Optimal Screening and Discovery of Sparse Signals with Applications to Multistage High-throughput Studies

T. Tony Cai

University of Pennsylvania, Philadelphia, USA

Wenguang Sun

University of Southern California, Los Angeles, USA

Summary. A common feature in large-scale scientific studies is that signals are sparse and it is desirable to significantly narrow down the focus to a much smaller subset in a sequential manner. In this paper, we consider two related data screening problems: One is to find the smallest subset such that it virtually contains all signals and another is to find the largest subset such that it essentially contains only signals. These screening problems are closely connected to but distinct from the more conventional signal detection or multiple testing problems. We develop data-driven screening procedures which control the error rates with near optimality properties and study how to design the experiments efficiently to achieve the goals in data screening. A class of new phase diagrams is developed to characterize the fundamental limitations in simultaneous inference. An application to multistage high-throughput studies is given to illustrate the merits of the proposed screening methods.

Keywords: adaptive design, classification, data screening, false discovery rate, false negative rate, phase transition

1. Introduction

A challenging and important problem in large-scale scientific studies is to recover sparse signals from massive amount of data. Multistage design provides a cost-effective way to glean significance from data by adaptively reducing a large set of variables to a much smaller subset in a sequential manner. The general strategy is to use information acquired from the previous measurements to adjust the subsequent measurements and focus resources on study units that are more likely to contain signals of interest. For example, Satagopan et al. (2004) proposed a two-stage design for genome-wide association studies and showed that the new design provides a substantial reduction in the study costs for a minimal loss of power compared to single-stage approaches. Haupt et al. (2009, 2011) proposed the distilled sensing method for large-scale signal processing problems. It was shown that with a fixed study cost, the distilled sensing method requires remarkably weaker condition for reliable recovery of sparse signals. In geostatistical analysis, Bloma et al. (2002) showed that a two-stage adaptive sampling approach leads to great savings in study costs. In the context of microarray, RNA-seq, and protein array experiments, Müller et al. (2004) and Rossell and Müller (2013) proposed simulation-based algorithms for the design and analysis of multi-stage experiments under a class of prespecified utility functions. Optimal stopping rules in multi-stage experiments are also studied by Lai (2000); Bartroff (2007); Durrieu and Briollais (2009) for various applications.

The analysis of large-scale multistage experiments poses new challenges that are not present in conventional small-scale and single-stage analyses. One critical issue is the control of decision errors at various stages in the screening process. At each stage of screening, both

false positive and false negative decisions may occur: a high false positive rate will increase the study costs in the next stage and may result in misleading scientific conclusions; meanwhile, since undetected signals will not be revisited in subsequent analyses, a high false negative rate may lead to an overall inefficient design and inevitable financial losses. To illustrate the key issues, we discuss in detail an important application, the high-throughput screening (HTS) of chemical compounds in drug discovery. The terms in HTS are adopted in later sections to facilitate the presentation, but the discussions apply to more general settings.

HTS is a large-scale hierarchical process (Figure 1) that conducts millions of chemical tests in multiple stages to identify active compounds and generate candidates for drug design and development. In the initial *primary screen*, an integrated robot system is used to rapidly collect data on a large library of chemical compounds. The compounds with desirable effect size (labeled "hits") will be followed up by a *secondary screen* which collects additional data on the narrowed subset. The results are further refined by a careful analysis to confirm their statistical significance and biological relevance. The confirmed hits with an established biomedical activity are termed as "leads," which may be developed into drug candidates and used for clinical testing. The advances in robotics and parallel data processing technique have dramatically increased the throughput, with more than 100,000 compounds sampled and measured per day in some *ultra high-throughput* experiments (Agresti et al., 2010). However, few analytical tools are available for dealing with such massive data sets, see Bleicher et al. (2003); Malo et al. (2006); Birmingham et al. (2009); Zhang (2011) for reviews of statistical methods currently used in the analysis of HTS. Moreover, the inferential process in most HTS analyses is mainly based on informal "rules of thumb," which are proposed for small studies with only a few dozens chemical compounds. The overwhelming number of targets in modern HTS has resulted in soaring costs in clinical testing and declined drug approval rate (Dove et al., 2003). It is imperative to develop flexible and cost-effective strategies for large-scale multi-stage inference problems to control the error rates accurately and optimize the discovery process.

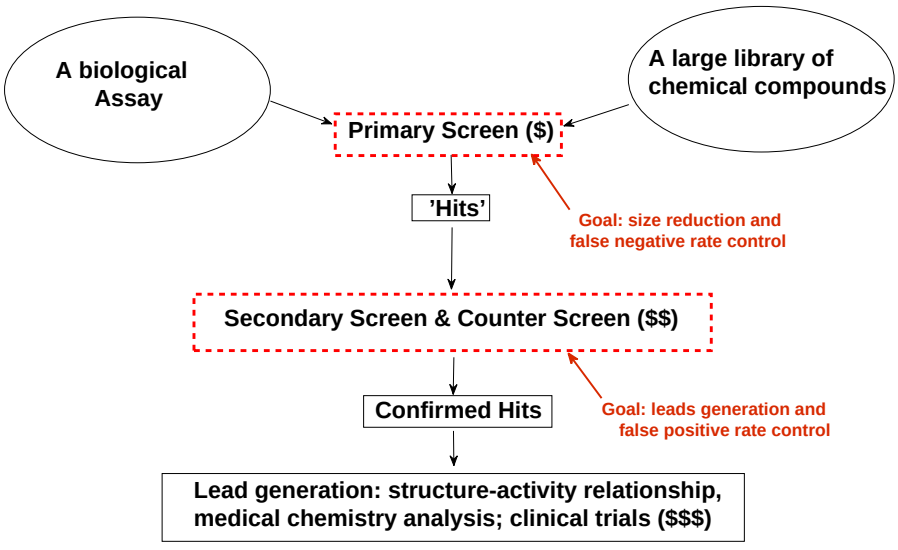

Fig. 1. Flowchart for HTS and drug development process.

Now we discuss two types of data screening problems which may arise at different stages of HTS. In primary screens, the goal is to reduce the size of the library significantly to meet the laboratory constraints such as capacity limitations in the more expensive secondary screens. In the current practice of HTS, compounds are measured with one or more replicates and a small fraction (say, top 1%) with highest activities are selected as "hits" to enter secondary screens. The important tasks at this stage include: (i) to construct a subset with negligible false negative rate (since undetected compounds will not be revisited); and (ii) to determine the number of replicates to ensure that the subset achieves a significant size reduction. In secondary screens, the goal is to confirm the "hits" selected by primary screens and use them to generate "leads." The far more complex and costly leads generation process calls for precise control of the false positive rate. The current practice in HTS is to obtain the z-scores using a few replicates and then threshold the z-scores at a pre-specified value (e.g. selecting cases with $|z| > 3$, Malo et al., 2006). However, these ad-hoc rules with prefixed thresholds provide no control of probabilistic error rates and can be either too conservative or too liberal. At this stage, the important tasks include (i) to construct a subset with negligible false positive rate; (ii) to determine the number of replicates to ensure that some useful signals can be reliably identified for leads generation.

The statistical issues in HTS commonly exists in other large-scale multistage experiments. In summary, the accurate and effective signal recovery via a multistage analysis requires the study of two inter-related data screening problems:

(i) to find the smallest subset such that it nearly contains all signals; and

(ii) to find the largest subset such that it virtually contains only signals.

In both screening problems, we need to address two issues: how to control the decision errors accurately and how to design the sample size efficiently.

The error control issue in multi-stage and sequential testing problems has been investigated in Lin (2006), Dmitrienko et al. (2007), Benjamini and Heller (2007), Goeman and Mansmann (2008), Yekutieli (2008), and Posch et al. (2009), among others. Blanchard and Geman (2005), Meinshausen (2008) and Goeman and Solari (2010) considered the control of family wise error rate in hierarchical variable selection/testing problems. However, these works essentially focus on the control of the false positive rate, and in particular the adjustment of statistical significance in hierarchical inference. The control of the false negative rate has not been considered and the issues on sample size design still remain unknown. Fan and Lv (2008), Wasserman and Roeder (2009) and Ji and Jin (2012) proposed multi-stage methods for high-dimensional regression problems. However, their settings are very different from ours. In addition, the issues on decision error control, efficient design and optimal subset construction have not been established by existing works.

The goals of this article include: (i) to develop data-driven screening procedures which control the error rates with near optimality properties, and (ii) to study how the multistage experiments can be designed efficiently to achieve the goals in data screening. We formulate a decision-theoretic framework for large-scale inference problems and develop asymptotically optimal data-driven screening procedures that control the false positive and false negative rates, respectively. To address the related design problems, we employ the technique of phase diagram (e.g. Donoho and Jin, 2004; Cai et al., 2007) to study the phase transition in optimal screening. The resulting *classification boundary*, *discovery boundary* and *screening boundary*, which characterize the precise conditions under which respective goals in data screening are achievable, lead to useful formulae for calculating the minimum number of replicates needed at different stages of HTS. We establish the optimality of the proposed data-driven procedures

by showing that they successfully attain the respective phase transition boundaries in a two-point normal mixture model. Compared to other popular screening schemes such as the distilled sensing method, the proposed data-driven procedures can be implemented for a more general class of mixture models and are capable of reducing the size of a large data set in a much faster and more reliable way. The numerical results demonstrate that the proposed methods control the error rates at the desired level with significantly improved power compared to existing methods. An application to HTS is also given to illustrate the merits of the proposed screening methods.

Under the two-component sparse mixture models, the study of optimality via phase transition has been limited to the *global inference* problems such as signal detection and sparsity estimation. See, for example, Donoho and Jin (2004); Meinshausen and Rice (2006); Cai et al. (2007); Cai and Wu (2014). This article develops new optimality theory for a class of important and closely related *simultaneous inference* problems including classification, signal discovery and data screening. Aiming to make many decisions at finer (individual) levels, the analysis in simultaneous inference involves very different techniques compared to that in global inference: one requires greater precision in decision making and needs to control the inflation of both false positive and negative errors. The main theoretical contribution of our work is the development of different phase diagrams, including the new classification, discovery and screening boundaries, which together characterize the fundamental limitations and hence the optimality benchmarks in simultaneous inference. The main methodological contributions include new results on false negative rate control and rates of screening levels; we show how the thresholding sequences may be chosen so that the phase transition boundaries can be successfully attained.

The rest of the article is organized as follows. Section 2 develops oracle and data-driven screening procedures for analysis of multistage experiments. Section 3 studies the phase transition in optimal screening and derives the sample size formulae to address related design problems in practice. We employ the phase transition theory to evaluate the effectiveness of different screening strategies in Section 4. Simulation studies are conducted in Section 5 to investigate the numerical performance of the proposed methods. An application to HTS is presented in Section 6 and a brief discussion is given in Section 7. The main theorems are proved in Section 8 and the proofs of some additional technical results are given in the supplemental material.

2. Optimal Screening: Theoretical Framework and Data-driven Procedures

We study the data screening problem in a decision theoretic framework. Let $X_1, \cdots, X_n$ be observations from a random mixture model

$$X_i|\theta_i \sim (1-\theta_i)F_0 + \theta_i F_1, \tag{2.1}$$

where $\theta_1, \cdots, \theta_n$ are independent Bernoulli(ϵ_n) random variables, and F_0 and F_1 are the null and non-null distributions, respectively. Here $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n) \in \{0,1\}^n$ denotes the true states of nature, with $\theta_i = 0$ indicating a null case and $\theta_i = 1$ indicating a signal of interest. The random mixture model (2.1) provides a powerful and convenient framework for large-scale inference problems and has been widely used in the literature; see, for example, Efron et al. (2001), Storey (2002), and Genovese and Wasserman (2002). There are two goals in data screening: One is to find the smallest subset such that it virtually contains the set of all signals and another is to find the largest subset such that it essentially contains only signals. Intuitively, it is clear that the difficulty of achieving these goals depends on the "distance" between the null distribution F_0 and non-null distribution F_1.

An important special case of (2.1) is the following two-point normal mixture model:

$$X_i|\theta_i \sim (1-\theta_i)N(0,1) + \theta_i N(\mu_n, \sigma^2), \quad i = 1, \cdots, n. \tag{2.2}$$

This simple and more concrete model is suitable for many applications and has been extensively studied in the literature. It has played a fundamental role in understanding the detection and classification problems in high-dimensional sparse inference; see, for example, Ingster (1998), Donoho and Jin (2004, 2006) and Cai et al. (2007).

Our methodological and theoretical development is divided into two steps. We first consider the general random mixture model (2.1) and propose oracle and data-driven screening procedures for global error rate control and study their optimality properties in the rest of this section. The focus is then turned to the efficient design of a multistage experiment under the more concrete two-point normal mixture model (2.2) in Section 3, with the aim of finding the *minimum* number of replicates with which it is possible to achieve the goals in data screening. These two steps together give the complete solution and both are indispensable: the conclusions would be invalid without effective error rate control, and the study would not attain the desired power without a reasonable sample size. In practice, one can first use the simple model (2.2) to determine the sample size, and then implement the screening procedures under the general model (2.1) to analyze the collected data without the parametric assumptions. This framework will be illustrated in Section 6.

2.1. Problem formulation

Consider a decision rule $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_n) \in \{0,1\}^n$, where $\delta_i = 1$ if case i is selected as an interesting case and $\delta_i = 0$ otherwise. The index sets of true signals, null cases and selected cases are denoted by $\mathcal{I} = \{i : \theta_i = 1\}$, $\mathcal{N} = \{i : \theta_i = 0\}$ and $\mathcal{S}_{\boldsymbol{\delta}} = \{i : \delta_i = 1\}$, respectively. Then the expected size of $\mathcal{S}_{\boldsymbol{\delta}}$ can be decomposed as

$$E[\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}})] = E[\mathrm{Card}(\mathcal{I} \cap \mathcal{S}_{\boldsymbol{\delta}})] + E[\mathrm{Card}(\mathcal{N} \cap \mathcal{S}_{\boldsymbol{\delta}})] = \mathrm{ETP}_{\boldsymbol{\delta}} + \mathrm{EFP}_{\boldsymbol{\delta}},$$

where $\mathrm{ETP}_{\boldsymbol{\delta}}$ and $\mathrm{EFP}_{\boldsymbol{\delta}}$ represent the **e**xpected numbers of **t**rue **p**ositives and **e**xpected numbers of **f**alse **p**ositives of $\boldsymbol{\delta}$, respectively. Let α_n and α_n' be two sequences of positive numbers converging to 0 slowly. A decision rule $\boldsymbol{\delta}$ is *valid for false positive rate (FPR) control* if

$$\mathrm{FPR}_{\boldsymbol{\delta}} = \mathrm{EFP}_{\boldsymbol{\delta}} / E[\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}})] \le \alpha_n, \tag{2.3}$$

and *valid for missed discovery rate (MDR, or non-discovery rate, NDR) control* if

$$\mathrm{MDR}_{\boldsymbol{\delta}} = 1 - \mathrm{ETP}_{\boldsymbol{\delta}} / E[\mathrm{Card}(\mathcal{I})] \le \alpha_n'. \tag{2.4}$$

Remark 1. The FPR (also referred to as the marginal false discovery rate, mFDR) is asymptotically equivalent to the well-known false discovery rate (FDR, Benjamini and Hochberg, 1995) under independence (Genovese and Wasserman, 2002) and weak dependence (Storey et al., 2004). The MDR, also called the "missed rate" (Taylor et al., 2005), is equivalent to the non-discovery rate (NDR) in Haupt et al. (2011) under the random mixture model (2.1). An alternative measure to the MDR is the false non-discovery rate or false negative rate (FNR, Genovese and Wasserman, 2002; Sarkar, 2004). Under the sparse signal settings, the FNR is close to zero and hence less sensitive. There is no essential difference between the FPR and FDR in large-scale testing problems; the use of FPR is mainly for technical considerations – its simplicity makes it possible to establish sharp optimality in data screening via an equivalent weighted classification problem (see proof of Theorem 1).

Let $\mathcal{D}_{d,\alpha_n}$ and $\mathcal{D}_{s,\alpha'_n}$ denote the collections of all screening procedures fulfilling conditions (2.3) and (2.4), respectively. We call a procedure *valid* if the FPR/MDR is controlled at the nominal level, and *optimal* if it constructs the largest (smallest) subset among all valid procedures. Following the standard notation in decision theory (e.g., Berger (1985)), we denote the optimal FPR procedure by $\boldsymbol{\delta}_d^\pi$ (subscript "d" indicates "discovery" and superscript "π" indicates "optimal"), which satisfies

$$\boldsymbol{\delta}_d^\pi \in \mathcal{D}_{d,\alpha_n} \text{ and } E\{\text{Card}(\mathcal{S}_{\boldsymbol{\delta}_d^\pi})\} \geq E\{\text{Card}(\mathcal{S}_{\boldsymbol{\delta}})\} \text{ for all } \boldsymbol{\delta} \in \mathcal{D}_{d,\alpha_n}. \tag{2.5}$$

The optimal MDR procedure, denoted by $\boldsymbol{\delta}_s^\pi$ (subscript "s" indicates "screening"), satisfies

$$\boldsymbol{\delta}_s^\pi \in \mathcal{D}_{s,\alpha'_n} \text{ and } E\{\text{Card}(\mathcal{S}_{\boldsymbol{\delta}_s^\pi})\} \leq E\{\text{Card}(\mathcal{S}_{\boldsymbol{\delta}})\} \text{ for all } \boldsymbol{\delta} \in \mathcal{D}_{s,\alpha'_n}. \tag{2.6}$$

Optimal FPR and MDR procedures will be derived in the next section.

REMARK 2. The formulation of (2.5) and (2.6) naturally extends the optimality concepts in single hypothesis testing, where the Neyman-Pearson lemma provides the most powerful test at a given nominal level. A similar type of optimality theory is developed in the next section, where test size and power in simple hypothesis testing are extended to the global error rate (FPR or MDR) and the expected subset size in data screening correspondingly.

2.2. Oracle and adaptive screening procedures

We begin by considering the random mixture model (2.1) in the oracle setting where all model parameters are assumed to be known. Let f_0 and f_1 denote the null and non-null densities, and $f = (1-\epsilon_n)f_0 + \epsilon_n f_1$ the marginal density. Define $T_i^\pi = \frac{(1-\epsilon_n)f_0(X_i)}{f(X_i)}$ [the local false discovery rate (Lfdr, Efron et al., 2001)] and $\boldsymbol{T}^\pi = (T_1^\pi, \cdots, T_n^\pi)$. Consider a decision rule of the form $\boldsymbol{\delta}(\boldsymbol{T}^\pi, t) = [I(T_1^\pi < t), \cdots, I(T_n^\pi < t)]$.

THEOREM 1. *Consider model (2.1). Let $FPR^\pi(t)$ and $MDR^\pi(t)$ be the FPR and MDR levels of decision rule $\boldsymbol{\delta}(\boldsymbol{T}^\pi, t)$. Then we have:*

(i) The optimal FPR procedure is $\boldsymbol{\delta}_d^\pi = (\delta_{d,1}^\pi, \cdots, \delta_{d,n}^\pi)$, where

$$\delta_{d,i}^\pi = I(T_i^\pi < t_d^\pi) \text{ and } t_d^\pi = \sup\{t : FPR^\pi(t) = \alpha_n\},\ i = 1, \cdots, n. \tag{2.7}$$

(ii) The optimal MDR procedure is $\boldsymbol{\delta}_s^\pi = (\delta_{s,1}^\pi, \cdots, \delta_{s,n}^\pi)$, where

$$\delta_{s,i}^\pi = I(T_i^\pi < t_s^\pi) \text{ and } t_s^\pi = \inf\{t : MDR^\pi(t) = \alpha'_n\},\ i = 1, \cdots, n. \tag{2.8}$$

REMARK 3. A result on optimal FPR control has been obtained in Sun and Cai (2007). Theorem 1 extends the result to the optimal screening problem defined by (2.5) and (2.6). In contrast with the higher criticism (HC) method (Donoho and Jin, 2004) that tests a global null hypothesis, the construction of a screening subset involves making many simultaneous decisions at individual levels. This important difference is demonstrated by the phase transition theory developed in Section 4. The general strategy in constructing the subsets is to first rank the observations from the most significant to the least significant and then choose a cutoff along the rankings. Two important questions are: (i) What is the optimal ranking? (ii) What is the optimal cutoff? Theorem 1 reveals that the optimal ranking is determined by T_i^π, with the optimal thresholds being given by t_d^π and t_s^π, respectively.

The optimal thresholds can be obtained using stepwise procedures. The derivation involve adaptive estimation of the FPR and MDR. More explicitly, to construct the desired subsets, we first order T_i^π from the smallest to the largest as $T_{(1)}^\pi, \cdots, T_{(n)}^\pi$. The following method was proposed in Sun and Cai (2007).

PROCEDURE 1. *FPR procedure at level α_n. Let $k_d = \max\left\{j : \frac{1}{j}\sum_{i=1}^{j} T_{(i)}^\pi \le \alpha_n\right\}$. Then the discovery subset can be constructed as $\hat{\mathcal{S}}_d^\pi = \{i : T_i^\pi \le T_{(k_d)}^\pi\}$.*

Next we develop an MDR procedure, which constructs a subset via a "backward elimination" scheme. Specifically, the method starts with the full set, and leaves out one by one the least significant observation in the subset until there is evidence that a non-negligible proportion of signals have been missed.

PROCEDURE 2. *MDR procedure at level α_n'. Let $k_s = \min\left\{j : \sum_{i=n}^{j}(1 - T_{(i)}^\pi) \le n\epsilon_n\alpha_n'\right\}$. Then the screening subset can be constructed as $\hat{\mathcal{S}}_s^\pi = \{i : T_i^\pi \le T_{(k_s)}^\pi\}$.*

The next theorem shows that both procedures are valid for error rates control.

THEOREM 2. *Consider random mixture model (2.1). Let α_n and α_n' be positive sequences converging to zero slowly, say, at the rate $(\log n)^{-1}$. Denote by FPR_d the FPR level of Procedure 1 and MDR_s the MDR level of Procedure 2. Then we have*

$$FPR_d \le \alpha_n \text{ and } MDR_s \le \alpha_n'.$$

REMARK 4. Theorem 2 also holds in a non-asymptotic setting with fixed $\alpha_n = \alpha$ and $\alpha_n' = \alpha'$. To maintain notational consistency, we state the result with vanishing α_n and α_n' because both will be used in the next two sections for studying the phase transition theory, which is formulated in an asymptotic setting. The convergence rates of α_n and α_n' are important and will be analyzed rigorously in later sections (see the proofs of Theorems 4 and 7). At present, we only give a practical recommendation. Roughly speaking, the rate of $\alpha_n = \alpha_n' = (\log n)^{-1}$ ensures that the optimality in phase transition can be attained with high probability and hence serves as a suitable choice for many applications.

Our stepwise procedures can be easily implemented for the general mixture model (2.1). Specifically, we can use the method in Jin and Cai (2007) to estimate the null density f_0 and non-null proportion ϵ_n, and a standard kernel method to estimate the marginal density f (e.g., Silverman, 1986). The corresponding estimates are denoted by $\hat{\epsilon}_n, \hat{f}_0$ and $\hat{f}$. Then the oracle statistic T_i^π can be estimated as

$$\hat{T}_i^\pi = (1 - \hat{\epsilon}_n)\hat{f}_0/\hat{f}. \tag{2.9}$$

Finally the plug-in estimates $\hat{T}_i^\pi$ and $\hat{\epsilon}_n$ are used in Procedures 1 and 2 to construct the desired subsets. If the estimated non-null proportion $\hat{\epsilon}_n$ is consistent for a fixed $\epsilon > 0$ (non-vanishing), it can be shown that the plug-in procedures are asymptotically valid. The claim follows similar arguments as those in Theorem 2.

COROLLARY 1. *Consider the random mixture model (2.1). Let α_n and α_n' be positive sequences converging to zero slowly. Let $\hat{\epsilon}_n$, $\hat{f}_0$, $\hat{f}$ be estimates of ϵ, f_0 and f such that $\hat{\epsilon}_n \xrightarrow{p} \epsilon$, $E\|\hat{f} - f\|^2 \to 0$ and $E\|\hat{f}_0 - f_0\|^2 \to 0$. Denote by FPR_d the FPR level of Procedure 1, and MDR_s the MDR level of Procedure 2 with plug-in estimates $\hat{T}_i^\pi$ and $\hat{\epsilon}_n$. Then we have*

$$FPR_d \le \alpha_n(1 + o(1)) \text{ and } MDR_s \le \alpha_n'(1 + o(1)).$$

3. Design of Multistage High-throughput Studies: Phase Transition and Sample Size Calculation

Section 2 derives data screening procedures under the general random mixture model (2.1), and establishes their optimality in the sense of (2.5) and (2.6). However, having the most powerful test only provides a partial solution to the data screening problem. We also need a sufficiently large sample size to attain the desired power. In this section, we consider the closely related problem of *optimal design*, which involves finding the minimum number of replicates such that the goals on error control and power can be achieved simultaneously.

Our analysis in the rest of this section is carried out for the two-point mixture model (2.2). This makes it possible to give a simple and precise characterization of the phase transition boundaries. We shall explain in Sections 3.1 and 4 that the result obtained under this simple model is both practically relevant and theoretically important. Meanwhile, it is important to point out that the FPR and MDR procedures proposed in Section 2 can be easily implemented and enjoy desirable properties under the general model (2.1). Therefore the two-point model (2.2) may not be viewed as a limitation of the proposed data screening procedures.

3.1. Sample size and phase transition in a two-point model

From a practical point of view, an efficient design of high-throughput experiments can greatly improve the screening accuracy and lead to savings in study costs. However, most existing design strategies are ad-hoc and can be highly inefficient. For example, in the common practice of HTS, compounds are measured only *once* in primary screens and *twice* in secondary screens. These arbitrary choices on sample size can be problematic. One major concern is that a small number of replicates would yield a low signal to noise ratio (SNR). As a consequence, the screening procedure may have a low power. To achieve the desired power, one needs to increase the SNR by obtaining more replicates at each testing unit. Meanwhile, the study costs will soar if too many replicates were obtained. The key issue in the design is to find the "right" number of replicates. Motivated by this, we develop the theory on phase transition in optimal screening to characterize the precise conditions under which the goals on error control and power are simultaneously achievable. The theory yields formulae for calculating the minimum sample sizes needed in the screening process. In Section 3.1, we discuss general considerations on problem formulation. Sections 3.2 and 3.3 are devoted to the design problems in primary and secondary screens, respectively.

To conceptualize the design issues properly, it is helpful to first closely examine the framework under which the sample size problem is formulated in the context of single hypothesis testing. Suppose we want to test the hypotheses $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$. Then the power of a rejection rule depends on the sample size, test size and effect size $\Delta\mu = |\mu_1 - \mu_0|$. Consequently, given the test size and desired power, the sample size is determined by the prespecified $\Delta\mu$. Similarly, the sample size in data screening problems must be calculated with regards to a fixed point alternative. Therefore it is natural to focus on the two-point normal mixture model (2.2). Denote by $\Delta\mu$ the biological meaningful effect that we wish to discover. Let N be the number of replicates that we obtain for every testing unit i, $i = 1, \cdots, n$. Then the signal strength of interest can be expressed as $\mu_n = \sqrt{N}\Delta\mu$, which connects model (2.2) to the sample size N. As in Donoho and Jin (2004) and Cai et al. (2007), we adopt the calibration $\epsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r \log n}$ in our theoretical analysis, with β and r denoting the model sparsity and signal strength, respectively. This calibration enables us to describe phase transition precisely, and yields simple sample size formulae.

We would like to comment here that the two-point model (2.2) is suitable to handle the design problem in a more general random mixture model. In applications the signal

strengths are likely to vary across different testing units. Suppose the alternative distribution is a normal mixture with K-components: $F_1 = \sum_{k=1}^{K} p_k N(\mu_k, \sigma_k^2)$. Let μ^* be the biologically meaningful effect size (set by the users) one wishes to discover. Then a sample size analysis can be carried out for the two-point model $(1-p)N(0,1) + \theta_i N(\mu^*, \sigma^2)$. When implementing the MDR procedure to the collected data under the general model (2.1), our choice of the sample size guarantees that the signals from components with $\mu_k \geq \mu^*$ can be identified reliably (Theorem 4). Although the signals from components with $\mu_k < \mu^*$ are likely to be missed by our MDR procedure, such missed findings are considered to be inconsequential. Therefore, model (2.2) provides a useful practical guidance in design; we called (2.2) a working model because it does not represent the true population distribution. (A similar situation arises from the sample size problem in simple hypothesis testing, where the user-specified $\Delta\mu$ typically differs from the true effect size.)

3.2. Phase transition in data screening

In primary screens, the goal is to eliminate a large proportion of null cases to meet the laboratory constraints such as capacity limitations in the more expensive and time-consuming secondary screens. In the current practice of HTS, only a small fraction (say, top 1%) of compounds with highest activities ("hits") can be accommodated for further investigation. To avoid a high MDR, it has been advocated that more replicates of measurements should be obtained to increase the signal to noise ratio. The question of interest is: how many replicates are sufficient so that it is possible to reduce the size of the compound library significantly without losing many signals?

Suppose we wish to eliminate $100(1-\gamma_n^{-1})\%$ null cases in the data set, where $\gamma_n \to \infty$ is the desired shrinkage level. The specific requirements in primary screens are:

(S1) $\boldsymbol{\delta}$ keeps most of the signals with high probability. That is, $P(|\mathcal{S}_{\boldsymbol{\delta}} \cap \mathcal{I}| \geq (1-\eta)|\mathcal{I}|) \to 1$ as $n \to \infty$ for any $\eta > 0$.

(S2) $\boldsymbol{\delta}$ eliminates a significant proportion of null cases with high probability. That is, $P(|\mathcal{S}_{\boldsymbol{\delta}} \cap \mathcal{N}| \leq \gamma_n^{-1}|\mathcal{N}|) \to 1$ as $n \to \infty$.

Let $\gamma_n = n^{\kappa}$ with $\kappa < 1/2$. The next theorem gives the precise condition under which a discovery subset with properties (S1) and (S2) can be constructed.

THEOREM 3. *Consider model (2.2) with $\epsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r\log n}$. Then the screening boundary is given by $r = \kappa$. Specifically, the boundary implies that*

(i) If $r > \kappa$, then we can find a screening procedure which fulfills both (S1) and (S2).

(ii) If $r < \kappa$, it is impossible to find a screening procedure which fulfills both (S1) and (S2).

The screening boundary characterizes the optimality benchmark of all data screening procedures. In fact, our result indicates that existing screening methods can be substantially improved, and our new procedures promise to lead to great savings in study costs in multistage experiments. See Section 5.5 for a detailed numerical analysis. The next theorem shows that our sure screening procedure (Procedure 1) is fully efficient in the sense that it achieves the screening boundary in phase transition.

THEOREM 4. *Consider model (2.2) with $\epsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r\log n}$. Let $\gamma_n = n^{\kappa}$ be the desired shrinkage level in primary screens. Assume that the sure screening condition is fulfilled, i.e. $r > k$. Let η_0 be a positive constant satisfying $\sqrt{2\eta_0} < \min\{\sqrt{r} - \sqrt{\kappa}, \frac{1-\beta}{2}\}$. Consider subset $\mathcal{S}_{\boldsymbol{\delta}}$ that is constructed by Procedure 2 at MDR level α_n', where α_n' converges to 0 slowly such that $n^{\eta_0}\alpha_n' \to \infty$. Then $P(|\mathcal{S}_{\boldsymbol{\delta}} \cap \mathcal{N}| \leq \gamma_n^{-1}|\mathcal{N}|) \to 1$.*

The screening boundary can be used to determine the sample size needed in primary screens. Suppose that we wish to eliminate $100(1-\gamma_n^{-1})\%$ null cases while keeping most active compounds with effect size greater than or equal to $\Delta\mu$. Then it follows from $\mu_n = \sqrt{N}\Delta\mu$, $\mu_n = \sqrt{2r\log n}$ and Theorem 3 that the required sample size N_S should satisfy

$$N_S > \frac{2\kappa \log n}{\Delta\mu^2}. \tag{3.1}$$

This formula will be used in Section 6 for HTS design.

3.3. Phase transition in signal discovery

In secondary screens, the goal is to confirm the "hits" in primary screens and use the confirmed hits to generate "leads." The complex leads generation process, which involves a comprehensive assessment of chemical integrity, synthetic accessibility and structure-activity relationship, calls for precise control of the false positive rate. In particular, when the number of the testing units is overwhelming and the signals are weak and sparse, it is possible that the highest activities are mostly noisy observations. In this case it is necessary to obtain more replicates to reduce the noise level. The question of interest is: what is the minimum number of replicates that is needed so that the true signals can be separated from noise reliably? This is a nontrivial question in large-scale inference.

Let $\mathcal{S}_{\boldsymbol{\delta}}$ be a subset of "confirmed hits" constructed by $\boldsymbol{\delta}$. It is required that $\mathcal{S}_{\boldsymbol{\delta}}$ is signal-dominant and ideally we hope that $\mathcal{S}_{\boldsymbol{\delta}}$ contains virtually *all* active compounds. The signals and noises can be nearly perfectly classified into two subsets if

(D1) the FPR is vanishingly small, i.e. $E(|\mathcal{S}_{\boldsymbol{\delta}} \cap \mathcal{N}|)/E(|\mathcal{S}_{\boldsymbol{\delta}}|) \to 0$; and

(D2) the MDR is vanishingly small, i.e. $1 - E[\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}} \cap \mathcal{I})]/E[\mathrm{Card}(\mathcal{I})] \to 0$ as $n \to \infty$.

Now we derive the precise condition under which both (D1) and (D2) can be fulfilled. The line in β-r plane which demarcates the possibility of achieving (D1) and (D2) is called the *classification boundary*.

THEOREM 5. *Consider model (2.2). Define the misclassification rate* $L(\boldsymbol{\theta}, \boldsymbol{\delta}) = n^{-1}\sum_{i=1}^{n}\{(1-\theta_i)\delta_i + \theta_i(1-\delta_i)\}$. *The classification boundary is given by* $r = \beta$ *for all* $0 < \beta \le 1$ *and* $\sigma > 0$. *The boundary divides the* β-r *plane into two parts: the classifiable region* ($r > \beta$) *and unclassifiable region* ($r < \beta$). *Specifically, we have*

(i) In the classifiable region, $\min_{\boldsymbol{\delta}} E[L(\boldsymbol{\theta}, \boldsymbol{\delta})]/\epsilon_n = o(1)$, *and it is possible to find* $\boldsymbol{\delta}$ *which fulfills (D1) and (D2) simultaneously.*

(ii) In the unclassifiable region, $\min_{\boldsymbol{\delta}} E[L(\boldsymbol{\theta}, \boldsymbol{\delta})]/\epsilon_n = 1 + o(1)$, *and it is impossible to find* $\boldsymbol{\delta}$ *which fulfills (D1) and (D2) simultaneously.*

REMARK 5. The classification boundary $r = \beta$ in a two-point homoscedastic model has been obtained in Haupt et al. (2011), and was stated informally in Donoho and Jin (2006) and Meinshausen and Rice (2006) without proofs. Theorem 5 extends the result to a heteroscedastic model. It is well known that the heteroscedasticity has critical impacts on the phase diagrams in signal detection and discovery problems. For example, Theorems 6 and 8, and Figures 3 and 4 show that the detection and discovery boundaries vary significantly according to the value of σ. In contrast, Theorem 5 shows that *the classification boundary is always* $r = \beta$ *for all* σ. This observation reveals significant differences between the classification problem and other related large-scale inference problems.

Achieving the classification boundary often entails obtaining a very large number of replicates at every testing unit, which can be unrealistic in practice. A less ambitious goal is to ensure that we can separate some, if not all, useful compounds reliably, with which we can carry out the next stage analysis. More precisely, we require (D1) and

(D3) A *non-empty* subset is constructed with high probability, i.e. $P(|\mathcal{S}_{\boldsymbol{\delta}}| \geq 1) \to 1$.

The next theorem derives the *discovery boundary* that characterizes the phase transition in optimal discovery. The boundary gives the minimum condition under which we can achieve (D1) and (D3) simultaneously.

THEOREM 6. *(Discovery boundary). Consider model (2.2). Denote by* $\rho_{\text{dis}}(\beta)$ *the discovery boundary,*

- *for* $\sigma = 1$, *let* $\rho_{\text{dis}}(\beta) = (1 - \sqrt{1-\beta})^2$;
- *for* $0 < \sigma < 1$, *let* $\rho_{\text{dis}}(\beta) = \begin{cases} (1 - \sigma\sqrt{1-\beta})^2 & \text{if } 1 - \sigma^2 < \beta < 1 \\ (1-\sigma^2)\beta & \text{if } 0 < \beta \leq 1 - \sigma^2 \end{cases}$; *and*
- *for* $\sigma > 1$, *let* $\rho_{\text{dis}}(\beta) = \begin{cases} (1 - \sigma\sqrt{1-\beta})^2 & \text{if } 1 - \frac{1}{\sigma^2} < \beta < 1 \\ 0 & \text{if } 0 < \beta \leq 1 - \frac{1}{\sigma^2} \end{cases}$.

The above discovery boundary divides the β*-r plane into two parts:*

(i) If $r > \rho_{\text{dis}}(\beta)$, *then it is possible to find* $\boldsymbol{\delta}$ *which fulfills (D1) and (D3) simultaneously.*

(ii) If $r < \rho_{\text{dis}}(\beta)$, *then it is impossible to find* $\boldsymbol{\delta}$ *which fulfills (D1) and (D3) simultaneously.*

Now we study the effectiveness of our data-driven procedure using the discovery boundary as a theoretical measure of optimality. The next theorem shows that Procedure 1 is fully efficient in the sense that it achieves the boundary in phase transition when applied at appropriate screening levels.

THEOREM 7. *Consider model (2.2). Suppose we apply Procedure 1 at screening level* $\alpha_n \to 0$ *slowly (e.g.* $\alpha_n = (\log n)^{-1}$*). If* $r > \rho_{\text{dis}}(\beta)$, *then both (D1) and (D3) hold.*

Let N_C and N_D be the number of replicates needed to discover all signals and some useful signals, respectively. Denote by n' the number of testing units in secondary screens. It follows from Theorems 5 and 6 that N_C and N_D should satisfy

$$N_C > \frac{2\beta \log(n')}{\Delta\mu^2} \text{ and } N_D > \frac{2\log(n')}{\Delta\mu^2}\rho_{\text{dis}}(\beta), \tag{3.2}$$

respectively. As one would expect, the goal of discovering all signals is very ambitious and N_C is usually much larger than N_D.

4. Large-scale Inference: Signal Detection, Classification and Screening

So far we have derived classification, discovery and screening boundaries to characterize the benchmark performance of optimal data screening procedures. The boundaries are of great importance from both practical and theoretical perspectives. Section 3 has shown that the boundaries can be used in practice to determine the optimal sample size needed in the screening process. This section further discusses how the boundaries can be employed as a theoretical measure to assess the difficulty of related large-scale inference problems (Section 4.1) and to evaluate the effectiveness of existing screening procedures (Section 4.2).

4.1. Detection, classification and discovery boundaries

In this section, we focus on the following two-point normal mixture model

$$X_i \overset{i.i.d.}{\sim} (1-\epsilon_n)N(0,1) + \epsilon_n N(\mu_n, 1), \quad i = 1, \cdots, n,$$

and consider a sequence of closely related large-scale inference problems: (i) Are there any signals (signal detection)? (ii) Can any signals be separated from noise (signal discovery)? (iii) Can all signals be separated from noise (classification)? It is clear that the task becomes more and more challenging along this sequence of problems. The increased difficulty can be conveniently illustrated by the varied boundaries in phase transition. A comparison of the phase diagrams of signal detection, discovery and classification is given in Figure 2. The rest of this section gives a detailed explanation of each boundary. Important insights and interesting connections/distinctions of related concepts are provided along the discussion.

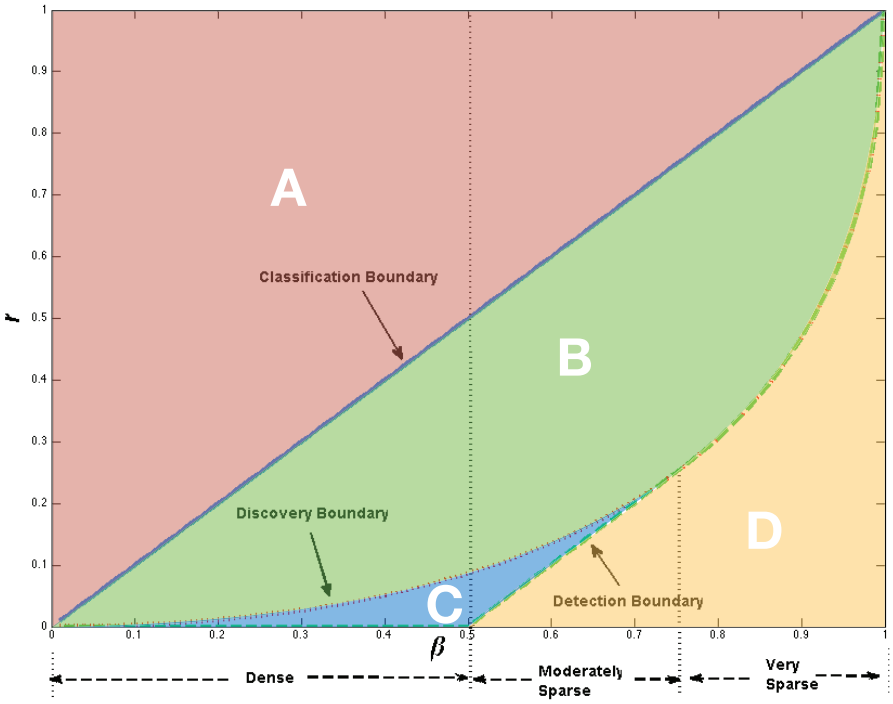

Fig. 2. Phase diagrams in a homoscedastic normal mixture model. The detection boundary $\rho_{\text{det}}(\beta)$, classification boundary $r = \beta$ and discovery boundary $\rho_{\text{dis}}(\beta)$ divide the β-r plane into four parts: (A) fully classifiable region; (B) partially classifiable and discoverable region; (C) undiscoverable but detectable region; (D) undetectable region.

The *detection boundary* (Ingster, 1998; Donoho and Jin, 2004, dashed line), defined by

$$\rho_{\text{det}}(\beta) = \begin{cases} \beta - 1/2 & 1/2 < \beta \le 3/4 \\ (1-\sqrt{1-\beta})^2 & 3/4 < \beta < 1 \end{cases},$$

is concerned with the possibility of reliably detecting the existence of any signal. The detection boundary divides the β-r plane into two parts: the undetectable region (region D in Fig. 2)

and detectable region. Specifically, let H_0^n be the global null hypothesis that there is no signal and H_1^n its alternative. In the interior of the undetectable region, H_0^n and H_1^n merge asymptotically and no statistical procedure would be successful in testing the global null with negligible error rate. In the interior of the detectable region, H_0^n and H_1^n separate asymptotically and the signals can be detected reliably with the sum of the Type I and Type II error rates converging to 0. Moreover, in this region the non-null proportion ϵ_n can be estimated consistently (Cai et al., 2007). The detection boundary provides an optimality benchmark that characterizes the fundamental limitation on the performance of all statistical procedures in testing the global hypotheses H_0^n vs. H_1^n. The higher criticism (HC, Donoho and Jin, 2004) procedure achieves the detection boundary and is hence *fully efficient* for the global testing problem.

The *classification boundary* $\rho_{\text{cls}}(\beta) = \beta$ (Theorem 5), gives the precise condition under which the observations can be separated into signals and noises with negligible misclassification rate. It divides the detectable region into two parts: *classifiable region* (region A in Fig. 2) and *partially classifiable region*. In the interior of the classifiable region, we can construct a subset with all signals and only signals (asymptotically); however in the partially classifiable region, a clear-cut separation of signal and noise is impossible and we must suffer from inflated false positive errors, or false negative errors, or both.

The *discovery boundary* (Theorem 6) further divides the region between the classification boundary and detection boundary into two parts: the *discoverable region* and *undiscoverable region*. In the region where $\rho_{\text{det}}(\beta) < r < \rho_{\text{dis}}(\beta)$ (detectable but undiscoverable, region C in Fig. 2), we can detect the existence of signals reliably but it is impossible to separate any individual signals from the noises. In the region where $\rho_{\text{dis}}(\beta) < r < \rho_{\text{cls}}(\beta)$ (discoverable but unclassifiable, region B in Fig. 2), we can identify some individual signals reliably with probability tending to 1 but it is impossible to separate all signals from the noises with negligible misclassification rate. The discovery boundary serves as a fundamental concept in simultaneous inference by providing the minimum condition for separating any individual signal from noise with high precision. Theorem 7 shows that this boundary is attained by the proposed FPR procedure with slowly converging screening levels.

4.2. Heteroscedasticity and connection to multiple testing theory

Thresholding is a useful technique in significance testing and subset selection. This section investigates, using the discovery boundary as a theoretical measure, the effectiveness of two thresholding strategies which are respectively based on the p-value and Lfdr. Our theory reveals that Lfdr thresholding is superior to p-value thresholding in large-scale inference.

In a two-point mixture model, define the p-value as $p_i = 1 - \Phi(X_i)$, where Φ is the cumulative distribution function of a standard normal variable. We consider two methods for comparison: minP and minL. The former selects the case with the smallest p-value and the latter selects the case with the smallest Lfdr value. Let $\rho_\delta(\beta)$ denote the *effective discovery boundary* of a given thresholding procedure δ. Let E_n^δ be the event that at least one true signal is identified by δ correctly. Then $\rho_\delta(\beta)$ divides the β-r plane into two regions: for the region where $r > \rho_\delta(\beta)$, we have $P(E_n^\delta) \to 1$, and for the region where $r < \rho_\delta(\beta)$, we have $P(E_n^\delta) \to 0$. The effective discovery boundaries for the minP and minL methods are summarized in the next theorem.

THEOREM 8. *Consider normal mixture model (2.2). Denote by* $\rho_{\text{minP}}(\beta)$ *and* $\rho_{\text{minL}}(\beta)$ *the effective discovery boundaries of the minP and minL methods, respectively. Then we have*

- *When* $\sigma = 1$, $\rho_{\text{minP}}(\beta) = \rho_{\text{minL}}(\beta) = \rho_{\text{dis}}(\beta) = (1 - \sqrt{1-\beta})^2$.

- *When* $0 < \sigma < 1$, $\rho_{\mathrm{minP}}(\beta) = (1 - \sigma\sqrt{1-\beta})^2$ *and*

$$\rho_{\mathrm{minL}}(\beta) = \rho_{\mathrm{dis}}(\beta) = \begin{cases} (1 - \sigma\sqrt{1-\beta})^2 & \text{if } 1 - \sigma^2 < \beta < 1 \\ (1-\sigma^2)\beta & \text{if } 0 < \beta \le 1 - \sigma^2 \end{cases}.$$

- *When* $\sigma > 1$,

$$\rho_{\mathrm{minP}}(\beta) = \rho_{\mathrm{minL}}(\beta) = \rho_{\mathrm{dis}}(\beta) = \begin{cases} (1 - \sigma\sqrt{1-\beta})^2 & \text{if } 1 - \frac{1}{\sigma^2} < \beta < 1 \\ 0 & \text{if } 0 < \beta \le 1 - \frac{1}{\sigma^2} \end{cases}.$$

REMARK 6. The minL approach selects the entries with small component-wise likelihood ratio (LR), whereas the minP approach picks the entries with large $|X_i|$. The two methods are equivalent if $\sigma \ge 1$ due to the monotonicity of the LR. However, this is no longer the case when $0 < \sigma < 1$. The minL procedure is fully efficient since it takes into account the distribution of the alternative hypothesis. In contrast, such information is completely ignored by the minP procedure. A more technical discussion of this point can be found in Section 9.8 of the supplemental material.

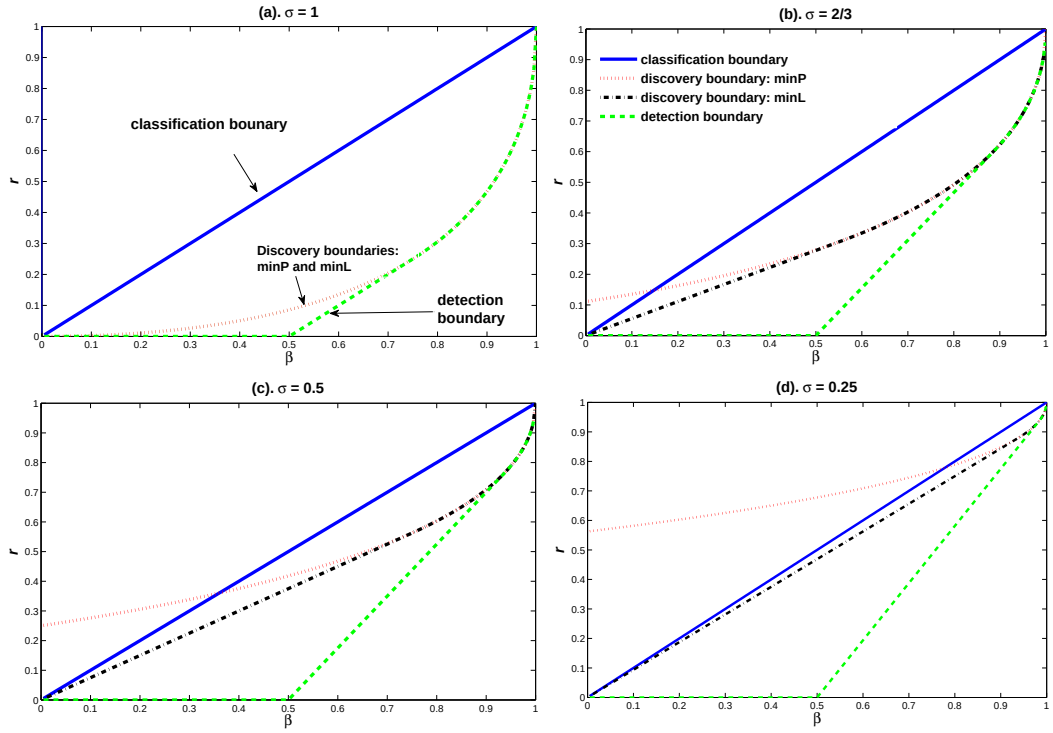

Fig. 3. The detection boundary $\rho_{\mathrm{det}}(\beta)$, classification boundary $r = \beta$ and discovery boundaries in a normal mixture model with $\sigma < 1$.

The situation with $0 < \sigma \le 1$ is illustrated by Figure 3. We can see that when $\sigma = 1$, the effective boundaries of the minP and minL methods, respectively denoted by $\rho_{\mathrm{minP}}(\beta)$ and $\rho_{\mathrm{minL}}(\beta)$, overlap with the optimality benchmark $\rho_{\mathrm{dis}}(\beta)$. However, in the heteroscedastic case with $0 < \sigma < 1$, the effective boundaries of minP and minL differ dramatically. The

minP method fails to achieve the discovery boundary whereas the minL method always works. The loss of efficiency of the minP method becomes larger as the non-null distribution becomes more concentrated (i.e. σ becomes smaller).

The boundaries for the case of $\sigma > 1$ are shown in Figure 4. We can see that $\rho_{\text{minP}}(\beta)$ and $\rho_{\text{minL}}(\beta)$ always overlap with the discovery boundary $\rho_{\text{dis}}(\beta)$; hence both the minP and minL methods are fully efficient in the signal discovery problem when $\sigma > 1$. It is interesting to note that the discovery boundary $\rho_{\text{dis}}(\beta)$ approaches the detection boundary $\rho_{\text{det}}(\beta)$ as σ approaches $\sqrt{2}$ from below. For $\sigma \geq \sqrt{2}$, all boundaries $\rho_{\text{det}}(\beta)$, $\rho_{\text{minP}}(\beta)$, $\rho_{\text{minL}}(\beta)$ and $\rho_{\text{dis}}(\beta)$ overlap completely with each other. The effective region (for signal detection and discovery) expands as σ increases. Actually, for $\sigma > \sqrt{2}$, it is even possible to detect signals with vanishingly small μ_n in the region where $0 < \beta < 1 - \sigma^{-2}$. An informal derivation of $\rho_{\text{det}}(\beta)$ is given in Section 9.8 in the supplemental material.

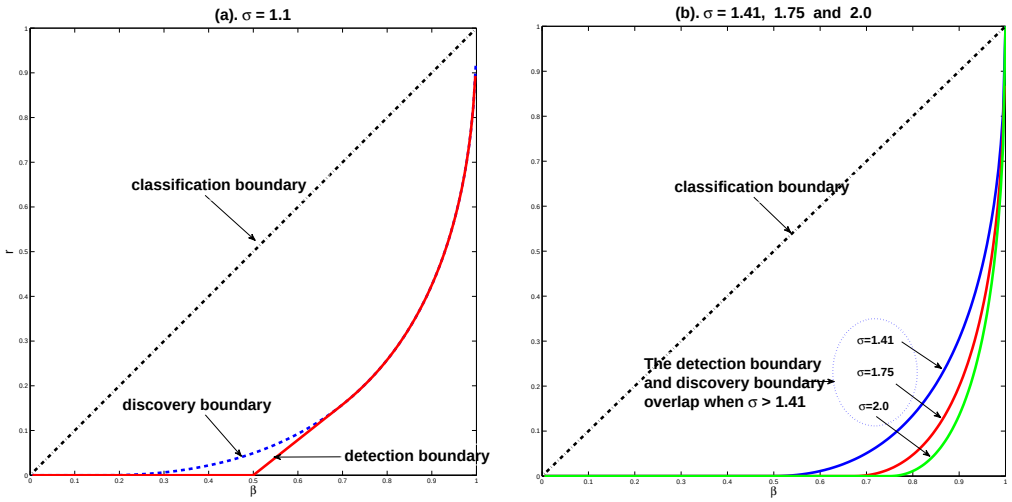

Fig. 4. The classification, detection and discovery boundaries when $\sigma > 1$. The discovery boundary approaches the detection boundary as σ increases and completely overlaps with the detection boundary when $\sigma \geq \sqrt{2}$.

5. Simulation Studies

We now investigate the empirical performance of different data screening procedures and compare the result with that predicted by theory. The R code for implementing the methods is available at `http://www-bcf.usc.edu/~wenguans/Papers/Optimal-Screening.html`.

5.1. Effective discovery boundaries: minP vs. minL

We conduct a simulation study to compare the minimum p-value (minP) and minimum Lfdr (minL) methods for their effectiveness in separating sparse signals from noise. The data are generated from the normal mixture model (2.2) with $\epsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r\log(n)}$. We set the parameter values as $\sigma = 0.3$, $\beta = 0.3$ and $r = 0.4$ to obtain a point in the region $\rho_{\text{minP}}(\beta) < r < \rho_{\text{minL}}(\beta)$. In view of the discovery boundary derived in Theorem 6 and also the curves in Figure 3, we expect that the minP method will fail whereas the minL method is likely to succeed in signal discovery. To get a sense of the rate of convergence, we vary

the sample sizes from $n = 10^3$ to $n = 10^6$. We simulate 1000 data sets. The minP method selects the location with largest $|X_i|$ and the Lfdr method picks the location with smallest Lfdr. The Lfdr is calculated using (2.9) as described in Section 2.2 without any parametric assumptions, making it a suitable and fair comparison.

In each data set, we construct two discovery subsets (each with one observation), respectively using the minP and minL methods, and then determine whether or not it is a true signal. The probabilities of accurate signal discovery is computed by counting the proportion of correct decisions among 1000 data sets. The results are summarized in Figure 5. We can see that the success rate of the minL method dominates that of the minP method at all sample sizes. As $n \to \infty$, the success rates of the minP method and minL method converge to 0 and 1, respectively. This is consistent with our theoretical prediction.

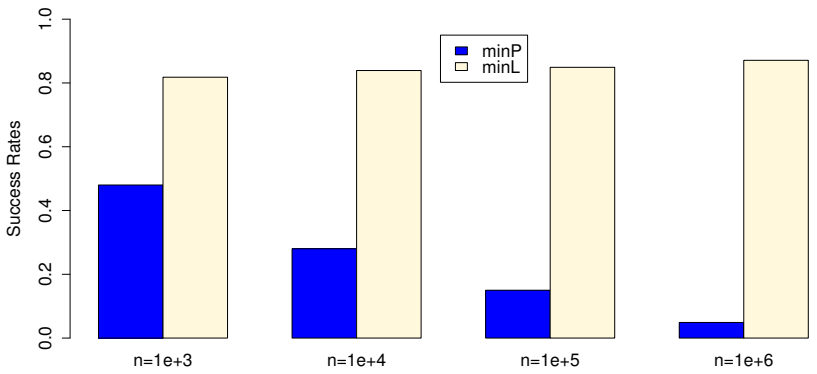

Fig. 5. The success probabilities of the minP and minL methods in separating sparse signals from noise when $\sigma = 0.3$, $\beta = 0.3$ and $r = 0.4$ for different sample sizes. As $n \to \infty$, The success rates of the minP method and minL method converge to 0 and 1, respectively.

5.2. Discovery of sparse signals and power analysis

We now turn to the performance of the FPR procedure (Procedure 1). It can be easily shown that Procedure 1 controls both the FPR and FDR. We consider FDR in this subsection and turn to FPR later (there is no essential difference between FDR and FPR in large-scale testing problems). The FDR level is set at γ. Consider the normal mixture model (2.2) and choose $r = 0.5$, $\beta = 0.3, 0.5, 0.7$. We apply the FPR procedure at level γ; the actual missed discovery rate (MDR), FDR and expected set size (ESS) are plotted as functions of γ. The results are summarized in Figure 6. It is clear that the sure discovery procedure controls the FDR at level γ precisely. As expected, the ESS of the discovery subset increases and the MDR decreases when the FDR level increases. The plot also suggests that a direct construction of signal-dominant subset may be unrealistic. For example, when $r = \beta = 0.5$, more than 90% signals can be missed at FDR level 0.05 in the homoscedastic model, and almost all signals are missed in the heteroscedastic model. This can be predicted from our theory on classification boundary. The result indicates that, in order to discover some useful signals, it is necessary to either increase the signal strength (e.g. by collecting more replicates at all testing units), or to reduce the original set to a much smaller subset (e.g. by applying

a data screening procedure and following up with a second stage analysis).

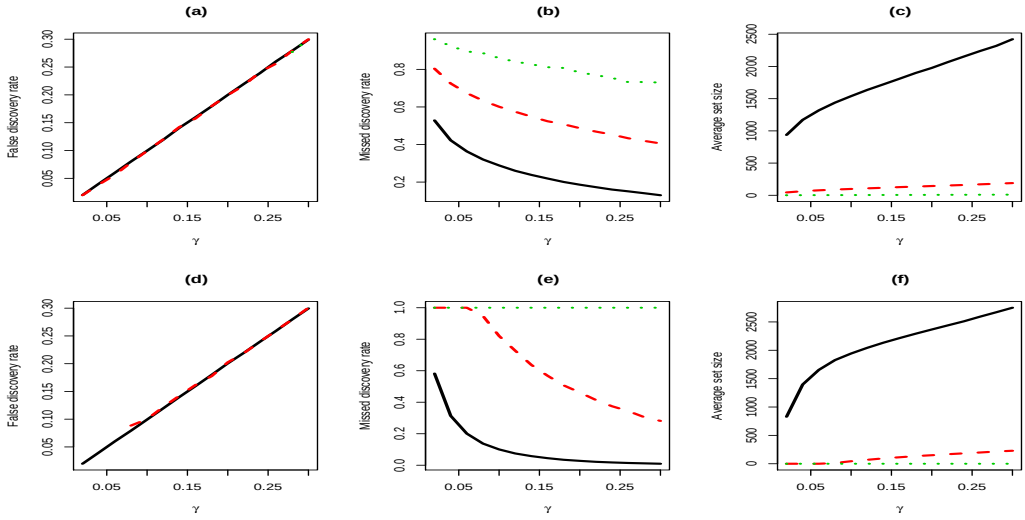

Fig. 6. Properties of the discovery procedure at different FDR levels. The goal is to make sure that the proportion of signals is at least $100(1-\gamma)\%$. The solid, dashed and dotted lines correspond to $\beta = 0.7$, $\beta = 0.5$, and $\beta = 0.3$, respectively. Top row considers $\sigma = 1$ and the bottom row considers $\sigma = 0.5$.

5.3. Data screening with sparse signals and power analysis

In the third simulation study we investigate the performance of MDR procedure (Procedure 2). The goal is to include at least $100(1-\alpha)\%$ signals while trying to eliminate as many noises as possible. We set the original set size to be $n = 50,000$ and generate observations from the normal mixture model (2.2). The screening subset is constructed at different screening levels. The actual MDR, FDR and ESS are plotted as functions of the nominal screening level α. The results are summarized in Figure 7, where the top and bottom rows consider homoscedastic case ($\sigma = 1$) and heteroscedastic case ($\sigma = 0.5$), respectively. We choose $r = 0.5$ and consider different sparsity levels $\beta = 0.3, 0.5$ and 0.7, which correspond to the solid, dashed and dotted lines in the plot, respectively. We can see that the MDR is controlled at level α precisely in all settings by the sure screening procedure. The FDR decreases with the screening level α. As the signals become sparser, we can achieve greater set size reduction and the corresponding FPR is also lower at the same screening level. The benefit of screening is clear; for example, when $r = \beta = 0.5$ and $\sigma = 0.5$, we can reduce the set size by more than 100 times (from $n = 50,000$ to $n < 500$) while only loosing about 5% signals.

5.4. FPR and MDR control under Gaussian and non-Gaussian alternative distributions

This section studies the effectiveness of our data-driven procedures for FPR and MDR control. Our methodology is essentially nonparametric and works well in the general mixture model (2.1). We illustrate this point by considering several simulation settings including models with departures from normality.

In our simulation the true effect sizes μ_i are observed with error ε_i:

$$x_i = \mu_i + e_i, i = 1, \cdots, n.$$

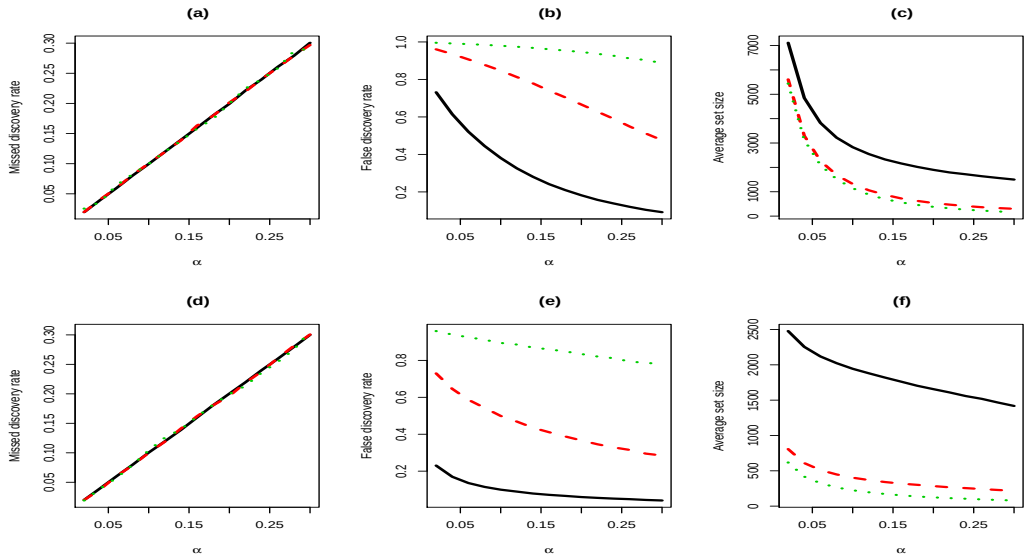

Fig. 7. Properties of the adaptive set S_U at different screening levels. The goal is to include at least $100(1-\alpha)\%$ signals. The solid, dashed and dotted lines correspond to $\beta = 0.7$, $\beta = 0.5$, and $\beta = 0.3$, respectively. Top row considers $\sigma = 1$ and the bottom row considers $\sigma = 0.5$.

We assume that μ_i's have a density function with a point mass at zero

$$f_\mu(\cdot) = (1-\epsilon_n)\delta_0(\cdot) + \epsilon_n g(\cdot),$$

where ϵ_n is the proportion of non-null cases as defined before, δ_0 is the dirac delta function and g is a continuous density function. Both g and the density function of e_i, denoted by f_e, will be specified later. We consider the following three models:

Model I: g is the density of $Y = 2 + 1.5|Z|$ where $Z \sim N(0,1)$, and f_ε is the standard Gaussian density.

Model II: g is the density of Uniform(2, 4) and f_e is the standard Gaussian density.

Model III: g is the density of Uniform(2, 4) and f_e is the t density with $df = 6$.

The FPR and MDR procedures in Section 2 are implemented with estimated Lfdr statistics. The estimation method is described by (2.9) in Section 2.2, with ϵ_n being estimated using the method in Jin and Cai (2007), f_0 being the density of a standard normal variable and $\hat{f}$ being a kernel density estimator with bandwidth chosen by cross validation. We vary the number of tests and sparsity levels and apply the FPR and MDR procedures at nominal level $\alpha_n = \alpha'_n = 0.05$. In each setting, the actual FPR and MDR levels are computed by averaging over 500 replications. The simulation results are summarized in Table 1. We can see that the FPR is controlled very well in all settings. In particular, the control is quite precise when n is large. The MDR control is more effective when n is large. This is consistent with our intuition since the MDR method relies on the accuracy of the estimators $\hat{\epsilon}_n$ and $\hat{f}$, and the precision of these estimators would improve in higher dimensions. The control become less effective in lower dimensional settings ($n = 10^3$), but the actual MDR levels are still acceptable. In both FPR and MDR control problems, the departure from Gaussian distributions seems to have little effect on the testing results.

Table 1. FPR and MDR control in Gaussian and non-Gaussian models

	FPR Control			MDR Control		
	Model I	Model II	Model III	Model I	Model II	Model III
$n = 10^3, \beta = 0.3$	0.051	0.053	0.054	0.057	0.052	0.050
$n = 10^3, \beta = 0.4$	0.053	0.052	0.053	0.064	0.056	0.057
$n = 10^4, \beta = 0.3$	0.052	0.050	0.051	0.049	0.045	0.045
$n = 10^4, \beta = 0.4$	0.051	0.051	0.052	0.061	0.052	0.056

5.5. Comparison with distilled sensing

We now compare our sure screening procedure with the distilled sensing (DS) method which was proposed by Haupt et al. (2011) under a homoscedastic normal mixture model. At each distillation step, the distilled sensing method keeps locations with positive observations and then obtain new observations for these locations. Due to the symmetry of a standard normal distribution and the sparsity of signals, DS eliminates roughly half of the noises in each distillation step. Interestingly, our theoretical result indicates that a much greater amount of shrinkage can be achieved. For example, when $n = 10^7$ and $\mu_n = 4$, we have $r \approx 1/2$. The DS method would keep about $n/2 = 5 \cdot 10^6$ locations for the next stage. In contrast, our theory suggests we can potentially shrink the data set down to a size as small as $n^{1/2} \approx 3,000$, more than a thousand times smaller than that of the DS method, and virtually without loosing more signals. Our result indicates that the DS method can be substantially improved. The other limitation of the DS method is that the MDR can be high if the signals are weak.

In our simulation we consider the following normal mixture model $X_1, \cdots, X_n \sim (1 - \epsilon_n)N(0, 1) + \epsilon_n N(3, 1)$ and apply the MDR procedure and the DS method for different n and at various sparsity levels. The results for the first two stages of screening are summarized in Table 2. For our MDR procedure, we set the nominal MDR level at $\alpha'_n = 0.02$. The Lfdr is estimated as $\widehat{\text{Lfdr}} = (1 - \hat{\epsilon}_n)f_0/\hat{f}$ as described in the previous subsection.

The following observations can be made: (i) our MDR method is more effective in reducing the size of ultra large data sets. For example, when $n = 10^6$ and $\beta = 0.3$, our MDR procedure shrinks the size of the data set down to about 30K in a single screening stage. In contrast, the DS method can only reduce the size to around 500K. (ii) The data reduction process of the MDR procedure is more dynamic, with high shrinkage level in the first stage and relatively low shrinkage level at the second stage. In contrast, the DS method has roughly the same shrinkage level at every screening stage. (iii) The MDR procedure controls the error rate at the nominal level more precisely. Note that with $\alpha = 0.02$, the actual MDR levels in the two stages are 0.98 and 0.96, respectively, which are quite close to our simulation results. In contrast, the DS method always has a very low false negative rate, but at the price of more measurements and hence higher study costs. Our theoretical analysis of the screening boundary shows that if the signal strength μ_n diverges at order $O((\log n)^\nu)$ with $0 < \nu < 1/2$, then the shrinkage rate of the DS method is asymptotically optimal and hence agrees with our MDR procedure. However, if μ_n diverges to infinity at a faster rate, then only eliminating half of the observations is too conservative and the DS method can be much improved. Finally, we want to point out that the MDR method relies on the accuracy of the estimators $\hat{\epsilon}_n$ and $\hat{f}$, and tends to work better in higher dimensions. We recommend applying the SS method when n is ultra large and switching to more stable method (such as DS method) at later stages of screening.

Table 2. Comparison of DS and SS methods. The first and second number in the parenthesis correspond to the results in the first and second stages, respectively.

Size	Sparsity	Card($\hat{\mathcal{S}}_{SS}$)	Card($\hat{\mathcal{S}}_{DS}$)	FNR$_{SS}$	FNR$_{DS}$
$n = 10^6$	$\beta = 0.2$	(98379, 61850)	(531999, 297350)	(0.97, 0.95)	(0.99, 0.99)
	$\beta = 0.3$	(29597, 15160)	(507715, 261756)	(0.95, 0.93)	(0.99, 0.99)
$n = 10^5$	$\beta = 0.2$	(13337, 9827)	(54936, 32617)	(0.98, 0.96)	(0.99, 0.99)
	$\beta = 0.3$	(7150, 3142)	(51645, 27435)	(0.98, 0.95)	(0.99, 0.99)
$n = 10^4$	$\beta = 0.2$	(1774, 1523)	(5702, 3691)	(0.97, 0.96)	(0.99, 0.99)
	$\beta = 0.3$	(1178, 612)	(5284, 2923)	(0.98, 0.95)	(0.99, 0.99)

6. Application

Alzheimer's disease (AD) is a progressive brain disorder with no effective treatments. Currently it is affecting six million Americans and is predicted to affect 1 in 85 people globally by 2050. The identification of small-molecule modulators of protein function, and the process of transforming these into informative leads for drug discovery, provide a promising direction towards the cure of AD. The HTS study has become a standard tool for improving the efficiency and speed of the identification process. To illustrate how our methodology can be implemented, we describe and analyze the HTS study conducted by McKoy et al. (2012). The goal of the study is to identify novel inhibitors of the amyloid beta peptide ($A\beta$), whose aggregation is believed to be a major underlying molecular culprit in AD. The inexpensive and effective isolation of novel inhibitors could lead to better molecular scaffolds for AD's therapy. In the study, 90 microplates are prepared, each with 24 by 24 wells containing carefully catalogued compounds to be tested. The size of the compound library is $n = 51,840$. The data set contains three z-scores for each compound, which are obtained from the raw data by respectively standardizing the three replicated measurements against the background. In the analysis, the informal "rule of three" was used to select candidate compounds. However, the prefixed threshold fails to control the probabilistic error rates: if we directly apply the rule of three to the first set of z-values, then both the FPR and MDR can be quite high; if we apply the rule of three repeatedly three times to the three sets of z-values, then only one compound would survive after three stages. In addition, the study design, which obtains *three* replicates for *all* compounds, can be highly inefficient.

In this section, we first implement our data screening procedures to analyze the HTS data, then discuss how to use the collected data set as a pilot data set to design a more effective multistage experiment. It is important to note that the general mixture model (2.1) is assumed when implementing our screening procedures to analyze the data sets, and the two-point model (2.2) is only used for the sample size calculation.

6.1. Data screening

To control the decision errors effectively, we adopt a three-stage "screen-clean" strategy in analysis, where the first two sets of z-values are used for "screening" and the last set of z-values are used for "cleaning." More explicitly, we first repeatedly apply the sure screening procedure at level 0.1 in the first two stages to reduce the size of the compound library, and then apply the sure discovery procedure at level 0.1 in the final stage to further eliminate the false positives. At the second and third stages, we only conduct analysis on testing units that are selected from the previous stage.

The implementation of our data screening procedures requires the estimation of unknown model parameters. We take the approach in Jin and Cai (2007) to estimate the *empirical null distribution* as $N(\hat{\mu}_0, \hat{\sigma}_0^2)$, where $\hat{\mu}_0 = 0.257$ and $\hat{\sigma}_0 = 0.76$. See Efron (2004) for more

Table 3. Summary of a three-stage analysis of the HTS data

	Lfdr Threshold	Subset size	Estimated FDR	Estimated MDR
Stage 1 (screen)	0.95	30,141	0.75	0.08
Stage 2 (screen)	0.95	21,311	0.75	0.12
Stage 3 (clean)	0.29	67	0.098	0.87

discussions on why the empirical null is superior to the theoretical null $N(0,1)$ in large-scale inference. We then proceed to estimate the proportion of the non-nulls as $\hat{\epsilon}_n = 0.0087$ based on the estimated empirical null. The marginal density f is estimated using a kernel density estimator $\hat{f}$ with the bandwidth chosen by cross validation. The test statistics $\hat{T}_i^{\pi}$ are then calculated based on (2.8). Finally we apply the three-stage "screen and clean" procedure with estimated $\hat{T}_i^{\pi}$. The results are summarized in Table 3.

We can see that after two stages of screening at level 0.1, the approximate size of the compound library is reduced from 50K to 20K. The estimated FPR and MDR are 0.75 and 0.12, respectively. In the last stage we eliminate the noises at FPR level 0.10. While the FPR is controlled at the desired level, most signals seem to have been missed (the estimated MDR is 0.87). It is clear that the statistical power is very low in this multistage analysis from two perspectives: (i) in the first two "screening" stages, while the MDR can be controlled effectively, we fail to achieve a significant size reduction for the compound library; (ii) in the final "cleaning" stage, while the FPR can be controlled precisely, a substantial percentage of signals have been missed. To increase the power, we need to increase the signal to noise ratio (SNR) by obtaining more replicates at each stage. Currently only one replicate is used to obtain the z-value, and our result indicates that the sample size is inadequate. In the next section, we use the observed data set as a pilot data set and discuss how to determine the sample sizes at different stages to achieve the desired power.

6.2. HTS design

We first consider the sample size problem in primary screens. In the previous section, it was estimated that 451 compounds in the data set are signals. Suppose that the lab capacity only allows $\tilde{n} = 4,000$ compounds to enter the more expensive second screens. The goal is to construct a subset which fulfills the size constraint while keeping most signals in the subset. The required shrinkage level is $\kappa = 1 - log(\tilde{n})/log(n) = 0.236$. The formula (3.1) can be used for sample size calculation. For example, if the goal is to keep all signals with effect sizes $\Delta\mu \geq 1.5\hat{\sigma}_0$, then the required sample size is 2.28 (rounded up to 3); if the goal is to keep in the subset all signals with $\Delta\mu \geq \hat{\sigma}_0$, then the required sample size is 5.12 (rounded up to 6).

Now suppose that we have reduced the size of the compound library to $\tilde{n} = 4,000$ and the primary screens have been successful in retaining most signals in the subset. Then the proportion of non-nulls in the subset is $\hat{\epsilon}^* = 0.11$ and the sparsity parameter $\beta^* = 0.266$. The goal in secondary screens is to construct a subset with only signals. If we wish to include in the discovery subset *all* signals with effect size $\Delta\mu \geq \hat{\sigma}_0$, then according to (3.2), the required sample size is 4.4 (rounded up to 5). If we only want to construct a *nonempty* discovery subset, then the discovery boundary is $\rho_{\text{dis}}(\beta) = 0.02$, and the required sample size is 0.33 (rounded up to 1).

In practice the sample size problem is complicated and it is not recommended to simply give a blind solution. We suggest that the investigators may use our sample size formulae to explore the efficacy of various designs. One possible approach is to utilize the information in a pilot study and create a table or a plot for decision support. For example, in Table 4, we summarize the estimates of respective study costs for various combinations of screening

Table 4. A summary of number of replicates and study budgets for decision support, with different combinations of lab capacities and effect sizes. N_1 is the minimum number of replicates needed to reduce the size from n to $\tilde{n}$ without losing important signals with effect size $\Delta\mu$. N_2 is the minimum number of replicates needed to discover most important signals with effect sizes $\Delta\mu$. c_1 and c_2 are the costs for obtaining one replicate in primary screens and secondary screens, respectively. The table is created using the HTS data set in McCoy et al. (2012) as a pilot data set. The situation $\tilde{n} = n$ corresponds to a single stage analysis.

	$\Delta\mu \geq \hat{\sigma}_0$			$\Delta\mu \geq 1.5\hat{\sigma}_0$		
	N_1	N_2	Cost	N_1	N_2	Cost
$\tilde{n} = n$	0	9.4	$c_1N_1n + c_2N_2\tilde{n}$	0	4.2	$c_1N_1n + c_2N_2\tilde{n}$
$\tilde{n} = 10K$	3.3	6.2	$c_1N_1n + c_2N_2\tilde{n}$	1.5	2.8	$c_1N_1n + c_2N_2\tilde{n}$
$\tilde{n} = 4K$	5.1	4.4	$c_1N_1n + c_2N_2\tilde{n}$	2.3	1.9	$c_1N_1n + c_2N_2\tilde{n}$
$\tilde{n} = 1K$	7.9	1.6	$c_1N_1n + c_2N_2\tilde{n}$	3.5	0.7	$c_1N_1n + c_2N_2\tilde{n}$

levels and effect sizes. Then the investigators can decide the best sample size carefully based on their experiences, budget constraints and biological insights.

7. Discussion

The present paper develops phase transition theory in optimal screening to characterize necessary conditions under which the goals on error control and power are simultaneously achievable in data screening problems. It was shown that our procedures achieve the boundaries in phase transition, implying that the conditions are also sufficient. The methods can be used in practice to calculate the optimal sample sizes at different stages of screening. The discovery boundary $\rho_{\text{det}}(\beta)$, derived as a part of our phase transition theory, lies in between the detection and classification boundaries (Figure 2). The discovery boundary can also be used as an optimality criterion to evaluate the effectiveness of different thresholding methods. We show that the Lfdr thresholding is fully efficient for signal discovery whereas p-value thresholding is inefficient in a heteroscedastic model with $0 < \sigma < 1$.

It is helpful to explain at intuitive levels why various phase diagrams differ so dramatically. The discussion would provide interesting insights on existing theories. More technical details are given in Section 9.8 in the supplemental material. The insights are that *the most informative part of the sample* depends on the goals in large-scale inference. To illustrate, consider the probability $P(X_i > \sqrt{2q\log n})$, where $0 < q \leq 1$ is a constant. For a given testing procedure δ, let q_δ denote the threshold for which the test has the largest power to reject the null (i.e. the most informative part of the sample). The key observation is that different thresholding procedures would yield different q_δ's. Specifically, the thresholding methods based on p-values always choose $q_{pv} = 1$, a scheme which virtually looks for non-nulls in the tail areas of the mixture density. In contrast, the screening procedures developed in Section 2, which are based on thresholding the Lfdr statistic, makes simultaneous decisions at individual levels by choosing q_{lf} that maximizes the likelihood ratio. In other words, the Lfdr looks for non-nulls in areas where the largest ratio of the non-null density and the null density occurs. The p-value method suffers from severe loss of power because the tail areas are not always the most informative parts of the sample when $0 < \sigma < 1$. In particular, as revealed by the analysis in the proof of Theorem 6, the most informative part of the sample for signal discovery is in the middle, not the tail areas of the mixture density. For testing the global null, the HC statistic uses threshold q_{hc}, which is chosen to maximize a normalized uniform empirical process (NUEP); hence HC looks for information where the values of the NUEP

under the global null and its alternative have the largest difference. In contrast with the popular tail thresholding methods which always choose $q_\delta = 1$, q_{hc} is adaptive to the sparsity level β and is not equal to 1 when $0 < \beta < \frac{3}{4}$, which indicates that the most informative part of the sample for testing global null may not be the tail areas. This phenomena, which has been observed in Donoho and Jin (2004), explains why the extreme value methods are inefficient in detection problems when the signals are weak and moderately sparse.

Under the setting of univariate thresholding, the main advantage of a multistage design is in the savings in study costs. Our analysis only provides a starting point for the optimal design of multistage experiments. Important open problems include: (i) optimization with a diverging number of distillation stages subject to a fixed budget constraints (as considered in Zehetmayer et al., 2008; Haupt et al., 2011); (ii) generalization of the sample size formulae to the non-Gaussian case; and (iii) development of phase transition theory under a more general setting.

8. Proofs of Main Theorems

In this section, we prove the main results of the paper. We first state a lemma, which is used in the proof of Theorem 1. The proof of the lemma is given in the supplemental material.

LEMMA 1. *Consider a weighted classification problem with loss function*

$$L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta}) = \lambda \sum_{i=1}^{n} (1 - \theta_i)\delta_i + \sum_{i=1}^{n} \theta_i (1 - \delta_i), \tag{8.1}$$

where λ is the inference loss of a false positive decision and $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_n) \in \{0, 1\}^n$ is a binary decision rule. Then the optimal rule which minimizes the classification risk $r(\lambda, \boldsymbol{\delta}) = E\{L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta})\}$ is $\boldsymbol{\delta}^{\pi,\lambda} = \{\delta_1^{\pi,\lambda}, \cdots, \delta_n^{\pi,\lambda}\}$, where

$$\delta_i^{\pi,\lambda} = I\{T_i^\pi < (1 + \lambda)^{-1}\},\ i = 1, \cdots, n. \tag{8.2}$$

8.1. Proof of Theorem 1

Let $\mathrm{EFP}_{\boldsymbol{\delta}} = E\{\sum_{i=1}^n (1 - \theta_i)\delta_i\}$ and $\mathrm{ETP}_{\boldsymbol{\delta}} = E(\sum_{i=1}^n \theta_i \delta_i)$ be the expected number of false positives and expected number of true positives when applying $\boldsymbol{\delta}$. Then we have $r(\lambda, \boldsymbol{\delta}) = n\epsilon_n + \lambda \mathrm{EFP}_{\boldsymbol{\delta}} - \mathrm{ETP}_{\boldsymbol{\delta}}$. According to Lemma 1, the optimal rule is $\delta_i^{\pi,\lambda} = I\{T_i^\pi < (1 + \lambda)^{-1}\},\ i = 1, \cdots, n$.

Let $\boldsymbol{\delta}^\pi(t) = \{\delta_1^\pi(t), \cdots, \delta_n^\pi(t)\}$ be a thresholding rule, where $\delta_i^\pi(t) = I(T_i^\pi < t),\ i = 1, \cdots, n$. Denote by $\mathrm{FPR}^\pi(t)$, $\mathrm{MDR}^\pi(t)$, $\mathrm{ETP}^\pi(t)$ and $\mathrm{EFP}^\pi(t)$ the FPR, MDR, ETP and EFP of $\boldsymbol{\delta}^\pi(t)$. Define $t_d^\pi = \sup\{t : \mathrm{FPR}^\pi(t) = \alpha_n\}$ and $t_s^\pi = \inf\{t : \mathrm{MDR}^\pi(t) = \gamma_n\}$. The goal is to show that $\boldsymbol{\delta}^\pi(t_d^\pi)$ satisfies (2.5) and $\boldsymbol{\delta}^\pi(t_d^\pi)$ satisfies (2.6). Now take $\lambda_d = 1/t_d^\pi - 1$ and consider a weighted classification problem with loss function $L_{\lambda_d}(\boldsymbol{\theta}, \boldsymbol{\delta})$. Then according to (8.2), the classification risk is minimized by $\boldsymbol{\delta}^\pi(t_d^\pi) = \{I(T_i^\pi < t_d^\pi) : i = 1, \cdots, n\}$. The minimum Bayes risk is

$$r\{\lambda_d, \boldsymbol{\delta}^\pi(t_d^\pi)\} = n\epsilon_n + \left(\frac{\alpha\lambda_d}{1 - \alpha} - 1\right) \mathrm{ETP}^\pi(t_d^\pi).$$

We claim $\left(\frac{\alpha\lambda_d}{1-\alpha} - 1\right) \le 0$ since the following decision rule "$\delta_i = 0$ for all i" must have a higher risk than $\boldsymbol{\delta}^\pi(t_d^\pi)$. Consider an arbitrary decision rule $\delta^* \in \mathcal{D}_{d,\alpha_n}$. The corresponding FPR, ETP and EFP are denoted by FPR^*, ETP^*, and EFP^*. It is easy to argue by contradiction that if $\mathrm{ETP}^* > \mathrm{ETP}^\pi(t_d^\pi)$, then we must have $\mathrm{FPR}^* > \alpha_n$. Therefore $\boldsymbol{\delta}^\pi(t_d^\pi)$ satisfies (2.5). Similarly we can show that $\boldsymbol{\delta}^\pi(t_s^\pi)$ satisfies (2.6). □

8.2. Proof of Theorem 2

Let $\boldsymbol{\delta}^d$ and $\boldsymbol{\delta}^s$ denote the stepwise sure discovery and screening procedures, respectively. Then the EFP of $\boldsymbol{\delta}^d$ is $\mathrm{EFP}_{\boldsymbol{\delta}^d} = E\left\{\sum_{i=1}^n (1-\theta_i)\delta_i^d\right\} = E\left\{Q(\boldsymbol{\delta}^d)\cdot k_d\right\}$, where $Q(\boldsymbol{\delta}^d) = k_d^{-1}\sum_{i=1}^{k_d} T_{(i)}^{\pi}$. The operation of $\boldsymbol{\delta}^d$ guarantees that $Q(\boldsymbol{\delta}^d) \le \alpha_n$ for all realizations of $\{X_1,\cdots,X_n\}$. Hence $\mathrm{EFP}_{\boldsymbol{\delta}^d} \le \alpha_n E\{\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}^d})\}$. It follows that $\mathrm{FPR}_{\boldsymbol{\delta}^d} \le \alpha_n$.

Next, the MDR of $\boldsymbol{\delta}^s$ is $\mathrm{MDR}_{\boldsymbol{\delta}^s} = E\{\sum_{i=1}^n \theta_i(1-\delta_i^s)\}/(n\epsilon_n) = E\{\tilde{Q}(\boldsymbol{\delta}^s)\}$, where $\tilde{Q}(\boldsymbol{\delta}^s) = (n\epsilon_n)^{-1}\sum_{i=n}^{j}\left(1-T_{(i)}^{\pi}\right)$. The operation of $\boldsymbol{\delta}^s$ guarantees that $\tilde{Q}(\boldsymbol{\delta}^s) \le \alpha_n'$ for all realizations of $\{X_1,\cdots,X_n\}$. It follows that $\mathrm{MDR}_{\boldsymbol{\delta}^s} \le \alpha_n'$. □

8.3. Proof of Theorem 3

Proof of Part (i). Consider threshold $t_n = \sqrt{2\kappa\log n}$ and subset $\mathcal{S}_{\boldsymbol{\delta}} = \{i : X_i > t_n\}$. We will show that both Conditions (S1) and (S2) are fulfilled by $\mathcal{S}_{\boldsymbol{\delta}}$ if $r > \kappa$. First, according to the standard bound on Gaussian tail, we have $q_n = P(Z_i > t_n) \le \frac{1}{2\sqrt{\pi\kappa\log n}} n^{-\kappa}$. Then $\gamma_n^{-1} - q_n = n^{-\kappa}(1+o(1))$. Next note that $|\mathcal{I}| = n^{1-\beta}$ and $|\mathcal{N}| = n(1+o(1))$, it follows from Hoeffding's inequality that

$$\begin{aligned} P\left\{\frac{\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}}\cap\mathcal{N})}{\mathrm{Card}(\mathcal{N})} > \gamma_n^{-1}\right\} &= P\left[\mathrm{Card}(\mathcal{N})^{-1}\sum_{i\in\mathcal{N}}\{I(X_i>t_n)-q_n\} > \gamma_n^{-1}-q_n\right] \\ &= P\left[\sqrt{\mathrm{Card}(\mathcal{N})}\left\{\mathrm{Ave}_{i\in\mathcal{N}} I(X_i>t_n)-q_n\right\} > n^{\frac{1}{2}-\kappa}(1+o(1))\right] \\ &\le \exp\{-2n^{1-2\kappa}(1+o(1))\} \to 0. \end{aligned}$$

Hence Condition (S1) is fulfilled.

Next, consider $Y_i \sim N(\mu_n,\sigma^2)$ and define $q_n' = P(Y_i < t_n)$. Then we have

$$q_n' = P\left\{Z_i > \frac{(\sqrt{r}-\sqrt{k})\sqrt{2\log n}}{\sigma}\right\} \le \frac{1}{2\sqrt{\pi\log n(\sqrt{r}-\sqrt{k})}} n^{-\frac{-(\sqrt{r}-\sqrt{k})^2}{\sigma^2}}.$$

Applying Hoeffding's inequality again we have

$$\begin{aligned} & P\left(\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}}\cap\mathcal{I}) < (1-\epsilon)\mathrm{Card}(\mathcal{I})|\right) \\ = \;& P\left[\sqrt{\mathrm{Card}}(\mathcal{I})\left\{\mathrm{Ave}_{i\in\mathcal{I}} I(X_i>t_n) - (1-q_n')\right\} < -\epsilon n^{\frac{1}{2}(1-\beta)}(1+o(1))\right] \\ \le \;& \exp\{-2\epsilon^2 n^{1-\beta}(1+o(1))\} \to 0. \end{aligned}$$

Hence Condition (S2) is fulfilled.

Proof of Part (ii). We focus on subsets of the form $\mathcal{S}_{\boldsymbol{\delta}} = \{i : X_i > t_n\}$ and show that there does not exist a threshold t_n such that both Conditions (S1) and (S2) are fulfilled. We consider the following threshold $t_n = \sqrt{2r\log n}$ and check condition (S2).

$$\begin{aligned} & P\left\{(\frac{1}{2}-\eta)\mathrm{Card}(\mathcal{I}) < \mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}}\cap\mathcal{I}) < (\frac{1}{2}+\eta)\mathrm{Card}(\mathcal{I})\right\} \\ = \;& P\left[\left|\sum_{i\in\mathcal{I}} I(X_i>t_n) - \frac{1}{2}\mathrm{Card}(\mathcal{I})\right| > \eta n^{1-\beta}(1+o(1))\right] \\ \le \;& 2\exp\{-2\eta^2 n^{1-\beta}(1+o(1))\} \to 0. \end{aligned}$$

We let $\eta \to 0$. The result indicates that with high probability, around half of the signals will be missed by the screening procedure and condition (S2) is violated. Therefore we must decrease t_n in order to include more signals in the screening set. However, t_n cannot be further decreased because we have already got too many noises even at this threshold level. Specifically, define $q_n = P(Z_i > t_n) \leq \frac{1}{2\sqrt{\pi r \log n}} n^{-r}$, then we have

$$\begin{aligned} P\left\{ \frac{|\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}} \cap \mathcal{N}) - \mathrm{Card}(\mathcal{N}) q_n|}{\mathrm{Card}(\mathcal{N})} > n^{-r} \right\} &= P\left\{ \frac{\left|\sum_{i \in \mathcal{N}} I(X_i > t_n) - q_n\right|}{\mathrm{Card}(\mathcal{N})} > n^{-r} \right\} \\ &\leq \exp\{-2n^{1-2r}(1+o(1))\} \to 0. \end{aligned}$$

Hence with overwhelming probability we have $\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}} \cap \mathcal{N})/\mathrm{Card}(\mathcal{N}) > q_n + n^{-r} = n^{-r}(1 + o(1)) > n^{-\kappa}$ for large n. Therefore Condition (S1) is violated. Therefore it is impossible to find a threshold which fulfill both Conditions (S1) and (S2) simultaneously. □

8.4. *Proof of Theorem 4*

We first state two lemmas. The first lemma summarizes the Bayes classification rule in a two point normal mixture model. The proof of the lemma follows some straightforward calculations and is omitted.

LEMMA 2. *Let θ_i, $i = 1, \ldots, n$, be independent Bernoulli(p_n) random variables. X_i are independent with $X_i|\theta_i = 0 \sim N(0,1)$ and $X_i|\theta_i = 1 \sim N(\mu_n, \sigma^2)$. For a classification rule $\boldsymbol{\delta}$ let the weighted misclassification rate be*

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = n^{-1} \sum_{i=1}^{n} \{\theta_i(1-\delta_i) + \lambda(1-\theta_i)\delta_i\}. \tag{8.3}$$

The optimal classification rule is summarized as follows.

(i) $0 < \sigma < 1$. *If* $\mu_n^2 \leq 2(1-\sigma^2)\log\frac{\sigma\lambda(1-\epsilon_n)}{\epsilon_n}$, *then the Bayes rule is* $\delta_i^\pi \equiv 0$. *If* $\mu_n^2 > 2(1-\sigma^2)\log\frac{\sigma\lambda(1-\epsilon_n)}{\epsilon_n}$, *then the Bayes rule is* $\delta_i^\pi = I(t_L < X_i < t_U)$, *where*

$$t_L = \frac{\mu_n - \sigma\sqrt{\mu_n^2 - 2(1-\sigma^2)\log\frac{\sigma\lambda(1-\epsilon_n)}{\epsilon_n}}}{1-\sigma^2} \text{ and } t_U = \frac{\mu_n + \sigma\sqrt{\mu_n^2 - 2(1-\sigma^2)\log\frac{\sigma\lambda(1-\epsilon_n)}{\epsilon_n})}}{1-\sigma^2}.$$

(ii) $\sigma = 1$. *The Bayes rule is* $\delta_i^\pi = I\left(X_i > \frac{\mu_n}{2} + \frac{1}{\mu_n}\log\frac{\lambda(1-\epsilon_n)}{\epsilon_n}\right)$.

(iii) $\sigma > 1$. *The Bayes rule is* $\delta_i^\pi = I\left(X_i < t_L\right) + I\left(X_i > t_U\right)$, *where*

$$t_L = \frac{-\mu_n - \sigma\sqrt{\mu_n^2 + 2(\sigma^2-1)\log\frac{\sigma\lambda(1-\epsilon_n)}{\epsilon_n}}}{\sigma^2-1} \text{ and } t_U = \frac{-\mu_n + \sigma\sqrt{\mu_n^2 + 2(\sigma^2-1)\log\frac{\sigma\lambda(1-\epsilon_n)}{\epsilon_n}}}{\sigma^2-1}.$$

Lemma 2 shows that the optimal classification rule has three possible forms: (a) $\delta_i = I(t_L < X_i < t_U)$ when $0 < \sigma < 1$; (b) $\delta_i = I(X_i > t)$ when $\sigma = 1$; (c) $\delta_i = I(X_i < t_L) + I(X_i > t_U)$ when $\sigma > 1$. The next lemma is proved in Section 9.

LEMMA 3. *Consider the homoscedastic case $\sigma = 1$. Let $\hat{t}_s^\pi$ be the threshold of Procedure 2 at screening level $\alpha_n' = n^{-\eta_0}$, where $\sqrt{2\eta_0} < \min\{\sqrt{r} - \sqrt{\kappa}, \sqrt{\frac{1-\beta}{2}}\}$. Then $\hat{t}_s^\pi$ is bounded below by a constant with probability tending to 1, i.e. $P(\hat{t}_s^\pi > (\sqrt{r} - \sqrt{\eta_0})\sqrt{2\log n}) \to 1$.*

Proof of Theorem 4. Some calculations reveal that, with the threshold adaptively chosen by Procedure 2, the optimal classification rule is asymptotically equivalent to a simple thresholding rule $\delta_i = I(X_i > t)$. It follows from Lemma 2 that we only need to focus on the homoscedastic case $\sigma = 1$. The heteroscedastic case $\sigma \neq 1$ can be proved similarly. Consider $\hat{t}_s^\pi$ defined in Lemma 3. The probability of interest is

$$\begin{aligned} & P(|\mathcal{S}_{\boldsymbol{\delta}} \cap \mathcal{N}| > \gamma_n^{-1}|\mathcal{N}|) = P(\mathrm{Ave}_{i\in\mathcal{N}} I(X_i > \hat{t}_s^\pi) > \gamma_n^{-1}) \\ \leq \;& P\left[\mathrm{Ave}_{i\in\mathcal{N}} I\left\{X_i > (\sqrt{r} - \sqrt{\eta_0})\sqrt{2\log n}\right\} > \gamma_n^{-1}\right] + P(\hat{t}_s^\pi \leq (\sqrt{r} - \sqrt{\eta_0})\sqrt{2\log n}) \end{aligned}$$

Applying the Hoeffding's inequality to the first term we have

$$P\left[\mathrm{Ave}_{i\in\mathcal{N}} I\left\{X_i > (\sqrt{r} - \sqrt{\eta_0})\sqrt{2\log n}\right\} > \gamma_n^{-1}\right] \leq e^{-2n^{1-2\kappa}(1+o(1))} \to 0.$$

The second term goes to 0 as $n \to \infty$ according to Lemma 3. Then the desired result follows. □

8.5. Proof of Theorem 5

We first state two lemmas, which are proved in Section 9.

LEMMA 4. *Consider a classification problem in the two-point normal mixture model (2.2).*

(i) If $\sigma = 1$, then the minimum expected misclassification rate satisfies

$$\inf_{\boldsymbol{\delta}} E[L(\boldsymbol{\theta}, \boldsymbol{\delta})] = \begin{cases} \frac{2r\sqrt{r}}{(r^2-\beta^2)\sqrt{\pi\log n}} n^{-\frac{(r+\beta)^2}{4r}}(1+o(1)). & \text{for } r > \beta \\ \frac{1}{2}\epsilon_n(1+o(1)) & \text{for } r = \beta \\ \epsilon_n(1+o(1)) & \text{for } r < \beta \end{cases}.$$

(ii) If $\sigma \neq 1$, then the minimum expected misclassification rate satisfies

$$\inf_{\boldsymbol{\delta}} E[L(\boldsymbol{\theta}, \boldsymbol{\delta})] = \begin{cases} c(r,\beta,\sigma)\cdot(\log n)^{-\frac{1}{2}} n^{-\frac{(\sqrt{r+(\sigma^2-1)\beta}-\sigma\sqrt{r})^2}{(\sigma^2-1)^2}}\epsilon_n(1+o(1)) & \text{for } r > \beta \\ \frac{1}{2}\epsilon_n(1+o(1)) & \text{for } r = \beta \\ \epsilon_n(1+o(1)) & \text{for } r \leq \beta \end{cases},$$

$$\text{where } c(r,\beta,\sigma) = \frac{(1-\sigma^2)(1-\sigma)(\sqrt{r}+\sqrt{r-(1-\sigma^2)\beta})}{2\sqrt{\pi}(\sqrt{r}-\sigma\sqrt{r-(1-\sigma^2)\beta})(\sqrt{r-(1-\sigma^2)\beta}-\sigma\sqrt{r})}. \tag{8.4}$$

LEMMA 5. *Consider model (2.2) and a screening subset $\mathcal{S}_{\boldsymbol{\delta}}$.*

(i) If $\sigma = 1$, then the expected size of the discovery set $\mathcal{S}_{\boldsymbol{\delta}}$ can be decomposed as $E[\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}})] = ETP + EFP$ with

$$\begin{aligned} EFP &= \frac{\sqrt{r}}{(r+\beta)\sqrt{\pi\log n}} n^{1-\frac{(r+\beta)^2}{4r}}(1+o(1)), \text{ and} \\ ETP &= \begin{cases} n\epsilon_n(1+o(1)) & \text{when } r > \beta \\ \frac{1}{2}n\epsilon_n(1+o(1)) & \text{when } r = \beta \\ \frac{\sqrt{r}}{(\beta-r)\sqrt{\pi\log n}} n^{1-\frac{(r+\beta)^2}{4r}}(1+o(1)) & \text{when } r < \beta \end{cases}. \end{aligned}$$

(ii) If $\sigma \neq 1$, then the expected size of the discovery set $\mathcal{S}_{\boldsymbol{\delta}}$ can be decomposed as $E[\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}})] = ETP + EFP$ with

$$EFP = \frac{\sigma^2-1}{2\sqrt{\pi \log n}(\sigma\sqrt{r+(\sigma^2-1)\beta}-\sqrt{r})} n^{1-\frac{(\sigma\sqrt{r+(\sigma^2-1)\beta}-\sqrt{r})^2}{(\sigma^2-1)^2}}(1+o(1)), \text{ and}$$

$$ETP = \begin{cases} n\epsilon_n(1+o(1)) & \text{when } r > \beta \\ \frac{1}{2}n\epsilon_n(1+o(1)) & \text{when } r = \beta \\ \frac{\sigma^2-1}{2\sqrt{\pi \log n}(\sqrt{r+(\sigma^2-1)\beta}-\sigma\sqrt{r})} \cdot n^{1-\frac{(\sigma\sqrt{r+(\sigma^2-1)\beta}-\sqrt{r})^2}{(\sigma^2-1)^2}}(1+o(1)) & \text{when } r < \beta \end{cases}.$$

Proof of Theorem 5 (i). It follows from Lemma 4 that, when $r > \beta$,

$$\frac{\inf_{\boldsymbol{\delta}} E[L(\boldsymbol{\theta},\boldsymbol{\delta})]}{\epsilon_n} = \begin{cases} O(n^{-\frac{(r-\beta)^2}{4r}}) & \text{if } \sigma = 1 \\ O(n^{-\frac{(\sqrt{r+(\sigma^2-1)\beta}-\sigma\sqrt{r})^2}{(\sigma^2-1)^2}}) & \text{if } \sigma \neq 1 \end{cases}.$$

In both cases we have $\inf_{\boldsymbol{\delta}} E[L(\boldsymbol{\theta},\boldsymbol{\delta})]/\epsilon_n \to 0$; hence the expected misclassification rate is negligible. Next, it follows from Lemma 5 and the equality $n^{\frac{-(\sigma\sqrt{r+(\sigma^2-1)\beta}-\sqrt{r})^2}{(\sigma^2-1)^2}} = n^{\frac{-(\sqrt{r+(\sigma^2-1)\beta}-\sigma\sqrt{r})^2}{(\sigma^2-1)^2}}\epsilon_n$ that, when $r > \beta$,

$$\frac{\mathrm{EFP}}{\mathrm{ETP}} = \begin{cases} O(n^{-\frac{(r-\beta)^2}{4r}}) & \text{if } \sigma = 1 \\ O(n^{-\frac{(\sqrt{r+(\sigma^2-1)\beta}-\sigma\sqrt{r})^2}{(\sigma^2-1)^2}}) & \text{if } \sigma \neq 1 \end{cases}.$$

This ratio is of the same order of magnitude of the previous ratio (but with different constants). For all $\sigma > 0$ we always have $\mathrm{EFP}/\mathrm{ETP} \to 0$; hence the signals are dominant in $\mathcal{S}_{\boldsymbol{\delta}}$ and the sure discovery property is established. Next, it again follows from Lemma 5 that, when $r > \beta$, $\mathrm{ETP}/(n\epsilon_n) = 1 + o(1)$ for all $\sigma > 0$ and $0 < \beta \leq 1$. Therefore virtually all signals are included in $\hat{\mathcal{S}}$ and the sure screening property is established.

Proof of Theorem 5 (ii). We first consider the case of $\sigma = 1$. In Lemma 5 we show that when $r < \beta$, the minimum expected misclassification rate is of the order $\epsilon_n(1+o(1))$; which is not negligible. Next, the Bayes threshold is shown to be $t_B = \frac{r+\beta}{\sqrt{2r}}\sqrt{\log n}$. The corresponding decision rule $\delta_i = I(X_i > t_B)$ yields a subset $\mathcal{S}_{\boldsymbol{\delta}}$ such that $\frac{\mathrm{EFP}}{\mathrm{ETP}} \to \frac{\beta-r}{\beta+r}$, as $n \to \infty$; hence the sure discovery property is violated. In order to construct a subset where the ETP dominates the EFP, we must choose a new threshold $t^* > t_B$. However, even with threshold t_B, the sure screening property is violated since $\mathrm{ETP}/\mathrm{Card}(\mathcal{I}) = O(n^{-\frac{(\beta-r)^2}{4r}}) \to 0$, and choosing a higher threshold implies losing even more signals. Hence it is impossible to construct a subset with both Properties (D1) and (D2).

Next we consider the heteroscedastic case. First it follows from Lemma 4 that the minimum misclassification rate must be of the order of $\epsilon_n(1+o(1))$; hence there is no classification rule yielding a negligible risk. It follows from Lemma 5 that the EFP and ETP are of the same order and $\frac{\mathrm{EFP}}{\mathrm{ETP}} \to \frac{\sqrt{r+(\sigma^2-1)\beta}-\sigma\sqrt{r}}{\sigma\sqrt{r+(\sigma^2-1)\beta}-\sqrt{r}}$, as $n \to \infty$. Hence we must alter the threshold so that the sure discovery property can be fulfilled. Take for example when $\sigma > 1$. Define $T_u = \frac{-\mu_n+\sigma\sqrt{\mu_n^2+2(\sigma^2-1)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{\sigma^2-1}$. The dominant parts of the ETP and EFP come from the following terms $n(1-\epsilon_n)\Phi(-T_u)$ and $n\epsilon_n\Phi\left(-\frac{T_u-\mu_n}{\sigma}\right)$, respectively. Thus we must increase T_u to fulfill condition (D1). However, a higher threshold implies that condition (D2) will be violated. Hence it is impossible to construct a subset with both properties (D1) and (D2). This completes the proof. □

8.6. Proof of Theorem 6

We prove the most complicated case with $0 < \sigma < 1$. The proofs for cases with $\sigma = 1$ and $\sigma > 1$ are provided in Section 9. We need to show that (i) when $r > \rho_{dis}(\beta)$, then we can construct a subset which fulfills conditions (D1) and (D2); and (ii) when $r < \rho_{dis}(\beta)$, then it is impossible construct a subset which fulfills both conditions (D1) and (D2). We first state a few lemmas, which are also proved in Section 9.

LEMMA 6. *For all* $0 < \beta < 1$ *and* $0 < \sigma < 1$, *we have* $(1 - \sigma\sqrt{1-\beta})^2 \geq \beta(1-\sigma^2)$.

LEMMA 7. *Consider the discovery boundary defined in Theorem 6 for the case of* $0 < \sigma < 1$. *If* $r > \rho_{\mathrm{dis}}(\beta)$, *then* $\beta + \left\{\frac{\sigma\sqrt{r} - \sqrt{r-(1-\sigma^2)\beta}}{1-\sigma^2}\right\}^2 < 1$.

Proof of part (i). The discovery boundary $\rho_{\mathrm{dis}}(\beta)$ and Lemma 6 together imply that if $r > \rho_{\mathrm{dis}}(\beta)$ then $r > (1-\sigma^2)\beta$. Therefore the following threshold

$$t^* = \left\{\frac{\sqrt{r} - \sigma\sqrt{r - \beta(1-\sigma^2)} + \sigma\epsilon_0}{1-\sigma^2}\right\}\sqrt{2\log n}$$

is always well-defined when we consider the region above the discovery boundary. Consider decision rule $\delta_i = I(X_i > t^*)$. It is easy to show that $t^* > \sqrt{2r\log n}$; hence the ETP and EFP can be calculated as

$$\begin{aligned}
\mathrm{ETP} &= \frac{1-\sigma^2}{2\sqrt{\pi\log n}\{\sigma\sqrt{r} + \epsilon_0 - \sqrt{r-(1-\sigma^2)\beta}\}} n^{-\beta - \left\{\frac{\sigma\sqrt{r}+\epsilon_0 - \sqrt{r-(1-\sigma^2)\beta}}{1-\sigma^2}\right\}^2}(1+o(1)) \text{ and} \\
\mathrm{EFP} &= \frac{1-\sigma^2}{2\sqrt{\pi\log n}\{\sqrt{r} + \sigma\epsilon_0 - \sigma\sqrt{r-\beta(1-\sigma^2)}\}} n^{-\left\{\frac{\sqrt{r}+\sigma\epsilon_0 - \sigma\sqrt{r-\beta(1-\sigma^2)}}{1-\sigma^2}\right\}^2}(1+o(1)),
\end{aligned}$$

respectively. Note that

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

It follows from some algebra that

$$\frac{\mathrm{ETP}}{\mathrm{EFP}} = \frac{\sqrt{r} + \sigma\epsilon_0 - \sigma\sqrt{r-\beta(1-\sigma^2)}}{\sigma\sqrt{r} + \epsilon_0 - \sqrt{r-(1-\sigma^2)\beta}} n^{\epsilon_0\left\{2\sqrt{r-(1-\sigma^2)\beta} - \epsilon_0\right\}} \to \infty.$$

Hence we can find $\epsilon_0 < 2\sqrt{r-(1-\sigma^2)\beta}$ such that the signal is dominant in the discovery set and condition (D1) is fulfilled.

To show that condition (D2) is fulfilled we need to show there exists a small ϵ_0 such that the discovery set is nonempty with probability tending to 1. Define $\zeta_n = P(X_i > t^*)$. The above arguments imply that

$$\zeta_n = \frac{1-\sigma^2}{2\sqrt{\pi\log n}\{\sigma\sqrt{r} + \epsilon_0 - \sqrt{r-(1-\sigma^2)\beta}\}} n^{-\beta - \left\{\frac{\sigma\sqrt{r}+\epsilon_0 - \sqrt{r-(1-\sigma^2)\beta}}{1-\sigma^2}\right\}^2}(1+o(1)).$$

Lemma 7 shows $r > \rho_{\mathrm{dis}}(\beta)$ always implies that $\beta + \left\{\frac{\sigma\sqrt{r} - \sqrt{r-(1-\sigma^2)\beta}}{1-\sigma^2}\right\}^2 < 1$. Therefore we can find $0 < \epsilon_0 < 2\sqrt{r-(1-\sigma^2)\beta}$ and $\kappa > 0$ such that

$$\beta + \left\{\frac{\sigma\sqrt{r} + \epsilon_0 - \sqrt{r-(1-\sigma^2)\beta}}{1-\sigma^2}\right\}^2 + \kappa \leq 1.$$

Then with threshold t^*, the probability of having an non-empty discovery set is

$$P\{|\mathcal{S}_{\boldsymbol{\delta}}| \geq 1\} = 1 - (1-\zeta_n)^n = 1 - e^{-c_n n^{-\kappa}}(1+o(1)) \to 0.$$

Therefore condition (D2) holds.

Proof of part (ii). If $r < (1-\sigma^2)\beta$, pick a q such that $r < q < 1$. The corresponding threshold is $\sqrt{2q\log n}$. Consider the ratio

$$\lambda_n = \frac{n^{1-\beta}P\{N(\mu_n,\sigma^2) > \sqrt{2q\log n}\}}{(n-n^{1-\beta})P\{N(0,1) > \sqrt{2q\log n}\}} = \frac{\sigma\sqrt{q}}{\sqrt{q}-\sqrt{r}} n^{q-\beta-\frac{(\sqrt{q}-\sqrt{r})^2}{\sigma^2}}(1+o(1)).$$

Now for the growth rate of the ratio we have

$$f(q) = q - \beta - \frac{(\sqrt{q}-\sqrt{r})^2}{\sigma^2} = -\frac{1-\sigma^2}{\sigma^2}\left(\sqrt{q} - \frac{\sqrt{r}}{1-\sigma^2}\right)^2 + \frac{r}{1-\sigma^2} - \beta.$$

It follows that $\lambda_n \to 0$ if $r < (1-\sigma^2)\beta$ for all values of $0 < q < 1$ (note that the case of $0 < q \leq r$ is trivial), i.e. the noises are dominant *everywhere*. Therefore if $r < (1-\sigma^2)\beta$, then it is impossible to construct a subset fulfills condition (D1).

Hence it is sufficient to only consider the case where $r > (1-\sigma^2)\beta$. The optimal decision rule must be of the form

$$\delta_i = I(t_L < X_i < t_U) \tag{8.5}$$

with the center of the interval $t_C = \frac{\sqrt{r}}{1-\sigma^2}\sqrt{2\log n}$. We first argue that the lower limit t_L should be at least as large as

$$t_* = \left\{\frac{\sqrt{r} - \sigma\sqrt{r-\beta(1-\sigma^2)}}{1-\sigma^2}\right\}\sqrt{2\log n}$$

in order for the signals to be dominant. As $t_L = t_*$, a rejection region of the form (8.5) satisfy

$$\begin{aligned}
\text{ETP} &= \frac{1-\sigma^2}{2\sqrt{\pi\log n}\{\sigma\sqrt{r} - \sqrt{r-(1-\sigma^2)\beta}\}} n^{-\beta-\left\{\frac{\sigma\sqrt{r}-\sqrt{r-(1-\sigma^2)\beta}}{1-\sigma^2}\right\}^2}(1+o(1)), \text{ and} \\
\text{EFP} &= \frac{1-\sigma^2}{2\sqrt{\pi\log n}\{\sqrt{r} - \sigma\sqrt{r-\beta(1-\sigma^2)}\}} n^{-\left\{\frac{\sqrt{r}-\sigma\sqrt{r-\beta(1-\sigma^2)}}{1-\sigma^2}\right\}^2}(1+o(1)),
\end{aligned}$$

respectively. It follows that

$$\frac{\text{ETP}}{\text{EFP}} = \frac{\sqrt{r} - \sigma\sqrt{r-\beta(1-\sigma^2)}}{\sigma\sqrt{r} - \sqrt{r-(1-\sigma^2)\beta}},$$

and hence Condition (D1) is violated. As $t_L < t_*$, a rejection region of the form (8.5) would have a even lower ratio of ETP/EFP. Therefore we must have $t_L > t_*$.

According to the definition of the discovery boundary, we only need to show that if $1-\sigma^2 < \beta < 1$ and $\sqrt{r} + \sigma\sqrt{1-\beta} < 1$, the choice of any $t_L > t*$ would lead to an empty discovery set with probability tending to 1. Note that if $\sqrt{r}+\sigma\sqrt{1-\beta} < 1$ and $1-\sigma^2 < \beta < 1$, then we have

$$\begin{aligned}
\sqrt{r} + \sigma\sqrt{1-\beta} < 1 &\iff r - 2\sqrt{r} + 1 > \sigma^2(1-\beta) \\
&\iff (1-\sigma^2)^2 - 2\sqrt{r}(1-\sigma^2) + r > \sigma^2\{r - \beta(1-\sigma^2)\} \\
&\iff \{r-(1-\sigma^2)\}^2 > (\sigma\sqrt{r-\beta(1-\sigma^2)})^2
\end{aligned}$$

Therefore we can find $\kappa > 0$ such that $\frac{\{\sqrt{r}-\sigma\sqrt{r-\beta(1-\sigma^2)}\}^2}{(1-\sigma^2)^2} = 1+\kappa$. Define $\zeta_n = P(t_L < X_i < t_U)$. Then we have $\zeta_n < P(X_i > t_*) = c(r,\beta,\sigma)n^{-\frac{\left\{\sqrt{r}-\sigma\sqrt{r-\beta(1-\sigma^2)}\right\}^2}{(1-\sigma^2)^2}}(1+o(1))$, where

$$c(r,\beta,\sigma) = \frac{(1-\sigma^2)(1+\sigma)(\sqrt{r}-\sqrt{r-(1-\sigma^2)\beta})}{2\sqrt{\pi\log n}(\sqrt{r}-\sigma\sqrt{r-\beta(1-\sigma^2)})(\sigma\sqrt{r}-\sqrt{r-\beta(1-\sigma^2)})}.$$

Let $\mathcal{S}_{\boldsymbol{\delta}}$ be the discovery set. Then

$$P(\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}}) \geq 1) \leq 1 - \exp\{-c_n n^{-\kappa}\}(1+o(1)) \to 0.$$

The desired result follows by combining (i) and (ii). □

8.7. Proof of Theorem 7

Consider the decision rule of the form $\delta_i = I(\mathrm{Lfdr}_i < \alpha_n)$. The goal is to show that the minL method is fully efficient in constructing a signal-dominant subset when $r > \rho_{\mathrm{dis}}(\beta)$. Define the discovery set $\hat{\mathcal{S}} = \{i : \mathrm{Lfdr}_i < \alpha_n\}$. Let $\zeta_n = P(\mathrm{Lfdr}_i < \alpha_n)$. Note that $\mathrm{Lfdr}_i < \alpha_n$ implies that

$$(1-\epsilon_n)e^{-X_i^2/2} < \alpha_n\left\{(1-\epsilon_n)e^{-X_i^2/2} + \frac{\epsilon_n}{\sigma}e^{-(X_i-\mu_n)^2/(2\sigma^2)}\right\}.$$

This further implies that

$$\frac{1-\sigma^2}{2\sigma^2}X_i^2 - \frac{\mu_n}{\sigma^2}X_i + \frac{\mu_n^2}{2\sigma^2} + \log\sigma + \log\frac{1-\epsilon_n}{\epsilon_n} + \log\frac{1-\alpha_n}{\alpha_n} < 0.$$

Let $\alpha_n = n^{-\eta_0}$. We shall specify the range of η_0 at a later time. For the present we take it as a small positive constant. It can be show that the above equation can be simplified as

$$\left\{\frac{1-\sigma^2}{2\sigma^2}\left(X_i - \frac{\mu_n}{1-\sigma^2}\right)^2 + \left(\beta+\eta_0-\frac{r}{1-\sigma^2}\right)\log n\right\}(1+o(1)) < 0. \tag{8.6}$$

We have shown that $\sqrt{r}+\sigma\sqrt{1-\beta} > 1$ implies that $r > (1-\sigma^2)\beta$ (Lemma 6). Hence we shall choose an η_0 such that $r > (1-\sigma^2)(\beta+\eta_0)$. Let $d_0 = r-(1-\sigma^2)(\beta+\eta_0)$. Then the above equation can be solved and it follows from the solution that

$$\begin{aligned}
\zeta_n &= P\left[\frac{\sqrt{2\log n}\{\sqrt{r}-\sigma\sqrt{d_0}\}}{1-\sigma^2} < X_i < \frac{\sqrt{2\log n}\{\sqrt{r}+\sigma\sqrt{d_0}\}}{1-\sigma^2}\right](1+o(1)) \\
&= P\left[X_i > \frac{\sqrt{2\log n}\{\sqrt{r}-\sigma\sqrt{d_0}\}}{1-\sigma^2}\right](1+o(1)) \\
&= \frac{1-\sigma^2}{2\sqrt{\pi\log n}(\sqrt{r}-\sigma\sqrt{d_0})}n^{-\frac{\left\{\sqrt{r}-\sigma\sqrt{d_0}\right\}^2}{(1-\sigma^2)^2}}(1+o(1)) \\
&\quad + \frac{1-\sigma^2}{2\sqrt{\pi\log n}\{\sigma\sqrt{r}-\sqrt{d_0}\}}n^{-\beta-\frac{\left\{\sigma\sqrt{r}-\sqrt{d_0}\right\}^2}{(1-\sigma^2)^2}}(1+o(1)) \\
&= \frac{1-\sigma^2}{2\sqrt{\pi\log n}\{\sigma\sqrt{r}-\sqrt{d_0}\}}n^{-\beta-\frac{\left\{\sigma\sqrt{r}-\sqrt{d_0}\right\}^2}{(1-\sigma^2)^2}}(1+o(1)).
\end{aligned}$$

The last equation holds since $\beta + \frac{\left\{\sigma\sqrt{r}-\sqrt{d_0}\right\}^2}{(1-\sigma^2)^2} - \frac{\left\{\sqrt{r}-\sigma\sqrt{d_0}\right\}^2}{(1-\sigma^2)^2} = -\frac{\eta_0}{1-\sigma^2} < 0$. Consider the regions of $1-\sigma^2 < \beta < 1$ and $0 < \beta \le 1-\sigma^2$, respectively. Similar as before, we can show that $r > \rho_{\rm dis}(\beta) \Longleftrightarrow \frac{\{\sqrt{r}-\sigma\sqrt{r-\beta(1-\sigma^2)}\}^2}{(1-\sigma^2)^2} < 1$ in both regions. Therefore we can choose $\eta_0 > 0$ such that both equations $r > (1-\sigma^2)(\beta+\eta_0)$ and $\frac{\{\sqrt{r}-\sigma\sqrt{r-(\beta+\eta_0)(1-\sigma^2)}\}^2}{(1-\sigma^2)^2} < 1$ are satisfied. Let $\alpha_n = n^{-\eta_0}$ and consider the thresholding rule $\delta_i = I(\mathrm{Lfdr}_i < \alpha_n)$. Denote by $\hat{\mathcal{S}}$ the discovery set. We need to establish the following results: (i) $P(\mathrm{Card}(\hat{\mathcal{S}}) \ge 1) \to 1$; and (ii) ETP/EFP $\to \infty$. To show (i), define $\frac{\{\sqrt{r}-\sigma\sqrt{r-\beta(1-\sigma^2)}\}^2}{(1-\sigma^2)^2} = 1-\kappa$, then $\kappa > 0$ and it follows that $P\{\mathrm{Card}(\hat{\mathcal{S}}) \ge 1\} = 1 - \exp\{-c(r,\beta,\sigma)n^{\kappa}\}(1+o(1)) \to 1$. To show (ii), we calculate the ETP and EFP of the discovery set:

$$\begin{aligned}
\mathrm{EFP} &= \frac{1-\sigma^2}{2\sqrt{\pi\log n}(\sqrt{r}-\sigma\sqrt{d_0})} n^{1-\frac{\left\{\sqrt{r}-\sigma\sqrt{d_0}\right\}^2}{(1-\sigma^2)^2}}(1+o(1)),\\
\mathrm{ETP} &= \frac{1-\sigma^2}{2\sqrt{\pi\log n}\{\sigma\sqrt{r}-\sqrt{d_0}\}} n^{1-\beta-\frac{\left\{\sigma\sqrt{r}-\sqrt{d_0}\right\}^2}{(1-\sigma^2)^2}}(1+o(1)).
\end{aligned}$$

It follows that $\frac{\mathrm{ETP}}{\mathrm{EFP}} = \frac{\sqrt{r}-\sigma\sqrt{d_0}}{\sigma\sqrt{r}-\sqrt{d_0}} n^{\eta_0/(1-\sigma^2)} \to \infty$. The desired result follows (i) and (ii). □

8.8. *Proof of Theorem 8*

The boundary for minL method follows Theorem 7. To establish the screening boundary for the minP method, we need to show the following

(i) If $\sqrt{r}+\sigma\sqrt{1-\beta} < 1$, then we can find a threshold t_n and corresponding decision rule $\delta_i = I(p_i < t_n)$ such that $P(|S| \ge 1) \to 1$ and $P\left(\min_{\{i:\theta_i=1\}} p_i \le t_n\right) \to 0$. This shows that the most significant observations are from the null distribution with probability tending to 1. Hence if we select the signals according to the ranking given by the p-values, then we will almost always start with a subset which only contains observations from the null distribution.

(ii) If $\sqrt{r}+\sigma\sqrt{1-\beta} > 1$, then we can find a $\tau > 1$ such that $\sqrt{r}+\sigma\sqrt{1-\beta} > \sqrt{\tau}$. Define $N_1(\tau) = \#\{i : X_i \ge \sqrt{2(1+\tau)\log n}\}$ and $N_2(\tau) = \#\{i : X_i \ge \sqrt{2(1+\tau)\log n}\ \&\ \theta_i = 0\}$. Then we can show that $P\{N_1(\tau) \ge 1\} \to 1$ and $P\{N_2(\tau) = 0\} \to 1$. Hence the smallest p-value comes from the non-null distribution with probability tending to 1.

Proof of (i). Let τ_1 and τ_2 be constants such that $\sqrt{r}+\sigma\sqrt{\tau_1-\beta} < \sqrt{r}+\sigma\sqrt{1-\beta} < \sqrt{\tau_2} < \sqrt{\tau_1} < 1$. If a non-null case $X_i \le \sqrt{\tau_2\log n}$, then $p_i \ge P\{X_i > \sqrt{2\tau_2\log n}\} \ge n^{-\tau_1}$. Hence

$$\begin{aligned}
P_{\theta_i=1}(p_i < n^{-\tau_1}) &\le P_{\theta_i=1}(X_i > \sqrt{2\tau_2\log n})\\
&= \frac{\sigma}{2\sqrt{\pi\log n}(\sqrt{\tau}_2-\sqrt{r})} n^{-(\sqrt{\tau}_2-\sqrt{r})^2/\sigma^2}(1+o(1)).
\end{aligned}$$

Let $\kappa = (\sqrt{\tau_2} - \sqrt{r})^2/\sigma^2 - (1-\beta)$, then $\kappa > 0$. Therefore

$$\begin{aligned} P(\min_{\{i:\theta_i=1\}} p_i \le n^{-\tau_1}) &= 1 - \prod_{\{i:\theta_i=1\}} \{1 - P_{\theta_i=1}(p_i < n^{-\tau_1})\} \\ &= 1 - \exp\left\{n^{1-\beta}\log\{1 - P_{\theta_i=1}(p_i < n^{-\tau_1})\}\right\} \\ &= 1 - \exp\left\{-n^{1-\beta}P_{\theta_i=1}(p_i < n_1^{-\tau})\}(1+o(1))\right\} \\ &\le 1 - \exp\left\{-c_n n^{1-\beta-(\sqrt{\tau}_2-\sqrt{r})^2/\sigma^2}(1+o(1))\right\} \\ &= 1 - \exp\{-c_n n^{-\kappa}(1+o(1))\} \to 0. \end{aligned}$$

Meanwhile, it follows from $\sqrt{r} + \sigma\sqrt{\tau_1 - \beta} < \sqrt{\tau_1} < 1$ that

$$\begin{aligned} & P(X_i > \sqrt{2\tau_1 \log n}) \\ = & \frac{1}{2\sqrt{\pi\tau_1 \log n}} n^{-\tau_1}(1+o(1)) + \frac{\sigma}{2\sqrt{\pi\tau_1 \log n}(\sqrt{\tau}_1 - \sqrt{r})} n^{-\beta-(\sqrt{\tau}_1-\sqrt{r})^2/\sigma^2}(1+o(1)) \\ = & \frac{1}{2\sqrt{\pi\tau_1 \log n}} n^{-\tau_1}(1+o(1)). \end{aligned}$$

It is easy to show that $P(|S| \ge 1) \to 1$. Therefore the threshold $n^{-\tau_1}$ for the p-value would yield a nonempty subset with all observations coming from the null distribution.

Proof of (ii). First it is important to note that the non-nulls are dominant at $\sqrt{(1+\tau)\log n}$ for $\tau > 1$. Specifically, define $f(x) = 1 - \beta - \frac{(\sqrt{1+x}-\sqrt{2r})^2}{2\sigma^2}$. Since $f(1) = 1 - \beta - \frac{(1-\sqrt{r})^2}{\sigma^2} > 0$ (by assumption), we can find $\tau > 1$ such that $f(\tau) > 0$. If $X_i > \sqrt{(1+\tau)\log n}$, then $p(X_i) < P\{N(0,1) > \sqrt{(1+\tau)\log n}\} < \alpha/n$. It follows that

$$\begin{aligned} P(p_i < \alpha/n) &> P(X_i > \sqrt{(1+\tau)\log n}) \\ &= \frac{\sigma}{\sqrt{2\pi \log n}(\sqrt{1+\tau} - \sqrt{2r})} n^{-\beta - \frac{(\sqrt{1+\tau}-\sqrt{2r})^2}{2\sigma^2}}(1+o(1)) \\ &\quad + \frac{1}{\sqrt{2\pi(1+\tau)\log n}} n^{-\frac{1+\tau}{2}}(1+o(1)) \\ &= \frac{\sigma}{\sqrt{2\pi \log n}(\sqrt{1+\tau} - \sqrt{2r})} n^{-\beta - \frac{(\sqrt{1+\tau}-\sqrt{2r})^2}{2\sigma^2}}(1+o(1)) \\ &> n^{-\beta - \frac{(\sqrt{\tau}-\sqrt{r})^2}{\sigma^2}}. \end{aligned}$$

The last equality holds by noting that $\beta + \frac{(\sqrt{1+\tau}-\sqrt{2r})^2}{2\sigma^2} < 1 < \frac{1+\tau}{2}$. Let $\kappa = 1 - \beta - \frac{(\sqrt{\tau}-\sqrt{r})^2}{\sigma^2} > 0$. Then we have $P(p_i < \alpha/n) > n^{-1+\kappa}$. Note that

$$\begin{aligned} P(p_{(1)} < \alpha_n) &= 1 - \left\{1 - P\left(p_i < \frac{\alpha}{n}\right)\right\}^n \\ &> 1 - (1 - n^{-1+\kappa})^n \\ &= 1 - \exp(-n^\kappa)(1+o(1)), \end{aligned}$$

we have $P(p_{(1)} < \alpha_n) \to 1$. It is easy to show that $P\{N_1(\tau) \ge 1\} \to 1$ and $P\{N_2(\tau) = 0\} \to 1$. Hence the smallest p-value comes from the non-null distribution with probability 1. Hence the BH procedure rejects a non-null with probability tending to 1 if $\sqrt{r} + \sigma\sqrt{1-\beta} > 1$. □

Acknowledgments
We thank the Editor, Associate Editor and two referees for their thorough and useful comments which have helped to improve the presentation of the paper. In particular, we are grateful to one of the referees whose comment provides important intuitions and insights behind the difference between the minL and minP methods stated in Remark 6. Tony Cai was supported in part by NSF Grant DMS-1208982 and NIH Grant R01 CA127334. Wenguang Sun's research was supported in part by NSF Grant DMS-CAREER-1255406.

References

Agresti, J. J., E. Antipov, A. R. Abate, K. Ahn, A. C. Rowat, J.-C. Baret, M. Marquez, A. M. Klibanov, A. D. Griffiths, and D. A. Weitz (2010). Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *P. Natl. A. Sci. 107*(9), 4004–4009.

Bartroff, J. (2007). Asymptotically optimal multistage tests of simple hypotheses. *Ann. Statist. 35*(5), 2075–2105.

Benjamini, Y. and R. Heller (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc. 102*, 1272–1281.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* ***57***, 289–300.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer.

Birmingham, A., L. M. Selfors, T. Forster, D. Wrobel, C. J. Kennedy, E. Shanks, J. Santoyo-Lopez, D. J. Dunican, A. Long, D. Kelleher, et al. (2009). Statistical methods for analysis of high-throughput rna interference screens. *Nat. Methods 6*(8), 569–575.

Blanchard, G. and D. Geman (2005). Hierarchical testing designs for pattern recognition. *Ann. Statist. 33*, 1155–1202.

Bleicher, K. H., H.-J. Böhm, K. Müller, and A. I. Alanine (2003). Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug. Discov. 2*(5), 369–378.

Bloma, P. E., S. J. Fleischerb, and Z. Smilowitzb (2002). Spatial and temporal dynamics of colorado potato beetle in fields with perimeter and spatially targeted insecticides. *Environ. Entomol. 31*(1), 149–159.

Cai, T. T., J. Jin, and M. G. Low (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist. 35*(6), 2421–2449.

Cai, T. T. and Y. Wu (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE T. Inform. Theory 60*(4), 2217–2232.

Dmitrienko, A., B. L. Wiens, A. C. Tamhane, and X. Wang (2007). Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Stat. Med. 26*(12), 2465–2478.

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

Donoho, D. and J. Jin (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist. 34*, 2980–3018.

Dove, A. et al. (2003). Screening for content-the evolution of high throughput. *Nat. Biotechnol. 21*(8), 859–864.

Durrieu, G. and L. Briollais (2009). Sequential design for microarray experiments. *J. Amer. Statist. Assoc. 104*(486), 650–660.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc. 99*(465), 96–104.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc. 96*, 1151–1160.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. B 70*(5), 849–911.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. B* ***64***, 499–517.

Goeman, J. J. and U. Mansmann (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics 24*(4), 537–544.

Goeman, J. J. and A. Solari (2010). The sequential rejection principle of familywise error control. *Ann. Statist. 38*(6), 3782–3810.

Haupt, J., R. M. Castro, and R. Nowak (2011). Distilled Sensing: Adaptive Sampling for Sparse Detection and Estimation. *IEEE T. Inform. Theory 57*(9), 6222–6235.

Haupt, J., R. Nowak, and R. Castro (2009). Adaptive sensing for sparse signal recovery. In *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. DSP/SPE 2009. IEEE 13th*, pp. 702–707. IEEE.

Ingster, Y. I. (1998). Minimax detection of a signal for l^n-balls. *Math. Methods Statist. 7*(4), 401–428 (1999).

Ji, P. and J. Jin (2012). Ups delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist. 40*(1), 73–103.

Jin, J. and T. T. Cai (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* ***102***, 495–506.

Lai, T. L. (2000). Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems. *IEEE T. Inform. Theory 46*(2), 595–608.

Lin, D. Y. (2006). Evaluating statistical significance in two-stage genomewide association studies. *Am. J. Hum. Genet. 78*, 505–509.

Malo, N., J. A. Hanley, S. Cerquozzi, J. Pelletier, and R. Nadon (2006). Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol. 24*(2), 167–175.

McKoy, A. F., J. Chen, T. Schupbach, and M. H. Hecht (2012). A novel inhibitor of amyloid β (aβ) peptide aggregation from high throughput screening to efficacy in an animal model of alzheimer disease. *J. Biol. Chem. 287*(46), 38992–39000.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika 95*(2), 265–278.

Meinshausen, N. and J. Rice (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* ***34***, 373–393.

Müller, P., G. Parmigiani, C. Robert, and J. Rousseau (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Amer. Statist. Assoc. 99*(468), 990–1001.

Posch, M., S. Zehetmayer, and P. Bauer (2009). Hunting for significance with the false discovery rate. *J. Amer. Statist. Assoc. 104*(486), 832–840.

Rossell, D. and P. Müller (2013). Sequential stopping for high-throughput experiments. *Biostatistics 14*(1), 75–86.

Sarkar, S. K. (2004). Fdr-controlling stepwise procedures and their false negatives rates. *J. Stat. Plan. Infer. 125*(1), 119–137.

Satagopan, J. M., E. Venkatraman, and C. B. Begg (2004). Two-stage designs for gene–disease association studies with sample size constraints. *Biometrics 60*(3), 589–597.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Volume 26. CRC press.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. B* ***64***, 479–498.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Statist. Soc. B 66*(1), 187–205.

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

Taylor, J., R. Tibshirani, and B. Efron (2005). The miss ratefor the analysis of gene expression data. *Biostatistics 6*(1), 111–117.

Wasserman, L. and K. Roeder (2009). High-dimensional variable selection. *Ann. Statist. 37*, 2178–2201.

Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology. *J. Amer. Statist. Assoc. 103*(481), 309–316.

Zehetmayer, S., P. Bauer, and M. Posch (2008). Optimized multi-stage designs controlling the false discovery or the family-wise error rate. *Stat. Med. 27*(21), 4145–4160.

Zhang, X. D. (2011). *Optimal high-throughput screening: practical experimental design and data analysis for genome-scale RNAi research.* Cambridge University Press.

9. Supplemental Materials: Proof of Other Results

9.1. Proof of Lemma 1

The Lemma is a restatement of Theorem 2 in Sun and Cai (2007). We provide the proof here for completeness. The joint distribution of $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$ is $\pi(\boldsymbol{\theta}) = \prod_i (1-\epsilon_n)^{1-\theta_i}\epsilon_n^{\theta_i}$. The posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{x}$ can be calculated as $P_{\boldsymbol{\theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{x}) = \prod_i P_{\theta_i|X_i}(\theta_i|x_i)$, where

$$P_{\theta_i|X_i}(\theta_i|x_i) = \frac{I(\theta_i=0)(1-\epsilon_n)f_0(x_i) + I(\theta_i=1)\epsilon_n f_1(x_i)}{(1-\epsilon_n)f_0(x_i) + \epsilon_n f_1(x_i)}.$$

The posterior risk is then given by

$$E_{\boldsymbol{\theta}|\mathbf{X}} L(\boldsymbol{\theta}, \boldsymbol{\delta}) = n^{-1}\sum_i \frac{\epsilon_n f_1(x_i)}{f(x_i)} + n^{-1}\sum_i \frac{\lambda(1-\epsilon_n)f_0(x_i) - \epsilon_n f_1(x_i)}{f(x_i)}\delta_i.$$

Then the Bayes classification rule is $\boldsymbol{\delta}^{\pi,\lambda} = (\delta_1^{\pi,\lambda}, \cdots, \delta_n^{\pi,\lambda})$, where

$$\delta_i^{\pi,\lambda} = I\{\lambda(1-\epsilon_n)f_0 < \epsilon_n f_1\}$$

and (8.2) follows.

9.2. Proof of Lemma 3

Let $Q(t)$ denote the MDR level of the thresholding rule $I(X_i > t)$. We argue by contradiction. Assume that there exists a $\varepsilon > 0$ such that $P(\hat{t}_s^\pi \le (\sqrt{r} - \sqrt{\eta_0})\sqrt{2\log n}) \ge \varepsilon$ for all n. Let A_n denote the event. It follows that on event A_n,

$$\begin{aligned} Q(\hat{t}_s^\pi) &\le Q\left((\sqrt{r} - \sqrt{\eta_0})\sqrt{2\log n}\right) \\ &= P_{f_1}\left\{X_i < (\sqrt{r} - \sqrt{\eta_0})\sqrt{2\log n}\right\} \\ &= \frac{1}{2\sqrt{\pi\eta_0 \log n}} n^{-\eta_0}(1+o(1)) \end{aligned}$$

Then on event A_n such that $P(A_n) \ge \varepsilon$,

$$\alpha_n' - Q(\hat{t}_s^\pi) = \left(1 - \frac{1}{2\sqrt{\pi\eta_0\log n}}\right) n^{-\eta_0}(1+o(1)) \ge \frac{1}{2} n^{-\eta_0} \tag{9.1}$$

for large n. Define $\hat{Q}(t)$ be the estimated MDR of thresholding rule $I(X_i > t)$. Then

$$\hat{Q}(\hat{t}_s^\pi) = (n\epsilon_n)^{-1}\sum_i I(X_i < \hat{t}_s^\pi)(1 - T_i^\pi) = Q(\hat{t}_s^\pi) + O_p\left(n^{-\frac{1-\beta}{2}}\right). \tag{9.2}$$

According to the definition of $\hat{Q}(\hat{t}_s^\pi)$ and the operation of Procedure 2, we always have

$$\alpha_n' - \frac{1}{n\epsilon_n} \le \hat{Q}(\hat{t}_s^\pi) \le \alpha_n'. \tag{9.3}$$

Combining (9.2) and (9.3) we conclude that $n^{2\eta_0}|\alpha_n' - Q(\hat{t}_s^\pi)| = O_p(1)$. This is a contradiction to (9.1). Hence the lemma is proved. □

9.3. Proof of Lemma 4

We first state a lemma, which will be used as the main technical tool in the proof of other lemmas. We omit the proof, which follows some standard calculations of the Bayes risk in a classification problem.

LEMMA 8. *Let θ_i, $i = 1, \ldots, n$, be independent Bernoulli(ϵ_n) random variables. X_i are independent observations from model $X_i|\theta_i = 0 \sim N(0,1)$ and $X_i|\theta_i = 1 \sim N(\mu_n, \sigma^2)$. For a classification rule $\boldsymbol{\delta}$ based on X let the misclassification rate be $L(\boldsymbol{\theta}, \boldsymbol{\delta}) = n^{-1}\sum_{i=1}^{n} I(\theta_i \neq \delta_i)$. Let $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. Then the minimum expected misclassification rate satisfies:*

(i) For $\sigma = 1$, $\inf_{\boldsymbol{\delta}} E(L(\boldsymbol{\theta}, \boldsymbol{\delta})) = (1-\epsilon_n)\Phi(-\frac{\mu_n}{2\sigma} - \frac{\sigma}{\mu_n}\log\frac{1-\epsilon_n}{\epsilon_n}) + \epsilon_n\Phi(-\frac{\mu_n}{2\sigma} + \frac{\sigma}{\mu_n}\log\frac{1-\epsilon_n}{\epsilon_n})$.

(ii) For $0 < \sigma < 1$, if $\mu_n^2 \leq 2(1-\sigma^2)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}$, $\inf_{\boldsymbol{\delta}} E(L(\boldsymbol{\theta}, \boldsymbol{\delta})) = \epsilon_n$; if $\mu_n^2 > 2(1-\sigma^2)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}$,

$$\begin{aligned}
&\inf_{\boldsymbol{\delta}} E(L(\boldsymbol{\theta}, \boldsymbol{\delta})) \\
= \;& (1-\epsilon_n)\left(\Phi(\frac{\mu_n + \sigma\sqrt{\mu_n^2 - 2(1-\sigma^2)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{1-\sigma^2}) - \Phi(\frac{\mu_n - \sigma\sqrt{\mu_n^2 - 2(1-\sigma^2)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{1-\sigma^2})\right) \\
&+ \epsilon_n\left(\Phi(\frac{\sigma\mu_n - \sqrt{\mu_n^2 - 2(1-\sigma^2)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{1-\sigma^2}) + \Phi(\frac{-\sigma\mu_n - \sqrt{\mu_n^2 - 2(1-\sigma^2)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{1-\sigma^2})\right).
\end{aligned}$$

(iii) For $\sigma > 1$,

$$\begin{aligned}
&\inf_{\boldsymbol{\delta}} E(L(\boldsymbol{\theta}, \boldsymbol{\delta})) = \\
&(1-\epsilon_n)\left\{\Phi(\frac{-\mu_n - \sigma\sqrt{\mu_n^2 + 2(\sigma^2-1)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{\sigma^2-1}) + \Phi(\frac{\mu_n - \sigma\sqrt{\mu_n^2 + 2(\sigma^2-1)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{\sigma^2-1})\right\} \\
&+ \epsilon_n\left\{\Phi(\frac{-\sigma\mu_n + \sqrt{\mu_n^2 + 2(\sigma^2-1)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{\sigma^2-1}) - \Phi(\frac{-\sigma\mu_n - \sqrt{\mu_n^2 + 2(\sigma^2-1)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{\sigma^2-1})\right\}.
\end{aligned}$$

Proof of Lemma 4 (i). We shall only consider the case $r > \beta$. The other two cases follow similar but simpler arguments. By plugging in $\epsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r\log n}$ into the result in part (i) of Lemma 8, we have

$$\inf_{\boldsymbol{\delta}} E(L(\boldsymbol{\theta}, \boldsymbol{\delta})) \quad = \quad (1-\epsilon_n)\Phi(-\frac{r+\beta}{\sqrt{2r}}\sqrt{\log n}) + \epsilon_n\Phi(-\frac{r-\beta}{\sqrt{2r}}\sqrt{\log n}).$$

It then follows the standard approximation for the Gaussian tail probability, $\Phi(-x) = \frac{1}{\sqrt{2\pi}x}e^{-\frac{1}{2}x^2}(1+o(1))$ as $x \to \infty$, that

$$\begin{aligned}
\inf_{\boldsymbol{\delta}} E(L(\boldsymbol{\theta}, \boldsymbol{\delta})) &= \frac{\sqrt{r}}{(r+\beta)\sqrt{\pi\log n}}n^{-\frac{(r+\beta)^2}{4r}}(1+o(1)) + \frac{\sqrt{r}}{(r-\beta)\sqrt{\pi\log n}}n^{-\frac{(r+\beta)^2}{4r}}(1+o(1)) \\
&= \frac{2r\sqrt{r}}{(r^2-\beta^2)\sqrt{\pi\log n}}n^{-\frac{(r+\beta)^2}{4r}}(1+o(1)).
\end{aligned}$$

Proof of Lemma 4 (ii). We shall only consider the case $0 < \sigma < 1$ and $r > (1-\sigma^2)\beta$. The case for $\sigma > 1$ can be proved similarly. By plugging in $\epsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r\log n}$ into the result in part (ii) of Lemma 8, we have

$$\begin{aligned}\inf_{\boldsymbol{\delta}} E(L(\boldsymbol{\theta},\boldsymbol{\delta})) &= \Phi(-\frac{\sqrt{r}-\sigma\sqrt{r-(1-\sigma^2)\beta}}{1-\sigma^2}\sqrt{2\log n})(1+o(1)) \\ &\quad + \epsilon_n\Phi(-\frac{\sqrt{r-(1-\sigma^2)\beta}-\sigma\sqrt{r}}{1-\sigma^2}\sqrt{2\log n})(1+o(1)).\end{aligned}$$

It follows that $\inf_{\boldsymbol{\delta}} E(L(\boldsymbol{\theta},\boldsymbol{\delta})) = \frac{1}{2}\epsilon_n(1+o(1))$ when $r=\beta$ and $\inf_{\boldsymbol{\delta}} E(L(\boldsymbol{\theta},\boldsymbol{\delta})) = \epsilon_n(1+o(1))$ when $r<\beta$. For $r>\beta$ the standard approximation for the Gaussian tail probability yields

$$\begin{aligned}\inf_{\boldsymbol{\delta}} E(L(\boldsymbol{\theta},\boldsymbol{\delta})) &= \frac{1-\sigma^2}{2\sqrt{\pi}(\sqrt{r}-\sigma\sqrt{r-(1-\sigma^2)\beta})}(\log n)^{-\frac{1}{2}} n^{-\frac{(\sqrt{r}-\sigma\sqrt{r-(1-\sigma^2)\beta})^2}{(1-\sigma^2)^2}}(1+o(1)) \\ &\quad + \frac{1-\sigma^2}{2\sqrt{\pi}(\sqrt{r-(1-\sigma^2)\beta}-\sigma\sqrt{r})}(\log n)^{-\frac{1}{2}} n^{-\frac{(\sqrt{r-(1-\sigma^2)\beta}-\sigma\sqrt{r})^2}{(1-\sigma^2)^2}}\epsilon_n(1+o(1)) \\ &= c(r,\beta,\sigma)\cdot(\log n)^{-\frac{1}{2}} n^{-\frac{(\sqrt{r-(1-\sigma^2)\beta}-\sigma\sqrt{r})^2}{(1-\sigma^2)^2}}\epsilon_n(1+o(1)),\end{aligned}$$

where $c(r,\beta,\sigma)$ is given by (8.4). Note that in the last step we have used the fact that $n^{\frac{-(\sigma\sqrt{r+(\sigma^2-1)\beta}-\sqrt{r})^2}{(\sigma^2-1)^2}} = n^{\frac{-(\sqrt{r+(\sigma^2-1)\beta}-\sigma\sqrt{r})^2}{(\sigma^2-1)^2}}\epsilon_n$. □

9.4. Proof of Lemma 5

Part (i). The threshold in the Bayes rule is $t_B = \frac{\mu_n}{2} + \frac{\sigma^2}{\mu_n}\log\frac{1-\epsilon_n}{\epsilon_n} = \frac{r+\beta}{\sqrt{2r}}\sqrt{\log n}$. Then

$$\begin{aligned}E[\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}})] &= nP(X_i > t_B) = n\epsilon_n P(X_i > T_B|\theta_i=1) + n(1-\epsilon_n)P(X_i > T_B|\theta_i=0) \\ &= n\epsilon_n\Phi(\frac{r-\beta}{\sqrt{2r}}\sqrt{\log n}) + n(1-\epsilon_n)\Phi(-\frac{r+\beta}{\sqrt{2r}}\sqrt{\log n}) \\ &\equiv \mathrm{ETP}+\mathrm{EFP}.\end{aligned}$$

The result follows from straightforward calculations using Gaussian tail approximation.

Part (ii). We only consider the case $\sigma > 1$. Set

$$T_l = \frac{-\mu_n - \sigma\sqrt{\mu_n^2 + 2(\sigma^2-1)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{\sigma^2-1}$$

and

$$T_u = \frac{-\mu_n + \sigma\sqrt{\mu_n^2 + 2(\sigma^2-1)\log\frac{\sigma(1-\epsilon_n)}{\epsilon_n}}}{\sigma^2-1},$$

then the expected size of the discovery set is

$$\begin{aligned}E[\mathrm{Card}(\mathcal{S}_{\boldsymbol{\delta}})] &= n(P(X_i < T_l) + P(X_i > T_u)) \\ &= n\epsilon_n\{P(X_i < T_l|\theta_i=1) + P(X_i > T_u|\theta_i=1)\} \\ &\quad + n(1-\epsilon_n)\{P(X_i < T_l|\theta_i=0) + P(X_i > T_u|\theta_i=0)\} \\ &= n\epsilon_n\left\{\Phi(\frac{T_l-\mu_n}{\sigma}) + \Phi(-\frac{T_u-\mu_n}{\sigma})\right\} + n(1-\epsilon_n)(\Phi(T_l)+\Phi(-T_u)) \\ &\equiv \mathrm{ETP}+\mathrm{EFP}.\end{aligned}$$

The rest of the proof follows similar arguments as in the proof of Part (i). □

9.5. Proof of Lemma 6

Consider function $f(\beta) = (1-\sigma\sqrt{1-\beta})^2 - \beta(1-\sigma^2)$. It is easy to show that $f'(\beta) = \frac{\sigma}{\sqrt{1-\beta}} - 1$. We claim that $f(\beta)$ achieves its minimum value at $1-\sigma^2$ since $f'(\beta) > 0$ if $1-\sigma^2 < \beta < 1$ and $f'(\beta) < 0$ if $0 < \beta < 1-\sigma^2$. The desired result follows by noting that $f(\beta)|_{\beta=1-\sigma^2} = 0$. □

9.6. Proof of Lemma 7

(i). Consider the region $\sqrt{r} + \sigma\sqrt{1-\beta} > 1$ and $1-\sigma^2 < \beta < 1$. Then we have

$$\begin{aligned}
\sqrt{r} + \sigma\sqrt{1-\beta} > 1 &\iff r - 2\sqrt{r} + 1 < \sigma^2(1-\beta) \\
&\iff (1-\sigma^2)^2 - 2\sqrt{r}(1-\sigma^2) + r < \sigma^2\{r - \beta(1-\sigma^2)\} \\
&\iff \{r - (1-\sigma^2)\}^2 < (\sigma\sqrt{r - \beta(1-\sigma^2)})^2
\end{aligned}$$

The result follows from the fact that $\beta + \left\{\frac{\sigma\sqrt{r} - \sqrt{r-(1-\sigma^2)\beta}}{1-\sigma^2}\right\}^2 = \left\{\frac{\sqrt{r} - \sigma\sqrt{r-\beta(1-\sigma^2)}}{1-\sigma^2}\right\}^2$.

(ii). In the region $r > (1-\sigma^2)\beta$ and $0 < \beta < 1-\sigma^2$, consider two situations:

(a) $\sqrt{r} \le 1-\sigma^2$. It follows that

$$\begin{aligned}
&\sqrt{r} - (1-\sigma^2) < \sigma\sqrt{r - (1-\sigma^2)\beta} \\
\iff\ &\sqrt{r} - \sigma\sqrt{r - (1-\sigma^2)\beta} < (1-\sigma^2) \\
\iff\ &\frac{\{\sqrt{r} - \sigma\sqrt{r - \beta(1-\sigma^2)}\}^2}{(1-\sigma^2)^2} < 1.
\end{aligned}$$

(b) $\sqrt{r} > 1-\sigma^2$. Note that $0 < \beta \le 1-\sigma^2$ implies that $1-\beta \ge \sigma^2$, we have $\sigma^2(1-\beta) \ge \sigma^4 > (1-\sqrt{r})^2$. Finally, note that

$$\begin{aligned}
&\sigma^2(1-\beta) > (1-\sqrt{r})^2 \\
\iff\ &\{\sqrt{r} - (1-\sigma^2)\}^2 < \{\sigma\sqrt{r - \beta(1-\sigma^2)}\}^2 \\
\iff\ &\sqrt{r} - \sigma\sqrt{r - \beta(1-\sigma^2)} < 1-\sigma^2 \\
\iff\ &\frac{\{\sqrt{r} - \sigma\sqrt{r - \beta(1-\sigma^2)}\}^2}{(1-\sigma^2)^2} < 1.
\end{aligned}$$

The result follows by combining (i) and (ii). □

9.7. Proof of Theorem 6: other cases

We will need the following lemmas in the proof.

LEMMA 9. *Let $X_1, \cdots, X_n$ be iid rv with $0 \le X_i \le 1$ and $E(X_i) = \mu_i$, for $i = 1, \cdots, n$. Let $\bar{\mu} = n^{-1}\sum_i \mu_i$. Then for $0 < \bar{\mu} < 1/2$,*

$$P\{\sqrt{n}(\bar{X} - \bar{\mu}) \ge \lambda\} \le e^{-2\lambda^2}.$$

This is standard Hoeffding's inequality and the proof is omitted.

LEMMA 10. *Consider the discovery boundary defined in Theorem 6 for the case of $\sigma > 1$. If $r > \rho_{\rm dis}(\beta)$, then $\beta + \left\{\frac{\sqrt{r+(\sigma^2-1)\beta} - \sigma\sqrt{r}}{\sigma^2-1}\right\}^2 < 1$.*

Proof of Lemma 10: Consider the region $0 < \beta \le 1 - \sigma^{-2}$. It is easy to see that

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

Next consider the region $1 - \sigma^{-2} < \beta < 1$, if $r > \rho_{\text{dis}}(\beta)$, then we have $\sqrt{r} + \sigma\sqrt{1-\beta} > 1$. The rest of the proof follows similar lines of that of Lemma 7.

LEMMA 11. *If $\sqrt{r} + \sqrt{1-\beta} < 1$, then $q < \beta + (\sqrt{q} - \sqrt{r})^2$ for all $\beta < q < 1$.*

Proof of Lemma 11: Define $f(x) = \sqrt{x} - \sqrt{x-\beta} - \sqrt{r}$ for $\beta < x < 1$. Then

$$f'(x) = \frac{1}{2\sqrt{x}} - \frac{1}{2\sqrt{x-\beta}} < 0.$$

Therefore for all $\beta < q < 1$ we have

$$f(q) = \sqrt{q} - \sqrt{q-\beta} - \sqrt{r} > f(1) = 1 - \sqrt{1-\beta} - \sqrt{r} > 0$$

and the result follows. □

9.7.1. Proof of the homoscedastic case: $\sigma = 1$

First we note that it is sufficient to only consider the case where $r < \beta$ since if $r > \beta$ the signals and noises can be nearly perfectly separated.

(i) When $r > \rho_{\text{dis}}(\beta) = (1 - \sqrt{1-\beta})^2$, we have $(r+\beta)^2/4r < 1$. Therefore we can find $\epsilon_0 > 0$ such that $(r+\beta+\epsilon_0)^2/4r < 1$. Consider the threshold $t^* = \frac{(\beta+r+\epsilon_0)\sqrt{\log n}}{\sqrt{2r}}$. The goal is to show that both properties (D1) and (D2) are fulfilled by the decision rule $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_n)$, where $\delta_i = I(X_i > t^*)$. Let $\mathcal{S}_{\boldsymbol{\delta}} = \{i : \delta_i = 1\}$ be the discovery set and define $\zeta_n = P(X_i > t^*)$. When $r < \beta$, we have

$$\begin{aligned}
\zeta_n &= \frac{\sqrt{r}}{(\beta+r+\epsilon_0)\sqrt{\pi\log(n)}} n^{-\frac{(\beta+r+\epsilon_0)^2}{4r}}(1+o(1)) + \frac{\sqrt{r}}{(\beta-r+\epsilon_0)\sqrt{\pi\log(n)}} n^{-\beta-\frac{(\beta-r+\epsilon_0)^2}{4r}}(1+o(1)) \\
&= \frac{2\sqrt{r}}{(\beta-r+\epsilon_0)\sqrt{\pi\log(n)}} n^{-\frac{(r+\beta+\epsilon_0)^2}{4r}+\epsilon_0}(1+o(1)).
\end{aligned}$$

Let $c_n = \frac{2\sqrt{r}}{(\beta-r+\epsilon_0)\sqrt{\pi\log(n)}}$. It follows, by the choice of ϵ_0, that

$$\begin{aligned}
P\{\text{Card}(\mathcal{S}_{\boldsymbol{\delta}}) \ge 1\} &= 1 - (1-\zeta_n)^n \\
&\ge 1 - \{1 - c_n n^{-1+\epsilon_0}(1+o(1))\}^n \\
&= 1 - e^{-c_n n^{\epsilon_0}}(1+o(1)) \to 1.
\end{aligned}$$

Hence Condition (A) is fulfilled. Next note that in the discovery set S, the ratio of the expected numbers of true positives and false positives can be calculated as

$$\frac{\text{ETP}}{\text{EFP}} = \frac{\beta+r+\epsilon_0}{\beta-r+\epsilon_0} n^{\epsilon_0}(1+o(1)) \to \infty.$$

Hence the signals are dominant and Condition (B) is fulfilled.

(ii) When $r > \rho_{\text{dis}}(\beta) = (1 - \sqrt{1-\beta})^2$, the goal is to show that there does not exist a threshold such that both conditions (A) and (B) are fulfilled. We first argue that the threshold

cannot be smaller than $t_* = \frac{\beta+r}{\sqrt{2r}}\sqrt{\log n}$. As before we define $\zeta_n = P(X_i > t_*)$. It can be shown that the ratio of the ETP and EFP can be calculated as $\frac{\text{ETP}}{\text{EFP}} = \frac{\beta+r}{\beta-r}(1+o(1))$ when t_* is chosen. In order for the signals to be dominant in the discovery set, we must choose a threshold that is greater than t_*. However, the assumption that $r < \rho_{\text{dis}}(\beta)$ indicates that any threshold that is great than t_* would result in an empty discovery set with probability tending to 1. Specifically, if $r < \rho_{\text{dis}}(\beta)$, then $(r+\beta)^2/4r > 1$. Define $\kappa = (r+\beta)^2/4r - 1$. Then $\kappa > 0$, and with the most conservative threshold t_*, the probability of having an non-empty discovery set is

$$\begin{aligned} P\{|\mathcal{S}_{\boldsymbol{\delta}}| \geq 1\} &= 1-(1-\zeta_n)^n \\ &= 1-\{1-c_n n^{-1-\kappa}(1+o(1))\}^n \\ &= 1-e^{-c_n n^{-\kappa}}(1+o(1)) \to 0, \end{aligned}$$

and we end up with an empty subset with high probability. Therefore we conclude that conditions (A) and (B) cannot be fulfilled simultaneously when $r < \rho_{\text{dis}}(\beta)$. The desired result follows by combining (i) and (ii). □

9.7.2. Heteroscedastic case: $\sigma > 1$

Since much of the proof is similar to that of the case of $0 < \sigma < 1$, we shall only outline the main steps.

(i) According to Lemma 10, if $r > \rho_{\text{dis}}(\beta)$, then there exists $\epsilon_0 > 0$ such that $\beta + \left\{\frac{\sqrt{r+(\sigma^2-1)\beta}+\epsilon_0-\sigma\sqrt{r}}{\sigma^2-1}\right\}^2 < 1$. Correspondingly we can choose the following threshold $t^* = \frac{\sigma\sqrt{r+(\sigma^2-1)\beta}-\sqrt{r}+\sigma\epsilon_0}{\sigma^2-1}$. The goal is to show that the decision rule $\delta_i = I(X_i > t^*)$ fulfills both Conditions (A) and (B). First, note that $t^* > \sqrt{r}$, the ETP and EFP can be calculated as Gaussian tail probabilities:

$$\begin{aligned} \text{ETP} &= \frac{\sigma^2-1}{2\sqrt{\pi\log n}\{\sqrt{r+(\sigma^2-1)\beta}+\epsilon_0-\sigma\sqrt{r}\}} n^{-\beta-\left\{\frac{\sqrt{r+(\sigma^2-1)\beta}+\epsilon_0-\sigma\sqrt{r}}{\sigma^2-1}\right\}^2}(1+o(1)), \\ \text{EFP} &= \frac{\sigma^2-1}{2\sqrt{\pi\log n}\left\{\sigma\sqrt{r+(\sigma^2-1)\beta}+\sigma\epsilon_0-\sqrt{r}\right\}} n^{-\left\{\frac{\sigma\sqrt{r+(\sigma^2-1)\beta}+\sigma\epsilon_0-\sqrt{r}}{\sigma^2-1}\right\}^2}(1+o(1)). \end{aligned}$$

It follows from $\beta + \frac{\left\{\sqrt{r+(\sigma^2-1)\beta}-\sigma\sqrt{r}\right\}^2}{(\sigma^2-1)^2} = \frac{\left\{\sigma\sqrt{r+\beta(\sigma^2-1)}-\sqrt{r}\right\}^2}{(\sigma^2-1)^2}$ that

$$\frac{\text{ETP}}{\text{EFP}} = \frac{\sigma\sqrt{r+(\sigma^2-1)\beta}+\sigma\epsilon_0-\sqrt{r}}{\sqrt{r+(\sigma^2-1)\beta}+\epsilon_0-\sigma\sqrt{r}} n^{\frac{\sigma^2\epsilon_0\left\{\epsilon_0+2\sqrt{r+(\sigma^2-1)\beta}\right\}}{(\sigma^2-1)^2}} \to \infty.$$

Therefore Condition (A) is fulfilled. It can be shown similarly as the case of $0 < \sigma < 1$ that the choice of ϵ_0 ensures that $P(\text{Card}(\mathcal{S}_{\boldsymbol{\delta}}) \geq 1) \to 1$; hence Condition (B) is fulfilled.

(ii) It is sufficient to consider the case where $1-\sigma^{-2} < \beta < 1$. The optimal decision rule is of the form $\delta_i = I(X_i < t_L) + I(X_i > t_U)$. We shall show that t_U would be at least as large as $t_* = \frac{\sigma\sqrt{r+(\sigma^2-1)\beta}-\sqrt{r}}{\sigma^2-1}$. Otherwise the noises will be dominant. However, it can be shown similarly as before that, when $\sqrt{r}+\sigma\sqrt{1-\beta} < 1$, even with the conservative threshold t_*, we will essentially end up with an empty discovery set with probability tending to 1. The details of the arguments are omitted. Therefore both Conditions (A) and (B) cannot be fulfilled simultaneously. The desired result follows by combining (i) and (ii). □

9.8. The most informative part of the sample

In this section, we provide more details for the results in Section 7. The following three test statistics will be discussed in turn: the p-value, the Lfdr statistic, and the HC statistic.

(1). The p-value (PV) procedure. Consider $0 < q \le 1$. Let $\mu_n = \sqrt{2r\log n}$. It is easy to show that the numbers of nulls and non-nulls on the right hand side of $\sqrt{2q\log n}$ are

$$\begin{aligned} n \cdot P\left\{N(0,1) > \sqrt{2q\log n}\right\} &= \frac{1}{2\sqrt{q\pi\log n}} n^{1-q}(1+o(1)), \text{ and} \\ n^{1-\beta} \cdot P\left\{N(\mu_n,\sigma^2) > \sqrt{2q\log n}\right\} &= \frac{\sigma}{2\sqrt{\pi\log n}(\sqrt{q}-\sqrt{r})} n^{1-\beta-(\sqrt{q}-\sqrt{r})^2/\sigma^2}(1+o(1)), \end{aligned}$$

respectively. The PV procedure rejects a non-null with probability 1 if the number of non-nulls grows to ∞ and if the non-nulls are dominant in the tails. Equivalently, the following two conditions

(i) $1-\beta-\frac{(\sqrt{q}-\sqrt{r})^2}{\sigma^2} > 0$; and

(ii) $1-q < 1-\beta-\frac{(\sqrt{q}-\sqrt{r})^2}{\sigma^2}$ are satisfied simultaneously for $q=1$.

The second condition is implied by the first condition if $0 < q \le 1$. Hence when $q=1$, the conditions reduce to $\sqrt{r} > 1-\sigma\sqrt{1-\beta}$, for all $\sigma > 0$. Therefore

- For $0 < \sigma < 1$, the PV rejection boundary is $\rho_{PV}(\beta) = (1-\sigma\sqrt{1-\beta})^2$.
- For $\sigma > 1$, the PV rejection boundary is $\rho_{PV}(\beta) = \begin{cases} (1-\sigma\sqrt{1-\beta})^2 & \text{if } 1-\frac{1}{\sigma^2} < \beta < 1 \\ 0 & \text{if } 0 < \beta \le 1-\frac{1}{\sigma^2} \end{cases}$.

(2). The Lfdr procedure. In order for Lfdr method to work, we require that the number of non-nulls grows to ∞ and the non-nulls are dominant. First we need to find the most informative q that optimizes the growth rate of the ratio

$$\lambda_n = \frac{n^{1-\beta}\cdot P\left\{N(\mu_n,\sigma^2) > \sqrt{2q\log n}\right\}}{n\cdot P\left\{N(0,1) > \sqrt{2q\log n}\right\}} = \frac{\sigma\sqrt{q}}{\sqrt{q}-\sqrt{r}} n^{q-\beta-\frac{(\sqrt{q}-\sqrt{r})^2}{\sigma^2}}(1+o(1)).$$

Let $f(q) = q-\beta-\frac{(\sqrt{q}-\sqrt{r})^2}{\sigma^2}$.

(a) If $0 < \sigma < 1$, then $f(q) = -\frac{1-\sigma^2}{\sigma^2}\left(\sqrt{q}-\frac{\sqrt{r}}{1-\sigma^2}\right)^2 + \frac{r}{1-\sigma^2} - \beta$.

(i) If $\sqrt{r} \le 1-\sigma^2$, then the growth rate of the ratio is optimized at $q_{\text{Lfdr}} = \frac{r}{(1-\sigma^2)^2}$. It is easy to see that if $\frac{r}{1-\sigma^2} - \beta > 0$, then the number of non-nulls goes to ∞ and are dominant at q_{Lfdr}. We can solve β from the following equation

$$1-\sigma^2 = 1-\sigma\sqrt{1-\beta}$$

to obtain the changing point is $1-\sigma^2$. Therefore the rejection boundary is $\rho_{Lfdr}(\beta) = (1-\sigma^2)\beta$ for $0 < \beta < 1-\sigma^2$.

(ii) If $\sqrt{r} > 1-\sigma^2$, we only require $f(1) = 1-\beta-\frac{(1-\sqrt{r})^2}{\sigma^2} > 0$ so that the number of non-nulls goes to ∞ and are dominant at $q=1$. Therefore the Lfdr rejection boundary is $\rho_{Lfdr}(\beta) = (1-\sigma\sqrt{1-\beta})^2$ for $1-\sigma^2 < \beta < 1$.

(b) Next we consider $\sigma > 1$. Note that now $f(q)$ can be written as

$$f(q) = \frac{\sigma^2 - 1}{\sigma^2}\left(\sqrt{q} - \frac{\sqrt{r}}{1-\sigma^2}\right)^2 + \frac{r}{1-\sigma^2} - \beta.$$

We only require $f(1) = 1 - \beta - \frac{(1-\sqrt{r})^2}{\sigma^2} > 0$ so that the number of non-nulls goes to ∞ and are dominant at $q = 1$, where the conditions reduce to $\sqrt{r} > 1 - \sigma\sqrt{1-\beta}$. Therefore the Lfdr rejection boundary is

$$\rho_{Lfdr}(\beta) = \begin{cases} (1 - \sigma\sqrt{1-\beta})^2 & \text{if } 1 - \frac{1}{\sigma^2} < \beta < 1 \\ 0 & \text{if } 0 < \beta \le 1 - \frac{1}{\sigma^2} \end{cases}.$$

The Lfdr rejection boundary is overlapped with the PV rejection boundary.

(3). The HC procedure. First we need to find the most informative q that optimizes the growth rate of the normalized uniform empirical process

$$W_n \quad = \quad L_n n^{(1+q)/2 - \beta - (\sqrt{q}-\sqrt{r})^2/\sigma^2}(1 + o(1)).$$

Let $f(q) = (1+q)/2 - \beta - (\sqrt{q} - \sqrt{r})^2/\sigma^2$.

(a) We first consider $0 < \sigma < \sqrt{2}$. It is easy to show that

$$f(q) = -\frac{2 - \sigma^2}{2\sigma^2}\left(\sqrt{q} - \frac{2\sqrt{r}}{2-\sigma^2}\right)^2 + \frac{r}{2-\sigma^2} + \frac{1}{2} - \beta.$$

(i) If $\sqrt{r} \le 1 - \sigma^2/2$, then the growth rate of the ratio is optimized at $q_{\text{HC}} = \frac{4r}{(1-\sigma^2/2)^2}$. It is easy to see that if $\frac{2\sqrt{r}}{2-\sigma^2} + 1/2 - \beta > 0$, then the number of non-nulls goes to ∞ and are dominant at q_{HC}. The changing point of β can be solved from

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

Therefore the rejection boundary is $\rho_{HC}(\beta) = 0$ if $0 < \beta \le 1/2$, and $\rho_{HC}(\beta) = (2 - \sigma^2)(\beta - 1/2)$ if $1/2 < \beta \le 1 - \sigma^2/4$.

(ii) If $\sqrt{r} > 1 - \sigma^2/2$ (or equivalently, $1 - \sigma^2/4 < \beta < 1$), the most informative q is $q_{\text{HC}} = 1$. We only require that $f(1) = 1 - \beta - \frac{(1-\sqrt{r})^2}{\sigma^2} > 0$ so that the number of non-nulls goes to ∞ and are dominant at $q = 1$. Therefore the HC detection boundary is $\rho_{HC}(\beta) = (1 - \sigma\sqrt{1-\beta})^2$ for $1 - \sigma^2/4 < \beta < 1$.

(b) Next we consider $\sigma = \sqrt{2}$. Note that $f(q)$ can be written as $f(q) = \sqrt{rq} + \frac{1-r}{2} - \beta$. Obviously the most informative q is $q_{\text{HC}} = 1$. We only require $f(1) = 1 - \beta - \frac{(1-\sqrt{r})^2}{2} > 0$ so that the number of non-nulls goes to ∞ and are dominant at $q = 1$. Therefore the HC detection boundary is $\rho_{HC}(\beta) = \{1 - \sqrt{2(1-\beta)}\}^2$ for all $0 < \beta < 1$. The HC detection boundary is overlapped with the PV/Lfdr boundary.

(c) Finally we consider $\sigma > \sqrt{2}$. Note that now $f(q)$ can be written as

$$f(q) = \frac{\sigma^2 - 2}{2\sigma^2}\left(\sqrt{q} - \frac{2\sqrt{r}}{2-\sigma^2}\right)^2 + \frac{r}{2-\sigma^2} + \frac{1}{2} - \beta.$$

Again, the most informative q is $q_{\text{HC}} = 1$. We only require $f(1) = 1 - \beta - \frac{(1-\sqrt{r})^2}{\sigma^2} > 0$ so that the number of non-nulls goes to ∞ and are dominant at $q = 1$. The conditions reduce to $\sqrt{r} > 1 - \sigma\sqrt{1-\beta}$. Therefore the HC rejection boundary is

$$\rho_{HC}(\beta) = \begin{cases} (1 - \sigma\sqrt{1-\beta})^2 & \text{if } 1 - \frac{1}{\sigma^2} < \beta < 1 \\ 0 & \text{if } 0 < \beta \le 1 - \frac{1}{\sigma^2} \end{cases}.$$

Again, the HC detection boundary is overlapped with the PV/Lfdr discovery boundary. The reason is that all procedures look for non-nulls at the tail area.

- For $0 < \sigma < \sqrt{2}$, we have $\rho_{HC}(\beta) = \begin{cases} 0 & \text{if } 0 < \beta \le \frac{1}{2} \\ (2-\sigma^2)(\beta - \frac{1}{2}) & \text{if } \frac{1}{2} < \beta \le 1 - \frac{\sigma^2}{4} \\ (1 - \sigma\sqrt{1-\beta})^2 & \text{if } 1 - \frac{\sigma^2}{4} < \beta < 1 \end{cases}$.
- For $\sigma = \sqrt{2}$, we have $\rho_{HC}(\beta) = \begin{cases} 0 & \text{if } 0 < \beta \le \frac{1}{2} \\ \{1 - \sqrt{2(1-\beta)}\}^2 & \text{if } \frac{1}{2} < \beta < 1 \end{cases}$.
- For $\sigma > \sqrt{2}$, the detection boundary is $\rho_{HC}(\beta) = \begin{cases} (1 - \sigma\sqrt{1-\beta})^2 & \text{if } 1 - \frac{1}{\sigma^2} < \beta < 1 \\ 0 & \text{if } 0 < \beta \le 1 - \frac{1}{\sigma^2} \end{cases}$.