Optimal Detection of Heterogeneous and Heteroscedastic Mixtures

T. Tony Cai
Department of Statistics, University of Pennsylvania
X. Jessie Jeng*
Department of Biostatistics and Epidemiology, University of Pennsylvania
Jiashun Jin
Department of Statistics, Carnegie Mellon University

October 28, 2010

Abstract

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 reliably detect the presence of non-null effect, 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 the specific information of model parameters, is optimally adaptive to the unknown degrees of heterogeneity and heteroscedasticity in both the sparse and dense cases.

Keywords: Detection boundary, Higher Criticism, Likelihood Ratio Test (LRT), optimal adaptivity, sparsity.

AMS 2000 subject classifications: Primary-62G10; secondary 62G32, 62G20.

Acknowledgments: The authors would like to thank Mark Low for helpful discussion. Jeng and Jin were partially supported by NSF grant DMS-0639980 and DMS-0908613. Tony Cai was supported in part by NSF Grant DMS-0604954 and NSF FRG Grant DMS-0854973.

*Corresponding author. E-mail: xjeng@upenn.edu.

1 Introduction

The problem of detecting non-null components in Gaussian mixtures arises in many applications, where a large amount 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 disease outbreaks 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 Genome-Wide Association Studies (GWAS), a typical single-nucleotide polymorphism (SNP) data set consists of a very long sequence of measurements containing signals that are both sparse and weak. To better locate such signals, 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 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 two-fold: (a) Discover the *detection boundary* in the parameter space that separates the *detectable region*, where it is possible to reliably detect the existence of signals based on the noisy and mixed observations, from the *undetectable region*, where it is impossible to do so. (b) 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 the *optimal adaptivity*.

The problem is formulated as follows. Given n independent observation units $X = (X_1, X_2, \ldots, X_n)$. For each $1 \leq i \leq n$, we suppose that X_i has probability ϵ to be a non-null effect and probability $1-\epsilon$ to be 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 heterogeneous parameter, and σ the heteroscedastic 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. That is, one wishes to test the hypothesis $\epsilon = 0$ or equivalently, test the joint null hypothesis

$$H_0: \qquad X_i \overset{iid}{\sim} N(0,1), \qquad 1 \leq i \leq n, \tag{1.1}$$

against a specific alternative hypothesis in its complement

$$H_1^{(n)}: \qquad X_i \overset{iid}{\sim} (1-\epsilon)N(0,1) + \epsilon N(A, \sigma^2), \qquad 1 \leq i \leq n. \tag{1.2}$$

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 mentioned earlier, we focus on the case where ϵ is small. We adopt an asymptotic framework where n is the driving variable, while 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

$$\epsilon = \epsilon_n = n^{-\beta}. \tag{1.3}$$

The detection problem behaves very differently in two regimes: the *sparse regime* where $1/2 < \beta < 1$ and the *dense regime* where $0 < \beta \leq 1/2$. In the sparse regime, $\epsilon_n \ll 1/\sqrt{n}$, 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 $\epsilon_n \gg 1/\sqrt{n}$, 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:

$$A_n(r;\beta) = \sqrt{2r\log n}, \qquad 0 < r < 1, \qquad \text{if } 1/2 < \beta < 1 \text{ (sparse case)}, \tag{1.4}$$

$$A_n(r;\beta) = n^{-r}, \qquad 0 < r < 1/2, \qquad \text{if } 0 < \beta \leq 1/2 \text{ (dense case)}. \tag{1.5}$$

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

$$p_i = \bar{\Phi}(X_i) \equiv P\{N(0,1) \geq X_i\}, \tag{1.6}$$

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

$$HC_n^* = \max_{\{1 \leq i \leq n\}} HC_{n,i}, \qquad \text{where} \qquad HC_{n,i} = \sqrt{n}\left[\frac{i/n - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}\right], \tag{1.7}$$

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 $X \sim N(0, I_n)$, $p_i \overset{iid}{\sim} U(0,1)$ and so $HC_{n,i} \approx N(0,1)$. Therefore, by the well-known results from empirical processes (e.g. Shorack and Wellner (2009)), $HC_n^* \approx \sqrt{2\log\log n}$, which grows to ∞ very slowly. In contrast, if $X \sim N(\mu, I_n)$ where some of the coordinates of μ is nonzero, then $HC_{n,i}$ has an elevated mean for some i, and HC_n^* could grow to ∞ algebraically fast. Consequently, Higher Criticism is able to separate two hypotheses even in the very sparse case. We mention that (1.7) is only one variant of the 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

dense cases are considered. We believe that heteroscedasticity is a more natural assumption in many applications. For example, signals can often bring additional variations to the background. This phenomenon can be captured by the Gaussian hierarchical model:

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

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

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

which is heteroscedastic as $\sigma > 1$.

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 the 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, on the other hand, focuses on sparse signals that individually may be moderately strong. As a result, to derive the detection boundary for the latter, one usually needs unconventional analysis. Note that 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 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 tends to ∞, 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 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 the 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 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 of the signal sparsity is usually unknown, and the lack of a unified approach restricts the discovery of the full catalog of signals. The adaptivity of HC 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 on 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 discussions in Section 4.

The rest of the paper is organized as follows. Section 2 demonstrates the detection boundaries in the sparse and dense cases, respectively. Limiting behaviors 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 to other methods are also presented. Section 4 discusses other closely related problems including proportion estimation and signal identification. Simulation examples for finite n is given in Section 5. Further extensions and future work are discussed in Section 6. Main proofs are presented in Section 7. Appendix includes complementary technical details.

2 Detection boundary

The meaning of 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 tends to ∞. In the interior of the detectable region, the sum of Type I and Type II errors of Neyman-Pearson's Likelihood Ratio Test (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 (1.3)-(1.4). We find the exact expression of $\rho^*(\beta;\sigma)$ as follows,

$$\rho^*(\beta;\sigma) = \begin{cases} (2-\sigma^2)(\beta-1/2), & 1/2 < \beta \le 1-\sigma^2/4, \\ (1-\sigma\sqrt{1-\beta})^2, & 1-\sigma^2/4 < \beta < 1, \end{cases} \qquad 0 < \sigma < \sqrt{2}, \tag{2.8}$$

and

$$\rho^*(\beta;\sigma) = \begin{cases} 0, & 1/2 < \beta \le 1-1/\sigma^2, \\ (1-\sigma\sqrt{1-\beta})^2, & 1-1/\sigma^2 < \beta < 1, \end{cases} \qquad \sigma \ge \sqrt{2}. \tag{2.9}$$

Note that when $\sigma = 1$, the detection boundary $r = \rho^*(\beta;\sigma)$ reduces to the detection boundary in Donoho and Jin (2004) (see also Ingster (1997), Ingster (1999), and Jin (2004)). The curve $r = \rho^*(\beta;\sigma)$ is plotted in the left panel of Figure 1 for $\sigma = 0.6$, 1, $\sqrt{2}$ and 3. The detectable and undetectable regions correspond to $r > \rho^*(\beta;\sigma)$ and $r < \rho^*(\beta;\sigma)$, respectively.

When $r < \rho^*(\beta;\sigma)$, the Hellinger distance between the joint density of X_i under the null and that under the alternative tends to 0 as n tends to ∞, 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 Section 7.1.

Theorem 2.1 *Let ϵ_n and A_n be calibrated as in (1.3)-(1.4) and let $\sigma > 0$, $\beta \in (1/2, 1)$, and $r \in (0, 1)$ be fixed such that $r < \rho^*(\beta;\sigma)$, where $\rho^*(\beta;\sigma)$ is as in (2.8)-(2.9). Then for any test the sum of Type I and Type II error probabilities tends to 1 as $n \to \infty$.*

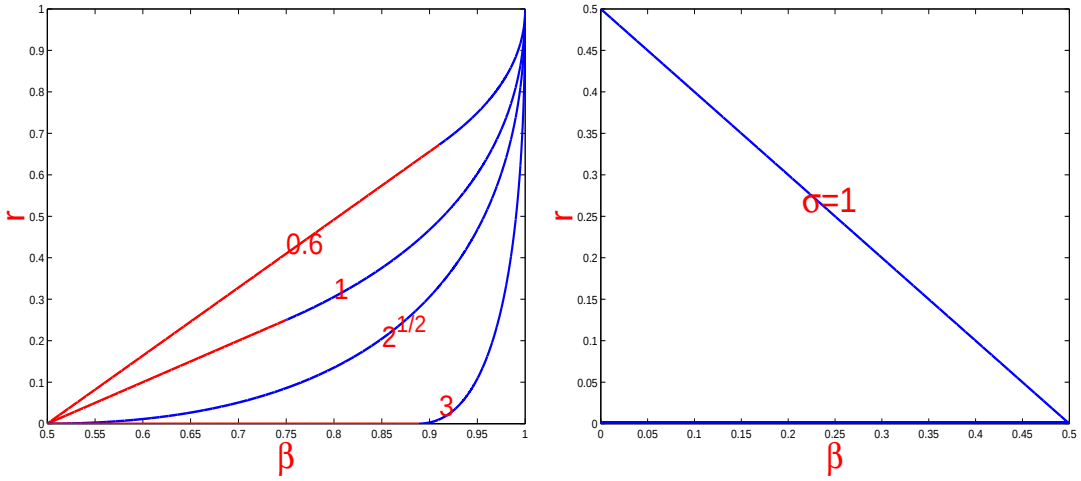

Figure 1: Left: Detection boundary $r = \rho^*(\beta;\sigma)$ in the sparse case for $\sigma = 0.6,\ 1,\ \sqrt{2}$ and 3. The detectable region is $r > \rho^*(\beta;\sigma)$, and the undetectable region is $r < \rho^*(\beta;\sigma)$. Right: Detection boundary $r = \rho^*(\beta;\sigma)$ in the dense case for $\sigma = 1$. The detectable region is $r < \rho^*(\beta;\sigma)$, and the undetectable region is $r > \rho^*(\beta;\sigma)$.

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

$$LR_n = LR_n(X_1, X_2, \ldots, X_n; \beta, r, \sigma),$$

and consider the LRT which rejects H_0 if and only if

$$\log(LR_n) > 0. \tag{2.10}$$

The following theorem, which is proved in Section 7.2, shows that when $r > \rho^*(\beta;\sigma)$, $\log(LR_n)$ converges to $\mp\infty$ in probability, under the null and the alternative, respectively. Therefore, asymptotically the alternative hypothesis can be perfectly separated from the null by the LRT.

Theorem 2.2 *Let ϵ_n and A_n be calibrated as in (1.3)-(1.4) and let $\sigma > 0$, $\beta \in (1/2, 1)$, and $r \in (0, 1)$ be fixed such that $r > \rho^*(\beta;\sigma)$, where $\rho^*(\beta;\sigma)$ is as in (2.8)-(2.9). As $n \to \infty$, $\log(LR_n)$ converges to $\mp\infty$ in probability, under the null and the alternative, 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 the left panel of Figure 1. As σ increases, the curve $r = \rho^*(\beta;\sigma)$ moves towards the south-east corner; the detectable region gets larger which implies that the detection problem gets 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 (1.3) and (1.5). We find the detection boundary as $r = \rho^*(\beta;\sigma)$, where

$$\rho^*(\beta;\sigma) = \begin{cases} \infty, & \sigma \neq 1, \\ 1/2 - \beta, & \sigma = 1, \end{cases} \qquad 0 < \beta < 1/2. \tag{2.11}$$

The curve $r = \rho^*(\beta;\sigma)$ is plotted in the right panel of Figure 1 for $\sigma = 1$ and $\sigma \neq 1$. Note that, unlike that 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 Section 7.3 and 7.4.

Theorem 2.3 *Let ϵ_n and A_n be calibrated as in (1.3) and (1.5) and let $\sigma > 0$, $\beta \in (0, 1/2)$, and $r \in (0, 1/2)$ be fixed such that $r > \rho^*(\beta;\sigma)$, where $\rho^*(\beta;\sigma)$ is defined in (2.11). Then for any test the sum of Type I and Type II error probabilities tends to 1 as $n \to \infty$.*

Theorem 2.4 *Let ϵ_n and A_n be calibrated as in (1.3) and (1.5) and let $\sigma > 0$, $\beta \in (0, 1/2)$, and $r \in (0, 1/2)$ be fixed such that $r < \rho^*(\beta;\sigma)$, where $\rho^*(\beta;\sigma)$ is defined in (2.11). Then, the sum of Type I and Type II error probabilities of the LRT tends to 0 as $n \to \infty$.*

Comparing (2.11) with (2.8)-(2.9), we see that the detection boundary in the dense case is very different from that in the sparse case. In particular, heteroscedasticity is more crucial in the dense case, and the non-null component is always detectable when $\sigma \neq 1$.

2.3 Limiting behavior of LRT on the detection boundary

In the preceding section, we examine the situation when the parameters (β, r) fall strictly in the interior of either the detectable or undetectable region. When these parameters get very close to the detection boundary, the behavior of the LRT becomes more subtle. In this section, we discuss the behavior 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, 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 the sparse case, but simpler. Due to limitation in space, we only present the details for the sparse case.

We introduce the following calibration:

$$A_n = \sqrt{2r \log n}, \qquad \epsilon_n = \begin{cases} n^{-\beta}, & 1/2 < \beta \leq 1 - \sigma^2/4, \\ n^{-\beta}(\log(n))^{(1-\sqrt{1-\beta}/\sigma)}, & 1 - \sigma^2/4 < \beta < 1. \end{cases} \tag{2.12}$$

Compared to the calibrations in (1.3)-(1.4), A_n remains the same but ϵ_n is modified slightly so that the limiting distribution of LRT would be non-degenerate. Denote

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

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

$$\psi^0_{\beta,\sigma}(t) = \frac{1}{2\sqrt{\pi}\sigma^{1/(\sigma^2-1)}(\sigma-\sqrt{1-\beta})}\int_{-\infty}^{\infty}\big[e^{it\log(1+e^y)}-1-ite^y\big]e^{(\frac{\sigma-2\sqrt{1-\beta}}{\sigma-\sqrt{1-\beta}}-2)y}dy$$

and

$$\psi^1_{\beta,\sigma}(t) = \frac{1}{2\sqrt{\pi}\sigma^{\sigma^2/(\sigma^2-1)}(\sigma-\sqrt{1-\beta})}\int_{-\infty}^{\infty}\big[e^{it\log(1+e^y)}-1\big]e^{(\frac{\sigma-2\sqrt{1-\beta}}{\sigma-\sqrt{1-\beta}}-1)y}dy,$$

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

Theorem 2.5 *Let A_n and ϵ_n be defined as in (2.12), and let $\rho^*(\beta;\sigma)$ be as in (2.8)-(2.9). Fix $\sigma \in (0,\sqrt{2})$, $\beta \in (1/2,1)$, and set $r = \rho^*(\beta,\sigma)$. As $n \longrightarrow \infty$,*

$$\log(LR_n) \xrightarrow{L} \begin{cases} N(-\frac{b(\sigma)}{2}, b(\sigma)), & 1/2 < \beta < 1-\sigma^2/4, \\ N(-\frac{b(\sigma)}{4}, \frac{b(\sigma)}{2}), & \beta = 1-\sigma^2/4, \\ \nu^0_{\beta,\sigma}, & 1-\sigma^2/4 < \beta < 1, \end{cases} \quad \text{under } H_0,$$

and

$$\log(LR_n) \xrightarrow{L} \begin{cases} N(\frac{b(\sigma)}{2}, b(\sigma)), & 1/2 < \beta < 1-\sigma^2/4, \\ N(\frac{b(\sigma)}{4}, b(\sigma)/2), & \beta = 1-\sigma^2/4, \\ \nu^1_{\beta,\sigma}, & 1-\sigma^2/4 < \beta < 1, \end{cases} \quad \text{under } H_1^{(n)},$$

where $\xrightarrow{L}$ denotes "converges in law".

Note that the limiting distribution is Gaussian when $\beta \leq 1-\sigma^2/4$ and non-Gaussian otherwise.

Next, we consider the case of $\sigma \geq \sqrt{2}$, where the range of interest is $\beta > 1-1/\sigma^2$.

Theorem 2.6 *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 (2.12). Then as $n \longrightarrow \infty$,*

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

In this case, the limiting distribution is always non-Gaussian. This phenomenon (i.e., the weak limits of the log-likelihood ratio might be nonGaussian) 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 Figure 2, we fix $(\beta,\sigma) = (0.75, 1.1)$, and plot the characteristic functions and the density functions corresponding to the limiting distribution of $\log(LR_n)$. Two density functions are generally overlapping with 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 tends to ∞.

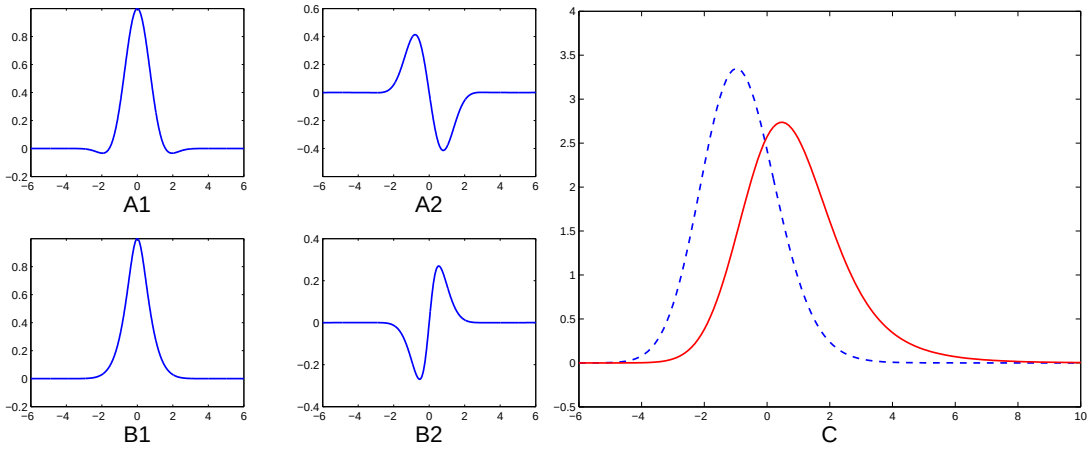

Figure 2: Characteristic functions and density functions of $\log(LR_n)$ for $(\beta, \sigma) = (0.75, 1.1)$. A1 and A2 show the real and imaginary parts of $e^{\psi^0_{\beta,\sigma}}$, B1 and B2 show the real and imaginary parts of $e^{\psi^0_{\beta,\sigma}+\psi^1_{\beta,\sigma}}$, and C shows the density functions of $\nu^0_{\beta,\sigma}$ (dashed) and $\nu^1_{\beta,\sigma}$ (solid).

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 the Higher Criticism, which is a non-parametric procedure, is successful in the entire detectable region for both the sparse and dense cases. This property is called the *optimal adaptivity* of Higher Criticism. Donoho and Jin (2004) discovered this property in the case of $\sigma = 1$ and $\beta \in (1/2, 1)$. 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 HC statistic by using the absolute value of $HC_{n,i}$:

$$HC^*_n = \max_{1\le i\le n} |HC_{n,i}|, \tag{3.13}$$

where $HC_{n,i}$ is defined as in (1.7). Recall that, under the null,

$$HC^*_n \approx \sqrt{2\log\log n}.$$

So a convenient critical point for rejecting the null is when

$$HC^*_n \ge \sqrt{2(1+\delta)\log\log n}, \tag{3.14}$$

where $\delta > 0$ is any fixed constant. The following theorem is proved in Section 7.5.

Theorem 3.1 *Suppose ϵ_n and A_n either satisfy (1.3) and (1.4) and $r > \rho^*(\beta;\sigma)$ with $\rho^*(\beta;\sigma)$ defined as in (2.8) and (2.9), or ϵ_n and A_n satisfy (1.3) and (1.5) and $r < \rho^*(\beta;\sigma)$ with $\rho^*(\beta;\sigma)$ defined as in (2.11). Then the test which rejects H_0 if and only if $HC^*_n \ge \sqrt{2(1+\delta)\log\log n}$ satisfies*

$$P_{H_0}\{\text{Reject } H_0\} + P_{H_1^{(n)}}\{\text{Reject } H_1^{(n)}\} \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty.$$

The above theorem states, somewhat surprisingly, that the optimal adaptivity of Higher Criticism continues to hold even when the data poses 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 higher threshold. Higher Criticism is able to 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, one 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 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, the simulation experiment can be finished reasonably fast. We find that, due to the slow convergence of the iterative logarithmic law, critical values determined in this way are usually much more accurate than $\sqrt{2(1+\delta)\log\log n}$.

3.1 How Higher Criticism works

We now illustrate how the 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 cdf and empirical survival function of X_i, respectively,

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i<t\}}, \qquad \bar{F}_n(t) = 1 - F_n(t),$$

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

$$W_n(t) = \sqrt{n}\left(\frac{\bar{F}_n(t) - \bar{\Phi}(t)}{\sqrt{\bar{\Phi}(t)(1-\bar{\Phi}(t))}}\right). \tag{3.15}$$

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

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

Comparing this with (3.13), we have

$$HC_n^* = \sup_{-\infty<t<\infty} |W_n(t)|. \tag{3.16}$$

The proof of (3.16), which we omit, is elementary.

Now, note that for any fixed t,

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

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

$$\left|\sqrt{n}\frac{\bar{F}(t_n)-\bar{\Phi}(t_n)}{\sqrt{\bar{\Phi}(t_n)(1-\bar{\Phi}(t_n))}}\right| \gg \sqrt{2\log\log n}, \tag{3.17}$$

then we can tell the alternative from the null by merely using $W_n(t_n)$. This guarantees the detection succuss of HC.

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

$$t_n^{Ideal}(\beta, r, \sigma) = \mathrm{argmax}_t \left|\sqrt{n}\frac{\bar{F}(t)-\bar{\Phi}(t)}{\sqrt{\bar{\Phi}(t)(1-\bar{\Phi}(t))}}\right|. \tag{3.18}$$

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

$$t_n^*(\beta, r, \sigma) = \begin{cases} \min\{\frac{2}{2-\sigma^2}A_n, \sqrt{2\log n}\}, & \sigma < \sqrt{2}, \\ \sqrt{2\log n}, & \sigma \geq \sqrt{2}. \end{cases} \tag{3.19}$$

The following lemma is proved in the appendix.

Lemma 3.1 *Let ϵ_n and A_n be calibrated as in (1.3)-(1.4). Fix $\sigma > 0$, $\beta \in (1/2, 1)$ and $r \in (0, 1)$ such that $r > \rho^*(\beta, r, \sigma)$, where $\rho^*(\beta, r, \sigma)$ is defined in (2.8) and (2.9). Then*

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

In the dense case when $0 < \beta < 1/2$, the analysis is much simpler. In fact, (3.17) holds under the alternative if $A_n \ll t \leq C$ for some constant C. To show the result, we can simply set the threshold as

$$t_n^*(\beta, r, \sigma) = 1, \tag{3.20}$$

then it follows that

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

One might have expected A_n to be the best threshold as it represents the signal strength. Interestingly, this turns out to be not the case: 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 $\frac{2}{2-\sigma^2}A_n$ 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 has to raise the threshold to counter the fact that there are merely too many noise than signals.

Finally, the optimal adaptivity of Higher Criticism comes from the "sup" part of its definition (see (3.16)). When the null 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

$$HC_n^* \geq W_n(t_n^{Ideal}(\sigma, \beta, r)),$$

the value of the 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 $W_n(t_n^{Ideal}(\sigma,\beta,r))$, despite that 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 to improve this method if it ceases to be optimal in such case? The question is interesting but the answer is not immediately clear. In principle, given the literature on empirical processes and law of iterative logarithm, it is possible to modify the normalizing term of $HC_{n,i}$ so that the resultant HC statistic has a better power. Such a study involves the second order asymptotic expansion of the HC statistic, which not only requires substantially more delicate analysis but also is comparably less important from a practical point of view than the analysis considered here. For these reasons, we leave the exploration along this line to the future.

3.2 Comparison to other testing methods

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

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

The approach is intrinsically related to multiple testing methods including that of Bonferroni and that of controlling the False Discovery Rate (FDR).

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

$$\lim_{n\longrightarrow\infty}\{\mathrm{Max}_n/\sqrt{2\log n}\}\longrightarrow 1, \qquad \text{in probability.}$$

Additionally, if we reject H_0 if and only if

$$\mathrm{Max}_n \ge \sqrt{2\log n}, \tag{3.21}$$

then the Type I error tends to 0 as n tends to ∞. For brevity, we call the test in (3.21) the Max_n.

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

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

So as n tends to ∞, the Type II error of the test (3.21) tends to 0 if and only if

$$\sqrt{r}+\sigma\cdot\sqrt{1-\beta} > 1.$$

Note that this is trivially satisfied when $\sigma\sqrt{1-\beta} > 1$. The discussion is recaptured in the following theorem, the proof of which is omitted.

Theorem 3.2 *Let ϵ_n and A_n be calibrated as in (1.3)-(1.4). Fix $\sigma > 0$ and $\beta \in (1/2, 1)$. As $n \longrightarrow \infty$, the sum of Type I and Type II error probabilities of the test in (3.21) tends to 0 if $r > ((1 - \sigma \cdot \sqrt{1-\beta})_+)^2$ and tends to 1 if $r < ((1 - \sigma \cdot \sqrt{1-\beta})_+)^2$.*

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 sub-optimal. While the comparison is for the sparse case, we note that the dense case is even more favorable for the Higher Criticism. In fact, as n tends to ∞, 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 sample mean, Hotelling's test, Fisher's combined probability test, etc.. These tests have the form of $\sum_{i=1}^n f(X_i)$ for some function f. In fact, Hotelling's test can be recast as $\sum_{i=1}^n X_i^2$, and Fisher's combined probability test can be recast as $-2\sum_{i=1}^n \bar{\Phi}(X_i)$. The key fact is that the standard deviations of such tests usually are of the order of $\sqrt{n}$. But, in the sparse case, the number of non-null effects $\ll \sqrt{n}$. Therefore, these tests are not able to separate the two hypotheses in the sparse case.

4 Detection and related problems

The detection problem 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) provides rate-optimal estimates of the signal proportion ϵ_n and signal mean A_n for the homoscedastic Gaussian mixture: $X_i \sim (1-\epsilon_n)N(0,1)+\epsilon_n N(A_n, 1)$. The techniques developed in that paper can be generalized to estimate the parameters ϵ_n, A_n, and σ in the current heteroscedastic Gaussian mixture setting, $X_i \sim (1-\epsilon_n)N(0,1)+\epsilon_n N(A_n, \sigma^2)$, 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, one wishes to test the hypotheses

$$H_{0,i}: \ X_i \sim N(0,1) \qquad \text{vs.} \qquad H_{1,i}: \ X_i \sim N(A_n, \sigma^2) \tag{4.22}$$

for $1 \le i \le n$. An immediate question is: when the signals are identifiable? It is intuitively clear that it is harder 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

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

We calibrate ϵ_n and A_n by

$$\epsilon_n = n^{-\beta} \ \text{ and } \ A_n = \sqrt{2r \log n}.$$

Note that the above calibration is the same as in (1.4)–(1.5) for the sparse case ($\beta > 1/2$) but is different for the dense case ($\beta < 1/2$). The following theorem is a straightforward

extension of (Ji and Jin, 2010, Theorem 1.1), so we omit the proof. See also Xie et al. (2010).

Theorem 4.1 *Fix $\beta \in (0,1)$ and $r \in (0,\beta)$. For any sequence of multiple testing procedure $\{\hat{T}_n\}_{n=1}^{\infty}$,*

$$\liminf_{n\longrightarrow\infty}\left[\frac{Err(\hat{T}_n)}{n\epsilon_n}\right] \geq 1.$$

Theorem 4.1 shows that if the signal strength is relatively weak, i.e., $A_n = \sqrt{2r\log n}$ for some $0 < r < \beta$, then it is impossible to successfully separate the signals from noise: 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 $n\epsilon_n$.

Theorems 3.1 and 4.1 together depict a picture as follows. Suppose

$$A_n < \sqrt{2\beta\log n}, \text{ if } 1/2 < \beta < 1; \quad n^{\beta-1/2} \ll A_n < \sqrt{2\beta\log n}, \text{ if } 0 < \beta < 1/2. \tag{4.23}$$

Then it is possible to reliably detect the presence of the signals 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 developed in Sun and Cai (2007). By viewing (4.22) as a multiple testing problem, one wishes to test the hypotheses $H_{0,i}$ versus $H_{1,i}$ for all $i = 1,..,n$. A commonly used criterion in multiple testing is to control the false discovery rate (FDR) at a given level, say, FDR $\leq \alpha$. Equipped with consistent estimates $(\hat{\epsilon}_n, \hat{A}_n, \hat{\sigma})$, we can specialize the general adaptive testing procedure proposed in Sun and Cai (2007) to solve the signal identification problem in the current setting. Define

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

The adaptive procedure has three steps. First calculate the observed $\widehat{\text{Lfdr}}(X_i)$ for $i = 1,..,n$. Then rank $\widehat{\text{Lfdr}}(X_i)$ in an increasing order: $\widehat{\text{Lfdr}}_{(1)} \leq \widehat{\text{Lfdr}}_{(2)} \leq \widehat{\text{Lfdr}}_{(n)}$. Finally reject all $H_0^{(i)}, i = 1,\ldots,k$ where $k = \max\{i : \frac{1}{i}\sum_{j=1}^{i}\widehat{\text{Lfdr}}_{(j)} \leq \alpha\}$. 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 there are n subjects that are labeled 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) show 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. While the scope in these works are 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, the Higher Criticism, the Max, and the SM (which stands for Sample Mean; to be defined below). The LRT is defined in (2.10); the Higher Criticism is defined in (3.14) 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; the Max is defined in (3.21). In addition, denoting

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

let the 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 H_0). The 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 the Higher Criticism, the Max, and the SM do not need such information.

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

Experiment 1. In this experiment, we investigate how the LRT performs and how relevant the theoretic detection 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 the left panel of Figure 3. Recall that Theorem 2.1-2.2 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 $r \geq 0.5$ and is large when $r \leq 0.1$. In addition, if we view the sum of Type I and Type II errors as a function of r, then as n gets larger, the function gets increasingly close to the indicator function $1_{\{r<0.3\}}$. This is consistent with Theorems 2.1-2.2.

For the dense case, we fix $(\beta, \sigma^2) = (0.2, 1)$, $n \in \{10^4, 10^5, 10^7\}$, and let r range from $1/30$ to 0.5 with an increment of $1/30$. The results are displayed in the right panel of Figure 3, where a similar conclusion can be drawn.

Experiment 2. In this experiment, we compare the Higher Criticism with the LRT, the Max, and the 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 the left panel of Figure 4. The figure illustrates that the Higher Criticism has a similar performance to that of the LRT, and outperforms the Max. We also note that SM usually does not work in the sparse case, so we leave it out for comparison.

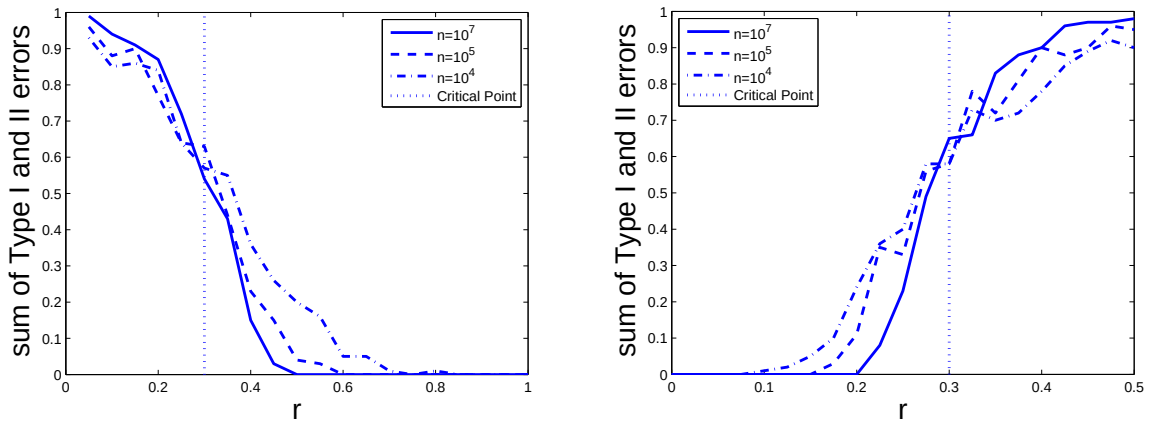

Figure 3: Sum of Type I and Type II errors of the LRT. Left: $(\beta, \sigma^2) = (0.7, 0.5)$, r ranges from 0.05 to 1 with an increment of 0.05, and $n = 10^4, 10^5, 10^7$ (dot-dashed, dashed, solid). Right: $(\beta, \sigma) = (0.2, 1)$, r ranges from 1/30 to 0.5 with an increment of 1/30, and $n = 10^4, 10^5, 10^7$ (dot-dashed, dashed, solid). In each panel, the vertical dot-dashed line illustrates the critical point of $r = \rho^*(\beta; \sigma)$. The results are based on 100 replications.

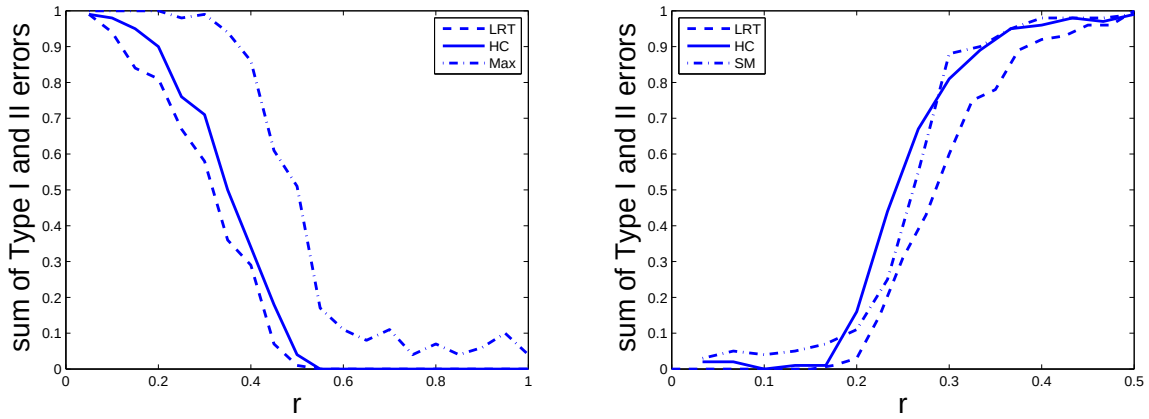

Figure 4: Sum of Type I and Type II errors of the Higher Criticism (solid), the LRT (dashed) and the Max (dot-dashed; left panel) or the SM (dot-dashed; right panel). Left: $(n, \beta, \sigma^2) = (10^6, 0.7, 0.5)$, and r ranges from 0.05 to 1 with an increment of 0.05. Right: $(n, \beta, \sigma^2) = (10^6, 0.2, 1)$, and r ranges from 1/30 to 0.5 with an increment of 1/30. The results are based on 100 replications.

We note that the LRT has optimal performance, but the implementation of which needs specific information of (β, r, σ). In contrast, the Higher Criticism is non-parametric and does not need such information. Nevertheless, Higher Criticism has comparable performance as 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, the Max usually does not work well, so we compare the Higher Criticism with the LRT and the SM only. The results are summarized in the right panel of Figure 4, where a similar conclusion can be drawn.

Experiment 3. In this experiment, we continue to compare the Higher Criticism with the LRT, the Max, and the 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 the left panel of Figure 5 (that of the SM is left out for it would not work well in the very sparse case), where the performance of each test gets 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, the Higher Criticism has a comparable performance to that of the LRT, and outperforms the Max except for large σ, where the Higher Criticism and Max perform comparatively.

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 the Higher Criticism with that of the LRT and the SM. The results are displayed in the right panel of Figure 5. It is noteworthy that the Higher Criticism and the LRT perform reasonably well when σ is bounded away from 1, and effectively fail when $\sigma = 1$. This is due to the fact that 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 strong enough in order for successful detection. Note that for the whole range of σ, the SM has poor performance.

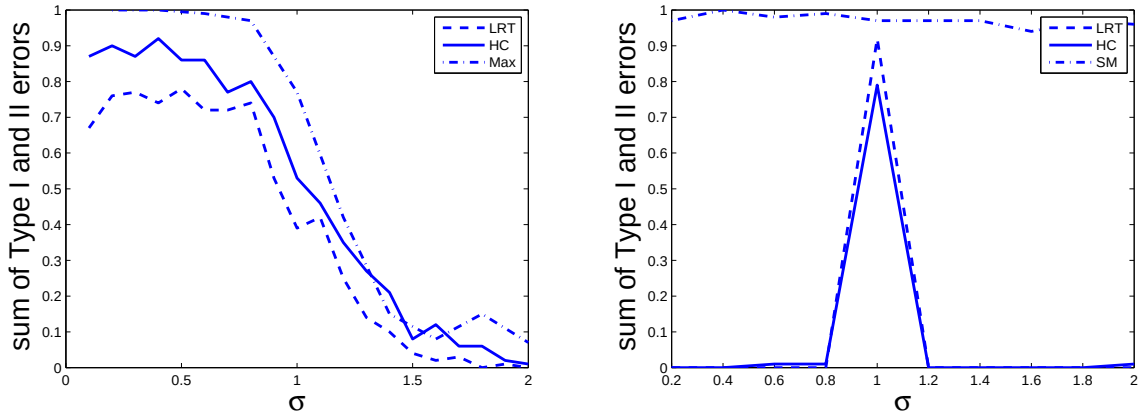

Figure 5: Sum of Type I and Type II errors of Higher Criticism (solid), the LRT (dashed) and the Max (dot-dashed; left panel) or the SM (dot-dashed; right panel). Left: $(n, \beta, r) = (10^6, 0.7, 0.25)$, and σ ranges from 0.2 to 2 with an increment of 0.2. Right: $(n, \beta, r) = (10^6, 0.2, 0.4)$, and σ ranges from 0.2 to 2 with an increment of 0.2. The visible spike is due to that, in the dense case, the detection problem is intrinsically different when $\sigma = 1$ and $\sigma \neq 1$. The results are based on 100 replications.

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

First, we investigate the case of $\beta > 1/2$. 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 the left panel of Figure 6. The figure illustrates that the detection problem becomes increasingly

more difficult when β increases and r is fixed. Nevertheless, the Higher Criticism has a comparable performance to that of the LRT and outperforms the Max.

Second, we investigate the case of $\beta < 1/2$. 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 to the previous case, a similar conclusion can be drawn if we replace the Max by the SM.

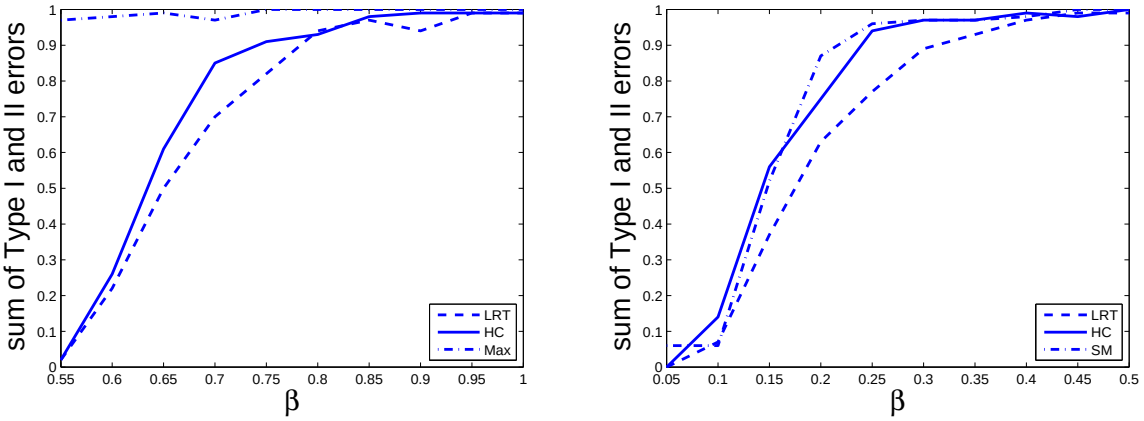

Figure 6: Sum of Type I and Type II errors of the Higher Criticism (solid), the LRT (dashed) and the Max (dot-dashed; left panel) or the SM (dot-dashed; right panel). Left: $(n, r, \sigma^2) = (10^6, 0.25, 0.5)$, and β ranges from 0.55 to 1 with an increment of 0.05. Right: $(n, r, \sigma^2) = (10^6, 0.3, 1)$, and β ranges from 0.05 to 0.5 with an increment of 0.05. The results are based on 100 replications.

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, then calculated the mean for each bootstrap sample. The estimated standard error is the standard deviation of the 500 bootstrap means. Due 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.

	Sparse			Dense		
σ	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)

Table 1: Means with their estimated standard errors in parentheses for different methods. Sparse: $(n, \beta, r) = (10^6, 0.7, 0.25)$. Dense: $(n, \beta, r) = (10^6, 0.2, 0.4)$.

In conclusion, the Higher Criticism has a comparable performance to that of the LRT. But unlike the LRT, the Higher Criticism is non-parametric. The Higher Criticism automatically adapts to different signal strengths, heteroscedasticity levels, and sparsity levels, and outperforms the Max and the 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 signal strengths may be unequal, the case where the noise maybe correlated or nonGaussian, and the case where the heteroscedasticity parameter σ has a more complicated source.

6.1 When the signal strength maybe unequal

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

$$H_1^{(n)}: \quad X_i \overset{iid}{\sim} (1-\epsilon_n)N(0,1) + \epsilon_n \int \frac{1}{\sigma}\phi(\frac{x-u}{\sigma})dG_n(u), \tag{6.24}$$

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

Interestingly, the Hellinger distance associated with testing problem is monotone with respect to G_n. In fact, fixing $n \geq 1$, if the support of G_n is contained in $[0, A_n]$, then the Hellinger distance between $N(0,1)$ and the density in (6.24) is no greater than that between $N(0,1)$ and $(1-\epsilon_n)N(0,1) + \epsilon_n N(A_n, \sigma^2)$. The proof is elementary so we omit it.

At the same time, similar monotonicity exists for the Higher Criticism. In detail, fixing n, we apply the Higher Criticism to n samples from $(1-\epsilon_n)N(0,1) + \epsilon_n \int \frac{1}{\sigma}\phi(\frac{x-u}{\sigma})dG_n(u)$, as well as to n samples from $(1-\epsilon_n)N(0,1) + \epsilon_n N(A_n, \sigma^2)$, 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 two random variables $X \leq Y$ stochastically if the cumulative distribution function of the former is no smaller than that of the latter point-wise). 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 $\epsilon_n = n^{-\beta}$ and $\tau_p = \sqrt{2r\log p}$. The following can be proved.

- Suppose $r < \rho^*(\beta; \sigma)$. Consider the problem of testing H_0 against $H_1^{(n)}$ as in (6.24). 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 $\longrightarrow 1$ as $n \longrightarrow \infty$).

- Suppose $r > \rho^*(\beta; \sigma)$. Consider the problem of testing H_0 against $H_1^{(n)}$ as in (6.24). 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 $\longrightarrow 0$ as $n \longrightarrow \infty$.

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 nonGaussian.

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 $supp(\mu)$ be the support of μ, and let $\Lambda = \Lambda(\mu)$ be an n by n diagonal matrix the

k-th coordinate of which is σ or 1 depending on $k \in supp(\mu)$ 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) and Hall and Jin (2010), where $\sigma = 1$ so that the model is homoscedastic in a sense. In these work, we found that the correlation structure among the noise is not necessarily a *curse* and could be a *blessing*. We showed that we could better the testing power of the 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 microarray, where X_i could be the Studentized t-scores of m different replicates for the i-th 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., 2010) where we learned that the 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 heteroscedastic 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 the support of $h(\sigma)$ is contained in an interval $[a,b]$, where $0 < a < b < \infty$. We consider a setting where under $H_1^{(n)}$, $X_i \overset{iid}{\sim} g(x)$, with

$$g(x) = g(x;\epsilon_n, A_n, h, a, b) = (1-\epsilon_n)\phi(x) + \epsilon_n \int_a^b \left[\frac{1}{\sigma}\phi(\frac{x-A_n}{\sigma})\right] h(\sigma)d\sigma. \tag{6.25}$$

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 $\sigma = \nu_b$, respectively. Note that $h(\sigma)$ is supported in $[a,b]$. In comparison, the detection problem in the current setting should be easier than the case of $\sigma = \nu_a$, and be harder than the case of $\sigma = \nu_b$. In other words, the "detection boundary" 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 nonzero at the point b, then there is a non-vanishing fraction of σ, say $\delta \in (0,1)$, that falls closely to b. Heuristically, the detection problem is at most as hard as the case where $g(x)$ in (6.25) is replaced by $\tilde{g}(x)$, where

$$\tilde{g}(x) = (1-\delta\epsilon_n)N(0,1) + \delta\epsilon_n N(A_n, b^2). \tag{6.26}$$

Since the constant δ only has a negligible effect on the testing problem, the detection boundary associated with (6.26) will the same as in the case of $\sigma = \nu_b$. For reasons of space, we omit the proof.

We briefly comment on using Higher Criticism for real data analysis. One interesting application of HC is for high dimensional feature selection and classification (see Section 4). In a related paper (Donoho and Jin, 2008), the method has been applied to several by now standard gene microarray data sets (Leukemia, Prostate cancer, and Colon cancer). The results reported are encouraging and the method is competitive to many widely used classifiers including the random forest and the Support Vector Machine (SVM). Another interesting application of the HC is for nonGaussian detection in the so-called WMAP data (stands for Wilkinson Microwave Anisotropy Probe) (Cayon et al., 2005). The method is competitive to the Kurtosis-based method, which is the most widely used one by cosmologists and astronomers. In these real data analysis, it is hard to tell whether the assumption of homoscedasticity is valid or not. However, the current paper suggests that the 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), which investigated Higher Criticism in the context of goodness-of-fit. It is also connected to Meinshausen and Buhlmann (2006) and Cai et al. (2007), which used Higher Criticism to motivate lower bounds for the proportion of non-null effects.

7 Proofs

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

7.1 Proof of Theorem 2.1

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 $(1-\epsilon_n)N(0,1)+\epsilon_n N(A_n,\sigma^2)$ 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

$$g_n(x) = g_n(x;r,\sigma) = \frac{\phi_\sigma(x-A_n)}{\phi(x)}. \tag{7.27}$$

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

Lemma 7.1 *Fix $\sigma > 1$, $\beta\in(1/2,1)$, and $r\in(0,\rho^*(\beta;\sigma))$. As n tends to ∞,*

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

We now proceed to show Theorem 2.1. First, since that $E\big[\sqrt{1-\epsilon_n+\epsilon_n g_n(X)\cdot 1_{\{D_n\}}}\big]\le E\big[\sqrt{1-\epsilon_n+\epsilon_n g_n(X)}\big]\le 1$, so all we need to show is

$$E\big[\sqrt{1-\epsilon_n+\epsilon_n g_n(X)\cdot 1_{\{D_n\}}}\big] = 1+o(1/n).$$

Now, note that for $x \geq -1$, $\left|\sqrt{1+x} - 1 - \frac{x}{2}\right| \leq Cx^2$. Applying this with $x = \epsilon_n(g_n(X) \cdot 1_{\{D_n\}} - 1)$ gives

$$E\sqrt{1 - \epsilon_n + \epsilon_n g_n(X) \cdot 1_{\{D_n\}}} = 1 - \frac{\epsilon_n}{2} E[g_n(X) \cdot 1_{\{D_n^c\}}] + err, \tag{7.28}$$

where, by Cauchy-Schwarz inequality,

$$|err| \leq C\epsilon_n^2 E[g_n(X) \cdot 1_{\{D_n\}} - 1]^2 \leq C\epsilon_n^2 \big(E[g_n^2(X) \cdot 1_{\{D_n\}}] + 1\big). \tag{7.29}$$

Recall $\epsilon_n^2 = n^{-2\beta} = o(1/n)$. Combining Lemma 7.1 with (7.28)-(7.29) gives the claim.

□

7.2 Proof of Theorem 2.2

Since the proofs are similar, we only show that under the null. By Chebyshev's inequality, to show that $-\log(LR_n) \longrightarrow \infty$ in probability, it is sufficient to show that as n tends to ∞,

$$-E[\log(LR_n)] \to \infty, \tag{7.30}$$

and

$$\frac{\mathrm{Var}[\log(LR_n)]}{(E[\log(LR_n)])^2} \longrightarrow 0. \tag{7.31}$$

Consider (7.30) first. Recalling that $g_n(x) = \phi_\sigma(x - A_n)/\phi(x)$, we introduce

$$LLR_n(X) = LLR_n(X; \epsilon_n, g_n) = \log(1 - \epsilon_n + \epsilon_n g_n(X)), \tag{7.32}$$

and

$$f_n(x) = f_n(x; \epsilon_n, g_n) = \log(1 + \epsilon_n g_n(x)) - \epsilon_n g_n(x). \tag{7.33}$$

By definitions and elementary calculus, $\log(LR_n) = \sum_{i=1}^n LLR_n(X_i)$, and $E[LLR_n(X)] = E[\log(1+\epsilon_n g_n(X)) - \epsilon_n g_n(X)] + O(\epsilon_n^2) = E[f_n(X)] + O(\epsilon_n^2)$. Recalling $\epsilon_n^2 = n^{-2\beta} = o(1/n)$,

$$E[\log(LR_n)] = nE[LLR_n(X)] = nE[f_n(X)] + o(1). \tag{7.34}$$

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

$$E[f_n(X)] \leq -C\bigg(\epsilon_n E[g_n(X) 1_{\{\epsilon_n g_n(X) > 1\}}] + \epsilon_n^2 E[g_n^2(X) 1_{\{\epsilon_n g_n(X) \leq 1\}}]\bigg). \tag{7.35}$$

The following lemma is proved in the appendix.

Lemma 7.2 *Fix $\sigma > 0$, $\beta \in (1/2, 1)$ and $r \in (0,1)$ such that $r > \rho^*(\beta; \sigma)$, then, as n tends to ∞, we have either*

$$n\epsilon_n E[g_n(X) 1_{\{\epsilon_n g_n(X) > 1\}}] \longrightarrow \infty \tag{7.36}$$

or

$$n\epsilon_n^2 E[g_n^2(X) 1_{\{\epsilon_n g_n(X) \leq 1\}}] \longrightarrow \infty. \tag{7.37}$$

Combine Lemma 7.2 with (7.34)-(7.35) gives the claim in (7.30).

Next, we show (7.31). Recalling $\log(LR_n) = \sum_{i=1}^n LLR_n(X_i)$, we have

$$\mathrm{Var}[\log(LR_n)] = n\mathrm{Var}(LLR_n(X)) = n\big(E[LLR_n^2] - (E[LLR_n])^2\big).$$

Comparing this with (7.31), it is sufficient to show that there is a constant $C > 0$ such that

$$E[LLR_n^2(X)] \le C\big|E[LLR_n(X)]\big|. \tag{7.38}$$

First, by Schwartz inequality, for all x,

$$\log^2(1-\epsilon_n+\epsilon_n g_n(x)) = \big[\log(1-\frac{\epsilon_n}{1+\epsilon_n g_n(x)})+\log(1+\epsilon_n g_n(x))\big]^2 \le C[\epsilon_n^2+\log^2(1+\epsilon_n g_n(x))].$$

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

$$E[LLR_n^2] \le CE[\log^2(1+\epsilon_n g_n(X))] + o(1/n).$$

Second, note that $\log(1+x) < C\sqrt{x}$ for $x > 1$ and $\log(1+x) < x$ for $x > 0$. By similar argument as in the proof of (7.35),

$$E[\log^2(1+\epsilon_n g_n(X))] \le C\bigg(\epsilon_n E[g_n(X)1_{\{\epsilon_n g_n(X)>1\}}] + \epsilon_n^2 E[g_n^2(X)1_{\{\epsilon_n g_n(X)\le 1\}}]\bigg).$$

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

$$E[LLR_n^2] \le C\bigg(\epsilon_n E[g_n(X)1_{\{\epsilon_n g_n(X)>1\}}] + \epsilon_n^2 E[g_n^2(X)1_{\{\epsilon_n g_n(X)\le 1\}}]\bigg).$$

Comparing this with (7.35) gives the claim. □

7.3 Proof of Theorem 2.3

By the similar argument as in Section 7.1, all we need to show is that when $\sigma = 1$ and $r > 1/2 - \beta$,

$$E[\sqrt{1-\epsilon_n+\epsilon_n g_n(X)}] = 1 + o(n^{-1}), \tag{7.39}$$

where $X \sim N(0,1)$, and $g_n(X)$ is as in (7.27). By Taylor expansion,

$$E[\sqrt{1-\epsilon_n+\epsilon_n g_n(X)}] \ge E[1 + \frac{\epsilon_n}{2}(g_n(X)-1) - \frac{\epsilon_n^2}{8}(g_n(X)-1)^2].$$

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

$$E[\sqrt{1-\epsilon_n+\epsilon_n g_n(X)}] \ge 1 - \frac{\epsilon_n^2}{8}(E[g_n^2(X)] - 1). \tag{7.40}$$

Write

$$E[g_n^2(X)] = \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{(\frac{1}{2}-\frac{1}{\sigma^2})x^2 + \frac{2A_n x}{\sigma^2} - \frac{A_n^2}{\sigma^2}} dx = \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{2-\sigma^2}{2\sigma^2}(x - \frac{2A_n}{2-\sigma^2})^2 + \frac{A_n^2}{2-\sigma^2}} dx.$$

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

$$\frac{\epsilon_n^2}{8}(E[g_n^2(X)] - 1) \sim \epsilon_n^2 A_n^2 = o(n^{-1}). \tag{7.41}$$

Inserting (7.40)-(7.41) into (7.39) gives the claim.

7.4 Proof of Theorem 2.4

Recall that $LLR_n(x) = \log(1+\epsilon_n(g_n(x)-1))$ and $\log(LR_n) = \sum_{j=1}^n LLR_n(X_j)$. By similar arguments as in Section 7.2, it is sufficient to show that for $X \sim N(0,1)$, when $n \longrightarrow \infty$,

$$nE[LLR_n(X)] \longrightarrow -\infty, \tag{7.42}$$

and

$$\frac{\mathrm{Var}[\log(LR_n)]}{(E[\log(LR_n)])^2} \longrightarrow 0. \tag{7.43}$$

Consider (7.42) first. Introduce the event $B_n = \{X : \epsilon_n g_n(X) \leq 1\}$. Note that $\log(1+x) \leq x$ for all x and $\log(1+x) \leq x - x^2/4$ when $x \leq 1$, and that $E[g_n(X)] = 1$. It follows that

$$E[LLR_n(X)] \leq E[\epsilon_n(g_n(X)-1)] - \frac{1}{4}E[\epsilon_n^2(g_n(X)-1)^2 \cdot 1_{B_n}] = -\frac{1}{4}\epsilon_n^2 E[(g_n(X)-1)^2 \cdot 1_{B_n}]. \tag{7.44}$$

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

$$E[(g_n(X)-1)^2 1_{B_n}] \geq E[g_n^2(X)1_{B_n}] - 2 + P(B_n) = E[g_n^2(X)1_{B_n}] - 1 - P(B_n^c). \tag{7.45}$$

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

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

$$P(B_n^c) = o(A_n^2), \qquad E[g_n^2(X)1_{B_n}] = \frac{e^{A_n^2}}{\sqrt{2\pi}} \int_{\{x:\epsilon_n g_n(x) \leq 1\}} e^{-(x-2A_n)^2/2} dx = 1 + A_n^2 \cdot (1+o(1)).$$

Combining this with (7.44)-(7.45), $E[LLR_n(X)] \lesssim -\frac{1}{4}\epsilon_n^2 A_n^2 = -\frac{1}{4}n^{-2(\beta+r)}$. The claim follows by the assumption $r < 1/2 - \beta$.

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

$$E[g_n^2(X)1_{B_n}] \sim \begin{cases} \frac{1}{\sigma\sqrt{2-\sigma^2}}, & \sigma < \sqrt{2}, \\ C\sqrt{\log(n)}, & \sigma = \sqrt{2}, \\ (C/\sqrt{\log n})n^{\beta(\sigma^2-2)/(\sigma^2-1)}, & \sigma > \sqrt{2}, \end{cases} \tag{7.46}$$

where we note that $\frac{1}{\sigma\sqrt{2-\sigma^2}} > 1$ when $\sigma < \sqrt{2}$. In fact, once this is shown, noting that $P(B_n^c) = o(1)$, it follows from (7.45) that there is a constant $c_0(\sigma) > 0$ such that for sufficiently large n, $E[(g_n(X)-1)^2 1_{B_n}] - 1 \geq 4c_0(\sigma)$. Combining this with (7.44), $E[LLR_n(X)] \leq -c_0(\sigma)\epsilon_n^2 = -c_0(\sigma)n^{-2\beta}$. The claim follows from the assumption $\beta < 1/2$.

We now show (7.46). Write

$$E[g_n^2(X)1_{B_n}] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\{x:\epsilon_n g_n(x) \leq 1\}} e^{(\frac{1}{2}-\frac{1}{\sigma^2})x^2 + \frac{2A_n x}{\sigma^2} - \frac{A_n^2}{\sigma^2}} dx. \tag{7.47}$$

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

$$E[g_n^2(X)1_{B_n}] \sim \frac{1}{\sqrt{2\pi\sigma^2}} \int e^{(\frac{1}{2}-\frac{1}{\sigma^2})x^2} dx = \frac{1}{\sigma\sqrt{2-\sigma^2}},$$

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

$$E[g_n^2(X) 1_{B_n}] = \frac{1}{\sqrt{2\pi}\sigma^2} \int_{x_-(n)}^{x_+(n)} e^{(\frac{1}{2} - \frac{1}{\sigma^2})x^2 + \frac{2A_n x}{\sigma^2} - \frac{A_n^2}{\sigma^2}} dx \sim \frac{1}{\sqrt{2\pi}\sigma^2} \int_{x_-(n)}^{x_+(n)} e^{(\frac{1}{2} - \frac{1}{\sigma^2})x^2} dx. \quad (7.48)$$

When $\sigma = \sqrt{2}$, $1/2 - 1/\sigma^2 = 0$. By (7.48), $E[g_n^2(X) 1_{B_n}] \sim (1/(\sqrt{2\pi}\sigma^2)) 2x_0(n) \sim \frac{2}{\sigma}\sqrt{\beta \log(n)/(\pi(\sigma^2 - 1))}$, which gives the claim. When $\sigma > \sqrt{2}$, $1/2 - 1/\sigma^2 > 0$. By (7.48) and elementary calculus,

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

and the claim follows.

We now show (7.43). By similar argument as in Section 7.2, it is sufficient to show that

$$E[LLR_n^2(X)] \leq C \big| E[LLR_n(X)] \big|. \quad (7.49)$$

Note that it is proved in (7.44) that

$$\big| E[LLR_n(X)] \big| \geq \frac{1}{4} E[\epsilon_n^2 (g_n(X) - 1)^2 \cdot 1_{B_n}]. \quad (7.50)$$

Recall that $LLR_n(x) = \log(1 + \epsilon_n(g_n(x) - 1))$. Since $\log^2(1 + a) \leq a$ for $a > 1$ and $|\log^2(1 + a)| \lesssim a^2$ for $-\epsilon_n \leq a \leq 1$,

$$E[LLR_n^2(X)] \lesssim E[\epsilon_n (g_n(X) - 1) \cdot 1_{B_n^c}] + E[\epsilon_n^2 (g_n(X) - 1)^2 \cdot 1_{B_n}]. \quad (7.51)$$

Compare (7.51) with (7.50). To show (7.49), it is sufficient to show that

$$E[\epsilon_n (g_n(X) - 1) \cdot 1_{B_n^c}] \leq C E[\epsilon_n^2 (g_n(X) - 1)^2 \cdot 1_{B_n}]. \quad (7.52)$$

Note that 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 $B_n = \{X : |X| \geq n^{\beta + r} \exp(A_n^2)\}$.

We now show (7.52) for the case $\sigma > 1$. By the proof of (7.42),

$$E[\epsilon_n^2 (g_n(X) - 1)^2 1_{B_n}] \geq \begin{cases} Cn^{-2\beta}, & 1 < \sigma < \sqrt{2}, \\ C\sqrt{\log(n)}\, n^{-2\beta}, & \sigma = \sqrt{2}, \\ (C/\sqrt{\log(n)})\, n^{-\beta\sigma^2/(\sigma^2 - 1)}, & \sigma > \sqrt{2}. \end{cases} \quad (7.53)$$

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

$$\epsilon_n E[g_n(X) \cdot 1_{B_n^c}] \sim 2\epsilon_n \int_{x_0(n)}^{\infty} \frac{1}{\sigma} \phi(\frac{x - A_n}{\sigma}) dx \leq \frac{C}{\sqrt{\log n}} n^{-\beta \frac{\sigma^2}{\sigma^2 - 1}}. \quad (7.54)$$

Note that $\sigma^2/(\sigma^2 - 1) \geq 2$ when $\sigma \leq \sqrt{2}$. Comparing (7.53) and (7.54) gives (7.52). □

7.5 Proof of Theorem 3.1

It is sufficient to show that as n tends to ∞,

$$P_{H_0}\left\{HC_n^* \geq \sqrt{2(1+\delta)\log\log n}\right\} \to 0, \tag{7.55}$$

and

$$P_{H_1^{(n)}}\left\{HC_n^* < \sqrt{2(1+\delta)\log\log n}\right\} \to 0. \tag{7.56}$$

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

$$\frac{HC_n^*}{\sqrt{2\log\log n}} \longrightarrow 1, \qquad \text{in probability.}$$

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

$$P_{H_1^{(n)}}\left\{|W_n(t_n^*(\sigma,\beta,r))| < \sqrt{2(1+\delta)\log\log n}\right\} \longrightarrow 0. \tag{7.57}$$

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

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

$$E[W_n(t)] = \frac{\sqrt{n}\epsilon_n[\bar{\Phi}(\frac{t-A_n}{\sigma} - \bar{\Phi}(t)]}{\sqrt{\bar{\Phi}(t)(1-\bar{\Phi}(t))}} \sim \sqrt{n}\epsilon_n[\bar{\Phi}(\frac{t-A_n}{\sigma}) - \bar{\Phi}(t)]/\sqrt{\bar{\Phi}(t)}, \tag{7.58}$$

and

$$\mathrm{Var}(W_n(t)) = \frac{\bar{F}(t)(1-\bar{F}(t))}{\bar{\Phi}(t)(1-\bar{\Phi}(t))} \sim \frac{\bar{F}(t)}{\bar{\Phi}(t)}. \tag{7.59}$$

By Mills' ratio (Wasserman, 2006),

$$\bar{\Phi}(\sqrt{2q\log n}) = PL(n)\cdot n^{-q}, \qquad \bar{\Phi}\left(\frac{\sqrt{2q\log n} - A_n}{\sigma}\right) = PL(n)\cdot n^{-(\sqrt{q}-\sqrt{r})^2/\sigma^2}. \tag{7.60}$$

Inserting (7.60) into (7.58) gives

$$\frac{\sqrt{n}\epsilon_n[\bar{\Phi}(\frac{t-A_n}{\sigma}) - \bar{\Phi}(t)]}{\sqrt{\bar{\Phi}(t)}} = \begin{cases} PL(n)\cdot n^{r/(2-\sigma^2)-(\beta-1/2)}, & \sigma < \sqrt{2}, r < (2-\sigma^2)^2/4, \\ PL(n)\cdot n^{1-\beta-(1-\sqrt{r})^2/\sigma^2}, & \text{otherwise.} \end{cases} \tag{7.61}$$

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

$$E[W_n(t)]/\sqrt{2(1+\delta)\log\log n} \longrightarrow \infty. \tag{7.62}$$

Combining (7.58) and (7.59), it follows from Chebyshev's inequality that

$$P_{H_1^{(n)}}\left\{|W_n(t_n^*(\sigma,\beta,r))| < \sqrt{2(1+\delta)\log\log n}\right\} \leq C\frac{\mathrm{Var}(W_n(t))}{(E[W_n(t)])^2} \leq C\frac{\bar{F}(t)}{n\epsilon_n^2[\bar{\Phi}(\frac{t-A_n}{\sigma}) - \bar{\Phi}(t)]^2}.$$

Applying (7.61), the above approximately equals to

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

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

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

$$E[W_n(1)] = \frac{\sqrt{n}\epsilon_n[\bar{\Phi}(\frac{1-A_n}{\sigma}) - \bar{\Phi}(1)]}{\sqrt{\bar{\Phi}(1)(1-\bar{\Phi}(1))}} \sim C\sqrt{n}\epsilon_n[\bar{\Phi}(\frac{1-A_n}{\sigma}) - \bar{\Phi}(1)],$$

and

$$\text{var}[W_n(1)] = \frac{\bar{F}(1)(1-\bar{F}(1))}{\bar{\Phi}(1)(1-\bar{\Phi}(1))} \sim \text{a constant.} \tag{7.63}$$

Furthermore,

$$\sqrt{n}\epsilon_n[\bar{\Phi}(\frac{1-A_n}{\sigma}) - \bar{\Phi}(1)] = -Cn^{\frac{1}{2}-\beta}[(\frac{1}{\sigma} - 1) - \frac{A_n}{\sigma}](1 + o(1)).$$

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

$$E[W_n(1)] \sim n^\gamma \tag{7.64}$$

for some $\gamma > 0$ and

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

On the other hand, when $\sigma < 1$,

$$E[W_n(1)] \sim -n^\gamma \tag{7.65}$$

for some $\gamma > 0$ and

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

Combining (7.63), (7.64), and (7.65), it follows from Chebyshev's inequality that

$$P_{H_1^{(n)}}\left\{|W_n(t_n^*(\sigma,\beta,r))| < \sqrt{2(1+\delta)\log\log n}\right\} \le C\frac{\text{Var}[W_n(1)]}{(E[W_n(1)])^2} \le Cn^{-2\gamma} \to 0.$$

□

8 Appendix

8.1 Proof of Theorem 2.5 and Theorem 2.6

We consider the case $\sigma \in (0, \sqrt{2})$ first. Since the proofs are similar, we only show that under the null. Recall that $\log(LR_n) = \sum_{j=1}^n LLR_n(X_j)$ (see Section 6.2). It is sufficient to show that

$$E[e^{itLLR_n(X)}] = \begin{cases} 1 + \big(-\frac{it+t^2}{2}\big)\frac{1}{\sigma\sqrt{2-\sigma^2}}\frac{1}{n}[1+o(1)], & \frac{1}{2} < \beta < 1 - \sigma^2/4, \\ 1 + \big(-\frac{it+t^2}{2}\big)\frac{1}{2\sigma\sqrt{2-\sigma^2}}\frac{1}{n}[1+o(1)], & \beta = 1 - \sigma^2/4, \\ 1 + \frac{1}{n}\psi_{\beta,\sigma}^0(t)[1+o(1)], & 1 - \sigma^2/4 < \beta < 1. \end{cases}$$

Note that $E[e^{itLLR_n(X)}] = e^{it\log(1-\epsilon_n)}E[e^{it\log(1+\epsilon_n g_n(X))}]+O(\epsilon_n^2)$, $e^{it\log(1-\epsilon_n)} = 1-it\epsilon_n+O(\epsilon_n^2)$, and $E[e^{it\log(1+\epsilon_n g_n(X))}] = 1 + it\epsilon_n + E[e^{it\log(1+\epsilon_n g_n(X))} - 1 - it\epsilon_n g_n(X)]$. Therefore,

$$E[e^{itLLR_n(X)}] = 1 + E[e^{it\log(1+\epsilon_n g_n(X))} - 1 - it\epsilon_n g_n(X)] + o(1/n). \tag{8.66}$$

We now analyze the limiting behavior of $E[e^{it\log(1-\epsilon_n+\epsilon_n\cdot g_n(X)} - 1 - i\epsilon_n t g_n(X)]$ for the case of $1 \le \sigma < \sqrt{2}$. The case $0 < \sigma < 1$ is similar to that of $1 \le \sigma < \sqrt{2}$, thus omitted.

In the case $1 \le \sigma < \sqrt{2}$, we discuss three sub-cases $\beta \le (1-\sigma^2/4)$, $\beta = (1-\sigma^2/4)$, and $\beta > (1-\sigma^2/4)$ separately.

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

$$r = (2-\sigma^2)\cdot(\beta - 1/2), \qquad \text{so} \qquad 0 < r < \frac{1}{4}(2-\sigma^2)^2. \tag{8.67}$$

Write

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

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

$$n^{1-(\beta+\frac{(1-\sqrt{r})^2}{\sigma^2})}. \tag{8.68}$$

Note that by (8.67), the exponent $1-(\beta+\frac{(1-\sqrt{r})^2}{\sigma^2}) < 0$. When $\sigma < 1$, the exponent is a concave function in x. We further consider two sub-sub-cases: $\sqrt{2\log n} \le A_n/(1-\sigma^2)$ and $\sqrt{2\log n} > A_n/(1-\sigma^2)$. For the first case, the maximum is reached at $x = \sqrt{2\log n}$ with the maximum value of (8.68), where the exponent < 0. For the second case, we have $\sqrt{r} < 1-\sigma^2$, and the maximum is reached at $x = A_n/(1-\sigma^2)$ with the maximum value of

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

Notice that, together, (8.67) and that $r < (1-\sigma^2)^2 < (1-\frac{\sigma^2}{2})(1-\sigma^2)$ imply that $\beta < 1-\sigma^2/2$. So, using (8.67) 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

$$\max_{\{|x|\le\sqrt{2\log n}\}}|\epsilon_n g_n(x)| = \exp\left(\max_{\{|x|\le\sqrt{2\log n}\}}\left\{(\frac{1}{2}-\frac{1}{2\sigma^2})x^2 + \frac{A_n x}{\sigma^2} - \frac{A_n^2}{2\sigma^2}\right\}\right) = o(1). \tag{8.69}$$

Now, introduce

$$f_n(x) = f(x;t,\beta,r) = e^{it\log(1+\epsilon_n\cdot g_n(X)} - 1 - it\epsilon_n g_n(x),$$

and the event $D_n = \{|X| \le \sqrt{2\log n}\}$. We have

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

On one hand, by (8.69) and Taylor expansion,

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

On the other hand,

$$|f_n(X)| \leq (1 + \epsilon_n g_n(X)).$$

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

$$E[\epsilon_n^2 g_n^2(X) \cdot 1_{\{D_n\}}] \sim \frac{1}{\sqrt{\sigma^2(2-\sigma^2)}} \cdot (1/n), \tag{8.70}$$

and that

$$E[(1 + \epsilon_n g_n(X)) \cdot 1_{\{D_n^c\}}] = o(1/n). \tag{8.71}$$

Consider (8.70) first. By similar argument as that in the proof of Lemma 7.1,

$$\epsilon_n^2 E[g_n^2(X) \cdot 1_{\{D_n\}}] = \frac{1}{\sqrt{2\pi}\sigma^2} n^{-2\beta+2r/(2-\sigma^2)} \int_{-\sqrt{2\log(n)}-A_n/(1-\sigma^2/2)}^{\sqrt{2\log(n)}-A_n/(1-\sigma^2/2)} e^{-(1/\sigma^2-1/2)y^2} dy. \tag{8.72}$$

Note that $\sqrt{2\log(n)} - A_n/(1-\sigma^2/2) = \sqrt{2\log n} \cdot (1 - \frac{2\sqrt{r}}{2-\sigma^2})$, where $(1 - \frac{2\sqrt{r}}{2-\sigma^2}) > 0$ as $r < \frac{1}{4}(2-\sigma^2)^2$. Therefore,

$$\int_{-\sqrt{2\log(n)}-A_n/(1-\sigma^2/2)}^{\sqrt{2\log(n)}-A_n/(1-\sigma^2/2)} e^{-(1/\sigma^2-1/2)y^2} dy \sim \sqrt{2\pi(\sigma^2/(2-\sigma^2))}.$$

Moreover, by (8.67), $2\beta - 2r/(2-\sigma^2) = 1$, so

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

and therefore,

$$E[f_n(X) \cdot 1_{\{D_n\}}] \sim (-t^2/2) \frac{1}{\sqrt{\sigma^2(2-\sigma^2)}} \cdot \frac{1}{n}, \tag{8.73}$$

which gives (8.70).

Consider (8.71). Recalling $g_n(x) = \phi_\sigma(x - A_n)/\phi(x)$,

$$E[(1 + \epsilon_n g_n(X)) \cdot 1_{\{D_n^c\}}] \leq \int_{|x|>\sqrt{2\log n}} [\phi(x) + \epsilon_n \phi_\sigma(x - A_n)] dx. \tag{8.74}$$

It is seen that

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

and that

$$\int_{|x|>\sqrt{2\log n}} \epsilon_n \phi_\sigma(x - A_n) dx = o(1) \cdot n^{-\beta} \cdot \phi((1-\sqrt{r})\sqrt{2\log n}) = o(n^{-\beta+\frac{(1-\sqrt{r})^2}{\sigma^2}}).$$

Moreover, by (8.67), $\beta + \frac{(1-\sqrt{r})^2}{\sigma^2} > 1$, so it follows (8.74) gives that

$$E[(1+\epsilon_n g_n(X)) \cdot 1_{\{D_n^c\}}] = o(1/n). \tag{8.75}$$

This gives (8.71) and concludes the claim in the case of $\beta < 1 - \sigma^2/4$.

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

$$\tilde{D}_n = \{|X| \le \sqrt{2\log n} - \frac{\log^{1/2}(\log(n))}{\sqrt{2\log n}}\}.$$

For reasons of space, we omit further discussion.

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

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

and

$$r = (1-\sigma\sqrt{1-\beta})^2, \qquad \text{so} \qquad \sqrt{r} > 1 - \sigma^2/2. \tag{8.76}$$

Equate $\epsilon_n \cdot \frac{\phi_\sigma(x-A_n)}{\phi_0(x)} = \frac{1}{\sigma}$. Direct calculations show that we have two solutions; using (8.76), it is seen that one of them $\sim \sqrt{2\log n}$ and we denote this solution by $x_0 = x_0(n) = \sqrt{2\log n} - \log(\log n)/\sqrt{2\log n}$. By the way ϵ_n is chosen, we have $\frac{1}{x_0}e^{-x_0^2/2} \sim 1/n$. Now, change variable with $x = x_0 + \frac{y}{x_0}$. It follows that

$$\epsilon_n g_n(x) = \frac{1}{\sigma} e^{(1-\frac{1-\sqrt{r}}{\sigma^2})y} e^{-\frac{y^2}{2x_0^2}(1/\sigma^2-1)}, \qquad \phi(x) = \frac{1}{\sqrt{2\pi}} x_0 \cdot (1/n) \cdot e^{-y} \cdot e^{-\frac{y^2}{2x_0^2}}.$$

Therefore,

$$E[f_n(X)] = \frac{1}{(\sqrt{2\pi})n} \int e^{it\log(1+\frac{1}{\sigma}e^{(1-\frac{1-\sqrt{r}}{\sigma^2})y} e^{-\frac{y^2}{2x_0^2}(1/\sigma^2-1)}) - 1 - \frac{it}{\sigma} e^{(1-\frac{1-\sqrt{r}}{\sigma^2})y} e^{-\frac{y^2}{2x_0^2}(1/\sigma^2-1)}] e^{-y} e^{-\frac{y^2}{2x_0^2}} dy.$$

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

$$h_n(y) = [e^{it\log(1+\frac{1}{\sigma}e^{(1-\frac{1-\sqrt{r}}{\sigma^2})y} e^{-\frac{y^2}{2x_0^2}(1/\sigma^2-1)}) - 1 - \frac{1}{\sigma} e^{(1-\frac{1-\sqrt{r}}{\sigma^2})y} e^{-\frac{y^2}{2x_0^2}(1/\sigma^2-1)}] e^{-y} e^{-\frac{y^2}{2x_0^2}}.$$

It is seen that point-wise, $h_n(u)$ converge to

$$h(y) = [e^{it\log(1+\frac{1}{\sigma}e^{(1-\frac{1-\sqrt{r}}{\sigma^2})y})} - 1 - \frac{1}{\sigma} e^{(1-\frac{1-\sqrt{r}}{\sigma^2})y}] e^{-y}.$$

At the same time, note that

$$|e^{it(1+e^y)} - 1 - ite^y| \le C \cdot \min\{e^y, e^{2y}\}.$$

It is seen that

$$|h_n(y)| \le Ce^{-y} \min\{e^{(1-\frac{1-\sqrt{r}}{\sigma^2})y}, e^{2(1-\frac{1-\sqrt{r}}{\sigma^2})y}\}.$$

The key fact here is that, by (8.76), $0 < \frac{1-\sqrt{r}}{\sigma^2} < 1/2$. Therefore,

$$e^{-y}\min\{e^{1-\frac{1-\sqrt{r}}{\sigma^2}}, e^{2(1-\frac{1-\sqrt{r}}{\sigma^2})y}\} = \begin{cases} e^{-\frac{1-\sqrt{r}}{\sigma^2}\cdot y}, & y \geq 0, \\ e^{(1-2\frac{1-\sqrt{r}}{\sigma^2})y}, & y < 0, \end{cases}$$

where the right hand side is integrable. It follows from the Dominated Convergence Theorem that

$$nE[f_n(X)] \longrightarrow (2\pi)^{-1/2}\int h(x)dx,$$

which proves the claim.

Consider the case $\sigma \geq \sqrt{2}$. The proof is similar to the case of $\sigma < \sqrt{2}$ and $\beta > (1-\sigma^2/4)$ so we omit it. This concludes the claim. □

8.2 Proof of Lemma 3.1

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

$$\bar{\Phi}(\sqrt{2q\log n}) = PL(n)\cdot n^{-q}, \qquad \bar{\Phi}\left(\frac{\sqrt{2q\log n} - A_n}{\sigma}\right) = PL(n)\cdot n^{-(\sqrt{q}-\sqrt{r})^2/\sigma^2}.$$

It follows that

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

where

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

It suffices to show that $\delta(q;\beta,r,\sigma)$ reaches its maximum at at $q = \min\{(\frac{2}{2-\sigma^2})^2 r, 1\}$ when $\sigma < \sqrt{2}$ and at $q = 1$ otherwise.

Towards this end, we note that, first, when $\sigma < \sqrt{2}$ 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 $\sigma < \sqrt{2}$ and $r \geq (2-\sigma^2)^2/4$ or $\sigma \geq \sqrt{2}$, $\delta(q;\beta,r,\sigma)$ is monotonically increasing. Combining these gives the claim.

8.3 Proof of Lemma 7.1

Consider the first claim. Direct calculations show that

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

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

$$C\epsilon_n\left(\phi(\frac{(1-\sqrt{r})}{\sigma}\sqrt{2\log n}) + \phi(\frac{(1+\sqrt{r})}{\sigma}\sqrt{2\log n})\right) = Cn^{-(\beta+\frac{(1-\sqrt{r})^2}{\sigma^2})}.$$

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

$$\beta + \frac{(1-\sqrt{r})^2}{\sigma^2} = 1 - [(1-\beta) - \frac{(1-\sqrt{r})^2}{\sigma^2}] > 1.$$

Consider the second claim. We discuss for the case $\sigma \geq \sqrt{2}$ and the case $\sigma < \sqrt{2}$ separately. When $\sigma \geq \sqrt{2}$, write

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

which is a convex function of x. Therefore, the extreme value over the range of $|x| \leq \sqrt{2\log n}$ assumes at the endpoints, which is seen to be

$$g_n^2(\sqrt{2\log n})\phi(\sqrt{2\log n}) = C \cdot n^{1-\frac{2}{\sigma^2}(1-\sqrt{r})^2}.$$

Therefore,

$$\epsilon_n^2 E[g_n^2(X) \cdot 1_{\{D_n\}}] \leq C \cdot \sqrt{\log n} \cdot n^{1-2(\beta+\frac{1}{\sigma^2}(1-\sqrt{r})^2)}.$$

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

When $\sigma < \sqrt{2}$, we similarly have

$$\epsilon_n^2 E[g_n^2(X) \cdot 1_{\{D_n\}}] \leq C\epsilon_n^2 \int_{x \leq \sqrt{2\log n}} e^{(\frac{1}{2}-\frac{1}{\sigma^2})x^2 + \frac{2A_n x}{\sigma^2} - \frac{A_n^2}{\sigma^2}} dx.$$

Write

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

By changing of variables,

$$\begin{aligned}\epsilon_n^2 E[g_n^2(X) \cdot 1_{\{D_n\}}] &\leq C n^{-2\beta+2r/(2-\sigma^2)} \int_{y \leq \sqrt{2\log n} - A_n/(1-\sigma^2/2)} e^{-(1/\sigma^2-1/2)y^2} dy \\ &= C n^{-2\beta+2r/(2-\sigma^2)} \Phi(\frac{\sqrt{2-\sigma^2}}{\sigma}(\sqrt{2\log n} - A_n/(1-\sigma^2/2)).\end{aligned}$$

Rewrite

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

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

$$\epsilon_n^2 E[g_n^2(X) \cdot 1_{\{D_n\}}] \leq C \begin{cases} n^{-2\beta+2r/(2-\sigma^2)}, & r \leq \frac{1}{4}(2-\sigma^2)^2, \\ n^{-2\beta+2r/(2-\sigma^2)-\frac{1}{\sigma^2}(2-\sigma^2)(1-\frac{2\sqrt{r}}{2-\sigma^2})^2}, & \text{otherwise.} \end{cases} \tag{8.77}$$

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

$$-2\beta + 2r/(2-\sigma^2) < -2\beta + 2(\beta - 1/2) = -1,$$

the claim follows directly from (8.77).

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

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

Applying (8.77) gives the claim. □

8.4 Proof of Lemma 7.2

Note that it is not necessary that (7.36) and (7.37) are simultaneously true. We prove the claim for three cases separately: (a) $1/2 < \beta < 1$ and $r > (1 - \sigma\sqrt{1-\beta})^2$ and $\sigma < \sqrt{2}$; or $1/2 < \beta < 1$ and $r > \rho^*(\beta;\sigma)$ and $\sigma \geq \sqrt{2}$, and (b) $1/2 < \beta < 1 - \sigma^2/4$ and $(2-\sigma^2)(\beta - 1/2) < r < (1 - \sigma\sqrt{1-\beta})^2$ and $1 < \sigma < \sqrt{2}$, and (c) $1/2 < \beta < 1 - \sigma^2/4$ and $(2-\sigma^2)(\beta-1/2) < r < (1-\sigma\sqrt{1-\beta})^2$ and $\sigma < 1$. The discussion for cases where (β, r, σ) fall right on the boundaries of the partition of these sub-regions is similar, so we omit it.

For (a), we show that (7.36) holds. For (β, r, σ) in this range, by elementary algebra and the definition of $\rho^*(\beta, \sigma)$,

$$1 - \beta - \frac{(1-\sqrt{r})^2}{\sigma^2} > 1. \tag{8.78}$$

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

$$n\epsilon_n E[g_n(X) 1_{\{\epsilon_n g_n(X) > 1\}}] \geq n\epsilon_n E[g_n(X) 1_{\{X \geq \sqrt{2\log n}\}}] = n\epsilon_n \int_{\sqrt{2\log n}}^{\infty} \frac{1}{\sigma} \phi(\frac{x - A_n}{\sigma}) dx.$$

By elementary calculus and Mills' ratio (Wasserman, 2006), the right hand side $= PL(n) n^{1-\beta-\frac{(1-\sqrt{r})^2}{\sigma^2}}$. The claim follows directly from (8.78).

For (b), we show (7.37) holds. It is seen that $\sup_{\{0 \leq x \leq \sqrt{2\log n}\}} \{\epsilon_n g_n(x)\} = o(1)$ for (β, r, σ) in this range, so

$$n\epsilon_n^2 E[g_n^2(X) 1_{\{\epsilon_n g_n(X) \leq 1\}}] \geq n\epsilon_n^2 E[g_n^2(X) 1_{\{0 \leq X \leq \sqrt{2\log n}\}}].$$

Direct calculations show that

$$n\epsilon_n^2 E[g_n^2(X) 1_{\{0 \leq X \leq \sqrt{2\log n}\}}] = n\epsilon_n^2 e^{\frac{A_n^2}{2-\sigma^2}} \Phi\left(\frac{\sqrt{2-\sigma^2}}{\sigma} (1 - \frac{\sqrt{r}}{1-\sigma^2/2}) \sqrt{2\log n} \right).$$

By basic algebra, for (β, r, σ) in the current range, $\frac{\sqrt{2-\sigma^2}}{\sigma}(1 - \frac{\sqrt{r}}{1-\sigma^2/2}) > 0$. Combining these gives

$$n\epsilon_n^2 E[g_n^2(X) 1_{\{\epsilon_n g_n(X) \leq 1\}}] \gtrsim n\epsilon_n^2 e^{\frac{A_n^2}{2-\sigma^2}} = n^{1-2\beta+\frac{2r}{2-\sigma^2}}.$$

The claim follows as $1 - 2\beta + \frac{2r}{2-\sigma^2} > 0$.

For (c), we consider two sub-cases separately: (c1) $1/2 < \beta < 1-\sigma^2/4$ and $r < (1-\sigma^2)\beta$ and $\sigma < 1$; or $1-\sigma^2 < \beta < 1-\sigma^2/4$ and $r \geq (1-\sigma^2)\beta$ and $\sigma < 1$, and (c2) $1/2 < \beta < 1-\sigma^2$ and $r \geq (1-\sigma^2)\beta$ and $\sigma < 1$. We show that (7.36) holds in cases (a) and (c2), whereas (7.37) holds in cases (b) and (c1).

For (c1), we show (7.37) holds. Similarly, for (β, r, σ) in this range, $\sup_{\{0 < x < \sqrt{2\log n}\}} \{\epsilon_n g_n(x)\} = o(1)$ and so

$$n\epsilon_n^2 E[g_n^2(X) 1_{\{\epsilon_n g_n(X) \leq 1\}}] \geq n\epsilon_n^2 E[g_n^2(X) 1_{\{0 < X \leq \sqrt{2\log n}\}}].$$

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

Consider (c2). Introduce

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

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

$$n\epsilon_n E[g_n(X) 1_{\{\epsilon_n g_n(X) > 1\}}] \geq n\epsilon_n \int_{\sqrt{2\Delta\log n}}^{\sqrt{2\Delta\log n}+\sqrt{\log\log n}} \frac{1}{\sigma}\phi(\frac{x - A_n}{\sigma})dx \gtrsim \frac{C}{\sqrt{\log n}} n^{1-\beta-\frac{(\sqrt{\Delta}-\sqrt{r})^2}{\sigma^2}},$$

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

$$1 - \beta - \frac{(\sqrt{\Delta} - \sqrt{r})^2}{\sigma^2} \geq 1 - \beta - \frac{(\sqrt{\Delta} - \sqrt{r})^2}{\sigma^2}\Big|_{\{r=(1-\sigma^2)\beta\}} = 1 - \frac{\beta}{1-\sigma^2},$$

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," *Theory 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," *Mon. Not. Roy. Astron. Soc*, 362, 826–832.

Delaigle, A., Hall, P., and Jin, J. (2010), "Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic," *J. Roy. Statist. Soc. B*, To Appear.

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

— (2008), "Higher Criticism thresholding: optimal feature selection when useful features are rare and week," *Proc. Natl. Acad. Sci.*, 105, 14790–14795.

— (2009), "Feature selection by Higher Criticism thresholding: optimal phase diagram," *Phil. Tran. Roy. Soc. A*, 367, 4449–4470.

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

— (2010), "Innovated Higher Criticism for detecting sparse signals in correlated noise," *Ann. Statist.*, 38(3), 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. Roy. Statist. Soc. B*, 70, 158–173.

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," *The Astronomical Journal*, 123(2), 1086–1094.

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

— (1999), "Minimax detection of a signal for l_n^p-balls," *Math. Methods 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," *Unpublished Manuscript*.

Jin, J. (2003), *Detecting and Estimating Sparse Mixtures*, Ph.D Thesis, Department of Statistics, Stanford University.

— (2004), "Detecting a target in very noisy data from multiple looks," *A festschrift for Herman Rubin, IMS Lecture Notes Monograph*, 45, Inst. Math. Statist, Beachwood, OH, 255–286.

— (2009), "Impossibility of successful classification when useful features are rare and weak," *Proc. Natl. Acad. Sci*, 106(22), 8856–8864.

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R., and Mostashari, F. (2005), "A space-time permutation scan statistic for disease outbreak detection," *PLoS Med*, 2(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*, SIAM, Philadelphia.

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

Wasserman, L. (2006), *All of Nonparametric Statistics*, Springer, NY.

Xie, J., Cai, T. T., and Li, H. (2010), "Sample size and power analysis for sparse signal recovery in Genome-Wide Association Studies," *Unpublished Manuscript*.

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