Journal of Machine Learning Research 25 (2024) 1-59 Submitted 11/22; Revised 9/24; Published 9/24

Sparse Recovery With Multiple Data Streams: An Adaptive Sequential Testing Approach

Weinan Wang WWANG@SNAP.COM
Snap Inc.
2850 Ocean Park Blvd.
Santa Monica, CA, 90405, USA

Bowen Gang BGANG@FUDAN.EDU.CN
Department of Statistics and Data Science
Fudan University
Shanghai, 200433, China

Wenguang Sun WGSUN@ZJU.EDU.CN
School of Management and Center for Data Science
Zhejiang University
Hangzhou, 310013, China

Editor: Pradeep Ravikumar

Abstract

Multistage design has been utilized across a variety of scientific fields, enabling the adaptive allocation of sensing resources to effectively eliminate null locations and localize signals. We present a decision-theoretic framework for multi-stage adaptive testing that minimizes the total number of measurements while ensuring pre-specified constraints on both the false positive rate (FPR) and the missed discovery rate (MDR). Our method, SMART, explicitly addresses the often-overlooked aspect of uncertainty quantification in machine learning algorithms, incorporating it at every decision stage. This enables SMART to respond adaptively to important patterns in the data streams, adjusting its decisions based on the strength of evidence at specific locations. By leveraging technical tools and key concepts from multiple testing, adaptive thresholding, and compound decision theory, SMART not only enhances the aggregation of information across individual tests but also allows for varying thresholds tailored to the observed data, thereby ensuring effective error rate control and resulting in significant savings on total study costs. Through comprehensive analyses of large-scale A/B tests, high-throughput screening, and image analysis, we demonstrate that our approach yields substantial efficiency gains and improved control over error rates compared to existing methodologies.

Keywords: Compound Decision Problem; Distilled Sensing; False Discovery Rate; Missed Discovery Rate; Sequential Probability Ratio Test

1. Introduction

In high-dimensional data analysis, recovering the support of a sparse vector is a fundamental challenge. We study the problem of estimating the support of a high-dimensional sparse

vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p) \in \mathbb{R}^p$ using measurements obtained from a set of variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$. Specifically, for the jth measurement of X_i (where $i \in [p] = \{1, 2, \ldots, p\}$), the observed value $x_{ij} = \mu_i + \epsilon_{ij}$, where ϵ_{ij} denotes the measurement error. Let $\mathcal{S} = \{i : \mu_i \neq 0\}$ represent the indices of the non-zero entries of $\boldsymbol{\mu}$, which constitute the support of the vector. We focus on a setup where measurements on variables are performed sequentially. The sequential framework not only facilitates the accumulation of information over time but also allows for adaptive strategies to enhance recovery accuracy and efficiency.

1.1 A multistage mixture model

We assume that the observed data obey a *multistage random mixture model*, which can be described through a hierarchical framework. Let $\mathbb{I}(\cdot)$ denote an indicator function, and define $\theta_i = \mathbb{I}(\mu_i \neq 0) = \mathbb{I}\{i \in \mathcal{S}\}$. The variable θ_i follows a Bernoulli distribution:

$$\theta_i \sim \text{Bernoulli}(\pi), \quad i = 1, \ldots, p, \tag{1.1}$$

where $\pi = \mathbb{P}(\mu_i \neq 0)$ denotes the expected proportion of non-zero entries within the vector. The unobserved effect sizes μ_i are generated conditionally based on the value of θ_i as follows:

$$\mu_i = \begin{cases} 0, & \text{if } \theta_i = 0, \\ \mu_i \sim G(\cdot), & \text{if } \theta_i = 1, \end{cases} \tag{1.2}$$

where $G(\cdot)$ represents the cumulative distribution function (CDF) of non-zero effect sizes. In our analysis, we do not assume that $G(\cdot)$ is known, nor do we impose any parametric constraints on it. In Section 3.2, we discuss a nonparametric approach to estimating $g(\cdot)$, the density function corresponding to $G(\cdot)$. Finally, the observation model is formulated as:

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, p; \ j = 1, 2, 3, \ldots, \tag{1.3}$$

where X_{ij} represents the observed measurement for the jth observation of variable i, and ϵ_{ij} denotes the random error associated with that measurement.

We now provide more explanations for the hierarchical mixture model (1.1)-(1.3), where measurements are collected in stages $j = 1, 2, \ldots$ in a sequential manner. The observed value X_{ij} follows the null distribution $F_0 = \mathcal{N}(0, \sigma^2)$ if $i \notin \mathcal{S}$, and the non-null distribution $F_{1i} = \mathcal{N}(\mu_i, \sigma^2)$ if $i \in \mathcal{S}$. The model posits that the null distribution F_0 is identical across all indices $i \notin \mathcal{S}$, while the non-null distribution F_{1i} can vary across $i \in \mathcal{S}$. Allowing heterogeneous F_{1i} offers great flexibility in applications where non-zero entries exhibit varying effect sizes. The fraction π is typically small. This model has been widely used in large-scale inference problems (e.g. Efron, 2001, 2016) by providing a flexible framework for a range of estimation and testing problems. In this article, we focus on the *sparse recovery* problem, which seeks to accurately identify the subset $\mathcal{S} \subset \{1, \ldots, p\}$ with negligible decision errors.

1.2 Non-adaptive vs. adaptive designs

Consider a scenario where each measurement incurs a fixed cost, making it prohibitively expensive to collect repeated measurements across *all* variables in large-scale studies, particularly when p is on the order of thousands or even millions. In this context, the fundamental

principle of an adaptive design is to strategically allocate sensing efforts, thus avoiding the necessity of evaluating all features at every stage of the process.

Let $\mathcal{A}_j \subseteq \mathcal{A}_1 \coloneqq \{1, \ldots, p\}$ for $j = 1, 2, \ldots$ denote a monotonic sequence of nested sets, where $\mathcal{A}_j$ represents the *active set* of coordinates at which measurements are conducted during stage j. This structure enforces the nestedness constraint, whereby $\mathcal{A}_j \subseteq \mathcal{A}_{j-1}$. The set $\mathcal{A}_j$ can be conceptualized as encompassing the locations where there is insufficient information to make confident decisions regarding the values of θ_i at stage j. In large-scale experiments involving millions of features, it is crucial to minimize additional sensing efforts. To this end, once a coordinate is selected or excluded, the decision remains fixed in subsequent stages; thus, we do not revisit earlier decisions. This approach effectively reduces the cognitive burden in the decision-making process and justifies the consideration of nested sets within the framework of sequential design. For a more comprehensive discussion of the nestedness constraint, please refer to Section 2.1.

Denote $X_{\mathcal{A}_j} = (X_{ij} : i \in \mathcal{A}_j)$ as the collection of observations corresponding to the active set $\mathcal{A}_j$. In a non-adaptive setting, the observations X_{ij} are sampled at each stage according to a pre-established policy, whereby each coordinate i is allocated an equal share of the measurement budget. This approach, while systematic, may not efficiently utilize available resources. In contrast, an *adaptive sampling design* (Malloy and Nowak, 2014a,b; Li and Haupt, 2015; Jain and Chang, 2004) tailors the sampling strategy to different coordinates. By allowing for flexibility in modifying $\mathcal{A}_j$ based on the information gathered from previous stages, adaptive designs improve the efficacy of the measurement process, particularly in scenarios where data uncertainty varies significantly across different features.

The adaptive sampling and inference framework provides a powerful approach to sparse estimation and testing problems. Intuitively, the sensing resources in later stages can be allocated in a more cost-effective way to reflect our updated contextual knowledge during the course of the study; hence greater precision in inference can be achieved with the same study budgets or computational costs. We discuss this in more detail in Section 1.6. A plethora of powerful multistage testing and estimation procedures have been developed under this flexible framework; some recent developments include the hierarchical testing procedures for pattern recognition (Blanchard and Geman, 2005; Meinshausen et al., 2009; Sun and Wei, 2015), distilled sensing and sequential thresholding methods for sparse detection (Haupt et al., 2011; Malloy and Nowak, 2011, 2014a), multi-scale search and open-loop feedback control algorithms for adaptive estimation (Bashan et al., 2011; Wei and Hero, 2013, 2015) and sequentially designed compressed sensing (Haupt et al., 2012; Malloy and Nowak, 2014b). These works demonstrate that methodologies adopting adaptive designs can substantially outperform those developed under non-adaptive settings.

1.3 Applications and statistical challenges

Multistage experiments have been widely used in many scientific fields including large-scale A/B testing (Johari et al., 2015), environmental sciences (Cormack, 1988; Thompson and Seber, 1996), microarray, RNA-seq, and protein array experiments (Müller et al., 2004; Rossell and Müller, 2013), geostatistical analysis (Roesch, 1993; Bloma et al., 2002), genome-wide association studies (Satagopan et al., 2004; Rothman et al., 2010) and astronomical

surveys (Meinshausen et al., 2009). We first describe some applications and then discuss related statistical issues.

High-throughput screening [HTS, Zhang et al. (1999); Bleicher et al. (2003)]. HTS is a large-scale hierarchical process that has played a crucial role in fast-advancing fields such as stem cell biology and drug discovery. In drug discovery, HTS involves testing a large number of chemical compounds in multiple stages including target identification, assay development, primary screening, confirmatory screening, and follow-up of hits. The accurate selection of useful compounds is an important issue at each aforementioned stage of HTS. For example, at the primary screening stage, a library of compounds is tested to generate an initial list of active compounds (hits). The goal of this stage is to reduce the size of the library significantly with negligible false negative rate. In the confirmatory screening stage, the hits are further investigated to generate a list of confirmed hits (leads), which will be used to develop drug candidates. As the lab costs for leads generation are very high, an important task at the confirmatory stage is to construct a subset with negligible false positive rate while keeping as many useful compounds as possible.

Large-scale A/B testing. A/B testing has begun to be widely deployed by various high-tech firms for identifying features/designs that work the best among a large number of potential candidates. It provides a powerful tool for testing new ideas for a wide range of real-world decision-making scenarios (Kohavi and Longbotham, 2017; Xu et al., 2015; Fabijan et al., 2018). For example, A/B testing can offer key insights on (a) how to improve the users' experiences, (b) what new features should be incorporated into the product to boost the profits, and (c) which design is the most effective for getting more people click the sign-up button. A significant challenge in large-scale A/B testing is the multiplicity issue, i.e. we need to conduct hundreds of experiments simultaneously for a long period, and each experiment may involve thousands of metrics to be evaluated throughout the entire duration. An effective multistage design can substantially save study costs, minimize risk exposures to customers and enable faster decision making.

Ultra-high dimensional testing in astronomical surveys. The fast and accurate localization of sparse signals poses a significant challenge in Astrophysics (Meinshausen and Rice, 2006). When many images are taken with high frequencies, the computational cost of an exhaustive search through every single image and pixel can be prohibitively expensive as the search often involves testing billions of hypotheses. A multistage decision process can lead to great savings in total sensing efforts by quickly narrowing down the focus to a much smaller subset of most promising spots in the images (Djorgovski et al., 2012).

These three applications will be discussed in detail in Section 8. Our goal is to develop a cost-effective multistage sampling and inference procedure to narrow down the focus in a sequential manner to identify the vector support reliably. The proposed strategy consists of (a) a stopping rule for selecting the active coordinates, and (b) a testing rule for deciding whether a coordinate contains a signal.

1.4 Problem formulation

In the design and analysis of large-scale multistage studies, the inflation of decision errors and soaring study costs are among the top concerns. First, to identify useful signals effectively, we need to control the *missed discovery rate* (MDR) to be small at all stages,

as missed signals will not be revisited in subsequent stages. Second, to reduce study costs or increase *sampling efficiency*, it is desirable to eliminate as many null locations as possible at each stage. Finally, to avoid misleading scientific conclusions, the final stage of our analysis calls for strict control of the *false positive rate* (FPR). Next, we formulate a decision-theoretic framework that integrally addresses the three issues: MDR control, FPR control, and sampling efficiency.

When many coordinate-wise sequential decisions are made simultaneously, controlling inflated decision errors becomes a critical issue. Let $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_p)$, where $\delta_i = 0/1$ indicates that i is classified as a null or non-null case. Define the false positive rate (FPR) and missed discovery rate (MDR) as

$$\mathrm{FPR}(\boldsymbol{\delta}) = \frac{\mathbb{E}\left\{\sum_{i=1}^p (1-\theta_i)\delta_i\right\}}{\mathbb{E}(\sum_{i=1}^p \delta_i)} \quad \text{and} \quad \mathrm{MDR}(\boldsymbol{\delta}) = \frac{\mathbb{E}\left\{\sum_{i=1}^p \theta_i(1-\delta_i)\right\}}{\mathbb{E}(\sum_{i=1}^p \theta_i)}. \tag{1.4}$$

Remark 1 *The FPR is also referred to as the marginal false discovery rate (mFDR; Genovese and Wasserman, 2002; Sun and Cai, 2007). In the non-sequential setting, the FPR and the widely used false discovery rate (FDR; Benjamini and Hochberg, 1995) are shown to be asymptotically equivalent for the Benjamini-Hochberg procedure (Genovese and Wasserman, 2002) and a general class of FDR procedures (Proposition 7 in Appendix A.2 of Cai et al. (2019)). Such asymptotic equivalence requires further research in the sequential setting. The main consideration for using FPR (as opposed to FDR) is to develop optimality results within a decision-theoretic framework (Berger, 1985). Theorem 6 shows that our proposed method controls both the FPR and FDR at user-specified levels. An alternative measure to the MDR is the false negative rate (FNR; Genovese and Wasserman, 2002; Sarkar, 2004). The concepts of FNR and MDR are different, with the MDR being a more appropriate error measure in the sparse setting; see Meinshausen and Rice (2006); Haupt et al. (2011); Cai and Sun (2017) for related discussions.*

We define the expected sample size (ESS) per unit as a criterion for comparing the efficiency of different multistage testing methods:

$$\mathrm{ESS}(\boldsymbol{N}) = \mathbb{E}\left(p^{-1}\sum_{i=1}^p N_i\right), \tag{1.5}$$

where $\boldsymbol{N} = (N_1, \cdots, N_p)$, with N_i being the sample size [also referred to as the stopping time (Berger, 1985; Siegmund, 1985), defined formally in Section 2.1] at coordinate i. Let α and γ be the user-specified FPR and MDR levels. We study the following constrained optimization problem:

$$\text{minimize } \mathrm{ESS}(\boldsymbol{N}) \text{ subject to } \mathrm{FPR}(\boldsymbol{\delta}) \le \alpha \text{ and } \mathrm{MDR}(\boldsymbol{\delta}) \le \gamma. \tag{1.6}$$

The formulation (1.6) reflects the key role of adaptive sampling, which effectively reduces the ESS by allowing one to stop sampling early at certain coordinates.

The formulation (1.6) naturally extends the classical sequential testing problem, considered in Wald (1945) and Siegmund (1985), to the compound decision setting (Robbins, 1951). Specifically, the sequential sparse recovery problem, which involves adaptive sampling at many coordinates simultaneously, can be viewed as a *compound decision problem*. Each component problem corresponds to the sequential testing of a single hypothesis

$H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ based on a data stream. In the seminal work of Wald (1945), the following constrained optimization problem is studied:

$$\text{minimize } \mathbb{E}(N) \text{ subject to } \mathbb{P}_{H_0}(\text{Reject}) \leq \alpha' \text{ and } \mathbb{P}_{H_1}(\text{Accept}) \leq \gamma'. \tag{1.7}$$

Here N is the stopping time, and α' and γ' are pre-specified Type I and Type II error rates. $\mathbb{E}(N)$, which represents the average sampling costs, characterizes the efficiency of a sequential procedure. The connection between the sparse recovery problem and sequential testing becomes clear by comparing (1.7) with (1.6).

The sequential probability ratio test (SPRT) is shown to be *optimal* (Wald, 1945; Siegmund, 1985) for the single sequential testing problem (1.7) in the sense that it has the smallest $\mathbb{E}(N)$ among all sequential procedures at level (α', γ'). This theory will be extended to the multiple sequential testing setup in Section 2.1.

1.5 A preview of the proposed algorithm

We formulate the sparse recovery problem as a sequential testing problem with multiple data streams. A new adaptive testing procedure is developed under a compound decision-theoretic framework. The proposed procedure, which employs a simultaneous multistage adaptive ranking and thresholding (SMART) approach, utilizes all information collected through multiple stages and exploits the compound structure of the decision problem to pool information across different coordinates. We show that SMART controls the FPR and MDR at user-specified levels and achieves the information-theoretic lower bounds. SMART is simple, fast, and capable of handling millions of tests. Numerical studies confirm the effectiveness of SMART in controlling error rates and in substantially saving study costs.

A schematic illustration of the SMART algorithm is provided in Figure 1. At each stage, SMART determines two data-driven thresholds, which divide the ordered test statistics into three types of decisions. The first type identifies non-null cases, forming the subset of rejections (the red area at the bottom); the second type eliminates null cases, forming the subset of acceptances (the blue area at the top); and the third type selects coordinates for further measurements, forming the active set $\mathcal{A}_j$ at stage j (the white area in the middle). The active set $\mathcal{A}_j$ shrinks over the sampling process until it becomes empty. The shaded areas (blue and red) correspond to the savings in study costs.

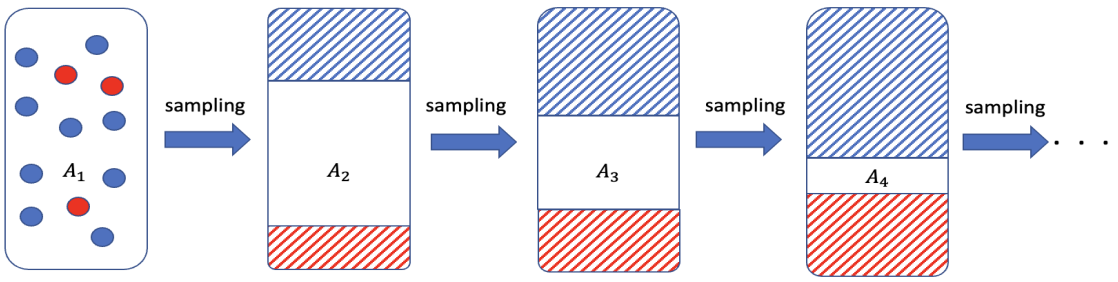


Figure 1: A schematic illustration of SMART: the blue, red and white areas respectively represent the subset of acceptances, rejections and indecisions. The active sampling set $\mathcal{A}_j$ shrinks over time.

1.6 Some theoretical insights

We first review the limits for fixed and adaptive designs in the literature and then compare them with our new limits derived based on adaptive sampling and SMART. Consider a two-point normal mixture model under a single-stage design:

$$X_i \overset{i.i.d.}{\sim} (1-\pi_p)N(0,1) + \pi_p N(\mu_p, \sigma^2), \tag{1.8}$$

where $\pi_p = p^{-\beta}$, $\mu_p = \sqrt{2r\log p}$, and $0 < \beta < 1$, $0 < r < 1$; σ^2 is unknown. The fundamental limits for a range of global and simultaneous inference problems have been derived under this setup (Donoho and Jin, 2004; Cai et al., 2007; Cai and Sun, 2017). Of particular interest is the *classification boundary* (Meinshausen and Rice, 2006; Haupt et al., 2011; Cai and Sun, 2017), which demarcates the possibility of constructing a subset with both the FPR and MDR tending to zero. Under model (1.8), the classification boundary is a straight line $r = \beta$ in the β-r plane for all values of $\sigma > 0$. The goal of $R^* \equiv \text{FPR}+\text{MDR} \to 0$ requires that the signal magnitude μ_p be at least on the order of $\sqrt{\log p}$. This establishes the fundamental limit of sparse recovery for fixed designs.

The rate $\sqrt{\log p}$ can be substantially improved in the adaptive setting. For example, Haupt et al. (2011) proposed the *distilled sensing* (DS) method, which achieves the classification boundary with much weaker signals. DS is a multistage testing procedure with a total measurement budget of $2p$. It assumes that observations follow a mixture model with noise distributed as standard normal. At each stage, DS keeps locations with positive observations and obtains new observations for these locations in the next stage. It was shown in Haupt et al. (2011) that after $k = \max\{\lceil \log_2 \log p \rceil, 0\} + 2$ steps, DS successfully constructs a subset with both FPR and MDR tending to zero, provided that μ_p *diverges at an arbitrary rate* in the dimension p. Malloy and Nowak (2014a) further utilizes the KL divergence and the average number of measurements per unit τ to characterize the boundary. Let s denote the cardinality of the support. It was shown that under fixed designs, reliable recovery requires that τ must be at least on the order of $\log p$, whereas under adaptive designs, the required τ is on the order of $\log s$, where $s = |\mathcal{S}|$ is the cardinality of the support of μ. This reveals the advantage of adaptive designs, noting that $\log s = (1-\beta)\log p$ under the calibration $\pi_p = p^{-\beta}$. The single sequential thresholding method (SST, Malloy and Nowak, 2014a) is optimal in the sense that it achieves the rate of $\log s$ asymptotically.

The theory in Malloy and Nowak (2014a) is developed under the stringent family-wise error rate (FWER) criterion. However, in large-scale testing problems, it is desirable to use a less stringent error rate, such as the FPR (1.4). We investigate the upper and lower bounds under the FPR-MDR paradigm and show that this paradigm requires fewer samples to guarantee that $R^* \to 0$. The theory complements the rate obtained by Haupt et al. (2011). SMART only requires a τ of the order $\log \xi(p)$, where $\xi(p)$ is a function that diverges to ∞ at an arbitrarily slow rate, improving the rate of $\log s$ achieved by SST.

1.7 Related work and our contributions

The error control in multi-stage testing has been studied extensively (Benjamini and Heller, 2007; Victor and Hommel, 2007; Dmitrienko et al., 2007; Goeman and Mansmann, 2008; Yekutieli, 2008; Liang and Nettleton, 2010; Benjamini and Bogomolov, 2014). However,

existing methods are usually designed for different types of applications and cannot be easily tailored to the problem at hand. The biggest limitation is that existing works focus only on controlling false positive errors, whereas the sparse recovery problem also requires the control of missed discoveries. Moreover, the ranking and stopping rules proposed in previous studies (Zehetmayer et al., 2008; Posch et al., 2009; Sarkar et al., 2013) are suboptimal, as the data combination methods that rely on p-values do not fully capitalize on the information provided by different stages and fail to exploit compound thresholding rules. Furthermore, recent research on sequential testing using SPRT rules (Bartroff and Song, 2013; Bartroff, 2014; Bartroff and Song, 2016) is impractical for large-scale studies due to computational complexity. The issue of optimality, in particular, has not been addressed in the aforementioned studies. SMART incorporates several key features that distinguish it from existing work. First, we aim to control both the FPR and MDR, unlike methods designed for smaller-scale studies that primarily target the control of the FWER. Second, our methodology explicitly tackles the uncertainty quantification of statistical decisions. The adaptability of our proposed framework, which hinges on evaluating the uncertainties of simultaneous decisions, represents a crucial statistical perspective that is notably absent in other machine learning approaches.

SMART offers several advantages over existing sparse recovery methods such as DS and SST. First, both DS and SST employ fixed thresholding rules that are not adaptive to the patterns in the observations. Additionally, due to their failure to effectively quantify the uncertainty of decisions at each stage, the theoretical foundations of DS and SST rely on strong assumptions about the unknown signal strengths. When signals are weak, both methods fail to control the error rate. In contrast, our algorithm is data-adaptive, allowing thresholds to vary according to the observed data. This adaptability enables researchers to gather additional data sequentially and utilize more stages to strengthen the evidence, ensuring effective control of error rates. Second, the stopping and testing rules in DS and SST depend solely on observations from the current stage, neglecting valuable data from prior stages. This limitation undermines efficiency, as illustrated in our numerical studies. SMART effectively addresses these shortcomings. We provide a detailed comparison with different methods in Section 6.2.

1.8 Organization of the paper

The rest of this article is organized as follows. Section 2 develops oracle rules for sparse recovery that control both the FPR and MDR. Section 3 introduces a data-driven SMART algorithm and discusses its practical implementation. Methodological insights and asymptotic theory are presented in Sections 4 and 5, respectively. Connections to existing work are explored in Section 6. The numerical performance of SMART is demonstrated using synthetic data (Section 7) and real-world applications (Section 8). Section 9 concludes with a discussion of future directions. Proofs are included in Section 10, with additional results available in the Supplementary Material.

2. Oracle Rules for Sparse Recovery

The sparse recovery problem in the adaptive setting involves the simultaneous testing of p hypotheses:

$$H_{0,i} : \mu_i = 0 \quad \text{vs.} \quad H_{1,i} : \mu_i \neq 0, \quad 1 \leq i \leq p, \tag{2.1}$$

based on p streams of observations $\{(X_{i1}, X_{i2}, \ldots) : 1 \leq i \leq p\}$. This section derives an oracle procedure (Section 2.1) and the corresponding thresholds (Section 2.2). Additionally, the data-driven procedure that mimics the oracle rule is developed in Section 3.

2.1 Oracle procedure

Let X_{ij} be the measurement of variable X_i at stage j and $\boldsymbol{X}_i^j = (X_{i1}, \ldots, X_{ij})$ the collection of measurements for X_i up to stage j. A *multistage decision procedure* involves choosing a *stopping rule* and a *testing rule* for each location. At location i, the stopping rule $\boldsymbol{\tau}_i$ consists of a series of functions $\tau_{i1}(\boldsymbol{X}_i^1)$, $\tau_{i2}(\boldsymbol{X}_i^2)$, $\ldots$, where τ_{ij} takes values in $\{0, 1\}$, with 0 and 1 standing for "taking another observation" and "stopping sampling and making a decision," respectively. We consider a class of multistage designs where the focus is sequentially narrowed down, i.e., the active sets satisfy $\mathcal{A}_1 = \{1, \ldots, p\}$, and $\mathcal{A}_j \subseteq \mathcal{A}_{j-1}$ for $j = 2, 3, \ldots$. At every coordinate $i \in \mathcal{A}_j$, there are three possible actions: (i) stop sampling and accept $H_{0,i}$ as true; (ii) stop sampling and claim $H_{0,i}$ is false; and (iii) do not make a decision and take another observation.

The monotonicity condition on $(\mathcal{A}_j : j = 1, 2, \ldots)$ is a natural requirement in multistage analysis and plays a crucial role in resource-sensitive scenarios such as drug discovery and resume screening. In drug discovery, screening steps aim to shortlist a few candidates from thousands or millions, making the monotonicity condition reasonable. In HTS, promising compounds are selected for costly further testing to confirm their effectiveness. The resource-sensitive nature of HTS necessitates the avoidance of revisiting unpromising compounds to save study costs, thus a monotone structure naturally emerges.

We follow the standard notations under the decision-theoretic framework detailed in Chapter 7 of Berger (1985). The stopping rule $\boldsymbol{\tau}_i$ can be equivalently described by the *stopping time* $N_i = \min\{n \geq 1 : \tau_{in}(\boldsymbol{X}_i^n) = 1\}$, the final stage at which we stop sampling and make a decision. If we collect one sample per stage, we use the terms *sample size* and *stopping time* interchangeably. The testing rule $\delta_i \in \{0, 1\}$ is applied at the terminal sampling stage N_i, where $\delta_i = 0$ (1) indicates that coordinate i is classified as a null (non-null) case. A multistage decision procedure is therefore denoted as $\boldsymbol{d} = (\boldsymbol{N}, \boldsymbol{\delta})$, where $\boldsymbol{N} = (N_1, \ldots, N_p)$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)$ are the sample sizes and terminal decisions, respectively. Consider the following test statistic

$$T_{OR}^{i,j} = \mathbb{P}(\theta_i = 0 | \boldsymbol{X}_i^j). \tag{2.2}$$

We view $T_{OR}^{i,j}$ as a significance index reflecting our confidence on claiming that $H_{0,i}$ is true based on all information available at stage j. Let t_l and t_u be constants satisfying $0 < t_l < t_u < 1$. Consider a class of sequential testing procedures $\boldsymbol{d}^*(t_l, t_u)$ of the form:

$$\begin{aligned} &\text{stop sampling for unit } i \text{ at } N_i = \min\{j \geq 1 : T_{OR}^{i,j} \leq t_l \text{ or } T_{OR}^{i,j} \geq t_u\}, \\ &\qquad \text{and decide } \delta_i = 1 \text{ if } T_{OR}^{i,j} \leq t_l \text{ and } \delta_i = 0 \text{ if } T_{OR}^{i,j} \geq t_u, \end{aligned} \tag{2.3}$$

where the superscript $*$ indicates that we are considering the idealized case where $T_{OR}^{i,j}$ is known. Denote by $\mathcal{D}_{\alpha,\gamma}$ the collection of all sequential decision procedures that simultaneously satisfy

$$(i)\ \mathrm{FPR}(\boldsymbol{d}) \leq \alpha \text{ and } (ii)\ \mathrm{MDR}(\boldsymbol{d}) \leq \gamma.$$

Let $f(\boldsymbol{X}_i^1|\theta_i)$ be the conditional density function evaluated at $\boldsymbol{X}_i^1$ given θ_i. The following assumption, which is a standard condition in the sequential analysis literature (cf. Berger, 1985; Siegmund, 1985), requires that $f(\boldsymbol{X}_i^1|\theta_i = 0)$ and $f(\boldsymbol{X}_i^1|\theta_i = 1)$ differ with some positive probability. This ensures that the sequential testing procedure will stop within a finite number of stages almost surely.

Assumption 1 $\mathbb{P}_{\theta_i}\left(\log\left\{\frac{f(\boldsymbol{X}_i^1|\theta_i=1)}{f(\boldsymbol{X}_i^1|\theta_i=0)}\right\} = 0\right) < 1$ *for* $i = 1, \cdots, p$.

The next theorem derives an oracle sequential testing procedure that provides the optimal solution to the constrained optimization problem (1.6).

Theorem 2 *Consider the multistage model* (1.1)–(1.3)*, and a class of sequential testing rules* $\boldsymbol{d}^\pi(t_l, t_u)$ *taking the form of* (2.3)*. Denote* $Q_{OR}(t_l, t_u)$ *and* $\tilde{Q}_{OR}(t_l, t_u)$ *the FPR and MDR levels of* $\boldsymbol{d}^\pi(t_l, t_u)$*, respectively. Then under Assumption 1, we have*

(a). $Q_{OR}(t_l, t_u)$ is non-decreasing in t_l for a fixed t_u, and $\tilde{Q}_{OR}(t_l, t_u)$ is non-increasing in t_u for a fixed t_l.

(b). Let $0 < \alpha, \gamma < 1$. Then there exists a pair of oracle thresholds (t_{OR}^l, t_{OR}^u), based on which we can define the oracle procedure

$$\boldsymbol{d}_{OR} \equiv \boldsymbol{d}^\pi(t_{OR}^l, t_{OR}^u) \tag{2.4}$$

such that $\boldsymbol{d}_{OR}$ is optimal in the sense that (i) FPR($\boldsymbol{d}_{OR}$) $\leq \alpha$; (ii) MDR($\boldsymbol{d}_{OR}$) $\leq \gamma$; and (iii) ESS($\boldsymbol{d}_{OR}$) $\leq$ ESS($\boldsymbol{d}_*$) for all $\boldsymbol{d}_* \in \mathcal{D}_{\alpha,\gamma}$.

Remark 3 *The development of a multiple testing procedure consists of two steps: deriving the optimal ranking statistic and finding a threshold along the ranking. Under the FPR criterion, the optimal ranking statistic is T_{OR}. The work by Heller and Rosset (2021) indicates that T_{OR} is also optimal under the FDR criterion. More comparisons and discussions of the FDR and FPR criteria are provided in Heller and Rosset (2021) and Appendix D.*

While the optimality theory in Theorem 2 may seem intuitive and naturally aligns with concepts presented in earlier works by Berger (1985) and Siegmund (1985), its proof incorporates several novel techniques that are inherently complex and highly nontrivial. The main goal of Theorem 2 (as well as some later theorems) is to motivate our methodology and provide insight into the potential merits of the framework. Similar to the theories in Johari et al. (2022), we have considered the ideal case with known $T_{OR}^{i,j}$, and the validity of the theorems does not depend on a specific family of non-null and null distributions. We develop data-driven procedures in Section 3.2 for practical implementations where the non-null density function $f(\boldsymbol{X}_i^1|\theta_i = 1)$, non-null proportion π, and σ^2 must be estimated.

2.2 Approximation of oracle thresholds and its properties

The oracle procedure $\boldsymbol{d}_{OR}$ is a thresholding rule based on the oracle statistic T_{OR} and the oracle thresholds t^l_{OR} and t^u_{OR}. However, Theorem 2 only establishes the existence of t^l_{OR} and t^u_{OR}, which remain unknown in practice. In this section, we derive approximations for t^l_{OR} and t^u_{OR} by framing the sparse recovery problem as an m-parallel sequential testing problems. These approximate thresholds are then refined using a compound thresholding technique and are subsequently utilized in the prototype algorithm presented in Section 3.1.

If we focus on a single location i, then the sequential probability ratio test (SPRT) (Wald, 1945) is a thresholding rule based on $\mathcal{L}_{i,j} = \frac{f(\boldsymbol{X}_i^j|\theta_i=1)}{f(\boldsymbol{X}_i^j|\theta_i=0)}$ and of the form

$$\begin{array}{ll} \text{if } \log \mathcal{L}_{i,j} \le a, & \text{stop sampling and decide } H_{0,i} \text{ is true;} \\ \text{if } \log \mathcal{L}_{i,j} \ge b, & \text{stop sampling and decide } H_{0,i} \text{ is false;} \\ \text{if } a < \log \mathcal{L}_{i,j} < b, & \text{take another observation.} \end{array}$$

Let $\alpha' = \mathbb{P}_{H_{0,i}}(\text{Reject } H_{0,i})$ and $\gamma' = \mathbb{P}_{H_{1,i}}(\text{Accept } H_{0,i})$ be prespecified Type I and Type II error rates. Applying Wald's identity, the thresholds a and b can be approximated as:

$$\tilde{a} = \log \frac{\gamma'}{1-\alpha'} \ \text{ and } \ \tilde{b} = \log \frac{1-\gamma'}{\alpha'}. \tag{2.5}$$

The oracle statistic

$$T^{i,j}_{OR} \coloneqq \mathbb{P}(\theta_i = 0|\boldsymbol{X}_i^j) = \frac{(1-\pi)f(\boldsymbol{X}_i^j|\theta_i=0)}{(1-\pi)f(\boldsymbol{X}_i^j|\theta_i=0)+\pi f(\boldsymbol{X}_i^j|\theta_i=1)} = \frac{1-\pi)}{(1-\pi)+\pi\mathcal{L}_{i,j}} \tag{2.6}$$

is monotone in $\mathcal{L}_{i,j}$. In multiple hypothesis setting, we can view $\boldsymbol{d}_{OR} = \boldsymbol{d}^{\pi}(t^l_{OR}, t^u_{OR})$ as m parallel SPRTs. The problem then boils down to how to obtain approximate formulas for the oracle thresholds t^l_{OR} and t^u_{OR} given pre-specified FPR and MDR levels (α, γ). In our derivation, we utilize classical techniques [cf. Section 7.5.2 in Berger (1985)], and the relationships between the FPR and MDR levels (α, γ) and the Type I and Type II error rates (α', γ') are exploited. The key steps in the derivation are provided in Appendix A. Assume that $\pi < 1/2$ and $\alpha < 1/2$, then approximated thresholds are given by

$$\tilde{t}^l_{OR} = \alpha \ \text{ and } \ \tilde{t}^u_{OR} = \frac{1-\pi}{\pi\gamma + 1 - \pi}. \tag{2.7}$$

The next theorem shows that the pair (2.7) is valid for FPR and MDR control.

Proposition 4 *Consider the multistage model* (1.1)–(1.3). *Denote* $\tilde{\boldsymbol{d}}_{OR} = \boldsymbol{d}^{\pi}(\tilde{t}^l_{OR}, \tilde{t}^u_{OR})$ *the thresholding procedure that operates according to* (2.3) *with upper and lower thresholds given by* (2.7). *Then we have*

$$FPR(\tilde{\boldsymbol{d}}_{OR}) \le \alpha,\ FDR(\tilde{\boldsymbol{d}}_{OR}) \le \alpha \text{ and } MDR(\tilde{\boldsymbol{d}}_{OR}) \le \gamma.$$

Remark 5 *In Section 4, we discuss the reasons for not recommending the use of* $\tilde{\boldsymbol{d}}_{OR}$ *in practice, despite its validated efficacy for error rate control, as established in Proposition 4. Specifically, we highlight the overshooting issue associated with* $\tilde{\boldsymbol{d}}_{OR}$ *in Section 4.1. To mitigate this issue, the thresholds derived from* $\tilde{\boldsymbol{d}}_{OR}$ *will be refined using a compound thresholding technique, which is detailed shortly in Algorithm 1 in Section 3.1. This refinement is shown to significantly improve statistical power.*

3. Sparse Recovery via SMART

This section introduces the SMART procedure utilizing compound thresholding to address the overshooting issue and enhance statistical power. To facilitate understanding, we present the core concepts of the algorithm upfront, allowing readers to grasp the essential ideas without delving into complex theoretical considerations. Technical details are provided in Section 4 for those interested in a more in-depth exploration.

3.1 A prototype algorithm under the oracle setup

This section begins by introducing a prototype algorithm under the oracle setup, which assumes knowledge of both π and g. Subsequently, we develop a data-driven procedure that can be implemented in real-world scenarios where π and g are unknown.

The operation of SMART can be described as follows. At stage j, we first calculate the oracle statistics in the active set $\mathcal{A}_j$ and sort them from smallest to largest. Then we carry out two thresholding procedures along the ranking: part (a) chooses a lower cutoff $\hat{t}^{l,j}_{OR}$ with selected locations claimed as signals; part (b) chooses an upper cutoff $\hat{t}^{u,j}_{OR}$ with selected locations claimed as nulls. We stop sampling on locations where definitive decisions are made, and take new observations on remaining locations for further investigation. The above steps will be iterated until convergence. The detailed operation is described in Algorithm 1.

Algorithm 1 A Prototype Algorithm

Input Target FPR level α, target MDR level γ, $\boldsymbol{X}_i^j$ for $i = 1, \ldots, p$, $j = 1, \ldots$, non null proportion π, conditional densities $f(\boldsymbol{X}_i^j|\theta_i)$ for $\theta_i = 0/1$, $j = 1, \ldots$

Preliminary Step Define the lower and upper thresholds $\tilde{t}^l_{OR} = \alpha$ and $\tilde{t}^u_{OR} = \frac{1-\pi}{\pi\gamma+1-\pi}$. Let $\mathcal{A}_j$ be the active set at stage j, $j = 1, 2, \cdots$.

Iterate Steps 1 to 3 until $\mathcal{A}_j = \emptyset$.

Step 1 (Ranking). For all $i \in \mathcal{A}_j$, compute $T^{i,j}_{OR}$ using (2.6) and sort them in ascending order

$T^{(1),j}_{OR} \le T^{(2),j}_{OR} \le \cdots \le T^{(k_j),j}_{OR}$, where $k_j = \mathrm{Card}(\mathcal{A}_j)$.

Step 2 (Thresholding).

(a) (Signal discovery). Let $k^s_j = \max\{r : r^{-1}\sum_{i=1}^r T^{(i),j}_{OR} \le \tilde{t}^l_{OR}\}$ and $\hat{t}^{l,j}_{OR} = T^{(k^s_j),j}_{OR}$.

Define $\mathcal{S}^s_j = \{i \in \mathcal{A}_j : T^{i,j}_{OR} \le \hat{t}^{l,j}_{OR}\}$. For all $i \in \mathcal{S}^s_j$, stop sampling and let $\delta_i = 1$.

(b) (Noise elimination). Let $k^e_j = \max\{r : r^{-1}\sum_{i=0}^{r-1}(T^{(k_j-i),j}_{OR}) \ge \tilde{t}^u_{OR}\}$ and $\hat{t}^{u,j}_{OR} = T^{(k_j-k^e_j+1),j}_{OR}$.

Define $\mathcal{S}^e_j = \{i \in \mathcal{A}_j : T^{i,j}_{OR} \ge \hat{t}^{u,j}_{OR}\}$. For all $i \in \mathcal{S}^e_j$, stop sampling and let $\delta_i = 0$.

Step 3 (Updating). Let $\mathcal{A}_{j+1} = \mathcal{A}_j \setminus (\mathcal{S}^s_j \cup \mathcal{S}^e_j)$. Take new observations on $\mathcal{A}_{j+1}$.

Output Estimated non-null positions $\{i : \delta_i = 1\}$.

The SMART algorithm uses stage-wise thresholds $\hat{t}^{l,j}_{OR}$ and $\hat{t}^{u,j}_{OR}$. The operation of the algorithm implies that we always have $\hat{t}^{l,j}_{OR} \ge \tilde{t}^l_{OR}$ and $\hat{t}^{u,j}_{OR} \le \tilde{t}^u_{OR}$ for all j. Hence SMART always uses fewer samples than $\tilde{\boldsymbol{d}}_{OR}$. The next theorem shows that SMART is valid for FPR and MDR control.

Theorem 6 *Denote $\boldsymbol{d}_{SM}$ the SMART procedure described in Algorithm 1 with pre-specified FPR and MDR levels (α, γ). Then*

$$\mathrm{FPR}(\boldsymbol{d}_{SM}) \leq \alpha,\ \mathrm{FDR}(\boldsymbol{d}_{\mathrm{SM}}) \leq \alpha \ \ \textit{and}\ \ \mathrm{MDR}(\boldsymbol{d}_{\mathrm{SM}}) \leq \gamma.$$

3.2 Numerical implementation of SMART

In practice the oracle statistic T_{OR} is unknown and needs to be estimated from data. This subsection develops an algorithm for computing T_{OR} based the model in introduced in Section 1.1, which is summarized in Algorithm 2. Note that the model in Section 1.1 can be equivalently written as

$$\mu_i \sim (1-\pi)\delta_0(\cdot) + \pi g(\cdot), \quad X_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1, \ldots, p, \quad j = 1, 2, \ldots \tag{3.8}$$

In this subsection we assume g is mixture of point masses $g(\cdot) = \sum_{s=1}^{m_1} w_s \delta_{\eta_s}(\cdot)$. Consequently, the oracle statistics $T_{OR}^{i,j}$ can be written as

$$T_{OR}^{i,j} := \mathbb{P}(\theta_i = 0 | \boldsymbol{X}_i^j) = \frac{(1-\pi)\prod_{t=1}^{j}\phi(x_{it}, 0, \sigma^2)}{(1-\pi)\prod_{t=1}^{j}\phi(x_{it}, 0, \sigma^2) + \pi\sum_{s=1}^{m_1} w_s \prod_{t=1}^{j}\phi(x_{it}, \eta_s, \sigma^2)}, \tag{3.9}$$

where $\phi(\cdot, \mu, \sigma^2)$ is the density function of $N(\mu, \sigma^2)$.

In the preliminary estimation step, we propose to use the method in Jin and Cai (2007) to estimate the unknown mixing proportion π. Next, the estimation of $(1-\pi)\delta_0 + \pi g(\cdot)$, referred to as the deconvolution problem, has been studied intensively in the Empirical Bayes literature (Koenker and Gu (2017), Efron (2016), Jiang and Zhang (2009)). The *GLmix* function in the R package **REBayes**, which is employed in our algorithm, is shown to deliver excellent performance in practice. *GLmix* first discretizes $(1-\pi)\delta_0 + \pi g(\cdot)$ as a mixture of point masses and then assigns weights to each point mass to approximate $(1-\pi)\delta_0 + \pi g(\cdot)$. To get the final estimate of g, we first calculate the minimum detectable signal strength using the classification boundary in Cai et al. (2011), i.e. calibrate $\hat{\beta} = -\frac{\log \hat{\pi}}{\log p}$ and set $A_\kappa = \max\{\sqrt{2\hat{\beta}\log p}, \kappa\}$, where κ is a pre-specified constant (empirically, the choice of $\kappa = 1.5$ works well in practice and has been used across all simulations). Then we filter out the locations with absolute signal amplitudes smaller than A_κ. Finally, we update $\hat{g}$ by re-weighting the remaining locations so that the sum of point mass probabilities equals 1. In later stages $j = 2, 3, \cdots$, we sequentially update $\hat{T}_{OR}^{i,j}$ using $\hat{g}$ and $\hat{\pi}$ that are estimated based on data collected at stage 1.

The asymptotic properties of $\hat{\pi}$ and $\hat{g}(\cdot)$ have been investigated in Jin and Cai (2007) and Saha and Guntuboyina (2020), respectively. The numerical experiments in Jin and Cai (2007) and Koenker and Mizera (2014) suggest these estimators perform quite well in practice. Together with the next proposition, the data-driven SMART provides a reasonable and sensible algorithm for the sparse recovery problem.

Proposition 7 *Suppose $g(\cdot) = \sum_{s=1}^{m_1} w_s \delta_{\eta_s}(\cdot)$ is a finite mixture of point masses, with $supp(g) \cap [-a, a] = \emptyset$ for some $a > 0$ and $w_s > \epsilon$ for all s and some fixed ϵ. Further assume that $\hat{g}(\cdot) = \sum_{s=1}^{m_2} \hat{w}_s \delta_{\hat{\eta}_s}(\cdot)$ with $\hat{w}_s > \epsilon$ for all s, and for every η_s there exist $\hat{\eta}_k$ such that*

$|\eta_s - \hat{\eta}_k| < a/2 - \epsilon$. *Let* $T_{OR}^{i,j}$ *be as defined in* (3.9) *and* $\hat{T}_{OR}^{i,j}$ *as defined in Algorithm 2, then for any* α*, there exist* K_α *such that*

$$\mathbb{P}\left\{\forall j > K_\alpha, |T_{OR}^{i,j} - \hat{T}_{OR}^{i,j}| < e^{-cj}\right\} > 1 - \alpha,$$

where $c > 0$ *is a constant depends on* α*,* π*,* $\hat{\pi}$*,* σ*,* $\hat{\sigma}$*,* a*.*

Proposition 7 establishes, with high probability, that $\hat{T}_{OR}^{i,j}$ converges to 0 or 1 at a rate of $O(e^{-cj})$. As a result, the number of iterations is at most $\log p$. Therefore, the computational complexity for Algorithm 2 does not exceed $O(p(\log p)^2)$. The proof of Proposition 7, provided in Section 10, employs techniques from Howard et al. (2021), which establish a time-uniform confidence sequence for the sample mean of subgaussian distributions. Our work builds upon this result, leveraging (a) the one-to-one correspondence between the oracle statistic $T_{OR}^{i,j}$ and the likelihood ratio $\mathcal{L}_{i,j}$, and (b) the fact that, given $\hat{g}$ and $\hat{\sigma}$, the likelihood ratio is a function of the sample mean. If the noise does not follow a Gaussian distribution but some subgaussian distribution Φ_σ instead, then we can replace the Gaussian density function $\phi(\cdot, \cdot, \sigma)$ in Algorithm 2 with $\varphi(\cdot, \cdot, \sigma)$, where φ is the density function of Φ_σ, and Proposition 7 remains valid.

Algorithm 2 The data-driven SMART algorithm.

Input Target FPR & MDR levels (α, γ), sequentially collected data

The preliminary estimation step ($j = 1$):

Step 1 Estimate π and σ^2. Denote the estimated value as $\hat{\pi}$ and $\hat{\sigma}^2$ respectively.

Step 2 Use *GLmix* in R package **REBayes** to estimate $\psi(\mu) = (1 - \pi)\delta_0 + \pi g(\cdot)$.

Step 3 Calibrate $\hat{\beta} = -\frac{\log \hat{\pi}}{\log p}$, then set $A_{1.5} = \max\{\sqrt{2\hat{\beta} \log p}, 1.5\}$.

Step 4 Define $\tilde{g}(\mu) = \psi(\mu)\mathbb{I}(|\mu| \geq A_{1.5})$. Set $\hat{g}(\mu) = \dfrac{\tilde{g}(\mu)}{\int_{-\infty}^{\infty} \tilde{g}(\mu) d\mu} = \sum_{s=1}^m w_s \delta_{\eta_s}(\cdot)$.

Step 5 Define the lower and upper thresholds $\tilde{t}_{OR}^l = \alpha$ and $\tilde{t}_{OR}^u = \dfrac{1 - \hat{\pi}}{\hat{\pi}\gamma + 1 - \hat{\pi}}$.

The SMART procedure ($j > 1$): Iterate Step 6 to Step 9 until $\mathcal{A}_j = \emptyset$.

Step 6 (Estimation). Compute $\hat{T}_{OR}^{i,j} := \dfrac{(1 - \hat{\pi}) \prod_{t=1}^j \phi(x_{it}, 0, \hat{\sigma}^2)}{(1 - \hat{\pi}) \prod_{t=1}^j \phi(x_{it}, 0, \hat{\sigma}^2) + \hat{\pi} \prod_{t=1}^j \phi(x_{it}, \hat{\mu}_i^j, \hat{\sigma}^2)}$,

where $\hat{\mu}_i^j := \sum_{s=1}^m \frac{w_s \prod_{t=1}^j \phi(x_{it}, \eta_s, \hat{\sigma}^2)}{\sum_{k=1}^m w_k \prod_{t=1}^j \phi(x_{it}, \eta_k, \hat{\sigma}^2)} \eta_s$ and $\phi(x_{it}, \eta_s, \hat{\sigma}^2) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left(-\frac{(x_{it} - \eta_s)^2}{2\hat{\sigma}^2}\right)$.

Step 7 (Thresholding).

(a) (Signal discovery) Let $k_j^d = \max\{r : r^{-1} \sum_{i=1}^r \hat{T}_{OR}^{(i),j} \leq \tilde{t}_{OR}^l\}$ and $\hat{t}_{OR}^{l,j} = \hat{T}_{OR}^{(k_j^d),j}$.

Define $\mathcal{S}_j^d = \{i \in \mathcal{A}_j : \hat{T}_{OR}^{i,j} \leq \hat{T}_{OR}^{(k_j^d),j}\}$. For all $i \in \mathcal{S}_j^d$, stop sampling and let $\delta_i = 1$.

(b) (Noise elimination) Let $k_j^e = \max\{r : \frac{1}{r} \sum_{i=1}^r \hat{T}_{OR}^{(k_j - i),j} \geq \tilde{t}_{OR}^u\}$ and $\hat{t}_{OR}^{u,j} = \hat{T}_{OR}^{(k_j - k_j^e + 1),j}$.

Define $\mathcal{S}_j^e = \{i \in \mathcal{A}_j : \hat{T}_{OR}^{i,j} \geq \hat{T}_{OR}^{(k_j^e),j}\}$. For all $i \in \mathcal{S}_j^e$, stop sampling and let $\delta_i = 0$.

Step 8 (Updating). Let $\mathcal{A}_{j+1} = \mathcal{A}_j \backslash (\mathcal{S}_j^d \bigcup \mathcal{S}_j^e)$. Take new observations on $\mathcal{A}_{j+1}$.

Output Estimated non-null locations $\{i : \delta_i = 1\}$.

Remark 8 *We have assumed that $g(\cdot)$ is a mixture of point masses. This assumption is mild, as a mixture of point masses is sufficiently complex to represent the underlying process that generates our observed data. The classical Carathéodory theorem establishes that the optimal $\hat{G}$ must be atomic, which is a discrete distribution with fewer than n atoms. The assumption that $supp(g) \cap [-a, a] = \emptyset$ for some $a > 0$ is reasonable, since the difference between the means of the signal and noise must exceed a certain threshold for the problem to be meaningful. Additionally, we assume that $\hat{g}$ is a finite mixture of point masses, which is a standard assumption in the deconvolution literature (Efron, 2016; Koenker and Mizera, 2014). The support of $\hat{g}$, represented by $\{\hat{\eta}_1, \ldots, \hat{\eta}_{m_2}\}$, is typically a pre-specified grid on the real line. If a lower bound a is known, a sufficiently dense grid can be chosen in $[\min X_{i1}, \max X_{i1}]\backslash(-\tilde{a}, \tilde{a})$ such that our condition is satisfied.*

Remark 9 *Developing a consistent algorithm for estimating g while fully exploiting the information from all samples is a non-trivial question in itself and requires further investigation. Such a method could be impractical for several reasons. Firstly, estimating g requires solving a complicated optimization problem that can be computationally intensive. Updating g at each stage would result in an algorithm that is prohibitively slow. Secondly, existing deconvolution techniques (Koenker and Mizera, 2014; Efron, 2016; Jiang and Zhang, 2009) require the mean parameters μ_i to be independent and identically distributed (i.i.d.) from g. However, once a location is identified as signal or noise, the SMART algorithm no longer samples from that location. Consequently, the samples collected in subsequent stages only reflect a biased portion of g.*

4. The Overshooting Problem and Compound Thresholding

In the thresholding step, SMART determines the lower and upper cutoffs using the moving averages of selected oracle statistics. We refer to SMART as a compound thresholding procedure because its stage-wise cutoffs, $\hat{t}^{l,j}_{OR}$ and $\hat{t}^{u,j}_{OR}$, are jointly derived from data across multiple locations. In contrast, we classify $\tilde{\boldsymbol{d}}_{OR}$, as defined in Theorem 4, as a simple thresholding rule since the decision at location i relies solely on its own data. A key finding from Robbins (1951) demonstrates that compound rules generally outperform simple rules, even when observations across different units are independent. Additionally, in the context of multiple testing, Sun and Cai (2007) showed that pooling information from independent tests leads to more powerful FDR procedures. In light of these results, this section elucidates why SMART offers advantages over $\tilde{\boldsymbol{d}}_{OR}$.

4.1 The overshooting problem

The approximations used in Equation (2.7) result in conservative FPR and MDR levels. This is primarily due to a phenomenon known as "overshooting." Specifically, the SPRTs can be conceptualized as a Brownian motion with an average step size determined by $E_{\theta_i}[\log\{f_1(X_{ik})/f_0(X_{ik})\}]$. Figure 2 displays sample paths for units with both large and small step sizes, illustrating how SPRT statistics fluctuate under the null and alternative hypotheses. The SPRT reaches a decision when $\log \mathcal{L}_{ij}$ exits the interval (a, b). To derive $\tilde{t}^{l}_{OR}$ and $\tilde{t}^{u}_{OR}$, we have employed Wald's approximation, assuming that the two boundaries are *precisely hit* by $\log \mathcal{L}_{ij}$. However, this assumption is highly idealized, as evidenced

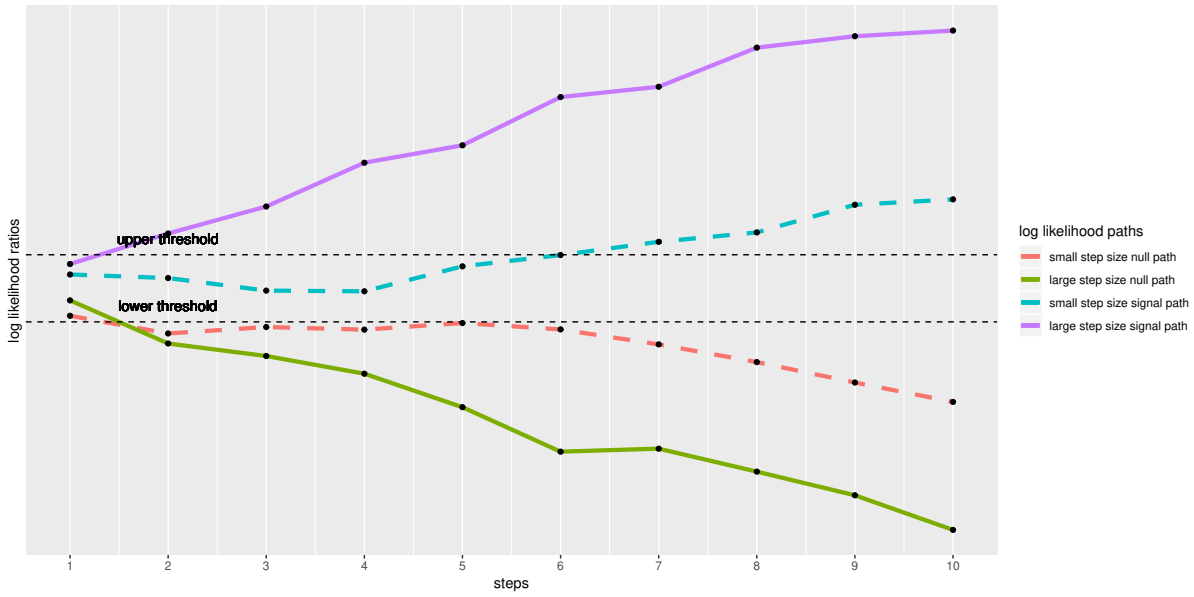


Figure 2: The sample paths for two SPRTs, with the null being $N(0,1)$, and alternative being $N(1,1)$ (dashed lines, smalle steps) and $N(2,1)$ (solid lines, large steps), respectively. L_{ij} change drastically with large steps, accentuating the overshooting issue.

by the sample paths in Figure 2, where the boundaries are seldom hit. This tendency toward overshooting often leads to an overestimation of true error probabilities and results in conservative error control, as outlined by Theorem 12 in Section 7.5.3 of Berger (1985). Our simulations show that the FPR of SPRTs based on Wald's approximations can be as low as 0.01 when the nominal FPR level is 0.05. Such conservativeness can lead to a substantial loss of efficiency, particularly in high-dimensional settings where most locations are only tested once or twice (i.e., when the SPRTs exhibit "large step sizes"). These concerns motivated us to develop new methodologies in Section 3.

4.2 Compound thresholding and knapsack problems

By adopting a compound thresholding scheme and pooling information across parallel SPRTs, SMART overcomes the overshoot problem of individual SPRTs. To illustrate the core idea, consider the following toy example.

Example 4.1 Let the FDR level $\alpha = 0.05$. Suppose at stage j the ordered T_{OR} values are $\{0.01, 0.055, 0.07, 0.10, \cdots\}$. If we use the simple thresholding rule $\tilde{\boldsymbol{d}}_{OR}$ with $\tilde{t}^l_{OR} = \alpha = 0.05$, then we can reject one hypothesis with $T^{(1),j}_{OR} = 0.01$; the gap between $T^{(1),j}_{OR}$ and α is 0.04. By contrast, the moving average of the top three statistics is $0.045 < \alpha$; hence SMART rejects three hypotheses. The gap between the moving average and the threshold is only 0.005. Thus the boundary α can be approximated more precisely by the moving average. In large-scale testing problems, the gap is almost negligible, implying very accurate error rate control; this has been corroborated by our simulation studies.

We provide further insights by viewing the decision process of SMART as a 0-1 knapsack problem with varying FPR capacities (Gang et al., 2023). Let $k_j^* = \max\{1 \le i \le m : T_{OR}^{(i),j} \le \alpha\}$. Then the discovery step in Algorithm 1 can be written as

$$\sum_{i=k_j^*+1}^{k_j^s} \left(T_{OR}^{(i),j} - \alpha\right) \le \sum_{i=1}^{k_j^*} \left(\alpha - T_{OR}^{(i),j}\right). \tag{4.1}$$

If we view the nominal FPR level α as the initial capacity in a knapsack problem, then Equation 4.1 shows that every time we reject a hypothesis with $T_{OR}^{i,j} < \alpha$, we will increase the capacity of the knapsack, enabling one to reject hypotheses with $T_{OR}^{i,j} > \alpha$. Hence the approximation errors (overshooting) can be effectively mitigated by aggregating the gaps between T_{OR} and α to increase the capacity and make more discoveries.

4.3 Connection to compound sequential decision theory

The sparse recovery problem (1.6) is closely related to compound sequential decision theory. Suppose we are interested in classifying θ_i based on observed data. Consider the following loss function:

$$L_{\lambda_1,\lambda_2}(\boldsymbol{\theta}, \boldsymbol{d}) = \lambda_1 \left\{\sum_{i=1}^p (1-\theta_i)\delta_i\right\} + \lambda_2 \left\{\sum_{i=1}^p \theta_i(1-\delta_i)\right\} + \sum_{i=1}^p N_i, \tag{4.2}$$

where λ_1 and λ_2 are the relative costs. The sum of the first two terms in (4.2) corresponds to the total decision errors, and the last term gives the total sampling costs. The optimal solution to this weighted classification problem is the *Bayes sequential procedure* $\boldsymbol{d}^\pi$ that minimizes the expected loss

$$E\{L_{\lambda_1,\lambda_2}(\boldsymbol{\theta}, \boldsymbol{d}^\pi)\} = \inf_{\boldsymbol{d}} E\{L_{\lambda_1,\lambda_2}(\boldsymbol{\theta}, \boldsymbol{d})\}. \tag{4.3}$$

Our oracle procedure in Theorem 2 is inspired by a classical result in Berger (1985) which states that $\boldsymbol{d}^\pi$ is a thresholding rule based on $T_{OR}^{i,j}$. The intuition is that with appropriately selected tuning parameters λ_1 and λ_2 the FPR/MDR problem in (1.6) is equivalent to the problem in (4.2). A proof that the FDR control problem is equivalent to a weighted classification problem can be found in Sun and Cai (2007).

Remark 10 *As pointed out by an insightful reviewer, Ouhamma et al. (2021) considered the following loss function*

$$L_T = \sum_{i=1}^p a_i \mathbb{I}(\theta_i \ne \delta_i), \tag{4.4}$$

where the subscript T denotes the total budget. The loss (4.4) *is similar to* (4.2) *in that the weight a_i can be selected to reflect the relative severity of making a false positive and a false negative decision. However,* (4.4) *is designed for the "fixed budget" regime, which aims to minimize the total number of errors while adhering to a budget constraint of T. Under such a constraint, it may be infeasible to achieve the objective of simultaneously controlling both the FPR and MDR at user-specified levels. In contrast, the loss function* (4.2) *considers the trade-off between the total sampling cost $\sum_{i=1}^p N_i$ and the total decision errors $\lambda_1 \left\{\sum_{i=1}^p (1-\theta_i)\delta_i\right\} + \lambda_2 \left\{\sum_{i=1}^p \theta_i(1-\delta_i)\right\}$. This explicit trade-off is crucial in deriving the SMART algorithm, and it has not been reflected in Equation* (4.4).

To gain more insights on the overall structure of the problem, we rewrite (4.2) as the sum of stage-wise losses:

$$L_{\lambda_1,\lambda_2}(\boldsymbol{\theta},\boldsymbol{d}) = \sum_{j=1}^{N}\left[\sum_{i\in\mathcal{S}_j}\{\lambda_1(1-\theta_i)\delta_i + \lambda_2\theta_i(1-\delta_i)\} + \mathrm{Card}(\mathcal{A}_j)\right], \tag{4.5}$$

where $N = \max\{N_i : 1 \le i \le p\}$ is the total number of stages and $\mathcal{S}_j = \{i : N_i = j\}$ is the subset of coordinates at which we stop sampling and make terminal decisions at stage j; the remaining locations will become the active set for the next stage $\mathcal{A}_{j+1} = \mathcal{A}_j \setminus \mathcal{S}_j$, on which new observations are taken. We then proceed to make further decisions on $\mathcal{A}_{j+1}$. The process will be continued until the active set becomes empty. This view from the simultaneous testing perspective motivates us to employ the idea of *ranking and thresholding* in designing the SMART procedure in Algorithm 1. Ranking and selection has been widely used in the multiple testing literature. For example, the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) first orders all p-values from the smallest to largest, and then uses a step-up method to choose a cutoff along the p-value ranking.

Remark 11 *We mention a few important distinctions between the methodology presented by Sun and Cai (2007) and our adaptation. Firstly, while Sun and Cai's focus is on controlling the FDR, the objective of our SMART algorithm is* signal recovery*. Specifically, Sun and Cai (2007) aims to maximize the power subject to constraints on the FDR, whereas SMART aims to minimize the expected sample sizes (ESS) subject to constraints on both the FPR and MDR. Secondly, SMART is a multi-stage method that is more intricate than Sun and Cai's one-stage approach. While Sun and Cai' method only conducts signal selection, SMART involves both noise elimination and signal selection. Finally, the optimality theories developed in Sun and Cai (2007) and our paper differ substantially. Sun and Cai (2007) demonstrated that their procedure has the largest power asymptotically amongst all FDR procedures at the nominal level α for* static data*. However, in the sequential setting with* dynamic data streams*, our primary focus is to optimize sampling efficiency, measured by ESS, subject to the FPR and MDR constraints. This has necessitated the development of new techniques for deriving a new optimality theory, which presents significantly greater theoretical challenges than those encountered by Sun and Cai (2007).*

5. Asymptotic Bounds and Optimality of SMART

This section discusses the notion of fundamental limits in sparse inference for both fixed and adaptive settings. The discussion serves two purposes. First, since the limits reflect the challenges of support recovery under different conditions, they demonstrate the advantages of using adaptive designs over fixed designs. Second, these limits provide an optimality benchmark for what can be achieved asymptotically by any inference procedure in the form of a lower bound; thus, we can establish the optimality of an inference procedure by demonstrating its capability of achieving this limit.

5.1 The fundamental limits in adaptive designs

In signal recovery with thousands, and even millions, of variables, we assume that $p \to \infty$ in the asymptotic analysis, where the FPR and MDR levels are denoted as α_p and γ_p.

The basic setup assumes that the null and alternative distributions F_0 and F_1 are identical across all locations. Let G and H be two distributions with corresponding densities g and h. Define

$$D(G\|H) = \int_{-\infty}^{\infty} g(x) \log\left(\frac{g(x)}{h(x)}\right) dx.$$

Consider a general multistage decision procedure $\boldsymbol{d}$. The performance of $\boldsymbol{d}$ is characterized by its total risk

$$R^*(\boldsymbol{d}) = \mathrm{FPR}(\boldsymbol{d}) + \mathrm{MDR}(\boldsymbol{d}).$$

Recall $\mathrm{ESS}(\boldsymbol{d})$ is the average number of measurements allocated to each unit. Intuitively, $D(F_0\|F_1)$ and $ESS(\boldsymbol{d})$ together characterize the possibility of constructing a $\boldsymbol{d}$ such that $R^*(\boldsymbol{d}) \to 0$. The fundamental limit described in the next theorem gives the minimum condition under which the goal $R^*(\boldsymbol{d}) \to 0$ can be achieved.

Theorem 12 Fundamental limits (lower bound). *Let $\boldsymbol{d}$ be a multistage adaptive testing rule. Assume that $\pi < \frac{1}{2}$. If*

$$\tau \equiv ESS(\boldsymbol{d}) \leq \frac{\log(1/4\eta)}{\max\{D(F_0\|F_1), D(F_1\|F_0)\}},$$

then we must have $R^*(\boldsymbol{d}) \geq \eta$ for all $\eta > 0$.

Remark 13 *The restrictive model with identical F_0 and F_1 is only for theoretical analysis, while our proposed SMART procedure works under the more general model given by equations* (1.1)-(1.3)*, which allows F_{1i} to vary across testing units.*

In asymptotic analyses we typically take an η that converges to 0 slowly. Theorem 12 shows that any adaptive procedure with total risk tending to zero must at least have an ESS (or the average sample size per dimension) in the order of $\log\{(4\eta)^{-1}\}$. This limit can be used as a theoretical measure to assess the efficiency of a multistage procedure.

Theorem 12 is closely related to Theorem 1 in Malloy and Nowak (2014a). Let $\mathcal{S} = \{i : \theta_i = 1\}$ and $\hat{\mathcal{S}} = \{i : \delta = 1\}$. Denote m the average number of samples per index and $\mathbb{P}_e = \mathbb{P}(\mathcal{S} \neq \hat{\mathcal{S}})$. With these definitions in place, we present below a reformulation of the theory in Malloy and Nowak (2014a) using our notation system.

Theorem 1 in Malloy and Nowak (2014a). Any uniform coordinate-wise sequential thresholding procedure with $m \leq \dfrac{\log s + \log(\frac{1}{4\delta})}{\max\{D(F_0\|F_1), D(F_1\|F_0)\}}$ must have $\mathbb{P}_e \geq 1 - e^{-\delta} \approx \delta$.

This fundamental limit for signal recovery was established under the FWER criterion, which can be too conservative for large-scale testing problems. Our Theorem 14 introduces a more appropriate framework for modern data applications. Specifically, our extension is based on a new statistic and a new criterion, resulting in a novel bound that accurately reflects the potential of reliable signal recovery at a reduced sampling cost, under the less conservative FPR and MDR paradigm.

Denote $\boldsymbol{d}_{SM} = (\boldsymbol{N}_{SM}, \boldsymbol{\delta}_{SM})$ the SMART procedure described in Algorithm 1. As we proceed, we need the following assumption that is essentially equivalent to Condition (8) in Malloy and Nowak (2014a):

$$\mathbb{E}(T_{OR}^{i,N_i} | T_{OR}^{i,N_i} > t_u) \leq \frac{t_u}{(1-t_u)e^{-C_1} + t_u}, \tag{5.1}$$

$$\mathbb{E}(T_{OR}^{i,N_i} | T_{OR}^{i,N_i} < t_l) \geq \frac{t_l}{(1-t_l)e^{C_2} + t_l} \tag{5.2}$$

for all possible thresholds $t_l < t_u$. The condition is satisfied when $\log \mathcal{L}_{i,1}$ follows a bounded distribution such as Gaussian and exponential distributions. A more detailed discussion on this issue can be found in Ghosh (1960). The following theorem derives the upper bound.

Theorem 14 Asymptotic optimality (Upper bound). *Consider the SMART procedure described in Algorithm 1 with lower and upper thresholds*

$$\tilde{t}_{OR}^{l} = \frac{1}{\xi(p)^{1+\epsilon}} \quad \textit{and} \quad \tilde{t}_{OR}^{u} = \frac{(1-\pi)\xi(p)^{1+\epsilon}}{\pi + (1-\pi)\xi(p)^{1+\epsilon}},$$

where $\epsilon > 0$ is an arbitrarily small constant and $\xi(p)$ is a function of p that grows to infinity at an arbitrarily slow rate. Then under (5.1)-(5.2), the SMART procedure satisfies $\lim_{p\to\infty} R^*(\boldsymbol{\delta}_{SM}) = 0$ and

$$\limsup_{p\to\infty} \left(ESS(\boldsymbol{N}_{SM}) - \frac{(1+\epsilon)\log \xi(p)}{\min\{D(F_0\|F_1), D(F_1\|F_0)\}} \right) \leq 0.$$

Consider the asymptotic regime presented in Model (1.8), where $F_0 = N(0,1)$, $F_1 = N(\mu_p, \sigma^2)$, $\pi_p = p^{-\beta}$ and $\mu_p = \sqrt{2r\log p}$. Some calculations reveal that

$$D(F_0\|F_1) = \log\sigma + \frac{1+\mu_p^2}{2\sigma^2} - \frac{1}{2} \quad \text{and} \quad D(F_1\|F_0) = -\log\sigma + \frac{\sigma^2+\mu_p^2}{2} - \frac{1}{2},$$

both of which are of order μ_p^2. Theorem 14 shows that SMART is capable of achieving the limit given in Theorem 12 up to a constant multiple. If we take $\eta = 1/\xi(p)$ with $\xi(p) \to \infty$ slowly, then the rates in Theorems 12 and 14 would match. Theorems 12 and 14 together show that SMART provides a good solution to Problem (1.6) under the asymptotic setup.

Remark 15 *Theorem 12 is a necessary intermediate step for establishing the optimality of SMART. The proofs of Theorems 12 and 14 adopt the theoretical framework in Malloy and Nowak (2014a) and mirror their approach. Malloy and Nowak (2014a) showed that when DS or SST are considered, the minimum condition under which $R^*(\boldsymbol{d}) \to 0$ requires the ESS per unit to be of the order of $\log s$, where s is the number of signals. In the asymptotic setup where $s \to \infty$ as $p \to \infty$, the result from Malloy and Nowak (2014a) effectively imposes a lower bound on the rate of the required ESS necessary for $R^*(\boldsymbol{d}) \to 0$. Our theory shows that, when the goal is to control FPR and MDR, it is possible to achieve $R^*(\boldsymbol{d}) \to 0$ with the ESS being of the order of $\xi(p)$, which is allowed to converge to infinity at an arbitrarily slow rate. This rate can be, of course, slower than $\log s$. Our theory is corroborated by numerical results that demonstrate the efficiency gain of SMART over DS and SST under the FPR-MDR paradigm, characterized by the ESS per unit, in many settings.*

6. Connection to Existing Works

In this section we compare SMART with existing works on sequential analysis.

6.1 Comparison with distilled sensing and single sequential thresholding

The distilled sensing (DS, Haupt et al. (2011)) and single sequential thresholding (SST, Malloy and Nowak (2014a)) methods provide efficient adaptive sampling schemes for sparse recovery. Both works consider the following setup:

$$X_{ij} = \mu_i + \gamma_{ij}^{-1/2}\epsilon_{ij}, \quad i = 1, \ldots, p, \quad j = 1, 2, \ldots,$$

where γ_{ij} may be conceptualized as the "sample size" collected at location i stage j. In the setting of our paper, $\gamma_{ij} = 0$ or 1, where $\gamma_{ij} = 0$ indicates that no observation is collected at (i, j). Both DS and SST aim to recover the support under the budget constraint $\sum_{j=1}^{n}\sum_{i=1}^{p}\gamma_{ij} \leq R(p)$, where $R(p)$ is an increasing function of p that represents the total measurement budget. A prototype DS algorithm operates as follows: it starts with the full set $\mathcal{A}_1 = \{1, \cdots, p\}$, then performs k stages of distillation, where k is determined by a pre-specified $R(p)$. At stage j, $x_{ij} = \mu_i + \epsilon_{ij}$ is observed and the active set is updated as $A_{j+1} \leftarrow \{i \in \mathcal{A}_j : x_{ij} > 0\}$. This simple operation indicates that only locations with positive observations are kept. The DS algorithm outputs A_k as the final "discovery set". If the signals are sparse, then each stage eliminates about half of the null locations. The operation of SST is similar: at each step the procedure takes $R(p)/(2p)$ samples from each active location. The summary statistic at location j is obtained as the sum of $R(p)/(2p)$ observations. If the sum is less than zero, then no observation will be collected for that coordinate in subsequent passes. In both DS and SST, the observations from previous stages are abandoned and the compound structure of the problem is ignored, which leads to efficiency loss.

SMART employs an adaptive sampling strategy that is equipped with a simultaneous inference framework to control user-specified error rates. It improves DS and SST in several ways. First, as opposed to DS and SST, which use a fixed thresholding rule $\mathbb{I}(x_{ij} > 0)$ at all units and through all stages, SMART uses data-driven thresholds that are *adaptive* to both data and pre-specified FPR and MDR levels. Second, the sampling and inference rules in DS and SST only use observations at the current stage. By contrast, the building block of SMART is $T_{OR}^{i,j}$, which uses all observations collected up to the current stage. This can greatly increase the signal to noise ratio and lead to more powerful testing and stopping rules. Third, SMART uses the powerful compound thresholding idea to exploit the multiple hypothesis structure, which can greatly improve the accuracy of inference and lead to more precise error rates control. Finally, DS and SST only have a stopping rule to eliminate null locations. Intuitively, such a scheme is inefficient as one should also stop sampling at locations where the evidence against the null is extremely strong. The proposed scheme in Step 2 of SMART involves two algorithms that can simultaneously identify signals and eliminate noise. In other words, SMART stops sampling once a definitive decision ($\delta_i = 0$ or $\delta_i = 1$) is reached; this more flexible operation is desirable and would further save study budgets.

6.2 Comparison with Bartroff and Song (2013)

We provide a description of the sequential procedure by (Bartroff and Song, 2013), denoted BS, and contrast it with the SMART algorithm. As the BS procedure is a bit complicated, we only discuss high level ideas and omit some technical details below.

Let Λ_j^i denote the test statistic for coordinate i at stage j, $1 \le i \le p$, $j = 1, 2, \cdots$. BS first assigns a sequence of critical values

$$A_1^i \le A_2^i \le \ldots \le A_p^i \le B_p^i \le B_{p-1}^i \le \ldots \le B_1^i \tag{6.3}$$

to each location i. Then the BS procedure conducts simultaneous tests at all coordinates in a sequential manner. It stops sampling until $\Lambda_j^i \notin (A_1^i, B_s^i)$ or $\Lambda_j^i \notin (A_s^i, B_1^i)$; a rejection/acceptance decision is made if some Λ_j^i crosses a certain critical value: if $\Lambda_i^j \notin (A_1^j, B_s^j)$[or $\Lambda_i^j \notin (A_s^j, B_1^j)$], i.e. Λ_i^j crosses B_s^j [or A_s^j] first, then the null hypothesis is rejected [or accepted]. The sequence of critical values $A_1^i \le A_2^i \le \ldots \le A_p^i \le B_p^i \le B_{p-1}^i \le \ldots \le B_1^i$ is assigned in such a way so that the decision rule describe above controls both Type I and Type II error probabilities. Bartroff and Song (2013) recommends using Monte Carlo simulation for computing the sequence of critical values.

Bartroff and Song (2013) employ a strategy that combines several techniques such as (a) the thresholds in the step-up procedure by Benjamini-Hochberg, (b) sequential (or group sequential) MC sampling methods of Gaussian data, and (c) the utilization of limiting distributions of the signed roots of complicated equations, for computing these critical values. It turns out that the computation of the sequence (6.3) can be quite complicated and intensive. As the computational burden is in the order of $O(p^3)$, the simulation study in Bartroff and Song (2013) only considers $p = 20$ data streams.

In contrast, SMART is built upon the compound decision theory and local false discovery rate (Efron, 2001; Sun and Cai, 2007) ideas. Its operation is fundamentally different from existing SPRTs, and is particularly suitable for large-scale testing problems. The thresholds can be computed quickly using a simple step-up procedure that employs elegant ideas similar to those in Sun and Cai (2007). This step-up algorithm, which involves ranking and thresholding, has computational complexity of $O(p \log(p))$. Hence SMART is capable of handling millions of data streams. The thresholds lead to effective error rates control and substantially outperform competing methods in expected sample size (power).

6.3 Connection to e-values

In various statistical applications, e-variables (Shafer, 2021; Wang and Ramdas, 2022) naturally arise as sequential likelihood ratios under the null: $\mathcal{L}_{i,j} = f(\boldsymbol{X}_i^j|\theta_i = 1)/f(\boldsymbol{X}_i^j|\theta_i = 0)$. One potential solution to our problem is to incorporate e-variables into sequential testing by using the e-BH procedure (Wang and Ramdas, 2022). However, this approach comes with limitations that require further investigation in future research. Firstly, the power of e-BH may be significantly lower than that of the SMART approach, which is likely due to the robustness of e-BH, which remains valid under arbitrary dependence. Secondly, e-BH only controls the FDR, while the signal recovery problem requires the control of both the FPR and MDR, with the latter being an unresolved issue under the e-value framework. Lastly, theoretical bounds on signal recovery using e-values require further exploration.

6.4 Connection to multi-armed bandits

We discuss a related line of research on multi-armed bandits. Both multi-armed bandits and SMART algorithms aim to efficiently identify a small number of arms/units from many potential candidates. The multi-arm bandits problem has many variations that require different approaches. For instance, should the focus be on "exploring" or "exploiting"? If the goal is to "explore," should the emphasis be on finding the top M arms or identifying arms with mean parameters that exceed some threshold? Is the agent given a fixed budget beforehand, or is the aim to identify "good" arms with high probability? While these problems may seem similar, each has its unique nuances. Adapting solutions from one problem to another is a highly non-trivial task. Consequently, we find that solutions developed for fixed budget settings or identifying the top M arms are unsuitable for our problem without significant modification. In what follows, we discuss several important differences between our work and existing literature on the problem.

First, we have tackled the problem from a statistical perspective; the focus has been on the risk quantification of sequential decisions, and particularly the control of the error rates in multiple hypothesis testing. By contrast, the multi-armed bandits formulation solves the problem from an operational perspective. For example, the algorithms in Degenne et al. (2019); Chen and Li (2015); Karnin et al. (2013); Mason et al. (2020); Katz-Samuels and Jamieson (2020) focus on identifying the top M arms, while the associated error rates or statistical risks in sequential decisions are not investigated along the way. In Hong et al. (2022) the objective is to minimize regret by emphasizing "exploitation". Our goal is to efficiently identify non-null locations while providing theoretical guarantees on both the FPR and MDR control. It is unclear how the algorithm and theory presented in Hong et al. (2022) can be adapted to address our specific problem. Moreover, under the fixed budget regime (Malloy and Nowak (2014a); Locatelli et al. (2016); Mukherjee et al. (2017)), an upper bound on the number of samples is pre-specified, and the theoretical analyses focus on whether it is possible to develop an algorithm that can identify arms above a certain threshold effectively. In contrast, we adopt the setup in sequential hypothesis testing, where the sampling process continues until definitive decisions are reached at all locations. An important feature of our algorithm is that the error rates (FPR and FNR) are guaranteed to fall below the target levels asymptotically in the entire sampling-decision process.

Second, the error rates considered in the two lines of works are different. The analysis of multi-armed bandits algorithms focuses on (a) the probability of finding the best arm (Chen and Li, 2015), (b) the probability of finding all "good" arms (Mason et al., 2020), or (c) the probability that all identified arms are "good" and none are "bad" (Locatelli et al., 2016; Katz-Samuels and Jamieson, 2020). However, the goals in (b) and (c) can be too stringent for the large-scale setting. It is extremely difficult to identify *all* good arms or providing guarantees that *all* arms are good when thousands or more arms are being considered. Concretely, in large-scale testing problems, researchers are happy as long as most, say 90% of the good arms (missed discovery rate MDR is controlled at 0.1) can be identified correctly, or 90% of the arms that we identified are good (false positive rate FPR is controlled at 0.1). In such scenarios the FPR and MDR adopted in our formulation seem to provide more appropriate error rate targets.

Finally, the operational characteristics of the algorithms in the two lines of works are different. Under the conventional framework of multi-armed bandits, it is difficult to tailor algorithms according to user-specified error rates/probabilities. By contrast, we develop a convenient user interface where data-driven thresholds can be determined adaptively according to user-specified error rates. Our algorithms aim to control the error rates below the pre-specified levels at all decision points, while minimizing the total study budget (in terms of the total samples to be collected).

7. Simulation

This section is organized as follows. Section 7.1 compares SMART with simultaneous SPRTs (with single thresholding) to illustrate the advantage of compound thresholding. Section 7.2 and 7.3 compare the ranking efficiency of SMART vs. existing methods for sparse recovery and error rate control. We show that SMART achieves the same FPR and MDR levels with a much smaller study cost. Section 7.4 investigates the robustness of SMART for error rate control under a range of dependence structures. Section 7.5 illustrates the use of SMART as a global testing procedure.

7.1 Simple thresholding vs. compound thresholding

This simulation study compares the following methods:

(a) the simple thresholding procedure $\tilde{\boldsymbol{d}}_{OR}$ presented before Theorem 4, which assumes an oracle knows the true parameters (OR.ST);

(b) the SMART procedure with known parameters (OR.SM);

(c) the simple thresholding procedure with estimated parameters (DD.ST), where DD refers to "data-driven";

(d) the SMART procedure with estimated parameters (DD.SM).

We generate data according to the following mixture model

$$X_{ij} \sim (1-\pi)N(0,1) + \pi N(\mu_i, 1),$$

for $i = 1, \ldots, p$ and $j = 1, 2, \ldots$, where π and μ_i will be specified shortly. The number of locations is $p = 10^5$. Let $(\alpha, \gamma) = (0.05, 0.05)$ be pre-specified FPR and MDR levels. The following settings are considered:

- Setting 1: Set $\pi = 0.01$ and $\mu_i = \mu$ for all i. Vary μ from 2 to 4 with step size 0.2.
- Setting 2: Set $\pi = 0.05$ and $\mu_i = \mu$ for all i. Vary μ from 2 to 4 with step size 0.2.
- Setting 3: Draw μ_i randomly from a uniform distribution $\mathcal{U}(2, 4)$. Vary π from 0.05 to 0.2 with step size 0.01.

We apply the four methods to the simulated data. The FPR, MDR and ESS (expected sample size) are computed based on the average of 100 replications, and are plotted as

functions of varied parameter values. The results are summarized in Figure 3. In Setting 3, the two oracle methods (OR.ST and OR.SM) are not implemented as recursive formulae for $T_{OR}^{i,j}$ are unavailable.

We can see that all four methods control the FPR at the nominal level. However, the two single thresholding methods (OR.ST and DD.ST) are very conservative (the actual FPR is only about half of the nominal level). Similarly, both OR.ST and DD.ST are very conservative for MDR control. Specifically, OR.ST and DD.ST operate as p parallel SPRTs and suffer from the overshoot problem. The approximation errors of SPRTs can be greatly reduced by SMART that employs the compound thresholding strategy. We can see that the SMART procedures (OR.SM and DD.SM) control the error rates more accurately and require smaller sample sizes. When signals are sparse and weak (top middle panel), the MDR level of DD.SM is slightly higher than the nominal level. This is due to the estimation errors occurred at stage 1. It is of interest to develop more accurate estimation procedures in such settings. The FDR and FPR levels are similar for OR.SM and DD.SM. See Figure 2 in the supplement for details.

7.2 SMART vs. Distilled Sensing

This simulation study compares the efficiency performance of SMART vs DS (Haupt et al., 2011). As DS does not provide precise error rates control, the simulation is designed in the following way to make the comparison on an equal footing: we first run DS with up to 10 stages and record its FDR and MDR levels. Then we apply SMART at the corresponding FDR and MDR levels so that the two methods have roughly equal error rates. The ESS is used to compare the efficiency.

The data are generated from the same model as that in Section 7.1. The number of locations is $p = 10^5$. The following two settings are considered:

- Setting 1: Let $\pi = 0.05$ and $\mu_i = \mu$ for all i. Vary μ from 2 to 4 with step size 0.2.
- Setting 2: Draw μ_i randomly from $\mathcal{U}(2, 4)$. Vary π from 0.05 to 0.2.

In both settings, the FDR, MDR and ESS are computed by averaging over 100 replications, and are plotted as functions of varied parameter values. The results are summarized in Figure 4.

We can see that the error rates of both OR.SM and DD.SM match well with those of DS, but they require fewer samples. OR.ST and DD.ST also outperform DS, achieving lower error rates with fewer samples. DD.SM controls the error rates more accurately compared to DD.ST, and requires fewer samples.

7.3 SMART vs. the Hunt procedure

This section conducts simulation studies to compare SMART with the multi-stage testing procedure proposed in Posch et al. (2009), which is denoted "Hunt" subsequently. We set the nominal FDR and MDR levels to be 0.05 and generate $m = 10^5$ observations from the model in Section 7.1 We consider two settings:

- Setting 1: Vary π from 0.05 to 0.2 and generate $\mu_i \sim \mathcal{U}(2, 4)$.

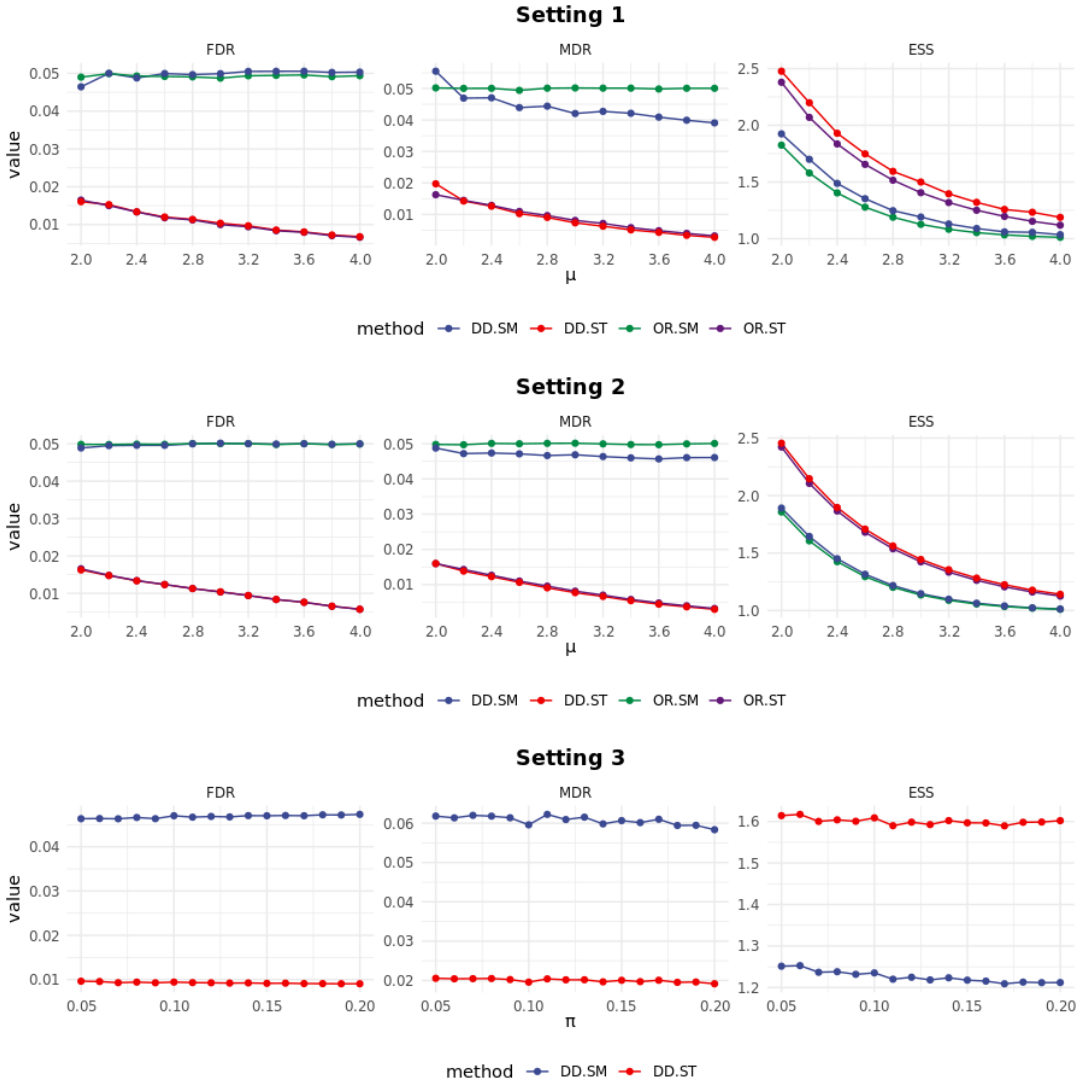


Figure 3: SMART vs. single thresholding: the displayed procedures are DD.SM, OR.SM, DD.ST, OR.ST.

- Setting 2: Fix $\pi = 0.05$. Let $\mu_i = \mu$, vary μ from 2 to 4.

We apply Hunt and data-driven SMART and compute the FDR, MDR and the total sample size by averaging results in 100 replications. The Hunt method operates under a fixed stage regime; we fix the number of stages to be 2 for illustration. The sample size required for the Hunt procedure is always $2m = 2 \times 10^5$. The simulation results are summarized in Figure 5. We can see that SMART effectively controls both the FDR and MDR. By contrast, Hunt controls the FDR but does not control the MDR. This is because Hunt is not designed to control the missed discovery rates. Moreover, SMART requires a smaller sample size to achieve effective error rate control. The intuition is that the group sequential design adopted by Hunt does not eliminate the null coordinates in the first stage; this leads to loss in efficiency and higher study costs.

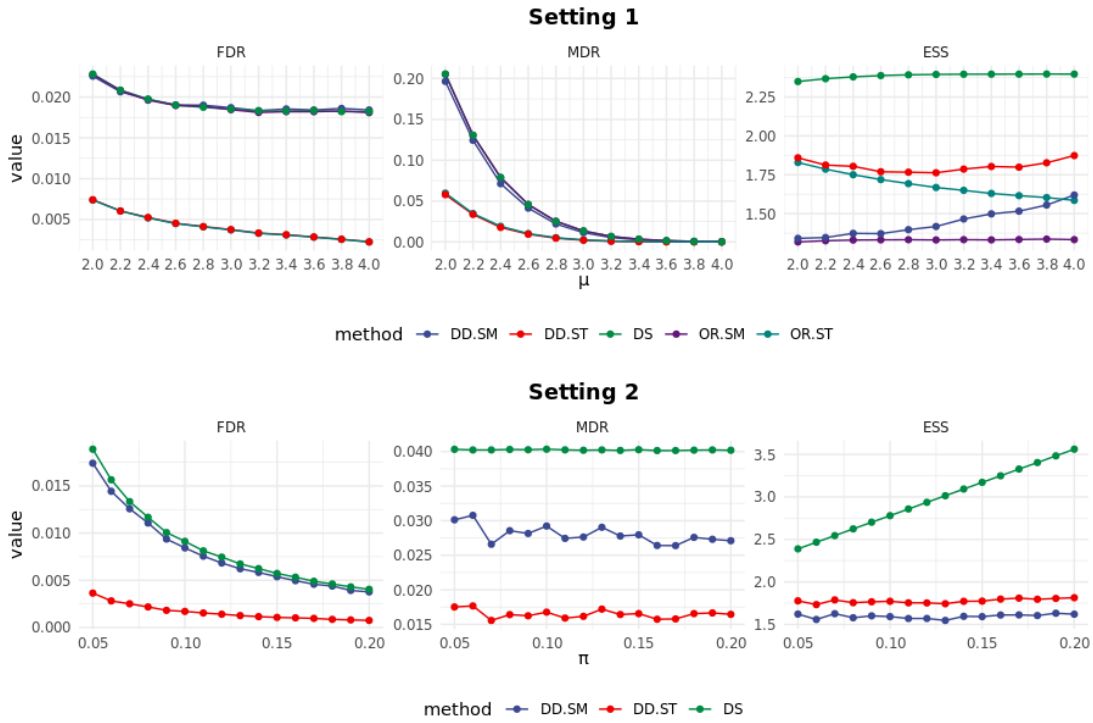


Figure 4: Comparison with DS. SMART outperforms DS in the sense that it achieves the same FDR and MDR levels with much smaller sample sizes.

7.4 Robustness of SMART under dependence

Our method and theory assume independence of observations across the coordinates. However, the independence assumption may be violated in practice. This section carries out additional numerical studies with correlated error terms across locations to investigate the robustness of the proposed SMART algorithm.

For the sequential testing setup, we first generate non-null effects from $\mathcal{U}(2,4)$ at $p = 10^4$ locations. Then we apply the single thresholding procedure DD.ST and SMART procedure DD.SM to simulated data with correlation structures described in Settings 1-3 below. We use the computational algorithm described in Section 3.2. The FPR, MDR and ESS (expected sample sizes) are then calculated based on averaging results from 200 data sets, and finally plotted as functions of the non-null proportion π. The simulation results are summarized in Fig. 2.

- Setting 1 (Gaussian process): the error terms are generated from a zero-mean Gaussian process with the covariance functions

$$C(r) = e^{-r^2}, \text{ with } r \geq 0 \text{ being the distance between two locations.}$$

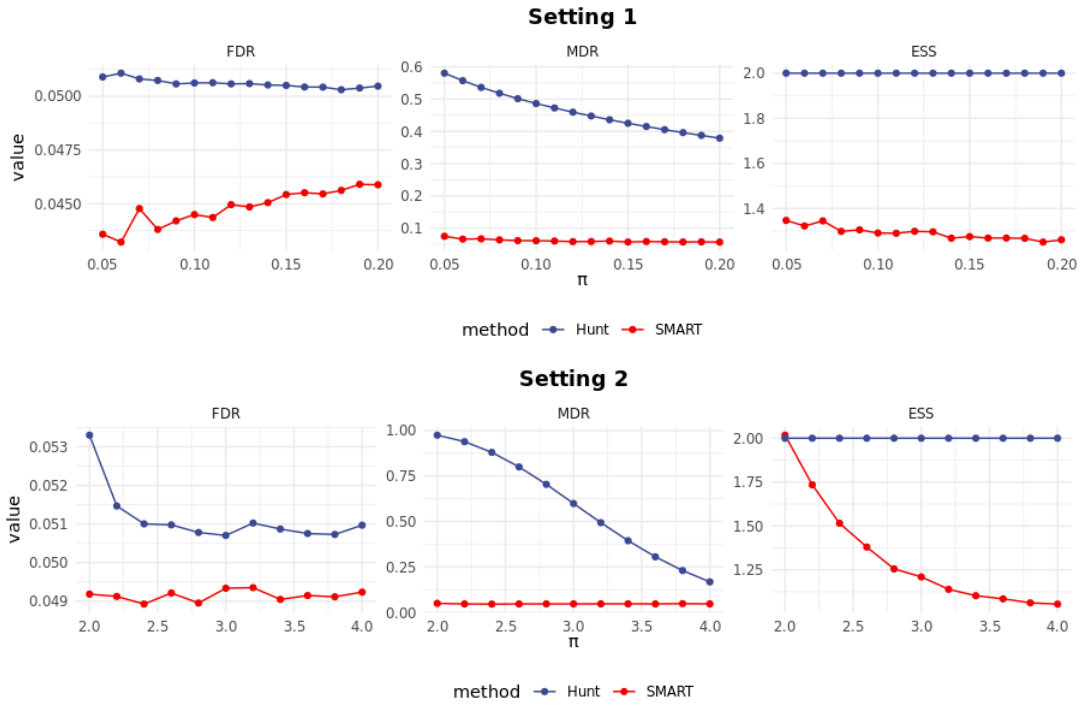


Figure 5: SMART vs Hunt. Top and bottom rows correspond to Settings 1 & 2.

- Setting 2 (Tridiagonal): the error terms are generated from a multivariate normal distribution with the following covariance matrix Σ:

$$\begin{aligned}
\Sigma_{i,i} &= 1,\ i = 1, \cdots, p, \\
\Sigma_{i,i+1} &= 0.5,\ i = 1, \cdots, p-1, \\
\Sigma_{i-1,i} &= 0.5,\ i = 2, \cdots, p, \\
\Sigma_{i,i+2} &= 0.4,\ i = 1, \cdots, p-2, \\
\Sigma_{i-2,i} &= 0.4,\ i = 3, \cdots, p.
\end{aligned}$$

  All other entries equal to zero.

- Setting 3 (Autoregressive model): the error terms are generated from a multivariate normal distribution with the following covariance matrix Σ:

$$\Sigma_{i,j} = 0.5^{|i-j|},\ i, j = 1, \cdots, p.$$

We observe some similar patterns as before. SMART (DD.SM) roughly controls the FPR at the nominal level when moderate correlations are present between locations. The single thresholding procedures with parallel SPRTs (DD.ST) are too conservative for error rates control in all settings. The resulting ESS of DD.ST is significantly larger than that is needed by DD.SM. In Setting 2 the FPR of SMART can be higher than the nominal level 0.05 but seems to be acceptable. SMART controls the MDR under the nominal level in all settings. Our simulation studies suggest that SMART has robust performance under a range

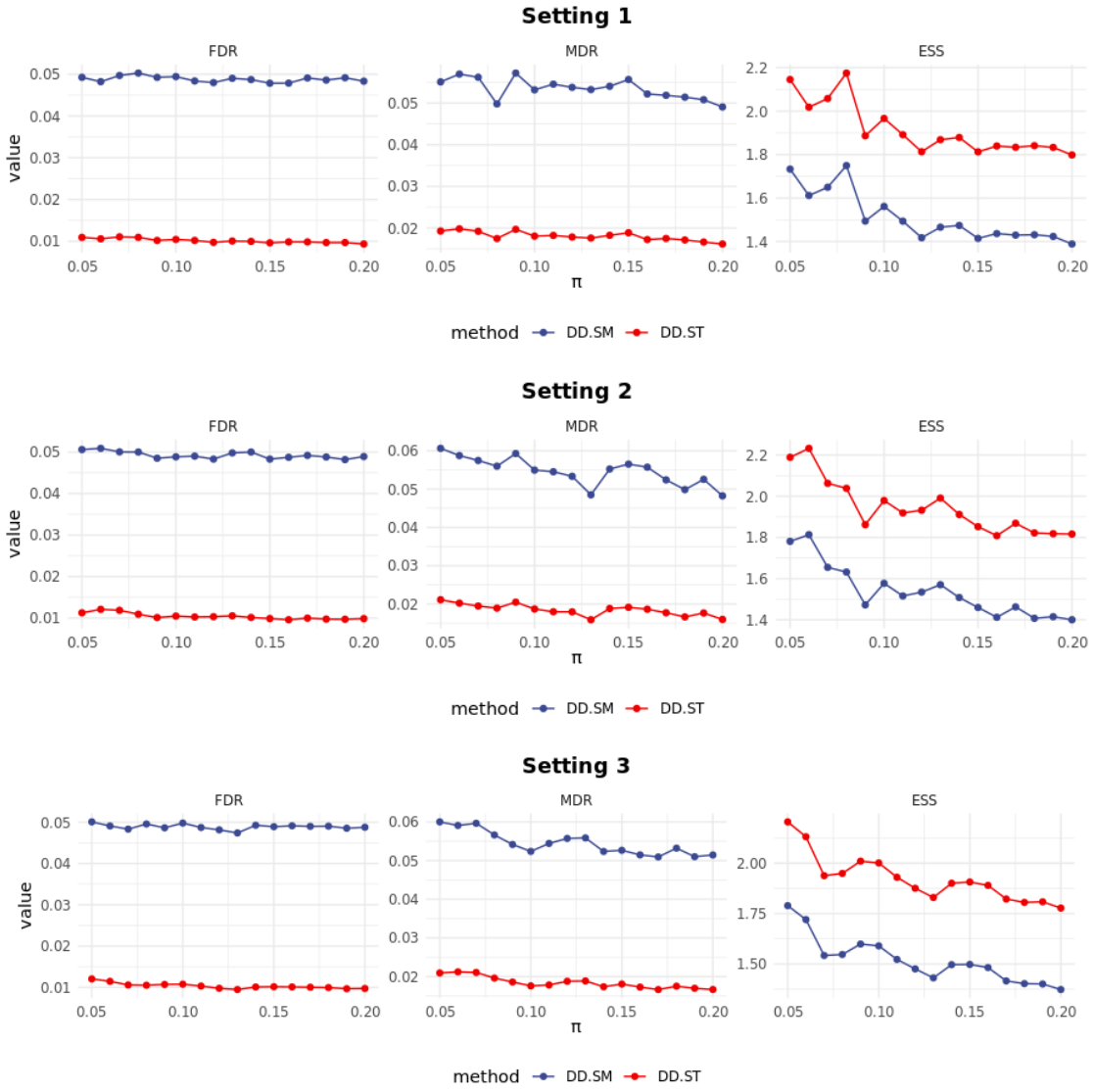


Figure 6: Dependent tests: Settings 1-3 are described in the main text. SMART controls the error rates at the nominal levels in most settings.

of dependence structures. However, the scope of our empirical studies is limited and more numerical and theoretical support are needed for making definitive conclusions. Finally, we stress that it is still an open issue to develop multiple testing methods that incorporates informative correlation structures to improve efficiency Sun and Cai (2009); Sun et al. (2015). The modeling of spatial dependence in the sequential setting is complicated because we have a lot of missing data in later stages.

7.5 SMART for signal detection

SMART can be employed as a method for signal detection Donoho and Jin (2004). The goal is to test the global null hypothesis

$$H_0^p : \{X_1, \cdots, X_p\} \sim F_0 \text{ vs. } H_1^p : \{X_1, \cdots, X_p\} \sim (1-\pi)F_0 + \pi F_{1i} \tag{7.1}$$

for some $\pi > 0$. This is a *global inference* problem that is very different from the *simultaneous inference* problem we have considered in previous sections. Comparing the detection boundaries derived in Donoho and Jin (2004) with the upper limit in Theorem 5, we can see that SMART requires fewer samples than non-adaptive testing schemes for separating H_0^p and H_1^p with negligible testing errors. Similar findings have been reported by Haupt et al. (2011).

This section carries out simulation studies to investigate the performance of SMART as a signal detection procedure. Specifically, SMART (Algorithm 1) eliminates noise locations stage by stage. If all coordinates are eliminated as noise locations eventually, then we do not reject the global null. If any coordinate is selected as signal, then we reject the global null. Under the global null $\pi = 0$, the false discovery proportion ($\text{FDP}(\boldsymbol{\delta}) = \frac{\sum_{i=1}^p (1-\theta_i)\delta_i}{\max\{(\sum_{i=1}^p \delta_i), 1\}}$) only takes two possible values: 0 (if no coordinate is selected as signal) and 1 (if at least one coordinate is selected as signal). It was argued by Benjamini and Hochberg (1995) that under the global null, the FPR cannot be controlled, whereas it is always possible to control the FDR. In our simulation, the FDR is calculated as the average of the FDPs over 200 replications; the average may be conceptualized as an estimate of the Type I error rate for testing the global null, i.e. the relative frequency for incorrectly rejecting the global null when $\pi = 0$ among all repeated experiments.

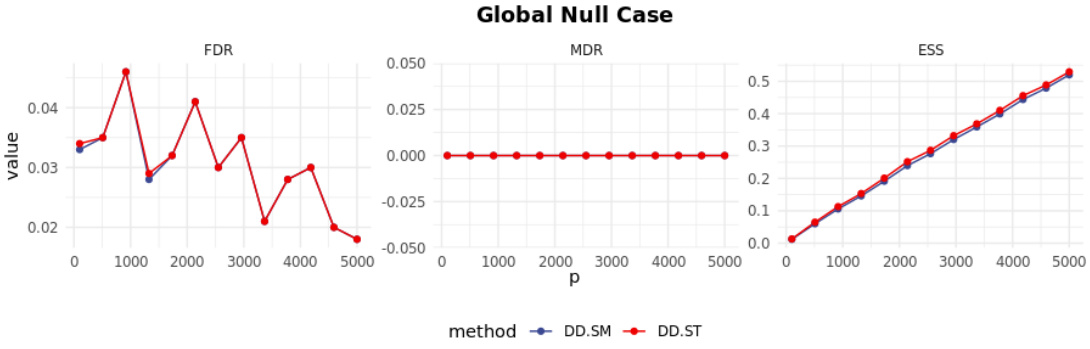


Figure 7: SMART for signal detection: the FDR is used as the Type I error rate.

We apply SMART by setting $\alpha = \gamma = 0.05$. The data are generated using the same models as those in previous sections except that we now let $\pi = 0$. We vary the dimension p from 100 to 5,000 and plot the corresponding FDRs. It can be seen that SMART controls the FDR at the nominal level 0.05 in all settings. The MDR and ESS are also reported. (If no true signals are discovered, then the MDR will be set as 0.) We conclude that SMART controls the Type I error rate for testing the global null.

8. Real Data Applications

In this section we apply the SMART procedure to A/B testing (Section 8.1), high-throughput screening (HTS, Section 8.2) and satellite image analysis (Section 8.3).

8.1 A/B Testing

A/B testing has become the gold standard for testing new ideas and building product launch plans in high-tech firms. This section illustrates the application of our SMART procedure to A/B testing using a real-world experiment at Snap Inc.. The quantity of interest is the average treatment effect (ATE) between the control and treatment arms (each arm consists of several hundred users). The ATEs are calculated on a daily basis for thousands of metrics. To save the study costs and speed up the decision process, we adopt an adaptive design to gradually narrow down (day by day) the set of metrics to be investigated. To check whether this application provides a suitable scenario for applying our SMART procedure, note that (a) it is reasonable to assume that the expected value of the ATE for a given metric is fixed over time, and (b) the ATEs, which are allowed to vary across testing units, are approximately normally distributed according to the central limit theorem. We conclude that the conditions (fixed μ_i over different stages and asymptotic normality of the summary statistic) required by our methodology have been fulfilled.

We analyze a data set from a multi-stage experiment with a running duration of 22 days. As the data are collected on a daily basis, it is natural to view that the experiment has 22 stages. At every stage, the ATEs are computed for a total of 1702 metrics and standardized as z-scores. We set the null distribution of the ATE statistic as $N(0, 1)$. Under the alternative, the ATE statistic at unit i has mean μ_i. We assume that μ_i is fixed at all stages $j = 1, 2, \cdots$, and allow μ_i to vary across i. The sparsity parameter $\hat{\pi}$ is estimated to be $\hat{\pi} \approx 0.15$ by applying Jin and Cai's method (Jin and Cai, 2007) to the data collected on Day 1. The pre-specified FDR and MDR levels are 0.05 and 0.2, respectively.

The solution path of the SMART procedure is shown in Figure 8. We can see that by using the adaptive design, definitive decisions on a large number of metrics can be made during earlier periods of the multi-stage experiment. This translates directly to great savings in data storage and pipeline calculations. Specifically, under the traditional A/B testing paradigm, one needs to collect data on 1,702 metrics throughout the 22-day duration. This requires a total of 44,252 measurements (each measurement corresponds to the cost of measuring one metric per day). By contrast, the SMART procedure only requires 3,541 measurements to localize important metrics with effective error rates control. We conclude that the new algorithm promises to save the study costs and greatly speed up the decision making process. Such benefits enable the company to quickly narrow down the list; hence

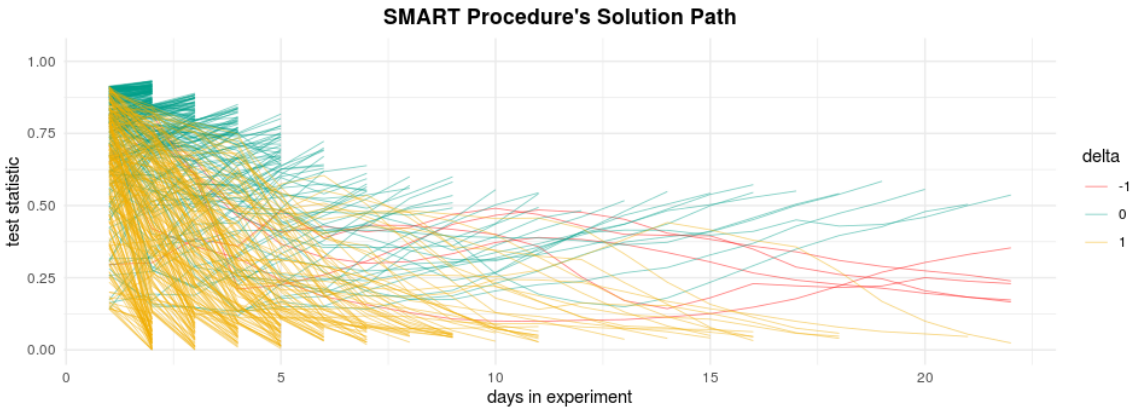


Figure 8: The solution path of our SMART procedure (data source: Snap Inc.): $\delta = 1/0/-1$ respectively indicates rejected/accepted/undecided hypotheses.

the efforts and resources can be focused on improving a few core metrics that are pivotal for business.

8.2 Applications to HTS studies

The goal of the HTS study conducted by McKoy et al. (2012) is to identify novel inhibitors of the amyloid beta peptide ($A\beta$), whose aggregation is believed to be a trigger of the Alzheimer's disease. In the study, a total of $p = 51,840$ compounds are tested, with three measurements recorded for each compound. We use the observed data set as a pilot data set and simulate observations in later stages to illustrate how to design a multistage sampling and inference procedure for identifying useful compounds.

We compare the performances of different methods in two ways: (i) the total sample sizes needed to achieve pre-specified error rates, and (ii) the actual error rates achieved for a fixed total sample size. We first obtain z-scores based on the average of the three measurements and then estimate the non-null proportion and null distribution using the method in Jin and Cai (2007). The estimated non-null proportion is $\hat{\pi} \approx 0.0007$, and the estimated null distribution (referred to as the *empirical null distribution*, Efron (2004)) is $N(\hat{\mu}_0, \hat{\sigma}_0^2)$ with $\hat{\mu}_0 = 0.2459, \hat{\sigma}_0 = 0.6893$. Next, we choose the largest $100\hat{\pi}\%$ of the data and use their average as the signal amplitude $\hat{\mu} = 3.194$. The observations in later stages will be generated based on the estimated parameters.

We set both the FDR and MDR at level 0.1, apply SMART and record the total sample size. We then apply DS with the recorded sample size by SMART. The results are summarized below. Since DS always eliminates half of the locations at each step, for this particular instance DS requires at least $1.5p$ observations. We can see that with the sample size of $1.5p$, the DS method does not offer proper error rate control. The false discovery proportion (FDP) is much higher than the nominal level.

Methods	FDP	MDP	Total Observation
SMART	0.083333	0.1081081	56926
DS	0.9971195	0	77641

Next, we run DS up to 10 stages. The recorded FDP and missed discovery proportion ($\text{MDP}(\boldsymbol{\delta}) = \frac{\sum_{i=1}^{p}\theta_i(1-\delta_i)}{(\sum_{i=1}^{p}\theta_i)}$) levels are 0.23 and 0.027, respectively. Next we apply SMART by setting the nominal FDR and MDR levels as 0.23 and 0.027 and compare the required sample sizes. The results are summarized below. We can see that SMART control the FDP and MDP below those of DS. Meanwhile, the required sample size is significantly smaller compared to DS.

Methods	FDP	MDP	Total Observation
DS	0.2340426	0.02702703	104308
SMART	0.1487284	0.02162162	67850

8.3 Application to the Phoenix Deep Survey

In astronomical surveys, a common goal is to separate sparse targets of interest (stars, supernovas, or galaxies) from background noise. We consider a dataset from Phoenix Deep Survey (PDS), a multi-wavelength survey of a region over 2 degrees diameter in the southern constellation Phoenix. The data set is publicly available. Fig. 5(a) shows a telescope image from the PDS. It has $p = 616 \times 536 = 330,176$ pixels, among which 1131 pixels exhibit signal amplitude of at least 2.98. In practice we monitor the same region for a fixed period of time. After taking high resolution images, it is of interest to narrow down the focus quickly using a sequential testing procedure so that we can use limited computational resources to explore certain regions more closely. The image is converted into gray-scale with signal amplitudes standardized. Fig. 5(b) depicts a contaminated image with simulated Gaussian white noise.

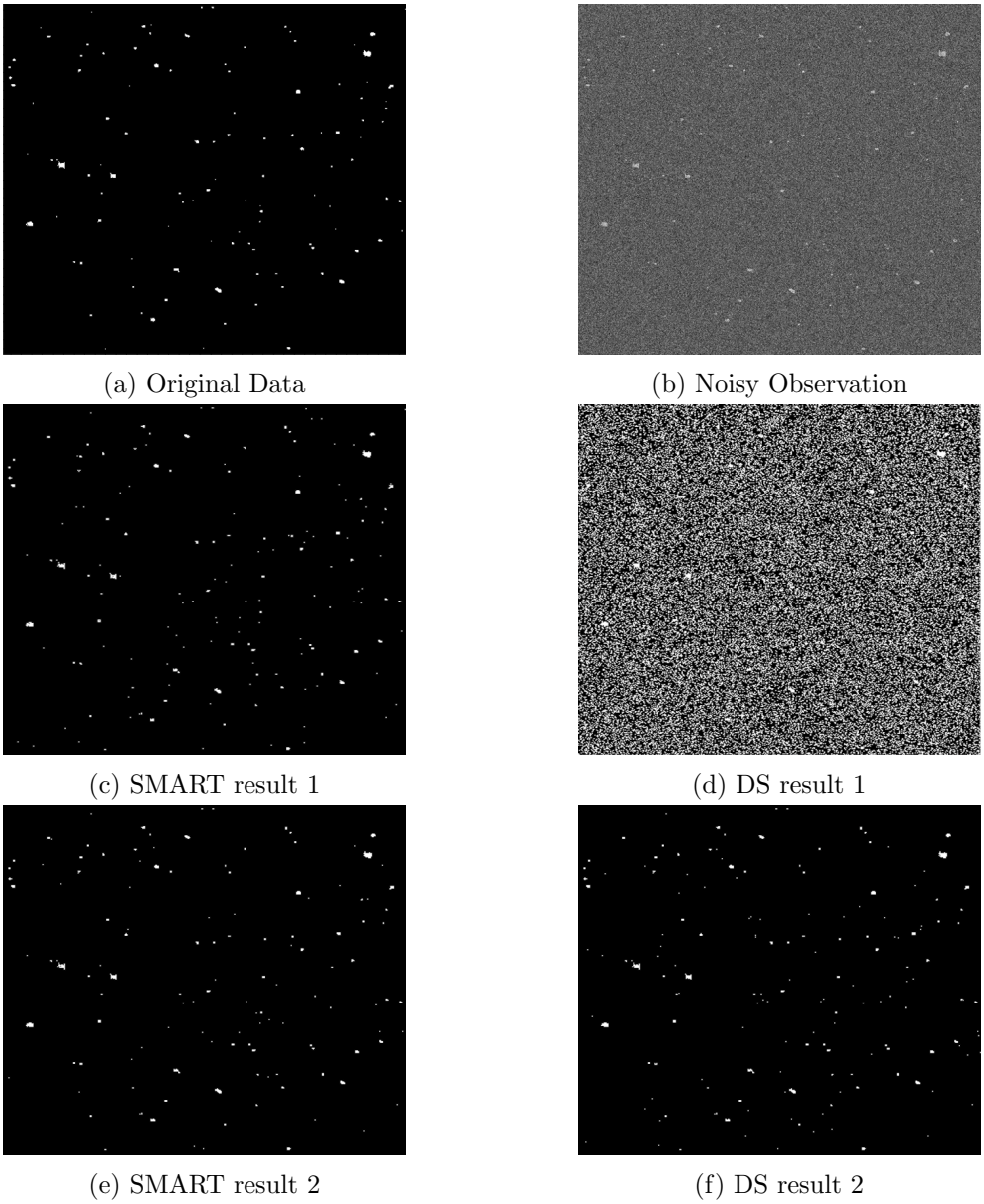
For the first comparison, we apply SMART by setting both the FPR and MDR at 5%. The total number of measurements is recorded as 368,796. As a comparison, we apply DS with the recorded sample size, which is approximately $1.5p$. We can see that SMART control the error rates precisely. The resulting images for SMART and DS are demonstrated in Fig. 5(c) and (d), respectively. We can see SMART produces much sharper images than DS.

Methods	FDP	MDP	Total Observation
SMART	0.06321335	0.05658709	368796
DS	0.9864338	0.00265252	495866

For the second comparison, we first implement DS up to 12 stages. The corresponding FDP and MDP levels are recorded as 0.07 and 0.014. We then apply SMART by setting the nominal FPR and MDR at these recorded error rates. The corresponding true error rates and required sample sizes of the two methods are summarized below.

Methods	FDP	MDP	Total Observation
DS	0.07237937	0.01414677	670331
SMART	0.03192407	0.00795756	479214

We can see that SMART controls both the FDP and MDP below the nominal levels (0.07 and 0.014). The required sample size is also much smaller than DS. The resulting images for SMART and DS are shown in Fig. 5(e) and (f), we can see SMART produces slightly sharper images than its competitor DS.



(a) Original Data (b) Noisy Observation

(c) SMART result 1 (d) DS result 1

(e) SMART result 2 (f) DS result 2

Figure 9: SMART and DS comparison. Fig. (a) and (b) show the original radio telescope image and tainted image with white noise, respectively. Fig. (c) and (d) compare SMART and DS when the total number of observations are about the same. Fig. (f) shows the resulting image when implementing DS for 12 stages, whereas Fig. (e) shows the image produced by SMART when using the recorded error rates from the 12-stage DS.

9. Discussion

We mention some potential limitations of the proposed SMART procedure and discuss directions for future research. First, SMART assumes the effect size at a certain coordinate is fixed over time. This is a reasonable assumption for the three applications we considered in this article, where repeated measurements are taken on the same unknown effect sizes over time. However, in application scenarios such as multiple stage clinical trials, the end points or treatment effect sizes may vary from one stage to another. Second, the issue on multiple testing dependence needs much research. Our limited simulation results show that SMART remains to be valid under weak and positive dependences. It would be of great interest to justify such results theoretically. Moreover, the optimality issue under dependence remains unknown. We hope to pursue these directions in future research.

The accuracy and effectiveness of sparse recovery, which involves both the FPR and MDR control, is a complicated issue that depends on several factors jointly. First, the reliable estimation of the stage-wise FPR/MDR requires a relatively large number of rejections, which in turn requires that the signals cannot be too sparse or too weak. Second, the consistent estimation of the non-null proportion (sparsity level) requires that the detection and discovery boundaries (Cai and Sun, 2017) must be achieved. The impact of sparsity on the finite sample performance of SMART is investigated in Appendix B of the Supplementary Material. Finally, the quality of density estimators depends on the sample size and smoothness of the underlying function. Our algorithm is designed to capitalize on copious data in ways not possible for procedures intended for moderate amounts of data, and SMART is most useful in large-scale testing scenarios where the structural information can be learned from data with good precision. An important future research direction is the development of precise estimation methods, which are instrumental for constructing powerful multi-stage testing procedures.

If we allow multiple samples at each stage and the number of samples are prefixed, then SMART operates essentially in the same way, and the theory on FPR and MDR control for the oracle SMART procedure still holds. However, the problem becomes complicated when the number of samples is allowed to be data-driven. Hence the optimal policy must be cast as a dynamic programming problem. We conjecture that the open-loop feedback control (OLFC) algorithm (Wei and Hero, 2013) may be incorporated into SMART but more research is needed. However, there are two complications. First, OLFC aims to optimize the estimation accuracy, whereas we aim to minimize the total sensing efforts subject to the constraints on the FPR and MDR. Second, OLFC can only handle two-stage designs and the extension to more stages is non-trivial.

10. Proofs

10.1 Proof of Theorem 2

Proof . **Part (a)** According to Assumption 1, $\mathbb{P}_{\theta_i}(N_i < \infty) = 1$ for all i; see Berger (1985) for a proof. Since $\mathbb{P}_{\theta_i}(N_i < \infty) = 1$ for all i, and p is finite, we claim that $\mathbb{P}(\max N_i < \infty) = 1$, i.e. the oracle procedure has a finite stopping time. Denote $\boldsymbol{d}^{\pi}(t_l, t_u) = \{(N_i, \delta_i) :$

$1 \le i \le p\}$. Using the definition of T_{OR}^{i,N_i}, we have

$$\mathbb{E}\left\{\sum_{i=1}^{p}(1-\theta_i)\delta_i\right\} = \mathbb{E}_{\boldsymbol{X}}\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X}}\left\{\sum_{i=1}^{p}(1-\theta_i)\delta_i\right\} = \mathbb{E}_{\boldsymbol{X}}\left(\sum_{i=1}^{p}T_{OR}^{i,N_i}\delta_i\right).$$

Then the FPR is $Q_{OR}(t_l,t_u) = \mathbb{E}\left(\sum_{i=1}^{p}T_{OR}^{i,N_i}\delta_i\right)/\mathbb{E}\left(\sum_{i=1}^{p}\delta_i\right)$. It follows that

$$\mathbb{E}\left[\sum_{i=1}^{p}\left\{T_{OR}^{i,N_i}-Q_{OR}(t_l,t_u)\right\}\mathbb{I}\left(T_{OR}^{i,N_i}\le t_l\right)\right] = \mathbb{E}\left[\sum_{i:T_{OR}^{i,N_i}\le t_l}\left\{T_{OR}^{i,N_i}-Q_{OR}(t_l,t_u)\right\}\right] = 0. \tag{10.1}$$

The above equation implies that $Q_{OR}(t_l,t_u) \le t_l$; otherwise every term on the LHS must be negative, resulting in a contradiction.

Next we prove that for a fixed t_u, $Q_{OR}(t_l,t_u)$ is non-decreasing in t_l. Let $Q_{OR}(t_{l,j},t_u) = \alpha_j$ for $j = 1,2$. We only need to show that if $t_{l,1} < t_{l,2}$, then $\alpha_1 \le \alpha_2$. Denote $N_{i,1}$ and $N_{i,2}$ the stopping times for location i corresponding to thresholds $(t_{l,1},t_u)$ and $(t_{l,2},t_u)$, respectively. If $t_{l,1} < t_{l,2}$, then it is easy to see that for any particular realization of the experiment, we must have $N_{i,1} \ge N_{i,2}$.

We shall show that if $t_{l,1} < t_{l,2}$ and $\alpha_1 > \alpha_2$, then we will have a contradiction. To see this, note that

$$\begin{aligned}
&\left(T_{OR}^{i,N_{i,2}}-\alpha_2\right)\mathbb{I}\left(T_{OR}^{i,N_{i,2}}\le t_{l,2}\right)\\
=\;&\left(T_{OR}^{i,N_{i,2}}-\alpha_2\right)\mathbb{I}\left(T_{OR}^{i,N_{i,2}}\le t_{l,1}\right)+\left(T_{OR}^{i,N_{i,2}}-\alpha_2\right)\mathbb{I}\left(t_{l,1}<T_{OR}^{i,N_{i,2}}\le t_{l,2}\right)\\
=\;&\left(T_{OR}^{i,N_{i,1}}-\alpha_2\right)\mathbb{I}\left(T_{OR}^{i,N_{i,1}}\le t_{l,1}\right)+\left(T_{OR}^{i,N_{i,2}}-\alpha_2\right)\mathbb{I}\left(t_{l,1}<T_{OR}^{i,N_{i,2}}\le t_{l,2}\right)\\
\ge\;&\left(T_{OR}^{i,N_{i,1}}-\alpha_1\right)\mathbb{I}\left(T_{OR}^{i,N_{i,1}}\le t_{l,1}\right)+(\alpha_1-\alpha_2)\mathbb{I}\left(T_{OR}^{i,N_{i,1}}\le t_{l,1}\right)\\
&+\left(T_{OR}^{i,N_{i,2}}-\alpha_1\right)\mathbb{I}\left(t_{l,1}<T_{OR}^{i,N_{i,2}}\le t_{l,2}\right).
\end{aligned}$$

The second equality holds because if $T_{OR}^{i,N_{i,2}} < t_{l,1}$, then we must have $N_{i,1} = N_{i,2}$. Taking expectations on both sides, we have

$$\begin{aligned}
\mathbb{E}\left\{\sum_{i=1}^{p}\left(T_{OR}^{i,N_{i,2}}-\alpha_2\right)\mathbb{I}\left(T_{OR}^{i,N_{i,2}}\le t_{l,2}\right)\right\} &= 0, \quad \text{and}\\
\mathbb{E}\left\{\sum_{i=1}^{p}\left(T_{OR}^{i,N_{i,1}}-\alpha_1\right)\mathbb{I}\left(T_{OR}^{i,N_{i,1}}\le t_{l,1}\right)\right\} &= 0.
\end{aligned}$$

However, since $\alpha_1 > \alpha_2$ by assumption and $\alpha_1 \le t_{l,1}$ as shown previously (right after (10.1) together with the definition of α_1 and $t_{l,1}$) we must have

$$\begin{aligned}
\mathbb{E}\left\{\sum_{i=1}^{p}(\alpha_1-\alpha_2)\mathbb{I}\left(T_{OR}^{i,N_{i,1}}\le t_{l,1}\right)\right\} &> 0, \qquad \text{and}\\
\mathbb{E}\left\{\sum_{i=1}^{p}\left(T_{OR}^{i,N_{i,2}}-\alpha_1\right)\mathbb{I}\left(t_{l,1}<T_{OR}^{i,N_{i,2}}\le t_{l,2}\right)\right\} &\ge 0.
\end{aligned}$$

This leads to a contradiction. Therefore, we conclude that $Q_{OR}(t_l, t_u)$ is non-decreasing in t_l for a fixed t_u.

Next, we prove that $\tilde{Q}_{OR}(t_l, t_u)$ is non-increasing in t_u for a fixed t_l. By the definition of MDR and similar arguments for the FPR part, we have

$$\mathbb{E}\left[\sum_{i=1}^{p}\left(1-T_{OR}^{i,N_i}\right)\left\{\mathbb{I}\left(T_{OR}^{i,N_i}\geq t_u\right)-\tilde{Q}_{OR}(t_l,t_u)\right\}\right]=0.$$

Since our model has a finite stopping time, naturally we have

$$\left\{i: T_{OR}^{i,N_i}\geq t_u\right\}\cup\left\{j: T_{OR}^{j,N_j}\leq t_l\right\}=\{1,2,3,\cdots,p\}.$$

It follows that

$$\mathbb{E}\left[\sum\nolimits_{\left\{i:T_{OR}^{i,N_i}\geq t_u\right\}}\left(1-T_{OR}^{i,N_i}\right)\left\{1-\tilde{Q}_{OR}(t_l,t_u)\right\}\right]=\mathbb{E}\left\{\sum\nolimits_{\left\{j:T_{OR}^{j,N_j}\leq t_l\right\}}\left(1-T_{OR}^{j,N_j}\right)\tilde{Q}_{OR}(t_l,t_u)\right\}.$$

We have

$$\frac{1-\tilde{Q}_{OR}(t_l,t_u)}{\tilde{Q}_{OR}(t_l,t_u)}=\frac{\mathbb{E}\left\{\sum_{\left\{j:T_{OR}^{j,N_j}\leq t_l\right\}}\left(1-T_{OR}^{j,N_j}\right)\right\}}{\mathbb{E}\left\{\sum_{\left\{i:T_{OR}^{i,N_i}\geq t_u\right\}}\left(1-T_{OR}^{i,N_i}\right)\right\}}. \tag{10.2}$$

Consider two thresholds $t_{u,1} > t_{u,2}$. Denote $N_{i,1}$ and $N_{i,2}$ the corresponding stopping times at location i. The operation of the thresholding procedure implies that $N_{i,1}\geq N_{i,2}$, $\left\{i: T_{OR}^{i,N_{i,1}}\geq t_{u,1}\right\}\subset\left\{i: T_{OR}^{i,N_{i,2}}\geq t_{u,2}\right\}$, and $\left\{j: T_{OR}^{j,N_{j,2}}\leq t_l\right\}\subset\left\{j: T_{OR}^{j,N_{j,1}}\leq t_l\right\}$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left\{\sum\nolimits_{\left\{i:T_{OR}^{i,N_{i,2}}\geq t_{u,2}\right\}}\left(1-T_{OR}^{i,N_{i,2}}\right)\right\}\\
=&\mathbb{E}\left\{\sum\nolimits_{\left\{i:T_{OR}^{i,N_{i,2}}\geq t_{u,1}\right\}}\left(1-T_{OR}^{i,N_{i,2}}\right)\right\}+\mathbb{E}\left\{\sum\nolimits_{\left\{i:t_{u,1}>T_{OR}^{i,N_2}\geq t_{u,2}\right\}}\left(1-T_{OR}^{i,N_{i,2}}\right)\right\}\\
\geq&\mathbb{E}\left\{\sum\nolimits_{\left\{i:T_{OR}^{i,N_{i,1}}\geq t_{u,1}\right\}}\left(1-T_{OR}^{i,N_{i,1}}\right)\right\}.
\end{aligned}$$

We have shown that $\left\{j: T_{OR}^{j,N_{j,2}}\leq t_l\right\}\subset\left\{j: T_{OR}^{j,N_{j,1}}\leq t_l\right\}$. Moreover, on the set $\left\{j: T_{OR}^{j,N_{j,2}}\leq t_l\right\}$, we have $N_{i,1}=N_{i,2}$. It follows that

$$\mathbb{E}\left\{\sum\nolimits_{\left\{j:T_{OR}^{j,N_{j,1}}\leq t_l\right\}}\left(1-T_{OR}^{j,N_{j,1}}\right)\right\}\geq\mathbb{E}\left\{\sum\nolimits_{\left\{j:T_{OR}^{j,N_{j,2}}\leq t_l\right\}}\left(1-T_{OR}^{j,N_{j,2}}\right)\right\}.$$

Combining the above results, we have

$$\frac{\mathbb{E}\left\{\sum_{\left\{j:T_{OR}^{j,N_{j,1}}\leq t_l\right\}}\left(1-T_{OR}^{j,N_{j,1}}\right)\right\}}{\mathbb{E}\left\{\sum_{\left\{i:T_{OR}^{i,N_{i,1}}\geq t_{u,1}\right\}}\left(1-T_{OR}^{i,N_{i,1}}\right)\right\}}\geq\frac{\mathbb{E}\left\{\sum_{\left\{j:T_{OR}^{j,N_{j,2}}\leq t_l\right\}}\left(1-T_{OR}^{j,N_{j,2}}\right)\right\}}{\mathbb{E}\left\{\sum_{\left\{i:T_{OR}^{i,N_{i,2}}\geq t_{u,2}\right\}}\left(1-T_{OR}^{i,N_{i,2}}\right)\right\}}.$$

Hence if $t_{u,1} > t_{u,2}$, then it follows from (10.2) that

$$\frac{1-\tilde{Q}_{OR}(t_l,t_{u,1})}{\tilde{Q}_{OR}(t_l,t_{u,1})} \geq \frac{1-\tilde{Q}_{OR}(t_l,t_{u,2})}{\tilde{Q}_{OR}(t_l,t_{u,2})}.$$

Therefore $\tilde{Q}_{OR}(t_l,t_{u,1}) \leq \tilde{Q}_{OR}(t_l,t_{u,2})$. We conclude that $\tilde{Q}_{OR}(t_l,t_u)$ is non-increasing in t_u for a fixed t_l.

Part (b). The proof is divided into two parts. The first part describes a process that identifies a unique pair of oracle thresholds (t^l_{OR}, t^u_{OR}). The second part shows that $\boldsymbol{d}^\pi(t^l_{OR}, t^u_{OR})$ has the largest power among all eligible procedures.

(1). Oracle thresholds. Let $Q_{OR}(1,1) = \bar{\alpha}$ be the theoretical upper bound corresponding to the FPR when *all* hypotheses are rejected. A prespecified FPR level $\alpha > 0$ is called *eligible* if $\alpha < \bar{\alpha}$. Let $\mathcal{R}_\alpha = \{t_u : Q_{OR}(t_u,t_u) > \alpha\}$. We can see that $\mathcal{R}_\alpha$ is nonempty if α is eligible, since $Q_{OR}(0,0) = 0$ and $Q_{OR}(1,1) = \bar{\alpha}$. Consider $t_u \in \mathcal{R}_\alpha$. Note that $Q_{OR}(0,t_u) = 0$ for all t_u, the following threshold is well defined:

$$t^l_{OR}(t_u) = \sup\{t_l : Q_{OR}(t_l,t_u) \leq \alpha\}. \tag{10.3}$$

We claim that $Q\{t^l_{OR}(t_u), t_u\} = \alpha$.

We prove this by contradiction. We first note $Q_{OR}(t_l,t_u)$ is continuous in t_l and t_u. To see that, note X_{ij} follows a continuous distribution and $T^{i,j}_{OR}$ is a continuous function of X_{ij}. Therefore, the distribution of $T^{i,j}_{OR}$ is also continuous. By definition $\delta_i = \mathbb{I}(T^{i,j}_{OR} \leq t_l)$. Hence $\mathbb{E}(\delta_i) = \mathbb{P}(T^{i,j}_{OR} \leq t_l)$ is a continuous function of t_l. Next, the distribution of $T^{i,j}_{OR}\delta_i$ is proportional to the distribution of $T^{i,j}_{OR}$ truncated at t_l. Since $T^{i,j}_{OR}$ follows a continuous distribution, it follows that $\mathbb{E}(T^{i,j}_{OR}\delta_i)$ is also continuous in t_l. Continuity of $Q_{OR}(t_l,t_u)$ in t_u can be shown similarly. Now, according to the continuity of $Q_{OR}(t_l,t_u)$, for every $t_u \in R_\alpha$, we can find $t^*_l(t_u)$ such that $Q\left(t^*_l(t_u), t_u\right) = \alpha$ [since $Q_{OR}(0,t_u) = 0$ and $Q_{OR}(t_u,t_u) > \alpha$]. If not, the equality does not hold, i.e. we have

$$Q\{t^l_{OR}(t_u), t_u\} < \alpha,$$

then the monotonicity of $Q_{OR}(t_l,t_u)$ implies that $t^*_l(t_u) > t^l_{OR}(t_u)$, which contradicts the definition of $t^l_{OR}(t_u)$. The above construction shows that, for every $t_u \in \mathcal{R}_\alpha$, we can always identify a unique $t^l_{OR}(t_u)$ such that $Q_{OR}\left\{t^l_{OR}(t_u), t_u\right\} = \alpha$.

We say (α,γ) constitute an *eligible pair* of prespecified error rates if α is eligible, and for this α, γ satisfies

$$0 < \gamma < \sup\left\{\tilde{Q}_{OR}\left(t^l_{OR}(t_u), t_u\right) : t_u \in R_\alpha\right\}.$$

In the above definition, the eligibility of (α,γ) only depends on the model, but not any given t_u. Now consider an eligible pair (α,γ). The continuity of $\tilde{Q}_u(t_u) \equiv \tilde{Q}\{t^l_{OR}(t_u), t_u\}$ implies that we can find t^*_u such that $\tilde{Q}\left\{t^l_{OR}(t^*_u), t^*_u\right\} = \gamma$. Let

$$t^u_{OR} = \inf\{t_u \in R_\alpha : \tilde{Q}_{OR}(t^l_{OR}(t_u), t_u) = \gamma\}.$$

The pair of oracle thresholds are thus given by

$$(t_{OR}^l, t_{OR}^u) \equiv \{t_{OR}^l(t_{OR}^u), t_{OR}^u\}.$$

(2). Proof of optimality. Denote $\boldsymbol{d}_* = (\boldsymbol{N}_*, \boldsymbol{\delta}_*)$ a sequential procedure that satisfies $\mathrm{FPR}(\boldsymbol{d}_*) = \alpha_* \leq \alpha$, $\mathrm{MDR}(\boldsymbol{d}_*) = \gamma_* \leq \gamma$, where $\boldsymbol{N}_* = (N_*^1, \cdots, N_*^p)$ and $\boldsymbol{\delta}_* = (\delta_*^1, \cdots, \delta_*^p)$ are the corresponding stopping times and decision rules. Denote $\mathrm{ESS}(\boldsymbol{d}_*)$ the expected average stopping times. By definition, we have

$$\mathbb{E}\left\{\sum_{i=1}^{p}(T_{OR}^{i,N_*^i} - \alpha_*)\delta_*^i\right\} = 0, \quad \mathbb{E}\left\{\sum_{i=1}^{p}(1 - T_{OR}^{i,N_*^i})(1 - \delta_*^i - \gamma_*)\right\} = 0.$$

Next we present a hypothetical scenario to justify the operational characteristics of all sequential rules and a mechanism aimed at enhancing any suboptimal rule that does not adhere to the desired order. Specifically, we demonstrate that if decisions on any two coordinates are inconsistent with the specified order, those decisions can be uniformly improved by swapping them.

Suppose we sort T_{OR}^{i,N_*^i} as

$$T_{OR}^{(1),N_*^{(1)}} \leq T_{OR}^{(2),N_*^{(2)}} \leq \cdots \leq T_{OR}^{(p),N_*^{(p)}}$$

with their corresponding decisions $\delta_*^{(1)}, \delta_*^{(2)}, \cdots, \delta_*^{(p)}$. If $\boldsymbol{\delta}_*$ does not take the following form of decision rule

$$\text{there exists a } k, \text{ such that } \delta_*^{(i)} = \begin{cases} 1 & i \leq k \\ 0 & k < i \leq p \end{cases}, \tag{10.4}$$

then we can always modify $\boldsymbol{\delta_*}$ into such a form with the same ESS and smaller FPR and MDR. Specifically, suppose that there exists $l_1 < l_2$ such that $\delta_*^{(l_1)} = 0$ and $\delta_*^{(l_2)} = 1$, then we swap these two decisions. Such operation can be iterated until the decision rule takes the form as (10.4). Denote the new decision rule by $\boldsymbol{d}'_* = (\boldsymbol{N}_*, \boldsymbol{\delta}'_*)$. Since $T_{OR}^{(l_1),N_*^{(l_1)}} \leq T_{OR}^{(l_2),N_*^{(l_2)}}$ in each swapping, we can reduce the FPR and MDR:

$$\begin{aligned}
\alpha'_* &= \frac{\sum_{i=1}^{p}\left(T_{OR}^{i,N_*^i}\delta_*'^i\right)}{\mathbb{E}\left(\sum_{i=1}^{p}\delta_*'^i\right)} \leq \frac{\sum_{i=1}^{p}\left(T_{OR}^{i,N_*^i}\delta_*^i\right)}{\mathbb{E}\left(\sum_{i=1}^{p}\delta_*^i\right)} = \alpha_*, \\
\gamma'_* &= \frac{\sum_{i=1}^{p}\left\{(1 - T_{OR}^{i,N_*^i})(1 - \delta_*'^i)\right\}}{p\pi} \\
&\leq \frac{\sum_{i=1}^{p}\left\{(1 - T_{OR}^{i,N_*^i})(1 - \delta_*^i)\right\}}{p\pi} = \gamma_*.
\end{aligned}$$

Expressing $\boldsymbol{d}'_*$ in the form of (2.3), we can find t'_l and t'_u such that

$$\delta_{i,N_i} = \begin{cases} 1 & T_{OR}^{i,N_i} \leq t'_l \\ 0 & T_{OR}^{i,N_i} \geq t'_u \end{cases},$$

where $Q_{OR}(t_l', t_u') = \alpha_*'$, $\tilde{Q}_{OR}(t_l', t_u') = \gamma_*'$.

Now we claim that $t_l' \le t_{OR}^l(t_{OR}^u)$ and $t_u' \ge t_{OR}^u$. We prove this by contradiction. First, if we have $t_{OR}^u > t_u'$, then by the definition of t_{OR}^u, we have

$$\tilde{Q}_{OR}\left(t_{OR}^l(t_u'), t_u'\right) > \gamma,\ Q_{OR}\left(t_{OR}^l(t_u'), t_u'\right) = \alpha.$$

However, we also have $\tilde{Q}_{OR}\left(t_l', t_u'\right) = \gamma_*' \le \gamma$. By the definition of $\tilde{Q}_{OR}(t_l, t_u)$, with the same t_u, only a larger t_l could result in a strictly smaller MDR level. Together with the monotonicity of $\tilde{Q}_{OR}(t_l, t_u)$ for a fixed t_u, we claim that $t_{OR}^l(t_u') < t_l'$. Since $Q_{OR}(t_l, t_u)$ is non-decreasing in t_l for a fixed t_u as shown in part (a) of the theorem, we have $Q\left(t_{OR}^l(t_u'), t_u'\right) = \alpha$. It follows that $Q_{OR}\left(t_l', t_u'\right) > \alpha$, contradicting the fact that $Q_{OR}(t_l', t_u') = \alpha_*' \le \alpha$. Therefore we must have $t_u' \ge t_{OR}^u$.

Next, assume that $t_l' > t_{OR}^l(t_{OR}^u)$. Then by the definition of $t_{OR}^l(t_u)$ and monotonicity of $Q_{OR}(t_l, t_u)$, we have $Q\left(t_l', t_{OR}^u\right) > \alpha$. Given the fact that $Q_{OR}(t_l', t_{OR}^u) \le t_l'$, we must have $\alpha < t_l'$, claiming that t_l' is always bounded below by α regardless of α_*', which leads to a contradiction since we always have $Q_{OR}(t_l', t_u') = \alpha_*' < t_l'$. Hence we must have $t_l' \le t_{OR}^l(t_{OR}^u)$. Therefore

$$\mathrm{ESS}(\boldsymbol{d}_*') = \mathrm{ESS}(\boldsymbol{d}_*) \ge \mathrm{ESS}(\boldsymbol{d}_{OR})$$

and the desired result follows.

■

10.2 Proof of Proposition 4

Proof . The goal is to show that the pair $t_{OR}^l = \alpha$ and $t_{OR}^u = \frac{1-\pi}{\pi\gamma+1-\pi}$ control the FPR and MDR. The FPR part is straightforward since according to the definition of T_{OR}^{i,N_i}, we have

$$\mathrm{FPR}(\tilde{\boldsymbol{d}}_{OR}) = \frac{\mathbb{E}\left\{\sum_{i=1}^p T_{OR}^{i,N_i}\mathbb{I}(T_{OR}^{i,N_i} \le \alpha)\right\}}{\mathbb{E}\left\{\sum_{i=1}^p \mathbb{I}(T_{OR}^{i,N_i} \le \alpha)\right\}} \le \frac{\mathbb{E}\left\{\sum_{i=1}^p \alpha \cdot \mathbb{I}(T_{OR}^{i,N_i} \le \alpha)\right\}}{\mathbb{E}\left\{\sum_{i=1}^p \mathbb{I}(T_{OR}^{i,N_i} \le \alpha)\right\}} = \alpha.$$

Similarly for the FDR control, we have

$$\mathrm{FDR}(\tilde{\boldsymbol{d}}_{OR}) = \mathbb{E}\left\{\frac{\sum_{i=1}^p T_{OR}^{i,N_i}\mathbb{I}(T_{OR}^{i,N_i} \le \alpha)}{\sum_{i=1}^p \mathbb{I}(T_{OR}^{i,N_i} \le \alpha) \vee 1}\right\} \le \frac{\mathbb{E}\left\{\sum_{i=1}^p \alpha \cdot \mathbb{I}(T_{OR}^{i,N_i} \le \alpha)\right\}}{\mathbb{E}\left\{\sum_{i=1}^p \mathbb{I}(T_{OR}^{i,N_i} \le \alpha) \vee 1\right\}} \le \alpha.$$

Remark 16 *From the proof we can see that the choice of $t_{OR}^l = \alpha$, derived based on Wald's approximation, can be conservative in practice.*

To show the MDR part, we first carry out an analysis of the false negative rate (FNR), which is defined as

$$\mathrm{FNR}(\tilde{\boldsymbol{d}}_{OR}) = \frac{\mathbb{E}\left\{\sum_{i=1}^p \theta_i(1-\delta_i)\right\}}{\mathbb{E}\left\{\sum_{i=1}^p (1-\delta_i)\right\}}. \tag{10.5}$$

According to the operation of $\tilde{\boldsymbol{d}}_{OR}$, the FNR can be further calculated as

$$\begin{aligned} \mathrm{FNR}(\tilde{\boldsymbol{d}}_{OR}) &= \frac{\mathbb{E}\left\{\sum_{i=1}^{p}(1-T_{OR}^{i,N_i})\mathbb{I}(T_{OR}^{i,N_i}\geq t_u)\right\}}{\mathbb{E}\left\{\sum_{i=1}^{p}\mathbb{I}(T_{OR}^{i,N_i}\geq t_u)\right\}} \\ &\leq 1-t_u=\frac{\pi\gamma}{\pi\gamma+1-\pi}. \end{aligned} \tag{10.6}$$

Denote $\tilde{\boldsymbol{d}}_{OR}^{i}=(\tilde{N}_{OR}^{i},\tilde{\delta}_{OR}^{i})$. We have shown that $\mathrm{FPR}(\tilde{\boldsymbol{d}}_{OR})\leq\alpha$. Suppose the actual FPR level is $\tilde{\alpha}\leq\alpha$. Then $\mathbb{E}\left\{\sum_{i=1}^{p}(1-\theta_i)\tilde{\delta}_{OR}^{i}\right\}=\tilde{\alpha}(\sum_{i=1}^{p}\tilde{\delta}_{OR}^{i})$. It follows that

$$(1-\tilde{\alpha})\mathbb{E}\left(\sum_{i=1}^{p}\tilde{\delta}_{OR}^{i}\right)=\mathbb{E}\left\{\sum_{i=1}^{p}\theta_i\tilde{\delta}_{OR}^{i}\right\}. \tag{10.7}$$

Meanwhile, our analysis of the FNR ((10.5) and (10.6)) shows that

$$\mathbb{E}\left\{\sum_{i=1}^{p}\theta_i(1-\tilde{\delta}_{OR}^{i})\right\}\leq\frac{\pi\gamma}{\pi\gamma+1-\pi}\cdot\mathbb{E}\left\{\sum_{i=1}^{p}(1-\tilde{\delta}_{OR}^{i})\right\}. \tag{10.8}$$

Combining (10.7) and (10.8), we obtain

$$(\pi\gamma+1-\pi)\left\{p\pi-(1-\tilde{\alpha})\mathbb{E}\left(\sum_{i=1}^{p}\tilde{\delta}_{OR}^{i}\right)\right\}\leq\pi\gamma\left\{p-\mathbb{E}\left(\sum_{i=1}^{p}\tilde{\delta}_{OR}^{i}\right)\right\}.$$

It follows that

$$\mathbb{E}\left(\sum_{i=1}^{p}\tilde{\delta}_{OR}^{i}\right)\geq\frac{p\pi(1-\pi)(1-\gamma)}{-\pi\gamma\tilde{\alpha}+(1-\pi)(1-\tilde{\alpha})}\geq\frac{p\pi(1-\gamma)}{(1-\tilde{\alpha})}$$

Using (10.7) and $p\pi=\mathbb{E}(\sum_i\theta_i)$, we have

$$\mathbb{E}(\textstyle\sum_{i=1}^{p}\theta_i\tilde{\delta}_{OR}^{i})\geq\mathbb{E}(\sum_{i=1}^{p}\theta_i)-\gamma\mathbb{E}(\sum_{i=1}^{p}\theta_i).$$

Therefore $\mathbb{E}\{\sum_{i=1}^{p}\theta_i(1-\tilde{\delta}_{OR}^{i})\}\leq\gamma\mathbb{E}(\sum_{i=1}^{p}\theta_i)$ and the desired result follows. ■

10.3 Proof of Theorem 6

Proof . **Part (i). The FPR and FDR control.** To show the validity of SMART for FPR control, we use the idea in Efron (2008); Cai and Sun (2009) for group-wise testing. Define stage-wise false positive rate sFPR_j as

$$\mathrm{sFPR}_j:=\frac{\mathbb{E}\left\{\sum_{i\in\mathcal{S}_j}(1-\theta_i)\delta_i\right\}}{\mathbb{E}\sum_{i\in\mathcal{S}_j}\delta_i},$$

where sFPR_j is the ratio of the expected number of false rejections at stage j over the expected number of all rejections at stage j. The first step is to show that SMART controls all stage-wise FDRs at level α. The second step is to show that the global FDR is controlled

at level α by combining hypotheses rejected from all stages. By our definition of sFPR_j, we have

$$\text{sFPR}_j = \frac{\mathbb{E}\left(\sum_{i=1}^{k_j^s} T_{OR}^{(i),j}\right)}{\mathbb{E}\left(k_j^s\right)} \leq \frac{\mathbb{E}\left(k_j^s \alpha\right)}{\mathbb{E}\left(k_j^s\right)} = \alpha.$$

Therefore SMART controls the sFPR at level α across all stages.

Next we show that if sFPR_j is controlled universally at pre-specified levels across all stages, then the global FPR will be controlled at the same level. Let N_{SM} denote the total number of stages that $\boldsymbol{d}_{SM}$ has. It follows that

$$\begin{aligned}
\text{FPR}(\boldsymbol{d}_{SM}) &= \frac{\mathbb{E}\left\{\sum_{j=1}^{N_{SM}} \sum_{i\in\mathcal{S}_j} (1-\theta_i)\, \delta_{SM}^i\right\}}{\mathbb{E}\left(\sum_{j=1}^{N_{SM}} \sum_{i\in\mathcal{S}_j} \delta_{SM}^i\right)} \\
&\leq \frac{\mathbb{E}\left(\sum_{j=1}^{N_{SM}} \alpha \sum_{i\in\mathcal{S}_j} \delta_{SM}^i\right)}{\mathbb{E}\left(\sum_{j=1}^{N_{SM}} \sum_{i\in\mathcal{S}_j} \delta_{SM}^i\right)} = \alpha.
\end{aligned}$$

Consider the quantity $\mathbb{E}\left\{\sum_{j=1}^{N_{SM}} \sum_{i\in\mathcal{S}_j} (1-\theta_i)\, \delta_{SM}^i\right\}$. Note that $i \in \mathcal{S}_j$ indicates that (i) a total of j data points $\boldsymbol{x}_i^j = (x_i^1, \cdots, x_i^j)$ are eventually collected for the ith unit and (ii) a decision for unit i is made at stage j. It follows that $\boldsymbol{\delta}_{SM}^i$ only depends on $\boldsymbol{x}_i^j$ and can be factored out:

$$\begin{aligned}
\mathbb{E}\left\{\sum_{j=1}^{N_{SM}} \sum_{i\in\mathcal{S}_j} (1-\theta_i)\, \delta_{SM}^i\right\} &= \mathbb{E}\left\{\sum_{j=1}^{N_{SM}} \sum_{i\in\mathcal{S}_j} \mathbb{E}\left[(1-\theta_i)\, \delta_{SM}^i | \boldsymbol{x}_i^j\right]\right\} \\
&= \mathbb{E}\left\{\sum_{j=1}^{N_{SM}} \sum_{i\in\mathcal{S}_j} \delta_{SM}^i \mathbb{E}\left[(1-\theta_i)\, | \boldsymbol{x}_i^j\right]\right\}
\end{aligned}$$

According to the definition of the oracle statistic, We have

$$\mathbb{E}\left\{\sum_{j=1}^{N_{SM}} \sum_{i\in\mathcal{S}_j} (1-\theta_i)\, \delta_{SM}^i\right\} = \mathbb{E}\left\{\sum_{j=1}^{N_{SM}} \sum_{i\in\mathcal{S}_j} \delta_{SM}^i T_{OR}^{i,j}\right\} \leq \mathbb{E}\left\{\sum_{j=1}^{N_{SM}} \alpha \sum_{i\in\mathcal{S}_j} \delta_{SM}^i\right\}.$$

The last inequality is due to the operation of SMART, which ensures that at every stage j, we always have $\sum_{i\in\mathcal{S}_j} T_{OR}^{i,j} \delta_{SM}^i \leq \alpha \sum_{i\in\mathcal{S}_j} \delta_{SM}^i$. Therefore we have

$$\text{FPR}(\boldsymbol{d}_{SM}) \leq \frac{\mathbb{E}\left(\sum_{j=1}^{N_{SM}} \alpha \sum_{i\in\mathcal{S}_j} \delta_{SM}^i\right)}{\mathbb{E}\left(\sum_{j=1}^{N_{SM}} \sum_{i\in\mathcal{S}_j} \delta_{SM}^i\right)} = \alpha.$$

Similarly, we can prove the global FDR control by jointly evaluating the performance SMART at separate stages

$$\begin{aligned}\mathrm{FDR}(\boldsymbol{d}_{SM}) &= \mathbb{E}\left\{\frac{\sum_{j=1}^{N_{SM}}\sum_{i\in\mathcal{S}_j}(1-\theta_i)\,\delta_{SM}^i}{(\sum_{j=1}^{N_{SM}}\sum_{i\in\mathcal{S}_j}\delta_{SM}^i)\vee 1}\right\}\\ &= \mathbb{E}\left[\frac{\sum_{j=1}^{N_{SM}}\sum_{i\in\mathcal{S}_j}T_{OR}^{i,j}\delta_{SM}^i}{\left(\sum_{j=1}^{N_{SM}}\sum_{i\in\mathcal{S}_j}\delta_{SM}^i\right)\vee 1}\right]\leq\alpha\end{aligned}$$

The last inequality is due to the operation of SMART.

Part (ii). MDR control. Unlike the FDR analysis, stage-wise MDR control does not imply global MDR control. We introduce an intermediate quantity, the false non-discovery rate (FNR) and divide the proof into two steps: the first step shows that stage-wise FNR control implies global FNR control; the second step establishes the relationship between the global MDR and global FNR.

Define stage-wise false non-discovery rate sFNR_j as

$$\mathrm{sFNR}_j := \frac{\mathbb{E}\left\{\sum_{i\in\mathcal{S}_j}\theta_i(1-\delta_i)\right\}}{\mathbb{E}\sum_{i\in\mathcal{S}_j}(1-\delta_i)},$$

where sFNR_j is the ratio of the expected number of false non-discoveries over the expected number of all non-discoveries. It follows that

$$\mathrm{sFNR}_j = \frac{\mathbb{E}\left\{\sum_{i=0}^{k_j^e-1}\left(1-T_{OR}^{k_j-i,j}\right)\right\}}{\mathbb{E}\left(k_j^e\right)} \leq \frac{\mathbb{E}\left(k_j^e\frac{\pi\gamma}{\pi\gamma+1-\pi}\right)}{\mathbb{E}\left(k_j^e\right)} = \frac{\pi\gamma}{\pi\gamma+1-\pi}.$$

Therefore SMART controls the sFNR at level $\frac{\pi\gamma}{\pi\gamma+1-\pi}$ across all stages. Next, we show that if sFNR_j is controlled universally at pre-specified levels across all stages, then the global FNR will be controlled at the same level.

$$\begin{aligned}\mathrm{FNR} &= \frac{\mathbb{E}\left\{\sum_{j=1}^{N}\sum_{i\in\mathcal{S}_j}\theta_i\left(1-\delta_{SM}^i\right)\right\}}{\mathbb{E}\left\{\sum_{j=1}^{N}\sum_{i\in\mathcal{S}_j}\left(1-\delta_{SM}^i\right)\right\}}\\ &\leq \frac{\mathbb{E}\left\{\sum_{j=1}^{N}\frac{\pi\gamma}{\pi\gamma+1-\pi}\sum_{i\in\mathcal{S}_j}\left(1-\delta_{SM}^i\right)\right\}}{\mathbb{E}\left\{\sum_{j=1}^{N}\sum_{i\in\mathcal{S}_j}\left(1-\delta_{SM}^i\right)\right\}}\\ &= \frac{\pi\gamma}{\pi\gamma+1-\pi}.\end{aligned}$$

Finally, according to the proof of Theorem 2, the MDR and FNR satisfy the following relationship

$$\mathrm{MDR}\leq\gamma \text{ if } \mathrm{FNR}\leq\tfrac{\pi\gamma}{\pi\gamma+1-\pi}.$$

We conclude that the MDR is controlled at level γ, completing the proof.

■

10.4 Proof of Theorem 12

Proof For a symmetric decision procedure $\boldsymbol{d}$, denote its Type I and Type II errors on unit i by $\alpha' = \mathbb{P}_{H_{i,0}}(\text{Reject } H_{i,0})$ and $\gamma' = \mathbb{P}_{H_{i,1}}(\text{Accept } H_{i,0})$. It can be shown that the corresponding global error rates are given by

$$\text{FPR}(\boldsymbol{d}) = \frac{(1-\pi)\alpha'}{(1-\pi)\alpha' + \pi(1-\gamma')} \quad \text{and} \quad \text{MDR}(\boldsymbol{d}) = \gamma', \tag{10.9}$$

respectively. Our result largely follows from the lower bound derived in Malloy and Nowak (2014a) on family-wise error rate (FWER); we only highlight the main steps on how to go from the FWER paradigm to the FPR/MDR paradigm, which essentially involves exploiting the relationship (10.9). From Thm. 2.39 in Siegmund (1985), we have

$$\begin{aligned}
\tau_1 &\geq \frac{\alpha' \log(\frac{\alpha'}{1-\gamma'}) + (1-\alpha')\log(\frac{1-\alpha'}{\gamma'})}{D(\mathbb{F}_0||\mathbb{F}_1)} \quad \text{and} \\
\tau_2 &\geq \frac{(1-\gamma')\log(\frac{(1-\gamma'}{\alpha'}) + \gamma' \log(\frac{\gamma'}{1-\alpha'})}{D(\mathbb{F}_1||\mathbb{F}_0)},
\end{aligned}$$

where τ_1 and τ_2 are the expected stopping times for null and non-null locations, respectively. Furthermore, from Malloy and Nowak (2014a), we have

$$\tau_1 \geq \frac{(1-\alpha')\log(\gamma')^{-1} - \log 2}{D(\mathbb{F}_0||\mathbb{F}_1)},\ \tau_2 \geq \frac{(1-\gamma')\log(\alpha')^{-1} - \log 2}{D(\mathbb{F}_1||\mathbb{F}_0)}.$$

Using the definition of KL divergence $D_{KL}(F_0, F_1) = \max\{D(\mathbb{F}_0|\mathbb{F}_1), D(\mathbb{F}_1|\mathbb{F}_0)\}$, the average stopping time of all locations satisfies

$$\begin{aligned}
\tau &= \frac{(p-p\pi)\tau_1 + p\pi\tau_2}{p} \\
&\geq \frac{(1-\pi)(1-\alpha')\log(\gamma')^{-1} + \pi(1-\gamma')\log(\alpha')^{-1} - \log 2}{D_{KL}(F_0, F_1)}.
\end{aligned} \tag{10.10}$$

We consider two situations. If $\alpha' \leq \gamma'$, then

$$\tau \geq (1-\pi)(1-\gamma')\log(\gamma')^{-1} + \pi(1-\gamma')\log(\gamma')^{-1} - \log 2.$$

Note that for $0 \leq x \leq 1$, $\{x \log x : x \in (0,1)\}$ reaches its minimum when $x = e^{-1}$. It follows that $2^{-1} < e^{-1/e} \leq \gamma'^{\gamma'} \leq 1$. Therefore

$$(1-\gamma')\log(\gamma')^{-1} \geq \log(2\gamma')^{-1}.$$

Together with (10.10), we have $\tau \geq \frac{\log(4\gamma')^{-1}}{D_{KL}(F_0,F_1)}$. According to our constraint on τ, we conclude that

$$\log(4\eta)^{-1} \geq \tau D_{KL}(F_0, F_1) \geq \log(4\gamma')^{-1}.$$

Therefore, $\gamma' \geq \eta$ and $R^*(\boldsymbol{d}) \geq \eta$.

If $\gamma' \leq \alpha'$, then we can similarly show that

$$\tau \geq \frac{(1-\alpha')\log \alpha'^{-1} - \log 2}{D_{KL}(F_0, F_1)} \geq \frac{\log \frac{1}{2\alpha'} - \log 2}{D_{KL}(F_0, F_1)} \geq \frac{\log \frac{1}{4\alpha'}}{D_{KL}(F_0, F_1)}.$$

It follows that

$$\log\left(\frac{1}{4\eta}\right) \geq \tau D_{KL}(F_0, F_1) \geq \log\left(\frac{1}{4\alpha'}\right),$$

which implies $\alpha' \geq \eta$. Under the assumption that $\pi < \frac{1}{3}$ and $\eta \leq \frac{1}{2}$, we have $\eta \leq \frac{1}{2} \leq \frac{1-2\pi}{1-\pi}$. Consider the function $k(x) = \frac{(1-\pi)x}{(1-\pi)x+\pi}$. It is easy to see that $k(x)$ is monotonically increasing in x. Hence

$$\begin{aligned} R^*(\boldsymbol{d}) &\geq \frac{(1-\pi)\alpha'}{(1-\pi)\alpha' + \pi(1-\gamma')} \geq \frac{(1-\pi)\alpha'}{(1-\pi)\alpha' + \pi} \\ &\geq \frac{(1-\pi)\eta}{(1-\pi)\eta + \pi} \geq \frac{(1-\pi)\eta}{(1-\pi)\frac{1-2\pi}{1-\pi} + \pi} = \eta, \end{aligned}$$

completing the proof. ■

10.5 Proof of Theorem 14

Proof In Theorem 6 we have already shown that when $t_l = \alpha$ and $t_u = \frac{1-\pi}{\pi\gamma+1-\pi}$, the SMART procedure controls the FDR and MDR at level α and γ, respectively. Let

$$\alpha = \gamma = \frac{1}{\xi(p)^{1+\epsilon}}.$$

With the choice of t_l and t_u mentioned above, we have

$$\lim_{p\to\infty} R^*(\boldsymbol{d}) = \lim_{p\to\infty} \frac{2}{\xi(p)^{1+\epsilon}} = 0,$$

which proves the first part of the theorem.

Next we establish the upper bound. Consider p simultaneous SPRTs with the same threshold t_l and t_u. The operation of our SMART procedure uses these thresholds for the moving averages; hence the SPRT approach with the same t_l and t_u will always take more samples. It is sufficient to show that the result holds for simultaneous SPRTs.

We first use the relationship (A.1) to convert thresholds t_l and t_u to the thresholds for SPRTs:

$$A = \frac{(1-\pi)(1-t_u)}{\pi t_u,}, \quad B = \frac{(1-\pi)(1-t_l)}{\pi t_l}.$$

Under our specifications, we further have

$$A = \frac{\gamma(1-\pi)}{\pi\gamma + 1 - \pi}, \quad B = \frac{(1-\pi)(1-\alpha)}{\pi\alpha}.$$

From Siegmund (1985) (Equations 9-10, P10), we have

$$\begin{aligned}\alpha' &\leq B^{-1}(1-\gamma') \leq B^{-1},\\ \gamma' &\leq A(1-\alpha') \leq A.\end{aligned}$$

According to Assumption (5.1), we have

$$\begin{aligned}\mathbb{E}(\log \mathcal{L}_{i,N_i} | \log \mathcal{L}_{i,N_i} < \log A) &\geq \log A - C_1,\\ \mathbb{E}(\log \mathcal{L}_{i,N_i} | \log \mathcal{L}_{i,N_i} > \log B) &\leq \log B + C_2\end{aligned}$$

for some positive constants C_1 and C_2. Consider $\mathbb{E}_{\theta_i=0}(\log \mathcal{L}_{i,N_i})$. Then

$$\begin{aligned}& -\mathbb{E}_{\theta_i=0}(\log \mathcal{L}_{i,N_i})\\ =\ & -(1-\alpha)\mathbb{E}_{\theta_i=0}(\log \mathcal{L}_{i,N_i} | \log \mathcal{L}_{i,N_i} < \log A) - \alpha \mathbb{E}_{\theta_i=0}(\log \mathcal{L}_{i,N_i} | \log \mathcal{L}_{i,N_i} > \log B)\\ \leq\ & (1-\alpha)(\log A^{-1} + C_1) - \alpha \log B\\ \leq\ & (1-\alpha)(\log A^{-1} + C_1) \leq \log A^{-1} + C_1.\end{aligned}$$

Likewise, we can show that

$$\mathbb{E}_{\theta_i=1}(\log \mathcal{L}_{i,N_i}) \leq \log B + C_2.$$

Let $C_3 = (1-\pi)C_1 + \pi C_2$. According to Wald's identity (P490, Berger (1985)), we have

$$\begin{aligned}\mathbb{E}_{\theta_i=1}(N_i) &= \frac{\mathbb{E}_{\theta_i=1}(\log \mathcal{L}_{i,N_i})}{\mathbb{E}_{\theta_i=1}(\log \mathcal{L}_{i,1})} = \frac{\mathbb{E}_{\theta_i=1}(\log \mathcal{L}_{i,N_i})}{D(F_1|F_0)},\\ \mathbb{E}_{\theta_i=0}(N_i) &= \frac{\mathbb{E}_{\theta_i=0}(\log \mathcal{L}_{i,N_i})}{\mathbb{E}_{\theta_i=0}(\log \mathcal{L}_{i,1})} = \frac{-\mathbb{E}_{\theta_i=0}(\log \mathcal{L}_{i,N_i})}{D(F_0|F_1)}.\end{aligned}$$

It follows that

$$\begin{aligned}\limsup_{p\to\infty} \frac{\tau}{\log \xi(p)} &= \limsup_{p\to\infty} \frac{(1-\pi)\mathbb{E}_{\theta_i=0}(N_i) + \pi \mathbb{E}_{\theta_i=1}(N_i)}{\log \xi(p)}\\ &\leq \limsup_{p\to\infty} \frac{(1-\pi)(\log A^{-1} + C_1)}{\log \xi(p) D(F_0|F_1)} + \limsup_{p\to\infty} \frac{\pi(\log B + C_2)}{\log \xi(p) D(F_1|F_0)}\\ &= \limsup_{p\to\infty} \frac{(1-\pi)\log \frac{\pi\gamma+1-\pi}{\gamma(1-\pi)} + \pi \log \frac{(1-\pi)(1-\alpha)}{\pi\alpha} + C_3}{\log \xi(p) \min\{(D(F_0|F_1), D(F_1|F_0)\}} \qquad (10.11)\\ &= \limsup_{p\to\infty} \frac{(1-\pi)\log(\gamma^{-1}) + \pi \log \frac{(1-\pi)}{\pi\alpha} + C_3}{\log \xi(p) \min\{(D(F_0|F_1), D(F_1|F_0)\}} \qquad (10.12)\\ &= \limsup_{p\to\infty} \frac{\log(\alpha^{-1})}{\log \xi(p) \min\{(D(F_0|F_1), D(F_1|F_0)\}} \qquad (10.13)\\ &= \frac{1+\epsilon}{\min\{(D(F_0|F_1), D(F_1|F_0)\}}.\end{aligned}$$

From (10.11) to (10.12), we have used the fact α and γ are error rates converging to zero; hence

$$\begin{aligned}\frac{\pi\gamma+1-\pi}{\gamma(1-\pi)} &= \frac{1}{\gamma}\{1+o(1)\}\\ \frac{(1-\pi)(1-\alpha)}{\pi\alpha} &= \frac{(1-\pi)}{\pi\alpha}\{1+o(1)\}.\end{aligned}$$

Equation (10.13) uses the fact that $\alpha = \gamma$. The desired result follows. ■

10.6 Proof of Proposition 7

Let $\phi(\cdot, \mu, \sigma^2)$ be the density function of $N(\mu, \sigma^2)$. Observe that

$$\frac{\prod_{t=1}^{j} \phi(x_{it}, \eta_k, \hat{\sigma}^2)}{\prod_{t=1}^{j} \phi(x_{it}, 0, \hat{\sigma}^2)} = \exp\left\{ \frac{1}{2\hat{\sigma}^2} \left[j\eta_k (2\bar{x}_i^j - \eta_k) \right] \right\},$$

where $\bar{x}_i^j = \frac{1}{j} \sum_{t=1}^{j} x_{it}$. We first consider the case $x_{it} \sim N(\mu_i, \sigma^2)$, Howard et al. (2021) proves the following result:

$$\mathbb{P}\left(\forall\, j : |\bar{x}_i^j - \mu_i| < V(j, \alpha) \right) > 1 - \alpha, \tag{10.14}$$

where

$$V(j, \alpha) = 1.7 \sqrt{\sigma \frac{\log\log(2j) + 0.72 \log(5.2/\alpha)}{j}}.$$

Hence, if $|\eta_k - \mu_i| < |\mu_i|/2 - \epsilon$ and $\eta_k \mu_i > 0$ then $\exists K$ depending on α such that

$$\mathbb{P}\left(\forall\, j > K : \frac{\prod_{t=1}^{j} \phi(x_{it}, \eta_k, \hat{\sigma}^2)}{\prod_{t=1}^{j} \phi(x_{it}, 0, \hat{\sigma}^2)} > \exp(c_\alpha j) \right) > 1 - \alpha, \tag{10.15}$$

for some c_α depends on $\sigma,\ \hat{\sigma},\ \alpha$. Recall $g(\cdot) = \sum_{s=1}^{m_1} w_s \delta_{\eta_s}(\cdot)$, we have

$$T_{OR}^{i,j} := \mathbb{P}(\theta_i = 0 | \boldsymbol{X}_i^j) = \frac{(1-\pi) \prod_{t=1}^{j} \phi(x_{it}, 0, \sigma^2)}{(1-\pi) \prod_{t=1}^{j} \phi(x_{it}, 0, \sigma^2) + \pi \sum_{s=1}^{m_1} w_s \prod_{t=1}^{j} \phi(x_{it}, \eta_s, \sigma^2)}.$$

Since $w_s > \epsilon$ for some ϵ, and $\mu_i = \eta_s$ for some η_s, by (10.14) we have

$$\mathbb{P}\left(\forall\, j > K : T_{OR}^{i,j} < \exp(-c_\alpha j) \right) > 1 - \alpha, \tag{10.16}$$

Let $\hat{g}(\cdot) = \sum_{s=1}^{m_2} \hat{w}_s \delta_{\hat{\eta}_s}(\cdot)$. Note that $\hat{\mu}_i^j$ can be written as

$$\hat{\mu}_i^j = \sum_{s=1}^{m} p_s^j \hat{\eta}_s, \quad \text{where } \ p_s^j = \frac{\hat{w}_s \prod_{t=1}^{j} \phi(x_{it}, \hat{\eta}_s, \hat{\sigma}^2)}{\sum_{k=1}^{m_2} \hat{w}_k \prod_{t=1}^{j} \phi(x_{it}, \hat{\eta}_k, \hat{\sigma}^2)}.$$

Let $\hat{\eta}_s$ be the point in $\{\hat{\eta}_1, \ldots, \hat{\eta}_{m_2}\}$ that is closest to μ_i then use similar argument as above, we can show that there exist K' depending on α such that

$$\mathbb{P}\left(\forall\, j > K : |1 - p_s^j| < \exp(-c_\alpha j) \right) > 1 - \alpha.$$

As a result, there exist K such that

$$\mathbb{P}\left(\forall\, j > K : |\hat{\mu}_i^j - \mu_i| < |\mu_i|/2 - \epsilon \right) > 1 - \alpha.$$

Recall

$$\hat{T}_{OR}^{i,j} := \frac{(1-\hat{\pi})\prod_{t=1}^{j}\phi(x_{it},0,\hat{\sigma}^2)}{(1-\hat{\pi})\prod_{t=1}^{j}\phi(x_{it},0,\hat{\sigma}^2)+\hat{\pi}\prod_{t=1}^{j}\phi(x_{it},\hat{\mu}_i^j,\hat{\sigma}^2)}.$$

By (10.15) we also have there $\exists K$ depending on α such that

$$\mathbb{P}\left(\forall\ j > K : \hat{T}_{OR}^{i,j} < \exp(-c_\alpha j)\right) > 1-\alpha.$$

Combine this with (10.16) we have the desired result. The case for $x_{it} \sim N(0,\sigma^2)$ is similar.

Acknowledgement

We thank the editor, AE and referees for their thorough and constructive feedback, which significantly contributes to the enhancement of the paper's presentation. W.Sun thanks Prof. Aaditya Ramdas for the helpful reference of Howard et al. (2021) and valuable insights into the problem formulation. B. Gang's research was supported by NSFC grant 12201123 and STCSM grant 22YF1403000.

References

Jay Bartroff. Multiple hypothesis tests controlling generalized error rates for sequential data. *arXiv:1406.5933*, 2014.

Jay Bartroff and Jinlin Song. Sequential tests of multiple hypotheses controlling false discovery and nondiscovery rates. *arXiv:1311.3350*, 2013.

Jay Bartroff and Jinlin Song. A rejection principle for sequential tests of multiple hypotheses controlling familywise error rates. *Scandinavian Journal of Statistics*, 43(1):3–19, 2016.

Eran Bashan, Gregory Newstadt, and Alfred O Hero. Two-stage multiscale search for sparse targets. *IEEE Transactions on Signal Processing*, 59(5):2331–2341, 2011.

Yoav Benjamini and Marina Bogomolov. Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 297–318, 2014.

Yoav Benjamini and Ruth Heller. False discovery rates for spatial signals. *J. Amer. Statist. Assoc.*, 102:1272–1281, 2007. ISSN 0162-1459.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B*, **57**:289–300, 1995. ISSN 0035-9246.

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

Gilles Blanchard and Donald Geman. Hierarchical testing designs for pattern recognition. *Ann. Statist.*, 33:1155–1202, 2005.

Konrad H Bleicher, Hans-Joachim Böhm, Klaus Müller, and Alexander I Alanine. Hit and lead generation: beyond high-throughput screening. *Nature reviews Drug discovery*, 2(5): 369–378, 2003.

Paul E. Bloma, Shelby J. Fleischerb, and Zane Smilowitzb. Spatial and temporal dynamics of colorado potato beetle in fields with perimeter and spatially targeted insecticides. *Environmental Entomology*, 31(1):149–159, 2002.

T Cai, X Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):629–662, 2011.

T. Tony Cai and Wenguang Sun. Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.*, 104:1467–1481, 2009.

T. Tony Cai, Jiashun Jin, and Mark G. Low. Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, 35(6):2421–2449, 2007. ISSN 0090-5364.

T Tony Cai, Wenguang Sun, and Weinan Wang. Cars: Covariate assisted ranking and screening for large-scale two-sample inference (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:1–31, 2019.

Tony Cai and Wenguang Sun. Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):197–223, 2017.

Lijie Chen and Jian Li. On the optimal sample complexity for best arm identification. *arXiv preprint arXiv:1511.03774*, 2015.

RM Cormack. Statistical challenges in the environmental sciences: a personal view. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 201–210, 1988.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

S George Djorgovski, AA Mahabal, Ciro Donalek, Matthew J Graham, Andrew J Drake, Baback Moghaddam, and Mike Turmon. Flashes in a star stream: Automated classification of astronomical transient events. In *E-Science (e-Science), 2012 IEEE 8th International Conference on*, pages 1–8. IEEE, 2012.

Alex Dmitrienko, Brian L Wiens, Ajit C Tamhane, and Xin Wang. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in medicine*, 26(12):2465–2478, 2007.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32:962–994, 2004. ISSN 0090-5364.

Bradely Efron. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

Bradley Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.*, 99(465):96–104, 2004. ISSN 0162-1459. doi: 10.1198/016214504000000089. URL http://dx.doi.org/10.1198/016214504000000089.

Bradley Efron. Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Stat.*, **2**:197–223, 2008.

Bradley Efron. Empirical bayes deconvolution estimates. *Biometrika*, 103(1):1–20, 2016.

Aleksander Fabijan, Pavel Dmitriev, Helena Holmstrom Olsson, and Jan Bosch. Online controlled experimentation at scale: an empirical survey on the current state of a/b testing. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 68–72. IEEE, 2018.

Bowen Gang, Wenguang Sun, and Weinan Wang. Structure–adaptive sequential testing for online false discovery rate control. *Journal of the American Statistical Association*, 118 (541):732–745, 2023.

Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. B*, **64**:499–517, 2002. ISSN 1369-7412.

MN Ghosh. Bounds for the expected sample size in a sequential probability ratio test. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 360–367, 1960.

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

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

Jarvis Haupt, Richard G Baraniuk, Rui M Castro, and Robert D Nowak. Sequentially designed compressed sensing. In *SSP*, pages 401–404, 2012.

Ruth Heller and Saharon Rosset. Optimal control of false discovery criteria in the two-group model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(1): 133–155, 2021.

Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7724–7741. PMLR, 2022.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2): 1055–1080, 2021.

Ankur Jain and Edward Y Chang. Adaptive sampling for sensor networks. In *Proceeedings of the 1st international workshop on Data management for sensor networks: in conjunction with VLDB 2004*, pages 10–16, 2004.

Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

Jiashun Jin and T. Tony Cai. Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.*, **102**:495–506, 2007. ISSN 0162-1459.

Ramesh Johari, Leo Pekelis, and David J Walsh. Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*, 2015.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, 70(3):1806–1821, 2022.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pages 1781–1791. PMLR, 2020.

Roger Koenker and Jiaying Gu. Rebayes: an r package for empirical bayes mixture methods. *Journal of Statistical Software*, 82(1):1–26, 2017.

Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109 (506):674–685, 2014.

Ron Kohavi and Roger Longbotham. Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining*, 7(8):922–929, 2017.

Xingguo Li and Jarvis Haupt. Identifying outliers in large matrices via randomized adaptive compressive sampling. *IEEE Transactions on Signal Processing*, 63(7):1792–1807, 2015.

Kun Liang and Dan Nettleton. A hidden markov model approach to testing multiple hypotheses on a tree-transformed gene ontology graph. *Journal of the American Statistical Association*, 105(492):1444–1454, 2010.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR, 2016.

Matthew Malloy and Robert Nowak. Sequential analysis in high-dimensional multiple testing and sparse recovery. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2661–2665. IEEE, 2011.

Matthew L Malloy and Robert D Nowak. Sequential testing for sparse recovery. *IEEE Transactions on Information Theory*, 60(12):7862–7873, 2014a.

Matthew L Malloy and Robert D Nowak. Near-optimal adaptive compressed sensing. *IEEE Transactions on Information Theory*, 60(7):4001–4012, 2014b.

Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all ϵ-good arms in stochastic bandits. *Advances in Neural Information Processing Systems*, 33:20707–20718, 2020.

Nicholai Meinshausen and John Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, **34**:373–393, 2006.

Nicolai Meinshausen, Peter Bickel, John Rice, et al. Efficient blind search: Optimal power of detection under computational cost constraints. *The Annals of Applied Statistics*, 3 (1):38–60, 2009.

Subhojyoti Mukherjee, Kolar Purushothama Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Thresholding bandits with augmented ucb. *arXiv preprint arXiv:1704.02281*, 2017.

Peter Müller, Giovanni Parmigiani, Christian Robert, and Judith Rousseau. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99(468):990–1001, 2004.

Reda Ouhamma, Odalric-Ambrym Maillard, and Vianney Perchet. Online sign identification: Minimization of the number of errors in thresholding bandits. *Advances in Neural Information Processing Systems*, 34:18577–18589, 2021.

Martin Posch, Sonja Zehetmayer, and Peter Bauer. Hunting for significance with the false discovery rate. *Journal of the American Statistical Association*, 104(486):832–840, 2009.

Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 131–148, Berkeley and Los Angeles, 1951. University of California Press.

Francis A Roesch. Adaptive cluster sampling for forest inventories. *Forest Science*, 39(4): 655–669, 1993.

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

Nathaniel Rothman, Montserrat Garcia-Closas, Nilanjan Chatterjee, Nuria Malats, Xifeng Wu, Jonine D Figueroa, Francisco X Real, David Van Den Berg, Giuseppe Matullo, Dalsu Baris, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature genetics*, 42(11):978–984, 2010.

Sujayam Saha and Adityanand Guntuboyina. On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *The Annals of Statistics*, 48(2):738–762, 2020.

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

Sanat K Sarkar, Jingjing Chen, and Wenge Guo. Multiple testing in a two-stage adaptive design with combination tests controlling fdr. *Journal of the American Statistical Association*, 108(504):1385–1401, 2013.

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

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431, 2021.

David Siegmund. *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media, 1985.

Wenguang Sun and T. Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.*, **102**:901–912, 2007. ISSN 0162-1459.

Wenguang Sun and T. Tony Cai. Large-scale multiple testing under dependence. *J. R. Stat. Soc. B*, 71:393–424, 2009.

Wenguang Sun and Zhi Wei. Hierarchical recognition of sparse patterns in large-scale simultaneous inference. *Biometrika*, 102:267–280, 2015.

Wenguang Sun, Brian J Reich, T Tony Cai, Michele Guindani, and Armin Schwartzman. False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):59–83, 2015.

Steven K. Thompson and George A. Seber. *Adaptive sampling*. John Wiley & Sons, Inc., New York, NY, 1996.

Anja Victor and Gerhard Hommel. Combining adaptive designs with control of the false discovery rate–a generalized definition for a global p-value. *Biometrical Journal*, 49(1): 94–106, 2007.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022.

Dennis Wei and Alfred O Hero. Multistage adaptive estimation of sparse signals. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):783–796, 2013.

Dennis Wei and Alfred O Hero. Performance guarantees for adaptive estimation of sparse signals. *IEEE Transactions on Information Theory*, 61(4):2043–2059, 2015.

Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2227–2236, 2015.

Daniel Yekutieli. Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316, 2008.

Sonja Zehetmayer, Peter Bauer, and Martin Posch. Optimized multi-stage designs controlling the false discovery or the family-wise error rate. *Statistics in medicine*, 27(21): 4145–4160, 2008.

Ji-Hu Zhang, Thomas DY Chung, and Kevin R Oldenburg. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of biomolecular screening*, 4(2):67–73, 1999.

Supplementary Material for "Sparse Recovery With Multiple Data Streams: A Sequential Adaptive Testing Approach"

Appendix A. Derivation of thresholds

In this section we provide details about the derivations of the approximation formulae in (2.7). The derivation in this section is less formal. We need the following assumption in our derivation. The assumption has been commonly adopted in the literature on SPRT (e.g. Berger (1985); Siegmund (1985)).

Assumption 2 *Let* $Z_{i,1} = \log \dfrac{f(X_{i1}|\theta_i = 1)}{f(X_{i1}|\theta_i = 0)}$. *For all* i, *we have* $\mathbb{P}_{\theta_i}(Z_{i,1} = 0) < 1$, $\mathbb{P}_{\theta_i}(|Z_{i,1}| < \infty) = 1$, $\mathbb{P}_{\theta_i}(Z_{i,1} < 0) > 0$, *and* $\mathbb{P}_{\theta_i}(Z_{i,1} > 0) > 0$. *Moreover,* $M_{\theta_i}(t) = \mathbb{E}_{\theta_i}[e^{tZ_{i,1}}]$ *exists for all* t.

We start our derivation by noting that $T_{OR}^{i,j}$ is a monotone function of the likelihood ratio statistic $\mathcal{L}_{i,j}$:

$$T_{OR}^{i,j} = \mathbb{P}\left(\theta_i = 0 | \boldsymbol{X}_i^j\right) = 1/\left(1 + \frac{\pi}{1-\pi}\mathcal{L}_{i,j}\right). \tag{A.1}$$

Hence $\boldsymbol{d}^{\pi}(t_l, t_u)$ can be expressed as a thresholding rule based on $\mathcal{L}_{i,j}$:

$$\text{stops sampling for unit } i \text{ at time } N_i = \min\{j \geq 1 : \mathcal{L}_{i,j} \leq A \text{ or } \mathcal{L}_{i,j} \geq B\},$$
$$\text{deciding } \delta_{i,N_i} = 0 \text{ if } \mathcal{L}_{i,j} \leq A \text{ and } \delta_{i,N_i} = 1 \text{ if } \mathcal{L}_{i,j} \geq B.$$

We first solve (A, B) for a given pair (α, γ), then transform (A, B) to (t_l, t_u). The technique used in our derivation is similar to the classical ideas when deriving the upper and lower thresholds for SPRT. Since all testing units operate independently and have the same thresholds, it is sufficient to focus on the operation of SPRT on a generic testing unit. Hence, for simplicity, we drop index i and denote $\mathcal{L}_{i,N_i}$, θ_i and $Z_{i,k}$ as $\mathcal{L}_N$, θ and $Z_{\cdot,k}$, respectively.

Under the random mixture model, the FPR and MDR of the SPRT with thresholds (A, B) can be calculated as

$$\text{FPR} = \frac{(1-\pi)\mathbb{P}\left(\mathcal{L}_N > B | \theta = 0\right)}{\mathbb{P}\left(\mathcal{L}_N > B\right)}, \ \text{MDR} = \mathbb{P}\left(\mathcal{L}_N < A | \theta = 1\right).$$

Let $S_N = \sum_{k=1}^{N} Z_{\cdot,k} = \log \mathcal{L}_N$. Denote $a = \log A$ and $b = \log B$. Under Assumption 2, $\mathbb{P}_\theta(N < \infty) = 1$ and all moment of N exist. There exists a unique nonzero number t_θ for which $M_\theta(t_\theta) = 1$ (Berger (1985)). This fundamental identity then implies

$$\begin{aligned} 1 &= \mathbb{E}_\theta\left\{\exp(t_\theta S_N) M_\theta(t_\theta)^{-N}\right\} \\ &= \mathbb{E}_\theta\left\{\exp(t_\theta S_N)\right\} \\ &\approx \exp(t_\theta a)\mathbb{P}_\theta\left(S_N \leq a\right) + \exp(t_\theta b)\mathbb{P}_\theta\left(S_N \geq b\right). \end{aligned} \tag{A.2}$$

In the above approximation, we ignore the overshoots and pretend that S_N hits the boundaries a and b exactly. In reality, $\mathbb{E}_\theta\{\exp(t_\theta S_N)|S_N \le a\} \le \exp(t_\theta a)$ and $\mathbb{E}_\theta\{\exp(t_\theta S_N)|S_N \ge b\} \ge \exp(t_\theta b)$. Despite this fact, our approximation works well in all our simulation settings. In this idealized situation, S_N has a two-point distribution $\mathbb{P}^*_\theta$:

$$\begin{aligned}\mathbb{P}^*_\theta(S_N = a) &= \mathbb{P}_\theta(S_N \le a),\\ \mathbb{P}^*_\theta(S_N = b) &= \mathbb{P}_\theta(S_N \ge b).\end{aligned}$$

Moreover, Assumption 2 implies that

$$\begin{aligned}1 &= \mathbb{P}_\theta(N < \infty) = \mathbb{P}_\theta(\mathcal{L}_N \le A) + \mathbb{P}_\theta(\mathcal{L}_N \ge B)\\ &= \mathbb{P}_\theta(S_N \le a) + \mathbb{P}_\theta(S_N \ge b).\end{aligned}$$

Thus we can solve from the above that

$$\mathbb{P}_\theta(\mathcal{L}_N \ge B) = \mathbb{P}_\theta(S_N \ge b) \approx \frac{1-\exp(t_\theta a)}{\exp(t_\theta b) - \exp(t_\theta a)}.$$

According to Assumption 2, $\mathbb{P}_\theta(|Z_{\cdot,k}| < \infty) = 1$ for $\theta = 0, 1$. Then $t_{\theta=0} = 1, t_{\theta=1} = -1$ (P493, Berger (1985)). It follows that

$$\begin{aligned}\text{FPR} &\approx \frac{(1-\pi)\frac{1-A}{B-A}}{\mathbb{P}(\mathcal{L}_N \ge B)} = \frac{(1-\pi)\frac{1-A}{B-A}}{(1-\pi)\frac{1-A}{B-A} + \pi\frac{1-1/A}{1/B-1/A}}\\ &= \frac{1-\pi}{1-\pi+\pi B},\\ \text{MDR} &\approx 1 - \frac{1-\exp(-a)}{\exp(-b)-\exp(-a)} = 1 - \frac{1-1/A}{1/B-1/A}\\ &= \frac{A(B-1)}{B-A}.\end{aligned}$$

Setting FPR $= \alpha$, MDR $= \gamma$ and solving for A and B, we have

$$A \approx \frac{(\alpha^{-1}-1)(1-\pi)\gamma}{(\alpha^{-1}-1)(1-\pi) - \pi + \pi\gamma},\ B \approx \frac{(\alpha^{-1}-1)(1-\pi)}{\pi}.$$

The relationship (A.1) implies that

$$A = \frac{(1-\pi)(1-t_u)}{\pi t_u},\ B = \frac{(1-\pi)(1-t_l)}{\pi t_l}. \tag{A.3}$$

Transforming from $\mathcal{L}_{i,j}$ to $T^{i,j}_{OR}$, the corresponding thresholds can be obtained as:

$$t^l_{OR} = \alpha \quad \text{and} \quad t^u_{OR} = \frac{\pi\alpha\gamma + 1 - \pi - \alpha}{\pi\gamma + 1 - \pi - \alpha}.$$

To ensure an effective MDR control, we choose a more stringent upper threshold:

$$t^u_{OR} = \frac{1-\pi}{\pi\gamma + 1 - \pi} \ge \frac{\pi\alpha\gamma + 1 - \pi - \alpha}{\pi\gamma + 1 - \pi - \alpha},\ \forall \alpha \ge 0.$$

Appendix B. The impact of sparsity levels on SMART

This section conducts simulation studies to investigate the effectiveness of FPR/MDR control at different sparsity levels. The numerical results suggest that sparsity has substantial impact on the finite-sample performance of the data-driven procedure.

Consider the two group mixture model $(1-p)N(0,1)+pN(\mu,1)$ with $\mu = 3.5$. We generate $m = 100,000$ observations and vary p from 0.01 to 0.1 with step-size 0.01. The FDR and MDR are computed by averaging the results in 100 data sets. The FPR, MDR and total sample sizes are at varied sparsity levels are displayed in 1. the estimation accuracy is negatively affected when signals become more sparse: SMART has mildly inflated FPR, conservative MDR, and requires a larger sample size to effectively recover the signals.

Finally, we emphasize that the conclusions from numerical studies are limited and preliminary. The effectiveness of FPR/MDR control is a complicated issue that depends on several factors jointly.

(i) The reliable estimation of the stage-wise FPR/MDR requires a relatively large number of rejections, which in turn requires that the signals cannot be too sparse or too weak.

(ii) The consistent estimation of the non-null proportion requires that the detection and discovery boundaries must be achieved.

(iii) The quality of density estimators depends on the sample size and smoothness of the underlying function.

We conclude that an important future research direction is the development of precise estimation methods, which is instrumental for constructing powerful multi-stage testing procedures.

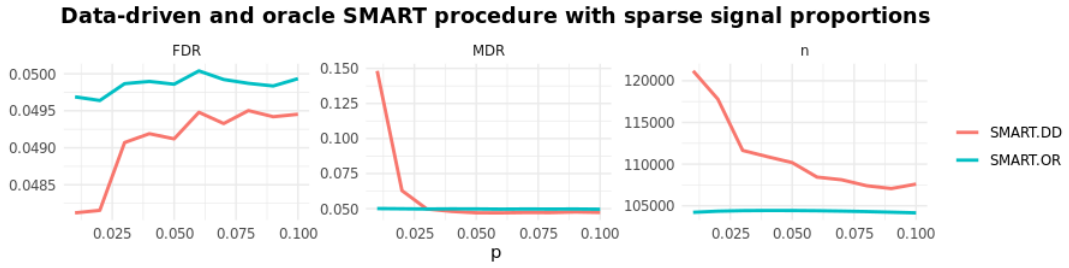


Figure 1: Error rates and total samples for data-driven and oracle SMART procedures with varying sparse signal proportions.

Appendix C. Comparison of realized FPR and FDR levels

We present the comparison of the realized FPR and FDR levels for the settings in Section 5.1. We can see that the FDR and FPR for each method (OR.SM and DD.SM) are very similar in all scenarios. However, the FPR and FDR can be very different when signals are rare and weak. See discussions in the next section.

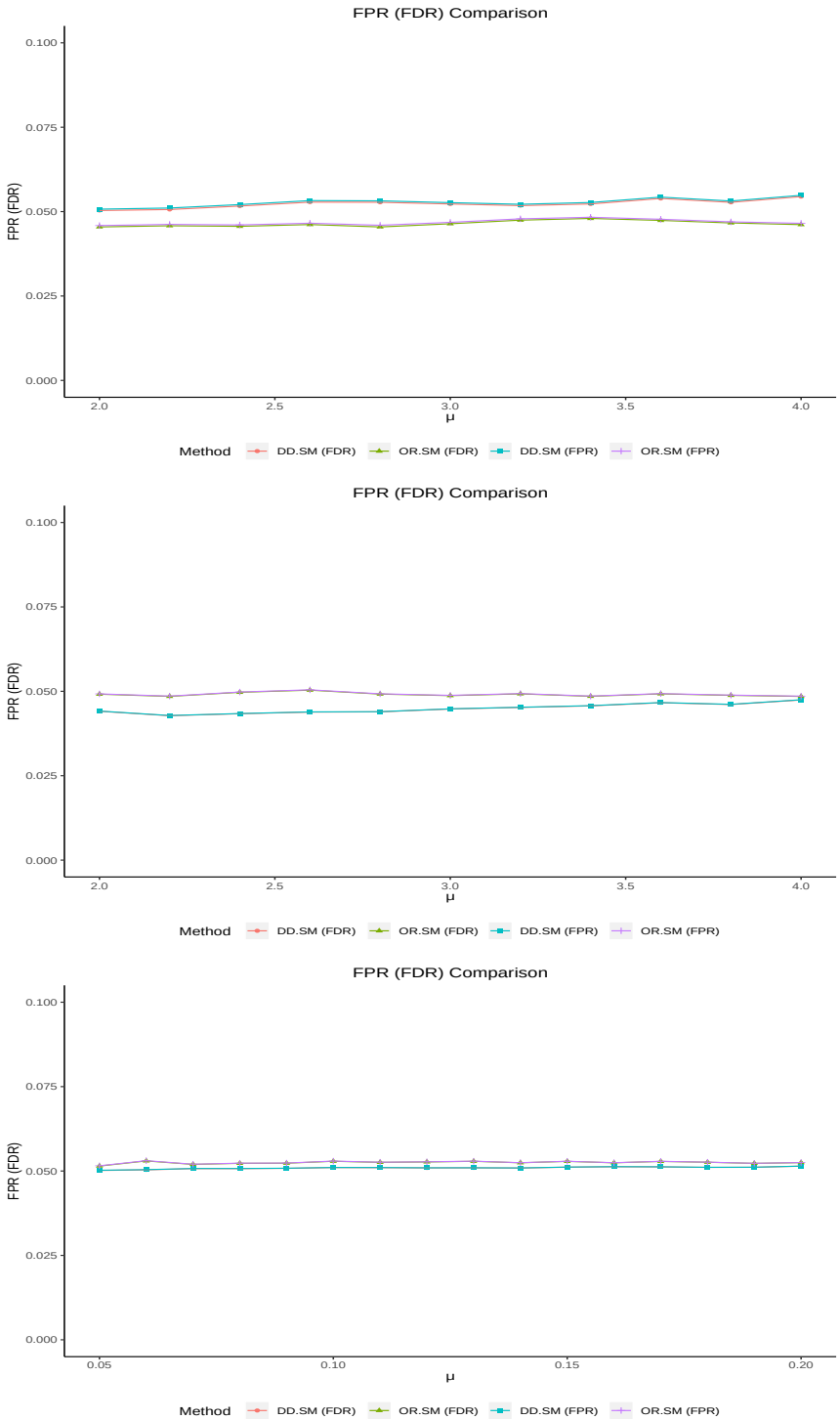


Figure 2: Dependent tests: Settings 1-3 are described in the main text. SMART controls the error rates at the nominal levels in most settings.

Appendix D. Some further investigations of FPR and FDR

We compare the FPR and FDR under the setting where signals are sparse and weak. We simulate 100 data sets, each with $m = 10000$ observations obeying the following two-group mixture model $X_i \sim (1-p)N(0,1) + pN(\mu, 1)$. In Setting 1, we fix $p = 0.05$ and vary μ from 0 to 2 with step size 0.1. In Setting 2, we vary p from 0.01 to 0.2 with step size 0.01 and fix $\mu = 1$. The range of 0 to 2 is considered as the "weak" signal setting ($r < 0.5$)..

We apply the BH procedure at $\alpha = 0.05$ at varied sparsity levels and inspect the corresponding mFDR and FDR levels. Let R^k and V^k be the number of rejections and number of false positives in data set k, $k = 1, \cdots, 100$. Then the FDR and mFDR are estimated as $\widehat{\text{FDR}} = \frac{1}{100}\sum_{k=1}^{100} \frac{V_k}{R_k \vee 1}$; $\widehat{\text{mFDR}} = \frac{\sum_{k=1}^{100} V_k}{\sum_{k=1}^{100} R_k}$. The simulation results are summarized in Figure 3, where we also display the total number of rejections ($\sum_{k=1}^{100} R_k$) and total number of false positives ($\sum_{k=1}^{100} V_k$). The following observations can be made.

(a) The FDR levels are equal to the nominal level approximately in all settings. The mFDR is very different from FDR when the signals are sparse and weak.

(b) From the left parts of the two panels on the right column, we can see that almost all rejections made by BH in the 100 data sets are false positives when signals are weak and sparse. However, this is not reflected by the FDR since $P(\text{FDP} = 0)$ is high. By contrast, the mFDR correctly reveals that most of our discoveries are false.

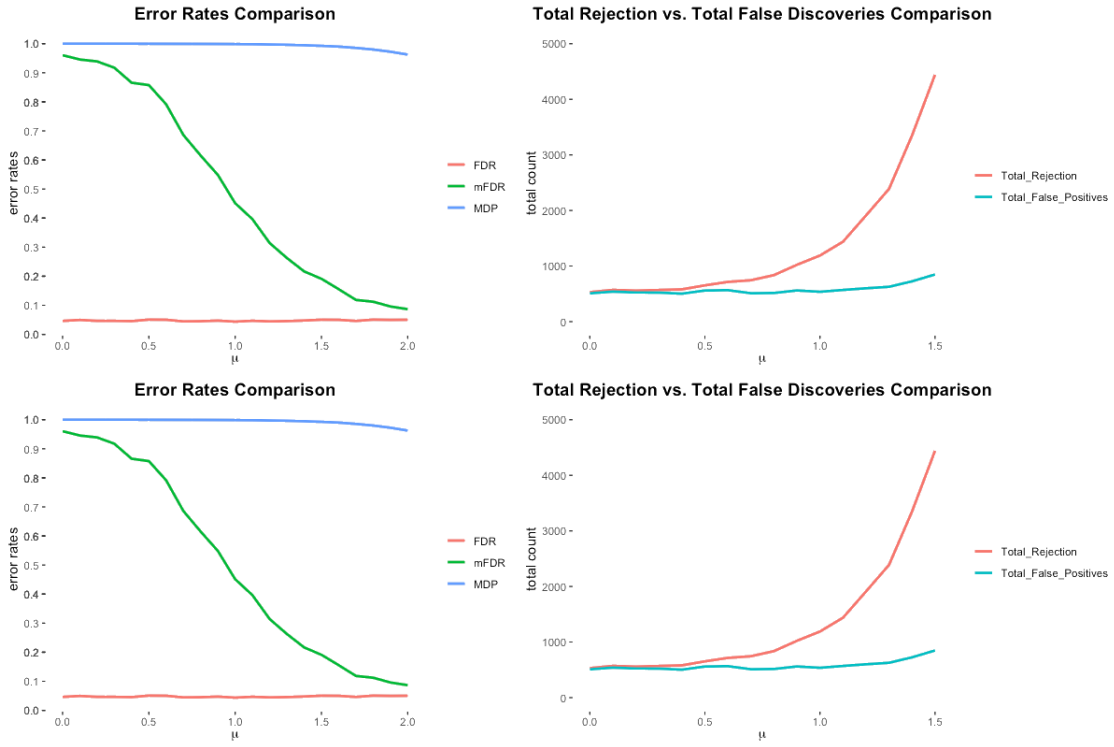


Figure 3: Comparison of the mFDR and FDR.