.

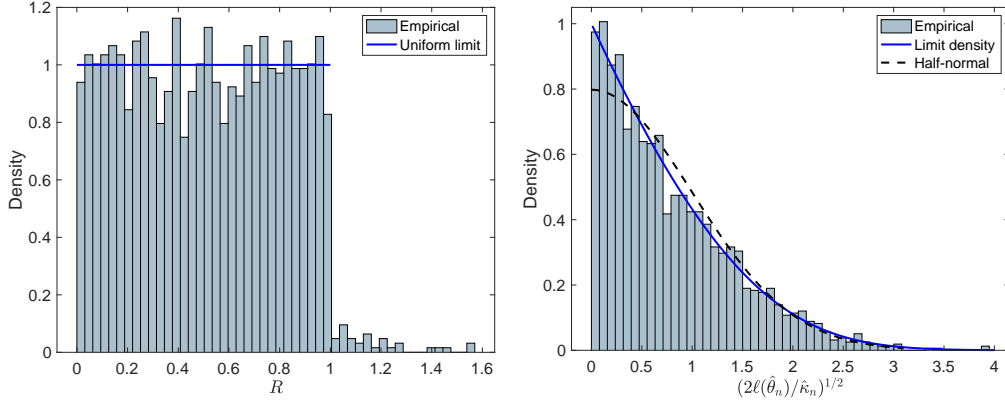


FIGURE 4. Histogram of $R = n\bar{Z}_n/Z_{(n)}$ (left) and $(2l(\hat{\theta}_n)/\hat{\kappa}_n)^{1/2}$ (right) for two thousand simulations of the Gauss-Cauchy mixture model with $n = 10^7$ observations restricted to samples for which $\bar{Z}_n > 0$.

The argument used in the proof of Theorem 4.1 suggests that the rate of convergence of Λ_n to G might be $(\log n)^{-1/2}$ or perhaps $(\log n)^{-1}$, but the simulations suggest a faster rate, particularly after Bartlett correction.

The local approximation (4.5) and ensuing argument is valid for any iid problem with a boundary at zero provided that the log likelihood derivative is a sum of zero mean random variables in the Cauchy domain of attraction. There may therefore be other statistical problems besides the two-component mixture problem in which the conditional limit distribution (4.4) arises; the only difficulty in establishing that $l(\theta) - \tilde{l}(\theta) = o_p(1)$, as the adequacy of the approximation (4.5) depends on the behaviour of higher derivatives, which is context specific.

4.4. Predictive activity rate. Suppose we reject the global null $X_i \sim \mathbb{P}_0$ in favor of a two-component mixture with extreme component f_1 , as if we knew the data were drawn from a latent variable model:

$$\begin{aligned} \mathcal{A}_i &\sim \text{Bernoulli}(\theta) \\ X_i \mid \mathcal{A}_i &\sim f_{\mathcal{A}_i} \quad i = 1, \dots, n \end{aligned} \tag{4.7}$$

independently, concluding that $\theta > 0$. Then it becomes natural to ask which observations arose from the f_1 component. To answer this question, we may compute for each observation X_i the local activity rate, defined

$$\mathbb{P}_\theta(\mathcal{A}_i = 1 \mid X_i) := \theta f_1(X_i) / f(X_i),$$

which measures for each observation how likely it is to be a draw from f_1 in the hierarchical model (4.7). In this section, we derive the asymptotic null distribution of the local activity rate for the observation with largest absolute value.

Given an estimate of θ , the fitted or predictive activity rate for unit i is

$$\mathbb{P}_{\hat{\theta}_n}(\mathcal{A}_i = 1 \mid X_i) := \frac{\hat{\theta}_n f_1(X_i)}{(1 - \hat{\theta}_n) f_0(X_i) + \hat{\theta}_n f_1(X_i)} = \frac{\hat{\theta}_n h(X_i)}{1 - \hat{\theta}_n + \hat{\theta}_n h(X_i)}. \tag{4.8}$$

If $\hat{\theta}_n = 0$, the fitted local activity rate is zero for every unit: every unit is deemed to have a null or negligible signal. Given $\hat{\theta}_n > 0$, the maximum predicted activity rate is

$$\frac{\hat{\theta}_n h(X_{\max})}{1 - \hat{\theta}_n + \hat{\theta}_n h(X_{\max})} = R + o_p(1)$$

where X_{\max} is the maximum absolute sample value, and $R := n\bar{Z}_n/Z_{(n)}$, where $Z_i := h(X_i)$. According to Theorem 3.2, R is uniformly distributed on $(0, 1)$ in the large-sample limit when $\theta = 0$. In this case, we also have $\hat{\theta}_n = 0$ with probability tending to 1. That is, the type 1 error of the MLE for testing the global null $H_0 : \theta = 0$ tends to zero, and conditionally on the type 1 error event $\{\hat{\theta}_n > 0\}$, the smallest fitted activity rate is an asymptotically valid p -value for testing H_0 . We record this implication as a corollary of Theorem 3.2 and Theorem 4.1 below.

Corollary 4.1. Consider the decision rule that rejects $H_0 : \theta = 0$ if the largest local activity rate exceeds $1 - \alpha$. Then the conditional probability that some unit is declared active or non-null at level α is equal to α in the large-sample limit. In other words,

$$\lim_{n \rightarrow \infty} \mathbb{P}_0 \left(\min_{i=1, \dots, n} \mathbb{P}_{\hat{\theta}_n}(\mathcal{A}_i = 0 \mid X_i) \leq \alpha \mid \hat{\theta}_n > 0 \right) = \alpha,$$

for any $\alpha \in [0, 1]$ where

$$\mathbb{P}_{\hat{\theta}_n}(\mathcal{A}_i = 0 \mid X_i) := \frac{1 - \hat{\theta}_n}{1 - \hat{\theta}_n + \hat{\theta}_n h(X_i)}.$$

Proof. Following the proof of Theorem 4.1, we have $R = \hat{\theta}_n h(X_{\max}) / (1 + \hat{\theta}_n h(X_{\max}))$. Together with Theorem 3.2, which implies that $h(X_{\max}) \rightarrow \infty$ and $\hat{\theta}_n \rightarrow 0$ in probability conditionally on $\hat{\theta}_n > 0$, we have

$$\begin{aligned} \frac{\hat{\theta}_n h(X_{\max})}{1 - \hat{\theta}_n + \hat{\theta}_n h(X_{\max})} &= R \times \frac{1 + \hat{\theta}_n h(X_{\max})}{\hat{\theta}_n h(X_{\max})} \times \frac{\hat{\theta}_n h(X_{\max})}{1 - \hat{\theta}_n + \hat{\theta}_n h(X_{\max})} \\ &= R \times \left(1 - \frac{\hat{\theta}_n}{1 + \hat{\theta}_n h(X_{\max})} \right)^{-1} = R (1 + o_p(1)). \end{aligned}$$

It follows that the complement is also Uniform(0, 1) distributed asymptotically, conditional on $\hat{\theta}_n > 0$. \square

4.5. Wald and Rao statistics. The conventional Wald and Rao statistics for testing $\theta = 0$ are

$$\hat{\theta}(i(\hat{\theta}))^{1/2}, \quad \text{and} \quad \frac{l'(0)}{i(0)^{1/2}},$$

where $l(\theta)$ is the log likelihood function and $i(\theta)$ is the Fisher information. Technically speaking, the Rao statistic does not exist in the mixture setting because $i(0)$ is not finite. For present purposes, however, we substitute $-l''(\theta)$ for $i(\theta)$ in both.

Given $\hat{\theta} > 0$, the approximate likelihood in section 4.3 implies

$$\begin{aligned} \tilde{l}'(\theta) &= \frac{Z_{(n)}}{1 + \theta Z_{(n)}} + n\bar{Z}_n - Z_{(n)}, \\ Z_{(n)}\tilde{\theta} &= R/(1 - R), \\ -\tilde{l}''(0) &= Z_{(n)}^2, \\ -\tilde{l}''(\tilde{\theta}) &= Z_{(n)}^2(1 - R)^2 = (Z_{(n)} - n\bar{Z}_n)^2, \end{aligned}$$

where $R = n\bar{Z}_n/Z_{(n)}$. It follows that the modified Wald and Rao statistics are

$$\begin{aligned} \tilde{\theta}(-\tilde{l}''(\tilde{\theta}))^{1/2} &= \frac{R}{(1 - R)Z_{(n)}} Z_{(n)}(1 - R) = R \\ \frac{\tilde{l}'(0)}{(-\tilde{l}''(0))^{1/2}} &= \frac{n\bar{Z}_n}{Z_{(n)}} = R. \end{aligned}$$

In other words, the two statistics are conditionally equivalent given $\hat{\theta} > 0$, both uniformly distributed on $(0, 1)$. The squared statistics are not approximately the same as the likelihood-ratio statistic, although all three are asymptotically equivalent in the sense that each is a monotone function of R . Conventional standard normal approximations are incorrect.

5. COMPOSITE MIXTURES

5.1. Definition. A composite Gaussian mixture is a family of distributions

$$\{F_\theta = (1 - \theta)F_0 + \theta F_1 : 0 \leq \theta \leq 1, F_1 \in \mathcal{F}\}$$

in which $F_0 = N(0, 1)$. The non-null distributions belong to some specified non-empty family \mathcal{F} , which need not be convex. Models of this type have been considered by Efron et al. (2001), Patra and Sen (2016). In the modern style, these are invariably nonparametric in the counter-semantic sense that \mathcal{F} is not a finite-dimensional manifold.

The emphasis in this section is on composite mixtures arising in the signal-detection setting where each distribution is necessarily a Gaussian convolution $F_\theta = N(0, 1) \star P$ with some symmetric signal distribution P . This condition is sufficient to ensure that each F_θ has a bounded continuous density $f_\theta(x) \leq \phi(0)$, and hence that each distribution in \mathcal{F} also has a bounded continuous density. Also, under suitable conditions (Proposition 4.1), the tails of F_θ are similar to those of P .

5.2. Equivalence classes. Let F_1, F_2 be two symmetric distributions having bounded continuous densities such that the ratio has a finite non-zero limit:

$$\lim_{x \rightarrow \infty} \frac{f_1(x)}{f_2(x)} = K. \quad (5.1)$$

Then $h_1(x) \sim Kh_2(x)$, where $h_r(x) = f_r(x)/\phi(x)$ is the density ratio arising in the elementary mixture model $(1 - \theta)\phi(x) + \theta f_r(x)$. On the assumption that $h_1(X)$ (and hence also $h_2(X)$), belongs to the Cauchy domain of attraction, it follows that

$$\begin{aligned} \frac{1}{B_{1,n}} \sum (h_1(X_i) - 1) &= A_{1,n} + \varepsilon_1 + o_p(1), \\ \frac{1}{B_{2,n}} \sum (h_2(X_i) - 1) &= A_{2,n} + \varepsilon_2 + o_p(1), \end{aligned}$$

where the stabilizing sequences satisfy

$$B_{1,n} \sim KB_{2,n}, \quad A_{2,n} \sim A_{1,n}, \quad T_{1,n} \sim KT_{2,n}.$$

If $\hat{\theta}_{1,n} > 0$, with high \mathbb{P}_0 -probability $h_1(X_{(n)}) > T_{1,n}$ implying $h_2(X_{(n)}) > T_{2,n}$, and hence $\hat{\theta}_{2,n} > 0$ with high probability. Moreover, given $\hat{\theta}_{1,n} > 0$, the ratios satisfy

$$\frac{1}{h_1(X_{(n)})} \sum (h_1(X_i) - 1) = \frac{1}{h_2(X_{(n)})} \sum (h_2(X_i) - 1) + o_p(1).$$

It follows that the likelihood-ratio statistics $\Lambda_{1,n}, \Lambda_{2,n}$ are asymptotically equal:

$$\lim_{n \rightarrow \infty} \mathbb{P}_0 \left(|\Lambda_{1,n} - \Lambda_{2,n}| > \epsilon \mid \hat{\theta}_n > 0 \right) = 0$$

for every $\epsilon > 0$. For example, if f_1, f_2 are Student t distributions on ν degrees of freedom with scale parameters σ_1, σ_2 , then (5.1) is satisfied with $K = (\sigma_1/\sigma_2)^\nu$, implying that the likelihood ratio statistics are asymptotically equal given $\hat{\theta}_n > 0$.

5.3. Implications. The first implication is as follows. Let $\mathcal{E}(F_1)$ be the set of distributions that are tail-equivalent to F_1 in the sense of (5.1), let $\Lambda_n(F_1)$ be the likelihood-ratio statistic in the elementary mixture $\text{conv}(F_0, F_1)$, and let $\Lambda_n(\hat{F})$ be the maximized likelihood-ratio statistic over θ and over the class $\mathcal{E}(F_1)$. Given that $\hat{\theta}_n > 0$, the difference $\Lambda_n(F) - \Lambda_n(F_1)$ is $o_p(1)$ for an arbitrary F in $\mathcal{E}(F_1)$. This implies that $\Lambda_n(\hat{F}) - \Lambda_n(F_1)$ is $o_p(1)$ so that the conditional distribution of the maximized likelihood-ratio statistic is

$$\lim_{n \rightarrow \infty} \mathbb{P}_0(\Lambda_n(\hat{F}) < x \mid \hat{\theta}_n > 0) = G(x).$$

From the perspective of statistical modelling and testing, no advantage is gained by extending the single distribution F_1 to the nonparametric composite mixture with $\mathcal{F} = \mathcal{E}(F_1)$.

The second implication is that the conditional limit distribution of the maximized likelihood-ratio statistic $\Lambda_n(\hat{F})$ given $\hat{\theta}_n > 0$, may depend on the topology of the quotient space \mathcal{F}/\mathcal{E} , i.e., the space of equivalence classes. In particular, if \mathcal{F}/\mathcal{E} is a finite-dimensional manifold of dimension d , the limit distribution G_d may depend on d . Otherwise, \mathcal{F} plays no role in the limit distribution.

There are some parallels between tail-equivalence as defined by (5.1), and sparse-equivalence as defined by McCullagh and Polson (2018). Signal distributions P and P' within the same sparse-equivalence class are statistically indistinguishable based on observations from the convolutions $P \star N(0, 1)$ and $P' \star N(0, 1)$, and have the same sparsity implications in the relevant statistical sense. Formally, two signal distributions P, P' are termed sparse-equivalent if their normalized exceedance measures are equal: $H(dx) = H'(dx)$. Two sparse signals having exceedance measures $H \neq H'$ give rise to mixtures that are tail-equivalent in the sense of (5.1) if the tail ratio has a finite non-zero limit $H(x^+)/H'(x^+) \sim K$. Thus sparse equivalence is finer than tail equivalence. Sparse equivalence is relevant for identifiability of mixtures generated by Gaussian convolution with a sparse signal; tail equivalence is relevant in the present setting where the focus is exclusively on the boundary behaviour of the likelihood-ratio and related statistics.

5.4. Example. For $0 < \nu < 2$, let $\zeta_\nu(x)$ be the inverse-power zeta-function defined in McCullagh and Polson (2018) and expressed as a convergent power series

$$\zeta_\nu(x) = \frac{\nu(2-\nu)}{\Gamma(2-\nu/2)} \sum_{r=1}^{\infty} \frac{2^{r-2} \Gamma(r-\nu/2) x^{2r}}{(2r)!}.$$

The product $\psi_\nu(x) = \phi(x)\zeta_\nu(x)$ is the density of a symmetric bimodal distribution Ψ_ν whose tails are regularly-varying with index $-\nu$. For purposes of likelihood maximization, it is convenient to extend the model by continuity to the upper boundary: $\zeta_2(x) = \lim_{\nu \rightarrow 2} \zeta_\nu(x) = x^2$ and $\psi_2(x) = x^2\phi(x)$.

One crucial feature of this family is that $\psi_\nu(X)$ belongs to the skew-Cauchy domain of attraction in all cases except $\nu = 2$, which belongs to the Gaussian domain. In the non-null sparse-signal-plus-noise setting, the maximum of the likelihood tends to occur near the upper boundary $\nu = 2$ only if the signal has finite variance; otherwise, if the variance is not finite, the maximum occurs at an interior point with high probability. To understand this phenomenon, observe that if Q_σ is symmetric with atoms at $\pm\sigma$, the convolution $Q_\sigma \star N(0, 1)$ has density

$$\begin{aligned} \frac{1}{2}\phi(x-\sigma) + \frac{1}{2}\phi(x+\sigma) &= \phi(x) e^{-\sigma^2/2} \cosh(\sigma x), \\ &= \phi(x) (1 - \sigma^2/2 + \sigma^2 x^2/2 + o(\sigma^2)). \end{aligned}$$

Every symmetric distribution P is a Q_σ scale mixture. If P has finite variance σ^2 , then $P \star N(0, 1)$ has exactly the same first-order small- σ expansion as Q_σ . In other words, every

symmetric finite-variance convolutional perturbation $P \star N(0, 1)$ is first-order equivalent to the top mixture $(1 - \theta + \theta x^2)\phi(x)$ with $\theta = \sigma^2/2$. In that sense, the class of finite-variance signals is reduced to a single mixture representative Ψ_2 .

The following theorem addresses the left-boundary behaviour of likelihood-based statistics for a composite mixture that includes a subset of the distributions Ψ_ν .

Theorem 5.1. Let $0 < \tau \leq 2$ be given, and let $\hat{\theta}_n, \hat{\nu}_n$ be the maximum-likelihood estimate in the composite mixture with $\mathcal{F} = \{\Psi_\nu : 0 < \nu \leq \tau\}$. Then, the large-sample boundary behaviour is as follows:

$$\begin{aligned} \mathbb{P}_0(\hat{\theta}_n > 0) &\sim \begin{cases} \tau/(2 \log n) & \tau < 2, \\ 1/2 & \tau = 2; \end{cases} \\ \lim_{n \rightarrow \infty} \mathbb{P}_0(\hat{\nu}_n = \tau \mid \hat{\theta}_n > 0) &= 1; \\ \lim_{n \rightarrow \infty} \mathbb{P}_0(\Lambda_n \leq x^2 \mid \hat{\theta}_n > 0) &= \begin{cases} G(x^2) & \tau < 2, \\ 1 - 2\bar{\Phi}(x) & \tau = 2. \end{cases} \end{aligned}$$

Remark 5.1. At the critical boundary $\theta = 0$, the parameter ν is indeterminate, which is a type of singularity that occurs in many non-regular problems. For $\hat{\theta}_n > 0$, the mixture likelihood has its maximum at the upper extremity $\hat{\nu} = \tau$ with high \mathbb{P}_0 -probability because these are the distributions that are closest to Gaussian in the natural metric. The limit behaviour for the composite mixture follows from the top elementary mixture $(1 - \theta)\phi + \theta\psi_\tau$.

Proof. Let $\hat{\theta}_\nu$ denote the maximum likelihood estimator of θ at a particular value of ν . Thus, $\hat{\theta}_n = \hat{\theta}_{\hat{\nu}}$, where $\hat{\nu} = \hat{\nu}_n$ is the maximum likelihood solution. The proof will establish containment on both sides of $\mathcal{E}_{\hat{\nu}} = \{\hat{\theta}_{\hat{\nu}} > 0\}$ and $\mathcal{E}_\tau = \{\hat{\theta}_\tau > 0\}$ modulo subsets of asymptotic probability 0 under \mathbb{P}_0 , i.e.

$$\lim_{n \rightarrow \infty} \mathbb{P}_0(\mathcal{E}_{\hat{\nu}} \setminus \mathcal{E}_\tau) = \lim_{n \rightarrow \infty} \mathbb{P}_0(\mathcal{E}_\tau \setminus \mathcal{E}_{\hat{\nu}}) = 0.$$

Since $\psi_\nu(x) \sim K_\nu/x^{\nu+1}$ for large x and $\nu < 2$, we have

$$\zeta_\nu(x) = \frac{\psi_\nu(x)}{\phi(x)} \sim \frac{\sqrt{2\pi}e^{x^2/2}K_\nu}{x^{\nu+1}},$$

and from the argument of §2.5, the density ratio satisfies

$$\mathbb{P}_0(\zeta_\nu(X) > \eta) \sim \frac{2\psi_\nu(\sqrt{2 \log \eta})}{\eta \sqrt{2 \log \eta}} \sim \frac{2K_\nu}{\eta (2 \log \eta)^{\nu/2+1}}.$$

Although the constants do not matter in this argument, we write the above upper probability, for consistency with §3, in the form $\bar{F}(\eta) \sim 2C_1/\eta L(\eta)$. Thus, from Corollary 3.1,

$$\begin{aligned} L(\eta) &= (2 \log \eta)^{\nu/2+1}/(\pi K_\nu) \sim 1/L^\dagger(\eta) \\ B_n \sim nL^\dagger(n) &\sim \frac{n\pi K_\nu}{(2 \log n)^{\nu/2+1}}, \\ A_n &\sim -4 \log n/(\pi \nu), \\ T_n = B_n|A_n| &\sim \frac{2nK_\nu}{\nu(2 \log n)^{\nu/2}}. \end{aligned}$$

For sufficiently large n , Appendix V shows that $\hat{\theta}_n > 0$ is equivalent to

$$\lim_{\theta \rightarrow 0} \frac{d}{d\theta} \ell(\theta; \hat{\nu}(\theta)) = \lim_{\theta \rightarrow 0} \sum_{i=1}^n (\zeta_{\hat{\nu}(\theta)}(X_i) - 1) > 0, \quad (5.2)$$

where $\ell(\theta; \hat{\nu}(\theta))$ is the profile log-likelihood function.

In view of this expression, consider the event $n^{-1} \sum \zeta_\nu(X_i) > 1$, which corresponds to $\hat{\theta}_\nu > 0$ for a given ν . For large n , this is equivalent to $\zeta_\nu(X_{(n)}) > \lambda T_n$ for some $\lambda > 1$, by Theorem 3.2 and ultimate monotonicity of ζ_ν . The condition $\zeta_\nu(X_{(n)}) > \lambda T_n$ is equivalent to

$$\begin{aligned} X_{(n)}^2 &> 2 \log(T_n X_{(n)}^{1+\nu}) + \text{const.} \\ &= 2 \log n - \nu \log(2 \log n) - 2 \log \nu + 2(1+\nu) \log X_{(n)} + \text{const.} \\ &= 2 \log n + \nu(2 \log X_{(n)} - \log(2 \log n)) - 2 \log \nu + 2 \log X_{(n)} + \text{const.} \end{aligned} \quad (5.3)$$

It follows from calculus that the right hand side above is non-increasing in ν over the interval

$$0 < \nu \leq \left(\log \frac{X_{(n)}}{\sqrt{2 \log n}} \right)^{-1} \quad (5.4)$$

and is otherwise increasing. Since $\tau \leq 2$, it suffices to show that, conditionally on $\hat{\theta}_n > 0$, $X_{(n)}/\sqrt{2 \log n}$ is bounded above by a sequence tending to 1, implying that (5.3) is decreasing in ν over $(0, \tau]$. To this end, define

$$\omega := \min\{x > 0 : x^2 \geq 2 \log n + \nu(2 \log x - \log(2 \log n)) - 2 \log \nu + 2 \log x\}.$$

First, we claim that for sufficiently large n , $\omega \leq \sqrt{2 \log n + 1.01 \log(2 \log n)}$. Indeed, plugging $x = \sqrt{2 \log n + 1.01 \log(2 \log n)}$ into the condition defining ω , we see it is satisfied:

$$\begin{aligned} &2 \log n + \nu[\log(2 \log n + 1.01 \log(2 \log n)) - \log(2 \log n)] - 2 \log \nu + \log(2 \log n) + o(1) \\ &\sim 2 \log n + (1 + o(1)) \log(2 \log n) \leq x^2 = 2 \log n + 1.01 \log(2 \log n), \end{aligned}$$

for n larger than a universal constant. Next, we will show that for some $c > 0$,

$$\mathbb{P}_0(X_{(n)} > \sqrt{2 \log n + c \log \log n} \mid X_{(n)} > \omega) \rightarrow 0.$$

Since $\omega \leq \sqrt{2 \log n + 1.01 \log(2 \log n)}$, the left hand side above is bounded by

$$\leq \frac{\mathbb{P}_0(X_{(n)} > \sqrt{2 \log n + c \log \log n})}{\mathbb{P}_0(X_{(n)} > \sqrt{2 \log n + 1.01 \log(2 \log n)})} \rightarrow 0,$$

which follows from Mills's ratio, for any $c > 1.01$, since $X_{(n)}$ is a maximum of n iid standard normal random variables. This shows that conditional on $\hat{\theta}_n > 0$, we have $X_{(n)} \leq \sqrt{2 \log n + c \log \log n}$ with probability tending to 1, so that the right hand side of (5.4) diverges to $+\infty$. Since the threshold is decreasing as a function of ν , the set of ν -values for which $\hat{\theta}_\nu$ is positive, i.e., the set of ν -values for which $\zeta_\nu(X_{(n)}) > T_n$, must be an upper interval $(\nu_0, \tau]$, which is empty if the threshold for $\lambda = 1$ is not exceeded at $\nu = \tau$. Thus, for sufficiently large n , \mathcal{E}_τ implies $\mathcal{E}_{\hat{\nu}}$. Conversely, given that the threshold is exceeded on the event $\mathcal{E}_{\hat{\nu}}$, the approximate likelihood-ratio statistic (4.6) is a monotone function of the ratio

$$\frac{1}{1-R} \simeq \frac{h(X_{(n)})}{T_n} \simeq \frac{\sqrt{2\pi} e^{X_{(n)}^2/2}}{2n X_{(n)}} \left(\frac{2 \log n}{X_{(n)}^2} \right)^{\nu/2} \times \nu.$$

For large n and $X_{(n)}^2 \sim 2 \log n$, the ratio increases linearly as a function of ν , which implies $\hat{\nu} = \tau$ whenever $\hat{\theta}_n > 0$. For sufficiently large n , $\mathcal{E}_{\hat{\nu}}$ implies \mathcal{E}_τ , and \mathcal{E}_τ implies $\mathcal{E}_{\hat{\nu}}$, thus the events are asymptotically equal under \mathbb{P}_0 . For $\tau < 2$, and by continuity for $\tau \leq 2$, the boundary behaviour is governed by the top elementary mixture with $\nu = \tau$. \square

Corollary 5.1. The same argument applies if we replace \mathcal{F} with the set of Student t distributions on $\nu \leq \tau$ degrees of freedom with scale parameter σ . In this one-dimensional

composite mixture, σ is fixed, $\tau < \infty$ is any positive real number, and the limit behavior exhibits no discontinuity at $\tau = 2$ or elsewhere. Section 5.2 implies that the scale parameter is nugatory for boundary behaviour. This argument fails for $\tau = \infty$.

Reproducibility. Code to reproduce all figures in the paper can be found at: <https://www.ma.imperial.ac.uk/~hbattey/BMX.html>.

APPENDICES

Appendix I: De Bruijn group. The de Bruijn group is the set (SV, \diamond) of slowly-varying functions together with the non-commutative binary operation $(L_1 \diamond L_2)(x) = L_1(x) L_2(x L_1(x))$. To see that this is an associative function $SV^2 \rightarrow SV$, observe that

$$\begin{aligned} (L_1 \diamond (L_2 \diamond L_3))(x) &= L_1(x) \times (L_2 \diamond L_3)(x L_1(x)), \\ &= L_1(x) \times L_2(x L_1(x)) \times L_3(x L_1(x) L_2(x L_1(x))); \\ ((L_1 \diamond L_2) \diamond L_3)(x) &= (L_1 \diamond L_2)(x) \times L_3(x (L_1 \diamond L_2)(x)), \\ &= L_1(x) \times L_2(x L_1(x)) \times L_3(x L_1(x) L_2(x L_1(x))), \\ &= (L_1 \diamond (L_2 \diamond L_3))(x). \end{aligned}$$

Associativity implies that the triple product is well-defined by the pairwise products. The identity element: $L \diamond 1 = 1 \diamond L = L$ is the unit constant function. If it exists, the inverse is a slowly-varying function L^\dagger such that $(L \diamond L^\dagger)(x) = (L^\dagger \diamond L)(x) = 1$ for all $x > 0$. Since slow variation is a characterization of the limiting behaviour for large x , it says little about the behaviour for general x beyond continuity or measurability. Thus, the existence of an inverse is not guaranteed. Nevertheless, the following theorem suffices for present purposes.

Theorem A.1 (de Bruijn, 1959). *To each $L \in SV$ there corresponds a function $L^\dagger \in SV$, satisfying $L(x)L^\dagger(xL(x)) = L^\dagger(x)L(xL^\dagger(x)) = 1$ for all x sufficiently large.*

Proof. Theorem 1.5.13 of Bingham, Goldie and Teugels (1987) proves existence in the sense of equivalence. Theorem 1.8.9 proves existence as stated above, i.e., for all x sufficiently large. \square

Although the product $xL(x)$ is ultimately monotone, it is not necessarily monotone for small x . Thus, SV contains functions for which no group inverse exists (as a function $(0, \infty) \rightarrow (0, \infty)$). The theorem states that an asymptotic inverse exists, which implies that the set of equivalence classes $(SV, \diamond)/\sim$ is a group. For the most part, it is the group of equivalence classes that is of interest here. However, we always work with a representative element.

Appendix II: Approximation of the sine-integral. Let F be the cumulative function of a finite-mean distribution on the positive real line, and let $\bar{F}(x) = 1/(xL(x))$ be the right-tail probability. Since $L(\lambda x)/L(x) \rightarrow 1$ uniformly in λ as $x \rightarrow \infty$, the derivative $xL'(x)/L(x)$ (with respect to λ at $\lambda = 1$) tends to zero. Hence

$$dF(x) = -d\bar{F}(x) = \left(1 + \frac{xL'(x)}{L(x)}\right) \frac{dx}{x^2 L(x)} \sim \frac{dx}{x^2 L(x)}.$$

We first split the integrand into two parts, $\sin(tx) = tx + (\sin tx - tx)$, and split the range into disjoint intervals $(0, T)$ and (T, ∞) , where $T = 1/|t|$ is large. Then

$$\int_0^\infty \sin(tx) dF(x) = t\mu + \int_0^T (\sin tx - tx) dF(x) + \int_T^\infty (\sin(tx) - tx) dF(x), \quad (.5)$$

where μ is the mean. The first integrand is bounded by $|\sin(tx) - tx| \leq |t|^3 x^3$ for $x \leq T$, so the contribution from the interval $(0, T)$ is bounded by

$$\begin{aligned} |t|^3 \int_0^T x^3 dF(x) &= |t|^3 \int_0^T x^3 \left(1 + \frac{xL'(x)}{L(x)}\right) \frac{dx}{x^2 L(x)}, \\ &\sim |t|^3 \int_0^T x^3 \frac{dx}{x^2 L(x)}, \\ &\leq |t|^3 T \frac{T}{L(T)} = \frac{1}{TL(T)}. \end{aligned}$$

Since $|\sin(tx)| \leq 1$, the first part of the second integral in (.5) has the same bound:

$$\left| \int_T^\infty \sin(tx) dF(x) \right| \leq \int_T^\infty dF(x) = \bar{F}(T) = \frac{1}{TL(T)}.$$

It follows that

$$\begin{aligned} \int_0^\infty \sin(tx) dx &= t\mu - t \int_T^\infty x dF(x) + O\left(\frac{t}{L(T)}\right), \\ &= t\mu - t \int_T^\infty \frac{dx}{xL(x)} + O\left(\frac{t}{L(T)}\right). \end{aligned}$$

Appendix III: Calculation of the asymmetric Cauchy density. By the Fourier inversion formula of the characteristic function for $\alpha = 1$ and $|\beta| \leq 1$, the asymmetric Cauchy density $f(x)$ in Fig. 1 is

$$f(x) = \frac{1}{\pi} \int_0^\infty e^{-t} \cos(tx + 2\beta t \log t / \pi) dt.$$

Substitute $u = g_x(t) := t(1 + \frac{2\beta}{\pi x} \log t)$, which is increasing over $(\exp(-\frac{\pi x}{2\beta}), \infty)$. As $x > 0$ becomes large, the integral over the interval $(0, \exp(-\frac{\pi x}{2\beta}))$ is exponentially small in x , and may be ignored:

$$\begin{aligned} \int_0^\infty e^{-t} \cos(xg_x(t)) dt &= \int_{\exp(-\frac{\pi x}{2\beta})}^\infty e^{-t} \cos(xg_x(t)) dt + O(e^{-\frac{\pi x}{2\beta}}), \\ &= \int_0^\infty e^{-g_x^{-1}(u)} \cos(xu) \frac{1}{1 + \frac{2\beta \log g_x^{-1}(u)}{\pi x} + \frac{2\beta}{\pi x}} du + O(e^{-\frac{\pi x}{2\beta}}), \end{aligned}$$

where $g_x^{-1} : (0, \infty) \rightarrow (\exp(-\frac{\pi x}{2\beta}), \infty)$ is the inverse map for the substitution. From here on, we ignore the exponentially small remainder term. It is easy to check that

$$\frac{u}{1 + \frac{2\beta}{\pi x} \log u} < g_x^{-1}(u) < u,$$

from which it follows that the integral is

$$= \int_{e^{-\frac{x}{\log x}}}^{\sqrt{x}} e^{-u} \cos(xu) \frac{1}{1 + \frac{2\beta \log u}{\pi x} + \frac{2\beta}{\pi x}} du (1 + o(1)) + O(e^{-\sqrt{x}}),$$

where $o(1)$ means a term that goes to zero as $x \rightarrow \infty$. Then the above is

$$\begin{aligned} &\sim \int_{e^{-\frac{x}{\log x}}}^{\sqrt{x}} e^{-u} \cos(xu) \left(1 - \frac{2\beta \log u}{\pi x}\right) du \\ &= \int_{e^{-\frac{x}{\log x}}}^{\sqrt{x}} e^{-u} \cos(xu) du - \frac{2\beta}{\pi x} \int_{e^{-\frac{x}{\log x}}}^{\sqrt{x}} e^{-u} \cos(xu) \log(u) du. \end{aligned}$$

Now using $\cos(xu) = \operatorname{Re}(e^{ixu})$, the above expression becomes

$$\begin{aligned} &= \int_{e^{-\frac{x}{\log x}}}^{\sqrt{x}} e^{-u} \operatorname{Re}(e^{ixu}) du - \frac{2\beta}{\pi x} \int_{e^{-\frac{x}{\log x}}}^{\sqrt{x}} e^{-u} \operatorname{Re}(e^{ixu}) \log(u) du \\ &= \operatorname{Re} \left(\frac{1}{1-ix} \int_{e^{-\frac{x}{\log x}(1-ix)}}^{\sqrt{x}(1-ix)} e^{-v} dv \right) - \operatorname{Re} \left(\frac{2\beta}{\pi x} \int_{e^{-\frac{x}{\log x}}}^{\sqrt{x}} e^{-u(1-ix)} \log(u) du \right). \end{aligned}$$

The first term gives

$$\operatorname{Re} \left(\frac{1}{1-ix} \int_{e^{-\frac{x}{\log x}(1-ix)}}^{\sqrt{x}(1-ix)} e^{-v} dv \right) \sim \operatorname{Re} \left(\frac{1}{1-ix} \right) = \frac{1}{1+x^2}.$$

The second term is

$$\begin{aligned} &= -\operatorname{Re} \left(\frac{2\beta}{\pi x} \int_{e^{-\frac{x}{\log x}(1-ix)}}^{\sqrt{x}(1-ix)} \frac{e^{-v}(\log(v) - \log(1-ix))}{1-ix} dv \right) \\ &= O(x^{-3} \log x) + \underbrace{\frac{2\beta}{\pi x} \operatorname{Re} \left(\frac{1}{1-ix} \int_{e^{-\frac{x}{\log x}(1-ix)}}^{\sqrt{x}(1-ix)} e^{-v} i \sin^{-1} \left(\frac{-x}{\sqrt{1+x^2}} \right) dv \right)}_{(*)}, \end{aligned}$$

since $1-ix = (1+x^2)^{1/2} e^{i \sin^{-1}(-x/\sqrt{1+x^2})}$. Finally, note that the above term $(*)$ is equal to

$$= \frac{2\beta}{\pi x} \operatorname{Re} \left(\frac{1+ix}{1+x^2} \int_{e^{-\frac{x}{\log x}(1-ix)}}^{\sqrt{x}(1-ix)} e^{-v} i \sin^{-1} \left(\frac{-x}{\sqrt{1+x^2}} \right) dv \right) \sim \frac{2\beta}{\pi x} \cdot \frac{\pi}{2x} = \frac{\beta}{x^2}.$$

To summarize, we have shown that for large $x > 0$,

$$f(x) \sim \frac{1}{\pi x^2} + \frac{\beta}{\pi x^2} = \frac{1+\beta}{\pi x^2},$$

which implies $\bar{F}(x) \sim (1+\beta)/(\pi x)$ for the right tail. For $x < 0$,

$$f(x) = \frac{1}{\pi} \int_0^\infty \cos(tx + 2\beta t \log t/\pi) dt = \frac{1}{\pi} \int_0^\infty \cos(t|x| - 2\beta t \log t/\pi) dt,$$

so the argument above implies that as $x \rightarrow -\infty$,

$$f(x) \sim \frac{1-\beta}{\pi x^2},$$

and $F(x) \sim (1-\beta)/(\pi|x|)$.

Appendix IV: Proof of Proposition 4.1. Suppose $F_1 = N(0, 1) \star G$, with G symmetric about zero with density function g having exponential index $2\kappa \in (0, 2)$ in the sense that:

$$-\log g(x) = |x|^{2\kappa} L(x)$$

for some slowly varying function $L(x)$. Then the Gaussian convolution $\phi \star g$ has exponential tail index 2κ .

Proof. Let $y > 0$ and $\beta := 2\kappa$. The convolution density $m(y) = \phi \star g$ is

$$\begin{aligned} m(y) &= \int g(y-\varepsilon) \phi(\varepsilon) d\varepsilon \\ &= \int_{|\varepsilon| < y^{\beta/2}} \exp(-|y-\varepsilon|^\beta L(y-\varepsilon)) \phi(\varepsilon) d\varepsilon + O(e^{-y^\beta L_1(y)}) \end{aligned}$$

for some slowly varying L_1 . Since $\beta < 2$, we have $y - \varepsilon = y(1 + o(1))$ as $y \rightarrow \infty$, so that

$$|y - \varepsilon|^\beta = y^\beta |1 - \varepsilon/y|^\beta \geq \begin{cases} y^\beta (1 - \beta\varepsilon/y) & \text{if } \beta > 1 \\ y^\beta (1 - \beta\varepsilon/y + \beta(\beta - 1)(\varepsilon/y)^2) & \text{if } \beta < 1, \end{cases}$$

for y sufficiently large. It follows that, over the range of the integral, we have

$$|y - \varepsilon|^\beta \geq y^\beta - \beta\varepsilon y^{\beta-1} + o(y^{2\beta-2}).$$

This implies $m(y)$ is bounded by

$$e^{-y^\beta L(y(1+o(1)))} \int_{|\varepsilon| < y^{\beta/2}} e^{-\beta\varepsilon y^{\beta-1} (1 + C y^{2\beta-2}/(\varepsilon y^{\beta-1})) L(y(1+o(1)))} \phi(\varepsilon) d\varepsilon + O(e^{-y^\beta L_1(y)})$$

for some constant $C > 0$. Now since $|y^{2\beta-2}/(\varepsilon y^{\beta-1})| = |y^{\beta-1}/\varepsilon| \geq y^{\beta/2-1} \rightarrow 0$, the above reduces to

$$m(y) \leq e^{-y^\beta L(y(1+o(1)))} \int_{|\varepsilon| < y^{\beta/2}} e^{-\beta\varepsilon y^{\beta-1} L_2(y)} \phi(\varepsilon) d\varepsilon + O(e^{-y^\beta L_1(y)}),$$

for some slowly varying function $L_2(y)$. The formula $E_{\varepsilon \sim N(0,1)}(e^{t\varepsilon}) = e^{t^2/2}$ implies

$$m(y) \leq e^{-y^\beta L(y(1+o(1)))} e^{\beta^2 y^{2\beta-2} L_2(y)^2/2} + O(e^{-y^\beta L_1(y)}).$$

Finally, $\beta < 2$ implies

$$y^\beta L(y(1+o(1))) + \beta^2 y^{2\beta-2} L_2(y)^2/2 = y^\beta L_3(y)$$

for some slowly varying $L_3(y)$, from which it follows that

$$-\log m(y) \geq y^\beta L_3(y).$$

A similar argument yields $-\log m(y) \leq y^\beta L_4(y)$ for some slowly varying $L_4(y)$, which implies

$$L_3(y) \leq \frac{-\log m(y)}{y^\beta} \leq L_4(y),$$

so that $\frac{-\log m(y)}{y^\beta}$ is also slowly varying, completing the proof. \square

Appendix V: Derivation of equation (5.2). For a log-likelihood function of the form

$$\ell(\theta, \nu) = \sum_{i=1}^n \log((1 - \theta)f_0(X_i) + \theta f_1(X_i; \nu)),$$

the profile log-likelihood derivative is

$$\frac{d}{d\theta} \ell(\theta; \hat{\nu}(\theta)) = \sum_{i=1}^n \frac{(h(X_i; \hat{\nu}(\theta)) - 1) + \theta t(X_i; \hat{\nu}(\theta))}{1 + \theta(h(X_i; \hat{\nu}(\theta)) - 1)}, \quad (.6)$$

where $h(x; \nu) = f_1(x/\nu)/f_0(x)$ and

$$t(x; \hat{\nu}(\theta)) = \frac{1}{f_0(x)} \frac{\partial f_1(x; \hat{\nu}(\theta))}{\partial \hat{\nu}} \frac{d\hat{\nu}(\theta)}{d\theta}. \quad (.7)$$

and for any given θ , $\hat{\nu}(\theta)$ solves

$$F(\theta, \hat{\nu}(\theta)) := \left. \frac{\partial \ell(\theta, \nu)}{\partial \nu} \right|_{\nu=\hat{\nu}(\theta)} = 0.$$

Total differentiation of this equation $F(\theta, \hat{\nu}(\theta)) = 0$ gives

$$\frac{d\hat{\nu}(\theta)}{d\theta} = -\frac{j_{\nu\theta}(\theta, \hat{\nu}(\theta))}{j_{\nu\nu}(\theta, \hat{\nu}(\theta))}, \quad (.8)$$

where $j_{\nu\theta}(\theta, \hat{\nu}(\theta))$ and $j_{\nu\nu}(\theta, \hat{\nu}(\theta))$ are the cross and double derivatives of the joint log-likelihood function evaluated at $(\theta, \hat{\nu}(\theta))$, given in the present context of $f_1(x; \nu) = \psi_\nu(x) \sim K_\nu/|x|^{\nu+1}$ by

$$j_{\nu\theta}(\theta, \hat{\nu}(\theta)) = \sum_{i=1}^n \frac{d\psi_\nu(X_i)}{d\nu} \Big|_{\nu=\hat{\nu}(\theta)}, \quad j_{\nu\nu}(\theta, \hat{\nu}(\theta)) = \theta \sum_{i=1}^n \frac{d^2\psi_\nu(X_i)}{d\nu^2} \Big|_{\nu=\hat{\nu}(\theta)},$$

where

$$\begin{aligned} \psi'_\nu(x) &:= \frac{d\psi_\nu(x)}{d\nu} \sim \frac{K'_\nu - K_\nu \log|x|}{|x|^{\nu+1}} \\ \psi''_\nu(x) &:= \frac{d^2\psi_\nu(x)}{d\nu^2} \sim \frac{K''_\nu - 2K'_\nu \log|x| + K_\nu(\log|x|)^2}{|x|^{\nu+1}}. \end{aligned}$$

On substituting (.8) in (.6) via (.7), and on making the replacement $f_0(x) = \phi(x)$, the profile log-likelihood derivative is

$$\frac{d}{d\theta} \ell(\theta; \hat{\nu}(\theta)) = \sum_{i=1}^n \frac{(\zeta_{\hat{\nu}(\theta)}(X_i) - r_{\hat{\nu}(\theta)}(X_i) - 1)}{1 + \theta(\zeta_{\hat{\nu}(\theta)}(X_i) - 1)}, \quad (.9)$$

where

$$\begin{aligned} r_{\hat{\nu}(\theta)}(X_i) &= \frac{\psi'_{\hat{\nu}(\theta)}(X_i)}{\phi(X_i)} \frac{\sum_{j=1}^n \psi'_{\hat{\nu}(\theta)}(X_j)}{\sum_{j=1}^n \psi''_{\hat{\nu}(\theta)}(X_j)} \\ &\sim \frac{K_{\hat{\nu}(\theta)} \log|X_i|}{|X_i|^{\hat{\nu}(\theta)+1} \phi(X_i)} \left(\frac{\sum_{j=1}^n K_{\hat{\nu}(\theta)} \log|X_j|/|X_j|^{\hat{\nu}(\theta)+1}}{\sum_{j=1}^n K_{\hat{\nu}(\theta)} (\log|X_j|)^2/|X_j|^{\hat{\nu}(\theta)+1}} \right) \rightarrow_{\mathbb{P}_0} 0, \quad (n \rightarrow \infty). \end{aligned}$$

Taking the limit of (.9) as $\theta \rightarrow 0$ gives (5.2). \square

Acknowledgement. P. McC is grateful to Titus Hilberdink for advice concerning regular-variation integrals and for pointing out the connection with de Bruijn conjugates.

REFERENCES

- [1] Bingham, N. H, Goldie, C. M. and Teugels, J. L. (1987). *Regular Variation*. Cambridge University Press, Cambridge, UK.
- [2] Brazzale, A. R. and Mameli, V. (2024). Likelihood asymptotics in nonregular settings: a review with emphasis on the likelihood ratio. *Statist. Sci.*, 39, 322–345.
- [3] Bickel, P. and Chernoff, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In: Ghosh, et al. (Eds.), *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, Wiley Eastern Limited, New Delhi, pp. 83–96.
- [4] de Bruijn, N. G. (1959). Pairs of slowly oscillating functions occurring in asymptotic problems concerning the Laplace transform. *Nieuw Arch. Wisk.*, 7, 20–26.
- [5] Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: the EM approach. *Ann. Statist.*, 37, 2523–2542.
- [6] Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.*, 25, 573–578.
- [7] Chow, T. L. and Teugels, J. L. (1979) The sum and the maximum of i.i.d. random variables. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, Petr Mandl and Marie Hušková, Editors, 81–92. North Holland Publishing Company.
- [8] Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96, 1151–1160.
- [9] Efron, B. (2012). Large-scale inference: empirical Bayes methods for estimation, testing, and prediction (Vol. 1). Cambridge University Press.
- [10] Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *Ann. Statist.*, 22, 1993–2010.
- [11] Ghosh, J. K. and Sen, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceeding of the Berkeley Conference in honour of Jerzy Neyman and Jack Kiefer*, 789–806.

- [12] Gnedenko, B. V and Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables*. Translated from the Russian and annotated by K. L. Chung; with an appendix by J. L. Doob. Addison-Wesley Pub. Co., Cambridge, Massachusetts.
- [13] Li, P., Chen, J. and Marriott, P. (2009). Non-finite Fisher information and homogeneity: an EM approach *Biometrika*, 96, 411–426.
- [14] Liu, X. and Shao, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J. Statist. Plann. Inference*, 123, 61–81.
- [15] McCullagh, P. and Polson, N. (2018). Statistical sparsity. *Biometrika*, 105, 797–814.
- [16] Patra, R. K. and Sen, B. (2016). Estimation of a two-component mixture model with applications to multiple testing. *J. R. Statist. Soc. B*, 78, 869–893.
- [17] Self, S. G. and Liang, K-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.*, 82, 605–610.
- [18] Shi, H. and Drton, M. (2024). On universal inference in Gaussian mixture models. *arXiv:2407.19361v1*.
- [19] Wassererman, L., Ramdas, A. and Balakrishnan, S. (2020). Universal inference. *Proc. Nat. Acad. Sci. USA*, 117, 16880–16890.
- [20] Vu, H. T. V. and Zhou, S. (1997). Generalization of likelihood ratio tests under nonstandard conditions. *Ann. Statist.*, 22, 1993–2010.
- [21] Zolotarev, V. M. (1986). *One-dimensional Stable Distributions*. American Mathematical Society, Providence.

HEATHER BATTEY: DEPARTMENT OF MATHEMATICS, IMPERIAL COLLEGE LONDON.

Email address: `h.battey@imperial.ac.uk`

PETER McCULLAGH AND DANIEL XIANG: DEPARTMENT OF STATISTICS, UNIVERSITY OF CHICAGO.

Email address: `pmcc@galton.uchicago.edu`, `dxiang@uchicago.edu`