Electronic Journal of Statistics
Vol. 16 (2022) 4982–5018
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2053

Statistical limits of sparse mixture detection

Subhodh Kotekal

Department of Statistics,
University of Chicago,
5747 S. Ellis Avenue,
Chicago, IL 60637, USA
e-mail: skotekal@uchicago.edu

Abstract: We consider the problem of detecting a general sparse mixture and obtain an explicit characterization of the phase transition under some conditions, generalizing the univariate results of Cai and Wu. Additionally, we provide a sufficient condition for the adaptive optimality of a Higher Criticism type testing statistic formulated by Gao and Ma. In the course of establishing these results, we offer a unified perspective through the large deviations theory. The phase transition and adaptive optimality we establish are direct consequences of the large deviation principle of the normalized log-likelihood ratios between the null and the signal distributions.

MSC2020 subject classifications: Primary 62G10, 62C10; secondary 60F10.
Keywords and phrases: Hypothesis testing, sparse mixture detection, phase transition, higher criticism, large deviations.

Received February 2022.

1. Introduction

Modern technological advancements have ushered in a new scientific regime in which researchers simultaneously take measurements of a very large number of units with only a small fraction of units potentially exhibiting a signal. Typical examples include microarrays in genomics [13] and microwave probes in cosmology [39]; this new regime is ubiquitous in modern science [14]. Moreover, in many applications the signal is believed to be not only sparse but also sufficiently weak such that consistent identification of signal exhibiting units is impossible. In such a regime, two statistical problems immediately come to mind. First is the detection problem: for which sparsity levels can the presence of a signal be consistently detected? Second is the adaptation problem: does there exist a test which can detect a detectable signal without knowledge of the signal sparsity?

The detection problem is formally stated as a sparse mixture testing problem

$$H_0^{(n)} : X_1, \dots, X_n \overset{iid}{\sim} P_n, \tag{1}$$

$$H_1^{(n)} : X_1, \dots, X_n \overset{iid}{\sim} (1 - \varepsilon_n)P_n + \varepsilon_n Q_n \tag{2}$$

where $\varepsilon_n \in (0,1)$ and $\{P_n\}$ and $\{Q_n\}$ are collections of probability distributions on, say for convenience, a separable metric space $\mathcal{X}$. A consistent sequence of tests for testing (1)-(2) is a sequence of measurable functions $\varphi_n : \mathcal{X}^n \to \{0,1\}$ such that

$$\lim_{n\to\infty} P_{H_0^{(n)}}(\varphi_n(X_1,\dots,X_n) = 1) + P_{H_1^{(n)}}(\varphi_n(X_1,\dots,X_n) = 0) = 0.$$

To model the sparsity of the signal, the sparse mixture detection literature has adopted the calibration

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

with $0 < \beta < 1$. This calibration is used throughout, and henceforth we drop the subscript n from ε_n. The detection problem is to characterize, for fixed $\{P_n\}$ and $\{Q_n\}$, the values β such that there exists a consistent sequence of tests for the testing problem (1)-(2). For such β, the collection of mixtures $\{(1-n^{-\beta})P_n + n^{-\beta}Q_n\}$ is said to be detectable. Note that by the Neyman-Pearson lemma, the likelihood ratio test is consistent whenever the collection of mixtures $\{(1-n^{-\beta})P_n + n^{-\beta}Q_n\}$ is detectable. However, the likelihood ratio test is not a solution to the adaptation problem as it requires knowledge of β. The adaptation problem remains of practical interest.

Arguably the prototypical sparse mixture detection problem is the sparse normal mixture detection problem considered by Ingster [23] and Jin [24, 25]. Specifically, this is the testing problem (1)-(2) under calibration (3) with $P_n = N(0,1)$ and $Q_n = N(\mu_n, 1)$ where $\mu_n = \sqrt{2r\log n}$ and $r \in (0,1)$. Ingster [23] and Jin [24, 25] independently derived a subtle phase transition in this seemingly simple detection problem. A delicate asymptotic analysis showed that if $\beta < \beta^*(r)$, then there exists a consistent sequence of tests to test (1)-(2) where

$$\beta^*(r) = \begin{cases} \frac{1}{2} + r & \textit{if } r \le \frac{1}{4}, \\ 1 - (1-\sqrt{r})_+^2 & \textit{if } r > \frac{1}{4}. \end{cases}$$

Here, the notation $(x)_+ := \max\{x, 0\}$ for $x \in \mathbb{R}$ is used. Additionally, Ingster and Jin independently showed that if $\beta > \beta^*(r)$, then no sequence of tests is consistent for testing (1)-(2). The existence of a subtle phase transition in an apparently simple detection problem sparked subsequent research interest in sparse mixture detection. Phase transitions have been discovered in a variety of other sparse mixture detection problems beyond the sparse normal mixture setting [19, 2, 10, 5, 12, 18]. In other words, the literature has established in various settings the existence of some β^* which characterizes when it is possible to consistently test (1)-(2). We generically refer to β^* as a *detection boundary.* In investigating the asymptotic consequences of signal rarity and strength on various statistical tasks, a theoretical framework called the *Asymptotic Rare/Weak* (ARW) model has been introduced [27, 11]. The framework's introduction has been followed by an active research program arguably spearheaded by Jin and collaborators [11, 27, 20, 5, 28, 26, 29]. We refer the reader to the review articles [11, 27] for a detailed treatment.

In the context of sparse mixture detection, the phase transitions were initially obtained through delicate asymptotic analyses of the likelihood ratio test. Later, a unified approach to deriving phase transitions for general sparse mixtures on $\mathbb{R}$ was put forth by Cai and Wu [6]. Cai and Wu characterized the exact asymptotic order of the Hellinger distance between P_n and $(1-n^{-\beta})P_n + n^{-\beta}Q_n$ in terms of β, and it turns out the exact asymptotic order fundamentally determines the phase transition (under some regularity conditions). Many of the phase transitions in the literature follow directly from the results of Cai and Wu (see Section V of [6]). Ditzhaus [9] extended the results of Cai and Wu [6] to a larger class of univariate sparse mixtures beyond those satisfying the regularity conditions of [6].

In the adaptation problem, Donoho and Jin [10] delivered a key construction when investigating the sparse normal mixture detection problem. Donoho and Jin formulated a sequence of tests based on Tukey's Higher Criticism statistic that is consistent whenever $\beta < \beta^*(r)$ and adapts to not only the sparsity level β but also to the signal strength r. Specifically, Donoho and Jin considered the sequence of tests

$$\psi_{\mathrm{HC}_n} := \mathbf{1}_{\left\{\mathrm{HC}_n > \sqrt{2(1+\delta)\log\log n}\right\}} \tag{4}$$

where $\delta > 0$ is an arbitrary constant and the Higher Criticism statistic is defined as

$$\mathrm{HC}_n := \sup_{t \in \mathbb{R}} \frac{\left|\sum_{i=1}^n \mathbf{1}_{\{X_i \leq t\}} - n\Phi(t)\right|}{\sqrt{n\Phi(t)(1-\Phi(t))}}. \tag{5}$$

Here, Φ is the cumulative distribution function of the standard normal distribution. Calculating $p_i = 1 - \Phi(X_i)$, a change of variable yields a more evocative form

$$\mathrm{HC}_n = \sup_{u \in (0,1)} \frac{\left|\sum_{i=1}^n \mathbf{1}_{\{p_i \leq u\}} - nu\right|}{\sqrt{nu(1-u)}}.$$

With this formulation in mind, Higher Criticism is attractive in that it can be widely applied to sparse mixture detection beyond the initial sparse normal mixture setting. One need only craft p-values $\{p_i\}_{1 \leq i \leq n}$ from the observations $\{X_i\}_{1 \leq i \leq n}$ to use in HC_n. We refer the reader to the review articles [11, 27] for a detailed discussion of Higher Criticism and its applications beyond the sparse normal mixture detection problem originally in mind.

Remarkably, Higher Criticism achieves the detection boundary in many other sparse mixture detection problems beyond the sparse normal mixture setting. For example, the optimality of Higher Criticism for signal detection in the heteroscedastic sparse normal mixture with $P_n = N(0,1)$ and $Q_n = N(\sqrt{2r\log n}, \sigma^2)$ was established by Cai, Jeng, and Jin [5]. In the case of Gaussian null $P_n = N(0,1)$ and general Q_n (under some conditions), Cai and Wu proved that Higher Criticism is optimal. Later, Ditzhaus [9] proved that Higher Criticism is optimal for general distributions P_n and Q_n on $\mathbb{R}$ (again, under some conditions).

1.1. Our contributions

Our main contribution is a unified perspective through which we answer the detection problem and the adaptation problem in the general case where P_n and Q_n are probability distributions on an abstract space $\mathcal{X}$. Each is discussed in the context of the existing literature.

1.1.1. Detection problem

The existing literature has largely focused on the setting where the probability distributions P_n and Q_n are on $\mathbb{R}$. However, the data measured from each unit in many modern applications is typically multivariate or structured in some manner (e.g. graphs, partitions, and ranks). In particular, the sparse signal detection problem is of interest when P_n and Q_n are probability distributions on an abstract space $\mathcal{X}$. The existing literature does not offer a clean and unified solution in this general settings (such as that of [6] in the case $\mathcal{X} = \mathbb{R}$). Namely, a general framework is not available to derive the detection boundary β^*.

Of course, the Neyman-Pearson lemma asserts that the optimal testing procedure is the likelihood ratio test since (1)-(2) is a testing problem with simple null and alternative hypotheses. In principle, the detection boundary β^* can always be derived via a direct examination of the asymptotics of the likelihood ratio test. But even in the case $\mathcal{X} = \mathbb{R}$ the analysis is often difficult and is delicately tailored on a problem-by-problem basis. In fact, Cai and Wu [6] themselves note that existing analyses of the likelihood ratio test in the literature (such as [10, 23]) rely “on the normality assumption of the null distribution” in order to obtain “the limiting distribution of the log-likelihood ratio near the detection boundary” (page 2219 in [6]). The work of Cai and Wu [6] is groundbreaking precisely because it sidesteps such an analysis. Specifically, they obtain a unified solution in the case of $\mathcal{X} = \mathbb{R}$ by examining the Hellinger distance between P_n and $(1-\varepsilon)P_n + \varepsilon Q_n$ rather than examining the asymptotic distribution of the likelihood ratio. Thus, we aim to obtain a similar unified framework in the abstract $\mathcal{X}$ case which sidesteps the need for establishing the asymptotic distribution of the likelihood ratio.

One might argue that a new abstract framework is not needed to derive the detection boundary in the abstract $\mathcal{X}$ case. It might be suggested that the data be reduced to real-valued summary statistics, that is, to consider a function $T : \mathcal{X} \to \mathbb{R}$ and obtain the real-valued statistics $\{T(X_i)\}_{i=1}^n$. Letting P_n^T and Q_n^T denote the distribution of $T(X)$ when $X \sim P_n$ and $X \sim Q_n$ respectively, we then have the induced testing problem

$$\begin{aligned} H_0^{(n)} &: T(X_1), \dots, T(X_n) \overset{iid}{\sim} P_n^T, \\ H_1^{(n)} &: T(X_1), \dots, T(X_n) \overset{iid}{\sim} (1-\varepsilon)P_n^T + \varepsilon Q_n^T. \end{aligned}$$

Since the $\{T(X_i)\}_{i=1}^n$ are real-valued, one might then attempt to apply the result of Cai and Wu [6] to derive the detection boundary in this induced problem.

One might then hope this corresponds to the detection boundary for the original problem (1)-(2).

This line of thinking is attractive but it is very unclear how to choose T, especially in the our general setup with abstract $\mathcal{X}$. A poor choice of T will result in summary statistics $\{T(X_i)\}_{i=1}^n$ which lose too much information compared to the original data $\{X_i\}_{i=1}^n$. Then the detection boundary in the induced problem will not match the detection boundary in (1)-(2). Of course, in light of the Neyman-Pearson lemma, one can always choose the likelihood ratio $T(x) = \frac{dM_n}{dP_n}(x)$ where $M_n = (1-\varepsilon)P_n + \varepsilon Q_n$. Though this choice does not lose us information in the context of testing, an attempt to apply Theorem 3 of [6] requires us to understand the asymptotic behavior of the null distribution of the likelihood ratio $\frac{dM_n}{dP_n}$ (null distribution quantiles are needed to apply Theorem 3 of [6]). This is precisely the difficulty we are trying to avoid in the first place! There is no other obvious choice for T, and so there is a substantial gap between the result of Cai and Wu [6] and solution for the detection problem for abstract $\mathcal{X}$. Hence, we seek to generalize the framework of Cai and Wu [6] to the abstract setting.

As mentioned before, Cai and Wu [6] characterized the exact asymptotic order of the Hellinger distance between P_n and $(1-n^{-\beta})P_n + n^{-\beta}Q_n$ to derive the detection boundary β^*. Their results and proofs rely on examining the asymptotic behavior of the function $x \mapsto \frac{dQ_n}{dP_n}(x)$ evaluated at specific quantiles of P_n (e.g. Theorem 3 in [6]). The fact that P_n and Q_n are distributions on $\mathbb{R}$ is put to good use by virtue of examining quantiles. Since we are considering the abstract $\mathcal{X}$ case, the notion of quantiles is ill-defined and there is no immediately natural analogue. To address the abstract $\mathcal{X}$ case, our first contribution is a unified perspective through the theory of large deviations. The core idea of Cai and Wu [6] in characterizing the sharp Hellinger asymptotics is crucial to our analysis; the large deviations theory gives suitable tools to treat the general case with Cai and Wu's idea in hand.

1.1.2. *Adaptation problem*

Just as the detection problem is of interest when P_n and Q_n are probability distributions on an abstract space $\mathcal{X}$, so too is the adaptation problem. Given the success of Higher Criticism (5) in the case $\mathcal{X} = \mathbb{R}$, one would hope to modify Higher Criticism in some way to the abstract setting. The natural idea is to use a function $T : \mathcal{X} \to \mathbb{R}$ to obtain the real-valued statistics $\{T(X_i)\}_{i=1}^n$ and apply Higher Criticism.

While one cannot select T to be the the likelihood ratio $x \mapsto \frac{dM_n}{dP_n}(x)$, where $M_n = (1-\varepsilon)P_n + \varepsilon Q_n$, since it relies on knowledge of β, we are able to select $T(x) = \frac{dQ_n}{dP_n}(x)$. Applying Higher Criticism to both of these choices is exactly equivalent since the definition (5) involves a supremum over $t \in \mathbb{R}$. Though generically requiring knowledge of both P_n and Q_n, the statistic is adaptive to the signal sparsity.

This Higher Criticism type testing statistic was proposed by Gao and Ma (Section 3.2 of [18]). While the statistic's formulation is general, Gao and Ma studied the behavior of this statistic in a specific sparse mixture detection problem. It is not known in the literature whether this statistic is optimal in our abstract setting, that is to say, it is unknown whether it achieves the detection boundary. Our second main contribution is the formulation of a sufficient condition ensuring the optimality of Gao and Ma's statistic.

Organization

The remainder of the paper is organized as follows. Section 2 reviews the connection between Hellinger asymptotics and phase transitions established by Cai and Wu [6], states some requisite definitions from large deviations theory, and presents our first main result characterizing the phase transition in general sparse mixture testing problems (under some technical conditions). Section 3 reviews the Higher Criticism type testing statistic proposed by Gao and Ma (Section 3.2 of [18]) and presents a sufficient condition under which this statistic furnishes optimal tests that are adaptive to the signal sparsity. Section 4 illustrates our results through some examples; derivations for these results are found in the supplement [30]. These derivations in the supplement importantly showcase the typical methods of using the results presented here in the main text. Section 5 discusses our results and a few directions of further work. Proofs not presented in the main text are found in the supplement [30].

Notation

We use the following notation throughout the paper. For $a, b \in [-\infty, \infty]$, denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For sequences $\{a_n\}, \{b_n\} \subset [-\infty, \infty]$, denote $a_n = o(b_n)$ if $\frac{a_n}{b_n} \to 0$ as $n \to \infty$. Further, denote $a_n = \omega(b_n)$ if $b_n = o(a_n)$. For positive sequences $\{a_n\}, \{b_n\}$, denote $a_n \lesssim b_n$ if there exists a constant $C > 0$ not depending on n such that $a_n \leq Cb_n$. Denote $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, and denote $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For $a \in [-\infty, \infty]$, denote $(a)_+ = \max\{a, 0\}$. Denote $\mathbb{R}_+ = [0, \infty)$. For a probability measure P, let P^n denote the n-fold product measure of P. A probability measure Q on a measurable space is said to be absolutely continuous with respect to a probability measure P on the same measurable space if $P(A) = 0$ implies $Q(A) = 0$ for every measurable set A. We denote this as $Q \ll P$. The total variation distance between P and Q is given by $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$. The Hellinger distance between P and Q is given by $H(P, Q) = \left(\int \left(\sqrt{dP/d\nu} - \sqrt{dQ/d\nu}\right)^2 d\nu\right)^{1/2}$ where ν is a measure such that $P, Q \ll \nu$. We also use $P \otimes Q$ to denote the product measure with marginals P and Q respectively.

2. A large deviations perspective on detection limits

To determine the fundamental detection limits, we follow the approach laid out by Cai and Wu [6] in characterizing the Hellinger asymptotics. Cai and Wu [6] only obtain results in the univariate case, namely where $\{P_n\}$ and $\{Q_n\}$ are probability distributions on $\mathbb{R}$. We generalize their results and offer a unified perspective through the theory of large deviations.

2.1. Preliminaries

Without loss of generality, we will take $Q_n \ll P_n$ for all $n \geq 1$. No generality is lost as argued in Section III.C in [6]. Assume further that $\{P_n\}$ and $\{Q_n\}$ are dominated by a common measure on $\mathcal{X}$ and so admit densities $\{p_n\}$ and $\{q_n\}$. We bundle these assumptions together as Assumption 2.1, which will be in force for the remainder of the paper.

Assumption 2.1. *The probability distributions $\{P_n\}$ and $\{Q_n\}$ are dominated by a common measure on a separable metric space $\mathcal{X}$ and admit densities $\{p_n\}$ and $\{q_n\}$. Furthermore, $Q_n \ll P_n$ for all $n \geq 1$.*

The following definition introduces precise quantities $\underline{\beta}^*$ and $\overline{\beta^*}$ which enable a precise statement of the detection problem. Specifically, the detection problem, for fixed $\{P_n\}$ and $\{Q_n\}$, is the problem of explicitly characterizing $\underline{\beta}^*$ and $\overline{\beta}^*$.

Definition 2.1. Consider the testing problem (1)-(2) with calibration (3). Define

$$\begin{aligned}
\overline{\beta}^* &:= \inf\left\{\beta \geq 0 : \lim_{n\to\infty} \mathrm{TV}(P_n^n, ((1-n^{-\beta})P_n + n^{-\beta}Q_n)^n) = 0\right\}, \\
\underline{\beta}^* &:= \sup\left\{\beta \geq 0 : \lim_{n\to\infty} \mathrm{TV}(P_n^n, ((1-n^{-\beta})P_n + n^{-\beta}Q_n)^n) = 1\right\}
\end{aligned}$$

where TV denotes the total variation distance.

By the Neyman-Pearson lemma, $\overline{\beta}^*$ is the smallest number such that if $\beta > \overline{\beta}^*$, then every sequence of tests for testing (1)-(2) has a sum of Type I and Type II errors converging to one. Likewise, $\underline{\beta}^*$ is the largest number such that if $\beta < \underline{\beta}^*$, then there exists a sequence of tests for testing (1)-(2) with vanishing sum of Type I and Type II errors. The quantities $\overline{\beta}^*$ and $\underline{\beta}^*$ can be equivalently characterized by the asymptotics of the Hellinger distance. The analysis is much more amenable due to the tensorization of the Hellinger distance, and the following lemma gives a clean characterization [6].

Lemma 2.1 (Equations (25) and (26) - [6]). *Consider the testing problem (1)-(2) with calibration (3). Let $H_n^2(\beta) := H^2(P_n, (1-n^{-\beta})P_n + n^{-\beta}Q_n)$ where H is the Hellinger distance. Then*

$$\begin{aligned}
\overline{\beta}^* &= \inf\{\beta \geq 0 : H_n^2(\beta) = o(n^{-1})\}, \\
\underline{\beta}^* &= \sup\{\beta \geq 0 : H_n^2(\beta) = \omega(n^{-1})\}.
\end{aligned}$$

Lemma 1 of [6] establishes the result $0 \leq \underline{\beta}^* \leq \overline{\beta}^* \leq 1$, confirming the intuition that $\underline{\beta}^* \leq \overline{\beta}^*$ and additionally establishing that any phase transition must occur in $[0, 1]$.

2.2. Main results

We present our main results in this section. First, we briefly state a few definitions which are special cases of definitions formulated in the general large deviations theory [8].

To motivate the forthcoming definitions, we sketch our approach to the detection problem. In the univariate case ($\mathcal{X} = \mathbb{R}$), Cai and Wu [6] examine $\frac{q_n}{p_n}(x)$ at various quantiles of the null distribution P_n. Roughly speaking, the intuition is that the value of $\frac{q_n}{p_n}(x)$ quantifies how much evidence observing the realization $X = x$ gives us towards determining whether X comes from the null distribution P_n or the signal distribution Q_n. To elaborate, for each $s > 0$ let x_s denote the n^{-s} quantile of P_n. As described in [6], in many situations a function $t : \mathbb{R}_+ \to \mathbb{R}$ such that $\frac{q_n}{p_n}(x_s) = n^{t(s)(1+o(1))}$ can be found. The value $t(s)$ thus indicates how much evidence observing $X = x_s$ provides for favoring Q_n over P_n. Of course, the right tail of P_n (that is, $1-n^{-s}$ quantiles) is also examined [6]. The function t turns out to be fundamental in that Cai and Wu [6] are able to establish an explicit formula for the detection boundary β^* in terms of t.

Their intuition is attractive, namely to quantify the evidence for which observing the quantile x_s gives for favoring Q_n over P_n. The issue is that the notion of quantiles is not sensible for an abstract $\mathcal{X}$. Our approach is to invert the idea. We seek a function $s : \mathbb{R} \to [0,\infty]$ such that for an "evidence" level $t \in \mathbb{R}$, the probability a draw $X \sim P_n$ from the null yields evidence level t is approximately $n^{-s(t)}$. Roughly speaking, we wish to be able to say something to the effect of $P\left\{\frac{q_n}{p_n}(X) \approx n^t\right\} \approx n^{-s(t)}$ where $X \sim P_n$. If such a statement could be made, then perhaps an explicit formula for the detection boundary β^* could be derived in terms of the function s.

The large deviations theory enables us to make this intuition rigorous. The first definition is that of a rate function, which essentially plays the role of the function s described in our sketch.

Definition 2.2. Let $\mathcal{Y}$ be a separable metric space. A *rate function* $I : \mathcal{Y} \to [0,\infty]$ is a lower semicontinuous function. A rate function I is *good* if the sublevel sets $\{y \in \mathcal{Y} : I(y) \leq \alpha\}$ are compact for all $\alpha \geq 0$.

We have stated the general, standard definition of the rate function. For our main purpose, we will take $\mathcal{Y} = \mathbb{R}$ since an evidence level t is a real number. However, there will be occasions to make other choices of $\mathcal{Y}$, mainly in intermediate steps of proofs in the study of various applications. The next definition enables us to make precise probability statements of the kind we desire. Again, it is stated in its general, standard form.

Definition 2.3. Let $\{\mu_n\}$ be a family of probability measures on $(\mathcal{Y}, \mathcal{B})$ where $\mathcal{Y}$ is a separable metric space and $\mathcal{B}$ is the completed Borel field on $\mathcal{Y}$. We say that $\{\mu_n\}$ satisfies the *large deviation principle* with speed $\{a_n\}$ and rate function I if for all $\Gamma \in \mathcal{B}$,

$$-\inf_{y \in \Gamma^\circ} I(y) \leq \liminf_{n\to\infty} a_n \log \mu_n(\Gamma) \leq \limsup_{n\to\infty} a_n \log \mu_n(\Gamma) \leq -\inf_{y \in \overline{\Gamma}} I(y).$$

Here, $\{a_n\}$ is a sequence of reals with $a_n \to 0$. Additionally, Γ° and $\overline{\Gamma}$ denote the interior and closure of Γ respectively.

We specialize the definition of the large deviation principle further to a form most frequently used in our arguments. Note that the following definition requires that I be a good rate function, whereas the general definition of the large deviation principle does not have such a requirement.

Definition 2.4. Suppose $\{P_n\}$ and $\{Q_n\}$ satisfy Assumption 2.1. We say that the sequence of (normalized) log-likelihood ratios $\left\{\frac{\log \frac{q_n}{p_n}}{\log n}\right\}$ satisfies the *large deviation principle under the null* if there exists a good rate function $I : \mathbb{R} \to [0, \infty]$ and for all Borel sets $\Gamma \subset \mathbb{R}$ we have

$$\begin{aligned} -\inf_{t \in \Gamma^\circ} I(t) &\leq \liminf_{n\to\infty} \frac{1}{\log n} \cdot \log P\left(\frac{\log \frac{q_n}{p_n}(X_n)}{\log n} \in \Gamma\right) \\ &\leq \limsup_{n\to\infty} \frac{1}{\log n} \cdot \log P\left(\frac{\log \frac{q_n}{p_n}(X_n)}{\log n} \in \Gamma\right) \leq -\inf_{t\in\overline{\Gamma}} I(t) \end{aligned}$$

where $X_n \sim P_n$.

With these definitions in place, we have a framework in place such that the rough statement $P\{\frac{q_n}{p_n}(X_n) \approx n^t\} \approx n^{-I(t)}$ can be made precise and manipulated rigorously. Beyond just this technical utility, there is also some intuition. The rate function I quantifies the asymptotic order by which Q_n deviates from P_n. For example, if $t \in \mathbb{R}$ is an evidence level such that $I(t)$ is large, then it is relatively unlikely to observe $\frac{q_n}{p_n}(X_n) \approx n^t$ when the observation X_n is truly from the null, i.e. $X_n \sim P_n$. In other words, observing $\frac{q_n}{p_n}(X_n) \approx n^t$ constitutes evidence against the null. In our main result, the rate function I is the fundamental object determining the phase transition.

Theorem 2.1. *Suppose $\{P_n\}$ and $\{Q_n\}$ are probability distributions that satisfy Assumption 2.1 for the testing problem (1)-(2) with calibration (3). Suppose there exists some $\gamma > 1$ such that the tail condition*

$$\limsup_{n\to\infty} \frac{1}{\log n} \cdot \log E\left[\left(\frac{q_n}{p_n}(X_n)\right)^\gamma\right] < \infty \tag{6}$$

holds, where $X_n \sim P_n$. Suppose further that $\left\{\frac{\log \frac{q_n}{p_n}}{\log n}\right\}$ satisfies the large deviation principle under the null. Let $I : \mathbb{R} \to [0, \infty]$ be the associated good rate function. Then

$$\overline{\beta}^* \leq \frac{1}{2} + \left(\sup_{t\geq 0}\left\{t - I(t) + \frac{1 \wedge I(t)}{2}\right\}\right)_+.$$

and

$$\underline{\beta}^* \geq \frac{1}{2} + \sup_{t>0}\left\{t - I(t) + \frac{1 \wedge I(t)}{2}\right\}.$$

The most interesting situation is when the upper and lower bounds meet, yielding a detection boundary.

Corollary 2.1. *Consider the setting of Theorem 2.1. If the conditions of Theorem 2.1 hold and*

$$\left(\sup_{t\geq 0}\left\{t - I(t) + \frac{1\wedge I(t)}{2}\right\}\right)_+ = \sup_{t>0}\left\{t - I(t) + \frac{1\wedge I(t)}{2}\right\},$$

then $\underline{\beta}^* = \overline{\beta}^* = \beta^*$ *where*

$$\beta^* := \frac{1}{2} + \left(\sup_{t\geq 0}\left\{t - I(t) + \frac{1\wedge I(t)}{2}\right\}\right)_+. \tag{7}$$

Corollary 2.1 is our main result concerning phase transitions in the general sparse mixture detection problem. As mentioned before, the rate function I is fundamental in that it fully determines the detection boundary (provided the conditions of Corollary 2.1 hold).

2.3. Comparison to Cai and Wu

It is worth pausing to compare the role of the rate function with an analogous function in the framework of Cai and Wu [6]. We state the main theorem regarding phase transitions from [6] below with notational modifications to fit our context. Recall that Cai and Wu work in the $\mathcal{X} = \mathbb{R}$ setting.

Theorem 2.2 (Theorem 3 - [6])**.** *Consider the testing problem (1)-(2) with calibration (3). Suppose Assumption 2.1 holds. Let F_n and z_n denote the cumulative distribution function and quantile function of P_n respectively, i.e. $z_n(p) = \inf\{y \in \mathbb{R} : F_n(y) \geq p\}$ for $p \in [0,1]$. Assume that the log-likelihood ratio $\ell_n := \log \frac{q_n}{p_n}$ satisfies*

$$\lim_{n\to\infty} \sup_{s\geq(\log_2 n)^{-1}} \left|\frac{\ell_n(z_n(n^{-s})) \vee \ell_n(z_n(1-n^{-s}))}{\log n} - \gamma(s)\right| = 0$$

for some measurable function $\gamma : \mathbb{R}_+ \to \mathbb{R}$. If $\gamma > 0$ on a set of positive Lebesgue measure, then

$$\beta^* = \frac{1}{2} + \left(\operatorname*{ess\,sup}_{s\geq 0}\left\{\gamma(s) - s + \frac{s\wedge 1}{2}\right\}\right)_+.$$

Here, ess sup *denotes the essential supremum with respect to Lebesgue measure on $\mathbb{R}$.*

In Theorem 3 of [6], the fundamental object determining the phase transition is the function $\gamma : \mathbb{R}_+ \to \mathbb{R}$, which is determined by the asymptotics of the log likelihood ratio $\frac{\log \frac{q_n}{p_n}}{\log n}$ evaluated at the n^{-s} and $1 - n^{-s}$ quantiles of P_n

across $s \geq (\log_2 n)^{-1}$. Intuitively, the function γ quantifies the order at which Q_n deviates from P_n. Roughly speaking, the points s at which $\gamma(s) > 0$ are those P_n-quantiles which are "relatively more likely" under Q_n compared to P_n; roughly, γ quantifies the order of "more likely".

As discussed earlier, it is clear that the univariate nature of P_n and Q_n is heavily exploited as it is the likelihood ratio's asymptotic behavior at *quantiles* of P_n that determines the fundamental object γ. In the abstract setting, it is the rate function of a large deviation principle that precisely quantifies the asymptotic order of the likelihood ratio at various regions in the sample space. The rate function gives us a way to measure how much Q_n "deviates" from P_n asymptotically, thus allowing us to lift the core idea of Cai and Wu (namely the idea to sharply characterize Hellinger asymptotics) to the abstract setting. We investigate the relationship between the rate function I and the function γ below in Proposition 2.1. In fact, we can show that our condition that the normalized log-likelihood ratios satisfy the large deviation principle under the null (Definition 2.4) is implied, under some constraints, by the condition formulated in Theorem 3 of [6].

Proposition 2.1. *Suppose $\{P_n\}$ and $\{Q_n\}$ are probability distributions on $\mathbb{R}$ satisfying Assumption 2.1 for the testing problem (1)-(2). Let $z^{(n)}$ denote the quantile function of P_n. Assume there exist measurable functions $\alpha_0 : \mathbb{R}_+ \to \mathbb{R}$ and $\alpha_1 : \mathbb{R}_+ \to \mathbb{R}$ such that*

$$\lim_{n\to\infty} \sup_{s \geq (\log_2 n)^{-1}} \left| \frac{\log \frac{q_n}{p_n}(z^{(n)}(n^{-s}))}{\log n} - \alpha_0(s) \right| = 0, \tag{8}$$

$$\lim_{n\to\infty} \sup_{s \geq (\log_2 n)^{-1}} \left| \frac{\log \frac{q_n}{p_n}(z^{(n)}(1-n^{-s}))}{\log n} - \alpha_1(s) \right| = 0. \tag{9}$$

If α_0 and α_1 are continuous, then $\left\{ \frac{\log \frac{q_n}{p_n}}{\log n} \right\}$ satisfies the large deviation principle under the null with good rate function

$$I(t) = I_0(t) \wedge I_1(t)$$

where I_0 and I_1 are good rate functions given by

$$I_0(t) = \inf\{s \geq 0 : t = \alpha_0(s)\},$$
$$I_1(t) = \inf\{s \geq 0 : t = \alpha_1(s)\}.$$

We use the convention that $\inf \emptyset = \infty$*.*

Note that if the conditions of Corollary 2.1 also hold, then the detection boundary is

$$\beta^* = \frac{1}{2} + \left(\sup_{t \geq 0} \left\{ t - I(t) + \frac{1 \wedge I(t)}{2} \right\} \right)_+$$

$$
\begin{aligned}
&= \frac{1}{2} + \left(\sup_{t \geq 0} \sup_{s : s = I(t)} \left\{ t - s + \frac{1 \wedge s}{2} \right\} \right)_+ \\
&= \frac{1}{2} + \left(\sup_{s \in I(\mathbb{R})} \sup_{t : I(t) = s} \left\{ t - s + \frac{1 \wedge s}{2} \right\} \right)_+ \\
&= \frac{1}{2} + \left(\sup_{s \in I(\mathbb{R})} \left\{ \alpha_0(s) \vee \alpha_1(s) - s + \frac{1 \wedge s}{2} \right\} \right)_+ \\
&= \frac{1}{2} + \left(\sup_{s \geq 0} \left\{ \alpha_0(s) \vee \alpha_1(s) - s + \frac{1 \wedge s}{2} \right\} \right)_+
\end{aligned}
$$

where the final equality follows from the observation that $-s + \frac{1 \wedge s}{2} \leq 0$ for all $s \geq 0$. Observe that this is the same formula that appears in Theorem 3 of [6] with $\gamma(s) = \alpha_0(s) \vee \alpha_1(s)$. Note that essential supremum and supremum coincide as we assumed α_0 and α_1 are continuous. However, note that the conditions of Theorem 3 in [6] are not exactly the same as the conditions in Proposition 2.1, and so Proposition 2.1 does not fully import the results of Theorem 3 in [6].

3. A Higher Criticism type statistic

Beyond the detection problem, the adaptation problem is of practical interest. Namely, it is of interest to furnish an optimal sequence of tests that adapts to the unknown signal sparsity β. The existing literature of sparse mixture detection has focused on the setting where P_n and Q_n are distributions on $\mathbb{R}$, and a useful idea in this setting is to compare the empirical distribution of the data to the null distribution, usually after some transformation. Typically, this transformation takes the form of calculating a p-value p_i from each univariate data point X_i. In the abstract setting, this idea can be mimicked by calculating a p-value p_i from some univariate statistic of the abstract observation X_i. As mentioned in the introduction, Higher Criticism can be interpreted as evaluating the goodness-of-fit between the empirical p-value distribution and the null distribution. Recall that the associated test is given by $\psi_{\mathrm{HC}_n} = \mathbf{1}_{\left\{ \mathrm{HC}_n > \sqrt{2(1+\delta) \log \log n} \right\}}$ where $\delta > 0$ is an arbitrary constant and the Higher Criticism statistic is

$$
\mathrm{HC}_n = \sup_{u \in (0,1)} \frac{\left| \sum_{i=1}^n \mathbf{1}_{\{p_i \leq u\}} - nu \right|}{\sqrt{nu(1-u)}}.
$$

Ditzhaus [9] established that Higher Criticism is adaptively optimal (meaning it achieves the detection boundary without needing knowledge of β) in a wide class of univariate sparse mixture detection problems beyond the original sparse normal mixture problem considered by Donoho and Jin [10].

To apply Higher Criticism to a sparse mixture detection problem with abstract $\mathcal{X}$, a univariate p-value p_i from a univariate summary statistic of $X_i \in \mathcal{X}$ must be constructed. Gao and Ma, in Section 3.2 of [18], propose to apply

Higher Criticism to the univariate statistics $\left\{\frac{q_n}{p_n}(X_i)\right\}_{i=1}^{n}$. In this section, we first review Gao and Ma's construction and then give a sufficient condition for optimality.

3.1. Gao and Ma's construction

For notational convenience, we will refer to the null and alternative hypotheses in (1)-(2) as H_0 and H_1, suppressing the dependence on n as the context makes it clear. Gao and Ma formulate a Higher Criticism type testing statistic for the testing problem (1)-(2) when the distributions $\{P_n\}$ and $\{Q_n\}$ are known but β is unknown. We reproduce their formulation here (see Sections 3.2 and A.5 of [18]). For a collection of events $\mathscr{A}$, define

$$\mathrm{HC}_n(\mathscr{A}) := \sup_{A \in \mathscr{A}} |T_n(A)|$$

where

$$T_n(A) := \frac{\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A\}} - nP_n(A)}{\sqrt{nP_n(A)(1-P_n(A))}}.$$

As Gao and Ma note, the supremum in the definition of $\mathrm{HC}_n(\mathscr{A})$ is largely for the sake of adapting to the unknown sparsity parameter β. Focusing attention on T_n, fix $A \in \mathscr{A}$ and consider the test

$$\varphi_A := \mathbf{1}_{\{|T_n(A)| > c_n\}}$$

where c_n is a positive diverging sequence to be specified. Observing that $E_{H_0}(T_n(A)) = 0$ and $\mathrm{Var}_{H_0}(T_n(A)) = 1$ for all $n \geq 1$, consider

$$P_{H_0}\{\varphi_A = 1\} = P_{H_0}\{|T_n(A)| > c_n\} \leq \frac{1}{c_n^2}$$

by Chebyshev's inequality, and so the Type I error goes to zero. Turning attention to the Type II error, observe that if c_n diverges at a slower order than $|E_{H_1}(T_n(A))|$ diverges, then for all sufficiently large n, it follows that

$$\begin{aligned} P_{H_1}\{\varphi_A = 0\} &= P_{H_1}\{|T_n(A)| \leq c_n\} \\ &\leq P_{H_1}\{|E_{H_1}(T_n(A))| - |T_n(A) - E_{H_1}(T_n(A))| \leq c_n\} \\ &= P_{H_1}\{|E_{H_1}(T_n(A))| - c_n \leq |T_n(A) - E_{H_1}(T_n(A))|\} \\ &\leq \frac{\mathrm{Var}_{H_1}(T_n(A))}{(|E_{H_1}(T_n(A))| - c_n)^2} \\ &\asymp \frac{\mathrm{Var}_{H_1}(T_n(A))}{(E_{H_1}(T_n(A)))^2}. \end{aligned}$$

Therefore, if c_n diverges sufficiently slowly and $\frac{(E_{H_1}(T_n(A)))^2}{\mathrm{Var}_{H_1}(T_n(A))} \to \infty$, then both the Type I and Type II error go to zero and so the test φ_A is consistent. Thus, characterizing the consistency of the test φ_A boils down to characterizing when $\frac{(E_{H_1}(T_n(A)))^2}{\mathrm{Var}_{H_1}(T_n(A))} \to \infty$.

Gao and Ma directly calculate (equation (A.36) in Section A.5 of [18])

$$\frac{(E_{H_1}(T_n(A)))^2}{\mathrm{Var}_{H_1}(T_n(A))} \asymp \frac{(n\varepsilon(Q_n(A) - P_n(A)))^2}{nP_n(A) + n\varepsilon Q_n(A)}$$

and so if $\frac{(n\varepsilon Q_n(A))^2}{nP_n(A)+n\varepsilon Q_n(A)} \to \infty$, then $\frac{(E_{H_1}(T_n(A)))^2}{\mathrm{Var}_{H_1}(T_n(A))} \to \infty$. (As Gao and Ma note, the condition $n\varepsilon^2 P_n(A) \to \infty$ is also sufficient but has the strong and uninteresting requirement $\beta < \frac{1}{2}$.) Equivalently, if both

$$\frac{n\varepsilon^2 Q_n(A)^2}{P_n(A)} \to \infty,$$
$$n\varepsilon Q_n(A) \to \infty$$

hold, then $\frac{(E_{H_1}(T_n(A)))^2}{\mathrm{Var}_{H_1}(T_n(A))} \to \infty$. The conditions are equivalent to

$$\beta < \frac{1}{2} + \frac{\log Q_n(A)}{\log n} + \frac{1}{2}\min\left(1, -\frac{\log P_n(A)}{\log n}\right). \tag{10}$$

In other words, if β satisfies condition (10) for all n sufficiently large for some sequence of events $\{A_n\} \subset \mathscr{A}$ and c_n is a positive sequence diverging sufficiently slowly, then the sequence of tests φ_{A_n} is consistent for testing (1)-(2). To maximize the set of β satisfying condition (10) for all n sufficiently large, one should select an event A_n that maximizes the right hand side of (10) for each n. Since the right hand side of (10) is increasing in $Q_n(A)$ and decreasing in $P_n(A)$, Gao and Ma argue that the Neyman-Pearson lemma implies that the maximum is achieved by the event $A_n = \left\{x \in \mathcal{X} : \frac{q_n}{p_n}(x) > t_n\right\}$ for some $t_n > 0$. With this observation in mind, Gao and Ma naturally select the collection of events

$$\mathscr{A}_n^* := \left\{\left\{x \in \mathcal{X} : \frac{q_n}{p_n}(x) > t\right\} : t > 0\right\} \tag{11}$$

and define the general HC-type statistic

$$\mathrm{HC}_n^* := \mathrm{HC}_n(\mathscr{A}_n^*) = \sup_{t>0} \frac{\left|\sum_{i=1}^n \mathbf{1}_{\left\{\frac{q_n}{p_n}(X_i)>t\right\}} - nP\left(\frac{q_n}{p_n}(Y_n) > t\right)\right|}{\sqrt{nP\left(\frac{q_n}{p_n}(Y_n) > t\right) P\left(\frac{q_n}{p_n}(Y_n) \le t\right)}} \tag{12}$$

where $Y_n \sim P_n$ are independent of the data $\{X_i\}_{i=1}^n$. The corresponding higher criticism test is

$$\psi_{\mathrm{HC}_n^*} := \mathbf{1}_{\left\{\mathrm{HC}_n^* > \sqrt{2(1+\delta)\log\log(n)}\right\}} \tag{13}$$

where $\delta > 0$ is an arbitrary constant. Note that the cutoff $\sqrt{2(1+\delta)\log\log n}$ is the same cutoff used in Donoho and Jin's original formulation of the Higher Criticism test (4); this choice of cutoff is not at all surprising since HC_n^* is precisely Higher Criticism applied to univariate statistics.

3.2. A sufficient condition for optimality

With the review of Gao and Ma's construction complete, we now proceed to giving a sufficient condition for its optimality. We formally define a quantity $\underline{\beta}^{\text{HC}}$ which demarcates the sparsity levels for which $\psi_{\text{HC}_n^*}$ consistently tests (1)-(2).

Definition 3.1. Consider the testing problem (1)-(2) with calibration (3). Define

$$\underline{\beta}^{\text{HC}} := \frac{1}{2} + \sup \liminf_{n\to\infty} \left\{ \frac{\log Q_n(A_n)}{\log n} + \frac{1}{2}\min\left(1, -\frac{\log P_n(A_n)}{\log n}\right)\right\}$$

where the supremum runs over sequences of events $\{A_n\}$ with $A_n \in \mathscr{A}_n^*$. Here, $\mathscr{A}_n^*$ is given by (11).

Proposition 3.1. *Consider the testing problem (1)-(2) with calibration (3). If $\beta < \underline{\beta}^{\text{HC}}$, then $\psi_{\text{HC}_n^*}$ is consistent. Here, $\psi_{\text{HC}_n^*}$ is given by (13).*

Proof. The choice of threshold $\sqrt{2(1+\delta)\log\log(n)}$ is given by Theorem 1.1 of [10]. Theorem 1.1 of [10] implies that this choice of threshold results in a vanishing Type I error of $\psi_{\text{HC}_n^*}$.

Turning attention to the Type II error, if β in the calibration (3) satisfies condition (10) for all n sufficiently large for some sequence of events $\widetilde{A}_n \in \mathscr{A}_n^*$, then it immediately follows that

$$P_{H_1}\{\psi_{\text{HC}_n^*} = 0\} = P_{H_1}\left\{\sup_{A_n \in \mathscr{A}_n^*} |T_n(A_n)| \leq \sqrt{2(1+\delta)\log\log(n)}\right\} \tag{14}$$

$$\leq P_{H_1}\left\{|T_n(\widetilde{A}_n)| \leq \sqrt{2(1+\delta)\log\log(n)}\right\} \tag{15}$$

$$= P_{H_1}\{\varphi_{\widetilde{A}_n} = 0\} \tag{16}$$

where $c_n = \sqrt{2(1+\delta)\log\log(n)}$. Since β satisfies condition (10) for $\widetilde{A}_n$ for all sufficiently large n, it immediately follows that $|E_{H_1}(T_n(\widetilde{A}_n))|$ diverges at a polynomial rate. Moreover, it follows that c_n diverges sufficiently slowly as c_n grows at a sub-polynomial rate. Hence, $P_{H_1}\{\varphi_{\widetilde{A}_n} = 0\}$ converges to zero and so the Type II error of $\psi_{\text{HC}_n^*}$ vanishes. Therefore, it has been shown that if β in the calibration (3) satisfies condition (10) for some sequence of events $\widetilde{A}_n \in \mathscr{A}_n^*$ for all n sufficiently large, then $\psi_{\text{HC}_n^*}$ is consistent. □

At first glance, it seems that HC_n^* requires full knowledge of both the null $\{P_n\}$ and signal $\{Q_n\}$ distributions. In contrast, Donoho and Jin's formulation of Higher Criticism (5) for the sparse normal mixture detection problem does not require knowledge of the signal strength r. The key observation is that the computation of HC_n^* actually only requires knowledge of the collection $\mathscr{A}_n^*$ so that one may calculate $\sup_{A\in\mathscr{A}_n^*} |T_n(A)|$. In some cases, it is possible to calculate the supremum without knowing $\{Q_n\}$ explicitly. For example, consider the sparse normal mixture detection problem with $P_n = N(0,1)$ and $Q_n = N(\sqrt{2r\log n}, 1)$ with $0 < r \leq 1$. A direct calculation shows

$$\begin{aligned}\mathscr{A}_n^* &= \left\{\left\{x \in \mathbb{R} : \frac{q_n}{p_n}(x) > t\right\} : t > 0\right\} \\ &= \left\{\left\{x \in \mathbb{R} : \exp\left(x\sqrt{2r\log n} - r\log n\right) > t\right\} : t > 0\right\} \\ &= \left\{\left\{x \in \mathbb{R} : x > t\right\} : t \in \mathbb{R}\right\}.\end{aligned}$$

Therefore, HC_n^* reduces to Donoho and Jin's Higher Criticism statistic HC_n given in (5), and so knowledge of the signal strength r is not required. In many problems, it may be the case that computation of HC_n^* does not require full knowledge of the signal distribution Q_n even though Gao and Ma's construction gives that impression.

With the HC_n^* testing statistic in hand, the next challenge is to investigate $\underline{\beta}^{\mathrm{HC}}$. When $\left\{\frac{\log \frac{q_n}{p_n}}{\log n}\right\}$ satisfies the large deviation principle under the null, a lower bound can be derived for $\underline{\beta}^{\mathrm{HC}}$.

Proposition 3.2. *Consider the testing problem (1)-(2) with calibration (3). Suppose there exists some $\gamma > 1$ such that the tail condition (6) holds for $X_n \sim P_n$. Suppose further that $\left\{\frac{\log \frac{q_n}{p_n}}{\log n}\right\}$ satisfies the large deviation principle under the null. Let $I : \mathbb{R} \to [0,\infty]$ be the associated good rate function. Then,*

$$\underline{\beta}^* \geq \underline{\beta}^{\mathrm{HC}} \geq \frac{1}{2} + \sup_{c \geq 0}\left\{\sup_{t > c}\{t - I(t)\} + \frac{1 \wedge \inf_{t \geq c} I(t)}{2}\right\}. \tag{17}$$

When the rate function associated to the large deviation principle is convex (and satisfies some further constraints) and the conditions of Corollary 2.1 hold, it can be shown that the lower and upper bounds in (17) match.

Theorem 3.1. *Consider the setting of Proposition 3.2 and suppose that the conditions of Corollary 2.1 hold. Suppose I is convex. Let $D := \{t \in \mathbb{R} : I(t) < \infty\}$ and note that D is an interval with some left endpoint $\underline{d}$ and some right endpoint $\overline{d}$. Suppose further that I is such that if $\underline{d} \in D$, we have that I is right-continuous at $\underline{d}$ and if $\overline{d} \in D$, we have that I is left-continuous at $\overline{d}$. Let $I'_-(t)$ be the left derivative of I (see Definition 6.2) with the domain of definition extended as in the statement of Theorem 6.3. Define $t_0 := \sup\{t \geq 0 : I'_-(t) \leq 0\}$ and set $t_0 = 0$ if $\{t \geq 0 : I'_-(t) \leq 0\} = \emptyset$. Likewise, define $t_1 := \sup\{t \geq 0 : I'_-(t) \leq 1\}$ and set $t_1 = 0$ if $\{t \geq 0 : I'_-(t) \leq 1\} = \emptyset$. If $t_0 \vee t_1 < \infty$, then*

$$\underline{\beta}^{\mathrm{HC}} = \underline{\beta}^* = \overline{\beta}^* = \beta^* = \frac{1}{2} + \left(\sup_{t \geq 0}\left\{t - I(t) + \frac{1 \wedge I(t)}{2}\right\}\right)_+.$$

The Gärtner-Ellis Theorem gives general conditions ensuring the convexity of the rate function. We state a special case of the Gärtner-Ellis Theorem (Theorem 2.3.6 in [8]) specialized for our use in the sparse mixture detection problem.

In many problems, the Gärtner-Ellis Theorem greatly simplifies the work needed in determining whether the large deviation principle under the null holds and computing the corresponding rate function. First, a regularity condition (Definition 2.3.5 from [8]) is needed.

Definition 3.2. Let $\Lambda : \mathbb{R} \to (-\infty, \infty]$ be a convex function and let $D_\Lambda := \{\lambda \in \mathbb{R} : \Lambda(\lambda) < \infty\}$. We say Λ is *essentially smooth* if $D_\Lambda^\circ \neq \emptyset$, Λ is differentiable on D_Λ°, and $\lim_{n\to\infty} |\Lambda'(\lambda_n)| = \infty$ for any sequence $\{\lambda_n\} \subset D_\Lambda^\circ$ converging to a point on the boundary of D_Λ°.

The following statement of the Gärtner-Ellis Theorem follows the presentation of Theorem 2.3.6 in [8] with modifications to suit our setting.

Theorem 3.2 (Gärtner-Ellis)**.** *Suppose $\{P_n\}$ and $\{Q_n\}$ are probability measures for the testing problem (1)-(2). For $\lambda \in \mathbb{R}$, define*

$$\Lambda_n(\lambda) := \frac{1}{\log n} \cdot \log E\left[\left(\frac{q_n}{p_n}(X_n)\right)^\lambda\right]$$

where $X_n \sim P_n$. Assume that the limit

$$\lim_{n\to\infty} \Lambda_n(\lambda) =: \Lambda(\lambda) \tag{18}$$

exists in $[-\infty, \infty]$ for $\lambda \in \mathbb{R}$. If Λ is essentially smooth, is a lower semicontinuous function, and $0 \in D_\Lambda^\circ$, then $\left\{\frac{\log \frac{q_n}{p_n}}{\log n}\right\}$ satisfies the large deviation principle under the null with good, convex rate function

$$\Lambda^*(t) := \sup_{\lambda \in \mathbb{R}} \{\lambda t - \Lambda(\lambda)\}.$$

Proof. Since $\lim_{n\to\infty} \Lambda_n(\lambda)$ exists in $[-\infty, \infty]$ and $0 \in D_\Lambda^\circ$, it follows that Assumption 2.3.2 of [8] is satisfied. Then, Lemma 2.3.9 of [8] yields the convexity of Λ as well as establishing that $\Lambda(\lambda) > -\infty$ for all λ and that Λ^* is a good convex rate function. Finally, Theorem 2.3.6 (Gärtner-Ellis) in [8] implies that $\left\{\frac{\log \frac{q_n}{p_n}}{\log n}\right\}$ satisfies the large deviation principle under the null with rate function Λ^*. □

Remark 3.1. The Gärtner-Ellis Theorem simplifies calculating the rate function of the large deviation principle in some exponential families. Suppose $\{f_\theta : \theta \in \Theta\}$ is an exponential family in the natural parametrization on a separable metric space with $\Theta \subset \mathbb{R}^d$. Write

$$f_\theta(x) = c(\theta)h(x)\exp\left(\langle \theta, T(x)\rangle\right).$$

Taking $p_n = f_\theta$ for some $\theta \in \Theta$ and $q_n = f_{\theta_n}$ for some sequence $\{\theta_n\}$ in Θ, the limit (18) becomes

$$\Lambda(\lambda) := \lim_{n\to\infty} \frac{\lambda \log c(\theta_n) + (1-\lambda)\log c(\theta) - \log c(\lambda(\theta_n - \theta) + \theta)}{\log n}$$

where we take $\log c(\lambda(\theta_n - \theta) + \theta) = -\infty$ if $\lambda(\theta_n - \theta) + \theta \notin \Theta$. Of course, one must check that the limit exists and the remaining conditions of Theorem 3.2 hold.

4. Examples

To illustrate our results and typical applications, we consider a few examples and state explicit detection boundaries. Details of our calculations are found in the supplementary material [30]. The derivations in the supplement highlight the typical methods of calculation when using the results of Sections 2 and 3.

4.1. Ingster-Donoho-Jin

4.1.1. Univariate

Consider the testing problem (1)-(2) with calibration (3) and distributions $P_n = N(0,1)$ and $Q_n = N(\mu_n, 1)$. The detection boundary for this testing problem with calibration $\mu_n = \sqrt{2r\log n}$ for $0 < r \leq 1$ was obtained by Ingster [23] and then independently by Jin [24, 25]. Donoho and Jin [10] introduced the Higher Criticism testing statistic and established its optimality in this sparse mixture detection problem. Following [6, 18], we refer to the detection boundary as the Ingster-Donoho-Jin detection boundary.

We illustrate how the large deviations perspective delivers both the Ingster-Donoho-Jin detection boundary

$$\beta^*_{IDJ}(r) := \begin{cases} \frac{1}{2} + r & \textit{if } 0 < r \leq \frac{1}{4} \\ 1 - (1-\sqrt{r})^2_+ & \textit{if } r > \frac{1}{4} \end{cases} \tag{19}$$

and the optimality of the Higher Criticism statistic. We use the Gärtner-Ellis Theorem and Remark 3.1. Before we begin the main computation, note that it is easily verified that the tail condition (6) is satisfied. Adopting the notation of Remark 3.1, consider that $\{N(\theta,1) : \theta \in \mathbb{R}\}$ is an exponential family with natural parameter θ and $\log c(\theta) = -\frac{\theta^2}{2}$. Taking $\theta = 0$ and $\theta_n = \mu_n$, observe that

$$\Lambda(\lambda) := \lim_{n\to\infty} \frac{\lambda \log c(\theta_n) + (1-\lambda)\log c(\theta) - \log c(\lambda(\theta_n - \theta) + \theta)}{\log n} = r(\lambda^2 - \lambda).$$

Noting that $D_\Lambda = \mathbb{R}$, Λ is essentially smooth, and Λ is continuous, it follows from the Gärtner-Ellis Theorem that $\left\{\frac{\log \frac{q_n}{p_n}}{\log n}\right\}$ satisfies the large deviation principle under the null with good convex rate function $\Lambda^*(t) := \sup_{\lambda\in\mathbb{R}}\{\lambda t - \Lambda(\lambda)\}$. Direct calculation yields $\Lambda^*(t) = \frac{(t+r)^2}{4r}$. It is easily checked that conditions of Corollary 2.1 are satisfied. Therefore, the detection boundary is given by

$$\beta^*(r) = \frac{1}{2} + \left(\sup_{t\geq 0}\left\{t - \frac{(t+r)^2}{4r} + \frac{1 \wedge \frac{(t+r)^2}{4r}}{2}\right\}\right)_+.$$

Solving the optimization problem yields $\beta^*(r) = \beta^*_{IDJ}(r)$.

Turning our attention to the Higher Criticism statistic, observe that the general HC-type statistic (12) of Gao and Ma reduces to the original Higher Criticism statistic introduced by Donoho and Jin [10]

$$\begin{aligned}\mathrm{HC}^*_n &= \sup_{t>0} \frac{\left|\sum_{i=1}^n \mathbf{1}_{\{\exp(X_i\mu_n - \mu_n^2/2)>t\}} - nP(\exp(X\mu_n - \mu_n^2/2) > t)\right|}{\sqrt{nP(\exp(X\mu_n - \mu_n^2/2) > t)P(\exp(X\mu_n - \mu_n^2/2) \leq t)}} \\ &= \sup_{t\in\mathbb{R}} \frac{\left|\sum_{i=1}^n \mathbf{1}_{\{X_i>t\}} - nP(X > t)\right|}{\sqrt{nP(X > t)P(X \leq t)}}\end{aligned}$$

where $X \sim N(0, 1)$ is independent of $\{X_i\}_{i=1}^n$. The conditions of Theorem 3.1 hold, and so the sequence of tests $\psi_{\mathrm{HC}^*_n}$ given by (13) achieves the detection boundary while adapting to the parameters r and β. In other words, $\underline{\beta}^{\mathrm{HC}} = \beta^*$.

4.1.2. Multivariate

The detection boundary for a multivariate version of the sparse normal mixture testing problem can be obtained in exactly the same fashion. Consider the testing problem (1)-(2) with $P_n = N(0, \Sigma)$, $Q_n = N(\mu_n, \Sigma)$ where $\Sigma \in \mathbb{R}^{d\times d}$ is a positive definite matrix. Further consider the calibration $\mu_n = \sqrt{2r \log n} \cdot u$ where $u \in \mathbb{R}^d$ with $||u|| = 1$. We can use the Gärtner-Ellis Theorem and Remark 3.1.

It can be shown that the large deviation principle under the null is satisfied with rate function $\Lambda^*(t) = \frac{\left(t+r\langle u,\Sigma^{-1}u\rangle\right)^2}{4r}$. Repeating the reasoning from Section 4.1.1 yields the detection boundary

$$\beta^*(r) = \begin{cases} \frac{1}{2} + r\langle u, \Sigma^{-1}u\rangle & \text{if } r\langle u, \Sigma^{-1}u\rangle \leq \frac{1}{4}, \\ 1 - \left(1 - \sqrt{r\langle u, \Sigma^{-1}u\rangle}\right)_+^2 & \text{otherwise.} \end{cases} \tag{20}$$

Details for the relevant calculations are found in the supplement [30].

The term $\langle u, \Sigma^{-1}u\rangle$ captures how the signal direction u and the covariance Σ interact to influence the detection boundary; the boundary can be summarized as $\beta^*(r) = \beta^*_{IDJ}(\langle u, \Sigma^{-1}u\rangle r)$. Notably when $\Sigma = I_d$ we have $\beta^*(r) = \beta^*_{IDJ}(r)$, meaning neither the direction u nor the dimension d has an impact on the detection boundary.

One's initial intuition might raise a red flag here when $\Sigma = I_d$, thinking that more structured $u \in \mathbb{R}^d$ (say, k-sparse) should yield more favorable detection boundaries than unstructured $u \in \mathbb{R}^d$. However, this intuition is applicable when u is not known exactly but rather is known to belong to a structured set. In contrast, the problem (1)-(2) is a simple null versus simple alternative testing problem. With respect to the fundamental limit of the testing problem, u (along with r and β) are known to the statistician. It is precisely because it is a simple versus simple testing problem that the likelihood ratio test is optimal as discussed earlier. Since the likelihood ratio is a function of the univariate statistics $\langle X_i, u\rangle$ when $\Sigma = I_d$, intuitively the dimension d does not affect the detection boundary. Additionally, since $\Sigma = I_d$ an appeal to symmetry implies u should also not affect the detection boundary.

To illustrate further, consider the one-dimensional problem by forming $W_i = \langle X_i, u\rangle$. This transformation intuitively captures all of the relevant information by an appeal to the likelihood ratio. The transformed data $\{W_i\}$ is distributed as

$$\begin{aligned} &H_0 : W_1, \dots, W_n \overset{iid}{\sim} N(0,1), \\ &H_1 : W_1, \dots, W_n \overset{iid}{\sim} (1-\varepsilon)N(0,1) + \varepsilon N(\sqrt{2r\log n}, 1). \end{aligned}$$

The intuition for why $\beta^*(r) = \beta^*_{IDJ}(r)$ is clear after examining this transformed problem. The application of the Gärtner-Ellis Theorem and Remark 3.1 proves that this reduction is indeed valid and delivers the correct detection boundary. Though easy, this multivariate extension of Section 4.1.1 is included to provide a simple illustration of how the paper's general theory can quickly yield the detection boundary in a unified manner. Problem specific transformations can be completely avoided.

The next natural thought is to compare this multivariate analysis to what is considered the usual Higher Criticism analysis [11]. For simplicity, consider $\Sigma = I_d$ and $u \in \mathbb{R}^d$ a fixed unit vector. In this case, a typical analysis may combine all of the coordinates of the data $\{X_i\}_{i=1}^n$ to obtain a collection of nd univariate data points; Higher Criticism (5) is then applied to the nd points. To understand how this approach differs from the methodology proposed in this paper, consider the following natural sparse mixture detection problem corresponding to the typical Higher Criticism analysis,

$$H_0 : Y_1, \dots, Y_{nd} \overset{iid}{\sim} N(0,1), \tag{21}$$

$$H_1 : Y_1, \dots, Y_{nd} \overset{iid}{\sim} (1-\varepsilon)N(0,1) + \varepsilon(H_u * N(0,1)) \tag{22}$$

where $\varepsilon = n^{-\beta}$ and $H_u = \frac{1}{d}\sum_{i=1}^n \delta_{\sqrt{2r\log n}\cdot u_i}$. Here, $*$ denotes convolution and δ_t denotes the probability measure placing full mass at t.

Now that the data are all univariate, the results of [6] can be applied directly. Corollary 2 in [6] applies generally to signal distributions which are convolutional form (here we have $H_u * N(0,1)$) and so the detection boundary is

$$\beta^*_{conv}(r) = \begin{cases} \frac{1}{2} + r||u||_\infty^2 & \textit{if } r \le \frac{1}{4||u||_\infty^2}, \\ 1 - (1 - ||u||_\infty\sqrt{r})_+^2 & \textit{if } r > \frac{1}{4||u||_\infty^2}. \end{cases}$$

The keen will note $\beta^*_{conv}(r) = \beta^*_{IDJ}(r||u||_\infty^2)$. In contrast, since $\Sigma = I_d$ and u is a unit vector, we have $\beta^*(r) = \beta^*_{IDJ}(r)$ where $\beta^*(r)$ is the detection boundary (20) coming from the multivariate setup. Notably instead of $||u||_\infty^2$ showing up as in $\beta^*_{conv}(r)$, the term $||u||^2 = 1$ takes its place in $\beta^*(r)$. Note $||u||_\infty \le 1$, and so it is easily checked that $\beta^*(r) \ge \beta^*_{conv}(r)$ for all $r > 0$.

The inequality $\beta^* \ge \beta^*_{conv}$ can be interpreted in plain language as follows. For any signal strength level $r > 0$, there exists a consistent sequence of tests in the multivariate setup at more severe sparsities than what is possible in the typical Higher Criticism analysis. Said another way, for a given sparsity level β, one can successfully test for lower signal strengths r. In other words, there is nontrivial structure in the multivariate setup which is ignored when one goes about the usual univariate Higher Criticism analysis.

With the multivariate setup, one is guaranteed that those data points coming from Q_n are, on average, near the point $\mu_n = \sqrt{2r \log n} \cdot u$. The structure is explicitly exploited when applying Gao and Ma's Higher Criticism type statistic as one essentially calculates $\langle X_i, u \rangle$ (see Section A.1.1 in the supplement [30]). When going about the usual Higher Criticism analysis by putting together all of the coordinates and considering (21)-(22), one is only guaranteed that a point coming from $H_u * N(0,1)$ is near *some* point in the collection $\{\sqrt{2r \log n} \cdot u_i\}_{i=1}^d$. The authors of [6] note in Remark 3 that in such convolutional signal distributions the detection boundary is driven only by the maximal component $||u||_\infty$, and notably the heterogeneity amongst the coordinates in u is immaterial. In contrast, this heterogeneity is critically exploited in the multivariate setup.

4.1.3. Brownian motion

With the large deviations perspective in hand, a phase transition can be derived in a stylized sparse mixture detection problem where our observations are sample paths of Brownian motion with possible drift. In particular, let $\mathcal{X} = C([0,1])$ be the space of all real-valued continuous functions on $[0,1]$, let P_n be the probability measure on $\mathcal{X}$ associated with standard Brownian motion $\{B_t\}_{t \in [0,1]}$, and let Q_n be the probability measure on $\mathcal{X}$ associated with the Brownian motion with drift $\{m_n(t) + B_t\}_{t \in [0,1]}$. Here, we take $m_n(t) = \sqrt{2r \log n} \cdot f(t)$ for some fixed continuously differentiable $f : [0,1] \to \mathbb{R}$ with $f(0) = 0$ and $\int_0^1 |f'(t)|^2 \, dt = 1$. With observations $X_1, \dots, X_n \in \mathcal{X}$, the problem is to test (1)-(2). Note that the observations are themselves real-valued functions on $[0,1]$. Further note that since $f' \in L^2([0,1])$, the measures P_n and Q_n are mutually absolutely continuous (Example 4 in [42]). The normalized log-likelihood ratio is given by [42]

$$\begin{aligned}\frac{\log \frac{dQ_n}{dP_n}(X)}{\log n} &= -\frac{1}{2\log n} \int_0^1 |m_n'(t)|^2 \, dt + \frac{1}{\log n} \int_0^1 m'(t) \, dX_t \\ &= -r + \frac{\sqrt{2r}}{\sqrt{\log n}} \int_0^1 f'(t) \, dX_t.\end{aligned}$$

It can be shown that the large deviations principle under the null is satisfied with good rate function $I(t) = \frac{(t+r)^2}{4r}$, and so the detection boundary is exactly given by $\beta^*_{IDJ}(r)$ from Section 4.1.1. Furthermore, it is clear that the rate function I is indeed convex and that the other conditions of Theorem 3.1 hold. Details for the relevant calculations are found in the supplement [30].

4.2. Heteroscedastic normal mixture

Cai, Jeng, and Jin [5] consider the testing problem (1)-(2) in a heteroscedastic normal mixture setting. More specifically, the setting where $P_n = N(0,1)$ and $Q_n = N(\mu_n, \sigma^2)$ is considered with calibration $\mu_n = \sqrt{2r \log n}, r > 0$, and fixed $\sigma^2 > 0$. Through an analysis of the likelihood ratio, they obtain the detection boundary

$$\beta^*(r,\sigma^2) := \begin{cases} \frac{1}{2} + \frac{r}{2-\sigma^2} & \textit{if } 2\sqrt{r} + \sigma^2 \leq 2, \\ 1 - \frac{(1-\sqrt{r})_+^2}{\sigma^2} & \textit{if } 2\sqrt{r} + \sigma^2 > 2. \end{cases}$$

Note that the detection boundary stated in [5] is in terms of r as a function of β and σ, whereas the above boundary is in terms of β as a function of r and σ. The boundaries are equivalent (in [6], see (21) and Section V.C). While one can straightforwardly obtain the detection boundary through Theorem 1 of [6], we illustrate a typical calculation under the large deviations perspective. For ease of calculation, let us take $\sigma^2 \neq 1$ without loss of generality. The case of $\sigma^2 = 1$ is just the Ingster-Donoho-Jin problem.

To derive the detection boundary, we establish that the large deviation principle under the null holds, derive the rate function, and apply Corollary 2.1. Note that to apply Corollary 2.1, it must be checked that the tail condition (6) holds; this is verified in the supplement [30]. It can be shown that the large deviation principle under the null holds with rate function

$$I(t) = \begin{cases} \frac{\sigma^2}{(\sigma^2-1)^2}\left(\sqrt{(\sigma^2-1)t+r} - \sqrt{\frac{r}{\sigma^2}}\right)^2 & \textit{if } t(\sigma^2-1)+r \geq 0, \\ \infty & \textit{if } t(\sigma^2-1)+r < 0. \end{cases}$$

Corollary 2.1 can be applied (after checking the conditions hold) with the rate function I to obtain the same detection boundary proved by Cai, Jeng, and Jin [5]. Additionally, it can be checked that Theorem 3.1 holds and so the Higher Criticism test defined in (13) achieves the detection boundary. See the supplement [30] for details.

4.3. Mixture of a mixture I

A sparse mixture detection problem making more use of the multivariate setting is the following. Consider the testing problem (1)-(2) with $P_n = N(0, I_d)$ and $Q_n = \frac{1}{2}N(\mu_1, I_d) + \frac{1}{2}N(\mu_2, I_d)$ where $\mu_1 = \sqrt{2r\log n} \cdot u_1$ and $\mu_2 = \sqrt{2r \log n} \cdot u_2$ where u_1, u_2 are fixed and linearly independent unit vectors in $\mathbb{R}^d$ and $r > 0$. Note that the tail condition (6) is easily verified. It can be shown that the large deviation principle under the null is satisfied with good rate function

$$I(t) = \begin{cases} \frac{(t+r)^2}{4r} & \textit{if } t \geq -r, \\ \frac{(t+r)^2}{2r(1+\langle u_1,u_2\rangle)} & \textit{if } t < -r. \end{cases}$$

Given the form of I, Corollary 2.1 immediately implies that the detection boundary $\beta^*(r)$ is exactly the same as the Ingster-Donoho-Jin detection boundary

$\beta^*_{IDJ}(r)$. It is also readily checked that Gao and Ma's Higher Criticism statistic furnishes a test which achieves the detection boundary. The details of the calculations yielding these results are found in the supplement [30].

4.4. Mixture of a mixture II

The following sparse mixture detection problem is inspired by the testing equivalence of clustering problem considered by Gao and Ma (Section 2.3 in [18]). Let $u, v \in \mathbb{R}^d$ be orthogonal unit vectors, i.e. $||u|| = ||v|| = 1$ and $\langle u, v \rangle = 0$. Let $\mu_n = \sqrt{2r \log n} \cdot u$ and $\nu_n = \sqrt{2r \log n} \cdot v$ with fixed $0 < r \leq 1$. Consider the testing problem (1)-(2) with $P_n = \frac{1}{2}N(\mu_n, I_d) + \frac{1}{2}N(-\mu_n, I_d)$ and $Q_n = \frac{1}{2}N(\nu_n, I_d) + \frac{1}{2}N(-\nu_n, I_d)$. As in previous examples, we establish a large deviations principle under the null, derive the rate function, and apply Corollary 2.1 to obtain the detection boundary. The tail condition (6) is verified to hold in the supplement [30].

It can be shown the the large deviation principle under the null is satisfied with good rate function

$$I(t) = \begin{cases} \frac{(t+2r)^2}{4r} & \textit{if } t < -2r, \\ \frac{(t+2r)^2}{8r} & \textit{if } |t| \leq 2r, \\ \frac{(t+2r)^2}{4r} - t & \textit{if } t > 2r. \end{cases}$$

Since the tail condition (6) is satisfied and since the conditions of Corollary 2.1 can be verified to hold, it follows that the detection boundary is given by an application of Corollary 2.1. A calculation can be done (see the supplement [30] for details) to obtain the detection boundary

$$\beta^*(r) = \begin{cases} \frac{3}{2}r + \frac{1}{2} & \textit{if } r \leq \frac{1}{5}, \\ \sqrt{1 - (1-2r)_+^2} & \textit{if } r > \frac{1}{5}. \end{cases}$$

We have exactly recovered the detection boundary stated in Theorem 2.3 of [18]. In fact, this detection boundary holds in a more general setting than that originally considered by Gao and Ma [18]. Additionally, it can be checked that Theorem 3.1 holds.

4.5. Detection of a low-rank perturbation

Consider the testing problem (1)-(2) with calibration (3) and distributions $P_n = N(0, I_p), Q_n = N(0, I_p + H)$ where H is a rank $k < p$ symmetric matrix with its k nonzero eigenvalues equal to $r > 0$. Note that we can write $H = Q \cdot rA_k \cdot Q^\intercal$ where $Q \in \mathbb{R}^{p \times p}$ is an orthogonal matrix and A_k is a diagonal matrix with the first k entries on the diagonal equal to one and the remaining diagonal entries equal to zero.

It can be shown that the tail condition (6) holds and that the large deviations principle under the null is satisfied with good rate function

$$I(t) = \begin{cases} \frac{r+1}{r} t & \textit{if } t \geq 0, \\ \infty & \textit{otherwise.} \end{cases}$$

It can be verified that Corollary 2.1 can be applied, yielding the detection boundary

$$\beta^*(r) = \begin{cases} \frac{1}{2} & \textit{if } r \leq 1, \\ 1 - \frac{1}{1+r} & \textit{if } r > 1. \end{cases}$$

Note the detection boundary is exactly the same as in the heteroscedastic normal mixture testing problem with $\mu = 0$ and $\sigma^2 = 1 + r$. It can also be checked that Theorem 3.1 holds. Details for the calculations yielding the above results are found in the supplement [30].

4.6. Detection of sparse correlated pairs

In Section 5 of the review article [11], Donoho and Jin consider the problem of detecting the presence of a small collection of correlated pairs. More specifically, the testing problem (1)-(2) with $P_n = N(0, I_2)$ and $Q_n = N(\mu_n \mathbf{1}_2, \Sigma)$ is considered where $\mu_n = \sqrt{r \log n}$ for $r > 0$ and

$$\mathbf{1}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

with $-1 < \rho < 1$. Without loss of generality, we take $\rho \neq 0$. The case $\rho = 0$ reduces to a special case of the sparse multivariate normal mixture studied in Section 4.1.2. Donoho and Jin illustrate the applicability of their original formulation of the Higher Criticism statistic and perform some simulations. We will deduce the detection boundary and investigate the behavior of Gao and Ma's Higher Criticism type testing statistic. It turns out that the detection boundary is highly related to the detection boundary in Section 4.2. A large deviation principle under the null can be established with good rate function

$$I(t) = \begin{cases} \frac{(\rho-1)(r+2\rho t)}{2\rho^2} & \textit{if } \rho t + r \leq \frac{(1+\rho)r}{4}, \\ \frac{1+\rho}{\rho^2}\left(\sqrt{\rho t + r} - \sqrt{\frac{r}{1+\rho}}\right)^2 & \textit{if } \rho t + r > \frac{(1+\rho)r}{4}. \end{cases}$$

An application of Corollary 2.1 along with a lengthy calculation yields the detection boundary

$$\beta^*(r, \rho) = \begin{cases} \frac{1}{2} + \frac{r}{1-\rho} & \textit{if } 2\sqrt{r} + \rho \leq 1, \\ 1 - \frac{(1-\sqrt{r})_+^2}{1+\rho} & \textit{if } 2\sqrt{r} + \rho > 1. \end{cases}$$

This is precisely the detection boundary of Cai, Jeng, and Jin [5] with $1+\rho = \sigma^2$. Note that we have parametrized the mean $\mu = \sqrt{r \log n}$ here, whereas the mean has parametrization $\sqrt{2r \log n}$ in Section 4.2. Theorem 3.1 can be verified to hold. The details for the calculations yielding these results are found in the supplement [30].

We presented the case $d = 2$ as this was the problem considered by [11]. However, our techniques can be used to treat the general $d > 2$ case. For proper setup, let us parametrize $\mu_n = \sqrt{\frac{2r}{d} \log n}$. Set $P_n = N(0, I_d)$ and $Q_n = N(\mu_n \mathbf{1}_d, \Sigma)$ where $\mathbf{1}_d \in \mathbb{R}^d$ is the vector with all entries equal to one and $\Sigma = (1-\rho)I_d + \rho \mathbf{1}_d \mathbf{1}_d^\intercal$ is the matrix with 1 on the diagonal and ρ everywhere else. Since it is anyway required that $\rho \in (-\frac{1}{d-1}, 1)$ in order for Σ to be positive definite, let us only consider $\rho > 0$ for simplicity. Though the details become more complicated in the $d > 2$ case, the broad path of the calculation for the $d = 2$ case can be followed to derive the rate function,

$$I(t) = \begin{cases} \frac{(\rho-1)(r(d-1)+\rho d t)}{\rho^2 d(d-1)} & \text{if } t\rho(d-1) + r \leq \frac{1-\rho+\rho d}{d^2} r, \\ \frac{1-\rho+\rho d}{\rho^2 (d-1)^2} \left(\sqrt{t\rho(d-1)+r} - \sqrt{\frac{r}{1-\rho+\rho d}} \right)^2 & \text{if } t\rho(d-1) + r > \frac{1-\rho+\rho d}{d^2} r. \end{cases}$$

Just as before, the resulting detection boundary is an instantiation of the detection boundary from Section 4.2. The detection boundary is

$$\beta^*(r, \rho) = \begin{cases} \frac{1}{2} + \frac{r}{1+\rho-\rho d} & \text{if } 2\sqrt{r} + \rho(d-1) \leq 1, \\ 1 - \frac{(1-\sqrt{r})_+^2}{1-\rho+\rho d} & \text{if } 2\sqrt{r} + \rho(d-1) > 1. \end{cases}$$

This is precisely the detection boundary from Section 4.2 with $\sigma^2 = 1 - \rho + \rho d$. After a moment's thought, this quantity is natural as it is the eigenvalue of Σ corresponding to the eigenvector $\frac{1}{\sqrt{d}} \mathbf{1}_d$, which is the direction of the signal.

The explicit expression for the detection boundary reveals a subtle interaction between the dimension d and the correlation level ρ. Observe that if $d > 1 + \frac{1}{\rho}$, then the condition $2\sqrt{r} + \rho(d-1) > 1$ is satisfied for all $r \geq 0$. Hence, if $d > 1 + \frac{1}{\rho}$, the detection boundary becomes

$$\beta^*(r, \rho) = 1 - \frac{(1-\sqrt{r})_+^2}{1-\rho+\rho d}.$$

Strikingly, $\beta^*(r, \rho)$ is actually increasing in d. In other words, at a given sparsity level, weaker signals can be detected for larger d compared to smaller d. One's intuition might sound the alarm here as it seems the problem should get harder for larger d.

However, a comparison to Section 4.2 yields clarity. Since the signal is in the direction of $\frac{1}{\sqrt{d}} \mathbf{1}_d$, the noise level under the signal distribution is effectively $1 - \rho + \rho d$. The noise level under the signal distribution grows with d while the noise level under the null distribution is constant. This is analogous to σ^2 increasing in the detection boundary of Section 4.2. As remarked in [5], "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.)" In our setting, the same phenomenon is attributed to the dimension d. In the heteroscedastic problem of Section 4.2, $\sigma^2 = 2$ is the critical variance. Namely, heteroscedasticity alone enables successful detection if and only if $\sigma^2 > 2$. In the same way, $d = 1 + \frac{1}{\rho}$ is the critical dimension (equivalently, one can speak in terms of critical correlation). The supplement [30] outlines how the rate function and detection boundary are obtained.

4.7. Stochastic block model: Detection of a change in a node's community

In some areas (such as sociology, political science, and neuroscience), the observational units exhibit relationships amongst one another thereby forming a network. The field of network analysis deals with addressing statistical questions regarding the observed network, such as determining whether there exist latent communities in the network, identifying the communities if they do exist, and possibly estimating parameters of a statistical model. We refer the reader to a recent survey [17] covering fundamental statistical limits in a number of estimation and testing tasks in network analysis.

The stochastic block model (SBM) [21] is a popular model for capturing the presence of communities in a network. We refer the reader to the survey [1] for further background. In the simplest case of two communities, consider n nodes and let $z \in \{-1, 1\}^n$ denote the community membership for the n nodes, i.e. z_i denotes to which community node i belongs. The statistician's observation is the random symmetric matrix $A \in \{0, 1\}^{n \times n}$, where $A_{ij} = A_{ji} = 1$ denotes the presence of an edge between nodes i and j. Likewise, $A_{ij} = A_{ji} = 0$ denotes the absence of an edge. The data generating process is determined by the community structure, namely for $1 \leq i < j \leq n$, we have independent draws

$$A_{ij} \sim \begin{cases} \text{Bernoulli}(p) & \textit{if } z_i = z_j, \\ \text{Bernoulli}(q) & \textit{if } z_i \neq z_j. \end{cases}$$

We set $A_{ji} = A_{ij}$ to enforce symmetry and $A_{ii} = 0$ to disallow self-loops. Here, $p, q \in (0, 1)$ are parameters. Note that p gives the probability of an edge between two nodes in the same community and q gives the probability of an edge between two nodes in different communities. When $p > q$ the SBM is said to be assortative; otherwise, the SBM is said to be disassortative. There is a litany of statistical tasks associated with the SBM; perhaps the most popular is the community detection problem (that is, estimation of z under a suitable loss function) [1, 41].

To illustrate an application of our results, we will consider a sparse mixture detection problem related to two sample testing of SBMs. Suppose we observe two independent SBMs on the same set of n nodes. We have a node of interest, say node i^*, and the statistical problem is to determine whether i^* is in a community with the same members in each observed network. As an example of a practical application, suppose we observe brain networks consisting of n neurons from healthy individuals and diseased individuals, and the scientific question of interest is to determine whether the connectivity of neuron i^* to neighboring neurons differs between healthy and diseased individuals. Another example entails detecting whether the connectivity of gene i^* in a gene regulatory network is different between healthy and diseased individuals (or control and treatment groups, etc.). An extension of this question is to test whether the community structure of the full network is the same between the two observed SBMs rather than exclusively focusing on node i^*. In this line of work, [16] has addressed the two community case.

We now describe the two-sample testing problem with a specific node i^* of interest. Let $z, \sigma \in \{-1, 1\}^{n+1}$ denote community membership for a common set of $n+1$ nodes (considering $n+1$ nodes versus n nodes is only for convenience). The statistician observes the two independent SBMs denoted by A and B, namely the independent random symmetric matrices $A, B \in \{0, 1\}^{(n+1)\times(n+1)}$ where $A_{ii} = B_{ii} = 0$ for all $1 \leq i \leq n+1$ and

$$A_{ij} \sim \begin{cases} \text{Bernoulli}(p) & \textit{if } \sigma_i = \sigma_j, \\ \text{Bernoulli}(q) & \textit{if } \sigma_i \neq \sigma_j, \end{cases}$$

$$B_{ij} \sim \begin{cases} \text{Bernoulli}(p) & \textit{if } z_i = z_j, \\ \text{Bernoulli}(q) & \textit{if } z_i \neq z_j. \end{cases}$$

for $1 \leq i < j \leq n+1$ are drawn independently. Here, $p, q \in (0, 1)$ are parameters. To ensure symmetry, set $A_{ji} := A_{ij}$ and $B_{ji} := B_{ij}$. As mentioned, the problem of interest is to determine whether node i^* has the same community members between the two SBMs A and B. Without loss of generality, let us take $i^* = 1$. To formally state the testing problem, let us define $S_A := \{2 \leq j \leq n+1 : \sigma_j = \sigma_1\}$ and $S_B := \{2 \leq j \leq n+1 : z_j = z_1\}$. Concretely, the testing problem is to test, for $\varepsilon > 0$,

$$H_0 : S_A = S_B, \tag{23}$$

$$H_1 : \frac{|S_A \,\Delta\, S_B|}{n} > \varepsilon. \tag{24}$$

Here, Δ denotes symmetric difference and so $S_A \,\Delta\, S_B = \{2 \leq j \leq n+1 : \sigma_j = \sigma_1, z_j \neq z_1 \text{ or } \sigma_j \neq \sigma_1, z_j = z_1\}$. For ease of notation, let us denote $B(\pi) := \text{Bernoulli}(\pi)$ for $\pi \in (0, 1)$. By sufficiency, only the observed connectivity of node 1 in both SBMs is relevant to the hypothesis testing problem. Under the null hypothesis,

$$\begin{pmatrix} A_{1j} \\ B_{1j} \end{pmatrix} \sim \mathbf{1}_{\{\sigma_1 = \sigma_j\}} B(p) \otimes B(p) + \mathbf{1}_{\{\sigma_1 \neq \sigma_j\}} B(q) \otimes B(q)$$

independently for $2 \leq j \leq n+1$. Under the alternative, there are two cases,

$$\begin{pmatrix} A_{1j} \\ B_{1j} \end{pmatrix} \sim \begin{cases} \mathbf{1}_{\{\sigma_1 = \sigma_j\}} B(p) \otimes B(p) + \mathbf{1}_{\{\sigma_1 \neq \sigma_j\}} B(q) \otimes B(q) & \textit{if } j \in (S_A \,\Delta\, S_B)^c, \\ \mathbf{1}_{\{\sigma_1 = \sigma_j\}} B(p) \otimes B(q) + \mathbf{1}_{\{\sigma_1 \neq \sigma_j\}} B(q) \otimes B(p) & \textit{if } j \in S_A \,\Delta\, S_B. \end{cases}$$

Recall that under the alternative, $n^{-1}|S_A \,\Delta\, S_B| > \varepsilon$. Assuming the two communities in A are of roughly equal size, we can formulate the related sparse mixture detection problem

$$H_0 : \begin{pmatrix} A_{1j} \\ B_{1j} \end{pmatrix} \overset{iid}{\sim} \frac{1}{2} \cdot B(p) \otimes B(p) + \frac{1}{2} \cdot B(q) \otimes B(q), \tag{25}$$

$$\begin{aligned} H_1 : \begin{pmatrix} A_{1j} \\ B_{1j} \end{pmatrix} &\overset{iid}{\sim} (1-\varepsilon) \left[\frac{1}{2} \cdot B(p) \otimes B(p) + \frac{1}{2} \cdot B(q) \otimes B(q) \right] \\ &+ \varepsilon \left[\frac{1}{2} \cdot B(p) \otimes B(q) + \frac{1}{2} \cdot B(q) \otimes B(p) \right]. \end{aligned} \tag{26}$$

Note that the indices run over $2 \leq j \leq n+1$. Adopting an asymptotic perspective, we use the calibration $\varepsilon = n^{-\beta}$ as in (3). Furthermore, set $p = \frac{n^r}{1+n^r}$ and $q = \frac{1}{1+n^r}$ where $r > 0$. Taking $P_n = \frac{1}{2}B(p) \otimes B(p) + \frac{1}{2}B(q) \otimes B(q)$ and $Q_n = \frac{1}{2}B(p) \otimes B(q) + \frac{1}{2}B(q) \otimes B(p)$, we are exactly in the setting of testing (1)-(2).

For the sparse mixture detection problem (25)-(26) it can be shown that the large deviation principle under the null holds with good rate function

$$I(t) = \begin{cases} 0 & \textit{if } t = -r, \\ r & \textit{if } t = r, \\ \infty & \textit{otherwise.} \end{cases}$$

Moreover, it can be checked that Corollary 2.1 can be applied, and doing so yields the detection boundary $\beta^*(r) = \frac{1+1\wedge r}{2}$. Note that I is not convex, and so Theorem 3.1 cannot be invoked. It can also be shown that one can reduce the bivariate observations (A_{1j}, B_{1j}) down to the univariate observations $A_{1j} + B_{1j} \bmod 2$ without losing information (in terms of the detection boundary). The details for the derivation of these results are given in the supplement [30]. We also mention to the reader that the problem (25)-(26) is related to the problem of sparse binary regression [35].

4.8. Detection with side information

Occasionally in applications, there is additional side information that may be useful in testing the global null hypothesis (1)-(2). The testing problem (1)-(2) admits a Bayesian interpretation as the two-groups model [13]. In particular, we have n individual hypotheses $\{H_i\}$ corresponding to each observation X_i. We are testing the global null, i.e. whether the individual hypotheses H_i are null with probability one (that is, $X_i \sim P_n$) for all $1 \leq i \leq n$, against the alternative in which each H_i has probability ε of being non-null (that is, $X_i \sim (1-\varepsilon)P_n + \varepsilon Q_n$). With this interpretation, we are in the setting of multiple testing with n individual hypotheses. This reformulation is usually the situation in which researchers find themselves; for example, each individual hypothesis might correspond to a different gene in a microarray or a different study in a meta-analysis. In such applications, there is usually detailed contextual information attached to each hypothesis and it is desirable to leverage this side information for various hypothesis testing tasks [33, 34, 22, 31, 15, 32, 40].

We consider a stylized problem to investigate how side information affects fundamental statistical limits. Consider a sequence of z-scores W_i for $1 \leq i \leq n$ which are $N(0,1)$ under the null. Under the alternative, an ε fraction of the W_i exhibit elevated mean, that is, $W_i \sim N(\mu, 1)$ where $\mu > 0$. The signal detection problem (1)-(2) with this setup is precisely the sparse normal mixture detection problem considered by Ingster [23] as well as Donoho and Jin [10]. However, suppose for each $1 \leq i \leq n$, we have additional side information (independent of the z-scores) that provides a clue as to whether W_i follows the null distribution or the signal distribution. To represent this side information, we will let A_i denote a Bernoulli random variable in which the outcome $A_i = 1$ denotes evidence that W_i follows the signal distribution and the outcome $A_i = 0$ denotes evidence that W_i follows the null distribution. For simplicity, say that under the null $A_i \sim \text{Bernoulli}(1-p)$ and under the alternative $A_i \sim \text{Bernoulli}(p)$. In other words, the side information correctly identifies both the null and the signal with probability p. While this is a stylized setup, one can think of A_i as the outcome of a well-trained classifier applied to the side information or an expert's judgement derived from existing scientific knowledge. Stated formally, we have the testing problem

$$H_0 : \begin{pmatrix} A_i \\ W_i \end{pmatrix} \overset{iid}{\sim} B(1-p) \otimes N(0,1), \tag{27}$$

$$H_1 : \begin{pmatrix} A_i \\ W_i \end{pmatrix} \overset{iid}{\sim} (1-\varepsilon) B(1-p) \otimes N(0,1) + \varepsilon B(p) \otimes N(\mu, 1) \tag{28}$$

for $1 \leq i \leq n$. Here, we use the notation $B(\pi) = \text{Bernoulli}(\pi)$ for $\pi \in (0,1)$. Adopting the asymptotic perspective, let us calibrate $\varepsilon = n^{-\beta}$ as in (3), let us take $p = \frac{n^r}{1+n^r}$ for $r > 0$, and let us take $\mu = \sqrt{2\rho \log n}$ for $0 < \rho \leq 1$. Furthermore, let us take $P_n = B(1-p) \otimes N(0,1)$ and $Q_n = B(p) \otimes N(\mu, 1)$. Thus, we are in the setting of testing (1)-(2).

It can be shown that the large deviation principle under the null is satisfied with good rate function

$$I(t) = \left[\frac{(t-r+\rho)^2}{4\rho} + r \right] \wedge \frac{(t+r+\rho)^2}{4\rho}.$$

It can be directly checked that $I(t) = \frac{(t-r+\rho)^2}{4\rho} + r$ whenever $t \geq 0$ since $r, \rho > 0$. After an application of Corollary 2.1 and a lengthy calculation, we obtain the detection boundary

$$\beta^*(r, \rho) = \begin{cases} 1 - (\sqrt{(1-r)_+} - \sqrt{\rho})_+^2 & \textit{if } \frac{1-r}{4} < \rho \leq 1, \\ \frac{1}{2} + \rho + \frac{r}{2} & \textit{if } 0 < \rho \leq \frac{1-r}{4}. \end{cases} \tag{29}$$

The detection boundary $\beta^*(r, \rho)$ given by (29) is exactly the Ingster-Donoho-Jin detection boundary when we naively "plug-in" $r = 0$ without care (see

Section 4.1). This is entirely as expected since the case $r = 0$ corresponds to the setting where the sequence of Bernoulli random variables $\{A_i\}$ does not provide strong enough information on the location of the sparse signals. Only the sequence of Gaussian variables $\{W_i\}$ exhibits strong enough signal, and so one can simply throw out the $\{A_i\}$ sequence without loss of power. On the other hand, consider that simply "plugging in" $\rho = 0$ into the detection boundary formula yields $\beta^*(r, 0) = \frac{1+1\wedge r}{2}$. In this setting, the detection boundary is exactly the boundary one would obtain by throwing out the Gaussian variables $\{W_i\}$ (which now have no signal) and only using the Bernoulli sequence $\{A_i\}$ for detection. As seen in the calculations presented in the supplement [30], $\beta^*(r, 0) = \frac{1+1\wedge r}{2}$ is precisely the boundary one obtains through the rate function of the large deviations principle under the null associated with just $\{A_i\}$.

In the intermediate regimes $0 < r < 1$ and $0 < \rho < 1$, the detection boundary $\beta^*(r, \rho)$ is larger than the Ingster-Donoho-Jin detection boundary and larger than the boundary $\frac{1+1\wedge r}{2}$ corresponding to using only the Bernoulli sequence $\{A_i\}$. In words, using both the Gaussian and the Bernoulli data yields higher detection boundaries (meaning we are able to detect weaker signals in sparser settings) compared to using only the Gaussian data or only the Bernoulli data. Furthermore, $\beta^*(r, \rho)$ gives a precise description of how the signal strengths r and ρ in the Bernoulli and Gaussian sequences relate to one another and affect the phase transition.

4.9. Curie-Weiss model

We briefly introduce a sparse mixture detection problem in the larger context of Ising models. There is an existing literature focused on inferential tasks given multiple independent and identically distributed samples from an unknown probabilistic graphical model, such as graph selection [38, 3] and property testing/goodness-of-fit testing [7, 4]. Note that in the context of the sparse mixture detection problems (1)-(2), we have n observations in which possibly $(1 - \varepsilon)n$ are drawn from the Ising model P_n and εn are drawn from the Ising model Q_n. The problem, of course, is to detect the presence of observations drawn from the Ising model Q_n.

We will consider the sample space $\mathcal{X} = \bigcup_{n=1}^{\infty} \{-1, 1\}^n$. Consider the testing problem (1)-(2) where for $x \in \mathcal{X}$,

$$p_n(x) = \frac{1}{Z_N(\theta, 0)} \exp\left(\frac{\theta}{N} \sum_{1 \le i < j \le N} x_i x_j \right) \cdot \mathbf{1}_{\{x \in \{-1,1\}^N\}},$$

$$q_n(x) = \frac{1}{Z_N(\theta, \mu)} \exp\left(\frac{\theta}{N} \sum_{1 \le i < j \le N} x_i x_j + \theta\mu \sum_{i=1}^{N} x_i \right) \cdot \mathbf{1}_{\{x \in \{-1,1\}^N\}}$$

where $N = \lceil \log n \rceil$. Further, $\theta > 0$ and $\mu > 0$ are parameters we take to be fixed and unchanging with n. In the parlance of statistical mechanics, p_n and q_n are the Curie-Weiss model on N particles. Each particle takes one of two states $x_i \in \{-1, 1\}$ for all $1 \leq i \leq N$. Additionally q_n models the existence of an external magnetic field with strength μ and p_n models the absence of an external magnetic field. The quantity $Z_N(\theta, \mu)$ is the normalizing constant or "partition function" in the statistical mechanics convention. Note that under both p_n and q_n, all of the random variables $x_1, \dots, x_N$ are all correlated with each other and the parameter θ controls the strength of this correlation. Intuitively, the correlation is strong when θ is large and the correlation is weak when θ is small. Additionally, the parameter μ modulates the probability that each particle takes value 1 instead of -1. Further background about the Curie-Weiss model (and related Ising models) can be found in Chapter 2 of [36].

A large deviation principle under the null can be established and the associated rate function can be derived. Interestingly, well-known phase transitions in the behavior of the Curie-Weiss model appear to have an effect on the rate function and thereby have an effect on the detection boundary. As it is outside the scope of the present article, we do not undertake a full analysis to obtain the detection boundary nor evaluate the applicability of Gao and Ma's statistic. The situation is discussed in more detail in the supplement [30].

5. Discussion

We have offered a unified perspective on deriving phase transitions in general sparse mixture detection problems via the large deviations theory. The fundamental object determining the phase transition is the the rate function associated to the large deviation principle of the normalized log likelihood ratios. The core phenomenon behind the phase transition lies in the asymptotics of the Hellinger distance between P_n and $(1 - n^{-\beta})P_n + n^{-\beta}Q_n$ as identified by Cai and Wu [6]; the large deviations theory provides suitable machinery to relate Hellinger asymptotics to phase transitions beyond the univariate sparse mixture case.

Additionally, we have obtained sufficient conditions on the rate function to guarantee the optimality of a sequence of tests based on a Higher Criticism type statistic formulated by Gao and Ma (Section 3.2 of [18]). This statistic HC_n^* adapts to the signal sparsity β and can be used "off-the-shelf". Moreover, as we discussed in Section 3, computation of HC_n^* need not require full knowledge of the signal distributions $\{Q_n\}$. Rather, in some problems it may suffice to consider a certain statistic of the data; of course, considerations will vary on a problem-to-problem basis.

We imagine that the large deviations perspective offered here will be useful in deriving phase transitions in more complicated and structured sparse mixture detection problems beyond what can be derived with the existing univariate theory. Further, we imagine that Gao and Ma's testing statistic will be practically useful in light of Theorem 3.1. We conclude with a few remarks.

5.1. Deriving large deviation principles and rate functions

The main results regarding detection boundaries we have presented only specify how the detection boundary is determined by the rate function when the normalized log likelihood ratios satisfy a suitable large deviation principle. These results have nothing to say about how to deduce a large deviation principle and calculate the associated rate function. This is not so surprising given the broad setting and the fundamental role of the rate function. Indeed, the main technical work in specific problems is to deduce the large deviation principle and the associated rate function. In the few examples we presented in Section 4, we have illustrated a small number of techniques useful to establishing the large deviation principle. Indispensable are the contraction principle (Theorem 6.1), exponential equivalence (Definition 6.1), and the indistinguishability of the large deviation principle for exponentially equivalent probability measures (Theorem 6.2). Lemma 3 in [6] was also quite useful in our examples in calculating the order of some exponential integrals.

5.2. Sparse mixture of exponentials: necessity of a tail condition

Unfortunately, the tail condition (6) in Theorem 2.1 can preclude calculation of a detection boundary in some problems. For example, consider the testing problem (1)-(2) with calibration (3) and $P_n = \text{Exponential}(1)$, $Q_n = \text{Exponential}(1+n^r)$ for $r > 0$ under the scale parameterization (i.e. P_n has mean 1 and Q_n has mean $1+n^r$).

It can be shown that $\limsup_{n\to\infty} \frac{1}{\log n} \log E\left[\left(\frac{q_n}{p_n}(X)\right)^\gamma\right] = \infty$ for all $\gamma > 1$ where $X \sim P_n$, and so the tail condition (6) fails to hold (see the supplement [30] for details). However, the testing problem indeed exhibits a detection boundary; Corollary 4.4 of [9] indicates that $\beta^* = \frac{1+1\wedge r}{2}$. One might be tempted to argue that the tail condition is simply too strong and that the conclusions of Theorem 2.1 and Corollary 2.1 might still hold. However, it can be shown that this is not the case. Specifically, it can be shown that the large deviation principle under the null is satisfied with rate function I with $I(t) = t + r$ for $t \geq -r$ and $I(t) = \infty$ otherwise (see the supplement [30] for details). When $r \leq \frac{1}{2}$, the conclusions of Theorem 2.1 and Corollary 2.1 suggest the detection boundary $\beta^* = 1 - r$, which does not match the detection boundary of [9]. Thus, a tail condition like (6) is indeed necessary for our large deviations approach. It is not clear to us if the sparse exponential mixture testing problem can be treated through alternative means within the large deviations framework.

5.3. Further generalizations

Theorem 2.1 only presents upper and lower bounds on $\overline{\beta}^*$ and $\underline{\beta}^*$ respectively. In this paper, we were only interested in when these bounds meet (Corollary 2.1). It is an open problem to give tight characterizations of $\overline{\beta}^*$ and $\underline{\beta}^*$. Likewise, it is of interest to furnish an example where normalized log likelihood ratios satisfy a large deviation principle under the null and where $\underline{\beta}^*$ and $\overline{\beta}^*$ do not meet.

Finally, it is of interest to develop results in the setting where the observations are correlated rather than independent and identically distributed. The approach of characterizing the Hellinger asymptotics is no longer tenable as this method exploited the tensorization property of the Hellinger distance over product measures. Both problems of determining the phase transitions and developing optimal procedures are open. In the normal mixture setting, Hall and Jin [20] develop the Innovated Higher Criticism. Remarkably, Hall and Jin show that signal detection can actually be easier in some cases; the independent noise case is statistically the hardest. We refer the interested reader to further discussion in that paper as well as the review article [27].

6. Auxiliary definitions and results

We state below some instrumental results and definitions from the large deviations theory and convex analysis. We follow the presentation found in Chapter 4 of [8] as well as the presentation of [37].

Theorem 6.1 (Contraction principle, Theorem 4.2.1 - [8])**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Hausdorff topological spaces and $f : \mathcal{X} \to \mathcal{Y}$ a continuous function. Consider a good rate function $I : \mathcal{X} \to [0, \infty]$.*

(a) For each $y \in \mathcal{Y}$, define $I'(y) := \inf\{I(x) : x \in \mathcal{X},\, y = f(x)\}$. Then I' is a good rate function on $\mathcal{Y}$. Here, we adopt the convention $\inf \emptyset = \infty$.

(b) If I controls the large deviation principle with a family of probability measures $\{\mu_n\}$ on $\mathcal{X}$, then I' controls the large deviation principle associated with the family of probability measures $\{\mu_n \circ f^{-1}\}$ on $\mathcal{Y}$.

Definition 6.1 (Exponential equivalence, Definition 4.2.10 - [8])**.** Let $(\mathcal{Y}, d)$ be a metric space. The probability measures $\{\mu_n\}$ and $\{\widetilde{\mu}_n\}$ are called *exponentially equivalent* with respect to speed $\{a_n\}$ if there exist probability spaces $\{\Omega, \mathcal{B}_n, P_n\}$ and two families of $\mathcal{Y}$-valued random variables $\{Z_n\}$ and $\{\widetilde{Z}_n\}$ with joint laws $\{P_n\}$ and marginals $\{\mu_n\}$ and $\{\widetilde{\mu}_n\}$ respectively such that the following condition is satisfied. For each $\delta > 0$, the set $\{\omega \in \Omega : d(\widetilde{Z}_n, Z_n) > \delta\}$ is $\mathcal{B}_n$ measurable, and

$$\limsup_{n \to \infty} a_n \log P_n\left(d(\widetilde{Z}_n, Z_n) > \delta\right) = -\infty.$$

Here, $\{a_n\}$ is a sequence of reals with $a_n \to 0$.

Theorem 6.2 (Indistinguishability of the large deviation principle, Theorem 4.2.13 - [8])**.** *If a large deviation principle with speed $\{a_n\}$ and good rate function I holds for the probability measures $\{\mu_n\}$, which are exponentially equivalent to $\{\widetilde{\mu}_n\}$, then the same large deviation principle holds for $\{\widetilde{\mu}_n\}$.*

Definition 6.2. Let $f : \mathbb{R} \to [-\infty, \infty]$ be a convex function. A real number x^* is said to be a *subgradient* of f at x if $f(z) \geq f(x) + x^* \cdot (z - x)$ for all $z \in \mathbb{R}$. The set of all subgradients of f at x is called the *subdifferential of f at x*, and is denoted by $\partial f(x)$. The mapping $x \mapsto \partial f(x)$ is called the *subdifferential* of f. If $\partial f(x)$ is not empty, f is said to be *subdifferentiable* at x.

Theorem 6.3 (Theorem 24.1 - [37])**.** *Let $f : \mathbb{R} \to [-\infty, \infty]$ be a closed proper convex function. For convenience, extend the right and left derivatives f'_+ and f'_- beyond the interval D on which f is finite as follows. For points to the right of D, set f'_+ and f'_- equal to ∞. For points the left of D, set f'_+ and f'_- equal to $-\infty$. Then f'_+ and f'_- are increasing functions on $\mathbb{R}$, finite on the interior of D, such that*

$$f'_+(z_1) \leq f'_-(x) \leq f'_+(x) \leq f'_-(z_2)$$

when $z_1 < x < z_2$. Moreover, for every x, we have $\lim_{z \downarrow x} f'_+(z) = f'_+(x)$, $\lim_{z \uparrow x} f'_+(z) = f'_-(x)$, $\lim_{z \downarrow x} f'_-(z) = f'_+(x)$, and $\lim_{z \uparrow x} f'_-(z) = f'_-(x)$.

Acknowledgments

The author thanks Chao Gao for suggesting this problem, providing detailed and helpful comments on multiple drafts, and offering substantial encouragement.

Supplementary Material

Supplement to "Statistical limits of sparse mixture detection"
(doi: 10.1214/22-EJS2053SUPP; .pdf). This supplement includes proofs for the results presented in the main text. Additionally, this supplement contains details for the calculations relevant to the examples discussed in the main text.

References

[1] Abbe, E. (2018). Community Detection and Stochastic Block Models: Recent Developments. *J. Mach. Learn. Res.* **18** 1–86. MR3827065

[2] Arias-Castro, E. and Wang, M. (2015). The Sparse Poisson Means Model. *Electron. J. Statist.* **9** 2170–2201. MR3406276

[3] Barber, R. F. and Drton, M. (2015). High-Dimensional Ising Model Selection with Bayesian Information Criteria. *Electron. J. Statist.* **9** 567–607. MR3326135

[4] Bezáková, I., Blanca, A., Chen, Z., Štefankovič, D. and Vigoda, E. (2020). Lower Bounds for Testing Graphical Models: Colorings and Antiferromagnetic Ising Models. *J. Mach. Learn. Res.* **21** 1–62. MR4071208

[5] Cai, T. T., Jeng, X. J. and Jin, J. (2011). Optimal Detection of Heterogeneous and Heteroscedastic Mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 629–662. MR2867452

[6] Cai, T. T. and Wu, Y. (2014). Optimal Detection of Sparse Mixtures against a given Null Distribution. *IEEE Trans. Inform. Theory* **60** 2217–2232. MR3181520

[7] Canonne, C. L., Diakonikolas, I., Kane, D. M. and Stewart, A. (2020). Testing Bayesian Networks. *IEEE Trans. Inform. Theory* **66** 3132–3170. MR4089774

[8] Dembo, A. and Zeitouni, O. (2010). *Large Deviations Techniques and Applications*, second ed. *Stochastic Modelling and Applied Probability* **38**. Springer-Verlag, Berlin Heidelberg. MR2571413

[9] Ditzhaus, M. (2019). Signal Detection via Phi-divergences for General Mixtures. *Bernoulli* **25** 3041–3068. MR4003573

[10] Donoho, D. and Jin, J. (2004). Higher Criticism for Detecting Sparse Heterogeneous Mixtures. *Ann. Statist.* **32** 962–994. MR2065195

[11] Donoho, D. and Jin, J. (2015). Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects. *Statist. Sci.* **30** 1–25. MR3317751

[12] Donoho, D. L. and Kipnis, A. (2022). Higher criticism to compare two large frequency tables, with sensitivity to possible rare and weak differences. *Ann. Statist.* **50** 1447–1472. MR4441127

[13] Efron, B. (2008). Microarrays, Empirical Bayes and the Two-Groups Model. *Statist. Sci.* **23** 1–22. MR2431866

[14] Efron, B. (2010). *Large-scale inference. Institute of Mathematical Statistics (IMS) Monographs* **1**. Cambridge University Press, Cambridge. MR2724758

[15] Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G. and Kong, A. (2008). Unsupervised Empirical Bayesian Multiple Testing with External Covariates. *Ann. Appl. Stat.* **2** 714–735. MR2524353

[16] Gangrade, A., Nazer, B. and Saligrama, V. (2018). Two-Sample Testing Can Be as Hard as Structure Learning in Ising Models: Minimax Lower Bounds. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 6931–6935. ISSN: 2379-190X.

[17] Gao, C. and Ma, Z. (2021). Minimax Rates in Network Analysis: Graphon Estimation, Community Detection and Hypothesis Testing. *Statist. Sci.* **36** 16–33. MR4194201

[18] Gao, C. and Ma, Z. (2022). Testing Equivalence of Clustering. *Ann. Statist.* **50** 407–429. MR4382022

[19] Gao, Z. and Stoev, S. (2021). *Concentration of Maxima and Fundamental Limits in High-Dimensional Testing and Inference*, 1 ed. *SpringerBriefs in Probability and Mathematical Statistics*. Springer, Cham. MR3967073

[20] Hall, P. and Jin, J. (2010). Innovated Higher Criticism for Detecting Sparse Signals in Correlated Noise. *Ann. Statist.* **38** 1686–1732. MR2662357

[21] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic Blockmodels: First Steps. *Soc. Networks* **5** 109–137. MR0718088

[22] Ignatiadis, N. and Huber, W. (2021). Covariate Powered Cross-Weighted Multiple Testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 720–751. MR4319999

[23] Ingster, Y. (1997). Some Problems of Hypothesis Testing Leading to Infinitely Divisible Distributions. *Math. Methods Statist.* **6** 47–69. MR1456646

[24] Jin, J. (2003). Detecting and Estimating Sparse Mixtures, PhD Thesis, Stanford University. MR2704645

[25] Jin, J. (2004). Detecting a Target in Very Noisy Data from Multiple Looks. In *A Festschrift for Herman Rubin. IMS Lecture Notes Monogr. Ser.* **45** 255–286. Inst. Math. Statist., Beachwood, OH. MR2126903

[26] Jin, J. (2009). Impossibility of Successful Classification When Useful Features Are Rare and Weak. *Proc. Natl. Acad. Sci. USA* **106** 8859–8864. MR2520682

[27] Jin, J. and Ke, Z. T. (2016). Rare and Weak Effects in Large-Scale Inference: Methods and Phase Diagrams. *Statist. Sinica* **26** 1–34. MR3468343

[28] Jin, J., Ke, Z. T. and Wang, W. (2017). Phase Transitions for High Dimensional Clustering and Related Problems. *Ann. Statist.* **45** 2151–2189. MR3718165

[29] Jin, J., Zhang, C.-H. and Zhang, Q. (2014). Optimality of Graphlet Screening in High Dimensional Variable Selection. *J. Mach. Learn. Res.* **15** 2723–2772. MR3270749

[30] Kotekal, S. Supplement to "Statistical Limits of Sparse Mixture Detection".

[31] Lei, L. and Fithian, W. (2018). AdaPT: An Interactive Procedure for Multiple Testing with Side Information. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 649–679. MR3849338

[32] Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J. and Thomas, D. C. (2007). Hierarchical Bayes Prioritization of Marker Associations from a Genome-Wide Association Scan for Further Investigation. *Genet. Epidemiol.* **31** 871–882.

[33] Li, A. and Barber, R. F. (2017). Accumulation Tests for FDR Control in Ordered Hypothesis Testing. *J. Amer. Statist. Assoc.* **112** 837–849. MR3671774

[34] Li, A. and Barber, R. F. (2019). Multiple Testing with the Structure-Adaptive Benjamini-Hochberg Algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 45–74. MR3904779

[35] Mukherjee, R., Pillai, N. S. and Lin, X. (2015). Hypothesis Testing for High-Dimensional Sparse Binary Regression. *Ann. Statist.* **43** 352–381. MR3311863

[36] Mézard, M. and Montanari, A. (2009). *Information, Physics, and Computation. Oxford Graduate Texts.* Oxford University Press, Oxford. MR2518205

[37] Rockafellar, R. T. (1970). *Convex Analysis. Princeton Mathematical Series, No. 28.* Princeton University Press, Princeton, NJ. MR0274683

[38] Santhanam, N. P. and Wainwright, M. J. (2012). Information-Theoretic Limits of Selecting Binary Graphical Models in High Dimensions. *IEEE Trans. Inform. Theory* **58** 4117–4134. MR2943079

[39] Vielva, P. (2010). A Comprehensive Overview of the Cold Spot. *Adv. Astron.* **2010**.

[40] Zablocki, R. W., Schork, A. J., Levine, R. A., Andreassen, O. A., Dale, A. M. and Thompson, W. K. (2014). Covariate-Modulated Local False Discovery Rate for Genome-Wide Association Studies. *Bioinformatics* **30** 2098–2104. MR3743296

[41] Zhang, A. Y. and Zhou, H. H. (2016). Minimax Rates of Community Detection in Stochastic Block Models. *Ann. Statist.* **44** 2252–2280. MR3546450

[42] Statistical Problems in the Theory of Stochastic Processes - Encyclopedia of Mathematics. *Encyclopedia of Mathematics*.