Journal of the
Royal Statistical Society
SERIES B
Statistical
Methodology

J. R. Statist. Soc. B (2011)
73, *Part* 5, *pp.* 629–662

Optimal detection of heterogeneous and heteroscedastic mixtures

T. Tony Cai and X. Jessie Jeng

University of Pennsylvania, Philadelphia, USA

and Jiashun Jin

Carnegie Mellon University, Pittsburgh, USA

[Received March 2010. Final revision March 2011]

Summary. The problem of detecting heterogeneous and heteroscedastic Gaussian mixtures is considered. The focus is on how the parameters of heterogeneity, heteroscedasticity and proportion of non-null component influence the difficulty of the problem. We establish an explicit detection boundary which separates the detectable region where the likelihood ratio test is shown to detect the presence of non-null effects reliably from the undetectable region where no method can do so. In particular, the results show that the detection boundary changes dramatically when the proportion of non-null component shifts from the sparse regime to the dense regime. Furthermore, it is shown that the higher criticism test, which does not require specific information on model parameters, is optimally adaptive to the unknown degrees of heterogeneity and heteroscedasticity in both the sparse and the dense cases.

Keywords: Detection boundary; Higher criticism; Likelihood ratio test; Optimal adaptivity; Sparsity

1. Introduction

The problem of detecting non-null components in Gaussian mixtures arises in many applications, where a large number of variables are measured and only a small proportion of them possibly carry signal information. In disease surveillance, for instance, it is crucial to detect outbreaks of disease in their early stage when only a small fraction of the population is infected (Kulldorff *et al.*, 2005). Other examples include astrophysical source detection (Hopkins *et al.*, 2002) and covert communication (Donoho and Jin, 2004).

The detection problem is also of interest because detection tools can be easily adapted for other purposes, such as screening and dimension reduction. For example, in genomewide association studies, a typical single-nucleotide polymorphism data set consists of a very long sequence of measurements containing signals that are both sparse and weak. To locate such signals better, one could break the long sequence into relatively short segments and use the detection tools to filter out segments that contain no signals.

In addition, the detection problem is closely related to other important problems, such as large-scale multiple testing, feature selection and cancer classification. For example, the detection problem is the starting point for understanding estimation and large-scale multiple testing (Cai *et al.*, 2007). The fundamental limit for detection is intimately related to the fundamental

Address for correspondence: X. Jessie Jeng, Department of Biostatistics and Epidemiology, University of Pennsylvania, Room 207, Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, USA.
E-mail: xjeng@upenn.edu

 1369–7412/11/73629

limit for classification, and the optimal procedures for detection are related to optimal procedures in feature selection. See Donoho and Jin (2008, 2009), Hall *et al.* (2008) and Jin (2009).

In this paper we consider the detection of heterogeneous and heteroscedastic Gaussian mixtures. The goal is twofold:

(a) to discover the *detection boundary* in the parameter space that separates the *detectable region*, where it is possible to detect reliably the existence of signals on the basis of the noisy and mixed observations, from the *undetectable region*, where it is impossible to do so;
(b) to construct an adaptively optimal procedure that works without the information of signal features but is successful in the whole detectable region; such a procedure has the property of what we call *optimal adaptivity*.

The problem is formulated as follows. Given n independent observation units $X=(X_1, X_2, \ldots, X_n)$. For each $1 \leqslant i \leqslant n$, we suppose that X_i has probability ε of being a non-null effect and probability $1-\varepsilon$ of being a null effect. We model the null effects as samples from $N(0,1)$ and non-null effects as samples from $N(A, \sigma^2)$. Here, ε can be viewed as the proportion of non-null effects, A the heterogeneity parameter and σ the heteroscedasticity parameter. A and σ together represent signal intensity. Throughout this paper, all the parameters ε, A and σ are assumed unknown.

The goal is to test whether any signals are present, i.e. we wish to test the hypothesis $\varepsilon=0$ or, equivalently, to test the joint null hypothesis

$$H_0 : X_i \overset{\text{IID}}{\sim} N(0,1), \qquad 1 \leqslant i \leqslant n, \tag{1}$$

against a specific alternative hypothesis in its complement

$$H_1^{(n)} : X_i \overset{\text{IID}}{\sim} (1-\varepsilon)\, N(0,1) + \varepsilon\, N(A, \sigma^2), \qquad 1 \leqslant i \leqslant n. \tag{2}$$

The setting here turns out to be the key to understanding the detection problem in more complicated settings, where the alternative density itself may be a Gaussian mixture, or where the X_i may be correlated. The underlying reason is that the Hellinger distance between the null density and the alternative density displays certain monotonicity. See Section 6 for further discussion.

Motivated by the examples that were mentioned earlier, we focus on the case where ε is small. We adopt an asymptotic framework where n is the driving variable, whereas ε and A are parameterized as functions of n (σ is fixed throughout the paper). In detail, for a fixed parameter $0<\beta<1$, we let

$$\varepsilon = \varepsilon_n = n^{-\beta}. \tag{3}$$

The detection problem behaves very differently in two regimes: the *sparse regime* where $\frac{1}{2}<\beta<1$ and the *dense regime* where $0<\beta \leqslant \frac{1}{2}$. In the sparse regime, $\varepsilon_n \ll 1/\sqrt{n}$, and the most interesting situation is when $A=A_n$ grows with n at a rate of $\sqrt{\log(n)}$. Outside this range either it is too easy to separate the two hypotheses or it is impossible to do so. Also, the proportion ε_n is much smaller than the standard deviation of typical moment-based statistics (e.g. the sample mean), so these statistics would not yield satisfactory testing results. In contrast, in the dense case where $\varepsilon_n \gg 1/\sqrt{n}$, the most interesting situation is when A_n degenerates to 0 at an algebraic order, and moment-based statistics could be successful. However, from a practical point, moment-based statistics are still not preferred as β is in general unknown.

In light of this, the parameter $A=A_n(r;\beta)$ is calibrated as follows: for the sparse case,

$$A_n(r;\beta) = \sqrt{\{2r\log(n)\}}, \qquad 0<r<1, \;\; \text{if } \tfrac{1}{2}<\beta<1, \tag{4}$$

and, for the dense case,

$$A_n(r;\beta)=n^{-r}, \qquad 0<r<\tfrac{1}{2}, \text{ if } 0<\beta\leqslant\tfrac{1}{2}. \tag{5}$$

A similar setting has been studied in Donoho and Jin (2004), where the scope is limited to the case $\sigma=1$ and $\beta\in(\frac{1}{2},1)$. Even in this simpler setting, the testing problem is non-trivial. A testing procedure called *higher criticism*, which contains three simple steps, was proposed. First, for each $1\leqslant i\leqslant n$, obtain a p-value by

$$p_i=\bar{\Phi}(X_i)\equiv P\{N(0,1)\geqslant X_i\}, \tag{6}$$

where $\bar{\Phi}=1-\Phi$ is the survival function of $N(0,1)$. Second, sort the p-values in the ascending order $p_{(1)}<p_{(2)}<\ldots<p_{(n)}$. Last, define the higher criticism statistic as

$$\mathrm{HC}_n^*=\max_{\{1\leqslant i\leqslant n\}}(\mathrm{HC}_{n,i}), \qquad \mathrm{HC}_{n,i}=\frac{i/n-p_{(i)}}{\sqrt{\{p_{(i)}(1-p_{(i)})\}}}\sqrt{n}, \tag{7}$$

and reject the null hypothesis H_0 when HC_n^* is large. Higher criticism is very different from the more conventional moment-based statistics. The key ideas can be illustrated as follows. When $X\sim N(0,I_n)$, $p_i\sim^{\mathrm{IID}}U(0,1)$ and so $\mathrm{HC}_{n,i}\approx N(0,1)$. Therefore, by the well-known results from empirical processes (e.g. Shorack and Wellner (2009)), $\mathrm{HC}_n^*\approx\sqrt{[2\log\{\log(n)\}]}$, which grows to ∞ very slowly. In contrast, if $X\sim N(\mu,I_n)$ where some of the co-ordinates of μ are non-zero, then $\mathrm{HC}_{n,i}$ has an elevated mean for some i, and HC_n^* could grow to ∞ algebraically fast. Consequently, higher criticism can separate two hypotheses even in the very sparse case. We mention that expression (7) is only one variant of higher criticism. See Donoho and Jin (2004, 2008, 2009) for further discussions.

In this paper, we study the detection problem in a more general setting, where the Gaussian mixture model is both heterogeneous and heteroscedastic and both the sparse and the dense cases are considered. We believe that heteroscedasticity is a more natural assumption in many applications. For example, signals can often bring additional variation to the background. This phenomenon can be captured by the Gaussian hierarchical model:

$$X_i|\mu\sim N(\mu,1), \qquad \mu\sim(1-\varepsilon_n)\delta_0+\varepsilon_n\,N(A_n,\tau^2),$$

where δ_0 denotes the point mass at 0. The marginal distribution is therefore

$$X_i\sim(1-\varepsilon_n)\,N(0,1)+\varepsilon_n\,N(A_n,\sigma^2), \qquad \sigma^2=1+\tau^2,$$

which is heteroscedastic as $\sigma>1$. In this paper, we consider the general heteroscedastic setting including $\sigma\geqslant1$ and $\sigma<1$.

In these detection problems a major focus is to characterize the so-called *detection boundary*, which is a curve that partitions the parameter space into two regions: the *detectable* region and the *undetectable* region. The study of the detection boundary is related to classical contiguity theory but is different in important ways. Adapting to our terminology, classical contiguity theory focuses on dense signals that are individually weak; the current paper, in contrast, focuses on sparse signals that individually may be moderately strong. As a result, to derive the detection boundary for the latter, we usually need unconventional analysis. In the case $\sigma=1$, the detection boundary was first discovered by Ingster (1997, 1999), and later independently by Donoho and Jin (2004) and Jin (2003, 2004).

In this paper, we derive the detection boundaries for both the sparse and the dense cases. It is shown that, if the parameters are known and are in the detectable region, the likelihood ratio test (LRT) has the sum of type I and type II error probabilities that tends to 0 as $n\to\infty$, which means that the LRT can asymptotically separate the alternative hypothesis from the null. We are particularly interested in understanding how the heteroscedastic effect may influence the detection boundary. Interestingly, in a certain range, the heteroscedasticity *alone* can separate the null and alternative hypotheses (i.e. even if the non-null effects have the same mean as that of the null effects).

The LRT is useful in determining the detection boundaries. It is, however, not practically useful as it requires knowledge of the parameter values. In this paper, in addition to the detection boundary, we also consider the practically more important problem of adaptive detection where the parameters β, r and σ are unknown. It is shown that a higher-criticism-based test is optimally adaptive in the whole detectable region in both the sparse and the dense cases, in spite of the very different detection boundaries and heteroscedasticity effects in the two cases. Classical methods treat the detections of sparse and dense signals separately. In real practice, however, the information on the signal sparsity is usually unknown, and the lack of a unified approach restricts discovery of the full catalogue of signals. The adaptivity of higher criticism that is found in this paper for both sparse and dense cases is a practically useful property. See further discussion in Section 3.

The detection of the presence of signals is of interest in its own right in many applications where, for example, the early detection of unusual events is critical. It is also closely related to other important problems in sparse inference such as estimation of the proportion of non-null effects and signal identification. The latter problem is a natural next step after detecting the presence of signals. In the current setting, both the proportion estimation problem and the signal identification problem can be solved by extensions of existing methods. See more discussion in Section 4.

The rest of the paper is organized as follows. Section 2 demonstrates the detection boundaries in the sparse and dense cases. Limiting behaviours of the LRT on the detection boundary are also presented. Section 3 introduces the modified higher criticism test and explains its optimal adaptivity through asymptotic theory and explanatory intuition. Comparisons with other methods are also presented. Section 4 discusses other closely related problems including estimation of proportions and signal identification. Simulation examples for finite n are given in Section 5. Further extensions and future work are discussed in Section 6. Main proofs are presented in Appendix A. Appendix B includes complementary technical details.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

http://www.blackwellpublishing.com/rss

2. Detection boundary

The meaning of the detection boundary can be elucidated as follows. In the β–r-plane with some σ fixed, we want to find a curve $r=\rho^*(\beta;\sigma)$, where $\rho^*(\beta;\sigma)$ is a function of β and σ, to separate the detectable region from the undetectable region. In the interior of the undetectable region, the sum of type I and type II error probabilities of any test tends to 1 as $n\to\infty$. In the interior of the detectable region, the sum of type I and type II errors of the Neyman–Pearson LRT with parameters (β,r,σ) specified tends to 0. The curve $r=\rho^*(\beta;\sigma)$ is called the detection boundary.

2.1. Detection boundary in the sparse case

In the sparse case, ε_n and A_n are calibrated as in expressions (3) and (4). We find the exact expression of $\rho^*(\beta;\sigma)$ as follows:

$$\rho^*(\beta;\sigma)=\begin{cases}(2-\sigma^2)(\beta-\frac{1}{2}), & \frac{1}{2}<\beta\leqslant 1-\sigma^2/4,\\ \{1-\sigma\surd(1-\beta)\}^2, & 1-\sigma^2/4<\beta<1,\end{cases}\qquad 0<\sigma<\surd 2, \tag{8}$$

and

$$\rho^*(\beta;\sigma)=\begin{cases}0, & \frac{1}{2}<\beta\leqslant 1-1/\sigma^2,\\ \{1-\sigma\surd(1-\beta)\}^2, & 1-1/\sigma^2<\beta<1,\end{cases}\qquad \sigma\geqslant\surd 2. \tag{9}$$

When $\sigma=1$, the detection boundary $r=\rho^*(\beta;\sigma)$ reduces to the detection boundary in Donoho and Jin (2004) (see also Ingster (1997, 1999) and Jin (2004)). The curve $r=\rho^*(\beta;\sigma)$ is plotted in Fig. 1(a) for $\sigma=0.6, 1, \surd 2, 3$. The detectable and undetectable regions correspond to $r>\rho^*(\beta;\sigma)$ and $r<\rho^*(\beta;\sigma)$ respectively.

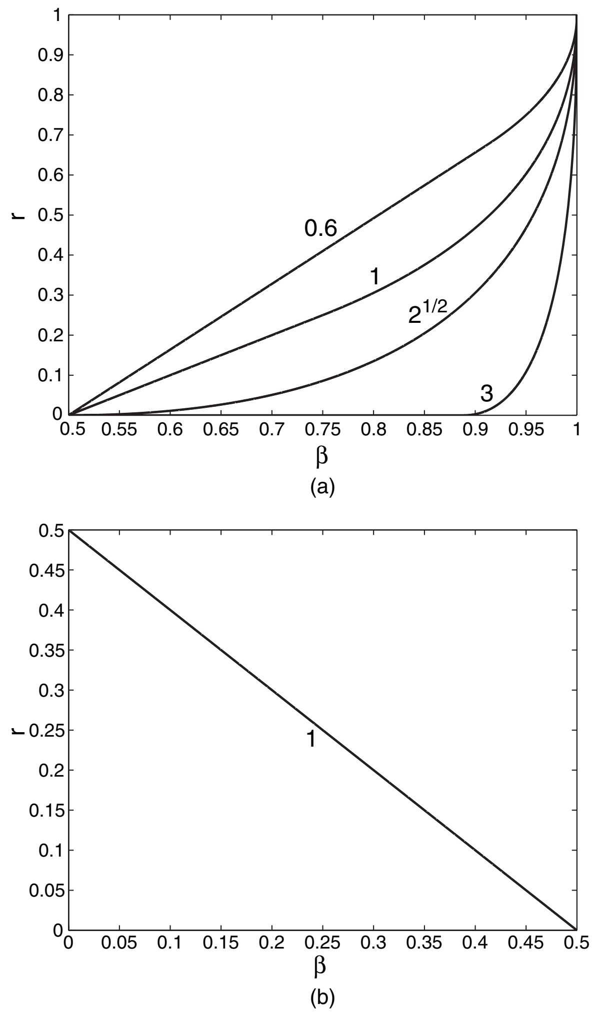


Fig. 1. (a) Detection boundary $r=\rho^*(\beta;\sigma)$ in the sparse case for $\sigma=0.6,1,\surd 2,3$ (the detectable region is $r>\rho^*(\beta;\sigma)$ and the undetectable region is $r<\rho^*(\beta;\sigma)$) and (b) detection boundary $r=\rho^*(\beta;\sigma)$ in the dense case for $\sigma=1$ (the detectable region is $r<\rho^*(\beta;\sigma)$ and the undetectable region is $r>\rho^*(\beta;\sigma)$; note that $\rho(\beta;\sigma)=\infty$ in the dense case when $\sigma\neq 1$)

When $r<\rho^*(\beta;\sigma)$, the Hellinger distance between the joint density of X_i under the null hypothesis and that under the alternative tends to 0 as $n\to\infty$, which implies that the sum of type I and type II error probabilities for any test tends to 1. Therefore no test could successfully separate these two hypotheses in this situation. The following theorem is proved in Appendix A.1.

Theorem 1. Let ε_n and A_n be calibrated as in expression (3) and (4) and let $\sigma>0$, $\beta\in(\frac{1}{2},1)$, and $r\in(0,1)$ be fixed such that $r<\rho^*(\beta;\sigma)$, where $\rho^*(\beta;\sigma)$ is as in expressions (8) and (9). Then for any test the sum of type I and type II error probabilities tends to 1 as $n\to\infty$.

When $r>\rho^*(\beta;\sigma)$, it is possible to separate the hypotheses successfully, and we show that the classical LRT can do so. In detail, denote the likelihood ratio by

$$\mathrm{LR}_n=\mathrm{LR}_n(X_1,X_2,\ldots,X_n;\beta,r,\sigma),$$

and consider the LRT which rejects H_0 if and only if

$$\log(\mathrm{LR}_n)>0. \tag{10}$$

The following theorem, which is proved in Appendix A.2, shows that, when $r>\rho^*(\beta;\sigma)$, $\log(\mathrm{LR}_n)$ converges to $\mp\infty$ in probability, under the null and the alternative hypotheses respectively. Therefore, asymptotically the alternative hypothesis can be perfectly separated from the null by the LRT.

Theorem 2. Let ε_n and A_n be calibrated as in expressions (3) and (4) and let $\sigma>0$, $\beta\in(\frac{1}{2},1)$, and $r\in(0,1)$ be fixed such that $r>\rho^*(\beta;\sigma)$, where $\rho^*(\beta;\sigma)$ is as in expressions (8) and (9). As $n\to\infty$, $\log(\mathrm{LR}_n)$ converges to $\mp\infty$ in probability, under the null and the alternative hypotheses respectively. Consequently, the sum of type I and type II error probabilities of the LRT tends to 0.

The effect of heteroscedasticity is illustrated in Fig. 1(a). As σ increases, the curve $r=\rho^*(\beta;\sigma)$ moves towards the bottom right-hand corner; the detectable region becomes larger which implies that the detection problem becomes easier. Interestingly, there is a 'phase change' as σ varies, with $\sigma=\sqrt{2}$ being the critical point. When $\sigma<\sqrt{2}$, it is always undetectable if A_n is 0 or very small, and the effect of heteroscedasticity alone would not yield successful detection. When $\sigma>\sqrt{2}$, it is, however, detectable even when $A_n=0$, and the effect of heteroscedasticity alone may produce successful detection.

2.2. Detection boundary in the dense case

In the dense case, ε_n and A_n are calibrated as in expressions (3) and (5). We find the detection boundary as $r=\rho^*(\beta;\sigma)$, where

$$\rho^*(\beta;\sigma)=\begin{cases}\infty, & \sigma\neq 1,\\ \frac{1}{2}-\beta, & \sigma=1,\end{cases}\qquad 0<\beta<\tfrac{1}{2}. \tag{11}$$

The curve $r=\rho^*(\beta;\sigma)$ is plotted in Fig. 1(b) for $\sigma=1$. Unlike in the sparse case, the detectable and undetectable regions now correspond to $r<\rho^*(\beta;\sigma)$ and $r>\rho^*(\beta;\sigma)$ respectively.

The following results are analogous to those in the sparse case. We show that, when $r>\rho^*(\beta;\sigma)$, no test could separate H_0 from $H_1^{(n)}$, and, when $r<\rho^*(\beta;\sigma)$, asymptotically the LRT can perfectly separate the alternative hypothesis from the null. Proofs for the following theorems are included in Appendices A.3 and A.4.

Theorem 3. Let ε_n and A_n be calibrated as in expressions (3) and (5) and let $\sigma > 0$, $\beta \in (0, \frac{1}{2})$ and $r \in (0, \frac{1}{2})$ be fixed such that $r > \rho^*(\beta;\sigma)$, where $\rho^*(\beta;\sigma)$ is defined in expression (11). Then for any test the sum of type I and type II error probabilities tends to 1 as $n \to \infty$.

Theorem 4. Let ε_n and A_n be calibrated as in expressions (3) and (5) and let $\sigma > 0$, $\beta \in (0, \frac{1}{2})$, and $r \in (0, \frac{1}{2})$ be fixed such that $r < \rho^*(\beta;\sigma)$, where $\rho^*(\beta;\sigma)$ is defined in expression (11). Then, the sum of type I and type II error probabilities of the LRT tends to 0 as $n \to \infty$.

Comparing expression (11) with expressions (8) and (9), we see that the detection boundary in the dense case is very different from that in the sparse case. In particular, the non-null component is always detectable for any $r \in (0, \frac{1}{2})$ when $\sigma \neq 1$. In the dense case, the proportion of non-null components is so large that a small heteroscedastic effect can be amplified to make the non-null component detectable. In contrast, when $\sigma = 1$, a small heterogeneous effect also makes a big difference. This is essentially why the calibrations of ε_n and A_n, and the detection boundary are very different in the dense case from those in the sparse case.

The dividing line between the sparse and dense case is $\beta = \frac{1}{2}$. In the case when β exactly equals $\frac{1}{2}$, A_n can be calibrated as a constant. Then, by a similar analysis, it can be shown that in such a setting the Hellinger distance between the joint density of observations under the null and that under the alternative hypothesis tends to some constant between 0 and 1. Furthermore, the sum of type I and type II error probabilities of the LRT also tends to some constant between 0 and 1. Therefore, the non-null effect is only partially detectable when $\beta = \frac{1}{2}$ and A_n is a constant. In contrast, if $A_n \to \infty$ at any rate, then the signals can be reliably detected.

2.3. *Limiting behaviour of the likelihood ratio test on the detection boundary*

In the preceding section, we examined the situation when the parameters (β, r) fall strictly in the interior of either the detectable or the undetectable region. When these parameters become very close to the detection boundary, the behaviour of the LRT becomes more subtle. In this section, we discuss the behaviour of the LRT when σ is fixed and the parameters (β, r) fall exactly on the detection boundary. We show that, up to some lower order term corrections of ε_n, the LRT converges to different non-degenerate distributions under the null and under the alternative hypothesis, and, interestingly, the limiting distributions are not always Gaussian. As a result, the sum of type I and type II errors of the optimal test tends to some constant $\alpha \in (0,1)$. The discussion for the dense case is similar to that for the sparse case, but simpler. For brevity, we present only the details for the sparse case.

We introduce the following calibration:

$$A_n = \surd\{2r \log(n)\}, \qquad \varepsilon_n = \begin{cases} n^{-\beta}, & \frac{1}{2} < \beta \leqslant 1 - \sigma^2/4, \\ n^{-\beta} \log(n)^{1-\surd(1-\beta)/\sigma}, & 1 - \sigma^2/4 < \beta < 1. \end{cases} \tag{12}$$

Compared with the calibrations in expressions (3) and (4), A_n remains the same but ε_n is modified slightly so that the limiting distribution of the LRT would be non-degenerate. Denote

$$b(\sigma) = \{\sigma\surd(2-\sigma^2)\}^{-1}.$$

We introduce two characteristic functions $\exp(\psi^0_{\beta,\sigma})$ and $\exp(\psi^1_{\beta,\sigma})$, where

$$\psi^0_{\beta,\sigma}(t) = \frac{1}{2\surd\pi\sigma^{1/(\sigma^2-1)}\{\sigma - \surd(1-\beta)\}} \int_{-\infty}^{\infty} (\exp[it\log\{1+\exp(y)\}] - 1 \\ - it\exp(y)) \exp\left\{\frac{\sigma - 2\surd(1-\beta)}{\sigma - \surd(1-\beta)} - 2\right\} y \, dy$$

and

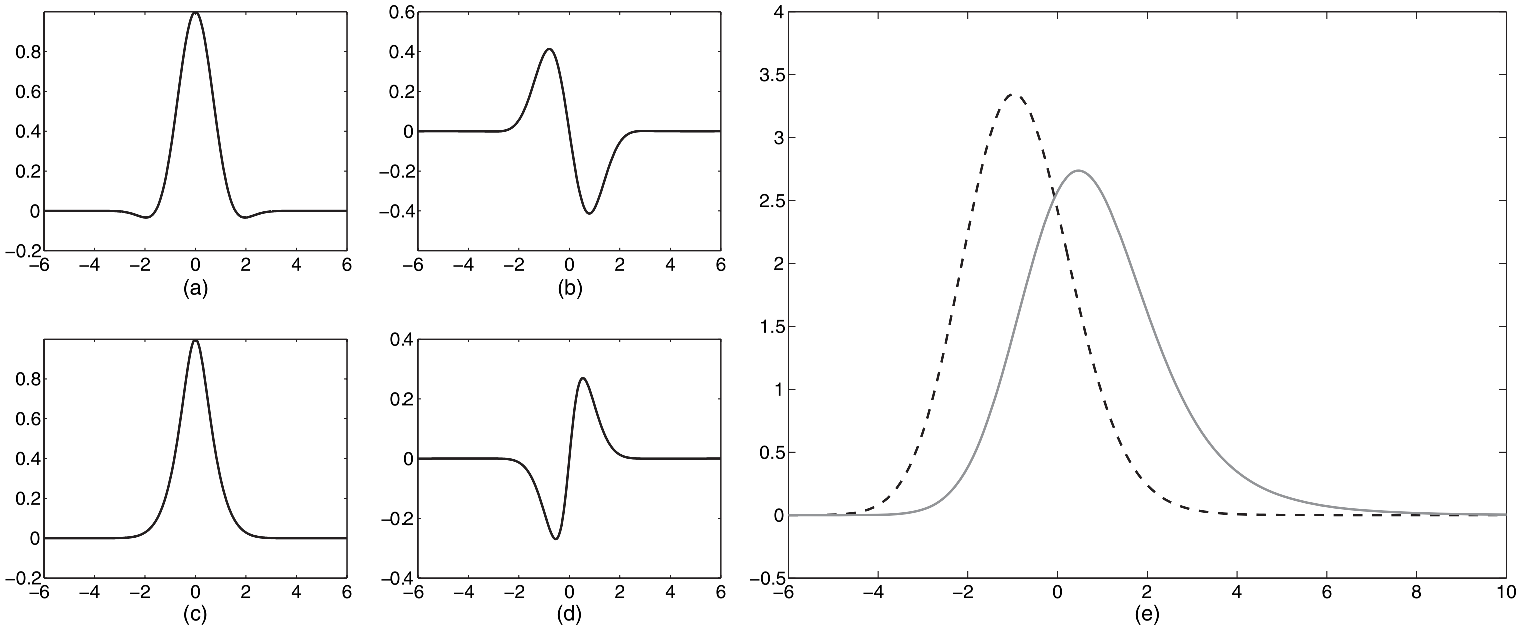


Fig. 2. Characteristic functions and density functions of $\log(\mathrm{LR}_n)$ for $(\beta,\sigma) = (0.75, 1.1)$: (a) real part of $\exp(\psi^0_{\beta,\sigma})$; (b) imaginary part of $\exp(\psi^0_{\beta,\sigma})$; (c) real part of $\exp(\psi^0_{\beta,\sigma} + \psi^1_{\beta,\sigma})$; (d) imaginary part of $\exp(\psi^0_{\beta,\sigma} + \psi^1_{\beta,\sigma})$; (e) density functions of $\nu^0_{\beta,\sigma}$ (-------) and $\nu^1_{\beta,\sigma}$ (———)

$$\psi^1_{\beta,\sigma}(t) = \frac{1}{2\sqrt{\pi\sigma^{\sigma^2/(\sigma^2-1)}}\{\sigma-\sqrt{(1-\beta)}\}} \int_{-\infty}^{\infty} (\exp[it\log\{1+\exp(y)\}]-1) \times \exp\left\{\frac{\sigma-2\sqrt{(1-\beta)}}{\sigma-\sqrt{(1-\beta)}}-1\right\} y \,\mathrm{d}y,$$

and let $\nu^0_{\beta,\sigma}$ and $\nu^1_{\beta,\sigma}$ be the corresponding distributions. We have the following theorems, which address the case of $\sigma<\sqrt{2}$ and the case of $\sigma\geqslant\sqrt{2}$.

Theorem 5. Let A_n and ε_n be defined as in expression (12), and let $\rho^*(\beta;\sigma)$ be as in expressions (8) and (9). Fix $\sigma\in(0,\sqrt{2})$ and $\beta\in(\frac{1}{2},1)$, and set $r=\rho^*(\beta,\sigma)$. As $n\to\infty$, under hypothesis H_0,

$$\log(\mathrm{LR}_n) \xrightarrow{\mathrm{L}} \begin{cases} N\{-b(\sigma)/2, b(\sigma)\}, & \frac{1}{2}<\beta<1-\sigma^2/4, \\ N\{-b(\sigma)/4, b(\sigma)/2\}, & \beta=1-\sigma^2/4, \\ \nu^0_{\beta,\sigma}, & 1-\sigma^2/4<\beta<1, \end{cases}$$

and, under hypothesis $H_1^{(n)}$,

$$\log(\mathrm{LR}_n) \xrightarrow{\mathrm{L}} \begin{cases} N\{b(\sigma)/2, b(\sigma)\}, & \frac{1}{2}<\beta<1-\sigma^2/4, \\ N\{b(\sigma)/4, b(\sigma)/2\}, & \beta=1-\sigma^2/4, \\ \nu^1_{\beta,\sigma}, & 1-\sigma^2/4<\beta<1, \end{cases}$$

where $\to^{\mathrm{L}}$ denotes 'converges in law'.

The limiting distribution is Gaussian when $\beta\leqslant 1-\sigma^2/4$ and non-Gaussian otherwise. Next, we consider the case of $\sigma\geqslant\sqrt{2}$, where the range of interest is $\beta>1-1/\sigma^2$.

Theorem 6. Let $\sigma\in[\sqrt{2},\infty)$ and $\beta\in(1-1/\sigma^2,1)$ be fixed. Set $r=\rho^*(\beta,\sigma)$ and let A_n and ε_n be as in expression (12). Then, as $n\to\infty$,

$$\log(\mathrm{LR}_n) \xrightarrow{\mathrm{L}} \begin{cases} \nu^0_{\beta,\sigma}, & \text{under hypothesis } H_0, \\ \nu^1_{\beta,\sigma}, & \text{under hypothesis } H_1^{(n)}. \end{cases}$$

In this case, the limiting distribution is always non-Gaussian. This phenomenon (i.e. the weak limits of the log-likelihood ratio might be non-Gaussian) was repeatedly discovered in the literature. See for example Ingster (1997, 1999), Jin (2003, 2004) for the case $\sigma=1$, and Burnashev and Begmatov (1991) for a closely related setting.

In Fig. 2, we fix $(\beta,\sigma)=(0.75,1.1)$ and plot the characteristic functions and the density functions corresponding to the limiting distribution of $\log(\mathrm{LR}_n)$. Two density functions are generally overlapping each other, which suggests that, when (β,r,σ) falls on the detection boundary, the sum of type I and type II error probabilities of the LRT tends to a fixed number in $(0,1)$ as $n\to\infty$.

3. Higher criticism and its optimal adaptivity

In real applications, the explicit values of model parameters are usually unknown. Hence it is of great interest to develop adaptive methods that can perform well without information on model parameters. We find that higher criticism, which is a non-parametric procedure, is successful in the entire detectable region for both the sparse and the dense cases. This property is called the *optimal adaptivity* of higher criticism. Donoho and Jin (2004) discovered this property in the case $\sigma = 1$ and $\beta \in (\frac{1}{2}, 1)$. Here, we consider more general settings where β ranges from 0 to 1 and σ ranges from 0 to ∞. Both parameters are fixed but unknown.

We modify the higher criticism statistic by using the absolute value of $\mathrm{HC}_{n,i}$:

$$\mathrm{HC}_n^* = \max_{1 \leqslant i \leqslant n} |\mathrm{HC}_{n,i}|, \tag{13}$$

where $\mathrm{HC}_{n,i}$ is defined as in expression (7). Recall that, under the null hypothesis,

$$\mathrm{HC}_n^* \approx \surd[2 \log\{\log(n)\}].$$

So a convenient critical point for rejecting the null hypothesis is when

$$\mathrm{HC}_n^* \geqslant \surd[2(1+\delta) \log\{\log(n)\}], \tag{14}$$

where $\delta > 0$ is any fixed constant. The following theorem is proved in Appendix A.5.

Theorem 7. Suppose that ε_n and A_n either satisfy expressions (3) and (4) and $r > \rho^*(\beta;\sigma)$ with $\rho^*(\beta;\sigma)$ defined as in expressions (8) and (9), or ε_n and A_n satisfy expressions (3) and (5) and $r < \rho^*(\beta;\sigma)$ with $\rho^*(\beta;\sigma)$ defined as in expression (11). Then the test which rejects H_0 if and only if $\mathrm{HC}_n^* \geqslant \surd[2(1+\delta) \log\{\log(n)\}]$ satisfies

$$P_{H_0}(\text{reject } H_0) + P_{H_1^{(n)}}(\text{reject } H_1^{(n)}) \to 0 \qquad \text{as } n \to \infty.$$

Theorem 7 states, somewhat surprisingly, that the optimal adaptivity of higher criticism continues to hold even when the data pose an unknown degree of heteroscedasticity, both in the sparse regime and in the dense regime. It is also clear that the type I error tends to 0 faster for a higher threshold. Higher criticism can successfully separate two hypotheses whenever it is possible to do so, and it has full power in the region where LRT has full power. But, unlike the LRT, higher criticism does not need specific information of the parameters σ, β and r.

In practice, we would like to pick a critical value so that the type I error is controlled at a prescribed level α. A convenient way to do this is as follows. Fix a large number N such that $N\alpha \gg 1$ (e.g. $N\alpha = 50$). We simulate the HC_n^*-scores under the null hypothesis for N times, and let $t(\alpha)$ be the top α percentile of the simulated scores. We then use $t(\alpha)$ as the critical value. With a typical office desktop computer, the simulation experiment can be finished reasonably fast. We find that, owing to the slow convergence of the iterative logarithmic law, critical values determined in this way are usually much more accurate than $\surd[2(1+\delta) \log\{\log(n)\}]$.

3.1. *How higher criticism works*

We now illustrate how higher criticism manages to capture the evidence against the joint null hypothesis without information on model parameters (σ, β, r).

To begin with, we rewrite the higher criticism in an equivalent form. Let $F_n(t)$ and $\bar{F}_n(t)$ be the empirical cumulative distribution function and empirical survival function of X_i respectively,

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i < t\}},$$

$$\bar{F}_n(t) = 1 - F_n(t),$$

and let $W_n(t)$ be the standardized form of $\bar{F}_n(t) - \bar{\Phi}(t)$,

$$W_n(t) = \frac{\bar{F}_n(t) - \bar{\Phi}(t)}{\surd[\bar{\Phi}(t)\{1 - \bar{\Phi}(t)\}]} \surd n. \tag{15}$$

Consider the value t that satisfies $\bar{\Phi}(t)=p_{(i)}$. Since there are exactly i p-values less than or equal to $p_{(i)}$, so there are exactly i samples from $\{X_1, X_2, \ldots, X_n\}$ that are greater than or equal to t. Hence, for this particular t, $\bar{F}_n(t)=i/n$, and so

$$W_n(t)=\frac{i/n-p_{(i)}}{\sqrt{\{p_{(i)}(1-p_{(i)})\}}}\sqrt{n}.$$

Comparing this with equation (13), we have

$$\mathrm{HC}_n^* = \sup_{-\infty<t<\infty} |W_n(t)|. \tag{16}$$

The proof of equation (16), which we omit, is elementary. Now, note that, for any fixed t,

$$E[W_n(t)]=\begin{cases} 0, & \text{under } H_0, \\ \sqrt{n}\,\dfrac{\bar{F}(t)-\bar{\Phi}(t)}{\sqrt{[\bar{\Phi}(t)\{1-\bar{\Phi}(t)\}]}}, & \text{under } H_1^{(n)}. \end{cases}$$

The idea is that, if, for some threshold t_n,

$$\left|\sqrt{n}\,\frac{\bar{F}(t_n)-\bar{\Phi}(t_n)}{\sqrt{[\bar{\Phi}(t_n)\{1-\bar{\Phi}(t_n)\}]}}\right| \gg \sqrt{[2\log\{\log(n)\}]} \tag{17}$$

then we can test H_0 against $H_1^{(n)}$ by merely applying thresholding on $W_n(t_n)$. This guarantees the success of detection of higher criticism.

For the case $\frac{1}{2}<\beta<1$, we introduce the notion of the ideal threshold, $t_n^{\mathrm{ideal}}(\beta, r, \sigma)$, which is a functional of (β, r, σ, n) that maximizes $|E[W_n(t)]|$ under the alternative:

$$t_n^{\mathrm{ideal}}(\beta, r, \sigma)=\arg\max_t \left|\sqrt{n}\,\frac{\bar{F}(t)-\bar{\Phi}(t)}{\sqrt{[\bar{\Phi}(t)\{1-\bar{\Phi}(t)\}]}}\right|. \tag{18}$$

The leading term of $t_n^{\mathrm{ideal}}(\beta, r, \sigma)$ turns out to have a rather simple form. In detail, let

$$t_n^*(\beta, r, \sigma)=\begin{cases} \min\left[\dfrac{2}{2-\sigma^2}A_n, \sqrt{\{2\log(n)\}}\right], & \sigma<\sqrt{2}, \\ \sqrt{\{2\log(n)\}}, & \sigma \geqslant \sqrt{2}. \end{cases} \tag{19}$$

The following lemma is proved in Appendix B.

Lemma 1. Let ε_n and A_n be calibrated as in expressions (3) and (4). Fix $\sigma>0$, $\beta\in(\frac{1}{2}, 1)$ and $r\in(0, 1)$ such that $r>\rho^*(\beta, r, \sigma)$, where $\rho^*(\beta, r, \sigma)$ is defined in expressions (8) and (9). Then

$$\frac{t_n^{\mathrm{ideal}}(\beta, r, \sigma)}{t_n^*(\beta, r, \sigma)} \to 1 \qquad \text{as } n\to\infty.$$

In the dense case when $0<\beta<\frac{1}{2}$, the analysis is much simpler. In fact, condition (17) holds under the alternative if $A_n \ll t \leqslant C$ for some constant C. To show the result, we can simply set the threshold as

$$t_n^*(\beta, r, \sigma)=1; \tag{20}$$

then it follows that

$$|E[W_n(1)]| \gg \sqrt{[2\log\{\log(n)\}]}.$$

We might have expected A_n to be the best threshold as it represents the strength of the signal. Interestingly, this turns out to be not so: the ideal threshold, as derived in the oracle situation when the values of (σ, β, r) are known, is nowhere near A_n. In fact, in the sparse case, the ideal threshold is either near $\{2/(2-\sigma^2)\}A_n$ or near $\sqrt{\{2\log(n)\}}$; both are larger than A_n. In the dense case, the ideal threshold is near a constant, which is also much larger than A_n. The elevated threshold is due to sparsity (note that, even in the dense case, the signals are outnumbered by noise): one must raise the threshold to counter the fact that there is merely much more noise than signals.

Finally, the optimal adaptivity of higher criticism comes from the 'sup' part of its definition (see expression (16)). When the null hypothesis is true, by the study on empirical processes (Shorack and Wellner, 2009), the supremum of $W_n(t)$ over all t is not substantially larger than that of $W_n(t)$ at a single t. But, when the alternative is true, simply because

$$\mathrm{HC}_n^* \geqslant W_n\{t_n^{\mathrm{ideal}}(\sigma, \beta, r)\},$$

the value of higher criticism is no smaller than that of $W_n(t)$ evaluated at the ideal threshold (which is unknown to us!). In essence, higher criticism mimics the performance of $W_n\{t_n^{\mathrm{ideal}}(\sigma, \beta, r)\}$, although the parameters (σ, β, r) are unknown. This explains the optimal adaptivity of higher criticism.

Does the higher criticism continue to be optimal when (β, r) falls exactly on the boundary, and how do we improve this method if it ceases to be optimal in such a case? The question is interesting but the answer is not immediately clear. In principle, given the literature on empirical processes and law of iterative logarithms, it is possible to modify the normalizing term of $\mathrm{HC}_{n,i}$ so that the resultant higher criticism statistic has a better power. Such a study involves the second-order asymptotic expansion of the higher criticism statistic, which not only requires substantially more delicate analysis but also is comparably less important from a practical point of view than the analysis that is considered here. For these reasons, we leave the exploration along this line to the future.

3.2. Comparison with other testing methods

A classical and frequently used approach for testing is based on the extreme value

$$\mathrm{Max}_n = \mathrm{Max}_n(X_1, X_2, \ldots, X_n) = \max_{\{1 \leqslant i \leqslant n\}} \{X_i\}.$$

The approach is intrinsically related to multiple-testing methods including that of Bonferroni and that of controlling the false discovery rate.

Recall that, under the null hypothesis, X_i are independent and identically distributed (IID) samples from $N(0,1)$. It is well known (e.g. Shorack and Wellner (2009)) that

$$\lim_{n\to\infty} [\mathrm{Max}_n / \sqrt{\{2\log(n)\}}] \to 1, \qquad \text{in probability.}$$

Additionally, if we reject H_0 if and only if

$$\mathrm{Max}_n \geqslant \sqrt{\{2\log(n)\}}, \tag{21}$$

then the type I error tends to 0 as $n \to \infty$. For brevity, we call the test in inequality (21) Max_n.

Now, suppose that the alternative hypothesis is true. In this case, X_i splits into two groups, where one contains $n(1-\varepsilon_n)$ samples from $N(0,1)$ and the other contains $n\varepsilon_n$ samples from $N(A_n, \sigma^2)$. Consider the sparse case first. In this case, $A_n = \sqrt{\{2r\log(n)\}}$ and $n\varepsilon_n = n^{1-\beta}$. It follows that, except for a negligible probability, the extreme value of the first group is approximately $\sqrt{\{2\log(n)\}}$, and that of the second group approximately $\sqrt{\{2r\log(n)\}} + \sigma\sqrt{\{2(1-\beta)\log(n)\}}$. Since Max_n equals the larger of the two extreme values,

$$\mathrm{Max}_n \approx \surd\{2\log(n)\}\max\{1, \surd r + \sigma\surd(1-\beta)\}.$$

So, as $n \to \infty$, the type II error of test (21) tends to 0 if and only if

$$\surd r + \sigma\surd(1-\beta) > 1.$$

This is trivially satisfied when $\sigma\surd(1-\beta) > 1$. The discussion is summarized in the following theorem, the proof of which is omitted.

Theorem 8. Let ε_n and A_n be calibrated as in expressions (3) and (4). Fix $\sigma > 0$ and $\beta \in (\frac{1}{2}, 1)$. As $n \to \infty$, the sum of type I and type II error probabilities of test (21) tends to 0 if $r > \{(1-\sigma\surd(1-\beta))_+\}^2$ and tends to 1 if $r < \{(1-\sigma\surd(1-\beta))_+\}^2$.

Note that the region where Max_n is successful is substantially smaller than that of higher criticism in the sparse case. Therefore, the extreme value test is only suboptimal. Although the comparison is for the sparse case, we note that the dense case is even more favourable for higher criticism. In fact, as $n \to \infty$, the power of Max_n tends to 0 as long as A_n is algebraically small in the dense case.

Other classical tests include tests based on the sample mean, Hotelling's test and Fisher's combined probability test. These tests have the form of $\Sigma_{i=1}^n f(X_i)$ for some function f. In fact, Hotelling's test can be recast as $\Sigma_{i=1}^n X_i^2$, and Fisher's combined probability test can be recast as $-2\,\Sigma_{i=1}^n \bar{\Phi}(X_i)$. The key fact is that the standard deviations of such tests usually are of the order of $\surd n$. But, in the sparse case, the number of non-null effects is much less than $\surd n$. Therefore, these tests cannot separate the two hypotheses in the sparse case.

4. Detection and related problems

The detection problem that is studied in this paper has close connections to other important problems in sparse inference including estimation of the proportion of non-null effects and signal identification. In the current setting, both the proportion estimation problem and the signal identification problem can be solved easily by extensions of existing methods. For example, Cai *et al.* (2007) provided rate optimal estimates of the signal proportion ε_n and signal mean A_n for the homoscedastic Gaussian mixture $X_i \sim (1-\varepsilon_n)\,N(0,1) + \varepsilon_n\,N(A_n, 1)$. The techniques developed by Cai *et al.* (2007) can be generalized to estimate the parameters ε_n, A_n and σ in the current heteroscedastic Gaussian mixture setting, $X_i \sim (1-\varepsilon_n)\,N(0,1) + \varepsilon_n\,N(A_n, \sigma^2)$, for both sparse and dense cases.

After detecting the presence of signals, a natural next step is to identify the locations of the signals. Equivalently, we wish to test the hypotheses

$$H_{0,i}: X_i \sim N(0,1) \quad \textit{versus} \quad H_{1,i}: X_i \sim N(A_n, \sigma^2) \tag{22}$$

for $1 \leqslant i \leqslant n$. An immediate question is when are the signals identifiable? It is intuitively clear that it is more difficult to identify the locations of the signals than to detect the presence of the signals. To illustrate the gap between the difficulties of detection and signal identification, we study the situation when signals are detectable but not identifiable. For any multiple-testing procedure $\hat{T}_n = \hat{T}_n(X_1, X_2, \ldots, X_n)$, its performance can be measured by the misclassification error

$$\mathrm{Err}(\hat{T}_n) = E[\#\{i: H_{0,i} \text{ is either falsely rejected or falsely accepted }, 1 \leqslant i \leqslant n\}].$$

We calibrate ε_n and A_n by

$$\varepsilon_n = n^{-\beta},$$
$$A_n = \surd\{2r \log(n)\}.$$

The above calibration is the same as in the sparse case ($\beta > \frac{1}{2}$) (see expression (4)), but different from the dense case ($\beta < \frac{1}{2}$) (see expression (5)). The following theorem is a straightforward extension of Ji and Jin's (2010) theorem 1.1, so we omit the proof. See also Xie *et al.* (2011).

Theorem 9. Fix $\beta \in (0,1)$ and $r \in (0,\beta)$. For any sequence of multiple-testing procedures $\{\hat{T}_n\}_{n=1}^{\infty}$,

$$\liminf_{n\to\infty} \left\{ \frac{\mathrm{Err}(\hat{T}_n)}{n\varepsilon_n} \right\} \geqslant 1.$$

Theorem 9 shows that, if the signal strength is relatively weak, i.e. $A_n = \surd\{2r \log(n)\}$ for some $0 < r < \beta$, then it is impossible to separate the signals from noise successfully: no identification method can essentially perform better than the naive procedure which simply classifies all observations as noise. The misclassification error of the naive procedure is obviously $n\varepsilon_n$.

Theorems 7 and 9 together depict a picture as follows. Suppose that

$$\begin{aligned} A_n &< \surd\{2\beta \log(n)\}, && \text{if } \tfrac{1}{2} < \beta < 1, \\ n^{\beta-1/2} \ll A_n &< \surd\{2\beta \log(n)\}, && \text{if } 0 < \beta < \tfrac{1}{2}. \end{aligned} \qquad (23)$$

Then it is possible to detect the presence of the signals reliably but it is impossible to identify the locations of the signals simply because the signals are too sparse and weak. In other words, the signals are detectable, but not identifiable.

A practical signal identification procedure can be readily obtained for the current setting from the general multiple-testing procedure that was developed in Sun and Cai (2007). By viewing test (22) as a multiple-testing problem, we wish to test the hypotheses $H_{0,i}$ *versus* $H_{1,i}$ for all $i = 1, \ldots, n$. A commonly used criterion in multiple testing is to control the false discovery rate FDR at a given level, say, FDR $\leqslant \alpha$. Equipped with consistent estimates $(\hat{\varepsilon}_n, \hat{A}_n, \hat{\sigma})$, we can specialize the general adaptive testing procedure that was proposed in Sun and Cai (2007) to solve the signal identification problem in the current setting. Define

$$\widehat{\mathrm{Lfdr}}(x) = \frac{(1-\hat{\varepsilon}_n)\,\phi(x)}{(1-\hat{\varepsilon}_n)\,\phi(x) + \hat{\varepsilon}_n\,\phi\{(x-\hat{A}_n)/\hat{\sigma}\}}.$$

The adaptive procedure has three steps. First calculate the observed $\widehat{\mathrm{Lfdr}}(X_i)$ for $i = 1, \ldots, n$. Then rank $\widehat{\mathrm{Lfdr}}(X_i)$ in an increasing order: $\widehat{\mathrm{Lfdr}}_{(1)} \leqslant \widehat{\mathrm{Lfdr}}_{(2)} \leqslant \ldots \leqslant \widehat{\mathrm{Lfdr}}_{(n)}$. Finally reject all $H_0^{(i)}, i = 1, \ldots, k$, where $k = \max\{i : (1/i)\Sigma_{j=1}^{i} \widehat{\mathrm{Lfdr}}_{(j)} \leqslant \alpha\}$. This adaptive procedure asymptotically attains the performance of an oracle procedure and thus is optimal for the multiple-testing problem. See Sun and Cai (2007) for further details.

We conclude this section with another important problem that is intimately related to signal detection: feature selection and classification. Suppose that there are n subjects that are labelled into two classes, and for each subject we have measurements of p features. The goal is to use the data to build a trained classifier to predict the label of a new subject by measuring its feature vectors. Donoho and Jin (2008) and Jin (2009) showed that the optimal threshold for feature selection is intimately connected to the ideal threshold for detection in Section 3.1, and the fundamental limit for classification is intimately connected to the detection boundary.

Although the scope in these works is limited to the homoscedastic case, extensions to heteroscedastic cases are possible. From a practical point of view, the latter is in fact broader and more attractive.

5. Simulation

In this section, we report simulation results, where we investigate the performance of four tests: the LRT, higher criticism, Max and the sample mean SM, which is defined below. The LRT is defined in expression (10); higher criticism is defined in expression (14) where the tuning parameter δ is taken to be the optimal value in $0.2 \times [0, 1, \ldots, 10]$ that results in the smallest sum of type I and type II errors; Max is defined in expression (21). In addition, denoting

$$\bar{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j,$$

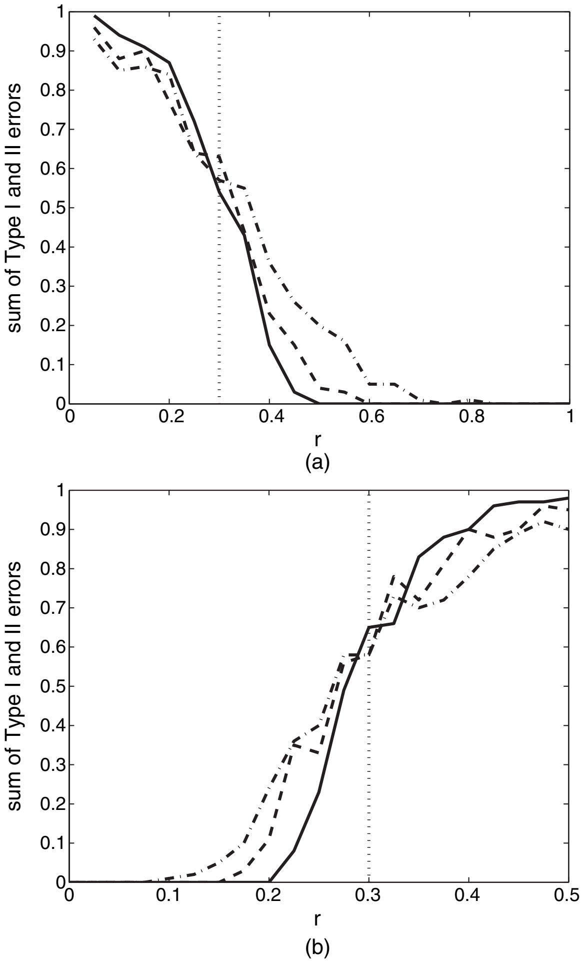


Fig. 3. Sum of type I and type II errors of the LRT, 100 replications (———, $n = 10^7$; – – –, $n = 10^5$; ·–·–·, $n = 10^4$; ⋮, critical point of $r = p^*(\beta; \sigma)$): (a) $(\beta, \sigma^2) = (0.7, 0.5)$, $r = 0.05, 0.10, \ldots, 1$; (b) $(\beta, \sigma) = (0.2, 1)$, $r = 1/30, 1/15, \ldots, 0.5$

let SM be the test that rejects H_0 when $\sqrt{n}\bar{X}_n > \sqrt{[\log\{\log(n)\}]}$ (note that $\sqrt{n}\bar{X}_n \sim N(0,1)$ under hypothesis H_0). SM is an example in the general class of moment-based tests. Note that the use of the LRT needs specific information of the underlying parameters (β, r, σ), but higher criticism, Max and SM do not need such information.

The main steps for the simulation are as follows. First, fixing parameters (n, β, r, σ), we let $\varepsilon_n = n^{-\beta}$, $A_n = \sqrt{\{2r\log(n)\}}$ if $\beta > \frac{1}{2}$ and $A_n = n^{-r}$ if $\beta < \frac{1}{2}$ as before. Second, for the null hypothesis, we drew n samples from $N(0,1)$; for the alternative hypothesis, we first drew $n(1-\varepsilon_n)$ samples from $N(0,1)$ and then draw $n\varepsilon_n$ samples from $N(A_n, 1)$. Third, we implemented all four tests for each of these two samples. Last, we repeated the whole process 100 times independently and then recorded the empirical type I error and type II errors for each test. The simulation contains four experiments below.

5.1. Experiment 1

In experiment 1, we investigate how the LRT performs and how relevant the theoretic detection

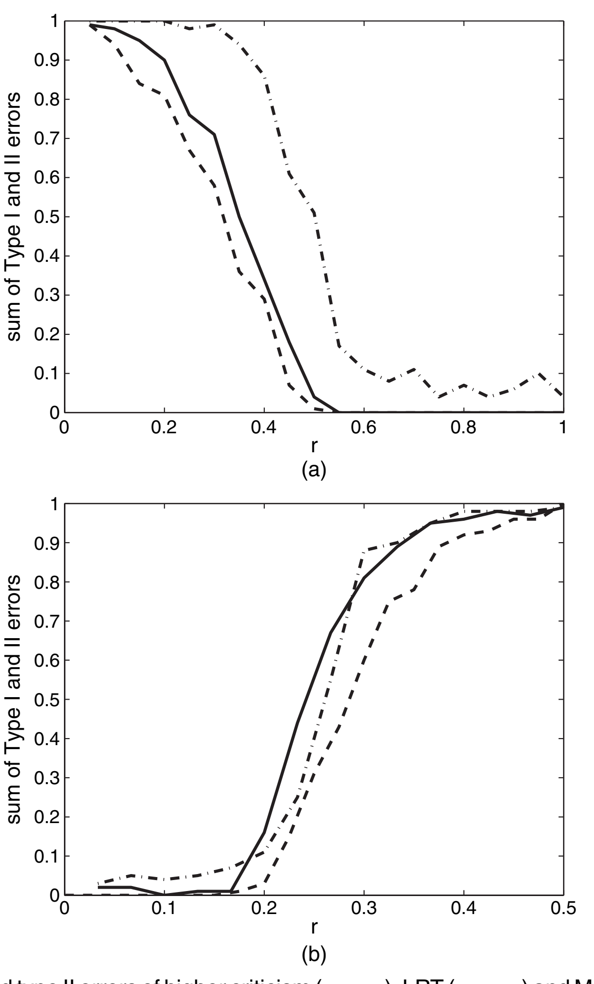
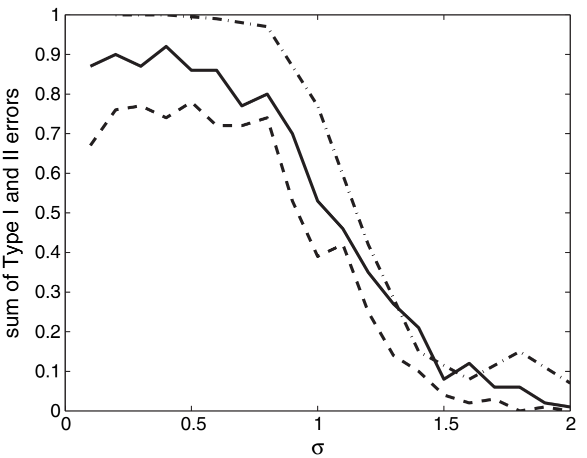


Fig. 4. Sum of type I and type II errors of higher criticism (———), LRT (– – –) and Max (·–·–·) (in (a)) or SM (·–·–·) (in (b)), 100 replications: (a) $(n, \beta, \sigma^2) = (10^6, 0.7, 0.5)$, $r = 0.05, 0.10, \ldots, 1$; (b) $(n, \beta, \sigma^2) = (10^6, 0.2, 1)$, $r = 1/30, 1/15, \ldots, 0.5$

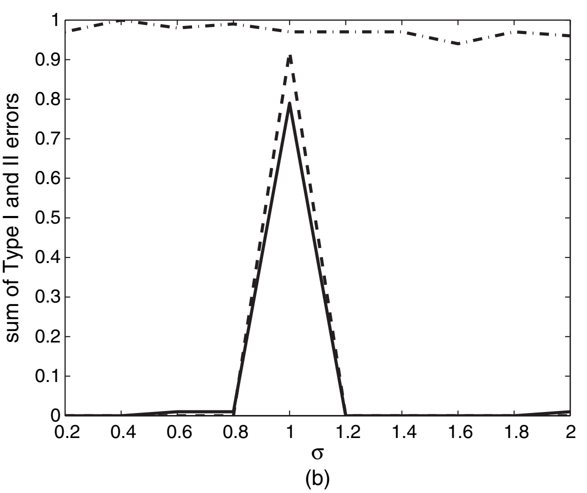
boundary is for finite n (the theoretic detection boundary corresponds to $n=\infty$). We investigate both a sparse case and a dense case.

For the sparse case, fixing $(\beta,\sigma^2)=(0.7,0.5)$ and $n\in\{10^4,10^5,10^7\}$, we let r range from 0.05 to 1 with an increment of 0.05. The sum of type I and type II errors of the LRT is reported in Fig. 3(a). Recall that theorems 1 and 2 predict that, for sufficiently large n, the sum of type I and type II errors of the LRT is approximately 1 when $r<\rho^*(\beta;\sigma)$ and is approximately 0 when $r>\rho^*(\beta;\sigma)$. In the current experiment, $\rho^*(\beta;\sigma)=0.3$. The simulation results show that, for each of $n\in\{10^4,10^5,10^7\}$, the sum of type I and type II errors of the LRT is small when $r\geqslant 0.5$ and is large when $r\leqslant 0.1$. In addition, if we view the sum of type I and type II errors as a function of r, then, as n grows larger, the function becomes increasingly close to the indicator function $\mathbf{1}_{\{r<0.3\}}$. This is consistent with theorems 2.

For the dense case, we fix $(\beta,\sigma^2)=(0.2,1)$ and $n\in\{10^4,10^5,10^7\}$, and let r range from 1/30 to



(a)



(b)

Fig. 5. Sum of type I and type II errors of higher criticism (———), the LRT (– – –) and Max (· – · – ·) (in (a)) or SM (· – · – ·) (in (b)), 100 replications: (a) $(n,\beta,r)=(10^6,0.7,0.25)$, $\sigma=0.2,0.4,\ldots,2$; (b) $(n,\beta,r)=(10^6,0.2,0.4)$, $\sigma=0.2,0.4,\ldots,2$ (the spike is because, in the dense case, the detection problem is intrinsically different when $\sigma=1$ and $\sigma\neq 1$)

0.5 with an increment of 1/30. The results are displayed in Fig. 3(b), where a similar conclusion can be drawn.

5.2. Experiment 2

In experiment 2, we compare higher criticism with the LRT, Max and SM, focusing on the effect of the signal strength (calibrated through the parameter r). We consider both a sparse case and a dense case.

For the sparse case, we fix $(n, \beta, \sigma^2) = (10^6, 0.7, 0.5)$ and let r range from 0.05 to 1 with an increment of 0.05. The results are displayed in Fig. 4(a), which illustrates that higher criticism has a similar performance with that of the LRT and outperforms Max. We also note that SM usually does not work in the sparse case, so we leave it out of the comparison.

We note that the LRT has optimal performance, but the implementation of which needs specific information of (β, r, σ). In contrast, higher criticism is non-parametric and does not

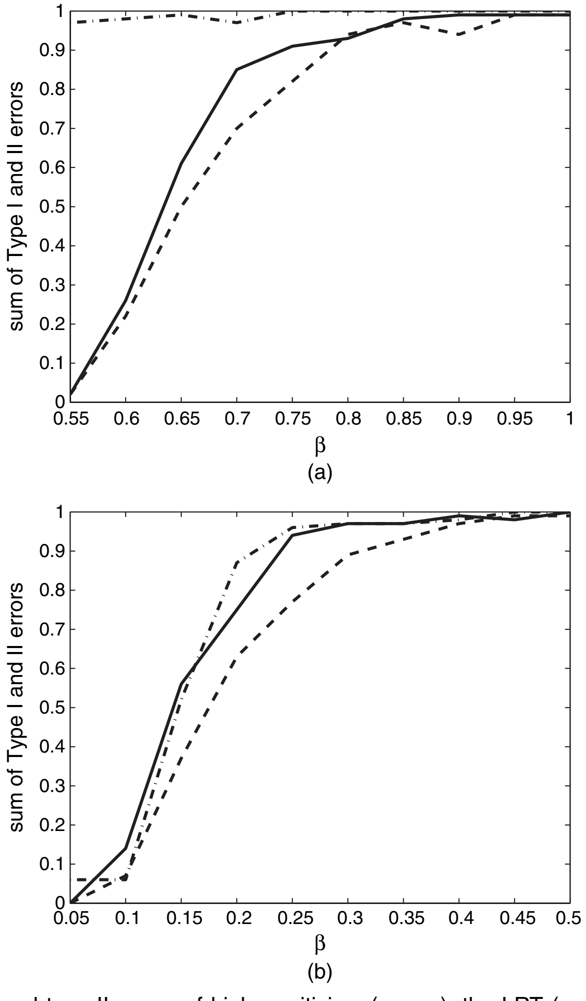


Fig. 6. Sum of type I and type II errors of higher criticism (———), the LRT (– – –) and Max (· – · – ·) (in (a)) or SM (· – · – ·) (in (b)), 100 replications: (a) $(n, r, \sigma^2) = (10^6, 0.25, 0.5)$, $\beta = 0.55, 0.60, \ldots, 1$; (b) $(n, r, \sigma^2) = (10^6, 0.3, 1)$, $\beta = 0.05, 0.10, \ldots, 0.5$

need such information. Nevertheless, higher criticism has comparable performance to that of the LRT.

For the dense case, we fix $(n, \beta, \sigma^2) = (10^6, 0.2, 1)$ and let r range from 1/30 to 0.5 with an increment of 1/30. In this case, Max usually does not work well, so we compare higher criticism with the LRT and SM only. The results are summarized in Fig. 4(b), where a similar conclusion can be drawn.

5.3. Experiment 3

In experiment 3, we continue to compare higher criticism with the LRT, Max and SM, but with the focus on the effect of the heteroscedasticity (calibrated by the parameter σ). We consider a sparse case and a dense case.

For the sparse case, we fix $(n, \beta, r) = (10^6, 0.7, 0.25)$ and let σ range from 0.2 to 2 with an increment of 0.2. The results are reported in Fig. 5(a) (that for SM is left out as it would not work well in the very sparse case), where the performance of each test becomes increasingly better as σ increases. This suggests that the testing problem becomes increasingly easier as σ increases, which fits well with the asymptotic theory in Section 2. In addition, for the whole region of σ, higher criticism has a comparable performance to that of the LRT, and it outperforms Max except for large σ, where higher criticism and Max perform comparably.

For the dense case, we fix $(n, \beta, r) = (10^6, 0.2, 0.4)$ and let σ range from 0.2 to 2 with an increment of 0.2. We compare the performance of higher criticism with that of the LRT and SM. The results are displayed in Fig. 5(b). It is noteworthy that higher criticism and the LRT perform reasonably well when σ is bounded away from 1 and effectively fail when $\sigma = 1$. This is because the detection problem is intrinsically different in the cases of $\sigma \neq 1$ and $\sigma = 1$. In the former, the heteroscedasticity alone could yield successful detection. In the latter, signals must be sufficiently strong for successful detection. Note that, for the whole range of σ, SM has poor performance.

5.4. Experiment 4

In experiment 4, we continue to compare the performance of higher criticism with that of the LRT, Max and SM, but with the focus on the effect of the level of sparsity (calibrated by the parameter β).

First, we investigate the case $\beta > \frac{1}{2}$. We fix $(n, r, \sigma^2) = (10^6, 0.25, 0.5)$ and let β range from 0.55 to 1 with an increment of 0.05. The results are displayed in Fig. 6(a), which illustrates that the detection problem becomes increasingly more difficult when β increases and r is fixed. Nevertheless, higher criticism has a comparable performance with that of the LRT and outperforms Max.

Second, we investigate the case $\beta < \frac{1}{2}$. We fix $(n, r, \sigma^2) = (10^6, 0.3, 1)$ and let β range from 0.05 to 0.5 with an increment of 0.05. Compared with the previous case, a similar conclusion can be drawn if we replace Max by SM.

In the simulation experiments, the estimated standard errors of the results are in general small. Recall that each point on the curves is the mean of 100 replications. To estimate the standard error of the mean, we use the following popular procedure (Zou, 2006). We generated 500 bootstrap samples out of the 100 replication results and then calculated the mean for each bootstrap sample. The estimated standard error is the standard deviation of the 500 bootstrap means. Owing to the large scale of the simulations, we pick several examples in both sparse and dense cases in experiment 3 and demonstrate their means with estimated standard errors in Table 1. The estimated standard errors are in general smaller than the differences between means. These results support our conclusions in experiment 3.

Table 1. Means with their estimated standard errors in parentheses for various methods†

σ	*Results for the sparse case*			*Results for the dense case*		
	LRT	*HC*	*Max*	*LRT*	*HC*	*SM*
0.5	0.84 (0.037)	0.91 (0.031)	1 (0)	0 (0)	0 (0)	0.98 (0.013)
1	0.52 (0.051)	0.62 (0.050)	0.81 (0.040)	0.93 (0.025)	0.98 (0.0142)	0.99 (0.010)

†Sparse, $(n,\beta,r)=(10^6,0.7,0.25)$; dense, $(n,\beta,r)=(10^6,0.2,0.4)$.

In conclusion, higher criticism has a comparable performance with that of the LRT. But, unlike the LRT, higher criticism is non-parametric. Higher criticism automatically adapts to different strengths of signal, levels of heteroscedasticity and levels of sparsity, and outperforms Max and SM.

6. Discussion

In this section, we discuss extensions of the main results in this paper to more general settings. We discuss the case where the strengths of signal may be unequal, the case where the noise may be correlated or non-Gaussian and the case where the heteroscedasticity parameter σ has a more complicated source.

6.1. When the signal strength may be unequal

In the preceding sections, the non-null density is a single normal $N(A_n,\sigma^2)$ distribution and the signal strengths are equal. More generally, we could replace the single normal distribution by a location Gaussian mixture, and the alternative hypothesis becomes

$$H_1^{(n)}: X_i \overset{\text{IID}}{\sim} (1-\varepsilon_n)\,N(0,1)+\varepsilon_n \int \frac{1}{\sigma}\phi\left(\frac{x-u}{\sigma}\right)\mathrm{d}G_n(u), \tag{24}$$

where $\phi(x)$ is the density of $N(0,1)$ and $G_n(u)$ is some distribution function.

Interestingly, the Hellinger distance that is associated with the testing problem is monotone with respect to G_n. In fact, fixing $n\geqslant 1$, if the support of G_n is contained in $[0,A_n]$, then the Hellinger distance between $N(0,1)$ and the density in expression (24) is no greater than that between $N(0,1)$ and $(1-\varepsilon_n)\,N(0,1)+\varepsilon_n\,N(A_n,\sigma^2)$. The proof is elementary so we omit it.

At the same time, similar monotonicity exists for higher criticism. In detail, fixing n, we apply higher criticism to n samples from

$$(1-\varepsilon_n)\,N(0,1)+\varepsilon_n \int \frac{1}{\sigma}\phi\left(\frac{x-u}{\sigma}\right)\mathrm{d}G_n(u),$$

as well as to n samples from $(1-\varepsilon_n)\,N(0,1)+\varepsilon_n\,N(A_n,\sigma^2)$, and obtain two scores. If the support of G_n is contained in $[0,A_n]$, then the former is stochastically smaller than the latter (we say that random variable X is less than or equal to random variable Y stochastically if the cumulative distribution function of the former is no smaller than that of the latter pointwise). The claim can be proved by elementary probability and mathematical induction, so we omit it.

These results shed light on the testing problem for general G_n. As before, let $\varepsilon_n=n^{-\beta}$ and $\tau_p=\sqrt{\{2r\log(p)\}}$. The following results can be proved.

(a) Suppose that $r<\rho^*(\beta;\sigma)$. Consider the problem of testing H_0 against $H_1^{(n)}$ as in expression (24). If the support of G_n is contained in $[0, A_n]$ for sufficiently large n, then two hypotheses are asymptotically indistinguishable (i.e., for any test, the sum of type I and type II errors tends to 1 as $n\to\infty$).
(b) Suppose that $r>\rho^*(\beta;\sigma)$. Consider the problem of testing H_0 against $H_1^{(n)}$ as in expression (24). If the support of G_n is contained in $[A_n,\infty)$, then the sum of type I and type II errors of the higher criticism test tends to 0 as $n\to\infty$.

6.2. When the noise is correlated or non-Gaussian

The main results in this paper can also be extended to the case where the X_i are correlated or non-Gaussian.

We discuss the correlated case first. Consider a model $X=\mu+Z$, where the mean vector μ is non-random and sparse, and $Z\sim N(0,\Sigma)$ for some covariance matrix $\Sigma=\Sigma_{n,n}$. Let $\text{supp}(\mu)$ be the support of μ, and let $\Lambda=\Lambda(\mu)$ be an $n\times n$ diagonal matrix the kth co-ordinate of which is σ or 1 depending on whether $k\in\text{supp}(\mu)$ or not. We are interested in testing a null hypothesis where $\mu=0$ and $\Sigma=\Sigma^*$ against an alternative hypothesis where $\mu\neq 0$ and $\Sigma=\Lambda\Sigma^*\Lambda$, where Σ^* is a known covariance matrix. Note that our preceding model corresponds to the case where Σ^* is the identity matrix. Also, a special case of the above model was studied in Hall and Jin (2008, 2010), where $\sigma=1$ so that the model is homoscedastic in a sense. In these works, we found that the correlation structure in the noise is not necessarily a *curse* and could be a *blessing*. We showed that we could better the testing power of higher criticism by combining the correlation structure with the statistic. The heteroscedastic case is interesting but has not yet been studied.

We now discuss the non-Gaussian case. In this case, how to calculate individual p-values poses challenges. An interesting case is where the marginal distribution of X_i is close to normal. An iconic example is the study of gene microarrays, where X_i could be the Studentized t-scores of m different replicates for the ith gene. When m is moderately large, the moderate tail of X_i is close to that of $N(0,1)$. Exploration along this direction includes Delaigle *et al.* (2011) where we learned that higher criticism continues to work well if we use bootstrapping correction on small p-values. The scope of this study is limited to the homoscedastic case, and extension to the heteroscedastic case is both possible and of interest.

6.3. When the heteroscedasticity has a more complicated source

In the preceding sections, we model the heteroscedasticity parameter σ as non-stochastic. The setting can be extended to a much broader setting where σ is random and has a density $h(\sigma)$. Assume that the support of $h(\sigma)$ is contained in an interval $[a,b]$, where $0<a<b<\infty$. We consider a setting where, under hypothesis $H_1^{(n)}$, $X_i\sim^{\text{IID}} g(x)$, with

$$\begin{aligned} g(x) &= g(x;\varepsilon_n, A_n, h, a, b) \\ &= (1-\varepsilon_n)\,\phi(x)+\varepsilon_n\int_a^b \frac{1}{\sigma}\phi\left(\frac{x-A_n}{\sigma}\right)h(\sigma)\,\mathrm{d}\sigma. \end{aligned} \tag{25}$$

Recall that, in the sparse case, the detection boundary $r=\rho^*(\beta;\sigma)$ is monotonically decreasing in σ when β is fixed. The interpretation is that a larger σ always makes the detection problem easier. Compare the current testing problem with two other testing problems: where $\sigma=\nu_a$ (point mass at a) and $\sigma=\nu_b$. Note that $h(\sigma)$ is supported in $[a,b]$. In comparison, the detection problem in the current setting should be easier than the case $\sigma=\nu_a$ and be more difficult than the case $\sigma = \nu_b$. In other words, the 'detection boundary' that is associated with the current case is sandwiched by two curves $r = \rho^*(\beta; a)$ and $r = \rho^*(\beta; b)$ in the β–r-plane.

If additionally $h(\sigma)$ is continuous and is non-zero at the point b, then there is a non-vanishing fraction of σ, say $\delta \in (0, 1)$, that falls close to b. Heuristically, the detection problem is at most as hard as the case where $g(x)$ in equation (25) is replaced by $\tilde{g}(x)$, where

$$\tilde{g}(x) = (1 - \delta\varepsilon_n)\, N(0, 1) + \delta\varepsilon_n\, N(A_n, b^2). \qquad (26)$$

Since the constant δ has only a negligible effect on the testing problem, the detection boundary that is associated with equation (26) will be the same as in the case $\sigma = \nu_b$. For brevity, we omit the proof.

We briefly comment on using higher criticism for real data analysis. One interesting application of higher criticism is for high dimensional feature selection and classification (see Section 4). In a related paper (Donoho and Jin, 2008), the method was applied to several by now standard gene microarray data sets (leukaemia, prostate cancer and colon cancer). The results that were reported are encouraging and the method is competitive with many widely used classifiers including random forests and the support vector machine. Another interesting application of higher criticism is for non-Gaussian detection in the so-called Wilkinson microwave anisotropy probe data (Cayon *et al.*, 2005). The method is competitive with the kurtosis-based method, which is the most widely used method by cosmologists and astronomers. In these real data analyses, it is difficult to tell whether the assumption of homoscedasticity is valid or not. However, the current paper suggests that higher criticism may continue to work well even when the assumption of homoscedasticity does not hold.

To conclude this section, we mention that this paper is connected to that by Jager and Wellner (2007), who investigated higher criticism in the context of goodness of fit. It is also connected to Meinshausen and Buhlmann (2006) and Cai *et al.* (2007), who used higher criticism to motivate lower bounds for the proportion of non-null effects.

Acknowledgements

The authors thank the Associate Editor and two referees for their helpful comments which have led to a better presentation of the paper. We also thank Mark Low for helpful discussion. Tony Cai was supported in part by National Science Foundation grant DMS-0604954 and National Science Foundation Focused Research Group grant DMS-0854973. Jeng and Jin were partially supported by National Science Foundation grants DMS-0639980 and DMS-0908613.

Appendix A: Proofs

We now prove the main results. In this section we shall use $\mathrm{PL}(n) > 0$ to denote a generic poly-log-term which may be different from one occurrence to another, satisfying $\lim_{n\to\infty}\{\mathrm{PL}(n)n^{-\delta}\} = 0$ and $\lim_{n\to\infty}\{\mathrm{PL}(n)n^{\delta}\} = \infty$ for any constant $\delta > 0$.

A.1. Proof of theorem 1

By the well-known theory on the relationship between the L^1-distance and the Hellinger distance, it suffices to show that the Hellinger affinity between $N(0, 1)$ and $(1 - \varepsilon_n)\, N(0, 1) + \varepsilon_n\, N(A_n, \sigma^2)$ behaves asymptotically as $1 + o(1/n)$. Denote the density of $N(0, \sigma^2)$ by $\phi_\sigma(x)$ (we drop the subscript when $\sigma = 1$), and introduce

$$g_n(x) = g_n(x; r, \sigma) = \frac{\phi_\sigma(x - A_n)}{\phi(x)}. \qquad (27)$$

The Hellinger affinity is then $E[\surd\{1-\varepsilon_n+\varepsilon_n\, g_n(X)\}]$, where $X\sim N(0,1)$. Let D_n be the event of $|X|\leqslant\surd\{2\,\log(n)\}$. The following lemma is proved in Appendix B.

Lemma 2. Fix $\sigma>1$, $\beta\in(\frac{1}{2},1)$, and $r\in(0,\rho^*(\beta;\sigma))$. As $n\to\infty$,

$$\varepsilon_n\, E[g_n(X)\,\mathbf{1}_{\{D_n^c\}}]=o(1/n),$$
$$\varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{D_n\}}]=o(1/n).$$

We now proceed to show theorem 1. First, since $E[\surd\{1-\varepsilon_n+\varepsilon_n\, g_n(X)\,\mathbf{1}_{\{D_n\}}\}]\leqslant E[\surd\{1-\varepsilon_n+\varepsilon_n\, g_n(X)\}]\leqslant 1$, all we need to show is that

$$E[\surd\{1-\varepsilon_n+\varepsilon_n\, g_n(X)\,\mathbf{1}_{\{D_n\}}\}]=1+o(1/n).$$

Now, note that, for $x\geqslant -1$, $|\surd(1+x)-1-x/2|\leqslant Cx^2$. Applying this with $x=\varepsilon_n\{g_n(X)\,\mathbf{1}_{\{D_n\}}-1\}$ gives

$$E\surd\{1-\varepsilon_n+\varepsilon_n\, g_n(X)\,\mathbf{1}_{\{D_n\}}\}=1-\frac{\varepsilon_n}{2}E[g_n(X)\,\mathbf{1}_{\{D_n^c\}}]+\text{err}, \tag{28}$$

where, by the Cauchy–Schwarz inequality,

$$|\text{err}|\leqslant C\varepsilon_n^2\, E[g_n(X)\,\mathbf{1}_{\{D_n\}}-1]^2\leqslant C\varepsilon_n^2\{E[g_n^2(X)\,\mathbf{1}_{\{D_n\}}]+1\}. \tag{29}$$

Recall that $\varepsilon_n^2=n^{-2\beta}=o(1/n)$. Combining lemma 2 with expressions (28) and (29) gives the claim.

A.2. Proof of theorem 2

Since the proofs are similar, we show only that under the null hypothesis. By Chebyshev's inequality, to show that $-\log(\text{LR}_n)\to\infty$ in probability, it is sufficient to show that, as $n\to\infty$,

$$-E[\log(\text{LR}_n)]\to\infty, \tag{30}$$

and

$$\frac{\text{var}\{\log(\text{LR}_n)\}}{E[\log(\text{LR}_n)]^2}\to 0. \tag{31}$$

Consider assumption (30) first. Recalling that $g_n(x)=\phi_\sigma(x-A_n)/\phi(x)$, we introduce

$$\text{LLR}_n(X)=\text{LLR}_n(X;\varepsilon_n,g_n)=\log\{1-\varepsilon_n+\varepsilon_n\, g_n(X)\}, \tag{32}$$

and

$$f_n(x)=f_n(x;\varepsilon_n,g_n)=\log\{1+\varepsilon_n\, g_n(x)\}-\varepsilon_n\, g_n(x). \tag{33}$$

By definitions and elementary calculus, $\log(\text{LR}_n)=\Sigma_{i=1}^n \text{LLR}_n(X_i)$, and $E[\text{LLR}_n(X)]=E[\log\{1+\varepsilon_n g_n(X)\}-\varepsilon_n\, g_n(X)]+O(\varepsilon_n^2)=E[f_n(X)]+O(\varepsilon_n^2)$. Recalling that $\varepsilon_n^2=n^{-2\beta}=o(1/n)$,

$$E[\log(\text{LR}_n)]=n\,E[\text{LLR}_n(X)]=n\,E[f_n(X)]+o(1). \tag{34}$$

Here, X and X_i are IID $N(0,1)$, $1\leqslant i\leqslant n$. Moreover, since there is a constant $c_1\in(0,1)$ and a generic constant $C>0$ such that $\log(1+x)\leqslant c_1 x$ for $x>1$ and $\log(1+x)-x\leqslant -Cx^2$ for $x\leqslant 1$, there is a generic constant $C>0$ such that

$$E[f_n(X)]\leqslant -C\{\varepsilon_n E[g_n(X)\,\mathbf{1}_{\{\varepsilon_n g_n(X)>1\}}]+\varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{\varepsilon_n g_n(X)\leqslant 1\}}]\}. \tag{35}$$

The following lemma is proved in Appendix B.

Lemma 3. Fix $\sigma>0$, $\beta\in(\frac{1}{2},1)$ and $r\in(0,1)$ such that $r>\rho^*(\beta;\sigma)$; then, as $n\to\infty$, we have either

$$n\varepsilon_n\, E[g_n(X)\,\mathbf{1}_{\{\varepsilon_n g_n(X)>1\}}]\to\infty \tag{36}$$

or

$$n\varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{\varepsilon_n g_n(X)\leqslant 1\}}]\to\infty. \tag{37}$$

Combining lemma 3 with expressions (34) and (35) gives the claim in assumption (30).

Next, we show assumption (31). Recalling that $\log(\text{LR}_n)=\Sigma_{i=1}^n \text{LLR}_n(X_i)$, we have

$$\text{var}\{\log(\text{LR}_n)\}=n\,\text{var}\{\text{LLR}_n(X)\}=n(E[\text{LLR}_n^2]-E[\text{LLR}_n]^2).$$

Comparing this with expression (31), it is sufficient to show that there is a constant $C>0$ such that

$$E[\mathrm{LLR}_n^2(X)] \leqslant C|E[\mathrm{LLR}_n(X)]|. \tag{38}$$

First, by the Schwartz inequality, for all x,

$$\log^2\{1-\varepsilon_n+\varepsilon_n g_n(x)\} = \left[\log\left\{1-\frac{\varepsilon_n}{1+\varepsilon_n g_n(x)}\right\}+\log\{1+\varepsilon_n g_n(x)\}\right]^2 \leqslant C[\varepsilon_n^2+\log^2\{1+\varepsilon_n g_n(x)\}].$$

Recalling that $\varepsilon_n^2 = o(1/n)$,

$$E[\mathrm{LLR}_n^2] \leqslant C E[\log^2\{1+\varepsilon_n g_n(X)\}] + o(1/n).$$

Second, note that $\log(1+x) < C\sqrt{x}$ for $x>1$ and $\log(1+x)<x$ for $x>0$. By a similar argument to that in the proof of result (35),

$$E[\log^2\{1+\varepsilon_n g_n(X)\}] \leqslant C\{\varepsilon_n E[g_n(X)\,\mathbf{1}_{\{\varepsilon_n g_n(X)>1\}}] + \varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{\varepsilon_n g_n(X)\leqslant 1\}}]\}.$$

Since the right-hand side has an order that is much larger than $o(1/n)$,

$$E[\mathrm{LLR}_n^2] \leqslant C\{\varepsilon_n E[g_n(X)\,\mathbf{1}_{\{\varepsilon_n g_n(X)>1\}}] + \varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{\varepsilon_n g_n(X)\leqslant 1\}}]\}.$$

Comparing this with inequality (35) gives the claim.

A.3. Proof of theorem 3

By a similar argument to that in Appendix A.1, all that we need to show is that, when $\sigma=1$ and $r>\frac{1}{2}-\beta$,

$$E[\sqrt{\{1-\varepsilon_n+\varepsilon_n g_n(X)\}}] = 1 + o(n^{-1}), \tag{39}$$

where $X \sim N(0,1)$, and $g_n(X)$ is as in equation (27). By Taylor series expansion,

$$E[\sqrt{\{1-\varepsilon_n+\varepsilon_n g_n(X)\}}] \geqslant E\left[1+\frac{\varepsilon_n}{2}\{g_n(X)-1\}-\frac{\varepsilon_n^2}{8}\{g_n(X)-1\}^2\right].$$

Note that $E[g_n(X)]=1$; then

$$E[\sqrt{\{1-\varepsilon_n+\varepsilon_n g_n(X)\}}] \geqslant 1-\frac{\varepsilon_n^2}{8}\{E[g_n^2(X)]-1\}. \tag{40}$$

Write

$$\begin{aligned} E[g_n^2(X)] &= \int \frac{1}{\sqrt{(2\pi)\sigma^2}} \exp\left\{\left(\frac{1}{2}-\frac{1}{\sigma^2}\right)x^2+\frac{2A_n x}{\sigma^2}-\frac{A_n^2}{\sigma^2}\right\}\mathrm{d}x \\ &= \int \frac{1}{\sqrt{(2\pi)\sigma^2}} \exp\left\{-\frac{2-\sigma^2}{2\sigma^2}\left(x-\frac{2A_n}{2-\sigma^2}\right)^2+\frac{A_n^2}{2-\sigma^2}\right\}\mathrm{d}x. \end{aligned}$$

In the current case, $\sigma=1$, and $A_n=n^{-r}$ with $r>\beta-\frac{1}{2}$. By direct calculations, $E[g_n^2(X)]=\exp(A_n^2)$, and

$$\frac{\varepsilon_n^2}{8}\{E[g_n^2(X)]-1\} \sim \varepsilon_n^2 A_n^2 = o(n^{-1}). \tag{41}$$

Inserting expressions (40) and (41) into equation (39) gives the claim.

A.4. Proof of theorem 4

Recall that $\mathrm{LLR}_n(x)=\log[1+\varepsilon_n\{g_n(x)-1\}]$ and $\log(\mathrm{LR}_n)=\Sigma_{j=1}^n \mathrm{LLR}_n(X_j)$. By similar arguments to those in Appendix A.2 it is sufficient to show that for $X\sim N(0,1)$, when $n\to\infty$,

$$n\,E[\mathrm{LLR}_n(X)] \to -\infty, \tag{42}$$

and

$$\frac{\mathrm{var}\{\log(\mathrm{LR}_n)\}}{E[\log(\mathrm{LR}_n)]^2} \to 0. \tag{43}$$

Consider assumption (42) first. Introduce the event $B_n = \{X : \varepsilon_n\, g_n(X) \leqslant 1\}$. Note that $\log(1+x) \leqslant x$ for all x and $\log(1+x) \leqslant x - x^2/4$ when $x \leqslant 1$, and that $E[g_n(X)] = 1$. It follows that

$$E[\mathrm{LLR}_n(X)] \leqslant E[\varepsilon_n\{g_n(X)-1\}] - \tfrac{1}{4} E[\varepsilon_n^2\{g_n(X)-1\}^2\, \mathbf{1}_{B_n}] = -\tfrac{1}{4}\varepsilon_n^2\, E[\{g_n(X)-1\}^2\, \mathbf{1}_{B_n}]. \tag{44}$$

Since $E[g_n(X)\, \mathbf{1}_{B_n}] \leqslant E[g_n(X)] = 1$, it is seen that

$$E[\{g_n(X)-1\}^2\, \mathbf{1}_{B_n}] \geqslant E[g_n^2(X)\, \mathbf{1}_{B_n}] - 2 + P(B_n) = E[g_n^2(X)\mathbf{1}_{B_n}] - 1 - P(B_n^{\mathrm{c}}). \tag{45}$$

We now discuss for the case of $\sigma = 1$ and $\sigma \neq 1$ separately.

Consider the case $\sigma = 1$ first. In this case, $g_n(x) = \exp(A_n x - A_n^2/2)$. By direct calculations,

$$P(B_n^{\mathrm{c}}) = o(A_n^2),$$

$$E[g_n^2(X)\, \mathbf{1}_{B_n}] = \frac{\exp(A_n^2)}{\surd(2\pi)} \int_{\{x:\varepsilon_n g_n(x)\leqslant 1\}} \exp\left\{-\frac{(x-2A_n)^2}{2}\right\} \mathrm{d}x = 1 + A_n^2\{1+o(1)\}.$$

Combining this with expressions (44) and (45), $E[\mathrm{LLR}_n(X)] \lesssim -\frac{1}{4}\varepsilon_n^2 A_n^2 = -\frac{1}{4} n^{-2(\beta+r)}$. The claim follows by the assumption $r < \frac{1}{2} - \beta$.

Consider the case $\sigma \neq 1$. It is sufficient to show that, as $n \to \infty$,

$$E[g_n^2(X)\mathbf{1}_{B_n}] \sim \begin{cases} 1/\sigma\surd(2-\sigma^2), & \sigma < \surd 2, \\ C\surd\log(n), & \sigma = \surd 2, \\ \{C/\surd\log(n)\} n^{\beta(\sigma^2-2)/(\sigma^2-1)}, & \sigma > \surd 2, \end{cases} \tag{46}$$

where we note that $1/\sigma\surd(2-\sigma^2) > 1$ when $\sigma < \surd 2$. In fact, once this has been shown, noting that $P(B_n^{\mathrm{c}}) = o(1)$, it follows from expression (45) that there is a constant $c_0(\sigma) > 0$ such that, for sufficiently large n, $E[\{g_n(X)-1\}^2\mathbf{1}_{B_n}] - 1 \geqslant 4\, c_0(\sigma)$. Combining this with expression (44), $E[\mathrm{LLR}_n(X)] \leqslant -c_0(\sigma)\varepsilon_n^2 = -c_0(\sigma) n^{-2\beta}$. The claim follows from the assumption $\beta < \frac{1}{2}$.

We now show result (46). Write

$$E[g_n^2(X)\, \mathbf{1}_{B_n}] = \frac{1}{\surd(2\pi)\sigma^2} \int_{\{x:\varepsilon_n g_n(x)\leqslant 1\}} \exp\left\{\left(\frac{1}{2} - \frac{1}{\sigma^2}\right)x^2 + \frac{2A_n x}{\sigma^2} - \frac{A_n^2}{\sigma^2}\right\} \mathrm{d}x. \tag{47}$$

Consider the case $\sigma < \surd 2$ first. In this case, $\frac{1}{2} - 1/\sigma^2 < 0$. Since $A_n = n^{-r}$, it is seen that

$$E[g_n^2(X)\mathbf{1}_{B_n}] \sim \frac{1}{\surd(2\pi)\sigma^2} \int \exp\left\{\left(\frac{1}{2} - \frac{1}{\sigma^2}\right)x^2\right\} \mathrm{d}x = \frac{1}{\sigma\surd(2-\sigma^2)},$$

and the claim follows. Consider the case $\sigma \geqslant \surd 2$. Let $x_\pm(n) = x_\pm(n; \sigma, \varepsilon_n, A_n)$, $x_- < x_+$, be the two solutions of $\varepsilon_n\, g_n(x) = 1$, and let $x_0(n) = x_0(n; \sigma, \beta) = \surd\{2\sigma^2\beta \log(n)/(\sigma^2-1)\}$. By elementary calculus, $\varepsilon_n\, g_n(x) \leqslant 1$ if and only if $x_-(n) \leqslant x \leqslant x_+(n)$ and $x_\pm(n) = \pm x_0(n) + o(1)$, where $o(1) \to 0$ algebraically fast as $n \to \infty$. It follows that

$$\begin{aligned} E[g_n^2(X)\mathbf{1}_{B_n}] &= \frac{1}{\surd(2\pi)\sigma^2} \int_{x_-(n)}^{x_+(n)} \exp\left\{\left(\frac{1}{2} - \frac{1}{\sigma^2}\right)x^2 + \frac{2A_n x}{\sigma^2} - \frac{A_n^2}{\sigma^2}\right\} \mathrm{d}x \\ &\sim \frac{1}{\surd(2\pi)\sigma^2} \int_{x_-(n)}^{x_+(n)} \exp\left\{\left(\frac{1}{2} - \frac{1}{\sigma^2}\right)x^2\right\} \mathrm{d}x. \end{aligned} \tag{48}$$

When $\sigma = \surd 2$, $\frac{1}{2} - 1/\sigma^2 = 0$. By equation (48),

$$E[g_n^2(X)\mathbf{1}_{B_n}] \sim \frac{1}{\surd(2\pi)\sigma^2} 2\, x_0(n) \sim \frac{2}{\sigma}\sqrt{\left\{\frac{\beta\log(n)}{\pi(\sigma^2-1)}\right\}},$$

which gives the claim. When $\sigma > \surd 2$, $\frac{1}{2} - 1/\sigma^2 > 0$. By equation (48) and elementary calculus,

$$E[g_n^2(X)\, \mathbf{1}_{B_n}] \sim \frac{1}{\surd(2\pi)\sigma^2(\frac{1}{2} - 1/\sigma^2)\, x_0(n)} \exp\left\{\left(\frac{1}{2} - \frac{1}{\sigma^2}\right)x_0^2(n)\right\} \sim \frac{\surd(\sigma^2-1)}{(\sigma^2-2)\sigma\surd\{\pi\beta\log(n)\}} n^{\beta(\sigma^2-2)/(\sigma^2-1)},$$

and the claim follows.

We now show assumption (43). By similar arguments to those in Appendix A.2, it is sufficient to show that

$$E[\mathrm{LLR}_n^2(X)] \leqslant C|E[\mathrm{LLR}_n(X)]|. \tag{49}$$

Note that it is proved in expression (44) that

$$|E[\mathrm{LLR}_n(X)]| \geqslant \tfrac{1}{4} E[\varepsilon_n^2 \{g_n(X)-1\}^2 \mathbf{1}_{B_n}]. \tag{50}$$

Recall that $\mathrm{LLR}_n(x) = \log[1+\varepsilon_n\{g_n(x)-1\}]$. Since $\log^2(1+a) \leqslant a$ for $a>1$ and $|\log^2(1+a)| \lesssim a^2$ for $-\varepsilon_n \leqslant a \leqslant 1$,

$$E[\mathrm{LLR}_n^2(X)] \lesssim E[\varepsilon_n\{g_n(X)-1\} \mathbf{1}_{B_n^c}] + E[\varepsilon_n^2\{g_n(X)-1\}^2 \mathbf{1}_{B_n}]. \tag{51}$$

Compare expression (51) with expression (50). To show inequality (49), it is sufficient to show that

$$E[\varepsilon_n\{g_n(X)-1\} \mathbf{1}_{B_n^c}] \leqslant C\, E[\varepsilon_n^2\{g_n(X)-1\}^2 \mathbf{1}_{B_n}]. \tag{52}$$

This follows trivially when $\sigma<1$, in which case $B_n^c = \emptyset$. This also follows easily when $\sigma=1$, in which case $g_n(x) = \exp(A_n x - A_n^2/2)$ and $B_n = \{X : |X| \geqslant n^{\beta+r} \exp(A_n^2)\}$.

We now show inequality (52) for the case $\sigma>1$. By the proof of assumption (42),

$$E[\varepsilon_n^2\{g_n(X)-1\}^2 \mathbf{1}_{B_n}] \geqslant \begin{cases} Cn^{-2\beta}, & 1<\sigma<\surd 2, \\ C\surd\log(n)\, n^{-2\beta}, & \sigma=\surd 2, \\ \{C/\surd\log(n)\}\, n^{-\beta\sigma^2/(\sigma^2-1)}, & \sigma>\surd 2. \end{cases} \tag{53}$$

At the same time, by the definitions and properties of $x_\pm(n)$ and Mills's ratio (Wasserman, 2006),

$$\varepsilon_n\, E[g_n(X)\, \mathbf{1}_{B_n^c}] \sim 2\varepsilon_n \int_{x_0(n)}^{\infty} \frac{1}{\sigma} \phi\left(\frac{x-A_n}{\sigma}\right) \mathrm{d}x \leqslant \frac{C}{\surd\log(n)} n^{-\beta\sigma^2/(\sigma^2-1)}. \tag{54}$$

Note that $\sigma^2/(\sigma^2-1) \geqslant 2$ when $\sigma \leqslant \surd 2$. Comparing expressions (53) and (54) gives inequality (52).

A.5. Proof of theorem 7

It is sufficient to show that, as $n \to \infty$,

$$P_{H_0}(\mathrm{HC}_n^* \geqslant \surd[2(1+\delta)\log\{\log(n)\}]) \to 0, \tag{55}$$

and

$$P_{H_1^{(n)}}(HC_n^* < \surd[2(1+\delta)\log\{\log(n)\}]) \to 0. \tag{56}$$

Recall that, under the null hypothesis, HC_n^* equals in distribution the extreme value of a normalized uniform empirical process and

$$\frac{\mathrm{HC}_n^*}{\surd[2\log\{\log(n)\}]} \to 1, \qquad \text{in probability.}$$

So, the first claim follows directly. Consider the second claim. By expressions (3.16), (3.19) and (3.20), $\mathrm{HC}_n^* = \sup_{-\infty<t<\infty} |W_n(t)| \geqslant |W_n\{t_n^*(\sigma,\beta,r)\}|$, so all we need to show is that, under the assumptions in theorem 7,

$$P_{H_1^{(n)}}(|W_n\{t_n^*(\sigma,\beta,r)\}| < \surd[2(1+\delta)\log\{\log(n)\}]) \to 0. \tag{57}$$

For this, we write for short $t = t_n^*(\sigma,\beta,r)$.

In the sparse case with $\frac{1}{2}<\beta<1$, direct calculations show that

$$E[W_n(t)] = \surd n\varepsilon_n \left\{\bar\Phi\left(\frac{t-A_n}{\sigma}\right) - \bar\Phi(t)\right\} \Big/ \surd[\bar\Phi(t)\{1-\bar\Phi(t)\}] \sim \surd n\varepsilon_n \left\{\bar\Phi\left(\frac{t-A_n}{\sigma}\right) - \bar\Phi(t)\right\} \Big/ \surd\bar\Phi(t), \tag{58}$$

and

$$\mathrm{var}\{W_n(t)\} = \frac{\bar F(t)\{1-\bar F(t)\}}{\bar\Phi(t)\{1-\bar\Phi(t)\}} \sim \frac{\bar F(t)}{\bar\Phi(t)}. \tag{59}$$

By Mills's ratio (Wasserman, 2006),

$$\bar{\Phi}(\surd\{2q\log(n)\} = \mathrm{PL}(n)n^{-q},$$
$$\bar{\Phi}\left[\frac{\surd\{2q\log(n)\}-A_n}{\sigma}\right] = \mathrm{PL}(n)n^{-(\surd q-\surd r)^2/\sigma^2}. \tag{60}$$

Inserting expression (60) into expression (58) gives

$$\surd n\varepsilon_n\left\{\bar{\Phi}\left(\frac{t-A_n}{\sigma}\right) - \bar{\Phi}(t)\right\}\Big/\surd\bar{\Phi}(t) = \begin{cases}\mathrm{PL}(n)n^{r/(2-\sigma^2)-(\beta-1/2)}, & \sigma<\surd 2, r<(2-\sigma^2)^2/4,\\ \mathrm{PL}(n)n^{1-\beta-(1-\surd r)^2/\sigma^2}, & \text{otherwise.}\end{cases} \tag{61}$$

It follows from $r>\rho^*(\sigma,\beta,r)$ and basic algebra that $E[W_n(t)]\to\infty$ algebraically fast. Especially,

$$E[W_n(t)]/\surd[2(1+\delta)\log\{\log(n)\}]\to\infty. \tag{62}$$

Combining expressions (58) and (59), it follows from Chebyshev's inequality that

$$P_{H_1^{(n)}}(|W_n\{t_n^*(\sigma,\beta,r)\}|<\surd[2(1+\delta)\log\{\log(n)\}])\leqslant C\frac{\mathrm{var}\{W_n(t)\}}{E[W_n(t)]^2}\leqslant C\bar{F}(t)\Big/n\varepsilon_n^2\left\{\bar{\Phi}\left(\frac{t-A_n}{\sigma}\right)-\bar{\Phi}(t)\right\}^2.$$

Applying expression (61), the above expression is approximately equal to

$$\begin{cases} n^{-2r/(2-\sigma^2)+2\beta-1}+n^{\sigma^2 r/(2-\sigma^2)^2+\beta-1}, & \sigma<\surd 2, r<(2-\sigma^2)^2/4,\\ n^{-1+\beta+(1-\surd r)^2/\sigma^2}, & \text{otherwise,}\end{cases}$$

which tends to 0 algebraically fast as $r>\rho^*(\sigma,\beta,r)$.

In the dense case with $0<\beta<\frac{1}{2}$, recall that $t_n^*(\sigma,\beta,r)=1$. Therefore,

$$E[W_n(1)] = \surd n\varepsilon_n\left\{\bar{\Phi}\left(\frac{1-A_n}{\sigma}\right)-\bar{\Phi}(1)\right\}\Big/\surd[\bar{\Phi}(1)\{1-\bar{\Phi}(1)\}]\sim C\surd n\varepsilon_n\left\{\bar{\Phi}\left(\frac{1-A_n}{\sigma}\right)-\bar{\Phi}(1)\right\},$$

and

$$\mathrm{var}\{W_n(1)\} = \frac{\bar{F}(1)\{1-\bar{F}(1)\}}{\bar{\Phi}(1)\{1-\bar{\Phi}(1)\}}\sim\text{constant}. \tag{63}$$

Furthermore,

$$\surd n\varepsilon_n\left\{\bar{\Phi}\left(\frac{1-A_n}{\sigma}\right)-\bar{\Phi}(1)\right\} = -Cn^{1/2-\beta}\left(\frac{1}{\sigma}-1-\frac{A_n}{\sigma}\right)\{1+o(1)\}.$$

So, when $\sigma>1$, or $\sigma=1$ and $r<\frac{1}{2}-\beta$,

$$E[W_n(1)]\sim n^{\gamma} \tag{64}$$

for some $\gamma>0$ and

$$E[W_n(1)]/\surd[2(1+\delta)\log\{\log(n)\}]\to\infty.$$

In contrast, when $\sigma<1$,

$$E[W_n(1)]\sim -n^{\gamma} \tag{65}$$

for some $\gamma>0$ and

$$E[W_n(1)]/\surd[2(1+\delta)\log\{\log(n)\}\to-\infty.$$

Combining expressions (63), (64) and (65), it follows from Chebyshev's inequality that

$$P_{H_1^{(n)}}\{|W_n\{t_n^*(\sigma,\beta,r)\}|<\surd[2(1+\delta)\log\{\log(n)\}]\leqslant C\frac{\mathrm{var}\{W_n(1)\}}{E[W_n(1)]^2}\leqslant Cn^{-2\gamma}\to 0.$$

Appendix B

B.1. Proof of theorem 5 and theorem 6

We consider the case $\sigma\in(0,\surd 2)$ first. Since the proofs are similar, we show only that under the null hypothesis. Recall that $\log(\mathrm{LR}_n)=\Sigma_{j=1}^n \mathrm{LLR}_n(X_j)$ (see Section 6.2). It is sufficient to show that

$$E[\exp\{\mathrm{i}t\,\mathrm{LLR}_n(X)\}]=\begin{cases} 1+\left(-\dfrac{\mathrm{i}t+t^2}{2}\right)\dfrac{1}{\sigma\surd(2-\sigma^2)}\dfrac{1}{n}\{1+o(1)\}, & \dfrac{1}{2}<\beta<1-\dfrac{\sigma^2}{4}, \\ 1+\left(-\dfrac{\mathrm{i}t+t^2}{2}\right)\dfrac{1}{2\sigma\surd(2-\sigma^2)}\dfrac{1}{n}\{1+o(1)\}, & \beta=1-\dfrac{\sigma^2}{4}, \\ 1+(1/n)\psi^0_{\beta,\sigma}(t)\{1+o(1)\}, & 1-\dfrac{\sigma^2}{4}<\beta<1. \end{cases}$$

Note that $E[\exp\{\mathrm{i}t\,\mathrm{LLR}_n(X)\}]=\exp\{\mathrm{i}t\log(1-\epsilon_n)\}E[\exp[\mathrm{i}t\log\{1+\epsilon_n\,g_n(X)\}]]+O(\epsilon_n^2)$, $\exp\{\mathrm{i}t\log(1-\epsilon_n)\}=1-\mathrm{i}t\epsilon_n+O(\epsilon_n^2)$, and $E[\exp[\mathrm{i}t\log\{1+\epsilon_n\,g_n(X)\}]]=1+\mathrm{i}t\epsilon_n+E[\exp[\mathrm{i}t\log\{1+\epsilon_n\,g_n(X)\}]-1-\mathrm{i}t\epsilon_n\,g_n(X)]$. Therefore,

$$E[\exp\{\mathrm{i}t\mathrm{LLR}_n(X)\}]=1+E[\exp[\mathrm{i}t\log\{1+\epsilon_n\,g_n(X)\}]-1-\mathrm{i}t\epsilon_n\,g_n(X)]+o(1/n). \tag{66}$$

We now analyse the limiting behaviour of $E[\exp[\mathrm{i}t\log\{1-\epsilon_n+\epsilon_n\,g_n(X)\}]-1-\mathrm{i}\epsilon_n t\,g_n(X)]$ for the case $1\leqslant\sigma<\surd 2$. The case $0<\sigma<1$ is similar to that of $1\leqslant\sigma<\surd 2$ and thus is omitted.

In the case $1\leqslant\sigma<\surd 2$, we discuss three subcases separately: $\beta\leqslant(1-\sigma^2/4)$, $\beta=(1-\sigma^2/4)$ and $\beta>(1-\sigma^2/4)$.

When $\beta<1-\sigma^2/4$, we have

$$r=(2-\sigma^2)(\beta-\tfrac{1}{2}), \qquad \text{so } 0<r<\tfrac{1}{4}(2-\sigma^2)^2. \tag{67}$$

Write

$$\epsilon_n g_n(x)=C\epsilon_n\exp\left\{\left(\frac{1}{2}-\frac{1}{2\sigma^2}\right)x^2+\frac{A_n x}{\sigma^2}-\frac{A_n^2}{2\sigma^2}\right\}.$$

We first show that $\max_{\{|x|\leqslant\surd\{2\log(n)\}\}}|\epsilon_n\,g_n(x)|\}=o(1)$. When $\sigma\geqslant 1$, the exponent is a convex function in x, and the maximum is reached at $x=\surd\{2\log(n)\}$ with the maximum value of

$$n^{1-\{\beta+(1-\surd r)^2/\sigma^2\}}. \tag{68}$$

By expression (67), the exponent $1-\{\beta+(1-\surd r)^2/\sigma^2\}<0$. When $\sigma<1$, the exponent is a concave function in x. We further consider two sub-subcases: $\surd\{2\log(n)\}\leqslant A_n/(1-\sigma^2)$ and $\surd\{2\log(n)\}>A_n/(1-\sigma^2)$. For the first case, the maximum is reached at $x=\surd\{2\log(n)\}$ with the maximum value of expression (68), where the exponent is less than 0. For the second case, we have $\surd r<1-\sigma^2$, and the maximum is reached at $x=A_n/(1-\sigma^2)$ with the maximum value of

$$n^{-\beta+r/(1-\sigma^2)}.$$

Together, expression (67) and the fact that $r<(1-\sigma^2)^2<(1-\sigma^2/2)(1-\sigma^2)$ imply that $\beta<1-\sigma^2/2$. So, using expression (67) again,

$$-\beta+\frac{r}{1-\sigma^2}=\frac{\beta}{1-\sigma^2}+\frac{2-\sigma^2}{2(1-\sigma^2)}<0.$$

Combining all these gives that

$$\max_{\{|x|\leqslant\surd\{2\log(n)\}\}}|\epsilon_n\,g_n(x)|=\exp\left[\max_{\{|x|\leqslant\surd\{2\log(n)\}\}}\left\{\left(\frac{1}{2}-\frac{1}{2\sigma^2}\right)x^2+\frac{A_n x}{\sigma^2}-\frac{A_n^2}{2\sigma^2}\right\}\right]=o(1). \tag{69}$$

Now, introduce

$$f_n(x)=f(x;t,\beta,r)=\exp[\mathrm{i}t\log\{(1+\epsilon_n g_n(X)\}]-1-\mathrm{i}t\epsilon_n\,g_n(x),$$

and the event $D_n=\{|X|\leqslant\surd\{2\log(n)\}\}$. We have

$$E[f_n(X)]=E[f_n(X)\mathbf{1}_{\{D_n\}}]+E[f_n(X)\mathbf{1}_{\{D_n^c\}}].$$

On one hand, by equation (69) and Taylor series expansion,

$$E[f_n(X)\mathbf{1}_{\{D_n\}}]\sim(-t^2/2)\,E[\epsilon_n^2\,g_n^2(X)\mathbf{1}_{\{D_n\}}].$$

On the other hand,

$$|f_n(X)| \leqslant 1 + \varepsilon_n g_n(X).$$

Compare this with the desired claim; it is sufficient to show that

$$E[\varepsilon_n^2 \, g_n^2(X) \, \mathbf{1}_{\{D_n\}}] \sim \frac{1}{\sqrt{\{\sigma^2(2-\sigma^2)\}}} \frac{1}{n}, \tag{70}$$

and that

$$E[\{1 + \varepsilon_n \, g_n(X)\} \, \mathbf{1}_{\{D_n^c\}}] = o(1/n). \tag{71}$$

Consider assumption (70) first. By a similar argument to that in the proof of lemma 2,

$$\varepsilon_n^2 \, E[g_n^2(X) \, \mathbf{1}_{\{D_n\}}] = \frac{1}{\sqrt{(2\pi)\sigma^2}} n^{-2\beta+2r/(2-\sigma^2)} \int_{-\sqrt{\{2\log(n)\}}-A_n/(1-\sigma^2/2)}^{\sqrt{\{2\log(n)\}}-A_n/(1-\sigma^2/2)} \exp\left\{-\left(\frac{1}{\sigma^2} - \frac{1}{2}\right) y^2\right\} \mathrm{d}y. \tag{72}$$

Note that

$$\sqrt{\{2\log(n)\}} - \frac{A_n}{1-\sigma^2/2} = \sqrt{\{2\log(n)\}}\left(1 - \frac{2\sqrt{r}}{2-\sigma^2}\right),$$

where $1 - 2\sqrt{r}/(2-\sigma^2) > 0$ as $r < \frac{1}{4}(2-\sigma^2)^2$. Therefore,

$$\int_{-\sqrt{\{2\log(n)\}}-A_n/(1-\sigma^2/2)}^{\sqrt{\{2\log(n)\}}-A_n/(1-\sigma^2/2)} \exp\left\{-\left(\frac{1}{\sigma^2} - \frac{1}{2}\right) y^2\right\} \mathrm{d}y \sim \sqrt{\left(\frac{2\pi\sigma^2}{2-\sigma^2}\right)}.$$

Moreover, by expression (67), $2\beta - 2r/(2-\sigma^2) = 1$, so

$$\varepsilon_n^2 \, E[g_n^2(X) \, \mathbf{1}_{\{D_n\}}] \sim \frac{1}{\sqrt{\{\sigma^2(2-\sigma^2)\}}} n^{-2\beta+2r/(2-\sigma^2)} = \frac{1}{\sqrt{\{\sigma^2(2-\sigma^2)\}}} \frac{1}{n},$$

and, therefore,

$$E[f_n(X) \, \mathbf{1}_{\{D_n\}}] \sim -\frac{t^2}{2} \frac{1}{\sqrt{\{\sigma^2(2-\sigma^2)\}}} \frac{1}{n}, \tag{73}$$

which gives expression (70).

Consider equation (71). Recalling that $g_n(x) = \phi_\sigma(x - A_n)/\phi(x)$,

$$E[\{1 + \varepsilon_n \, g_n(X)\} \, \mathbf{1}_{\{D_n^c\}}] \leqslant \int_{|x|>\sqrt{\{2\log(n)\}}} \{\phi(x) + \varepsilon_n \, \phi_\sigma(x - A_n)\} \, \mathrm{d}x. \tag{74}$$

It is seen that

$$\int_{|x|>\sqrt{\{2\log(n)\}}} \phi(x) = o(1) \, \phi[\sqrt{\{2 \, \log(n)\}}] = o(1/n),$$

and that

$$\int_{|x|>\sqrt{\{2\log(n)\}}} \varepsilon_n \, \phi_\sigma(x - A_n) \, \mathrm{d}x = o(1) n^{-\beta} \, \phi[(1-\sqrt{r})\sqrt{\{2\log(n)\}}] = o(n^{-\beta+(1-\sqrt{r})^2/\sigma^2}).$$

Moreover, by expression (67), $\beta + (1-\sqrt{r})^2/\sigma^2 > 1$, so it follows that inequality (74) gives that

$$E[\{1 + \varepsilon_n \, g_n(X)\} \, \mathbf{1}_{\{D_n^c\}}] = o(1/n). \tag{75}$$

This gives equation (71) and concludes the claim in the case $\beta < 1 - \sigma^2/4$.

Consider the case $\beta = 1-\sigma^2/4$. The claim can be proved similarly provided that we modify the event of D_n by

$$\tilde{D}_n = \left\{ |X| \leqslant \surd\{2\log(n)\} - \frac{\log^{1/2}\{\log(n)\}}{\surd\{2\log(n)\}} \right\}.$$

For brevity, we omit further discussion.

Consider the case $\beta > 1-\sigma^2/4$. In this case, we have

$$\varepsilon_n = n^{-\beta} \log(n)^{1-\surd(1-\beta)/\sigma},$$

and

$$r = \{1-\sigma\surd(1-\beta)\}^2, \qquad \text{so } \surd r > 1-\sigma^2/2. \tag{76}$$

Equate $\varepsilon_n \phi_\sigma(x-A_n)/\phi_0(x) = 1/\sigma$. Direct calculations show that we have two solutions; using expression (76), it is seen that one of them is approximately $\surd\{2\log(n)\}$ and we denote this solution by $x_0 = x_0(n) = \surd\{2\log(n)\} - \log\{\log(n)\}/\surd\{2\log(n)\}$. By the way that ε_n is chosen, we have $(1/x_0)\exp(-x_0^2/2) \sim 1/n$. Now, change variable with $x = x_0 + y/x_0$. It follows that

$$\varepsilon_n g_n(x) = \frac{1}{\sigma} \exp\left\{ \left(1 - \frac{1-\surd r}{\sigma^2}\right) y \right\} \exp\left\{ -\frac{y^2}{2x_0^2} \left(\frac{1}{\sigma^2} - 1\right) \right\},$$

$$\phi(x) = \frac{1}{\surd(2\pi)} x_0 \frac{1}{n} \exp(-y) \exp\left(-\frac{y^2}{2x_0^2} \right).$$

Therefore,

$$\begin{aligned} E[f_n(X)] = \frac{1}{\surd(2\pi)n} \int \Bigg\{ & \exp\Bigg(\mathrm{i}t \log\Bigg[1 + \frac{1}{\sigma} \exp\left\{ \left(1 - \frac{1-\surd r}{\sigma^2}\right) y \right\} \\ & \times \exp\left\{ -\frac{y^2}{2x_0^2} \left(\frac{1}{\sigma^2} - 1\right) \right\} \Bigg] \Bigg) - 1 - \frac{\mathrm{i}t}{\sigma} \exp\left\{ \left(1 - \frac{1-\surd r}{\sigma^2}\right) y \right\} \\ & \times \exp\left\{ -\frac{y^2}{2x_0^2} \left(\frac{1}{\sigma^2} - 1\right) \right\} \exp(-y) \exp\left(-\frac{y^2}{2x_0^2} \right) \mathrm{d}y. \end{aligned}$$

Denote the integrand (excluding $1/n$) by

$$\begin{aligned} h_n(y) = \Bigg\{ & \exp\Bigg(\mathrm{i}t \log\Bigg[1 + \frac{1}{\sigma} \exp\left\{ \left(1 - \frac{1-\surd r}{\sigma^2}\right) y \right\} \\ & \times \exp\left\{ -\frac{y^2}{2x_0^2} \left(\frac{1}{\sigma^2} - 1\right) \right\} \Bigg] \Bigg) - 1 - \frac{1}{\sigma} \exp\left\{ \left(1 - \frac{1-\surd r}{\sigma^2}\right) y \right\} \\ & \times \exp\left\{ -\frac{y^2}{2x_0^2} \left(\frac{1}{\sigma^2} - 1\right) \right\} \exp(-y) \exp\left(-\frac{y^2}{2x_0^2} \right). \end{aligned}$$

It is seen that, pointwise, $h_n(u)$ converges to

$$h(y) = \left\{ \exp\left(\mathrm{i}t \log\left[1 + \frac{1}{\sigma} \exp\left\{ \left(1 - \frac{1-\surd r}{\sigma^2}\right) y \right\} \right] \right) - 1 - \frac{1}{\sigma} \exp\left\{ \left(1 - \frac{1-\surd r}{\sigma^2}\right) y \right\} \exp(-y). \right.$$

At the same time, note that

$$|\exp[\mathrm{i}t\{1+\exp(y)\}] - 1 - \mathrm{i}t \exp(y)| \leqslant C \min\{\exp(y), \exp(2y)\}.$$

It is seen that

$$|h_n(y)| \leqslant C \exp(-y) \min\left[\exp\left\{ \left(1 - \frac{1-\surd r}{\sigma^2}\right) y \right\}, \exp\left\{ 2\left(1 - \frac{1-\surd r}{\sigma^2}\right) y \right\} \right].$$

The key fact here is that, by expression (76), $0<(1-\surd r)/\sigma^2<\frac{1}{2}$. Therefore,

$$\exp(-y)\min\left[\exp\left(1-\frac{1-\surd r}{\sigma^2}\right),\ \exp\left\{2\left(1-\frac{1-\surd r}{\sigma^2}\right)y\right\}\right]=\begin{cases}\exp\left(-\dfrac{1-\surd r}{\sigma^2}y\right), & y\geqslant 0,\\ \exp\left\{\left(1-2\dfrac{1-\surd r}{\sigma^2}\right)y\right\}, & y<0,\end{cases}$$

where the right-hand side is integrable. It follows from the dominated convergence theorem that

$$n\ E[f_n(X)]\to(2\pi)^{-1/2}\int h(x)\,\mathrm{d}x,$$

which proves the claim.

Consider the case $\sigma\geqslant\surd 2$. The proof is similar to the case $\sigma<\surd 2$ and $\beta>1-\sigma^2/4$ so we omit it. This concludes the claim.

B.2. Proof of lemma 1

Consider the first claim. Fix $r<q\leqslant 1$, by Mills's ratio (Wasserman, 2006),

$$\bar\Phi[\surd\{2q\log(n)\}]=\mathrm{PL}(n)n^{-q},$$
$$\bar\Phi\left[\frac{\surd\{2q\log(n)\}-A_n}{\sigma}\right]=\mathrm{PL}(n)n^{-(\surd q-\surd r)^2/\sigma^2}.$$

It follows that

$$\surd n\frac{\bar F(t)-\bar\Phi(t)}{\surd\{\bar\Phi(t)\ \Phi(t)\}}=\mathrm{PL}(n)n^{\delta(q;\beta,r,\sigma)},$$

where

$$\delta(q;\beta,r,\sigma)=(1+q)/2-\beta-(\surd q-\surd r)^2/\sigma^2.$$

It suffices to show that $\delta(q;\beta,r,\sigma)$ reaches its maximum at

$$q=\min\left\{\left(\frac{2}{2-\sigma^2}\right)^2 r,1\right\}$$

when $\sigma<\surd 2$ and at $q=1$ otherwise.

For this, we note that, first, when $\sigma<\surd 2$ and $r<(2-\sigma^2)^2/4$, $\delta(q;\beta,r,\sigma)$ maximizes at $q=4r/(2-\sigma^2)^2<1$ and is monotonically decreasing on both sides, and the claim follows. Second, when either $\sigma<\surd 2$ and $r\geqslant(2-\sigma^2)^2/4$ or $\sigma\geqslant\surd 2$, $\delta(q;\beta,r,\sigma)$ is monotonically increasing. Combining these gives the claim.

B.3. Proof of lemma 2

Consider the first claim. Direct calculations show that

$$\varepsilon_n\ E[g_n(X)\ \mathbf{1}_{\{D_n^c\}}]=\varepsilon_n\int_{|x|>\surd\{2\log(n)\}}\phi_\sigma(x-A_n)\,\mathrm{d}x$$
$$=\varepsilon_n\left(\bar\Phi\left[\frac{1-\surd r}{\sigma}\surd\{2\log(n)\}\right]+\bar\Phi\left[\frac{1+\surd r}{\sigma}\surd\{2\log(n)\}\right]\right).$$

Note that $\bar\Phi(x)\leqslant C\,\phi(x)$ for $x>0$; the last term is no greater than

$$C\varepsilon_n\left(\phi\left[\frac{1-\surd r}{\sigma}\surd\{2\log(n)\}\right]+\phi\left[\frac{1+\surd r}{\sigma}\surd\{2\log(n)\}\right]\right)=Cn^{-\{\beta+(1-\surd r)^2/\sigma^2\}}.$$

By the assumption, $r<\{1-\sigma\surd(1-\beta)\}^2$. The claim follows by

$$\beta+\frac{(1-\surd r)^2}{\sigma^2}=1-\left\{1-\beta-\frac{(1-\surd r)^2}{\sigma^2}\right\}>1.$$

Consider the second claim. We discuss the case $\sigma \geqslant \surd 2$ and the case $\sigma < \surd 2$ separately. When $\sigma \geqslant \surd 2$, write

$$g_n^2(x)\,\phi(x) = C\,\exp\left\{\left(\frac{1}{2}-\frac{1}{\sigma^2}\right)x^2 + \frac{2A_n x}{\sigma^2} - \frac{A_n^2}{\sigma^2}\right\},$$

which is a convex function of x. Therefore, the extreme value over the range of $|x| \leqslant \surd\{2\log(n)\}$ occurs at the end points, which is seen to be

$$g_n^2\{\surd\{2\log(n)\}\,\phi(\surd\{2\log(n)\} = Cn^{1-2(1-\surd r)^2/\sigma^2}.$$

Therefore,

$$\varepsilon_n^2\,E[g_n^2(X)\,\mathbf{1}_{\{D_n\}}] \leqslant C\surd\log(n) n^{1-2\{\beta+(1-\surd r)^2/\sigma^2\}}.$$

By the assumption of $r < \{1-\sigma\surd(1-\beta)\}^2$, $\beta + (1-\surd r)^2/\sigma^2 > 1$, and the claim follows.

When $\sigma < \surd 2$, we similarly have

$$\varepsilon_n^2\,E[g_n^2(X)\,\mathbf{1}_{\{D_n\}}] \leqslant C\varepsilon_n^2 \int_{x\leqslant\surd\{2\log(n)\}} \exp\left\{\left(\frac{1}{2}-\frac{1}{\sigma^2}\right)x^2 + \frac{2A_n x}{\sigma^2} - \frac{A_n^2}{\sigma^2}\right\}\mathrm{d}x.$$

Write

$$\left(\frac{1}{2}-\frac{1}{\sigma^2}\right)x^2 + \frac{2A_n x}{\sigma^2} - \frac{A_n^2}{\sigma^2} = -\left(\frac{1}{\sigma^2}-\frac{1}{2}\right)\left(x - \frac{A_n}{1-\sigma^2/2}\right)^2 + \frac{A_n^2}{2-\sigma^2},$$

By a change of variables,

$$\begin{aligned}\varepsilon_n^2\,E[g_n^2(X)\,\mathbf{1}_{\{D_n\}}] &\leqslant Cn^{-2\beta+2r/(2-\sigma^2)} \int_{y\leqslant\surd\{2\log(n)\}-A_n/(1-\sigma^2/2)} \exp\left\{-\left(\frac{1}{\sigma^2}-\frac{1}{2}\right)y^2\right\}\mathrm{d}y \\ &= Cn^{-2\beta+2r/(2-\sigma^2)}\,\Phi\left(\frac{\surd(2-\sigma^2)}{\sigma}\left[\surd\{2\log(n)\} - \frac{A_n}{1-\sigma^2/2}\right]\right).\end{aligned}$$

Rewrite

$$\surd\{2\log(n)\} - \frac{A_n}{1-\sigma^2/2} = \surd\{2\log(n)\}\left(1 - \frac{2\surd r}{2-\sigma^2}\right),$$

and note that $\Phi(x) \leqslant C\,\phi(x)$ when $x<0$ and $\Phi(x) \leqslant 1$ otherwise; we have

$$\varepsilon_n^2\,E[g_n^2(X)\,\mathbf{1}_{\{D_n\}}] \leqslant C\begin{cases} n^{-2\beta+2r/(2-\sigma^2)}, & r \leqslant \frac{1}{4}(2-\sigma^2)^2, \\ n^{-2\beta+2r/(2-\sigma^2)-(2-\sigma^2)/\sigma^2\{1-2\surd r/(2-\sigma^2)\}^2}, & \text{otherwise}. \end{cases} \tag{77}$$

We now discuss the two cases $r \leqslant \min\{\frac{1}{4}(2-\sigma^2)^2, \rho^*(\beta,\sigma)\}$ and $\frac{1}{4}(2-\sigma^2)^2 < r < \rho^*(\beta,\sigma)$ separately. In the first case, $r < (2-\sigma^2)(\beta-\frac{1}{2})$ and $r < \frac{1}{4}(2-\sigma^2)^2$, and so

$$-2\beta + 2r/(2-\sigma^2) < -2\beta + 2(\beta - \tfrac{1}{2}) = -1;$$

the claim follows directly from expression (77).

In the second case, note that this case is only possible when $\beta > 1-\sigma^2/4$. Therefore, $r < \{1-\sigma\surd(1-\beta)\}^2$, and

$$-2\beta + \frac{2r}{2-\sigma^2} - \frac{1}{\sigma^2}(2-\sigma^2)\left(1-\frac{2\surd r}{2-\sigma^2}\right)^2 = 1 - 2\left\{\beta + \frac{1}{\sigma^2}(1-\surd r)^2\right\} < -1.$$

Applying expression (77) gives the claim.

B.4. Proof of lemma 3

It is not necessary that expressions (36) and (37) are simultaneously true. We prove the claim for three cases separately:

(a) $\frac{1}{2}<\beta<1$ and $r>\{1-\sigma\surd(1-\beta)\}^2$ and $\sigma<\surd 2$; or $\frac{1}{2}<\beta<1$ and $r>\rho^*(\beta;\sigma)$ and $\sigma\geqslant\surd 2$,
(b) $\frac{1}{2}<\beta<1-\sigma^2/4$ and $(2-\sigma^2)(\beta-\frac{1}{2})<r<\{1-\sigma\surd(1-\beta)\}^2$ and $1<\sigma<\surd 2$, and
(c) $\frac{1}{2}<\beta<1-\sigma^2/4$ and $(2-\sigma^2)(\beta-\frac{1}{2})<r<\{1-\sigma\surd(1-\beta)\}^2$ and $\sigma<1$.

The discussion for cases where (β,r,σ) fall right on the boundaries of the partition of these subregions is similar, so we omit it.

For case (a), we show that expression (36) holds. For (β,r,σ) in this range, by elementary algebra and the definition of $\rho^*(\beta,\sigma)$,

$$1-\beta-\frac{(1-\surd r)^2}{\sigma^2}>0. \tag{78}$$

Also, $\varepsilon_n\, g_n\{\surd 2\log(n)\}=(1/\sigma)n^{1-\beta-(1-\surd r)^2/\sigma^2}$, which is larger than 1 for sufficiently large n, so

$$n\varepsilon_n\, E[g_n(X)\,\mathbf{1}_{\{\varepsilon_n\, g_n(X)>1\}}]\geqslant n\varepsilon_n\, E[g_n(X)\,\mathbf{1}_{\{X\geqslant\surd\{2\log(n)\}\}}]=n\varepsilon_n\int_{\surd\{2\log(n)\}}^{\infty}\frac{1}{\sigma}\phi\left(\frac{x-A_n}{\sigma}\right)\mathrm{d}x.$$

By elementary calculus and Mills's ratio (Wasserman, 2006), the right-hand side equals $\mathrm{PL}(n)n^{1-\beta-(1-\surd r)^2/\sigma^2}$. The claim follows directly from inequality (78).

For case (b), we show that expression (37) holds. It is seen that $\sup_{\{0\leqslant x\leqslant\surd\{2\log(n)\}\}}\{\varepsilon_n\, g_n(x)\}=o(1)$ for (β,r,σ) in this range, so

$$n\varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{\varepsilon_n\, g_n(X)\leqslant 1\}}]\geqslant n\varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{0\leqslant X\leqslant\surd\{2\log(n)\}\}}].$$

Direct calculations show that

$$n\varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{0\leqslant X\leqslant\surd\{2\log(n)\}\}}]=n\varepsilon_n^2\,\exp\left(\frac{A_n^2}{2-\sigma^2}\right)\Phi\left[\frac{\surd(2-\sigma^2)}{\sigma}\left(1-\frac{\surd r}{1-\sigma^2/2}\right)\surd\{2\log(n)\}\right].$$

By basic algebra, for (β,r,σ) in the current range,

$$\frac{\surd(2-\sigma^2)}{\sigma}\left(1-\frac{\surd r}{1-\sigma^2/2}\right)>0.$$

Combining these gives

$$n\varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{\varepsilon_n\, g_n(X)\leqslant 1\}}]\gtrsim n\varepsilon_n^2\,\exp\left(\frac{A_n^2}{2-\sigma^2}\right)=n^{1-2\beta+2r/(2-\sigma^2)}.$$

The claim follows as $1-2\beta+2r/(2-\sigma^2)>0$.

For case (c), we consider two subcases separately:

(i) $\frac{1}{2}<\beta<1-\sigma^2/4$ and $r<(1-\sigma^2)\beta$ and $\sigma<1$; or $1-\sigma^2<\beta<1-\sigma^2/4$ and $r\geqslant(1-\sigma^2)\beta$ and $\sigma<1$, and
(ii) $\frac{1}{2}<\beta<1-\sigma^2$ and $r\geqslant(1-\sigma^2)\beta$ and $\sigma<1$.

We show that expression (36) holds in cases (a) and (ii), whereas expression (37) holds in cases (b) and (i).

For case (i), we show that expression (37) holds. Similarly, for (β,r,σ) in this range, $\sup_{\{0<x<\surd\{2\log(n)\}\}}\{\varepsilon_n\times g_n(x)\}=o(1)$ and so

$$n\varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{\varepsilon_n\, g_n(X)\leqslant 1\}}]\geqslant n\varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{0<X\leqslant\surd\{2\log(n)\}\}}].$$

For (β,r,σ) in the current range, $n\varepsilon_n^2\, E[g_n^2(X)\,\mathbf{1}_{\{0<X\leqslant\surd\{2\log(n)\}\}}]\sim n^{1-2\beta+2r/(2-\sigma^2)}$, where the exponent is positive. The claim follows.

Consider case (ii). Introduce

$$\Delta=\Delta(\beta,r,\sigma)=\frac{[\surd r-\sigma\surd\{r-(1-\sigma^2)\beta\}]^2}{(1-\sigma^2)^2}.$$

For (β,r,σ) in this range elementary calculus shows that $\surd r<\Delta<1$, and that, for sufficiently large n, $\varepsilon_n\, g_n(x)\geqslant 1$ for $\surd\{2\Delta\log(n)\}\leqslant x\leqslant\surd\{2\Delta\log(n)\}+\surd\log\{\log(n)\}$. It follows that

$$n\varepsilon_n\, E[g_n(X)\,\mathbf{1}_{\{\varepsilon_n g_n(X)>1\}}] \geqslant n\varepsilon_n \int_{\surd\{2\Delta\log(n)\}}^{\surd\{2\Delta\log(n)\}+\surd\log\{\log(n)\}} \frac{1}{\sigma}\phi\left(\frac{x-A_n}{\sigma}\right)\mathrm{d}x \gtrsim \frac{C}{\surd\log(n)} n^{1-\beta-(\surd\Delta-\surd r)^2/\sigma^2},$$

where we have used $\Delta > r$. Fixing (β, σ), $\surd\Delta - \surd r$ is decreasing in r. So, for all $r \geqslant (1-\sigma^2)\beta$,

$$1-\beta-\frac{(\surd\Delta-\surd r)^2}{\sigma^2} \geqslant 1-\beta-\frac{(\surd\Delta-\surd r)^2}{\sigma^2}\bigg|_{\{r=(1-\sigma^2)\beta\}} = 1-\frac{\beta}{1-\sigma^2},$$

which is larger than 0 since $\beta < 1-\sigma^2$. Combining these gives the claim.

References

Burnashev, M. V. and Begmatov, I. A. (1991) On a problem of detecting a signal that leads to stable distributions. *Theor. Probab. Appl.*, **35**, 556–560.

Cai, T., Jin, J. and Low, M. (2007) Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, **35** 2421–2449.

Cayon, L., Jin, J. and Treaster, A. (2005) Higher Criticism statistic: detecting and identifying non-Gaussianity in the WMAP First Year data. *Mnthly Notes R. Astron. Soc.*, **362**, 826–832.

Delaigle, A., Hall, P. and Jin, J. (2011) Robustness and accuracy of methods for high dimensional data analysis based on Student's *t*-statistic. *J. R. Statist. Soc.* B, **73**, 283–301.

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

Donoho, D. and Jin, J. (2008) Higher Criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc. Natn. Acad. Sci. USA*, **105**, 14790–14795.

Donoho, D. and Jin, J. (2009) Feature selection by Higher Criticism thresholding: optimal phase diagram. *Phil. Trans. R. Soc. Lond.* A, **367**, 4449–4470.

Hall, P. and Jin, J. (2008) Properties of Higher Criticism under long-range dependence. *Ann. Statist.*, **36**, 381–402.

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

Hall, P., Pittelkow, Y. and Ghosh, M. (2008) Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *J. R. Statist. Soc.* B, **70**, 159–173.

Hopkins, A. M., Miller, C. J., Connolly, A. J., Genovese, C., Nichol, R. C. and Wasserman, L. (2002) A new source detection algorithm using the false-discovery rate. *Astron. J.*, **123**, 1086–1094.

Ingster, Y. I. (1997) Some problems of hypothesis testing leading to infinitely divisible distribution. *Math. Meth. Statist.*, **6**, 47–69.

Ingster, Y. I. (1999) Minimax detection of a signal for l_n^p-balls. *Math. Meth. Statist.*, **7**, 401–428.

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

Ji, P. and Jin, J. (2010) UPS delivers optimal phase diagram in high dimensional variable selection. *Manuscript*. (Available from `http://arxiv.org/abs/1010.5028`.)

Jin, J. (2003) Detecting and estimating sparse mixtures. *PhD Thesis*. Department of Statistics, Stanford University, Stanford.

Jin, J. (2004) Detecting a target in very noisy data from multiple looks. In *A Festschrift for Herman Rubin*, pp. 255–286. Beachwood: Institute of Mathematical Statistics.

Jin, J. (2009) Impossibility of successful classification when useful features are rare and weak. *Proc. Natn. Acad. Sci. USA*, **106**, 8856–8864.

Kulldorff, M., Heffernan, R., Hartan, J., Assuncao, R. and Mostashari, F. (2005) A space-time permutation scan statistic for disease outbreak detection. *PLOS Med.*, **2**, no. 3, e59.

Meinshausen, N. and Buhlmann, P. (2006) High dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**, 1436–1462.

Shorack, G. R. and Wellner, J. A. (2009) *Empirical Processes with Applications to Statistics*. Philadelphia: Society for Industrial and Applied Mathematics.

Sun, W. and Cai, T. T. (2007) Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Statist. Ass.*, **102**, 901–912.

Wasserman, L. (2006) *All of Nonparametric Statistics*. New York: Springer.

Xie, J., Cai, T. and Li, H. (2011) Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika*, **98**, 273–290.

Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.