Statistica Sinica **26** (2016), 955-978
doi:http://dx.doi.org/10.5705/ss.202015.0086

GENERALIZED LIKELIHOOD RATIO TEST FOR NORMAL MIXTURES

Wenhua Jiang and Cun-Hui Zhang

Soochow University and Rutgers University

Abstract: Let $X_1, \ldots, X_n$ be independent observations with $X_i \sim N(\theta_i, 1)$, where $(\theta_1, \ldots, \theta_n)$ is an unknown vector of normal means. Let $f_n(x) = \sum_{i=1}^n (d/dx) P_n\{X_i \le x\}/n$ be the average marginal density of observations. We consider the problem of testing $H_0\colon f_n \in \mathscr{F}_0$, where $\mathscr{F}_0$ is a family of mixture densities. This includes detecting nonzero normal means with $\mathscr{F}_0 = \{f_{\delta_0}\}$ and testing homogeneity in mixture models with $\mathscr{F}_0 = \{f_{\delta_\mu}\}$. We study a generalized likelihood ratio test (GLRT) based on the generalized maximum likelihood estimator (GMLE, Robbins (1950); Kiefer and Wolfowitz (1956)). We establish a large deviation inequality that provides a divergence rate ε_n of the GLRT under the null hypothesis. The inequality implies that the significance level of the test is of equal or smaller order than $n\varepsilon_n^2$. We show that the test can detect any alternative that is separated from the null by Hellinger distance ε_n. For the two-component Gaussian mixture, it turns out that the GLRT has full power asymptotically throughout the same region of amplitude sparsity where the Neyman-Pearson likelihood ratio test separates the two hypotheses completely (Donoho and Jin (2004)). We demonstrate the power of the GLRT for moderate samples with numerical experiments.

Key words and phrases: Detection boundary, generalized likelihood ratio test, generalized maximum likelihood estimator, normal mixture, sparse normal means.

1. Introduction

In this paper we study a generalized likelihood ratio test (GLRT) based on the generalized maximum likelihood estimator (GMLE) of the average of marginal densities of normal observations. Let $X_1, \ldots, X_n$ be independent observations with $X_i \sim N(\theta_i, 1)$, where $(\theta_1, \ldots, \theta_n)$ is an unknown vector of normal means. Let $f_n(x) = \sum_{i=1}^n (d/dx) P_n\{X_i \le x\}/n$ be the average marginal density of observations. We consider the problem of testing the null hypothesis $H_0\colon f_n \in \mathscr{F}_0$, where $\mathscr{F}_0$ is a family of mixture densities. It includes two important cases. One is detecting nonzero normal means with $\mathscr{F}_0 = \{f_{\delta_0}\}$ when $\theta_1, \ldots, \theta_n$ are deterministic conditional means, where δ_u is the probability distribution giving its entire mass to u. Another is testing homogeneity in mixture models with $\mathscr{F}_0 = \{f_{\delta_\mu}\}$ when $\theta_1, \ldots, \theta_n$ are i.i.d. conditional means.

Based on the GMLE, the GLRT assumes essentially no knowledge about the unknown means but still aims to approximate the usual Neyman-Pearson likelihood ratio test (LRT). The idea of GMLE was suggested in an abstract by Robbins (1950), and later received substantial theoretical development by Kiefer and Wolfowitz (1956). Lindsay (1995) gave a comprehensive overview of early works. Due to the prevalence of high-dimensional data and the rapid rise of computing power, recently there is a revival of interest in it, both in theory and computing. Zhang (2009) studied the convergence rate of the GMLE under the Hellinger distance. Jiang and Zhang (2009) considered estimation of a high-dimensional vector of normal means by GMLE. Jiang and Zhang (2010) investigated estimation of homoscedastic and heteroscedastic partial linear models. Koenker and Mizera (2014) studied convex optimization to compute the GMLE. They proposed an efficient R-package called REBayes. These works are all related to the compound estimation of normal means, where the oracle Bayes rule can be explicitly expressed in terms of the average of the marginal densities of the observations (Robbins (1956)). In view of these advances in estimation, a natural question has to do with the performance of GMLE in hypothesis testing. Recently, Gu, Koenker, and Volgushev (2013) considered testing homogeneity in mixture models.

One needs a careful analysis of the GLRT. Liu and Shao (2003) showed that under some general regularity conditions, the asymptotic distribution of the GLRT is the supremum of certain Gaussian processes. Azaïs, Gassiat, and Mercadier (2009) derived the distribution of the GLRT for a simple null hypothesis. However, these results cannot be directly applied in statistical inference. In this paper, we establish a large deviation inequality that provides a divergence rate ε_n of the GLRT under the null hypothesis. The inequality implies that the significance level of the test is of equal or smaller order than $n\varepsilon_n^2$. This type of result is new. We think it might be of independent interest as well. Meanwhile, the test statistic grows to infinity at a rate faster than $n\varepsilon_n^2$ under an alternative, provided the order of the Hellinger distance between the mixture densities under the null and the alternative is of larger order than ε_n. Consequently, the test can separate two hypotheses.

As an important case, we are interested in testing $H_0^*\colon f_n = f_{\delta_0}$. In the deterministic conditional means case, this amounts to testing if all normal means are zero. For the alternative, we consider testing against a sparse two-component Gaussian mixture; we wish to detect a sparse vector where a majority of normal means are zero and all the nonzero means are equal. There have recently been some important papers on this problem. See Donoho and Jin (2004) on the higher criticism approach (see also Ingster (1999, 2002)), Jager and Wellner (2007) on goodness-of-fit tests, Hall and Jin (2010) on innovated higher criticism for correlated data, Cai, Jeng, and Jin (2011) on detecting heteroscedastic mixtures, Greenshtein and Park (2012) on robust tests, Walther (2013) on the average likelihood ratio approach, and so on. It is known that for the two-component Gaussian mixture there is a threshold effect for the LRT: the sum of Type I and Type II errors tends to 0 or 1 depending on whether the value of nonzero means exceeds a detection boundary or not (Jin (2002)). However, the LRT requires the amplitude/sparsity parameters of the alternative to be known. This is not realistic in practice. The higher criticism is adaptive to the unknown degrees of heterogeneity in the detectable region. In this paper, we show that the GLRT has full power asymptotically throughout the same region of amplitude sparsity where the LRT separates the two hypotheses completely.

The rest of this paper is organized as follows. In Section 2 we introduce the statistical model and the GMLE. In Section 3. we propose the GLRT and study its significance level, power, and connection to the higher criticism. The generalization to a location-scale mixture model is also discussed. In Section 4 we provide some simulation results and data analysis. Section 5 contains conclusion. Proofs are given in Section 6.

2. The Generalized MLE

In this section, we introduce a general model in which the observations are independent and each observation is normally distributed given its latent conditional mean. This includes the inid case of deterministic conditional means and the i.i.d. case where the conditional means are themselves i.i.d., among other possible data generating models.

2.1. The inid location-mixture model

Let X_i be independent observations with

$$X_i|\theta_i \sim N(\theta_i, 1), \quad i = 1, \ldots, n, \tag{2.1}$$

where $(\theta_1, \ldots, \theta_n)$ is an unknown vector of normal means. We study the estimation of the average marginal density of the observations:

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{d}{dx} P_n\{X_i \le x\}. \tag{2.2}$$

This includes the cases of i.i.d. θ_i and completely deterministic θ_i under different choices of the probability measure P_n, since a deterministic sequence of constants can be treated as a sequence of degenerate random variables.

The average marginal density (2.2) can be explicitly written as a normal mixture density. Define the standardized normal location-mixture density as

$$f_G(x) = \int \varphi(x - u) dG(u), \tag{2.3}$$

where $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$ is the standard normal density. Let G_n be the average of the distribution functions of the unknowns $\{\theta_1, \ldots, \theta_n\}$ under P_n,

$$G_n(u) = \frac{1}{n}\sum_{i=1}^{n} P_n\{\theta_i \leq u\}. \tag{2.4}$$

Therefore, G_n is the distribution of $\theta_1, \ldots, \theta_n$ in the i.i.d. case and the empirical distribution of $\{\theta_1, \ldots, \theta_n\}$ in the deterministic conditional means case. With $\Phi(x) \equiv \int_{-\infty}^{x} \varphi(t)dt$, $P_n(X_i \leq x) = \int \Phi(x-u)dP\{\theta_i \leq u\}$, we have

$$f_n(x) = \int \varphi(x-u)dG_n(u) \tag{2.5}$$

as a location-mixture of standard normal densities.

2.2. The GMLE

Let $\mathscr{G}$ be the collection of all distributions in the real line $\mathbb{R}$ and take

$$\mathscr{F} = \big\{ f_G \colon G \in \mathscr{G} \big\} \tag{2.6}$$

as the family of all location-mixture of normal densities with unit variance. Given $X_1, \ldots, X_n$, the GMLE (Robbins (1950); Kiefer and Wolfowitz (1956)) of a normal mixture density is defined as

$$\hat{f}_n(x) = \arg\max_{f \in \mathscr{F}} \prod_{i=1}^{n} f(X_i). \tag{2.7}$$

Since the family $\mathscr{F}$ is indexed by the completely unknown mixing distribution G, the GMLE in (2.7) can be written as

$$\hat{f}_n(x) = f_{\hat{G}_n}(x), \quad \hat{G}_n = \arg\max_{G \in \mathscr{G}} \prod_{i=1}^{n} f_G(X_i). \tag{2.8}$$

That is, $\hat{f}_n$ itself is a mixture of normal density.

The GMLE in (2.7) and (2.8) is a sensible estimator for $f_n = f_{G_n}$, since the expectation of log-likelihood

$$E_n \log \prod_{i=1}^{n} f(X_i) = \sum_{i=1}^{n} E_n \log f(X_i) = n \int \big\{ \log f(x) \big\} f_{G_n}(x)dx$$

is uniquely maximized at $f = f_{G_n}$. The GMLE is a generalization of usual MLE in the following sense. Suppose θ_i's are i.i.d. variables with distribution G. To generate X_i's, it works to directly sample i.i.d. X_i's from density $f_G(x) = \int \varphi(x-u)dG(u)$. However, without the i.i.d. assumption, $\prod_{i=1}^{n} f(X_i)$ is not the likelihood of $X_1, \ldots, X_n$ for any $f \in \mathscr{F}$. So that to generate X_i's, we need first to generate the θ_i's and then to sample X_i from the normal distribution with mean θ_i and unit variance.

The maximization in (2.7) is done over an infinite-dimensional parameter space $\mathscr{F}$ and $\hat{G}_n$ is a complete nonparametric estimator in $\mathscr{G}$. It follows from (2.3) and the definition of $\hat{G}_n$ in (2.8) that the support of $\hat{G}_n$ is always within the range of the data $X_1, \ldots, X_n$ due to the monotonicity of $\varphi(x-u)$ in $|x-u|$. There exists a discrete solution of $\hat{G}_n$ with no more than n support points (e.g., Lindsay (1995)). The computation of the GMLE in (2.8) is typically carried out using iterative algorithms. For example, one may use the EM algorithm to maximize over the subfamily of all discrete distributions G supported on a fine grid in the range of data (Jiang and Zhang (2009)). Recently, Koenker and Mizera (2014) formulated the computation as a convex optimization problem and solved it by interior point methods. Our simulation presented in Section 4 used their R-package *REBayes*.

3. Generalized Likelihood Ratio Test

Let $X_1, \ldots, X_n$ be independent observations under model (2.1). Let f_n be the average marginal density and G_n be the average distribution function as in (2.2) and (2.4), respectively. We consider testing

$$H_0\colon f_n \in \mathscr{F}_0, \tag{3.1}$$

where $\mathscr{F}_0 \subset \mathscr{F}$. This amounts to testing if $(\theta_1, \ldots, \theta_n)$ are random samples from a certain distribution in $\mathscr{G}_0 \equiv \big\{G\colon f_G \in \mathscr{F}_0\big\}$ in the i.i.d. case, or if the empirical distribution of θ_i's is in $\mathscr{G}_0$ in the deterministic conditional means case. The GLRT is defined as

$$\Lambda_n = \sum_{i=1}^{n} \log \frac{\hat{f}_n(X_i)}{\hat{f}_{0,n}(X_i)}, \tag{3.2}$$

where $\hat{f}_n$ is given by (2.7) and $\hat{f}_{0,n} = \arg\max_{f\in\mathscr{F}_0} \prod_{i=1}^{n} f(X_i)$. It is natural to directly study the asymptotic null distribution of the proposed test. Liu and Shao (2003) showed that under some regularity conditions, the asymptotic distribution of Λ_n is the supremum of certain Gaussian processes. However, its properties for statistical inference are still unclear.

We divide this section into four subsections to study the significance level, the power, the connection to the higher criticism and the generalization to location-scale mixture model.

3.1. Main results

Our main result provides a large deviation inequality for the log-likelihood ratio Λ_n at a divergence rate ε_n that depends on the moments of elements in $\mathscr{G}_0$. The p-th weak moment of a distribution function G is

$$\mu_p(G) \equiv \Big\{ \sup_{x>0} x^p \int_{|u|>x} G(du) \Big\}^{1/p}. \tag{3.3}$$

Due to the Markov inequality, the p-th weak moment is no greater than the standard p-th absolute moment: $\big\{\mu_p(G)\big\}^p \le \int |u|^p G(du)$. The p-th weak moment of a distribution set $\mathscr{G}$ is defined as $\mu_p(\mathscr{G}) = \sup_{G\in\mathscr{G}} \mu_p(G)$. The divergence rate ε_n, as a function of the sample size n, the distribution set $\mathscr{G}_0$ and the power p of the weak moment, is defined as

$$\varepsilon(n,\mathscr{G},p) \equiv \max\Big\{ \sqrt{2\log n}, \Big\{ n^{1/p}\sqrt{\log n}\mu_p(\mathscr{G})\Big\}^{p/(2+2p)}\Big\}\sqrt{\frac{\log n}{n}}. \tag{3.4}$$

Theorem 1. *Let $X_1,\ldots,X_n$ be independent observations under* (2.1)*. For testing the null hypothesis H_0 in* (3.1)*, let $\hat f_n$ and $\hat f_{0,n}$ be defined as in* (3.2)*. Then under H_0, there exists a universal constant $k_* > 0$ such that for large n and all $k \ge k_*$,*

$$P_{H_0}\Big\{ \sum_{i=1}^n \log\frac{\hat f_n(X_i)}{\hat f_{0,n}(X_i)} \ge 3kn\varepsilon_n^2 \Big\} \le \exp\Big(-\frac{kn\varepsilon_n^2}{2\log n}\Big) \le n^{-k}, \tag{3.5}$$

where $\varepsilon_n \equiv \varepsilon(n,\mathscr{G}_0,p)$ is defined as in (3.4) *with $\mathscr{G}_0 \equiv \big\{G\colon f_G \in \mathscr{F}_0\big\}$. In particular,*

$$\varepsilon_n \asymp \begin{cases} n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)} & \textit{if } \mu_p(\mathscr{G}_0) = O(1) \textit{ for a fixed } p,\\ n^{-1/2}(\log n) & \textit{if } G([-M,M])=1 \textit{ for every } G\in\mathscr{G}_0,\ p=\infty. \end{cases}$$

To construct a level-α test, we must find a critical value $q(n,\alpha)$ such that

$$P_{H_0}\big\{\Lambda_n > q(n,\alpha)\big\} = \alpha. \tag{3.6}$$

It follows from (3.5) that $q(n,\alpha)$ is of equal or smaller order than $n\varepsilon_n^2$. In particular,

$$n\varepsilon_n^2 \asymp \begin{cases} n^{1/(1+p)}(\log n)^{(2+3p)/(2+2p)} & \text{if } \mu_p(\mathscr{G}_0) = O(1) \text{ for a fixed } p,\\ (\log n)^2 & \text{if } G([-M,M]) = 1 \text{ for every } G \in \mathscr{G}_0,\ p=\infty. \end{cases}$$

An important and special case of (3.1) is testing

$$H_0^*\colon f_n = f_{\delta_0}, \tag{3.7}$$

where δ_u is the probability distribution giving its entire mass to u. In the deterministic conditional means case, it amounts to testing if all normal means are zero, while its complement is that some of observations have nonnull mean.

Corollary 1. *Under H_0^* in* (3.7)*, $\varepsilon_n \asymp (\log n)/\sqrt{n}$. Consequently, $q(n,\alpha)$ is of equal or smaller order than $(\log n)^2$.*

The proof of Theorem 1 is enlightened by Zhang (2009), who studied the convergence rate of the GMLE under the Hellinger distance. Same technique can be employed in the divergence context. Theorem 1 depends on two elements. One is an entropy bound that controls the size of the normal location-mixture family $\mathscr{F}$ in (2.6) under the seminorm $\|h\|_{\infty,M} = \sup_{|x|\le M} |h(x)|$. For any semi-distance d, the ε-covering number $N(\varepsilon, \mathscr{F}, d)$ is the minimum number of balls of radius ε needed to cover $\mathscr{F}$. That is, with $\mathrm{Ball}(h_0, \varepsilon, d_0) \equiv \big\{h \colon d_0(h, h_0) < \varepsilon\big\}$,

$$N(\varepsilon, \mathscr{F}, d) \equiv \inf\big\{N \colon \mathscr{F} \subseteq \cup_{j=1}^{N} \mathrm{Ball}(h_j, \varepsilon, d)\big\}.$$

The other element is a large deviation inequality that bounds the likelihood ratio for large observations. We state these elements in two lemmas whose proofs are in Zhang (2009).

Lemma 1. *Let $\widetilde{L}(y) = \sqrt{-\log(2\pi y^2)}$ be the inverse of $y = \varphi(x)$. Then, for all $0 < \eta \le (2\pi e)^{-1/2}$,*

$$\log N(\eta^*, \mathscr{F}, \|\cdot\|_{\infty,M}) \le \Big\{4\big(6\widetilde{L}^2(\eta) + 1\big)\Big(\frac{2M}{\widetilde{L}(\eta)} + 3\Big) + 2\Big\}|\log \eta|, \tag{3.8}$$

where $\eta^* = 4\eta$. Consequently, there exists a universal constant C such that

$$\log N(\eta, \mathscr{F}, \|\cdot\|_{\infty,M}) \le C(\log \eta)^2 \max\Big(\frac{M}{\sqrt{|\log \eta|}}, 1\Big) \tag{3.9}$$

for all $0 < \eta \le 4(2\pi e)^{-1/2}$ and $M > 0$.

Lemma 2. *Let (X_i, θ_i) be independent random vectors with the conditional distribution $X_i|\theta_i \sim N(\theta_i, 1)$ under P_n. Let G_n and $\mu_p(G)$ be as in (2.4) and (3.3), respectively. Then for all constants $M \ge \sqrt{8 \log n}$, $0 < \lambda \le \min(1, p/2)$ and $a > 0$,*

$$E\Big\{\prod_{i=1}^{n} |aX_i|^{I\{|X_i| \ge M\}}\Big\}^{\lambda} \le \exp\Big\{2(aM)^{\lambda}\Big(\frac{2}{\sqrt{2\pi}M} + n\Big(\frac{2\mu_p(G_n)}{M}\Big)^p\Big)\Big\}.$$

3.2. Power

We now consider testing H_0 against a specific member in the complement. The Hellinger distance between two densities f_1 and f_2 is

$$d_H(f_1, f_2) = \Big(\int \big(f_1^{1/2}(x) - f_2^{1/2}(x)\big)^2 dx\Big)^{1/2}. \tag{3.10}$$

For two density sets $\mathscr{F}_1$ and $\mathscr{F}_2$, the Hellinger distance is

$$d_H(\mathscr{F}_1, \mathscr{F}_2) = \inf_{f_1 \in \mathscr{F}_1, f_2 \in \mathscr{F}_2} d_H(f_1, f_2). \tag{3.11}$$

Theorem 2. *Consider testing $H_0\colon f_n \in \mathscr{F}_0$ against $H_1^{(n)}\colon f_n \in \mathscr{F}_1^{(n)}$ where $\mathscr{F}_0 \cap \mathscr{F}_1^{(n)} = \emptyset$ for every $n \geq 1$. Let Λ_n and $q(n,\alpha)$ be defined as in* (3.2) *and* (3.6)*, respectively. Consider rejecting H_0 when $\Lambda_n > q(n,\alpha)$. Let $\eta_n = d_H(\mathscr{F}_0, \mathscr{F}_1^{(n)})$ be the Hellinger distance between $\mathscr{F}_0$ and $\mathscr{F}_1^{(n)}$. Let $\varepsilon_n \equiv \varepsilon(n, \mathscr{G}_0, p)$ be defined as in* (3.4) *with $\mathscr{G}_0 \equiv \big\{G\colon f_G \in \mathscr{F}_0\big\}$. If $\eta_n/\varepsilon_n \to \infty$, then the GLRT has full power asymptotically:*

$$P_{H_1^{(n)}}\big\{\textit{Reject } H_0\big\} \to 1, \quad n \to \infty.$$

Theorem 2 provides a sufficient condition under which the GLRT has full power to detect the alternative. Roughly speaking, the test is able to detect any alternative that is separated away from the null by distance ε_n.

Corollary 2. *Consider testing H_0^* in* (3.7) *against a simple alternative hypothesis $H_1^{(n)}\colon f_n \in \mathscr{F}_1^{(n)}$ where $\mathscr{F}_1^{(n)} = \big\{f_{G^{(n)}}\big\}$. Let Λ_n and $q(n,\alpha)$ be defined as in* (3.2) *and* (3.6) *respectively. Consider rejecting H_0^* when $\Lambda_n > q(n,\alpha)$. Let $\eta_n = d_H(f_{\delta_0}, f_{G^{(n)}})$ be the Hellinger distance between f_{δ_0} and $f_{G^{(n)}}$. If $n\eta_n^2/(\log n)^2 \to \infty$, then the GLRT has full power asymptotically:*

$$P_{H_1^{(n)}}\big\{\textit{Reject } H_0\big\} \to 1, \quad n \to \infty.$$

Theorem 2 is a consequence of the following.

Lemma 3. *Consider testing $H_0\colon f_n \in \mathscr{F}_0$ against $H_1\colon f_n \in \mathscr{F}_1$ where $\mathscr{F}_0 \cap \mathscr{F}_1 = \emptyset$. Let $\hat{f}_n$ and $\hat{f}_{0,n}$ be defined as in (3.2). Let $\eta = d_H(\mathscr{F}_0, \mathscr{F}_1)$ be the Hellinger distance between $\mathscr{F}_0$ and $\mathscr{F}_1$. Then for all $n \geq 1$,*

$$P_{H_1}\bigg\{\sum_{i=1}^n \log\frac{\hat{f}_n(X_i)}{\hat{f}_{0,n}(X_i)} > \frac{n}{2}\eta^2\bigg\} \geq 1 - \exp\Big(-\frac{n}{4}\eta^2\Big). \tag{3.12}$$

Thus the likelihood ratio is exponentially large with probability exponentially close to 1. The exponents are proportional to $n\eta^2$, where η is the Hellinger distance between $\mathscr{F}_0$ and $\mathscr{F}_1$. The inequality characterizes the divergence rate of the likelihood ratio of GMLE of a normal mixture density. It can be regarded as an extension of the likelihood ratio inequalities in Wong and Shen (1995).

3.3. GLRT in the two-component Gaussian mixture

In this subsection we focus on testing H_0^* in (3.7) against

$$H_1^{(n)}\colon f_n = (1-\xi_n)f_{\delta_0} + \xi_n f_{\delta_{\mu_n}}. \tag{3.13}$$

In the deterministic conditional means case, (3.13) means a small fraction ξ_n of θ_i's has nonzero value μ_n. In the i.i.d. case, this problem is called that of detecting sparse mixtures. There have been some important papers on this problem. An incomplete list of literature includes Donoho and Jin (2004) on the higher criticism approach, Jager and Wellner (2007) on goodness-of-fit tests, Hall and Jin (2010) on innovated higher criticism for correlated data, Cai, Jeng, and Jin (2011) on detecting heteroscedastic mixture, Greenshtein and Park (2012) on robust tests, Walther (2013) on the average likelihood ratio approach, and so on. Let

$$\xi_n = n^{-\beta}, \quad \mu_n = \sqrt{2r\log n}, \quad 0 < \beta < 1 \tag{3.14}$$

be the calibration. Define

$$\rho^*(\beta) = \begin{cases} 0, & 0 < \beta \le \frac{1}{2}, \\ \beta - \frac{1}{2}, & \frac{1}{2} < \beta \le \frac{3}{4}, \\ (1 - \sqrt{1-\beta})^2, & \frac{3}{4} < \beta < 1. \end{cases} \tag{3.15}$$

For the usual likelihood ratio test (LRT) , $\rho^*(\beta)$ defined in (3.15) partitions the amplitude/sparsity (β, r) plane into two regions: the detectable region $r > \rho^*(\beta)$ and the undetectable region $r < \rho^*(\beta)$. If the parameters are in the detectable region, the sum of type I and type II error probabilities of the likelihood ratio test vanishes asymptotically (Jin (2002)). It is well known that the higher criticism has asymptotically full power in the detectable region (Donoho and Jin (2004)). In the literature, $1/2 < \beta < 1$ and $0 < \beta \le 1/2$ are referred to as the sparse regime and dense regime, respectively. Donoho and Jin (2015) gave a comprehensive review on higher criticism for large-scale inference.

By analyzing the Hellinger distance between the mixture densities under H_0^* and $H_1^{(n)}$, we show that the GLRT also yields asymptotically full power for detection throughout the entire detectable region without the knowledge of parameters (β, r).

Theorem 3. *Consider testing* H_0^* *in* (3.7) *against* $H_1^{(n)}$ *in* (3.13). *Let* Λ_n *and* $q(n,\alpha)$ *be defined as in* (3.2) *and* (3.6) *respectively. Consider rejecting* H_0^* *when* $\Lambda_n > q(n,\alpha)$. *Let* $\xi_n = n^{-\beta}$ *for* $0 < \beta < 1$ *and* $\mu_n = \sqrt{2r\log n}$ *for* $0 < r < 1$. *Let* $G^{(n)} = (1-\xi_n)\delta_0 + \xi_n\delta_{\mu_n}$. *Then* $nd_H^2(f_{\delta_0}, f_{G^{(n)}})/(\log n)^2 \to \infty$ *if and only if* $r > \rho^*(\beta)$ *where* $\rho^*(\beta)$ *is defined in* (3.15). *Consequently, for every alternative* $H_1^{(n)}$ *in* (3.13) *with* r *exceeding the detection boundary* $\rho^*(\beta)$, *the GLRT has full power asymptotically.*

3.4. Location-scale mixture

The inid location-scale normal model is best described by a sequence of independent random vectors (X_i, ζ_i, τ_i) with the following conditional densities under P_n:

$$X_i|(\zeta_i,\tau_i)\sim N(\zeta_i,\tau_i^2),\quad \tau_i\ge\sigma,\quad i=1,\ldots,n, \tag{3.16}$$

where $\sigma>0$ is a known lower bound for the latent scale variables.

As far as the densities of the observations X_i are concerned, the location-scale mixture model (3.16) is equivalent to the location model (2.1). This can be seen as follows. Since the $N(\zeta_i/\sigma,\tau_i^2/\sigma^2)$ density is the convolution of the $N(0,1)$ and $N(\zeta_i/\sigma,\tau_i^2/\sigma^2-1)$ densities,

$$\frac{d}{dx}P_n\Big\{\frac{X_i}{\sigma}\le x\Big\}=E_n\varphi\Big(\frac{x-\zeta_i/\sigma}{\tau_i/\sigma}\Big)=E_n\int\varphi(x-u)d\Phi\Big(\frac{u-\zeta_i/\sigma}{\sqrt{\tau_i^2/\sigma^2-1}}\Big).$$

Thus, (2.1) holds with

$$G_n(u)=\frac{1}{n}\sum_{i=1}^n P_n\Big\{\frac{\theta_i}{\sigma}\le u\Big\}=\frac{1}{n}\sum_{i=1}^n E_n\Phi\Big(\frac{u-\zeta_i/\sigma}{\sqrt{\tau_i^2/\sigma^2-1}}\Big). \tag{3.17}$$

This gives the equivalence between the two models (2.1) and (3.16).

Let

$$h_n(x)=\frac{1}{n}\sum_{i=1}^n\frac{d}{dx}P_n\Big\{\frac{X_i}{\sigma}\le x\Big\} \tag{3.18}$$

be the average marginal density. Let

$$h_L(x)=\int\int\frac{1}{\tau}\varphi\Big(\frac{x-\zeta}{\tau}\Big)L(d\zeta,d\tau) \tag{3.19}$$

be the location-scale mixture where $L(\mathbb{R}\times[\sigma,\infty))=1$ with a known lower bound $\sigma>0$ for the scale. Let $\mathscr{L}$ be the collection of all distributions on $\mathbb{R}\times[\sigma,\infty)$ and take $\mathscr{H}=\big\{h_L\colon L\in\mathscr{L}\big\}$. Consider testing

$$H_0\colon h_n\in\mathscr{H}_0, \tag{3.20}$$

where $\mathscr{H}_0\subset\mathscr{H}$. This amounts to test if $(\zeta_1,\tau_1),\ldots,(\zeta_n,\tau_n)$ are random samples from a certain distribution $L_0\in\mathscr{L}_0\equiv\big\{L\colon h_L\in\mathscr{H}_0\big\}$ in the i.i.d. case, or to test if the empirical distribution of (ζ_i,τ_i)'s is a certain $L_0\in\mathscr{L}_0$ in the deterministic conditional means and variances case. The GLRT is defined as

$$\Lambda_n^*=\sum_{i=1}^n\log\frac{\hat h_n(X_i)}{\hat h_{0,n}(X_i)}, \tag{3.21}$$

where $\hat h_n=\arg\max_{h\in\mathscr{H}}\prod_{i=1}^n h(X_i)$ and $\hat h_{0,n}=\arg\max_{h\in\mathscr{H}_0}\prod_{i=1}^n h(X_i)$. Similarly, we take a critical value $q^*(n,\alpha)$ such that

$$P_{H_0}\big\{\Lambda_n^*>q^*(n,\alpha)\big\}=\alpha. \tag{3.22}$$

The equivalence given in (3.17) naturally leads to the following.

Theorem 4. *Let $X_1, \ldots, X_n$ be independent observations under location-scale model* (3.16)*. Consider testing the null hypothesis H_0 in* (3.20)*. Let $\hat{h}_n$ and $\hat{h}_{0,n}$ be defined as in* (3.21)*. Let $\varepsilon_n \equiv \varepsilon(n, \mathscr{G}_0, p)$, where $\mathscr{G}_0$ is the mapping of $\mathscr{L}_0 \equiv \big\{L \colon h_L \in \mathscr{H}_0\big\}$ under* (3.17)*. Then under H_0, there exists a universal constant $k_* > 0$ such that for large n and all $k \geq k_*$,*

$$P_{H_0}\bigg\{\sum_{i=1}^{n} \log \frac{\hat{h}_n(X_i)}{\hat{h}_{0,n}(X_i)} \geq 3kn\varepsilon_n^2\bigg\} \leq \exp\Big(-\frac{kn\varepsilon_n^2}{2\log n}\Big) \leq n^{-k}. \qquad (3.23)$$

Moreover, (3.23) *provides the divergence rate $\varepsilon_n \asymp n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)}$ if*

$$\sup_{x>0} \frac{x^p}{n} \sum_{i=1}^{n} P_n\big\{|\zeta_i| > x\big\} + \frac{1}{n}\sum_{i=1}^{n} E_n \tau_i^p = O(1) \qquad (3.24)$$

for a fixed $p > 0$.

Theorem 5. *Consider testing H_0 in* (3.20) *against $H_1^{(n)} \colon h_n \in \mathscr{H}_1^{(n)}$. Let Λ_n^* and $q^*(n, \alpha)$ be defined as in* (3.21) *and* (3.22) *respectively. Consider rejecting H_0 when $\Lambda_n^* > q^*(n, \alpha)$. Let $\eta_n = d_H(\mathscr{H}_0, \mathscr{H}_1^{(n)})$ be the Hellinger distance between $\mathscr{H}_0$ and $\mathscr{H}_1^{(n)}$. Let $\varepsilon_n \equiv \varepsilon(n, \mathscr{G}_0, p)$ where $\mathscr{G}_0$ is the mapping of $\mathscr{L}_0 \equiv \big\{L \colon h_L \in \mathscr{H}_0\big\}$ under* (3.17)*. If $\eta_n/\varepsilon_n \to \infty$, then the GLRT has full power asymptotically:*

$$P_{H_1^{(n)}}\big\{\textit{Reject } H_0\big\} \to 1, \quad n \to \infty.$$

4. Numerical Studies

In this section we report on comparisons of the power of the GLRT with that of other tests. We employed the R-package *REBayes* (Koenker and Mizera (2014)) to compute the GLRT.

4.1. Null distribution

We studied the asymptotic null distribution of GLRT Λ_n under $H_0^* \colon f_n = f_{\delta_0}$ by simulation. The top panels in Figure 1 display the histograms of $2\Lambda_n$ under H_0^* based on 10^4 replications with sample size $n = 1,000$ and 5,000. Two curves are added to the histograms: the density estimation curve and the pdf of χ^2-distribution with d.f. = 2 × Ave$\{\Lambda_n\}$. The bottom panels display the Q-Q plots of $2\Lambda_n$. These figures illustrate that the asymptotic null distribution of $2\Lambda_n$ is very close to χ^2. We further studied the distribution of $2\Lambda_n$. In Table 1, the mean, variance, skewness and kurtosis of $2\Lambda_n$ are displayed based on 10^4 replication. The latter three characteristics are compared with their counterparts of χ^2-distribution with the same mean. There are two messages from Table 1: the finite samples indicate that the asymptotic distribution of $2\Lambda_n$ is not exactly χ^2, and the discrepancy between $2\Lambda_n$ and χ^2 increases as sample size increases.

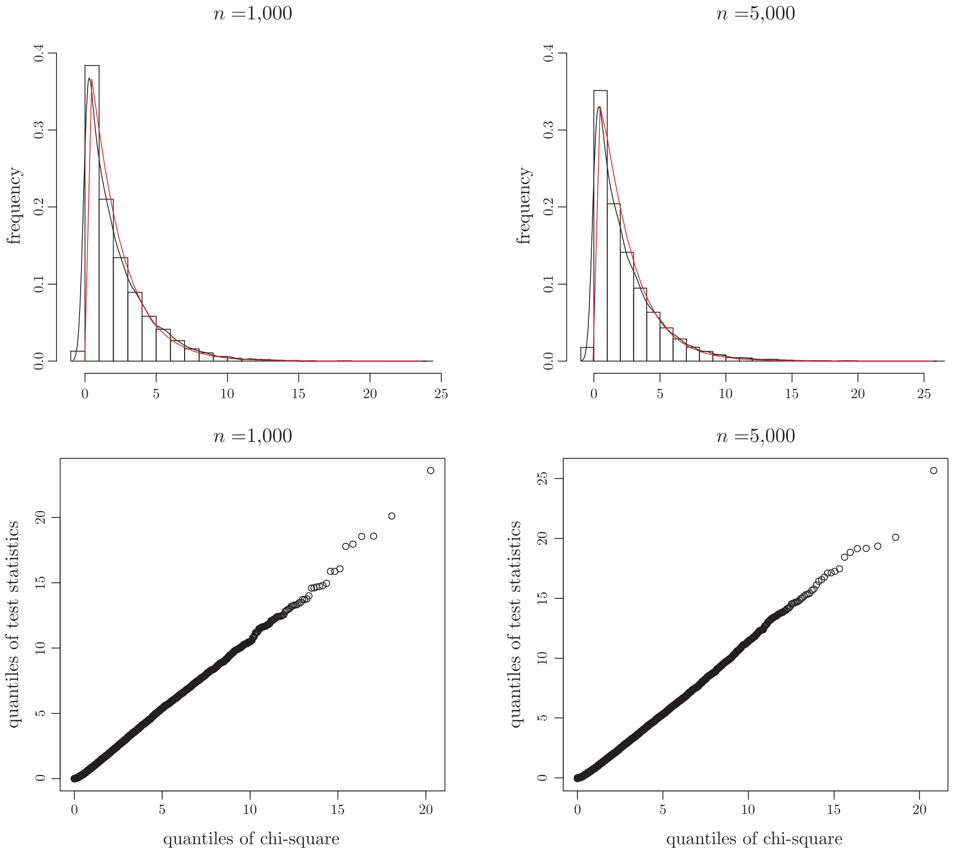

Figure 1. Top panels: histograms of $2\Lambda_n$. In each panel, the left curve represents the kernel density estimation of the distribution of $2\Lambda_n$. The right one represents the density curve of χ^2-distribution with d.f. $= 2 \times \text{Ave}\{\Lambda_n\}$. Bottom panels: Q-Q plots of $2\Lambda_n$. Left panels: $n = 1,000$; right panels: $n = 5,000$. Each panel is based on 10,000 replications.

4.2. Power comparison

We evaluated the power of the GLRT in testing $H_0^*: f_n = f_{\delta_0}$. The simulation includes the higher criticism (HC, Donoho and Jin (2004)), the HC+ (a variation of HC), the Berk-Jones test (Jager and Wellner (2007)) and the Kolmogorov-Smirnov test. Let p_i be the p-value for the ith component null hypothesis and $p_{(1)} < p_{(2)} < \ldots < p_{(n)}$ be the ordered p-values. The HC is defined as

$$\text{HC}_n^* = \max_{1 \le i \le \alpha \cdot n} \frac{\sqrt{n}\big[i/n - p_{(i)}\big]}{\sqrt{p_{(i)}(1 - p_{(i)})}}.$$

The Berk-Jones test is $\text{BJ}_n^+ = n \cdot \max_{1 \le i \le n/2} K^+(i/n, p_{(i)})$ where $K^+(t,x) = \big(t\log(t/x) + (1-t)\log((1-t)/(1-x))\big)I\{0 < x < t < 1\}$.

Table 1. Comparison between the empirical distribution of $2\Lambda_n$ under $H_0^*\colon X_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ and the χ^2-distribution with d.f. $= 2 \times \text{Ave}\{\Lambda_n\}$. Each entry is based on 10^4 replications.

		mean	variance	skewness	kurtosis
$n = 1,000$	$2\Lambda_n$	2.164	5.187	1.921	5.514
	χ^2	2.164	4.329	1.923	5.544
$n = 3,000$	$2\Lambda_n$	2.277	5.756	1.997	6.066
	χ^2	2.277	4.554	1.874	5.270
$n = 5,000$	$2\Lambda_n$	2.358	6.123	2.013	6.023
	χ^2	2.358	4.717	1.842	5.088

Table 2. Simulated critical values of various methods under $H_0^*\colon X_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ based on 10^4 replications. The HC, HC+ and the BJ are p-value based procedures. Both the one-side and two-side p-values are considered.

		$\alpha = 0.05$				$\alpha = 0.1$			
		GLRT	HC	HC+	BJ	GLRT	HC	HC+	BJ
$n = 1,000$	one-side	6.71	4.68	3.13	4.23	5.25	3.55	2.76	3.43
	two-side	—	4.66	3.16	4.30	—	3.61	2.80	3.48
$n = 3,000$	one-side	7.06	4.75	3.21	4.42	5.36	3.63	2.83	3.61
	two-side	—	4.83	3.23	4.42	—	3.66	2.85	3.62
$n = 5,000$	one-side	7.24	4.79	3.24	4.59	5.55	3.66	2.87	3.87
	two-side	—	4.92	3.22	4.54	—	3.65	2.84	3.68
$n = 10,000$	one-side	7.37	4.78	3.24	4.58	5.62	3.64	2.89	3.79
	two-side	—	4.83	3.25	4.60	—	3.64	2.88	3.78

In Table 2, we report simulated critical values of various methods based on 10^4 replications. Since the HC, HC+, and the BJ are all p-value based procedures, both critical values based on one-side and two-side p-values are reported.

In the first experiment we tested H_0^* against the two-component Gaussian mixtures model (3.13). We chose four combinations of parameters: $(n, \xi_n) = (1,000, 0.01)$, $(1,000, 0.005)$, $(5,000, 0.005)$, and $(5,000, 0.001)$. Let $\mu^* = \sqrt{2\rho^*(\beta)\log n}$ be the thresholding effect value of μ where $\rho^*(\beta)$ is the detection boundary in (3.15). The corresponding calibration gave $(\beta, \mu^*) = (0.667, 1.517)$, $(0.767, 1.923)$, $(0.622, 1.442)$, and $(0.811, 2.333)$. These settings are quite sparse. We let the amplitude parameter μ_n range from 1.25 to 3 with an increment of 0.25. This range always includes μ^*. We set the significance level $\alpha = 0.05$ and 0.1. Since the values of μ_n were always positive, we used the one-sided p-value for the HC, HC+, and the BJ tests. Table 3 displays the powers of the GLRT, the HC, and HC+ based on 1,000 replications. The boldface entries denote the power over 0.5. In the two-component Gaussian mixture setting, the power of GLRT is always between the HC and HC+.

Table 3. Power comparison in the setting where nonzero θ_i have common value μ. Each entry is based on 1,000 replications.

	α	μ_n	1.25	1.5	1.75	2	2.25	2.5	2.75	3
$n = 1,000$ $\xi_n = 0.01$	0.05	GLRT	0.104	0.151	0.200	0.358	**0.503**	**0.655**	**0.843**	**0.937**
		HC	0.096	0.138	0.228	0.356	**0.531**	**0.694**	**0.863**	**0.946**
		HC+	0.115	0.172	0.229	0.370	0.480	**0.611**	**0.782**	**0.875**
		BJ	0.115	0.177	0.231	0.390	**0.538**	**0.691**	**0.857**	**0.947**
		KS	0.090	0.090	0.092	0.111	0.101	0.073	0.094	0.113
	0.1	GLRT	0.179	0.258	0.296	0.463	**0.617**	**0.752**	**0.898**	**0.967**
		HC	0.168	0.260	0.358	**0.550**	**0.671**	**0.831**	**0.942**	**0.977**
		HC+	0.223	0.301	0.348	**0.515**	**0.618**	**0.729**	**0.859**	**0.923**
		BJ	0.221	0.305	0.372	**0.554**	**0.684**	**0.804**	**0.929**	**0.977**
		KS	0.165	0.164	0.173	0.182	0.204	0.157	0.166	0.193
$n = 1,000$ $\xi_n = 0.005$	0.05	GLRT	0.069	0.081	0.115	0.162	0.251	0.316	0.482	**0.629**
		HC	0.079	0.088	0.134	0.237	0.304	0.397	**0.576**	**0.751**
		HC+	0.099	0.107	0.115	0.162	0.215	0.255	0.341	0.420
		BJ	0.098	0.102	0.128	0.207	0.263	0.351	**0.515**	**0.667**
		KS	0.067	0.063	0.081	0.058	0.076	0.076	0.084	0.064
	0.1	GLRT	0.125	0.143	0.188	0.255	0.342	0.417	**0.582**	**0.735**
		HC	0.157	0.177	0.227	0.337	0.436	**0.546**	**0.691**	**0.851**
		HC+	0.173	0.189	0.202	0.260	0.327	0.389	0.477	**0.565**
		BJ	0.173	0.189	0.208	0.308	0.381	0.485	**0.615**	**0.785**
		KS	0.122	0.111	0.142	0.118	0.136	0.134	0.137	0.123
$n = 5,000$ $\xi_n = 0.005$	0.05	GLRT	0.111	0.163	0.261	0.455	**0.658**	**0.861**	**0.967**	**0.998**
		HC	0.077	0.145	0.230	0.368	**0.602**	**0.814**	**0.943**	**0.996**
		HC+	0.134	0.191	0.288	0.476	**0.667**	**0.863**	**0.968**	**0.991**
		BJ	0.119	0.192	0.277	0.479	**0.671**	**0.872**	**0.967**	**0.998**
		KS	0.078	0.089	0.101	0.111	0.110	0.099	0.107	0.112
	0.1	GLRT	0.189	0.273	0.390	**0.576**	**0.771**	**0.921**	**0.983**	**0.999**
		HC	0.170	0.253	0.385	**0.564**	**0.780**	**0.919**	**0.979**	**0.999**
		HC+	0.225	0.317	0.444	**0.631**	**0.778**	**0.925**	**0.982**	**0.998**
		BJ	0.192	0.303	0.410	**0.604**	**0.780**	**0.924**	**0.980**	**1.000**
		KS	0.142	0.155	0.175	0.197	0.218	0.184	0.185	0.183
$n = 5,000$ $\xi_n = 0.001$	0.05	GLRT	0.054	0.058	0.064	0.091	0.118	0.178	0.252	0.338
		HC	0.052	0.075	0.066	0.109	0.163	0.237	0.348	0.473
		HC+	0.060	0.060	0.075	0.099	0.117	0.168	0.190	0.227
		BJ	0.054	0.056	0.071	0.108	0.136	0.191	0.268	0.373
		KS	0.060	0.051	0.053	0.062	0.059	0.066	0.046	0.051
	0.1	GLRT	0.124	0.113	0.129	0.170	0.195	0.264	0.372	0.453
		HC	0.103	0.125	0.121	0.200	0.250	0.340	0.470	**0.589**
		HC+	0.122	0.130	0.154	0.166	0.181	0.249	0.282	0.361
		BJ	0.102	0.113	0.134	0.171	0.209	0.270	0.372	0.471
		KS	0.117	0.099	0.101	0.108	0.105	0.121	0.101	0.108

In the second experiment we tested H_0^* against the Gaussian hierarchical model. For $i = 1, \ldots, n$, we flipped a coin with probability ξ_n of landing heads. When the coin landed tails, we drew an observation X_i from $N(0, 1)$. When the coin landed heads, we drew an observation μ_i from $N(0, \tau^2)$ and then an observation X_i from $N(\mu_i, 1)$. We let τ range from 1 to 4 with an increment of 0.5. We still set $(n, \xi_n) = (1,000, 0.01)$, $(1,000, 0.005)$, $(5,000, 0.005)$ and $(5,000, 0.001)$. The average powers over 1,000 replications are displayed in Table 4, which demonstrate that the GLRT is competitive to the HC and the BJ. The results also suggest that the testing problem becomes easier as τ^2 increases.

4.3. Data analysis

We use leukemia gene microarray data (Golub (1999)) to illustrate the use of GLRT. We used the cleaned version published by Dettling (2004) that contains measurements for 3,571 genes. There are 72 samples coming from two classes: ALL (acute lymphoblastic leukemia) and AML (acute myeloid leukemia). Among these 72 samples, there are 38 (27 in ALL and 11 in AML) training samples and an independent collection of 34 (20 in ALL and 14 in AML) test samples.

We applied the proposed test to the gene microarray data. Let x_{ij} denote the expression level for the ith sample and the jth gene, $1 \le i \le n$, $1 \le j \le p$. Let C and D be the set of indices of samples from the training set and the test set, respectively. For notational consistency with later sections, we only used the data in the training set, but using the whole data gave similar results. Write $C = C_1 \cup C_2$, where C_1 and C_2 are the sets of indices of the training samples from classes 1 and 2, respectively. Let $\bar{x}_{jk} = \sum_{i \in C_k} x_{ij}/|C_k|$ be the average expression value of gene j for all samples in class k, $k = 1, 2$, and $s_j^2 = \big[\sum_{i \in C_1}(x_{ij} - \bar{x}_{j1})^2 + \sum_{i \in C_2}(x_{ij} - \bar{x}_{j2})^2\big]/(|C| - 2)$ be the pooled variance. Define the t-type statistic

$$z_j^* = \frac{1}{\sqrt{1/|C_1| + 1/|C_2|}} \frac{\bar{x}_{j1} - \bar{x}_{j2}}{s_j}, \quad j = 1, \ldots, p.$$

We followed Efron's suggestion (2004) to standardize z_j^*:

$$Z_j = \frac{z_j^* - \bar{z}^*}{\mathrm{sd}(z^*)}, \quad j = 1, \ldots, p,$$

where $\bar{z}^*$ and $\mathrm{sd}(z^*)$ represent the empirical mean and standard deviation of z_j^*'s, respectively.

We applied procedures to the Z_j's for the leukemia data. The resulting GLRT score was 15.3588. The p-values associated with the score is $\approx 10^{-3}$. This suggests the definite presence of signals scattered in the Z-vector. See Table 5 for other scores and p-values.

Table 4. Power comparison in the setting where nonzero θ_i are sampled from $N(0, \tau^2)$. Each entry is based on 1,000 replications.

	α	τ	1	1.5	2	2.5	3	3.5	4
$n = 1,000$ $\xi_n = 0.01$	0.05	GLRT	0.098	0.252	**0.523**	**0.746**	**0.888**	**0.957**	**0.981**
		HC	0.105	0.282	**0.562**	**0.788**	**0.902**	**0.962**	**0.982**
		HC+	0.070	0.141	0.268	0.381	0.489	**0.618**	**0.708**
		BJ	0.074	0.231	**0.506**	**0.751**	**0.872**	**0.953**	**0.978**
		KS	0.048	0.057	0.048	0.045	0.049	0.057	0.048
	0.1	GLRT	0.165	0.327	**0.593**	**0.788**	**0.910**	**0.962**	**0.988**
		HC	0.171	0.360	**0.632**	**0.832**	**0.932**	**0.973**	**0.991**
		HC+	0.135	0.222	0.369	0.493	**0.622**	**0.717**	**0.783**
		BJ	0.153	0.318	**0.590**	**0.792**	**0.910**	**0.969**	**0.984**
		KS	0.103	0.104	0.095	0.081	0.107	0.105	0.104
$n = 1,000$ $\xi_n = 0.005$	0.05	GLRT	0.074	0.158	0.303	**0.513**	**0.665**	**0.768**	**0.847**
		HC	0.081	0.160	0.345	**0.552**	**0.718**	**0.796**	**0.864**
		HC+	0.058	0.101	0.109	0.178	0.216	0.270	0.304
		BJ	0.054	0.132	0.292	0.484	**0.651**	**0.739**	**0.832**
		KS	0.060	0.045	0.051	0.056	0.050	0.058	0.058
	0.1	GLRT	0.143	0.222	0.376	**0.569**	**0.716**	**0.808**	**0.872**
		HC	0.141	0.237	0.417	**0.613**	**0.763**	**0.835**	**0.888**
		HC+	0.127	0.165	0.196	0.276	0.315	0.372	0.403
		BJ	0.135	0.209	0.375	**0.562**	**0.715**	**0.796**	**0.868**
		KS	0.109	0.093	0.090	0.110	0.097	0.114	0.115
$n = 5,000$ $\xi_n = 0.005$	0.05	GLRT	0.088	0.313	**0.750**	**0.948**	**0.993**	**1.000**	**1.000**
		HC	0.082	0.327	**0.762**	**0.945**	**0.991**	**1.000**	**1.000**
		HC+	0.076	0.193	0.488	**0.737**	**0.886**	**0.965**	**0.988**
		BJ	0.083	0.296	**0.723**	**0.939**	**0.990**	**0.999**	**1.000**
		KS	0.051	0.032	0.038	0.062	0.059	0.048	0.047
	0.1	GLRT	0.160	0.406	**0.806**	**0.963**	**0.995**	**1.000**	**1.000**
		HC	0.148	0.414	**0.836**	**0.967**	**0.994**	**1.000**	**1.000**
		HC+	0.150	0.302	**0.607**	**0.827**	**0.933**	**0.976**	**0.997**
		BJ	0.168	0.397	**0.794**	**0.963**	**0.993**	**1.000**	**1.000**
		KS	0.101	0.086	0.082	0.112	0.109	0.107	0.103
$n = 5,000$ $\xi_n = 0.001$	0.05	GLRT	0.064	0.099	0.204	0.420	**0.572**	**0.696**	**0.795**
		HC	0.059	0.113	0.248	0.446	**0.619**	**0.729**	**0.822**
		HC+	0.062	0.056	0.081	0.122	0.158	0.210	0.241
		BJ	0.057	0.094	0.197	0.406	**0.551**	**0.687**	**0.780**
		KS	0.057	0.047	0.042	0.041	0.058	0.054	0.055
	0.1	GLRT	0.128	0.176	0.285	0.480	**0.628**	**0.744**	**0.827**
		HC	0.118	0.179	0.328	**0.519**	**0.681**	**0.781**	**0.851**
		HC+	0.118	0.137	0.157	0.221	0.261	0.317	0.354
		BJ	0.113	0.165	0.293	0.459	**0.606**	**0.728**	**0.826**
		KS	0.112	0.091	0.084	0.099	0.122	0.099	0.106

Table 5. Analysis of leukemia microarray data.

	GLRT	HC	HC+	BJ
training	15.3588	10.9105	6.1057	11.1536
p-value	10^{-3}	0.01	5×10^{-5}	10^{-4}

5. Conclusion

As mentioned in the introduction, a primary motivation of our investigation is the compound estimation of normal means, where the oracle Bayes rule can be explicitly expressed in terms of the average of the marginal densities of the observations (2.2). In Jiang and Zhang (2009), an empirical Bayes estimator based on GMLE was demonstrated to have superb numerical performance in a wide range of situations, including sparse settings. Thus, a natural question is that of the performance of GMLE in testing. We found an upper bound for the significance level by establishing a large deviation inequality. This was a different approach than in Azaïs, Gassiat, and Mercadier (2009).

Under $H_0^*\colon f_n = f_{\delta_0}$, Theorem 1 implies that $q(n, \alpha)$ is of equal or smaller order than $(\log n)^2$ (Corollary 1). We believe that this rate can be improved. This remains as future work. For the higher criticism, an innovated procedure has been proposed for correlated data (Hall and Jin (2010)). It would also be interesting to explore the effect of correlation on the proposed test. This problem is beyond the scope of this paper and is an interesting topic for future research.

6. Proofs

Proof of Theorem 1. Let $\eta = 1/n^2$ and $M = 2n\varepsilon_n^2/(\log n)^{3/2}$. For positive functions h_1 and h_2, let $L_n(h_1, h_2) = \prod_{i=1}^n \{h_1(X_i)/h_2(X_i)\}$. Define

$$f^*(x) = \eta I\{|x| \le M\} + \frac{\eta M^2}{x^2} I\{|x| > M\}. \tag{6.1}$$

Let $\mathscr{F} = \{f_G\colon G \in \mathscr{G}\}$ be the family of all location-mixture of normal distributions with unit variance. Let $\{f_j, j \le N\}$ be an η-net of $\mathscr{F}$ under the seminorm $\|h\|_{\infty,M} \equiv \sup_{|x| \le M} |h(x)|$, with $N = N(\eta, \mathscr{F}, \|\cdot\|_{\infty,M})$. For any $f \in \mathscr{F}$, there exists $j \le N$ such that

$$f(x) \le \begin{cases} f_j(x) + \eta = f_j(x) + f^*(x) & \text{if } |x| \le M, \\ \varphi(0) = \frac{1}{\sqrt{2\pi}} & \text{if } |x| > M, \end{cases}$$

due to $f^*(x) = \eta$ for $|x| \le M$ and $\sup_{f \in \mathscr{F}} f(x) = \varphi(0)$. Under H_0, $f_n = f_{G_0}$ for some $f_{G_0} \in \mathscr{F}_0$. It follows that

$$\begin{aligned}
&\sup_{f\in\mathscr{F}} \prod_{i=1}^n f(X_i) / \sup_{f\in\mathscr{F}_0} \prod_{i=1}^n f(X_i) \\
&\quad \le \sup_{j\le N} \left\{ \prod_{i=1}^n \frac{f_j(X_i)+f^*(X_i)}{f_{G_0}(X_i)} \Big/ \prod_{|X_i|\ge M} \frac{f_j(X_i)+f^*(X_i)}{f_{G_0}(X_i)} \right\} \prod_{|X_i|\ge M} \frac{(2\pi)^{-1/2}}{f_{G_0}(X_i)} \\
&\quad \le \sup_{j\le N} L_n(f_j+f^*, f_{G_0}) \prod_{|X_i|\ge M} \frac{(2\pi)^{-1/2}}{f^*(X_i)}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&P_{G_0}\Big\{ \sup_{f\in\mathscr{F}} \prod_{i=1}^n f(X_i) / \sup_{f\in\mathscr{F}_0} \prod_{i=1}^n f(X_i) \ge \exp\big(3kn\varepsilon_n^2\big)\Big\} \\
&\le P_{G_0}\Big\{ \sup_{j\le N} L_n(f_j+f^*, f_{G_0}) \prod_{|X_i|\ge M} \frac{(2\pi)^{-1/2}}{f^*(X_i)} \ge \exp\big(3kn\varepsilon_n^2\big)\Big\} \\
&\le P_{G_0}\Big\{ \sup_{j\le N} \prod_{i=1}^n \frac{f_j(X_i)+f^*(X_i)}{f_{G_0}(X_i)} \ge \exp\big(\frac{kn\varepsilon_n^2}{3}\big)\Big\} \\
&\quad + P_{G_0}\Big\{ \prod_{|X_i|\ge M} \frac{(2\pi)^{-1/2}}{f^*(X_i)} \ge \exp\big(\frac{8kn\varepsilon_n^2}{3}\big)\Big\}. \qquad (6.2)
\end{aligned}$$

We derive large deviation inequalities for the right hind side of (6.2). For the first term,

$$\begin{aligned}
&P_{G_0}\Big\{ \prod_{i=1}^n \frac{f_j(X_i)+f^*(X_i)}{f_{G_0}(X_i)} \ge \exp\big(\frac{kn\varepsilon_n^2}{3}\big)\Big\} \\
&\le \exp\big(-\frac{kn\varepsilon_n^2}{6}\big) \prod_{i=1}^n E_{G_0}\Big(\frac{f_j(X_i)+f^*(X_i)}{f_{G_0}(X_i)}\Big)^{1/2} \\
&\le \exp\Big\{ -\frac{kn\varepsilon_n^2}{6} + \sum_{i=1}^n E_{G_0}\Big\{\Big(\frac{f_j(X_i)+f^*(X_i)}{f_{G_0}(X_i)}\Big)^{1/2} - 1\Big\}\Big\} \\
&= \exp\Big\{ -\frac{kn\varepsilon_n^2}{6} + n\Big(\int \sqrt{(f_j+f^*)f_{G_0}} - 1\Big)\Big\}. \qquad (6.3)
\end{aligned}$$

It follows from Jensen's inequality and the definition of Hellinger distance that

$$\begin{aligned}
\int \sqrt{(f_j+f^*)f_{G_0}} - 1 &\le \int \sqrt{f_j f_{G_0}} - 1 + \int \sqrt{f^* f_{G_0}} \\
&\le -\frac{1}{2} d_H^2(f_j, f_{G_0}) + \Big(\int f^*\Big)^{1/2}.
\end{aligned}$$

This and $\int f^* = 4\eta M$ by (6.1) yield

$$\int \sqrt{(f_j + f^*) f_{G_0}} - 1 \le \sqrt{4\eta M}. \tag{6.4}$$

It follows from (6.3), (6.4) and the entropy bound in (3.9) that

$$\begin{aligned} &P_{G_0}\Big\{ \sup_{j\le N} \prod_{i=1}^n \frac{f_j(X_i) + f^*(X_i)}{f_{G_0}(X_i)} \ge \exp\big(\frac{kn\varepsilon_n^2}{3}\big)\Big\} \\ &\quad \le \exp\Big(\log N + n\sqrt{4\eta M} - \frac{kn\varepsilon_n^2}{6}\Big). \end{aligned}$$

Since $\eta = 1/n^2$ and $M = 2n\varepsilon_n^2/(\log n)^{3/2} \ge 4\sqrt{\log n}$,

$$\begin{aligned} \log N + n\sqrt{4\eta M} &\le C(2\log n)^2 \max\Big(\frac{M}{\sqrt{2\log n}}, 1\Big) + \sqrt{4M} \\ &\le \Big(\frac{k_*}{24}\Big) M (\log n)^{3/2} \le \Big(\frac{k}{12}\Big) n\varepsilon_n^2 \end{aligned}$$

for large n and $k_* \le k$. Thus,

$$P_{G_0}\Big\{ \sup_{j\le N} \prod_{i=1}^n \frac{f_j(X_i) + f^*(X_i)}{f_{G_0}(X_i)} \ge \exp\big(\frac{kn\varepsilon_n^2}{3}\big)\Big\} \le \exp\Big(-\frac{kn\varepsilon_n^2}{12}\Big). \tag{6.5}$$

By (6.1), $1/f^*(x) = x^2/(\eta M^2) = (nx/M)^2$ for $|x| \ge M$. So that

$$\begin{aligned} &P_{G_0}\Big\{ \prod_{|X_i|\ge M} \frac{(2\pi)^{-1/2}}{f^*(X_i)} \ge \exp\big(\frac{8kn\varepsilon_n^2}{3}\big)\Big\} \\ &\quad \le \exp\Big(-\frac{4kn\varepsilon_n^2}{3\log n}\Big) E_{G_0}\Big(\prod_{|X_i|\ge M} \Big|\frac{nX_i}{M}\Big|\Big)^{1/\log n}. \end{aligned} \tag{6.6}$$

Since $M = 2n\varepsilon_n^2/(\log n)^{3/2} \ge 4\sqrt{\log n}$, Lemma 2 is applicable with $a = n/M$ and $\lambda = 1/\log n \le 1$. This yields

$$E_{G_0}\Big\{ \prod_{|X_i|\ge M} \Big|\frac{nX_i}{M}\Big|\Big\}^{1/\log n} \le \exp\Big\{ \frac{e}{\sqrt{2\pi \log n}} + 2en\Big(\frac{2\mu_p(G_0)}{M}\Big)^p\Big\}. \tag{6.7}$$

The definition of ε_n gives

$$\frac{n\varepsilon_n^2/\log n}{n\big(2\mu_p(G_0)/M\big)^p} \ge 1.$$

Therefore, (6.6) and (6.7) give

$$P_{G_0}\Big\{\prod_{|X_i|\ge M}\frac{(2\pi)^{-1/2}}{f^*(X_i)}\ge \exp\big(\frac{8kn\varepsilon_n^2}{3}\big)\Big\}$$
$$\le \exp\Big\{-\Big(\frac{4k}{3}-2e\Big)\frac{n\varepsilon_n^2}{\log n}+\frac{e}{\sqrt{2\pi\log n}}\Big\}. \tag{6.8}$$

Inserting (6.5) and (6.8) into (6.2), we find that for large n and $k\ge k_*$,

$$P_{G_0}\Big\{\sup_{f\in\mathscr{F}}\prod_{i=1}^n f(X_i)/\sup_{f\in\mathscr{F}_0}\prod_{i=1}^n f(X_i)\ge\exp\big(3kn\varepsilon_n^2\big)\Big\}$$
$$\le\exp\Big(-\frac{kn\varepsilon_n^2}{2\log n}\Big)\le\exp(-k\log n)=n^{-k}.$$

The rate $\varepsilon_n\asymp n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)}$ is clear from (3.4) under $\mu_p(\mathscr{G}_0)=O(1)$. The rate $\varepsilon_n\asymp n^{-1/2}(\log n)$ also follows immediately from (3.4) under $G([-M,M])=1$ for every $G\in\mathscr{G}_0$ and $p=\infty$. This completes the proof.

Proof of Lemma 3. By the definition of GMLE, $\sum_{i=1}^n\log\hat f_n(X_i)\ge\sum_{i=1}^n\log f_G(X_i)$ for every $f_G\in\mathscr{F}$. Under H_1, $f_n=f_{G_1}$ for some $f_{G_1}\in\mathscr{F}_1$. Denote $\hat f_{0,n}=f_{G_0}\in\mathscr{F}_0$, then

$$P_{G_1}\Big\{\sum_{i=1}^n\log\frac{\hat f_n(X_i)}{\hat f_{0,n}(X_i)}>\frac{n}{2}\eta^2\Big\}\ge P_{G_1}\Big\{\sum_{i=1}^n\log\frac{f_{G_1}(X_i)}{f_{G_0}(X_i)}>\frac{n}{2}\eta^2\Big\}$$
$$=1-P_{G_1}\Big\{\prod_{i=1}^n\Big(\frac{f_{G_0}(X_i)}{f_{G_1}(X_i)}\Big)^{1/2}\ge\exp\Big(-\frac{n}{4}\eta^2\Big)\Big\}.$$

It follows from the Chebyshev inequality and the definition of Hellinger distance that

$$P_{G_1}\Big\{\prod_{i=1}^n\Big(\frac{f_{G_0}(X_i)}{f_{G_1}(X_i)}\Big)^{1/2}\ge\exp\Big(-\frac{n}{4}\eta^2\Big)\Big\}\le\exp\Big(\frac{n}{4}\eta^2\Big)\Big\{E_{G_1}\Big(\frac{f_{G_0}}{f_{G_1}}\Big)^{1/2}\Big\}^n$$
$$\le\exp\Big(\frac{n}{4}\eta^2\Big)\Big(1-\frac{\eta^2}{2}\Big)^n$$
$$\le\exp\Big(-\frac{n}{4}\eta^2\Big).$$

The above two inequalities give (3.12).

Proof of Theorem 3. First of all, $f_{G^{(n)}}(x)=(1-\xi_n)\varphi(x)+\xi_n\varphi(x-\mu_n)$. We divide the analysis of the Hellinger distance into two parts:

$$\begin{aligned} d_H^2(f_{\delta_0}, f_{G^{(n)}}) &= \int \Big(\sqrt{\varphi(x)} - \sqrt{(1-\xi_n)\varphi(x) + \xi_n\varphi(x-\mu_n)}\Big)^2 dx \\ &= \int \Big(1 - \sqrt{1 - n^{-\beta} + n^{-\beta}\exp(\frac{x\mu_n - \mu_n^2}{2})}\Big)^2 \varphi(x)dx \\ &\stackrel{\triangle}{=} \int_{-\infty}^{(\beta/\mu_n)\log n + \mu_n/2} + \int_{(\beta/\mu_n)\log n + \mu_n/2}^{+\infty}. \end{aligned} \tag{6.9}$$

When $x < (\beta/\mu_n)\log n + \mu_n/2$, the Taylor series gives that

$$\begin{aligned} &\sqrt{1 - n^{-\beta} + n^{-\beta}\exp(x\mu_n - \frac{\mu_n^2}{2})} \\ &\quad = 1 - \frac{1}{2}\big(1 + o(1)\big)n^{-\beta}\Big\{1 - \exp\Big(x\mu_n - \frac{\mu_n^2}{2}\Big)\Big\}. \end{aligned} \tag{6.10}$$

Then for the first piece of integration in (6.9),

$$\begin{aligned} &\int_{-\infty}^{(\beta/\mu_n)\log n + (\mu_n/2)} \Big(1 - \sqrt{1 - n^{-\beta} + n^{-\beta}\exp(x\mu_n - \frac{\mu_n^2}{2})}\Big)^2 \varphi(x)dx \\ &= \big(1 + o(1)\big)\int_{-\infty}^{(\beta/\mu_n)\log n + \mu_n/2} \frac{1}{4}n^{-2\beta}\Big\{1 - \exp\Big(x\mu_n - \frac{\mu_n^2}{2}\Big)\Big\}^2 \varphi(x)dx \\ &= \big(1 + o(1)\big)\frac{1}{4}n^{-2\beta}\int_{-\infty}^{(\beta/\mu_n)\log n + \mu_n/2} \Big\{1 - 2\exp\Big(x\mu_n - \frac{\mu_n^2}{2}\Big) \\ &\quad + \exp(2x\mu_n - \mu_n^2)\Big\}\varphi(x)dx \\ &\stackrel{\triangle}{=} \big(1 + o(1)\big)(I_1 + I_2 + I_3). \end{aligned} \tag{6.11}$$

Notice that $I_1 \equiv n^{-2\beta}\Phi\big((\beta/\mu_n)\log n + \mu_n/2\big)/4 = O(n^{-2\beta})$,

$$I_1 \gg \frac{(\log n)^2}{n}, \quad 0 < \beta < \frac{1}{2}. \tag{6.12}$$

For the cross product, we have

$$\begin{aligned} I_2 &\equiv -\frac{1}{2}n^{-2\beta}\int_{-\infty}^{(\beta/\mu_n)\log n + \mu_n/2} \exp\Big(x\mu_n - \frac{\mu_n^2}{2}\Big)\varphi(x)dx \\ &= -\frac{1}{2}n^{-2\beta}\Phi\Big(\frac{\beta}{\mu_n}\log n - \frac{\mu_n}{2}\Big) \\ &\le O(n^{-2\beta}). \end{aligned} \tag{6.13}$$

The analysis of I_3 is a little complicated. Direct calculations show that

$$\begin{aligned} I_3 &\equiv \frac{1}{4} n^{-2\beta} \int_{-\infty}^{(\beta/\mu_n)\log n + \mu_n/2} \exp(2x\mu_n - \mu_n^2)\varphi(x)dx \\ &= \frac{1}{4} n^{2(r-\beta)} \Phi\Big(\big(\sqrt{\frac{\beta^2}{2r}} - 3\sqrt{\frac{r}{2}}\big)\sqrt{\log n}\Big). \end{aligned} \tag{6.14}$$

There are two cases. The first case is $r \le \beta/3$, in which $I_3 = O(n^{2(r-\beta)})$ by (6.14). Simple algebra shows that

$$I_3 \gg (\log n)^2/n \text{ iff } r > \beta - 1/2, \quad \text{for } 1/2 < \beta \le 3/4. \tag{6.15}$$

The other case is $r > \beta/3$. Due to $\Phi(-x) = (1+o(1))\varphi(x)/x$, $I_3 = O(n^{-(r+\beta)^2/(4r)} /\sqrt{\log n})$, and then

$$I_3 \gg \tfrac{(\log n)^2}{n} \text{ iff } r > (1 - \sqrt{1-\beta})^2, \quad \text{for } 3/4 < \beta < 1. \tag{6.16}$$

We turn to the second piece of integration in (6.9). When $x \ge (\beta/\mu_n)\log n + \mu_n/2$, $n^{-\beta}\exp(x\mu_n - \mu_n^2/2)$ in the square root is the main term. So that

$$\begin{aligned} I_4 &\equiv \int_{(\beta/\mu_n)\log n + \mu_n/2}^{+\infty} \Big(1 - \sqrt{1 - n^{-\beta} + n^{-\beta}\exp(x\mu_n - \mu_n^2/2)}\Big)^2 \varphi(x)dx \\ &= O(1) n^{-\beta} \int_{(\beta/\mu_n)\log n + \mu_n/2}^{+\infty} \exp\Big(x\mu_n - \frac{\mu_n^2}{2}\Big)\varphi(x)dx \\ &= O(1) n^{-\beta} \Phi\Big(\big(\sqrt{\frac{r}{2}} - \sqrt{\frac{\beta^2}{2r}}\big)\sqrt{\log n}\Big). \end{aligned} \tag{6.17}$$

There are still two cases. The first case is $r \ge \beta$, in which $I_4 = O(n^{-\beta}) \gg (\log n)^2/n$ by (6.17). The other case is $r < \beta$. Due to $\Phi(-x) = (1+o(1))\varphi(x)/x$, $I_4 = O(n^{-(r+\beta)^2/(4r)}/\sqrt{\log n})$, and then

$$I_4 \gg \tfrac{(\log n)^2}{n} \text{ iff } r > (1 - \sqrt{1-\beta})^2, \quad \text{for } \tfrac{1}{2} < \beta < 1. \tag{6.18}$$

Combining $d_H^2(f_{\delta_0}, f_{G^{(n)}}) = \big(1+o(1)\big)(I_1+I_2+I_3)+I_4$, (6.12), (6.13), (6.15), (6.16), and (6.18), we have that $nd_H^2(f_{\delta_0}, f_{G^{(n)}})/(\log n)^2 \to \infty$ if and only if (3.15) holds. It follows immediately from Corollary 2 that for every alternative $H_1^{(n)}$ with $r > \rho^*(\beta)$, the GLRT has full power asymptotically.

Proof of Theorem 4. Due to the equivalence between (3.16) and (2.1), it suffices to translate the rate $\varepsilon_n = \varepsilon(n, \mathscr{G}, p)$ into functionals of the moments of ζ_i and τ_i, where G_n is as in (3.17). We assume $\sigma = 1$ without loss of generality. By (3.17), we may write $\theta_i|(\zeta_i, \tau_i) \sim N(\zeta_i, \tau_i^2 - 1)$, so that $\theta_i = \zeta_i + Z_i\sqrt{\tau_i^2 - 1}$, where Z_i are i.i.d. $N(0,1)$ random variables independent of (ζ_i, τ_i). For the p-th weak moment, (3.24) implies

$$\begin{aligned}\big\{\mu_p(G_n)\big\}^p &\le \sup_{x>0} \frac{x^p}{n} \sum_{i=1}^n P_n\big\{|\theta_i + Z_i\tau_i| > x\big\} \\ &\le \sup_{x>0} \frac{x^p}{n} \sum_{i=1}^n P_n\Big\{|\theta_i| > \frac{x}{2}\Big\} + \frac{2^p}{n} \sum_{i=1}^n E_n|Z_i|^p E_n \tau_i^p = O(1).\end{aligned}$$

The conclusion follows from Theorem 1.

Acknowledgement

The research of Wenhua Jiang is supported by the National Nature Science Foundation of China (11201327) and Nature Science Foundation of Jiangsu Province (BK2012176). The research of Cun-Hui Zhang is supported by NSF Grant DMS-1106753 and DMS-1209014.

References

Azaïs, J.-M., Gassiat, E. and Mercadier, C. (2009). The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM Probab. Statist.* **13**, 301-327.

Cai, T., Jeng, J. and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. Roy. Statist. Soc. Ser. B* **73**, 629-662.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20**, 3583-3593.

Donoho, D. L. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 962-994.

Donoho, D. L. and Jin, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Science* **30**, 1-25.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99**, 96-104.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rate of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29**, 1233-1263.

Golub, T. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-536.

Greenshtein, E. and Park, J. (2012). Robust test for detecting a signal in a high dimensional sparse normal vector. *J. Statist. Plann. Inference* **142**, 1445-1456.

Gu, J., Koenker, R. and Volgushev, S. (2013). Testing for homogeneity in mixture models. Manuscript.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38**, 1686-1732.

Ingster, Y. I. (1999). Minimax detection of a signal for l_p^n-balls. *Math. Methods Statist.* **7**, 401-428.

Ingster, Y. I. (2002). Adaptive detection of a signal of growing dimension, I, II. *Math. Methods Statist.* **10**, 395-421.

Jager, L. and Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35**, 2018-2053.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37**, 1647-1684.

Jiang, W. and Zhang, C.-H. (2010). Empirical Bayes in-season prediction of baseball batting averages. In *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown,* IMS Collections **6**, 263-273.

Jin, J. (2002). Detection boundary for sparse mixtures. Unpublished manuscript.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887-906.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109**, 674-685.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications.* IMS, Hayward, CA.

Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist.* **31**, 807-832.

Robbins, H. (1950). A generalization of the method of maximum likelihood: Estimating a mixing distribution (abstract). *Ann. Math. Statist.* **21**, 314-315.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 157-163. University of California Press, Berkeley.

Walther, G. (2013). The average likelihood ratio for large-scale multiple testing and detecting sparse mixtures. In *From Probability to Statistics and Back: High Dimensional Models and Processes,* IMS Collections **9**, 317-326.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23**, 339-362.

Zhang, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statist. Sinica* **19**, 1297-1318.

School of Mathematical Sciences, Soochow University, P.O. Box 173, 1 Shizi Street, Suzhou, Jiangsu 215006, China.

E-mail: jiangwenhua@suda.edu.cn

Department of Statistics and Biostatistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, U.S.A.

E-mail: czhang@stat.rutgers.edu

(Received March 2015; accepted October 2015)