Ranking and Selection in Large-Scale Inference of Heteroscedastic Units

Bowen Gang*
Department of Statistics and Data Science, Fudan University
Luella Fu
Department of Mathematics, San Francisco State University
Gareth James
Goizueta Business School, Emory University
and
Wenguang Sun
School of Management and Center for Data Science, Zhejiang University

Abstract

The allocation of limited resources to a large number of potential candidates presents a pervasive challenge. In the context of ranking and selecting top candidates from heteroscedastic units, conventional methods often result in over-representations of subpopulations, and this issue is further exacerbated in large-scale settings where thousands of candidates are considered simultaneously. To address this challenge, we propose a new multiple comparison framework that incorporates a modified power notion to prioritize the selection of important effects and employs a novel ranking metric to assess the relative importance of units. We develop both oracle and data-driven algorithms, and demonstrate their effectiveness in controlling the error rates and achieving optimality. We evaluate the numerical performance of our proposed method using simulated and real data. The results show that our framework enables a more balanced selection of effects that are both statistically significant and practically important, and results in an objective and relevant ranking scheme that is well-suited to practical scenarios.

Key words and phrases: Compound decision theory; Composite null hypotheses; Deconvolution estimates; Empirical Bayes; False discovery rate; Weighted multiple testing

*B. Gang's research was supported by National Natural Science Foundation of China grant 12201123

1 Introduction

Allocating limited resources among numerous potential candidates is a common problem faced by both individuals and organizations. This dilemma is encountered by NBA basketball recruiters as they search for promising talent, public policy makers as they fund educational programs, and internet users on platforms such as Yelp as they decide which restaurants to visit. Such decision-making scenarios give rise to the ranking and selection problem, a fundamental statistical issue that requires the comparison of multiple unknown parameters.

Ranking and selection has been a classical topic in multiple comparisons (Mosteller, 1948; Paulson, 1949; Bechhofer, 1954; Gupta, 1965; Panchapakesan, 1971; Goel and Rubin, 1977), and its integration into other branches of statistics, operations research, and computing has made it a critical and constantly evolving area of study (Chen et al., 2000; Boyd et al., 2012; Luo et al., 2015; Ni et al., 2017; Kamiński and Szufel, 2018; Zhong et al., 2022). The decision process has two key components: first, establishing a meaningful criterion for ordering a pool of potential candidates, and second, selecting a subset of "most meritorious" candidates with a certain level of confidence. Properly accounting for the heteroscedasticity across data from diverse study units is essential for producing effective, sensible, and fair decisions in the ranking and selection process. In the following, we provide first an overview of conventional practices and identify relevant issues and then an exposition of our new framework for addressing the challenge of heteroscedasticity. Finally, we discuss related works and highlight the contributions of our approach.

1.1 Conventional practices and issues

Ranking is essential in multiple comparisons to evaluate and identify top-performers from a pool of potential candidates. While the importance of each candidate is linked to the magnitude of its associated parameter, the decision-making process also takes into account the associated uncertainties in order to ensure that the top candidates indeed belong to the "most meritorious" group. The two perspectives, namely the parameter magnitude and the confidence level in the assertions being made, are reflected by the estimated effect size and its associated statistical significance, respectively. In homoscedastic models, these two perspectives yield the same ranking. However, in cases where the data are heteroscedastic across study units, the rankings based on these two perspectives may disagree. As demonstrated shortly, the issue is further exacerbated in large-scale settings where thousands of candidates are being considered at once. Developing a sensible ranking and selection criterion that partially mitigates the conflict between the two perspectives poses a critical challenge in large-scale multiple comparison problems.

To demonstrate the drawbacks of conventional practices, we analyze the 2005 Annual Yearly Performance (AYP) data to identify K-12 schools with significant gaps in passing rates between socioeconomically advantaged (SEA) and disadvantaged (SED) students. The raw observations are the empirical differences in passing rates between the two groups, with standard errors (SEs) linked to the number of students in the schools. More details of the study are provided in Section F.7. We consider three selection strategies, which are respectively based on statistical significance (p-value), observed gap in passing rates (raw observation), and posterior mean (computed using Tweedie's formula). The results of our exploratory analysis are presented in Figure 1. Panel (a) shows the distribution of the SE. Panels (b), (c) and (d) show the distribution of 20 selected schools according to p-value,

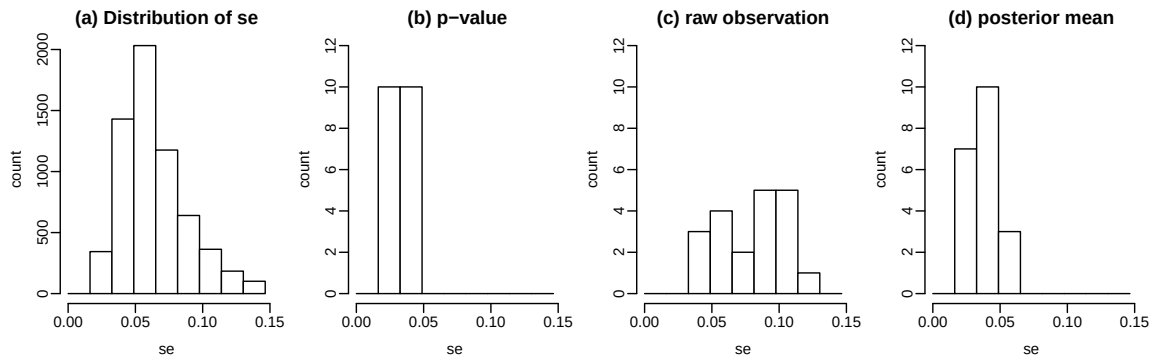

Figure 1: Panel (a): overall distribution of SEs of the AYP data set. Panel (b): distribution of SEs of the top 20 schools according to p-value. Panel (c): distribution of SEs of the top 20 schools according to raw observation. Panel (d): distribution of SEs of the top 20 schools according to posterior mean.

observed gap and posterior mean, respectively. Figure 1 reveals that selecting schools based on p-values and posterior means tends to result in an over-representation of schools with low SEs, while selecting based on raw observations may lead to an over-representation of those with high SEs. The design of this analysis draws on earlier works, including Sun and McLain (2012) and Henderson and Newton (2016), which have identified and provided initial insights into some perplexing phenomena that arise under heteroscedastic models. For example, Sun and McLain (2012) found that the largest 1% of K-12 schools are over-represented among the worst performing ten schools when the selection is based on p-values.

Although all three selection criteria have their advantages in capturing either the effect sizes or accounting for associated uncertainties, the over-representation of subgroups with high/low SEs is undesirable and runs counter to practical wisdom. We aim to develop a new ranking and selection framework that resides between the three polarized selection criteria, striking a balance between their respective advantages and disadvantages. The AYP data will be revisited under our new framework in Section F.7.

1.2 False discovery rate analysis under heteroscedasticity

We begin by examining the use of the false discovery rate (FDR) framework (Benjamini and Hochberg, 1995; Storey, 2002; Genovese and Wasserman, 2004) in the context of selecting important candidates. Most FDR methods operate in two steps: ranking and thresholding, where the building block of operation is the p-value (Benjamini and Hochberg, 1995) or the local false discovery rate (lfdr; Efron et al., 2001; Sun and Cai, 2007). Both the p-value and lfdr tend to prioritize the stability of data over effect sizes, which leads to the over-selection of schools with low SEs in the AYP analysis. To comprehend the limitations of the conventional formulation, we refer to the theory in Sun and Cai (2007), which shows that thresholding lfdr is optimal in the sense that it maximizes the average power subject to the constraint on the FDR. This perspective reveals two major issues that contribute to the difficulties of utilizing the FDR framework in heteroscedastic models.

The first issue is that the concept of average power, which is defined as the expected proportion of non-nulls that are correctly rejected, overlooks the severity of missed signals. This gives rise to a significant limitation that is particularly concerning in situations with substantial heteroscedasticity across units. Specifically, the identification of a large signal should be rewarded more than that of a small signal, even if both study units have the same level of statistical significance. However, this principle is not fulfilled by the conventional FDR formulation. To correct the inherent bias in conventional FDR analyses, it is desirable to modify the power concept such that selection of larger effects is prioritized with a higher reward.

The second issue pertains to the conventional multiple testing framework, which employs a thresholding procedure that is contingent on a fixed ordering determined by a predefined significance index, such as the p-value or lfdr. Our optimality theory reveals that such

an ordering to prioritize selections may not exist. The absence of a universally optimal ordering poses a significant challenge in developing an objective ranking, as the rankings can be inconsistent across users who may reasonably select different confidence or reference levels.

1.3 A preview of the proposed method

Our proposal presents a new multiple comparison framework that addresses the two aforementioned issues by incorporating a modified power notion to prioritize the selection of important effects and employing a novel ranking index to assess the relative importance of units.

We first study the prioritized selection problem by utilizing a constrained optimization formulation. The goal is to control a user-specified FDR while maximizing a modified power concept that assigns higher rewards to selections of larger effects. The solution leads to a selection method that carefully weighs the candidate's effect size against its significance. The new formulation reduces the bias inherent in commonly used significance indices that favor stability, ensuring that the effect size is more fairly represented in the selection process.

We then turn to the ranking issue by introducing a novel concept called the "r-value," which provides a measure of the relative importance of study units in a list. The importance of different units is captured by how early they are selected according to a varying target. The earlier a unit is selected, the more important it is considered to be relative to the other units. Thus an objective ranking of study units is generated.

1.4 Our contributions

In scenarios where study units display substantial heteroscedasticity, the proposed ranking and selection procedure offers a valuable alternative to conventional FDR analyses. Our method enables a more balanced selection of effects that are both statistically significant and practically important, resulting in a ranking that is objective and relevant for practical scenarios. To tackle the complexity that arises from our revised notion of power, we have devised an oracle procedure and developed theory to establish its optimality. The new theory offers a significant advance in contrast to the weighted FDR theory in Basu et al. (2018). Furthermore, we have developed a computational shortcut of the oracle procedure and rigorously established the asymptotic properties of the corresponding data-driven algorithm. Our work presents a unified framework that explicitly incorporates considerations of effect size, statistical significance, error control, theoretical guarantees, and computational efficiency for analyzing heteroscedastic data.

Previous studies have made progress in addressing some, but not all, of our challenges. Sun and McLain (2012) proposed a decision-theoretic framework that incorporates information about effect sizes, but their approach relies on standardization and does not resolve the issue of over-representation of small variances. Henderson and Newton (2016) put forth the maximal agreement method to avoid over-representation. However, their formulation differs significantly from ours in two aspects: firstly, the question of error rate control is left unaddressed, and secondly, the joint consideration of effect size and significance is absent. Gu and Koenker (2023) devised a robust set of ranking and selection methods within a compound decision-theoretic framework. Notably, they extended the maximal agreement method to include false discovery control. However, the challenge of balancing statistical significance and effect size has not been fully resolved. Finally, Fu et al. (2022)

demonstrated that standardization can distort structural information about the alternative distribution, but their analysis had focused on the conventional FDR framework.

1.5 Organization

The paper is structured as follows. Section 2 presents the problem formulation and an oracle procedure for prioritized selection. Section 3 develops a data-driven procedure and establishes its theoretical guarantees. Section 4 introduces the r-value and discusses its agreeability property. Sections 5 and 6 present results to illustrate the numerical performance of our proposed ranking and selection methods on both simulated and real data.

2 Prioritized Selection with FDR Control

This section first introduces the model, notation and problem formulation, and then proposes an oracle procedure for prioritized selection of important effects.

2.1 Problem formulation

Suppose X_i, $i \in [m] \equiv \{1, \cdots, m\}$ are independent observations from a random mixture model with possibly heteroscedastic errors:

$$X_i = \mu_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2), \tag{2.1}$$

where μ_i and σ_i are assumed to be mutually independent, and come from unspecified distributions with bounded supports:

$$\mu_i \sim g_\mu(\cdot), \qquad \sigma_i^2 \sim g_\sigma(\cdot), \quad i \in [m]. \tag{2.2}$$

To focus on the central idea, we assume that σ_i are known, a common practice pursued, for example, in Efron (2011), Xie et al. (2012), and Weinstein et al. (2018). The issue of

estimating unknown and heterogeneous σ_i has been considered in Gu and Koenker (2017a), Gu and Koenker (2017b), Banerjee et al. (2020), and Kwon and Zhao (2023).

Let $\mathcal{A}_i$ be a user-specified indifference region. Without loss of generality, suppose one wishes to test whether the effect size μ_i surpasses a given threshold μ_0, hence $\mathcal{A}_i = \mathcal{A} = \{\mu : \mu \leq \mu_0\}$. Upon observing (x_i, σ_i), the null and alternative hypotheses are

$$H_{0,i} : \mu_i \in \mathcal{A} \quad \text{vs.} \quad H_{1,i} : \mu_i \notin \mathcal{A}.$$

Denote $\theta_i = \mathbb{I}(\mu_i > \mu_0)$ the true state of the ith item, and $\delta_i \in \{0, 1\}$ the decision we make about that item, where $\delta_i = 1$ if the ith item is selected (or claimed as an important case) and $\delta_i = 0$ otherwise. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$.

In large-scale selection problems, a practical and effective goal is to control the false discovery rate (Benjamini and Hochberg, 1995)

$$\mathrm{FDR}(\boldsymbol{\delta}) = \mathbb{E}\left[\frac{\sum_i (1-\theta_i)\delta_i}{(\sum_i \delta_i) \vee 1}\right],$$

where $a \vee b = \max(a, b)$. A closely related quantity is the marginal false discovery rate

$$\mathrm{mFDR}(\boldsymbol{\delta}) = \frac{\mathbb{E}\left(\sum_i (1-\theta_i)\delta_i\right)}{\mathbb{E}\left(\sum_i \delta_i \vee 1\right)}.$$

Under certain first- and second-order conditions, the mFDR asymptotically equals the FDR (Genovese and Wasserman, 2002; Cai et al., 2019). For theoretical convenience we adopt the mFDR in our discussion.

In conventional FDR analysis, the goal is to find a decision rule $\boldsymbol{\delta}$ that controls the error rate at pre-specified level α with the largest power. A widely used metric for evaluating the power of a multiple testing procedure is the expected number of true positives

$$\mathrm{ETP}(\boldsymbol{\delta}) = \mathbb{E}\left(\sum_i \theta_i \delta_i\right) = \mathbb{E}\left\{\sum_i \mathbb{I}(\mu_i > \mu_0)\delta_i\right\}. \tag{2.3}$$

To prioritize the selection of large effects, we propose to modify the power concept as

$$\mathrm{ETP}^*(\boldsymbol{\delta}) = \mathbb{E}\left\{\sum_i (X_i - \mu_0)\delta_i\right\}. \tag{2.4}$$

The traditional power concept (2.3) has undergone two modifications: first, the indicator

$\mathbb{I}(\mu_i - \mu_0 > 0)$ is replaced by the actual difference $(\mu_i - \mu_0)$. Second, the unbiased estimate X_i is substituted in place of μ_i, resulting in the revised power concept (2.4). The first modification allows for the revised power metric to precisely capture the impact of signal magnitude, while the subsequent alteration is crucial for avoiding technical intricacies, as the unknown μ_i poses a significant difficulty in constructing the oracle rule in Section 2.2.

The above considerations give rise to the following constrained optimization problem, in which we aim to develop a selection rule $\boldsymbol{\delta} \in \{0,1\}^m$ to

$$\text{maximize } \text{ETP}^*(\boldsymbol{\delta}) \quad \text{subject to} \quad \text{mFDR}(\boldsymbol{\delta}) \leq \alpha. \tag{2.5}$$

Remark 1. Our formulation can be extended by replacing $(X_i - \mu_0)$ with a more general function $h_i(\boldsymbol{X}, \boldsymbol{\sigma})$. If one only cares about detecting the true state of nature and ignores the severity of missed signals, then we can take $h_i(\boldsymbol{X}, \boldsymbol{\sigma}) = \mathbb{I}(X_i > \mu_0)$. The other possible choice for $h_i(\boldsymbol{X}, \boldsymbol{\sigma})$ is $(X_i - \mu_0)_+$, which ensures that the weight is always positive. Moreover, the choice of $h_i(\boldsymbol{X}, \boldsymbol{\sigma}) = (X_i - \mu_0)_+$ simplifies subsequent analyses. However, we prefer $(X_i - \mu_0)$ over $(X_i - \mu_0)_+$, as the former penalizes the identification of small effects. This preference is in line with the objective of our formulation, which aims to allocate a more balanced representation to the effect size during the selection process. In Section 2.2, we demonstrate the critical role of the sign of $h_i(\boldsymbol{X}, \boldsymbol{\sigma})$. In contrast to the weighted FDR problem discussed in Basu et al. (2018), where the weights w_i are assumed to be non-negative and independent of X_i, the "weights" $h_i(\boldsymbol{X}, \boldsymbol{\sigma})$ in our formulation are allowed to be negative and depend on $(\boldsymbol{X}, \boldsymbol{\sigma})$. The difference poses new challenges in developing both oracle and data-driven procedures. We discuss related issues in subsequent sections.

2.2 Oracle selection procedure

This section considers an ideal scenario where an oracle knows g_μ and g_σ in (2.2). The oracle rule weighs the tradeoffs between α-investing and μ-investing processes, two concepts that we shall elaborate on shortly, and assesses their impacts on the modified power and FDR capacity, respectively. In what follows, we present a heuristic argument to explain how we arrive at the oracle rule, and rigorously prove its optimality in Theorem 1.

The process of α-investing (Foster and Stine, 2008; Gang et al., 2023), which is used to evaluate the gains and losses in making a discovery, relies on the conditional local false discovery rate statistic (Clfdr, Cai and Sun (2009); Efron (2012); Sun and McLain (2012)). The Clfdr statistic is defined as

$$\text{Clfdr}_i = \mathbb{P}(\mu_i \in \mathcal{A} | x_i, \sigma_i) = \frac{f_{0i}(x_i)}{f_i(x_i)}, \tag{2.6}$$

where $f_{0i}(x_i) = \int_{\mu\in\mathcal{A}} \phi_{\sigma_i}(x_i - \mu) g_\mu(\mu) d\mu$ and $f_i(x_i) = \int_{-\infty}^{\infty} \phi_{\sigma_i}(x_i - \mu) g_\mu(\mu) d\mu$. The ordered values of Clfdr statistics are denoted $\text{Clfdr}_{(1)}, \ldots, \text{Clfdr}_{(m)}$. As shown by Sun and Cai (2007), the following step-wise algorithm, which uses the Clfdr statistic as a basic operation unit, is asymptotically optimal in the sense that it maximizes the ETP subject to the constraint mFDR $\leq \alpha$.

$$\text{Let } k = \max\left\{ j : \tfrac{1}{j} \textstyle\sum_{i=1}^{j} \text{Clfdr}_{(i)} \leq \alpha \right\}, \text{ then reject } H_{(1)}, \ldots, H_{(k)}. \tag{2.7}$$

The Clfdr algorithm (2.7) can be interpreted as a varying-capacity knapsack process (Basu et al., 2018; Gang et al., 2023). Specifically, (2.7) can be viewed as an iterative decision process where the initial α-wealth is invested by rejecting hypotheses sequentially. The process adheres to the following constraint:

$$\text{Clfdr}_j - \alpha \leq C_j \equiv -\sum_{H_i \in \mathcal{R}_j} (\text{Clfdr}_i - \alpha), \text{ for } j = 1, 2, \ldots,$$

where $\mathcal{R}_j \subset \{H_1, H_2, \ldots, H_m\}$ is the collection of rejected hypotheses at step j, and C_j may be viewed as the *capacity* of the knapsack at step j, with the default choice $C_1 = 0$. Under

this view, the α-investing process corresponds to a knapsack problem whose capacity may either expand or shrink over time. If H_j with $\text{Clfdr}_j < \alpha$ ($\text{Clfdr}_j > \alpha$) is rejected, then C_j increases (decreases) by $|\alpha - \text{Clfdr}_j|$.

The μ-investing process, on the other hand, is relatively straightforward. When a hypothesis H_j with $X_j > \mu_0$ ($X_j < \mu_0$) is rejected, the return on investment increases (decreases) empirically by $|X_j - \mu_0|$.

Jointly considering the gains and losses in the α-investing and μ-investing processes, we divide the hypotheses into four groups:

0. $X_i - \mu_0 \geq 0$ and $\text{Clfdr}_i - \alpha \leq 0$;

1. $X_i - \mu_0 \geq 0$ and $\text{Clfdr}_i - \alpha > 0$;

2. $X_i - \mu_0 < 0$ and $\text{Clfdr}_i - \alpha \leq 0$;

3. $X_i - \mu_0 < 0$ and $\text{Clfdr}_i - \alpha > 0$.

Our problem formulation suggests that units in group 0 should always be selected, as their selection results in an increase in both α-wealth and power. Conversely, units in group 3 should never be selected, as their selection leads to decreases in both α-wealth and power. The tradeoffs involved in selecting units from groups 1 and 2 are nuanced. In the case of group 1, selecting units sacrifices capacity but also results in increased power. We hypothesize that the optimal strategy involves selecting units with a high value-to-cost ratio, defined as

$$T_i = \frac{X_i - \mu_0}{\text{Clfdr}_i - \alpha}, \tag{2.8}$$

By contrast, selecting units from group 2 involves trading power for increased capacity. Consequently, the T_i statistic can be viewed as a cost-to-value ratio. Therefore, it is desirable to select units with low values of T_i from group 2. We hypothesize that the

optimal decision rule can be expressed in the following form:

$$\delta(c_1, c_2)(T_i) = \begin{cases} 1 & \text{if } (X_i, \text{Clfdr}_i) \text{ belongs to group 0} \\ 1 & \text{if } (X_i, \text{Clfdr}_i) \text{ belongs to group 1 and } T_i > c_1 \\ 1 & \text{if } (X_i, \text{Clfdr}_i) \text{ belongs to group 2 and } T_i < c_2 \\ 0 & \text{otherwise} \end{cases}, \tag{2.9}$$

where c_1 and c_2 are thresholds to be determined.

Consider a class of decision rules of the form (2.9), with mFDR and modified power denoted as $\text{mFDR}(c_1, c_2)$ and $\text{ETP}^*(c_1, c_2)$, respectively. Define the oracle cutoffs

$$(c_1^{OR}, c_2^{OR}) = \arg\max_{(c_1,c_2)} \{\text{ETP}^*(c_1, c_2) : \text{mFDR}(c_1, c_2) = \alpha\}. \tag{2.10}$$

The next theorem shows that the decision rule given by (2.9) and (2.10) is optimal under the formulation (2.5).

Theorem 1. *The oracle procedure $\boldsymbol{\delta}^{OR} = \boldsymbol{\delta}(c_1^{OR}, c_2^{OR})$ proposed above controls mFDR at level α and is optimal in the sense that for any decision rule $\boldsymbol{\delta}$ that controls mFDR at level α, we always have $ETP^*(c_1^{OR}, c_2^{OR}) \geq ETP^*_{\boldsymbol{\delta}}$.*

2.3 Extension of the oracle procedure

We present an extension of the oracle procedure capable of solving the following problem

$$\text{Maximize } \mathbb{E}\left\{\sum_{i=1}^{m} h_i(\boldsymbol{X}, \boldsymbol{\sigma})\delta_i\right\} \quad \text{subject to} \quad \text{mFDR} \leq \alpha. \tag{2.11}$$

The previously stated formulation (2.5) may be recovered by setting $h_i(\boldsymbol{X}, \boldsymbol{\sigma}) = X_i - \mu_0$. In the Supplementary Material, we show that the previous oracle rule (2.9) is optimal under the formulation (2.11) by following two adjustments. First, the group membership is determined jointly based on the signs of the pair $\{h_i(\boldsymbol{X}, \boldsymbol{\sigma}), \text{Clfdr}_i - \alpha\}$. Second, the ranking statistic is modified as $T_{OR}^i = \frac{h_i(\boldsymbol{X}, \boldsymbol{\sigma})}{\text{Clfdr}_i - \alpha}$.

The proposed extension has several important implications. Firstly, it allows us to

consider more general indifference regions $\mathcal{A}_i$, such as general Borel sets instead of the restricted one-sided regions $\mathcal{A} = \{\mu : \mu \leq \mu_0\}$. We can design modified functions that reflect the distance of X_i from $\mathcal{A}_i$, such as $h_i(\boldsymbol{X}, \boldsymbol{\sigma}) = \min\{\|X_i - \mu\|, \mu \in \mathcal{A}_i\}$, where $\|\cdot\|$ represents some norm in a metric space. Secondly, alternative loss functions may be utilized to increase the flexibility of our framework. Finally, the extension enables researchers to use $\hat{\mu}_i$, such as the James-Stein estimator and Tweedie's estimator for μ_i (Efron, 2012), in place of X_i to modify the ETP*. However, one needs to proceed with caution as the use of such estimates may introduce additional variability and uncertainty, which may lead to unstable and counter-intuitive selections. The formulation $h_i(\boldsymbol{X}, \boldsymbol{\sigma}) = X_i - \mu_0$ still remains a straightforward, intuitive, and stable option.

2.4 ETP vs ETP*: an illustrative example

In the setting with homoscedastic errors, higher statistical significance is typically associated with larger effects. Consequently, selection procedures based on p-values or Clfdr statistics tend to automatically choose large effects. In such cases, practitioners are advised to follow conventional practice with the existing ETP notion (2.3). However, when units demonstrate high levels of heteroscedasticity, conventional practice tends to over-represent subgroups with lower variances. This outcome is undesirable as it may lead to the selection of small effects with minimal practical value. In such settings, we strongly recommend utilizing our modified power criterion ETP* as a preferred alternative. Next we provide an example to show that by assigning a higher reward to the selection of larger effects, the new formulation provides a more principled and balanced approach to the selection problem.

This illustrative example compares two oracle rules designed to maximize the conventional power (2.3) and modified power (2.4), respectively. The fundamental operational

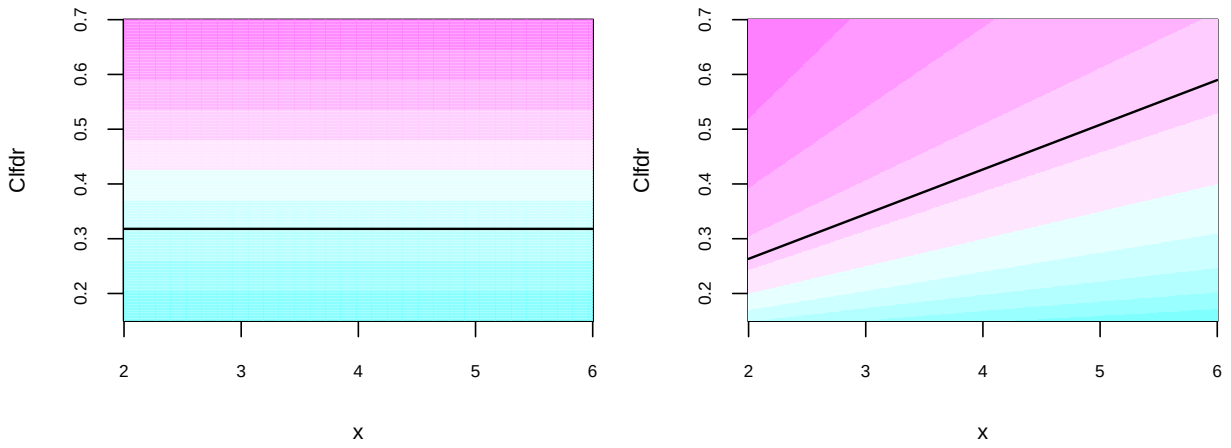

Figure 2: Left: heat map for Clfdr values. Right: heat map for T values. In both panels, the x-axis and y-axis represent raw observations X and Clfdr values, respectively. The rejection regions of the oracle rules $\boldsymbol{\delta}^C$ and $\boldsymbol{\delta}^T$ are the areas under the corresponding black lines.

units for the two oracle rules are Clfdr statistic and T, defined by (2.6) and (2.8), respectively. Suppose we are interested in testing $H_{0,i} : \mu_i \leq 0$, $i \in [m]$ based on data generated from the following model:

$$\theta_i \overset{i.i.d.}{\sim} \text{Bernoulli}(0.2), \quad \mu_i \overset{ind}{\sim} (1-\theta_i)U(-3,-1)+\theta_i U(1,2), \quad \sigma_i \overset{iid}{\sim} U(0.5,3), \quad X_i \overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma_i^2).$$

The oracle rule that maximizes the ETP in (2.3) is defined as $\boldsymbol{\delta}^C = (\delta_i^C : i \in [m])$, where $\delta_i^C = \mathbb{I}(\text{Clfdr}_i < c_\alpha)$ and c_α is determined by the desired FDR level α. For the oracle rule $\boldsymbol{\delta}^T = (\delta_i^T : i \in [m])$ that maximizes the ETP* in (2.4), Group 2 is empty, and the oracle rule for units in Group 1 is $\delta_i^T = \mathbb{I}(T_i > t_\alpha)$. Assuming known distributional information, we can determine $c_\alpha = 0.32$ and $t_\alpha = 12.21$ through numerical approximations such that the FDR levels of both oracle rules are exactly controlled at $\alpha = 0.1$.

Figure 2 illustrates the rejection regions of the oracle rules $\boldsymbol{\delta}^C$ and $\boldsymbol{\delta}^T$, depicted by the corresponding black lines. The left panel of the figure displays the Clfdr values via a heat map, with the x-axis representing raw observations X, and the y-axis representing Clfdr values. In the right panel, we present the heat map for T values, with the x-axis and

y-axis representing X and Clfdr values, respectively. It is worth noting that the new oracle rule $\boldsymbol{\delta}^T$ yields a rejection region where the threshold for Clfdr increases as X increases, which contrasts with the oracle rule $\boldsymbol{\delta}^C$ that uses a fixed Clfdr threshold. This observation indicates that the modified ETP* results in a selection rule that factors in both statistical significance and effect size simultaneously.

3 Data-driven procedure

This section presents the development of our data-driven procedure, including a non-parametric deconvoluting method (Section 3.1), a computational shortcut (Section 3.2), and a theoretical analysis (Section 3.3).

3.1 Nonparametric deconvolution

We propose a non-parametric g-modeling approach to estimating $g_\mu(\cdot)$, which plays a critical role in computing the value of T_i. Although prior research by Efron (2016), Gu and Koenker (2017b) and Gu and Koenker (2023) has tackled this issue, the theoretical properties of these methods remain largely unknown. Our new g-modeling method, which is based on the density matching idea, offers a fast and stable algorithm that performs comparably to competing methods, while having a form that greatly simplifies the theoretical analysis of the data-driven procedure.

Assume $supp(g_\mu) \subset [-M, M]$, $M < \infty$. The g-modeling approach (Jiang and Zhang, 2009; Koenker and Mizera, 2014) involves approximating $g_\mu(\cdot)$ using a mixture of point masses. We form a grid of size k evenly spaced between $-M$ and M:

$$\{s, s+\eta, s+2\eta, ..., s+(k-1)\eta\},$$

where $\eta = 2M/(k-1)$ and $s = -M$. Then $g_\mu(\cdot)$ can be approximated by $\hat{g}_\mu(\cdot) =$

$\sum_{j=1}^{k} w_j I_{s+(j-1)\eta}(\cdot)$, where $I_c(\cdot)$ is the Dirac delta function centered at c. The task at hand then boils down to determining the optimal weights $\boldsymbol{w} = (w_1, w_2, ..., w_k)$, which can be efficiently solved through a direct optimization approach.

We outline a "density matching" approach for formulating the optimization objective function. Specifically, two different techniques are employed to derive the density estimate, namely, $\hat{f}$, which is constructed based on a given $\hat{g}$, and $\hat{f}^m$, which is constructed using a weighted bivariate kernel estimator. The objective function is designed to ensure a high degree of similarity between the two density estimators.

First, upon obtaining $\hat{g}$, a natural estimate for $f_i(x_i)$ is readily provided by

$$\hat{f}_i(x) = \int_{-\infty}^{\infty} \phi_{\sigma_i}(x-y)\hat{g}_\mu(y)dy = \sum_{j=1}^{k} w_j \phi_{\sigma_i}\left\{x - s - (j-1)\eta\right\}.$$

Also, f_i can be estimated by employing a weighted bivariate kernel estimator:

$$\hat{f}_i^m(x) = \sum_{j=1}^{m} \frac{\phi_{h_\sigma}(\sigma_i - \sigma_j)}{\sum_{k=1}^{m} \phi_{h_\sigma}(\sigma_i - \sigma_k)} \phi_{h_{xj}}(x - x_j),$$

where $\boldsymbol{h} = (h_x, h_\sigma)$ is a pair of bandwidths, $\phi_{h_\sigma}(\sigma - \sigma_j)/\{\sum_{j=1}^{m} \phi_{h_\sigma}(\sigma - \sigma_j)\}$ that determine the contribution of (x_j, σ_j) based on σ_j, $h_{xj} = h_x\sigma_j$ is a bandwidth that varies across j, and $\phi_h(z) = (1/\sqrt{2\pi h^2}) \exp\{-z^2/(2h^2)\}$ is a Gaussian kernel. The motivation behind this approach is to leverage the smooth variation of $f_i(x_i)$ with respect to σ_i. The variable bandwidth h_{xj} is utilized to account for the heteroscedasticity inherent in the data, resulting in data points with higher variation being associated with flatter kernels.

To minimize the discrepancy between $\hat{f}_i$ and $\hat{f}_i^m$, we aim to find $\boldsymbol{w}$ that solves the following convex optimization problem:

$$\text{Minimize} \quad \sum_{i=1}^{m} \{\hat{f}_i(x_i) - \hat{f}_i^m(x_i)\}^2 \quad \text{subject to} \quad w_j \geq 0 \text{ for } 1 \leq j \leq k \text{ and } \sum_{j=1}^{k} w_j = 1. \tag{3.12}$$

Denote $\boldsymbol{w}^* = (w_1^*, \ldots, w_k^*)$ the optimizer of (3.12), then $\hat{f}_{0i}$ and $\hat{f}_i$ can be computed as

$$\begin{aligned}\hat{f}_{0i}(x) &= \int_{-\infty}^{\mu_0} \phi_{\sigma_i}(x-\mu)\hat{g}_\mu(\mu)d\mu = \sum_{s+(j-1)\eta\le\mu_0} w_j^*\phi_{\sigma_i}(x-s-j\eta) \text{ and} \\ \hat{f}_i(x) &= \int_{-\infty}^{\infty} \phi_{\sigma_i}(x-\mu)\hat{g}_\mu(\mu)d\mu = \sum_{j=1}^{k} w_j^*\phi_{\sigma_i}(x-s-(j-1)\eta).\end{aligned}$$

Finally, T_i can be estimated correspondingly using a plug-in method.

3.2 A computational shortcut and the step-wise algorithm

The oracle rule necessitates a search over a two-dimensional space for identifying the optimal cutoffs (c_1^{OR}, c_2^{OR}) defined in (2.10). This task can be computationally demanding. To overcome this challenge, we propose in this section a computational shortcut that leads to a considerable improvement in computational efficiency.

To maximize $\text{ETP}^*(c_1, c_2)$ for a given c_2, our strategy must reject as many hypotheses as possible from group 1. This involves the selection of the smallest c_1 that satisfies the mFDR constraint. Consequently, the optimal solution (c_1^{OR}, c_2^{OR}) must be located on the one-dimensional curve $L(c_2) = \{(c_1^*(c_2), c_2)\}$, where $c_1^*(c_2) = \inf\{c_1 : \text{mFDR}_{\boldsymbol{\delta}(c_1,c_2)} \le \alpha\}$. The problem boils down to determining the optimal c_2 on $L(c_2)$ such that $\text{ETP}^*(c_2; L) \equiv \text{ETP}^*(c_1^*(c_2), c_2)$ can be maximized. The following proposition establishes that if $\text{ETP}^*(c_2; L)$ starts to decrease along the curve $L(c_2)$ in the direction of increasing c_2, then it will continue to decrease in the direction of increasing c_2.

Proposition 1. *Consider three decision rules $\boldsymbol{\delta} = \boldsymbol{\delta}(c_1, c_2)$, $\boldsymbol{\delta}' = \boldsymbol{\delta}(c_1', c_2')$, $\boldsymbol{\delta}'' = \boldsymbol{\delta}(c_1'', c_2'')$ with (c_1, c_2), (c_1', c_2'), (c_1'', c_2'') all on L. If $c_1 \ge c_1' \ge c_1''$ and $ETP_{\boldsymbol{\delta}}^* \ge ETP_{\boldsymbol{\delta}'}^*$, then we must have $ETP_{\boldsymbol{\delta}'}^* \ge ETP_{\boldsymbol{\delta}''}^*$.*

Proposition 1 inspires us to adopt the following strategy: search along the curve L in the direction of increasing c_2 and stop when ETP^* begins to decrease. More precisely, we

first select as many units as possible from group 1 and record the resulting ETP* (Step 3). Next, we select a single hypothesis from group 2 (Step 4), which decreases the ETP* but increases the FDR capacity. We then return to Step 3 and select as many units as possible from group 1 using the additional FDR capacity and record the new ETP*. We compare the new ETP* with the previous ETP*. If the ETP* increases after the iteration, we repeat the aforementioned process (e.g. continue to Step 4 and return to Step 3), otherwise we stop the procedure and output the thresholds. The operation of the step-wise data-driven procedure is detailed in Algorithm 1. For the more general problem described in (2.11) we need only replace $\hat{T}_i$ with $\frac{h_i(\boldsymbol{X},\boldsymbol{\sigma})}{\widehat{\text{Clfdr}}_i-\alpha}$ in Algorithm 1. More details are given in the Supplementary Material A.4.

Algorithm 1: The data-driven procedure

Input: $\mathbf{x}$, $\widehat{\mathbf{Clfdr}}$, α.

Output: The estimated threshold for group 1 and group 2 ($\hat{c}_1$ and $\hat{c}_2$).

Step 1: Compute $\hat{T}_i = (x_i - \mu_0)/(\widehat{\text{Clfdr}}_i - \alpha)$. Form the 4 groups described in the oracle procedure using $\widehat{\mathbf{Clfdr}}$ and $\hat{\boldsymbol{T}}$ in place of $\mathbf{Clfdr}$ and $\boldsymbol{T}$.

Step 2: Let $\mathcal{R}$ denote the rejection set. Put the indices of hypotheses from group 0 into $\mathcal{R}$. Rank hypotheses in group 1 from largest to smallest according to $\hat{T}_i$. Rank hypotheses in group 2 from smallest to largest according to $\hat{T}_i$.

Step 3: Denote the ranked hypotheses in group 1 by $H^1_{(1)}, H^1_{(2)}, ...$ and the corresponding Clfdr values by $\text{Clfdr}_{(1)}, \text{Clfdr}_{(2)}, ...$ Let $k = \max\{j : \sum_{i=1}^{j}(\text{Clfdr}_{(i)} - \alpha) \leq -\sum_{i\in\mathcal{R}}(\text{Clfdr}_i - \alpha)\}$, reject $H_{(1)}, H_{(2)}, ...H_{(k)}$ and remove them from group 1. Compute and store ETP$^* = \sum_{i\in\mathcal{R}}(x_i - \mu_0)$.

Step 4: Denote the ranked hypotheses in group 2 by $H^2_{(1)}, H^2_{(2)}, ...$ Reject $H^2_{(1)}$ and remove it from group 2.

Step 5: Repeat step 3 and step 4. Terminate when ETP* starts to decrease or when either group 1 or group 2 is empty.

Step 6: Let $(\hat{c}_1, \hat{c}_2)$ to be the pair that maximizes ETP* and set $\boldsymbol{\delta}^{DD} = \boldsymbol{\delta}(\hat{c}_1, \hat{c}_2)$.

3.3 Theoretical properties of the data-driven procedure

In Section 2.2, we demonstrated that the oracle rule $\boldsymbol{\delta}^{OR}$ is both valid and optimal in the sense that it satisfies the FDR constraint and has the largest ETP* among all valid FDR rules. In this subsection, we aim to establish that the data-driven procedure $\boldsymbol{\delta}^{DD}$, defined in Algorithm 1, asymptotically approaches the performance of the oracle rule $\boldsymbol{\delta}^{OR}$ and therefore is asymptotically valid and optimal. Before we proceed with our theoretical analysis, we state the following regularity conditions.

(A1) $supp(g_\mu) \subset [-M, M]$ and $supp(g_\sigma) \in (M_1, M_2)$ for some $M_1 > 0$, $M_2 < \infty$, $M < \infty$.

(A2) The bandwidths $\boldsymbol{h} = (h_x, h_\sigma)$ satisfy $h_x \sim m^{-\eta_x}$, $h_\sigma \sim m^{-\eta_s}$ where η_x and η_s are small positive constants such that $0 < \eta_s + \eta_x < 1$.

(A3) The grid size satisfies $k \sim m^{1/3} \log m$.

Remark 2. *Assumption (A1) is a mild condition on boundedness of g_μ and g_σ, which is reasonable for most practical scenarios. Assumption (A2) is satisfied by commonly used bandwidth choices in Wand and Jones (1994). Assumption (A3) can be achieved by user's choice, and can be relaxed to $k \to \infty$ as $m \to \infty$. In theory, a larger k will not harm the quality of the deconvolution estimate, but it may lead to longer computational times. In Section C, we argue that the grid size k need not be of order greater than $m^{1/3} \log m$.*

We first state a crucial proposition that establishes the theoretical properties of the proposed density estimator $\hat{f}_{0i}$.

Proposition 2. *Suppose condition (A1), (A2), and (A3) hold, then $\widehat{Clfdr}_i \xrightarrow{p} Clfdr_i$ when $m \to \infty$.*

Now we present our theory on the asymptotic validity and optimality of the data-driven procedure.

Theorem 2. *Under Conditions (A1), (A2), and (A3) the data-driven procedure* $\boldsymbol{\delta}^{DD}$ *described in Algorithm 1 controls mFDR at level* $\alpha + o(1)$ *and* $ETP^*_{\boldsymbol{\delta}^{DD}}/ETP^*_{\boldsymbol{\delta}^{OR}} = 1 + o(1)$ *as* $m \to \infty$.

4 The R-Value in Multiple Comparisons

In this section, we investigate the integration of ranking and selection in a unified multiple comparison framework. Our proposed approach involves generating a ranking based on a suitable selection rule and utilizing a novel ranking metric called the r-value. This metric reflects the relative order in which different units are selected, thereby providing a practical criterion for assessing the relative importance of the units within a list.

We present two r-value notions. The first, presented in Section 4.1, assumes a fixed reference level μ_0, with the r-values generated by varying the confidence level α denoted as r_α. The second, presented in Section 4.2, assumes a fixed confidence level α, with the r-values generated by varying the reference level μ_0, denoted as r_{μ_0}. Important properties of the r-values are investigated in Section 4.4.

4.1 R-values generated by varying the confidence level

The conventional multiple testing framework relies on a thresholding procedure that assumes the presence of a significance index, such as the p-value or local false discovery rate, which provides a consistent ranking of study units that remains invariant across all FDR levels. However, in the case of a heteroscedastic setup, such a ranking cannot be provided. For instance, in the oracle rule (2.9), study units may be selected into the rejection set in different orders at varying FDR levels since the optimal statistic T depends on α. As a result, the ranking would be inconsistent across different users, who select different FDR

levels in their analysis. Furthermore, there is no natural order for the subjects in group 0, as all units in the whole group are selected simultaneously.

Next we propose the pivotal notion of the r-value, which has the ability to convert any selection procedure that controls the error probability into a meaningful and coherent ranking metric.

Definition 1. *Let $\mathcal{R}_\alpha^{\mathcal{D}}$ denote the set of units selected by a pre-defined selection procedure $\mathcal{D}$ that controls the error rate at level α. The r_α-value of a unit $i \in [m]$ linked with $\mathcal{D}$ is defined as*

$$r_{i,\alpha} = \inf\{\alpha : i \in \mathcal{R}_\alpha^{\mathcal{D}}\}. \tag{4.13}$$

Remark 3. *In Supplementary Material D, we demonstrate that the r-value defined by* (4.13) *encompasses the conventional p-value and q-value as special cases, provided that meaningful error concepts and their corresponding selection procedures are appropriately employed.*

When combined with the novel prioritized selection procedure that solves (2.5), the r-value corresponds to the minimum FDR level at which a study unit can be selected. This ranking metric addresses the inconsistency issue that may arise from the subjective specification of α values, offering an objective and consistent means of ranking across different users.

4.2 R-values generated by varying the reference level

The specification of μ_0 in practical scenarios hinges on prior domain expertise, which may be subjective and vary among users. Since the oracle statistic T (2.8) and the corresponding data-driven quantity are contingent on μ_0, divergent selections of μ_0 among analysts may yield inconsistent rankings. Assuming a consensus on the choice of the confidence parameter

α (e.g., 0.05), it is possible to generate r-values by varying the reference level μ_0. Suppose we vary μ_0 from ∞ to $-\infty$, then the earlier a unit is selected, the more important it is considered to be relative to the other units – thus an objective ranking of study units is generated.

Definition 2. *Let $\mathcal{R}^{\mathcal{D}}_{\mu_0}$ denote the set of units selected by a pre-defined selection procedure $\mathcal{D}$ that aims to select units with effect size larger than μ_0. The r_{μ_0}-value of a unit $i \in [m]$ associated with $\mathcal{D}$ is defined as follows:*

$$r_{i,\mu_0} = \sup\{\mu_0 : i \in \mathcal{R}^{\mathcal{D}}_{\mu_0}\}, \text{ or } r'_{i,\mu_0} = \frac{1 + \sum_{j \neq i} \mathbb{I}(r_j > r_i)}{m},$$

provided that no ties exist between r_i's.

Here, $r'_{i,\mu_0} \in \{i/m : i \in [m]\}$ is the standardized rank taking values in $(0, 1]$, which is suitable in situations where only the relative position of the study units is relevant. The non-standardized r_{μ_0}-value of a particular unit i corresponds to the largest predetermined reference value μ_0 at which unit i can be selected with confidence.

4.3 Which r-value to use

By integrating our r-value with the prioritized selection framework (2.5), we have developed a solution that is both intuitive and logically coherent for the challenging problem of ranking and selection under heteroscedastic setups. Both definitions of r-value attempt to strike a balance between statistical significance and practical relevance. However, the two nuanced concepts prioritize different aspects in the ranking process: r_α primarily aims to select the most prominent effects in terms of statistical significance, with a secondary objective of ensuring the practical relevance of the selected effects. Conversely, r_{μ_0} prioritizes selecting the most prominent effects in terms of observed magnitudes, with a secondary objective of ensuring reliable control over the uncertainty of the selection.

The choice between the two r-values hinges primarily on the specific practical goals of the study. For example, during the analysis of the AYP data (cf. Section 1.1 and Section F.7 of the Supplementary Material), as we were aware of impacts from social and economic discrepancies, we anticipated that there would always be some difference in academic performance between the SEA and SED groups. Our main goal was to identify schools where the observed gaps were most pronounced, while taking into consideration the associated heterogeneous variabilities. Consequently, the objective of r_{μ_0} was more closely aligned with our practical needs, as it improves the ability to allocate limited resources and budgets to schools that require the most assistance. By adopting r_{μ_0}, we can also obviate the need to establish a suitable μ_0, which can be subjective in practice due to, say, the variability in passing rates between schools and the absence of a clear or meaningful benchmark difference. By contrast, in high-stakes situations where minimizing decision risk or controlling the error probability is paramount, it may be appropriate to consider using r_α as it prioritizes making the safest choice.

4.4 Agreeability of ranking

The proposed framework of "selection to ranking" offers an appealing alternative to the conventional approach of "rank and then select," which is impractical in the presence of heteroscedastic data. For example, Definition 1 first tackles the selection issue through constrained optimization, resulting in an objective solution for any given α. Next, the ranking issue is handled using the r_α-value, which is determined by sequentially adjusting the selection level α without any user input. Consequently, the needs for a universally applicable test statistic and a potentially subjective choice of α can be eliminated, thus preventing the issue of inconsistent rankings.

To demonstrate the appropriateness of the ranking generated by r-values, we introduce the concept of *agreeability*. As previously mentioned, the ranking in heteroscedastic scenarios must consider two factors: effect size (captured by X) and statistical significance (captured by Clfdr statistic or its estimate $\widehat{\text{Clfdr}}$). The following theorem asserts that if unit i dominates unit j in terms of both effect size and statistical significance, then the use of the r-value ensures that unit i will be ranked higher than unit j.

Theorem 3. *Let r_i and $\hat{r}_i$ be the r-values produced by the oracle procedure* (2.10) *and the data-driven procedure (Algorithm 1), respectively, for $i \in [m]$. Then both the oracle and data-driven procedures are agreeable in the sense that if $X_i > X_j$ and $\text{Clfdr}_i < \text{Clfdr}_j$ (or $\widehat{\text{Clfdr}}_i < \widehat{\text{Clfdr}}_j$), then $r_i < r_j$ (or $\hat{r}_i < \hat{r}_j$). This assertion holds true for both Definition 1 and Definition 2 with r'_{i,μ_0}.*

Remark 4. Agreeability can be seen as a less stringent version of the nestedness notion. Gu and Koenker (2023) explore the notion of nestedness in ranking and selection, while Henderson and Newton (2016) suggest some potential issues regarding the nestedness requirement in the presence of heteroscedasticity. In Section E of the Supplementary Material, we precisely define the nestedness property and present counterexamples to demonstrate why nested selection may be infeasible under heteroscedastic setups.

5 Numeric experiments

We begin by presenting the implementation details of the data-driven procedure in Section 5.1. In Section 5.2, we investigate the performance of the prioritized selection procedure and compare it with competing methods in a scenario where both $g_\sigma(\cdot)$ and $g_\mu(\cdot)$ are continuous. In Section 5.3, we present additional results for the case where both $g_\sigma(\cdot)$ and $g_\mu(\cdot)$ are discrete, and where μ_i is correlated with σ_i. In all of our experiments, the FDR

and mFDR levels are numerically close. Therefore, we report only the more commonly used FDR levels. Simulation results in other scenarios, such as when σ_i must be estimated from the data, when the number of test is small, when the observations are weakly dependent, and when the target FDR levels vary, as well as a comparative analysis of the ETP* and ETP, are presented in Section F of the online Supplementary Material.

5.1 Some implementation details

The nonparametric deconvolution method discussed in Section 3.1 requires the estimation of $\hat{f}_i^m(x_i)$, $i \in [m]$, which involves specifying the tuning parameters $\boldsymbol{h} = (h_x, h_\sigma)$. In our analysis, we have employed the rule of thumb in Silverman (1986), given by $h_x = 0.9 \min\{\mathrm{sd}(\boldsymbol{x}), \mathrm{IQR}(\boldsymbol{x})\}/(1.34m^{1/5})$ and $h_\sigma = 0.9 \min\{\mathrm{sd}(\boldsymbol{\sigma}), \mathrm{IQR}(\boldsymbol{\sigma})\}/(1.34m^{1/5})$, where $\mathrm{sd}(\cdot)$ and $\mathrm{IQR}(\cdot)$ are the standard deviation and interquartile range of the input vector, respectively. In all our simulation studies, we use a grid size k of 50.

To ensure numerical stability, we recommend selecting the support of the grid to be $[\hat{F}^{-1}(0.01), \hat{F}^{-1}(0.99)]$, where $\hat{F}^{-1}(\cdot)$ represents the empirical quantile function of X_i. We solve the convex optimization problem (3.12) using the `CVXR` package in R (Fu et al., 2020). The source code for reproducing all the numerical results in this paper is available on our GitHub repository at `https://github.com/bgang92/rankingselection`.

5.2 Comparison for independent μ_i and σ_i

Next, we consider the following setting:

$$\theta_i \overset{iid}{\sim} \mathrm{Ber}(0.2), \quad \mu_i|\theta_i \sim (1-\theta_i)U(-3,-1) + \theta_i U(1,2), \quad \sigma_i \overset{iid}{\sim} U(0.5, \sigma_{max}),$$

$$X_i|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i^2), \quad i \in [5000].$$

We aim to test the hypotheses $H_{0,i} : \mu_i \leq \mu_0$ versus $H_{a,i} : \mu_i > \mu_0$, with $\mu_0 = 0$. In

addition to the three methods compared in Section F.1, we also incorporate the widely used Benjamini-Hochberg procedure (BH) procedure in the comparison. The p-values are computed as $1 - \Phi\{(X_i - \mu_0)/\sigma_i\}$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal variable. The nominal FDR level is set to $\alpha = 0.1$, while σ_{max} varies from 2 to 4 for different settings. The results are obtained by averaging the results in 100 replications and are presented in Fig 3. We can observe two important patterns. Firstly, BH appears to be excessively conservative, suggesting that p-value based methods may not be well-suited for testing composite hypotheses. Secondly, DD, OR, and Clfdr exhibit comparable levels of FDR but display noticeable differences in their ETP* and ETP values.

A more detailed comparison of the hypotheses rejected by DD and Clfdr underscores the marked differences between these two methods. In Fig 4 (a), we look at one particular run with $\sigma_{max} = 4$. The gray dots are hypotheses not rejected by either DD or Clfdr, green dots are hypotheses rejected by both DD and Clfdr, red dots are hypotheses rejected by DD but not Clfdr, and blue dots are hypotheses rejected by Clfdr but not DD.

Upon close examination, it is evident that DD is more likely to reject hypotheses with higher x_i values when compared to Clfdr. If we exclude the hypotheses that are rejected by both DD and Clfdr and assess the ETP* for the remaining hypotheses, a distinct contrast emerges, as depicted in Figure 4 (b). For the hypotheses that are rejected by only one method, DD has a superior ETP* in comparison to Clfdr. Additionally, the difference in ETP* becomes more pronounced as the degree of heteroscedasticity increases.

5.3 Comparison for correlated μ_i and σ_i

In this section, we present simulation results in a more complex scenario where σ_i and μ_i are correlated. Let I_c be an indicator function that takes the value of 1 at c and 0 elsewhere.

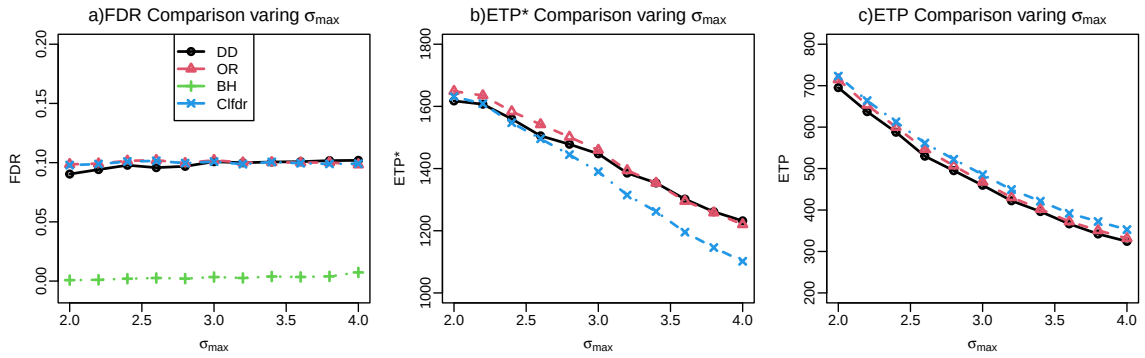

Figure 3: Comparison when σ_i and μ_i are uncorrelated and both are generated from a uniform distribution. We vary σ_{max} from 2 to 4. All methods control the FDR at the nominal level with BH being overly conservative.

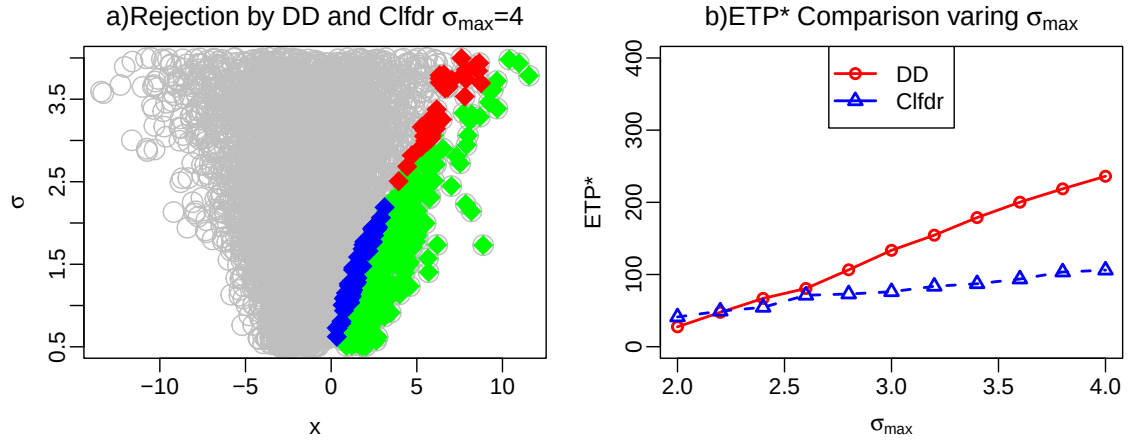

Figure 4: (a): A scatter plot of the hypotheses when $\sigma_{max} = 4$. The gray circles are hypotheses rejected by neither DD or Clfdr, green dots are hypotheses rejected by both DD and Clfdr, red dots are hypotheses rejected by DD but not Clfdr, blue dots are hypotheses rejected by Clfdr but not DD. (b): ETP* comparison of hypotheses rejected by either DD or Clfdr, but not both.

The model first generates σ_i from two groups and then generates μ_i in a manner that is dependent on σ_i, as described below:

$$\begin{aligned} X_i|\mu_i,\sigma_i &\sim N(\mu_i,\sigma_i^2), \quad \sigma_i \overset{iid}{\sim} \frac{1}{2}I_{0.25\sigma}(\cdot)+\frac{1}{2}I_{1.25\sigma}(\cdot), \\ \mu_i|\sigma_i=0.25\sigma &\sim 0.9N(-0.5,0.25^2)+0.1N(1.5,0.25^2), \\ \mu_i|\sigma_i=1.25\sigma &\sim 0.9N(-0.5,0.25^2)+0.1N(3,0.25^2). \end{aligned}$$

The hypotheses to be tested in our study are $H_{0,i} : \mu_i \leq \mu_0$ vs $H_{a,i} : \mu_i > \mu_0$, where $\mu_0 = 1, i \in [10000]$. The value of σ varies between 1.5 to 2 for different settings. It is

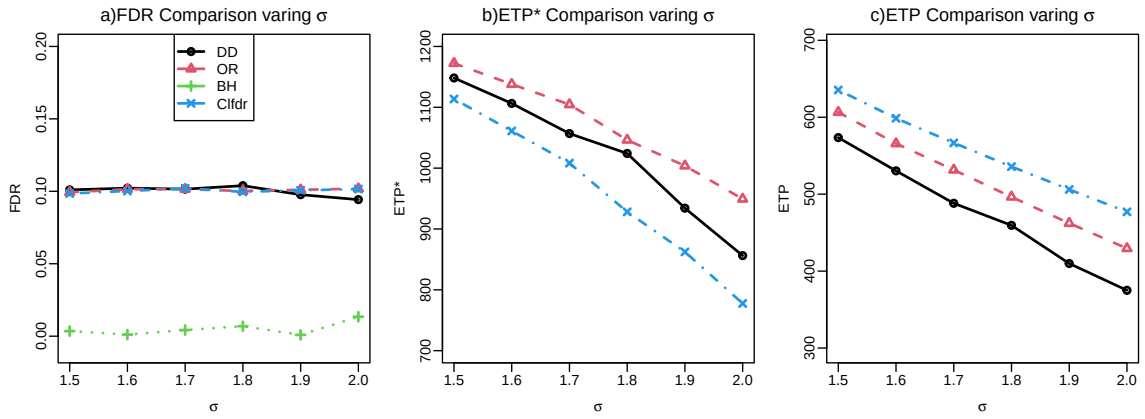

Figure 5: The comparison under the setting where σ_i and μ_i are correlated. Under the ETP* metric, DD and OR outperform Clfdr. Conversely, under the ETP metric, Clfdr exhibits a higher power than DD and OR.

important to note that in our design, μ_i is sampled from a mixture distribution, where a vast majority of μ_i values are generated from $N(-0.5, 0.25^2)$, corresponding to small effects. However, there is a small fraction of μ_i values that correspond to large effects. Furthermore, the effect sizes μ_i tend to increase as the variance becomes larger.

We conduct experiments on 100 datasets and apply DD, OR, Clfdr, and BH to select important units at FDR level $\alpha = 0.1$. The accuracy of the deconvolution method relies on the independence between σ_i and μ_i. Therefore, we initially partitioned the data into two groups based on whether $\sigma_i = 0.25\sigma$ or $\sigma_i = 1.25\sigma$, and estimated $g_\mu(\cdot)$ separately for each group. A summary of results for different values of μ is presented in Figure 5.

Our analysis reveals several patterns. Firstly, all methods maintain FDR control at the nominal level. Secondly, BH is excessively conservative, resulting in ETP* and ETP values that are significantly lower than the other three methods. Therefore, we exclude BH from the ETP* and ETP plots. Thirdly, DD and OR outperform Clfdr in terms of the ETP* criterion, whereas Clfdr outperforms DD and OR regarding the ETP criterion. To make a further comparison between Clfdr and DD, we present a scatter plot of rejected hypotheses when $\sigma = 2$ in Figure 6. In Figure 6 (a), the hypotheses rejected by DD but not Clfdr

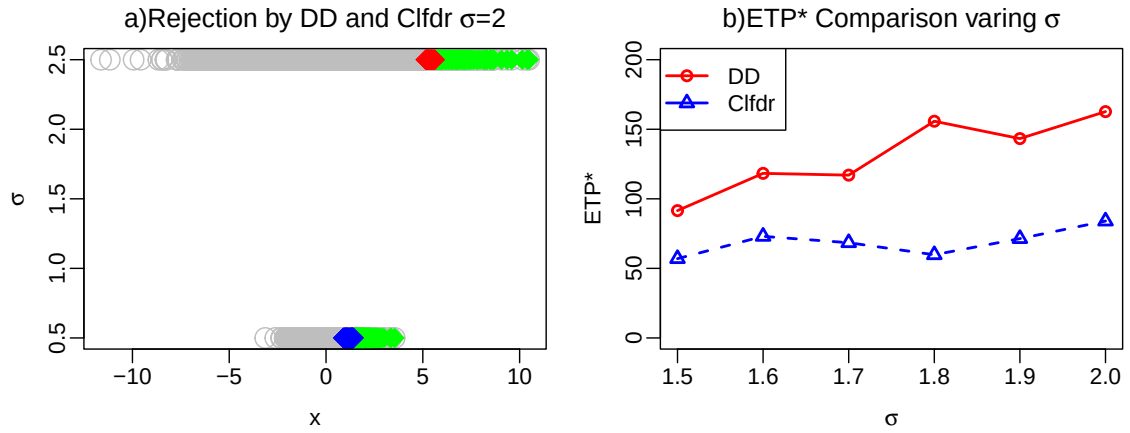

Figure 6: (a): A scatter plot of the hypotheses when $\sigma = 2$. The x-axis represents x_i and the y-axis represents σ_i. The gray circles are hypotheses rejected by neither DD or Clfdr, green dots are hypotheses rejected by both DD and Clfdr, red dots are hypotheses rejected by DD but not Clfdr, blue dots are hypotheses rejected by Clfdr but not DD. (b): ETP* comparison on hypotheses not rejected by both DD and Clfdr.

all have $\sigma_i = 2$. This implies that DD is more sensitive to larger variances than Clfdr. In Figure 6 (b), we observe that DD has a significantly higher ETP* on hypotheses where Clfdr and DD disagree.

6 Real Data Applications

In this section, we analyze mutual fund data obtained from CRSP accessed via the Wharton WRDS database at the University of Pennsylvania. The goal is to compare our ranking and selection method against Clfdr and BH that are solely based on significance indices. The analysis of the test performance data of K-12 schools from the 2005 Annual Yearly Performance (AYP) study is provided in Section F.7 of the Supplementary Material.

We analyze the estimated returns of mutual funds, which are denoted as x_i, over an average of 31 months of performance from the end of April 2006 to the end of October 2008. These estimated returns are obtained from the intercept term of Carhart's four-factor model (Carhart, 1997). The standard error of the returns, denoted as s_i, is computed as

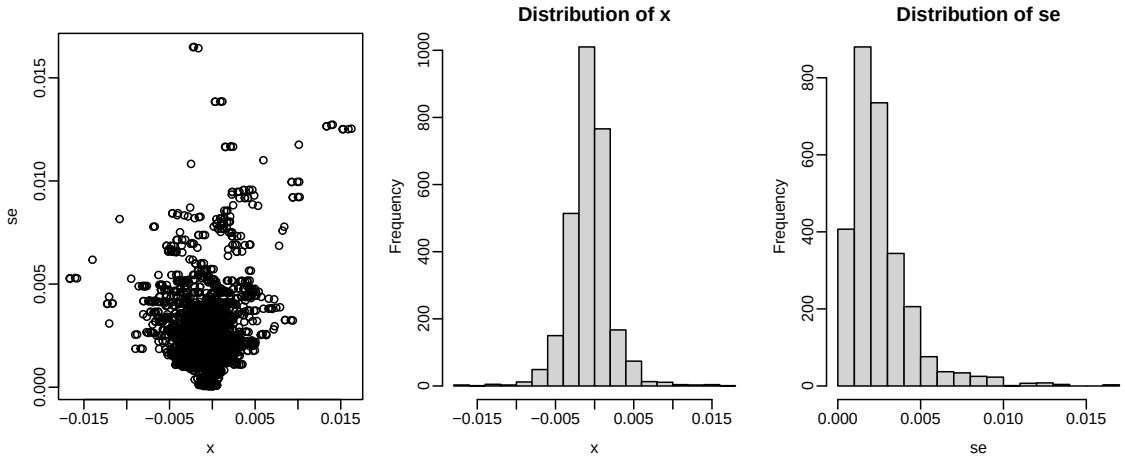

Figure 7: Left: scatter plot of the CRSP data: x-axis is the observation, y-axis is the standard error. Middle: histogram of the observations. Right: histogram of the standard errors.

the estimated standard error of the intercept term in the model. The dataset comprises 2796 pairs of observations (x_i, s_i). To ensure numerical stability, we exclude observations with standard errors below the 0.1% and above the 99.9% percentiles. Figure 7 visually depicts the distribution of the data.

Our objective is to identify mutual funds that exhibit positive returns. Therefore, we consider the following hypotheses: $H_{0,i} : \mu_i \leq \mu_0$ vs $H_{a,i} : \mu_i > \mu_0$, with $\mu_0 = 0$. We set the target FDR level at 0.1. The z-values and corresponding p-values are obtained as before.

The results, presented in Table 1, reveal that Clfdr rejects more hypotheses than DD. However, Clfdr exhibits negative modified power, demonstrating its tendency to select portfolios with negative returns. This phenomenon occurs because Clfdr does not consider the value of the returns, leading it to select hypotheses with estimated returns (x_i) slightly lower than the null hypothesis (μ_0), which corresponds to negative true returns. Clfdr selects such units because they do not increase the overall FDR above the target level. In practice, this type of "over-selection" is often undesirable, as demonstrated in this example.

We proceed by examining the units selected by DD but not Clfdr and vice versa. The

Table 1: Summary of power by each method on the CRSP data.

	DD	Clfdr	BH
Number of hypotheses rejected	491	546	46
Modified Power	1.307	1.060	0.024

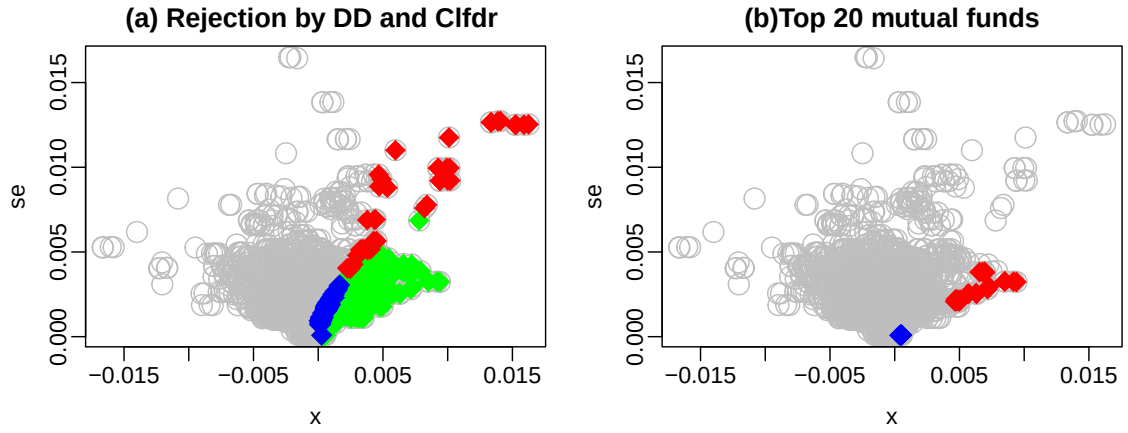

Figure 8: (a) The x-axis represents observations (x), while the y-axis corresponds to the SEs. The gray circles denote funds that were not selected by either DD or Clfdr. The green dots represent funds selected by both Clfdr and DD, while the blue dots are funds selected by Clfdr but not DD and the red dots are funds selected by DD but not Clfdr. (b) The scatter plot displays the top 20 mutual funds ranked according to the p-value and r-value (Definition 2 with $\alpha = 0.1$). The top 20 mutual funds ranked according to the r-value are denoted by red dots, while the top 20 mutual funds ranked according to the p-value are shown as blue dots.

results are presented in Figure 8 (a). We observe that the units selected by DD and Clfdr differ significantly. DD selects units with both high estimated returns (x_i) and high SEs, indicating its tendency to trade high variability for potentially high returns. Figure 8 (b) displays the top 20 mutual funds ranked according to p-values (blue dots) and r_{μ_0}-values (red dots). The results indicate that the r_{μ_0}-value places higher priority on selecting funds with higher returns, whereas the p-value favors the selection of funds with smaller SEs.

References

Banerjee, T., L. J. Fu, G. M. James, and W. Sun (2020). Nonparametric empirical bayes estimation on heterogeneous data. *arXiv preprint arXiv:2002.12586*.

Basu, P., T. T. Cai, K. Das, and W. Sun (2018). Weighted false discovery rate control in

large-scale multiple testing. *Journal of the American Statistical Association 113*(523), 1172–1183. PMID: 31011234.

Bechhofer, R. E. (1954). A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with known Variances. *The Annals of Mathematical Statistics 25*(1), 16 – 39.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society series b-methodological 57*(1), 289–300.

Benjamini, Y. and Y. Hochberg (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics 25*(1), 60–83.

Boyd, S., C. Cortes, M. Mohri, and A. Radovanovic (2012). Accuracy at the top. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 25. Curran Associates, Inc.

Cai, T. T. and W. Sun (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association 104*(488), 1467–1481.

Cai, T. T., W. Sun, and W. Wang (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of The Royal Statistical Society Series B-statistical Methodology 81*, 187–234.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance 52*(1), 57–82.

Chen, C.-H., J. Lin, E. Yücesan, and S. E. Chick (2000, jul). Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems 10*(3), 251–270.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Volume 1. Cambridge University Press.

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

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association 96*(456), 1151–1160.

Foster, D. P. and R. A. Stine (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(2), 429–444.

Fu, A., B. Narasimhan, and S. Boyd (2020). CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software 94*(14), 1–34.

Fu, L., B. Gang, G. M. James, and W. Sun (2022). Heteroscedasticity-adjusted ranking and thresholding for large-scale multiple testing. *Journal of the American Statistical Association 117*(538), 1028–1040.

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

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 499–517.

Genovese, C. and L. Wasserman (2004, 06). A stochastic process approach to false discovery control. *Ann. Statist. 32*(3), 1035–1061.

Goel, P. K. and H. Rubin (1977). On selecting a subset containing the best population-a bayesian approach. *The Annals of Statistics 5*(5), 969–983.

Gu, J. and R. Koenker (2017a). Empirical bayesball remixed: Empirical bayes methods for longitudinal data. *Journal of Applied Econometrics 32*(3), 575–599.

Gu, J. and R. Koenker (2017b). Unobserved heterogeneity in income dynamics: An empirical bayes perspective. *Journal of Business & Economic Statistics 35*(1), 1–16.

Gu, J. and R. Koenker (2023). Invidious comparisons: Ranking and selection as compound decisions. *Econometrica 91*(1), 1–41.

Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics 7*(2), 225–245.

Henderson, N. C. and M. A. Newton (2016). Making the cut: improved ranking and selection for large-scale inference. *Journal of the Royal Statistical Society. Series B, Statistical methodology 78*(4), 781.

Jiang, W. and C.-H. Zhang (2009). General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics 37*(4), 1647–1684.

Kamiński, B. and P. Szufel (2018). On parallel policies for ranking and selection problems. *Journal of Applied Statistics 45*(9), 1690–1713.

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

Kwon, Y. and Z. Zhao (2023). On f-modelling-based empirical bayes estimation of variances. *Biometrika 110*(1), 69–81.

Luo, J., L. J. Hong, B. L. Nelson, and Y. Wu (2015). Fully sequential procedures for large-scale ranking-and-selection problems in parallel computing environments. *Operations Research 63*(5), 1177–1194.

Mosteller, F. (1948). A k-Sample Slippage Test for an Extreme Population. *The Annals of Mathematical Statistics 19*(1), 58 – 65.

Ni, E. C., D. F. Ciocan, S. G. Henderson, and S. R. Hunter (2017). Efficient ranking and selection in parallel computing environments. *Operations Research 65*(3), 821–836.

Panchapakesan, S. (1971). On a subset selection procedure for the most probable event in a multinomial distribution**this research was supported in part by the office of naval research contract n00014-67-a-0226-00014 and the aerospace research laboratories contract af33 (615)67c1244 at purdue university. reproduction in whole or in part is permitted for any purposes of the united states government. In S. S. Gupta and J. Yackel (Eds.), *Statistical Decision Theory and Related Topics*, pp. 275–298. Academic Press.

Paulson, E. (1949). A Multiple Decision Procedure for Certain Problems in the Analysis of Variance. *The Annals of Mathematical Statistics 20*(1), 95 – 98.

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

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 479–498.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association 102*(479), 901–912.

Sun, W. and A. C. McLain (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association 107*(498), 673–687.

Wand, M. P. and M. C. Jones (1994). *Kernel Smoothing*, Volume 60 of *Chapman and Hall CRC Monographs on Statistics and Applied Probability*. Chapman and Hall CRC.

Weinstein, A., Z. Ma, L. D. Brown, and C.-H. Zhang (2018). Group-linear empirical bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association 0*(0), 1–13.

Xie, X., S. Kou, and L. D. Brown (2012). Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association 107*(500), 1465–1479.

Zhong, Y., S. Liu, J. Luo, and L. J. Hong (2022). Speeding up paulson's procedure for large-scale problems using parallel computing. *INFORMS Journal on Computing 34*(1), 586–606.

Online Supplementary Material for "Ranking and Selection in Large-Scale Inference of Heteroscedastic Units"

This Online Supplement contains the proofs of main theorems and propositions (Section A), proofs of technical lemmas (Section B), a discussion on the grid size (Section C), a discussion on special cases of the r-value notion (Section D), a discussion on the issues related to the nested selection (Section E) and additional numerical results (Section F).

A Proof of main Theorems and Propositions

A.1 Proof of Theorem 1

We consider the more general problem

$$\text{Maximize } \mathbb{E}\left\{\sum_{i=1}^{m} h_i(\boldsymbol{X},\boldsymbol{\sigma})\delta_i\right\} \quad \text{subject to} \quad \text{mFDR} \leq \alpha. \tag{A.14}$$

We divide the hypotheses into four groups:

0. $h_i(\boldsymbol{X},\boldsymbol{\sigma}) \geq 0$ and $\text{Clfdr}_i - \alpha \leq 0$;

1. $h_i(\boldsymbol{X},\boldsymbol{\sigma}) \geq 0$ and $\text{Clfdr}_i - \alpha > 0$;

2. $h_i(\boldsymbol{X},\boldsymbol{\sigma}) < 0$ and $\text{Clfdr}_i - \alpha \leq 0$;

3. $h_i(\boldsymbol{X},\boldsymbol{\sigma}) < 0$ and $\text{Clfdr}_i - \alpha > 0$.

Define $T_i = \frac{h_i(\boldsymbol{X},\boldsymbol{\sigma})}{\text{Clfdr}_i - \alpha}$, we then consider decision rules of the following form

$$\delta(c_1,c_2)(T_i) = \begin{cases} 1 & \text{if } (h_i(\boldsymbol{X},\boldsymbol{\sigma}), \text{Clfdr}_i) \text{ belongs to group 0} \\ 1 & \text{if } (h_i(\boldsymbol{X},\boldsymbol{\sigma}), \text{Clfdr}_i) \text{ belongs to group 1 and } T_i > c_1 \\ 1 & \text{if } (h_i(\boldsymbol{X},\boldsymbol{\sigma}), \text{Clfdr}_i) \text{ belongs to group 2 and } T_i < c_2 \\ 0 & \text{otherwise} \end{cases}. \tag{A.15}$$

To simplify the proof, we consider a bounded, continuous and monotone transformation $\xi(\cdot)$ of T_i. Let $S_i = \xi(T_i)$. An example of such a transformation could be the hyperbolic tangent function $\xi(x) \equiv \tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$. Then the following set of rules are equivalent to the set of rules described in (A.15).

$$\delta(c_1,c_2)(S_i) = \begin{cases} 1 & \text{if } (h_i(\boldsymbol{X},\boldsymbol{\sigma}), \text{Clfdr}_i) \text{ belongs to group 0} \\ 1 & \text{if } (h_i(\boldsymbol{X},\boldsymbol{\sigma}), \text{Clfdr}_i) \text{ belongs to group 1 and } S_i > c_1 \\ 1 & \text{if } (h_i(\boldsymbol{X},\boldsymbol{\sigma}), \text{Clfdr}_i) \text{ belongs to group 2 and } S_i < c_2 \\ 0 & \text{otherwise} \end{cases}, \tag{A.16}$$

Define

$$c_1^- = \inf_{(h(\boldsymbol{X},\boldsymbol{\sigma}),\text{Clfdr}) \in \text{group 1}} \xi(T), \quad c_1^+ = \sup_{(h(\boldsymbol{X},\boldsymbol{\sigma}),\text{Clfdr}) \in \text{group 1}} \xi(T),$$

$$c_2^- = \inf_{(h(\boldsymbol{X},\boldsymbol{\sigma}),\text{Clfdr}) \in \text{group 2}} \xi(T), \quad c_2^+ = \sup_{(h(\boldsymbol{X},\boldsymbol{\sigma}),\text{Clfdr}) \in \text{group 2}} \xi(T).$$

For a decision rule of the form (A.16), we denotes its mFDR and modified power by $\text{mFDR}(c_1,c_2)$ and $\text{ETP}^*(c_1,c_2)$ respectively. Note that $\text{mFDR}(c_1,c_2)$ and $\text{ETP}^*(c_1,c_2)$ are both continuous and bounded functions of (c_1,c_2). Moreover, both are constant outside of the rectangle $\left[c_1^-, c_1^+\right] \times \left[c_2^-, c_2^+\right]$. Hence, without loss of generality, we restrict $\text{mFDR}(c_1,c_2)$

and $\text{ETP}^*\boldsymbol{\delta}(c_1, c_2)$ to the compact set $\left[c_1^-, c_1^+\right] \times \left[c_2^-, c_2^+\right]$. Define

$$(c_1^{OR}, c_2^{OR}) = \underset{(c_1,c_2)\in\left[c_1^-,c_1^+\right]\times\left[c_2^-,c_2^+\right]}{\arg\max} \{\text{ETP}^*(c_1, c_2) : \text{mFDR}(c_1, c_2) = \alpha\}. \tag{A.17}$$

Then the following theorem implies Theorem 1

Theorem 4. *The oracle procedure* $\boldsymbol{\delta}^{OR} = \boldsymbol{\delta}(c_1^{OR}, c_2^{OR})$ *where* $\boldsymbol{\delta}^{OR}$ *is as defined in* (A.16) *and* (A.17) *controls mFDR at level* α *and is optimal in the sense that for any decision rule* $\boldsymbol{\delta}$ *that controls mFDR at level* α*, we always have* $ETP^*(c_1^{OR}, c_2^{OR}) \geq ETP^*_{\boldsymbol{\delta}}$.

We will prove the more general Theorem 4. Observe that solving the constrained optimization problem (A.14) is equivalent to solving the subsequent problem:

$$\text{maximize } \; E\left\{\sum_i (h_i(\boldsymbol{X}, \boldsymbol{\sigma}) - \mu_0)\delta_i\right\} \quad \text{subject to } \; E\left\{\sum_i \delta_i(\text{Clfdr}_i - \alpha)\right\} \leq 0.$$

We divide the discussion into the following scenarios.

1. Decisions for units in group 0.

Let $\boldsymbol{\delta}$ be a decision rule satisfying $E\left\{\sum_i \delta_i(\text{Clfdr}_i - \alpha)\right\} \leq 0$. Denote by $\mathcal{R}(\boldsymbol{\delta})$ the set of hypotheses rejected by $\boldsymbol{\delta}$. Suppose that the null hypothesis $H_{0,j}$ from group 0 is not rejected by $\boldsymbol{\delta}$. Consider another decision rule $\boldsymbol{\delta}'$ with $\mathcal{R}(\boldsymbol{\delta}') = \mathcal{R}(\boldsymbol{\delta}) \cup \{j\}$. It is clear that

$$E\left\{\sum_i \delta_i'(\text{Clfdr}_i - \alpha)\right\} \leq 0 \text{ and } \sum_i h_i(\boldsymbol{x}, \boldsymbol{\sigma})\delta_i' \geq \sum_i h_i(\boldsymbol{x}, \boldsymbol{\sigma})\delta_i.$$

Hence, the optimal procedure must reject all hypotheses from group 0.

2. Decisions for units in group 3.

Next, suppose $\boldsymbol{\delta}$ rejects the null hypothesis $H_{0,j}$ from group 3. Consider a new decision

rule $\boldsymbol{\delta}'$ with $\mathcal{R}(\boldsymbol{\delta}') = \mathcal{R}(\boldsymbol{\delta})\backslash\{j\}$. It is clear that

$$E\left\{\sum_i \delta_i'(\mathrm{Clfdr}_i - \alpha)\right\} \leq 0 \text{ and } \sum_i h_i(\boldsymbol{X}, \boldsymbol{\sigma})\delta_i' > \sum_i h_i(\boldsymbol{X}, \boldsymbol{\sigma})\delta_i.$$

Hence, the optimal procedure does not reject any hypothesis from group 3.

3. Decisions for units in group 1 and group 2.

Let $\mathcal{R}^+_{\boldsymbol{\delta}} = \{i \in \mathcal{R}(\boldsymbol{\delta}) : \mathrm{Clfdr}_i - \alpha > 0\}$, and $\mathcal{R}^-_{\boldsymbol{\delta}} = \{i \in \mathcal{R}(\boldsymbol{\delta}) : \mathrm{Clfdr}_i - \alpha \leq 0\}$. Then $\mathcal{R}^+_{\boldsymbol{\delta}}$ and $\mathcal{R}^+_{\boldsymbol{\delta}}$ respectively correspond to the decisions for units in group 1 and group 2.

Remark 5. *We pause momentarily to offer clarification on the key concepts that will be presented in the remainder of the proof. It is important to note that the α-investing and μ-investing processes are interdependent, which means that the optimal cutoff in group 1 is contingent on the cutoff chosen in group 2. As a result, the derivation of the optimal decision rule can be challenging. However, an important observation is that if any decision procedure deviates from the oracle rule for group 1, it can be uniformly enhanced by ranking hypotheses in group 1 based on the ordering of T_i in descending order and then selecting a suitable threshold. This argument applies similarly in the opposite direction for selection of units in group 2. Therefore, although the process of determining the optimal pairs of (c_1, c_2) may be complex, the format of the optimal decision rule can be determined.*

Subsequently, we will demonstrate separately that both $\mathcal{R}^+_{\boldsymbol{\delta}_{OR}}$ and $\mathcal{R}^-_{\boldsymbol{\delta}_{OR}}$, when holding the part fixed, correspond to optimal rejection sets that cannot be further improved.

Suppose $\boldsymbol{\delta}$ satisfies $E\left\{\sum_i \delta_i(\mathrm{Clfdr}_i - \alpha)\right\} \leq 0$ and $\mathcal{R}^-_{\boldsymbol{\delta}^{OR}} = \mathcal{R}^-_{\boldsymbol{\delta}}$. The oracle rule on

group 1 can be expressed as

$$\delta_i^{OR} = \begin{cases} 0 & \text{if } h_i(\boldsymbol{X}, \boldsymbol{\sigma}) \leq \xi^{-1}(c_1^{OR})(\text{Clfdr}_i - \alpha) \\ 1 & \text{if } h_i(\boldsymbol{X}, \boldsymbol{\sigma}) > \xi^{-1}(c_1^{OR})(\text{Clfdr}_i - \alpha). \end{cases} \tag{A.18}$$

Let $\mathcal{I}^+ = \{i : E(\delta_i^{OR} - \delta_i) > 0\}$ and $\mathcal{I}^- = \{i : E(\delta_i^{OR} - \delta_i) < 0\}$. For $i \in \mathcal{I}^+$, we have $\delta_i^{OR} = 1$ and hence $h_i(\boldsymbol{X}, \boldsymbol{\sigma}) > \xi^{-1}(c_1^{OR})(\text{Clfdr}_i - \alpha)$. Similarly for $i \in \mathcal{I}^-$, we have $\delta_i^{OR} = 0$ and $h_i(\boldsymbol{X}, \boldsymbol{\sigma}) < \xi^{-1}(c_1^{OR})(\text{Clfdr}_i - \alpha)$. Thus,

$$\sum_{i \in \mathcal{I}^+ \cup \mathcal{I}^-} E\{\delta_i^{OR} - \delta_i\}\{h_i(\boldsymbol{X}, \boldsymbol{\sigma}) - \xi^{-1}(c_1^{OR})(\text{Clfdr}_i - \alpha)\} \geq 0. \tag{A.19}$$

Given $\mathcal{R}^-_{\boldsymbol{\delta}^{OR}} = \mathcal{R}^-_{\boldsymbol{\delta}}$, c_1^{OR} is chosen as small as possible such that $E\left\{\sum_i \delta_i^{OR}(\text{Clfdr}_i - \alpha)\right\} = 0$ or until the entire group 1 is rejected. Such c_1^{OR} exists because $E\left\{\sum_i \delta_i^{OR}(\text{Clfdr}_i - \alpha)\right\}$ is a continuous and monotone function of c_1^{OR} when ignoring the decisions in the other group. The argument follows from that in (Cai et al., 2019). In particular, this implies

$$E\left\{\sum_i \delta_i(\text{Clfdr}_i - \alpha)\right\} \leq E\left\{\sum_i \delta_i^{OR}(\text{Clfdr}_i - \alpha)\right\}. \tag{A.20}$$

Recall the definition of the power function

$$ETP^*_{\boldsymbol{\delta}} = E\left\{\sum_{i=1}^m h_i(\boldsymbol{X}, \boldsymbol{\sigma})\delta_i\right\}.$$

Using (A.19) and (A.20), we conclude that $ETP^*_{\boldsymbol{\delta}^{OR}} \geq ETP^*_{\boldsymbol{\delta}}$.

By employing a similar line of reasoning as described above, it can be shown that the optimal procedure involves ranking hypotheses in group 2 in ascending order of T_i and selecting an appropriate threshold.

Finally, we combine the claims from the four groups, and claim that the oracle rule is optimal in the sense of (A.15).

A.2 Proof of Proposition 1

It is worth noting that if (a, b) and (c, d) are points on L and $a \geq c$, then we must have $c \leq d$. Additionally, if $b \leq c$, then we must have $\text{ETP}^*_{\boldsymbol{\delta}(a,b)} \geq \text{ETP}^*_{\boldsymbol{\delta}(a,c)}$.

Suppose $\text{mFDR}_{\boldsymbol{\delta}} > \text{mFDR}_{\boldsymbol{\delta}'}$, the we can find another point $(r_1', \tilde{r}_2') \in l$ such that $\tilde{r}_2' \in [r_2, r_2')$, $\text{mFDR}_{\boldsymbol{\delta}} = \text{mFDR}_{\boldsymbol{\delta}(r_1', \tilde{r}_2')}$ and $\text{ETP}^*_{\boldsymbol{\delta}'} \leq \text{ETP}^*_{\boldsymbol{\delta}(r_1', \tilde{r}_2')} \leq \text{ETP}^*_{\boldsymbol{\delta}}$. Similarly, suppose $\text{mFDR}_{\boldsymbol{\delta}'} > \text{mFDR}_{\boldsymbol{\delta}''}$, the we can find another point $(c_1'', \tilde{c}_2'') \in l$ such that $\tilde{c}_2'' \in [c_2', c_2'')$, $\text{mFDR}_{\boldsymbol{\delta}'} = \text{mFDR}_{\boldsymbol{\delta}(c_1'', \tilde{c}_2'')}$ and

$$\text{ETP}^*_{\boldsymbol{\delta}''} \leq \text{ETP}^*_{\boldsymbol{\delta}(c_1'', \tilde{c}_2'')} \leq \text{ETP}^*_{\boldsymbol{\delta}'}.$$

Thus, if we can show the claim holds under the assumption that $\text{mFDR}_{\boldsymbol{\delta}} = \text{mFDR}_{\boldsymbol{\delta}'} = \text{mFDR}_{\boldsymbol{\delta}''}$, then the desired result follows.

We introduce some notations:

- $\mathbf{CLfdr} = (\text{CLfdr}_1, \text{CLfdr}_2, ..., \text{CLfdr}_m)$.
- $\boldsymbol{T} = (T_1, T_2, ..., T_m)$.
- $\boldsymbol{x}_{I_k}$ is the vector $\boldsymbol{x}$ restricted to group k.
- $\boldsymbol{x}|_{\boldsymbol{y}}$ is the vector $\boldsymbol{x}$ restricted to the non-zero entries in $\boldsymbol{y}$.
- $\mathbb{1} = (1, 1, .., 1)$ a vector of 1's.
- $\text{ave}(\boldsymbol{x}/\boldsymbol{y}) = \boldsymbol{x}\mathbb{1}^t/\boldsymbol{y}\mathbb{1}^t$.
- if $\boldsymbol{a} = (a_1, \ldots, a_n)$ is a vector and b is a number, then $\boldsymbol{a} - b = (a_1 - b, \ldots, a_n - b)$.

By our ranking strategy for the units in group 1 and group 2 in the oracle rule, we have

$$\text{ave}\{(\boldsymbol{x}-\mu_0)_{I_2}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}/(\mathbf{CLfdr}-\alpha)_{I_2}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}\} \geq \text{ave}\{(\boldsymbol{x}-\mu_0)_{I_2}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}/(\mathbf{CLfdr}-\alpha)_{I_2}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}\}, \tag{A.21}$$

$$\text{ave}\{(\boldsymbol{x}-\mu_0)_{I_1}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}/(\mathbf{CLfdr}-\alpha)_{I_1}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}\} \leq \text{ave}\{(\boldsymbol{x}-\mu_0)_{I_1}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}/(\mathbf{CLfdr}-\alpha)_{I_1}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}\}. \tag{A.22}$$

Next, note that $\text{ETP}^*_{\boldsymbol{\delta}} \geq \text{ETP}^*_{\boldsymbol{\delta}'}$ implies

$$(\boldsymbol{\delta}'-\boldsymbol{\delta})_{I_1}(\boldsymbol{x}-\mu_0\mathbb{1})^t_{I_1} \leq -(\boldsymbol{\delta}'-\boldsymbol{\delta})_{I_2}(E(\boldsymbol{\mu})-\mu_0\mathbb{1})^t_{I_2},$$

which can be re-written as

$$\begin{aligned}&\text{ave}\{(\boldsymbol{x}-\mu_0)_{I_1}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}/(\mathbf{CLfdr}-\alpha)_{I_1}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}\} \times (\boldsymbol{\delta}'-\boldsymbol{\delta})_{I_1}(\mathbf{CLfdr}-\alpha\mathbb{1})^t_{I_1} \\ \leq &-\text{ave}\{(\boldsymbol{x}-\mu_0)_{I_2}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}/(\mathbf{CLfdr}-\alpha)_{I_2}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}\} \times (\boldsymbol{\delta}'-\boldsymbol{\delta})_{I_2}(\mathbf{CLfdr}-\alpha\mathbb{1})^t_{I_2}.\end{aligned} \tag{A.23}$$

The condition $\text{mFDR}_{\boldsymbol{\delta}} = \text{mFDR}_{\boldsymbol{\delta}'}$ implies that

$$(\boldsymbol{\delta}'-\boldsymbol{\delta})_{I_1}(\mathbf{CLfdr}-\alpha\mathbb{1})^t_{I_1} = -(\boldsymbol{\delta}'-\boldsymbol{\delta})_{I_2}(\mathbf{CLfdr}-\alpha\mathbb{1})^t_{I_2}.$$

According to (A.23), we have

$$\text{ave}\{(\boldsymbol{x}-\mu_0)_{I_2}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}/(\mathbf{CLfdr}-\alpha)_{I_2}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}\} \geq \text{ave}\{(\boldsymbol{x}-\mu_0)_{I_1}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}/(\mathbf{CLfdr}-\alpha)_{I_1}|_{\boldsymbol{\delta}'-\boldsymbol{\delta}}\}.$$

Combining (A.21) with (A.22), we have

$$\text{ave}\{(\boldsymbol{x}-\mu_0)_{I_2}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}/(\mathbf{CLfdr}-\alpha)_{I_2}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}\} \geq \text{ave}\{(\boldsymbol{x}-\mu_0)_{I_1}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}/(\mathbf{CLfdr}-\alpha)_{I_1}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}\}.$$

Similarly the condition $\text{mFDR}^*_{\boldsymbol{\delta}''} = \text{mFDR}^*_{\boldsymbol{\delta}'}$ implies that

$$(\boldsymbol{\delta}'' - \boldsymbol{\delta}')_{I_1}(\mathbf{CLfdr} - \alpha\mathbb{1})^t_{I_1} = (\boldsymbol{\delta}'' - \boldsymbol{\delta}')_{I_2}(\mathbf{CLfdr} - \alpha\mathbb{1})^t_{I_2}.$$

Hence we have

$$\begin{aligned}
&(\boldsymbol{\delta}'' - \boldsymbol{\delta}')_{I_2}(\boldsymbol{x} - \mu_0\mathbb{1})^t_{I_2} \\
=&\text{ave}\{(\boldsymbol{x} - \mu_0)_{I_2}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}/(\mathbf{CLfdr} - \alpha)_{I_2}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}\} \times (\boldsymbol{\delta}'' - \boldsymbol{\delta}')_{I_2}(\mathbf{CLfdr} - \alpha\mathbb{1})^t_{I_2} \\
\leq& -\text{ave}\{(\boldsymbol{x} - \mu_0)_{I_1}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}/(\mathbf{CLfdr} - \alpha)_{I_1}|_{\boldsymbol{\delta}''-\boldsymbol{\delta}'}\} \times (\boldsymbol{\delta}'' - \boldsymbol{\delta}')_{I_1}(\mathbf{CLfdr} - \alpha\mathbb{1})^t_{I_1} \\
=& -(\boldsymbol{\delta}'' - \boldsymbol{\delta}')_{I_1}(\boldsymbol{x} - \mu_0\mathbb{1})^t_{I_1}.
\end{aligned}$$

Now we can see that

$$(\boldsymbol{\delta}'' - \boldsymbol{\delta}')_{I_2}(\boldsymbol{x} - \mu_0\mathbb{1})^t_{I_2} + (\boldsymbol{\delta}'' - \boldsymbol{\delta}')_{I_1}(\boldsymbol{x} - \mu_0\mathbb{1})^t_{I_1} \leq 0,$$

which implies that $\text{ETP}^*_{\boldsymbol{\delta}'} \geq \text{ETP}^*_{\boldsymbol{\delta}''}$ and the desired result follows.

A.3 Proof of Proposition 2

We first state a useful lemma:

Lemma 1. *Suppose $\mu_i \overset{iid}{\sim} g(\cdot)$, for $i = 1, ..., m$. Let $\hat{g}$ be the empirical density function $\sum_{i=1}^n \delta_{\mu_i}(\cdot)$. Let $f(x) = \int_{-\infty}^{\infty} \phi_\sigma(\mu - x)g(\mu)d\mu$ and $\hat{f}(x) = \int_{-\infty}^{\infty} \phi_\sigma(\mu - x)\hat{g}(\mu)d\mu$. Then for every x, $E_{\boldsymbol{\mu}}|f(x) - \hat{f}(x)|^2 \to 0$ as $m \to \infty$.*

Lemma 1 implies it is possible to find a set $\{\mu_1, \ldots, \mu_m\}$ and $\hat{f}_\sigma(x) = \frac{1}{m}\sum_{i=1}^m \phi_\sigma(x - \mu_i)$

such that for all x, $|f_\sigma(x) - \hat{f}_\sigma(x)|^2 \to 0$. Consider the following set of functions

$$\left\{ \sum_{i=0}^{k-1} w_i \phi_\sigma(x - s - i\eta) | \sum_{i=0}^{k-1} w_i = 1, w_i \geq 0 \ \ \forall i \right\}.$$

We can make the grid fine enough so that for any $\epsilon > 0$ and i, there exists $s_i \in \{s, s + \eta, ..., s + (k-1)\eta\}$ such that $|\mu_i - s_i| < \epsilon$. Hence

$$\begin{aligned} \big|\frac{1}{m}\sum_{i=1}^{m} \phi_\sigma(x - \mu_i) - \frac{1}{m}\sum_{i=1}^{m} \phi_\sigma(x - s_i)\big|^2 &= \frac{1}{m^2}\big|\sum_{i=1}^{m} \phi_\sigma(x - \mu_i) - \sum_{i=1}^{m} \phi_\sigma(x - s_i)\big|^2 \\ &\leq \frac{1}{m}\sum_{i=1}^{m} |\phi_\sigma(x - s_j) - \phi_\sigma(x - \mu_i)|^2. \end{aligned}$$

If we let $\epsilon \to 0$, then $|\phi_\sigma(x - s_i) - \phi_\sigma(x - \mu_i)|^2 \to 0$ for $i = 1, ..., m$. It follows that there exists

$$\psi_\sigma \in \{\sum_{i=0}^{k-1} w_i \phi_\sigma(x - s - i\eta) | \sum_{i=0}^{k-1} w_i = 1, w_i \geq 0 \ \ \forall i\}$$

such that $|f_\sigma(x) - \psi_\sigma(x)|^2 \to 0$.

Using standard arguments in density estimation theory (e.g. Wand and Jones (1994)), we have

$$\mathbb{E}\|\hat{f}_\sigma^m - f_\sigma\|_2^2 = O\{(mh_x h_\sigma)^{-1} + h_x^4 + h_\sigma^4\}.$$

By assumption (A2) $(mh_x h_\sigma)^{-1} + h_x^4 + h_\sigma^4 \to 0$. It follows that

$$\frac{1}{m}\sum_{i=1}^{m} \{\hat{f}_i(x_i) - \hat{f}_i^m(x_i)\}^2 \xrightarrow{p} \frac{1}{m}\sum_{i=1}^{m} \{\hat{f}_i(x_i) - f_{\sigma_i}(x_i)\}^2.$$

By definition of the minimization problem, we have $\frac{1}{m}\sum_{i=1}^{m} \{\hat{f}_i(x_i) - f_{\sigma_i}(x_i)\}^2 \xrightarrow{p} 0$. Thus, $E_{\boldsymbol{x},\boldsymbol{\sigma}} E_{\sigma,x} |\hat{f}_\sigma(x) - f_\sigma(x)|^2 \to 0$ and $E_{\boldsymbol{x},\boldsymbol{\sigma}} E_{\sigma,x} |\hat{f}_{0,\sigma}(x) - f_{0,\sigma}(x)|^2 \to 0$, where $\hat{f}_{0,\sigma}(x) = \sum_{s_i < \mu_0} w_i \phi_\sigma(x - s_i)$ and $f_{0,\sigma}(x) = \int_{-\infty}^{\mu_0} \phi_\sigma(x - \mu) g(\mu) d\mu$. Here $E_{\sigma,x}$ is taken with respect to σ and x, $E_{\boldsymbol{x},\boldsymbol{\sigma}}$ is

taken with respect to the data that are used to construct $\hat{f}$.

Note that f_σ is continuous, then there exists $K_1 = [-M, M]$ such that $P(x \in K_1^c) \to 0$ as $M \to \infty$. Let $\inf_{x \in K_1} f_\sigma(x) = l_0$ and $A_{l_0} = \{x : |\hat{f}_\sigma(x) - f_\sigma(x)| \geq l_0/2\}$. Since

$$E_{\boldsymbol{x},\boldsymbol{\sigma}} E_{\sigma,x} |\hat{f}_{0,\sigma}(x) - f_{0,\sigma}(x)|^2 \geq (l_0/2)^2 P(A_\epsilon^f),$$

it follows that $P(A_{l_0}) \to 0$. Thus $\hat{f}_\sigma$ and f_σ are bounded below by a positive number for large n, m except for an event that has a low probability. Similar arguments can be applied to the upper bound of $\hat{f}_\sigma$ and f_σ, as well as to the upper and lower bounds for $\hat{f}_{0,\sigma}$ and $f_{0,\sigma}$. Therefore, we conclude that $\hat{f}_{0,\sigma}$, $\hat{f}_\sigma$, $f_{0,\sigma}$ and f_σ. are all bounded in the interval $[l_a, l_b]$, $0 < l_a < l_b < \infty$ for large n, m except for an event, say A_{l_0} that has low probability. Let $\widehat{\text{Clfdr}}(x, \sigma) = \hat{f}_{0,\sigma}(x)/\hat{f}_\sigma(x)$ and $\text{Clfdr}(x, \sigma) = f_{0,\sigma}(x)/f_\sigma(x)$. We have

$$\widehat{\text{Clfdr}}(x, \sigma) - \text{Clfdr}(x, \sigma) = \frac{\hat{f}_{0,\sigma}(x) f_\sigma(x) - f_{0,\sigma}(x) \hat{f}_\sigma(x)}{\hat{f}_\sigma(x) f_\sigma(x)}.$$

Since $|\widehat{\text{Clfdr}} - \text{Clfdr}|^2$ is bounded by 1, we have

$$\begin{aligned} & E_{\boldsymbol{x},\boldsymbol{\sigma}} E_{\sigma,x} \{\widehat{\text{Clfdr}}(x, \sigma) - \text{Clfdr}(x, \sigma)\}^2 \\ \leq \; & P(A_{l_0}) + c_1 E_{\boldsymbol{x},\boldsymbol{\sigma}} E_{\sigma,x} \{\hat{f}_{0,\sigma}(x) - f_{0,\sigma}(x)\}^2 + E_{\boldsymbol{x},\boldsymbol{\sigma}} E_{\sigma,x} \{\hat{f}_\sigma(x) - f_\sigma(x)\}^2. \end{aligned}$$

Thus, $E_{\boldsymbol{x},\boldsymbol{\sigma}} E_{\sigma,x} \{\widehat{\text{Clfdr}}(x, \sigma) - \text{Clfdr}(x, \sigma)\}^2 \to 0$. Let $B_\delta = \{x, \sigma : |\widehat{\text{Clfdr}}(x, \sigma)| - \text{Clfdr}(x, \sigma)| > \delta\}$. Then we have

$$\delta^2 P(B_\delta) \leq E_{\boldsymbol{x},\boldsymbol{\sigma}} E_{\sigma,x} \{\widehat{\text{Clfdr}}(x, \sigma) - \text{Clfdr}(x, \sigma)\}^2 \to 0,$$

and the desired result follows.

A.4 Proof of Theorem 2

We consider the following data-driven algorithm for solving the more general problem (A.14)

Algorithm 2: The data-driven procedure for problem (A.14)

Input: $\{h_i(\boldsymbol{x},\boldsymbol{\sigma})\}_{i=1}^m$, $\widehat{\mathbf{Clfdr}}$, α.

Output: The estimated threshold for group 1 and group 2 ($\hat{c}_1$ and $\hat{c}_2$).

Step 1: Compute $\hat{T}_i = h_i(\boldsymbol{x},\boldsymbol{\sigma})/(\widehat{\text{Clfdr}}_i - \alpha)$, set $\hat{S}_i = \xi(\hat{T}_i)$. Form the 4 groups described in the oracle procedure (A.16) using $\widehat{\mathbf{Clfdr}}$ and $\hat{\boldsymbol{S}}$ in place of $\mathbf{Clfdr}$ and $\boldsymbol{S}$.

Step 2: Let $\mathcal{R}$ denote the rejection set. Put the indices of hypotheses from group 0 into $\mathcal{R}$. Rank hypotheses in group 1 from largest to smallest according to $\hat{S}_i$. Rank hypotheses in group 2 from smallest to largest according to $\hat{S}_i$.

Step 3: Denote the ranked hypotheses in group 1 by $H^1_{(1)}, H^1_{(2)}, ...$ and the corresponding Clfdr by $\text{Clfdr}_{(1)}, \text{Clfdr}_{(2)}, ...$ Let $k = \max\{j : \sum_{i=1}^{j}(\text{Clfdr}_{(i)} - \alpha) \leq -\sum_{i\in\mathcal{R}}(\text{Clfdr}_i - \alpha)\}$, reject $H_{(1)}, H_{(2)}, ...H_{(k)}$ and remove them from group 1. Compute and store $\text{ETP}^* = \sum_{i\in\mathcal{R}} h_i(\boldsymbol{x},\boldsymbol{\sigma})$.

Step 4: Denote the ranked hypotheses in group 2 by $H^2_{(1)}, H^2_{(2)}, ...$ Reject $H^2_{(1)}$ and remove it from group 2.

Step 5: Repeat step 3 and step 4. Terminate when ETP^* starts to decrease or when either group 1 or group 2 is empty.

Step 6: Let $(\hat{c}_1, \hat{c}_2)$ to be the pair that maximizes ETP^* and set $\boldsymbol{\delta}^{DD} = \boldsymbol{\delta}(\hat{c_1}, \hat{c}_2)$.

We show that the above algorithm is asymptotically optimal for the problem (A.14).

We begin with a summary of notation used throughout the proof:

- $Q_m(c_1, c_2) = m^{-1}\sum_{i=1}^m(\text{Clfdr}_i - \alpha)\boldsymbol{\delta}(c_1, c_2)(S_i)$.
- $\hat{Q}_m(c_1, c_2) = m^{-1}\sum_{i=1}^m(\widehat{\text{Clfdr}}_i - \alpha)\boldsymbol{\delta}(c_1, c_2)(\hat{S}_i)$.
- $Q_\infty(c_1, c_2) = E\{(\text{Clfdr} - \alpha)\boldsymbol{\delta}(c_1, c_2)(S)\}$.
- $\hat{H}_m(c_1, c_2) = m^{-1}\sum_{i=1}^m h_i(\boldsymbol{X},\boldsymbol{\sigma})\boldsymbol{\delta}(c_1, c_2)(\hat{S}_i)$.
- $H_\infty(c_1, c_2) = E\{m^{-1}\sum_{i=1}^m h_i(\boldsymbol{x},\boldsymbol{\sigma})\boldsymbol{\delta}(c_1, c_2)(S_i)\}$.

- $(c_1^{OR}, c_2^{OR}) = \arg\max_{(c_1,c_2):Q_\infty(c_1,c_2)\leq 0} \{H_\infty(c_1,c_2)\}$.
- $(\hat{c}_1, \hat{c}_2) = \arg\max_{(c_1,c_2):\hat{Q}(c_1,c_2)\leq 0} \left\{\hat{H}(c_1,c_2)\right\}$.

Here $\boldsymbol{\delta}(c_1,c_2)(S_i)$ is as defined in (A.16). We first show $\hat{Q}_m(c_1.c_2) \xrightarrow{p} Q_\infty(c_1,c_2)$. Note that $Q_m(c_1,c_2) \xrightarrow{p} Q_\infty(c_1,c_2)$ by the WLLN, so that we only need to establish $\hat{Q}_m(c_1,c_2) \xrightarrow{p} Q_m(c_1,c_2)$. We state a lemma that will be useful for the proof:

Lemma 2. *Let* $U_i = (Clfdr_i - \alpha)\delta(c_1,c_2)(S_i)$ *and* $\hat{U}_i = (\widehat{Clfdr}_i - \alpha)\delta(c_1,c_2)(\hat{S}_i)$ *then* $E(U_i - \hat{U}_i)^2 = o(1)$.

By Lemma 2 and Cauchy-Schwartz inequality, we have

$$E\left\{\left(\hat{U}_i - U_i\right)\left(\hat{U}_j - U_j\right)\right\} = o(1).$$

Let $L_m = \sum_{i=1}^m \left(\hat{U}_i - U_i\right)$. It follows that

$$\text{Var}\left(m^{-1}L_m\right) \leq m^{-2}\sum_{i=1}^m E\left\{\left(\hat{U}_i - U_i\right)^2\right\} + O\left(\frac{1}{m^2}\sum_{i,j:i\neq j} E\left\{\left(\hat{U}_i - U_i\right)\left(\hat{U}_j - U_j\right)\right\}\right) = o(1).$$

By Lemma 2, $E(m^{-1}L_m) \to 0$, applying Chebyshev's inequality, we obtain

$$m^{-1}L_m = \hat{Q}(c_1,c_2) - Q(c_1,c_2) \xrightarrow{p} 0.$$

Next we show $\hat{Q}_m(c_1.c_2) \to Q_\infty(c_1,c_2)$ uniformly. Since $\hat{Q}_m(c_1,c_2) \xrightarrow{p} Q_\infty(c_1,c_2)$ for all (c_1,c_2) on the rectangle $\left[c_1^-, c_1^+\right] \times \left[c_2^-, c_2^+\right]$, where $c_1^-, c_1^+, c_2^-, c_2^+$ are as defined in the beginning of Section A. Given any $\epsilon > 0$ the Lebesgue measure of the set

$$\{(c_1,c_2) \in \left[c_1^-, c_1^+\right] \times \left[c_2^-, c_2^+\right] : |\hat{Q}_m(c_1,c_2) - Q_\infty(c_1,c_2)| > \epsilon\}$$

approaches 0. Suppose there exists $\epsilon > 0$ such that for any M there is $m > M$ and (c_1, c_2) with $\hat{Q}_m(c_1, c_2) - Q_\infty(c_1, c_2) > 2\epsilon$, since Q_∞ is smooth, there exists a square $S_\delta(c_1, c_2)$ centered at (c_1, c_2) with side length δ such that

$$\hat{Q}_m(c_1, c_2) - Q_\infty(c_1', c_2') > \epsilon, \ \forall (c_1', c_2') \in S_\delta(c_1, c_2).$$

Consider the triangle with vertices at $\{(c_1, c_2), \ (c_1 - \delta/2, c_2), (c_1, c_2 - \delta/2)\}$, call this triangle Δ. It is clear that by definition of $\hat{Q}$ we have

$$\min\{\hat{Q}_m(c_1, c_2), \hat{Q}_m(c_1 - \delta/2, c_2), \hat{Q}_m(c_1, c_2 - \delta/2)\} \leq \inf_{(a,b)\in\Delta} \hat{Q}_m(a, b),$$

$$\max\{\hat{Q}_m(c_1, c_2), \hat{Q}_m(c_1 - \delta/2, c_2), \hat{Q}_m(c_1, c_2 - \delta/2)\} \geq \sup_{(a,b)\in\Delta} \hat{Q}_m(a, b).$$

Since $\hat{Q}_m(c_1 - \delta/2, c_2) \geq \hat{Q}_m(c_1, c_2)$ and $\hat{Q}_m(c_1, c_2 - \delta/2) \geq \hat{Q}_m(c_1, c_2)$. It follows that $\hat{Q}_m(a, b) - Q_\infty(a, b) > \epsilon, \ \forall (a, b) \in \Delta$. Note that δ only depends on ϵ, hence the area of Δ does not go to 0 as $m \to \infty$, a contradiction. Similarly, there is no (c_1, c_2) and $\epsilon > 0$ such that for any M, there exists $m > M$ with

$$\hat{Q}_m(c_1, c_2) - Q_\infty(c_1, c_2) < -2\epsilon.$$

Hence, for all (c_1, c_2) and $\epsilon > 0$, there exists M such that if $m > M$ then

$$|\hat{Q}_m(c_1, c_2) - Q_\infty(c_1, c_2)| < \epsilon.$$

Use similar arguments, we can also show $\hat{H}_m(c_1, c_2) \to H_\infty(c_1, c_2)$ uniformly. Define

$$\hat{l} = \left\{ (c_1, \hat{S}_i) : \hat{S}_i \in \text{group } 2, c_1 = \max\{k : \hat{S}^{(k)} \in \text{group } 1 \text{ and } \sum_{j=1}^{m} (\widehat{\text{Clfdr}}_i - \alpha)\delta(\hat{S}^{(k)}, \hat{S}_i)(\hat{S}_j) \leq 0\} \right\}.$$

It is clear that the data-driven algorithm only searches among the points on $\hat{l}$. By uniform convergence, given any $\epsilon > 0$, we can find M such that for all $m > M$ $|\hat{Q}_m(c_1, c_2) \to Q_\infty(c_1, c_2)| < \epsilon$ for all $(c_1, c_2) \in \hat{l}$. This shows $\text{mFDR}_{\boldsymbol{\delta}^{DD}} = \alpha + o(1)$.

Next we show $\boldsymbol{\delta}^{DD}$ is asymptotically optimal. By uniform continuity of H_∞, given any $\epsilon > 0$, there exists $\delta > 0$ such that $|H_\infty(a, b) - H_\infty(c, d)| < \epsilon$ for all $\|(a, b) - (c, d)\| \leq \delta$. With probability goes to 1, there exists a point $(a, b) \in D_\delta(c_1^{OR}, c_2^{OR})$. By uniform convergence of $\hat{H}_m$ to H_∞, we can choose m big enough so that $|\hat{H}_m(a, b) - H_\infty(a, b)| < \epsilon$, thus

$$|\hat{H}(a, b) - H_\infty(c_1^{OR}, c_2^{OR})| \leq |\hat{H}_m(a, b) - H_\infty(a, b)| + |H_\infty(a, b) - H_\infty(c_1^{OR}, c_2^{OR})| < 2\epsilon.$$

Again by uniform convergence we have $|\hat{H}_m(\hat{c}_1, \hat{c}_2) - H_\infty(\hat{c}_1, \hat{c}_2)| < \epsilon$ for all m big enough. By definition $\hat{H}_m(\hat{c}_1, \hat{c}_2) > \hat{H}_m(a, b)$, thus

$$H_\infty(\hat{c}_1, \hat{c}_2) > \hat{H}_m(a, b) - \epsilon > H_\infty(c_1^{OR}, c_2^{OR}) - 3\epsilon.$$

It follows that $\text{ETP}^*_{\boldsymbol{\delta}^{DD}} / \text{ETP}^*_{\boldsymbol{\delta}^{OR}} \geq 1 + o(1)$, proving the desired result.

A.5 Proof of Theorem 3

Since the two definitions of r-value share the same selection procedure, it suffices to show that if $X_i > X_j$ and $\text{Clfdr}_i < \text{Clfdr}_j$ then the rejection of hypothesis j implies the rejection of hypothesis i. We break the proof into several cases:

case 1: If hypothesis j belongs to group 0, then by definition hypothesis i also belongs to group 0, hence it is also rejected.

case 2: If hypothesis j belongs to group 1, then hypothesis i is either in group 0 or in group 1 with $T_i > T_j$. By definition of the oracle procedure, hypothesis i is rejected.

case 3: If hypothesis j belongs to group 2, then hypothesis i is either in group 0 or in group 2 with $T_i < T_j$. By definition of the oracle procedure, hypothesis i is rejected.

The proof for the data-driven procedure with Clfdr substituted by $\widehat{\text{Clfdr}}$ and T replaced by $\hat{T}$ can be derived using the same argument.

B Proof of Lemmas

B.1 Proof of Lemma 1

We use the bias-variance decomposition:

$$E\{f(x) - \hat{f}(x)\}^2 = \{E\hat{f}(x) - f(x)\}^2 + \text{Var}\hat{f}(x).$$

Write $\hat{g} = \sum_{i=1}^{m} \frac{1}{m}\delta_{\mu_i}(\cdot)$ as a mixture of point mass where $\mu_i \overset{iid}{\sim} g$. By definition,

$$E\hat{f}(x) = E\sum_{i=1}^{m} \frac{1}{m}\phi_\sigma(x - \mu_i) = E\phi_\sigma(x - \mu) = \int_{-\infty}^{\infty} \phi_\sigma(x - \mu)g(\mu)d\mu = f(x).$$

$\{E\hat{f}(x) - f(x)\}^2 = 0$. Also since ϕ is bounded, it follows that $\text{Var}\{\phi_\sigma(x - \mu_i)\} < \infty$. Therefore

$$\begin{aligned}
\text{Var}\hat{f}(x) &= \text{Var}\left\{\int_{-\infty}^{\infty} \phi_\sigma(\mu - x)\hat{g}(\mu)d\mu\right\} \\
&= \text{Var}\left\{\frac{1}{m}\sum_{i=1}^{m} \phi_\sigma(x - \mu_i)\right\} \\
&= \frac{1}{m}\text{Var}\{\phi_\sigma(x - \mu_i)\} \to 0.
\end{aligned}$$

B.2 Proof of Lemma 2

We state a fact that will be helpful:

Lemma 3. $\hat{S}_i \xrightarrow{P} S_i$.

Lemma 3 is proved in section B.3. By definition of U_i and $\hat{U}_i$ we have the following:

$$\begin{aligned}
(U_i - \hat{U}_i)^2 = (\text{Clfdr}_i - \widehat{\text{Clfdr}}_i)^2 \mathbb{I}\{\delta(c_1, c_2)(S_i) = \delta(c_1, c_2)(\hat{S}_i) = 1\} \\
+ (\text{Clfdr}_i - \alpha)^2 \mathbb{I}\{\delta(c_1, c_2)(S_i) = 1, \delta(c_1, c_2)(\hat{S}_i) = 0\} \\
+ (\widehat{\text{Clfdr}}_i - \alpha)^2 \mathbb{I}\{\delta(c_1, c_2)(S_i) = 0, \delta(c_1, c_2)(\hat{S}_i) = 1\}
\end{aligned}$$

Denote the three sums on the RHS as I, II, and III respectively. By Proposition 2, $E(I) = o(1)$. To show $E(II + III) = o(1)$ we only need to show $P\{\delta(c_1, c_2)(S_i) \neq \delta(c_1, c_2)(\hat{S}_i)\} = o(1)$. We say S_i or $\hat{S}_i$ is from group a if (X_i, Clfdr_i) is from group a. $\delta(c_1, c_2)(S_i) \neq \delta(c_1, c_2)(\hat{S}_i)$ can only happen when at least one of the following holds:

1. S_i and $\hat{S}_i$ are not from the same group.
2. S_i and $\hat{S}_i$ both from group 1 but $\hat{S}_i > c_1$ and $S_i \leq c_1$.
3. S_i and $\hat{S}_i$ both from group 1 but $\hat{S}_i \leq c_1$ and $S_i > c_1$.

4. S_i and $\hat{S}_i$ both from group 2 but $\hat{S}_i < c_2$ and $S_i \geq c_1$.

5. S_i and $\hat{S}_i$ both from group 2 but $\hat{S}_i \geq c_1$ and $S_i < c_2$.

Since $P(\text{Clfdr}_i = \alpha) = P(\widehat{\text{Clfdr}}_i = \alpha) = 0$ The probability that S_i and $\hat{S}_i$ are not from the same group is bounded by

$$P(\text{Clfdr}_i < \alpha, \widehat{\text{Clfdr}}_i > \alpha) + P(\text{Clfdr}_i > \alpha, \widehat{\text{Clfdr}}_i < \alpha). \tag{B.24}$$

Note that

$$\begin{aligned} P\left(\text{Clfdr}_i < \alpha, \widehat{\text{Clfdr}}_i > \alpha\right) &\leq P\left(\text{Clfdr}_i < \alpha, \widehat{\text{Clfdr}}_i \in (\alpha, \alpha+\epsilon)\right) + P\left(\text{Clfdr}_i < \alpha, \widehat{\text{Clfdr}}_i \geq \alpha+\epsilon\right) \\ &\leq P\left(\widehat{\text{Clfdr}}_i \in (\alpha, \alpha+\epsilon)\right) + P\left(\left|\text{Clfdr}_i - \widehat{\text{Clfdr}}_i\right| > \epsilon\right). \end{aligned}$$

The first term on the right hand is vanishingly small as $\epsilon \to 0$ because $\widehat{\text{Clfdr}}_i$ is a continuous random variable. The second term converges to 0 by Proposition 2. We conclude that

$$P(\text{Clfdr}_i < \alpha, \widehat{\text{Clfdr}}_i > \alpha) = o(1).$$

Use similar argument, the remaining terms in (B.24) are $o(1)$, the probability of the first situation occurs is $o(1)$.

For situation 2, we have

$$\begin{aligned} P\left(\hat{S}_i > c_1, S_i \leq c_1\right) &\leq P\left(S_i \leq c_1, \hat{S}_i \in (c_1, c_1+\epsilon)\right) + P\left(S_i \leq c_1, \hat{S}_i \geq +\epsilon\right) \\ &\leq P\left(\hat{S}_i \in (c_1, c_1+\epsilon)\right) + P\left(\left|\hat{S}_i - S_i\right| > \epsilon\right). \end{aligned}$$

The first term on the right hand is vanishingly small as $\epsilon \to 0$ because $\hat{S}_i$ is a continuous

random variable. The second term converges to 0 by Lemma 3. we conclude that

$$P\left(\hat{S}_i > c_1, S_i \leq c_1\right) = o(1).$$

In a similar fashion, we can show that situation 3-5 are all $o(1)$. The lemma follows.

B.3 Proof of Lemma 3

Let $A_\epsilon = \{x : |\text{Clfdr}_i - \alpha| < \epsilon\}$. Then $P(A_\epsilon) \to 0$ as $\epsilon \to 0$. Let $l_0 = \inf_{x_i \in A_\epsilon^c} |\text{Clfdr}_i - \alpha|$ and $B_{l_0} = \{x_i : |\widehat{\text{Clfdr}}_i - \text{Clfdr}_i| > l_0/2\}$. Since $\widehat{\text{Clfdr}}_i \xrightarrow{P} \text{Clfdr}_i$. We have $P(B_{l_0}) \to 0$. Thus $|\text{Clfdr}_i - \alpha|$ and $|\widehat{\text{Clfdr}}_i - \text{Clfdr}_i|$ are bounded below by a positive number for large m except for an event that has a low probability. Note that

$$\hat{T}_i - T_i = \frac{(\text{Clfdr}_i - \widehat{\text{Clfdr}}_i)h_i(\boldsymbol{x}, \boldsymbol{\sigma})}{(\widehat{\text{Clfdr}}_i - \alpha)(\text{Clfdr}_i - \alpha)}.$$

It follows that $(\hat{T}_i - T_i)^2 = O\left\{(\widehat{\text{Clfdr}}_i - \text{Lfdr}_i)^2 (h_i(\boldsymbol{x}, \boldsymbol{\sigma}))^2\right\}$ on $A_\epsilon^c \cap B_{l_0}^c$. Note that

$$(\widehat{\text{Clfdr}}_i - \text{Lfdr}_i)^2 = O_P(\text{Lfdr}_i^2).$$

And since $g_\mu(\cdot)$ has bounded support and the noise is Gaussian, it follows that

$$\lim_{x_i \to \pm\infty} \text{Lfdr}_i^2 (h_i(\boldsymbol{x}, \boldsymbol{\sigma}))^2 = 0.$$

Hence $(\hat{T}_i - T_i)^2 = O_P\left\{(\widehat{\text{Clfdr}}_i - \text{Lfdr}_i)^2\right\}$. Since S_i and $\hat{S}_i$ are continuous function of T_i and $\hat{T}_i$ respectively and $\|S_i - \hat{S}_i\|^2$ is bounded it follows that $E\|S_i - \hat{S}_i\|^2 \to 0$ and $\hat{S}_i \xrightarrow{P} S_i$.

C Grid Size

In the proof of Proposition 2, we have used the fact that

$$\mathbb{E}\|\hat{f}_\sigma^m - f_\sigma\|_2^2 = O\{(mh_xh_\sigma)^{-1} + h_x^4 + h_\sigma^4\}.$$

The optimal rate of $(mh_xh_\sigma)^{-1} + h_x^4 + h_\sigma^4$ is $m^{-2/3}$ and is achieved when $h_x \sim h_\sigma \sim m^{-1/6}$. Since g_μ has bounded support, for any μ_j we can always find $u_{i(j)} \in \{u_1, \ldots, u_k\}$ such that $|u_{i(j)} - \mu_j| = O(1/k)$. Let $\epsilon = |u_{i(j)} - \mu_j|$, then

$$|\phi_\tau(x - \mu_j) - \phi_\tau(x - u_{i(j)})|^2 = \frac{1}{2\pi\tau^2} e^{-\frac{x^2}{\tau^2}} |1 - e^{\frac{2x\epsilon - \epsilon^2}{2\tau^2}}|^2. \tag{C.25}$$

We want the above to be of order $O(m^{-2/3})$ uniformly for any x. If x has order greater than $\sqrt{\log m}$ then it is clear that the RHS of (C.25) is $O(m^{-2/3})$. When x has order less than $\sqrt{\log m}$, since $e^{-\frac{x^2}{\tau^2}} = O(1)$ we focus on $|1 - e^{\frac{2x\epsilon - \epsilon^2}{2\tau^2}}|^2$. By Taylor's expansion, we have

$$|1 - e^{\frac{2x\epsilon - \epsilon^2}{2\tau^2}}|^2 = O\left\{ \left(\frac{2x\epsilon - \epsilon^2}{2\tau^2} \right)^2 \right\}.$$

It is clear that if $\epsilon = O\{1/(m^{1/3} \log m)\}$ then the above is $O(m^{-2/3})$, it follows that the a grid size of $k = O(m^{1/3} \log m)$ is sufficient.

D R-value, p-value and q-value

This section presents two examples that illustrate how to transform a selection procedure into an informative ranking metric using Definition 1 for the r_α-value.

Example 1. *Suppose that our objective is to identify significant cases among multiple*

candidate units while controlling the per-comparison error rate (PCER). To achieve this, we can employ a simple selection rule, denoted by $I(|T_i| > \alpha)$, $i \in [m]$, where T_i represents either a t-statistic or a z-statistic. By sequentially varying the PCER level α from 0 to 1, the study units can be selected in an ordered order. If we consider a scenario where the global null hypothesis is valid, and hypotheses are selected at a PCER level of α, then the minimum α required for a case to be chosen is equivalent to the familiar p-value. The p-value can subsequently be used as a ranking variable to signify a unit's position in the list.

Example 2. *In the second example, let us consider the application of the adaptive p-value procedure (Benjamini and Hochberg, 2000) to select units while controlling the positive false discovery rate (pFDR) at a given level of α. By gradually increasing α, an informative ranking of the units can be obtained. The minimum pFDR level α required for a unit to be selected is known as the q-value (Storey, 2002), which can be employed as a ranking variable to indicate the unit's relative position in the list. The earlier a unit is selected, the more crucial it is deemed to be in comparison to the remaining units.*

The r_α-value is a versatile concept that can be applied to a broad range of selection procedures, as illustrated by the two examples presented above. Specifically, we have shown that the p-value and q-value can be regarded as particular cases of the r_α-value.

Finally, our r_α-value draws inspiration from and is closely linked to the r-value presented in Henderson and Newton (2016). Nonetheless, the two definitions diverge significantly with regards to the optimization criterion and the intended goal of analysis.

E The Nestedness Property in Sequential Selection

The topic of nested selection has been previously addressed in Gu and Koenker (2023) and Henderson and Newton (2016). In an ideal scenario, if we relax the constraint by reducing

μ_0 or increasing α, we would expect that hypotheses rejected under the stricter condition would remain rejected under the relaxed constraint. However, the oracle procedure outlined in Section 2.2 may not satisfy the nestedness property defined in Definitions 3 or 4.

Section 4 introduced two notions of r-value, leading to the definition of two types of nestedness that will be discussed in the next two subsections respectively.

E.1 Nestedness induced by varying α

Definition 3. *Consider $\mathcal{R}_\alpha^{\mathcal{D}}$ and $\mathcal{R}_\alpha^{\mathcal{D}}$ as defined in Definition 1. A testing procedure $\mathcal{D}$ is* nested *if the rejection regions $\mathcal{R}_{\alpha'}^{\mathcal{D}}$ and $\mathcal{R}_\alpha^{\mathcal{D}}$ satisfy the inclusion property $\mathcal{R}_{\alpha'}^{\mathcal{D}} \subseteq \mathcal{R}_\alpha^{\mathcal{D}}$ for all $\alpha' < \alpha$.*

To illustrate why the oracle selection procedure is not nested according to Definition 3, consider the following example. Recall that $T_i = \dfrac{X_i - \mu_0}{\text{Clfdr}_i - \alpha}$. Suppose we have $T_1 \approx T_2 \approx \ldots \approx T_k$, with T_1 being slightly larger than $T_2, \ldots, T_k$. Additionally, assume that $\text{Clfdr}_1 - \alpha > 0$ and $X_1 - \mu_0 > 0$ are both relatively large, while $\text{Clfdr}_j - \alpha > 0$ and $X_j - \mu_0 > 0$ are relatively small for $j = 2, \ldots, k$. It is worth noting that there are more than k hypotheses in total, but we are focusing on these particular hypotheses for the sake of clarity and simlicity.

It is possible to select a value of α such that T_1 is rejected, while $T_2, \ldots, T_k$ are not. However, if we slightly increase the target FDR level to β, then $T_2, \ldots, T_k$ will exceed T_1 (assuming $\text{Clfdr}_i \geq \beta$ for $i = 1, \ldots, k$). Consequently, it is possible for hypothesis 1 to be rejected at FDR level α, but not at level β, violating the nestedness property.

E.2 Nestedness induced by varying μ_0

Definition 4. *Consider $\mathcal{R}^{\mathcal{D}}_{\mu_0'}$ and $\mathcal{R}^{\mathcal{D}}_{\mu_0}$ as defined in Definition 2. A testing procedure $\mathcal{D}$ is nested if for all $\mu_0 < \mu_0'$ we have $\mathcal{R}^{\mathcal{D}}_{\mu_0'} \subseteq \mathcal{R}^{\mathcal{D}}_{\mu_0}$.*

To further illustrate this point, consider the following example, where the observations are generated from the following model:

$$\mu_i \overset{iid}{\sim} U(0, 10), \quad X_i|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i^2).$$

Consider the scenario in which we have two data points, $z_1 = (x_1, \sigma_1) = (7.33, 1)$ and $z_2 = (x_2, \sigma_2) = (6.71, 0.5)$. We fix $\alpha = 0.1$ and set $\mu_0 = 6$. Numerical calculations reveal that both z_1 and z_2 belong to group 1, with $T_1 > T_2$. Consequently, it is possible that H_{01} is rejected while H_{02} is not. However, if we lower μ_0 to 5.9, z_1 still belongs to group 1, but z_2 now belongs to group 0. As a result, H_{02} is rejected, but H_{01} may not be. Therefore, the oracle selection procedure is not nested, as per Definition 4.

E.3 Conclusion

In summary, agreeability appears to be a more appropriate criterion for ranking procedures in the presence of heteroscedasticity. Our analysis has demonstrated that the r-values derived from Definitions 1 and 2 both meet the requirement of agreeability, which in turn results in ranking rules that are meaningful and valid.

F Supplementary Numerical Results

This section provides additional numerical results.

F.1 A comparative analysis of ETP* and ETP

In this section, we conduct numerical studies to demonstrate that maximizing ETP and ETP* are two distinct objectives. We generate $m = 10000$ observations that follow the hierarchical model described below.

$$\mu_i \sim N(5, 0.5^2), \qquad \sigma_i = 1, \quad X|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i^2), \quad 1 \le i \le 5000,$$
$$\mu_i \sim N(7, 0.5^2), \quad \sigma_i = \sigma, \quad X|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i^2), \quad 5001 \le i \le 10000.$$

We aim to test the hypotheses $H_{0,i} : \mu_i \le \mu_0$ versus $H_{a,i} : \mu_i > \mu_0$, for $i \in [m]$, where $\mu_0 = 6$. We compare the performance of the following three methods:

(a) the oracle procedure derived in Section 2.2, denoted as OR;

(b) the data-driven method proposed in Section 3, denoted as DD;

(c) The oracle procedure designed to maximize the conventional ETP while controlling the FDR, denoted as Clfdr (see Fu et al. (2022) for further details).

We repeat the experiment on 100 datasets, and set the nominal FDR level to $\alpha = 0.1$, and report the results based on the average of the 100 replications. The data-driven method requires the independence between σ_i and μ_i. To ensure the validity of the data-driven approach, we first partition the data into two groups based on whether $\sigma_i = 1$ or $\sigma_i \neq 1$. We then estimate $g_\mu(\cdot)$ separately for each of the two groups.

We calculate the FDR as the average of the FDPs over 100 replications. The FDP is defined as $\mathrm{FDP}(\boldsymbol{\delta}) = \sum_{i=1}^m \{(1-\theta_i)\delta_i\}/(\sum_{i=1}^m \delta_i \vee 1)$. Similarly, we compute the ETP and ETP* as the averages of $\sum_{i=1}^m \theta_i \delta_i$ and $\sum_{i=1}^m (x_i - \mu_0)\delta_i$, respectively, over 100 replications. The value of σ varies from 1.5 to 2.5 across different settings. We present the results in Fig 9.

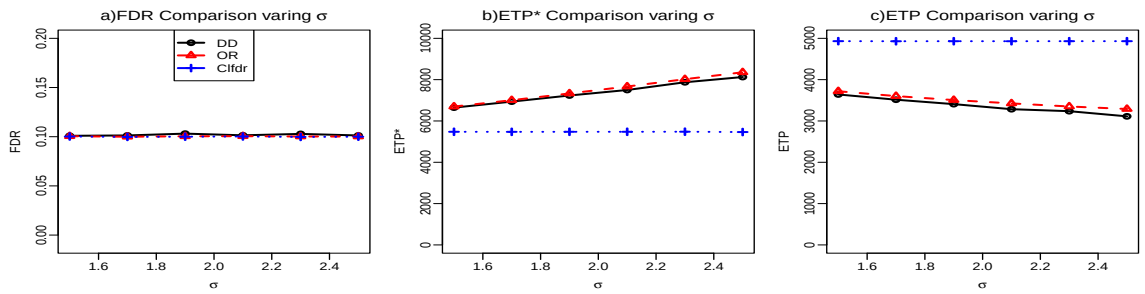

Figure 9: An example where ETP* and ETP are very different. DD and OR have much higher ETP* than Clfdr while Clfdr has much higher ETP than DD and OR.

The results indicate that all three methods effectively control the FDR at the nominal level. However, there are significant differences in their power performance. In particular, the ETP* values of the OR and DD methods are substantially higher than that of Clfdr, while Clfdr exhibits a significantly higher ETP than OR and DD. This observation aligns with the fact that the Clfdr method is designed to optimize traditional power, whereas OR and DD are developed with the objective of optimizing modified power.

F.2 Comparison when g_μ is uniform

We consider the following setting.

$$\mu_i \overset{iid}{\sim} U(0,10), \quad \sigma_i \overset{iid}{\sim} U(0.5,\sigma_{max}), \quad X_i \sim N(\mu_i,\sigma_i^2), \quad i \in [5000].$$

We aim to test the hypotheses $H_{0,i} : \mu_i \leq \mu_0$ versus $H_{a,i} : \mu_i > \mu_0$, with $\mu_0 = 6$. The nominal FDR level is set to 0.1, while σ_{max} varies from 2 to 4 for different settings. The final results are obtained by averaging the results in 100 replications and are presented in Figure 10. We can see that the data driven procedure performs reasonably well.

In Figure 11 (a), we look at one particular run with $\sigma_{max} = 4$. We can see that even though the ETP* and ETP of DD and Clfdr are close the rejection pattern is similar as before. Compare to Clfdr, DD still prefers hypotheses with larger X_i. In Figure 11 (b) we

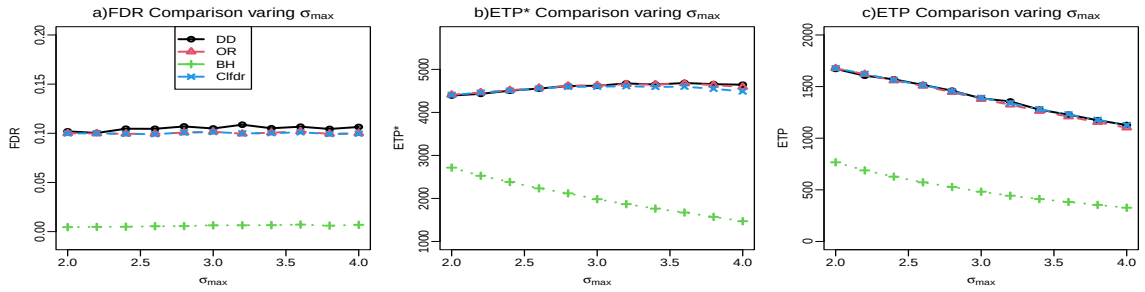

Figure 10: We vary σ_{max} from 2.0 to 4.0. The FDRs for DD are are close to the target level.

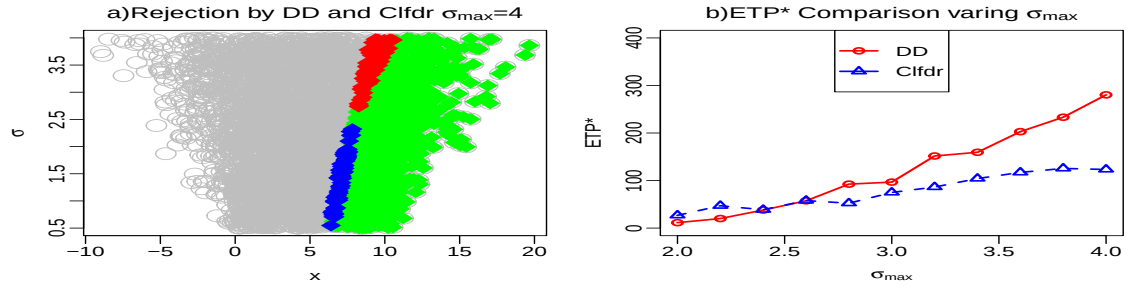

Figure 11: (a): A scatter plot of the hypotheses when $\sigma_{max} = 4$. The gray circles are hypotheses rejected by neither DD or Clfdr, green dots are hypotheses rejected by both DD and Clfdr, red dots are hypotheses rejected by DD but not Clfdr, blue dots are hypotheses rejected by Clfdr but not DD. (b): ETP* comparison of hypotheses rejected by either DD or Clfdr, but not both.

see that for the hypotheses that are rejected by only one method, DD has a superior ETP* in comparison to Clfdr when σ_{max} is large.

F.3 Comparison when σ_i are unknown

In the main text we assumed σ_i are known. However, in some applications σ_i is unknown and must be estimated from the data. In this subsection, we conduct experiments to study the effect of estimated σ_i of the performance of the data-driven method. The experiment considers the following hierarchical model to generate data:

$$\mu_i \overset{iid}{\sim} U(0, 10), \quad \sigma_i \overset{iid}{\sim} \sqrt{n}U(0.5, 4), \quad X_{i,1}, \ldots, X_{i,n} | \mu_i, \sigma_i \overset{iid}{\sim} N(\mu_i, \sigma_i^2), \quad i \in [5000].$$

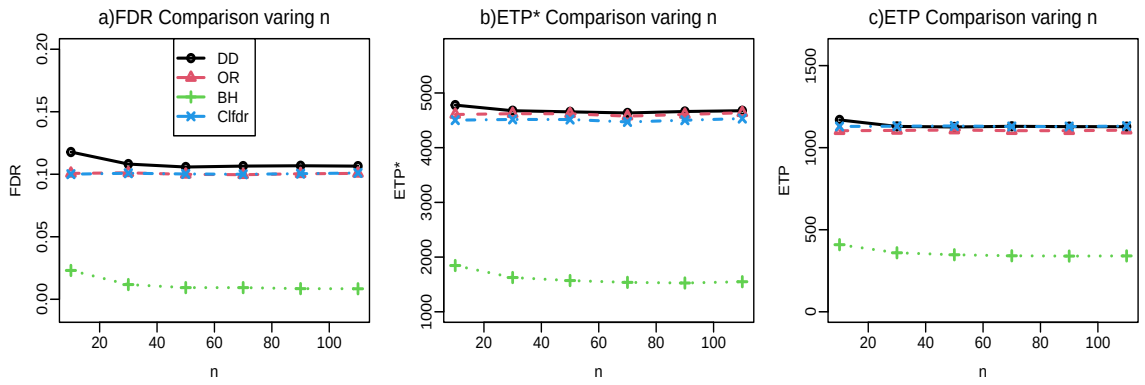

Figure 12: We vary n from10 to 110. The FDRs for DD are are close to the target level.

Our objective is to test the hypotheses

$$H_{0,i} : \mu_i \leq \mu_0 \text{ versus } H_{a,i} : \mu_i > \mu_0,\ i \in [5000],$$

where we choose $\mu_0 = 6$. The nominal FDR level is set to 0.1, while n varies from 10 to 110 for different settings. When implementing DD, we use $\bar{X}_i = \sum_{j=1}^{n} X_{i,j}$ as the raw observation, and $s_i/\sqrt{n}$ as the standard deviation for $\bar{X}_i$. Here s_i is the sample standard deviation of $X_{i,1}, \ldots, X_{i,n}$. The p-values are computed as $1 - \Phi\{(\bar{X}_i - \mu_0)/(s_i/\sqrt{n})\}$, where Φ is the distribution function of $N(0, 1)$. When implementing OR and Clfdr we use $\bar{X}_i$ as the raw observation and $\sigma_i/\sqrt{n}$ as the standard deviation. The final results are obtained by averaging the results in 100 replications and are presented in Figure 12. We can see that the FDR control for DD is reasonably good when n is of moderate size.

In Figure 13 (a), we look at one particular run with $n = 10$. We can see that the rejection pattern is similar as before, compare to Clfdr, DD still prefers hypotheses with larger X_i. In Figure 13 (b) we see that for the hypotheses that are rejected by only one method, DD still has a superior ETP* in comparison to Clfdr.

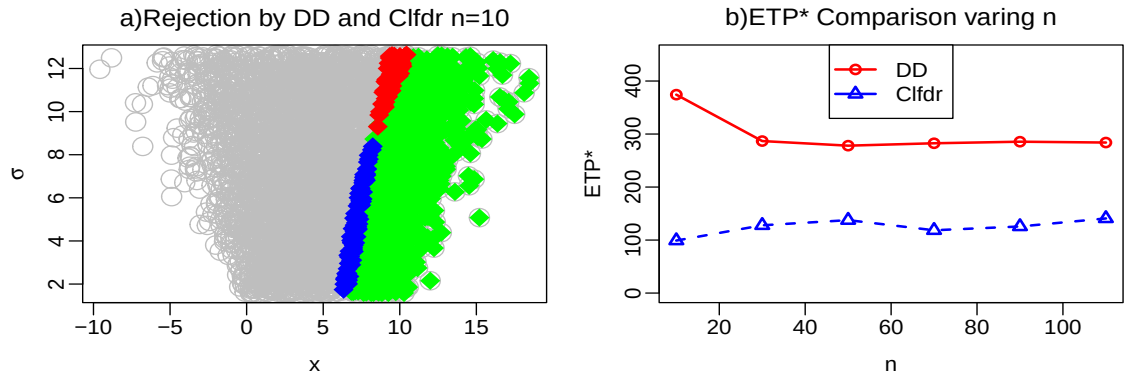

Figure 13: (a): A scatter plot of the hypotheses when $n = 0.5$. The gray circles are hypotheses rejected by neither DD or Clfdr, green dots are hypotheses rejected by both DD and Clfdr, red dots are hypotheses rejected by DD but not Clfdr, blue dots are hypotheses rejected by Clfdr but not DD. (b): ETP* comparison of hypotheses rejected by either DD or Clfdr, but not both.

F.4 Comparison when the number of hypotheses is small

Using a nonparametric method to estimate the distribution for μ is a challenging task. In this subsection we examine the performance of the data-driven method when the number of hypotheses is small.

We consider the following setting.

$$\theta_i \overset{iid}{\sim} \text{Ber}(0.2), \quad \mu_i|\theta_i \sim (1-\theta_i)U(-3,-1) + \theta_i U(1,2),$$

$$\sigma_i \overset{iid}{\sim} U(0.5,4), \quad X_i|\mu_i,\sigma_i \sim N(\mu_i,\sigma_i^2), \quad i \in [m].$$

We aim to test the hypotheses $H_{0,i} : \mu_i \leq \mu_0$ versus $H_{a,i} : \mu_i > \mu_0$, with $\mu_0 = 0$. The nominal FDR level is set to 0.1, while m varies from 200 to 2000 for different settings. The final results are obtained by averaging the results in 100 replications and are presented in Fig. 14. We can see that the data driven procedure performs well even when m is small, the FDRs only inflate slightly.

A plot of the hypotheses rejected by DD and Clfdr when $m = 200$ shows the same

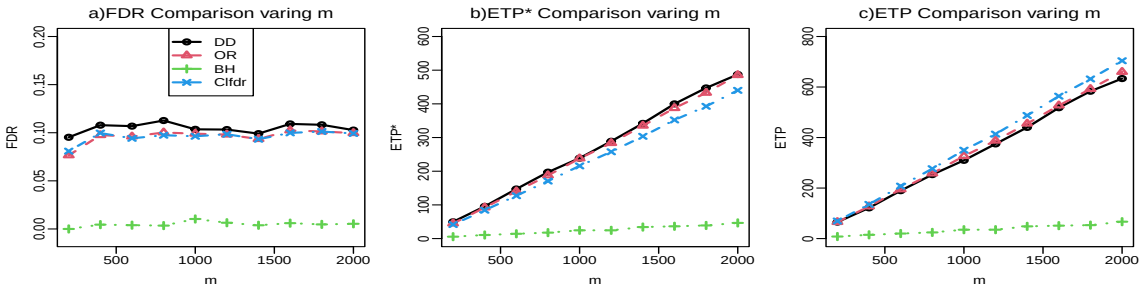

Figure 14: We vary m from 200 to 2000. The FDRs for DD are slightly inflated when m is small.

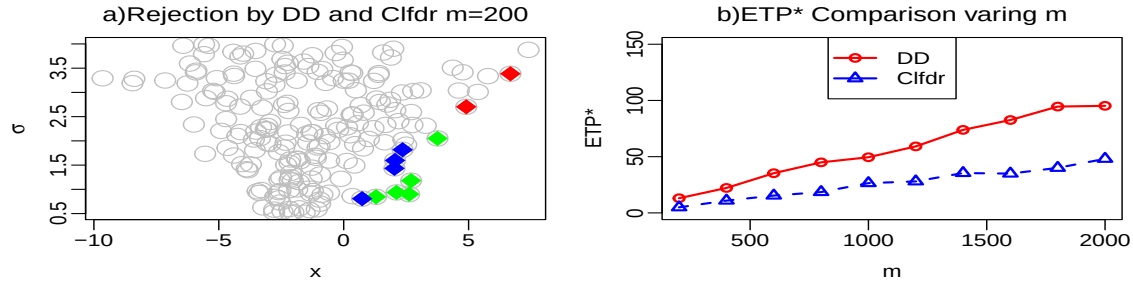

Figure 15: (a) A scatter plot of the hypotheses when $m = 200$. The gray circles are hypotheses rejected by neither DD or Clfdr, green dots are hypotheses rejected by both DD and Clfdr, red dots are hypotheses rejected by DD but not Clfdr, blue dots are hypotheses rejected by Clfdr but not DD. (b): ETP* comparison of hypotheses rejected by either DD or Clfdr, but not both.

pattern as before. In Figure 15 (a), we look at one particular run with $m = 200$. The gray dots are hypotheses not rejected by either DD or Clfdr, green dots are hypotheses rejected by both DD and Clfdr, red dots are hypotheses rejected by DD but not Clfdr, and blue dots are hypotheses rejected by Clfdr but not DD. It can be seen that DD is more likely to reject hypotheses with higher x_i values when compared to Clfdr. In Figure 15 (b) we see that for the hypotheses that are rejected by only one method, DD has a superior ETP* in comparison to Clfdr even when m is relatively small.

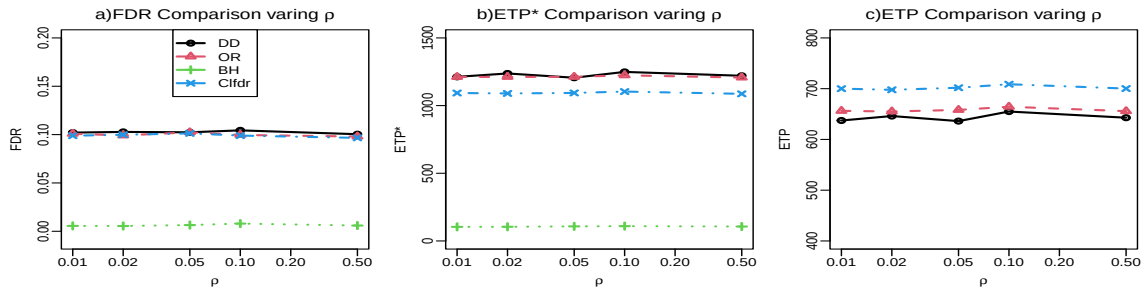

Figure 16: We vary ρ from 0.01 to 0.5. The FDRs for DD are are close to the target level.

F.5 Comparison when X_i are dependent

In this subsection we examine the performance of the data-driven procedure when X_i are dependent. We consider the following setting.

$$\theta_i \overset{iid}{\sim} \text{Ber}(0.2), \quad \mu_i|\theta_i \sim (1-\theta_i)U(-3,-1)+\theta_i U(1,2), \quad \sigma_i \overset{iid}{\sim} U(0.5,4), \quad \boldsymbol{X} = \boldsymbol{\mu}+\boldsymbol{\epsilon} \quad i \in [5000],$$

Here, Σ is a 5000×5000 block diagonal matrix, where the kth diagonal block A_k is a 10×10 matrix with its (i,j)th entry equals to $\sigma_{10k+i}\sigma_{10k+j}\rho^{|i-j|}$. We aim to test the hypotheses $H_{0,i} : \mu_i \leq \mu_0$ versus $H_{a,i} : \mu_i > \mu_0$, with $\mu_0 = 0$. The nominal FDR level is set to 0.1, while we set ρ to $0.01, 0.02, 0.05, 0.1, 0.5$. When implementing OR and Clfdr the Clfdr statistic is still computed as

$$\text{Clfdr}_i = \frac{f_{0i}(x_i)}{f_i(x_i)},$$

where $f_{0i}(x_i) = \int_{\mu \leq \mu_0} \phi_{\sigma_i}(x_i - \mu) g_\mu(\mu) d\mu$ and $f_i(x_i) = \int_{-\infty}^{\infty} \phi_{\sigma_i}(x_i - \mu) g_\mu(\mu) d\mu$, $\phi_{\sigma_i}(\cdot)$ is the density function of $N(0, \sigma_i^2)$, $g_\mu(\cdot)$ is the density function of the mixture distribution $0.8U(-3,-1) + 0.2U(1,2)$.

The final results are obtained by averaging the results in 100 replications and are presented in Figure 16, note that the x-axis is in logarithmic scale. We can see that the data-driven method still controls FDR reasonably well under weak dependence.

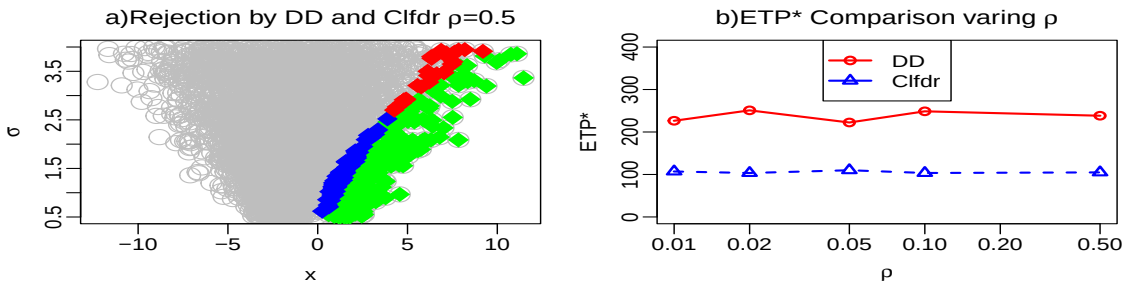

Figure 17: (a) A scatter plot of the hypotheses when $\rho = 0.5$. The gray circles are hypotheses rejected by neither DD or Clfdr, green dots are hypotheses rejected by both DD and Clfdr, red dots are hypotheses rejected by DD but not Clfdr, blue dots are hypotheses rejected by Clfdr but not DD. (b): ETP* comparison of hypotheses rejected by either DD or Clfdr, but not both.

In Figure 17 (a), we look at one particular run with $\rho = 0.5$. We can see that the rejection pattern is similar as before, compare to Clfdr, DD still prefers hypotheses with larger X_i. In Figure 17 (b) we see that for the hypotheses that are rejected by only one method, DD has a superior ETP* in comparison to Clfdr under weak dependence.

F.6 Comparison at various FDR levels

We consider the following setting.

$$\theta_i \overset{iid}{\sim} \text{Ber}(0.2), \quad \mu_i|\theta_i \sim (1-\theta_i)U(-3,-1) + \theta_i U(1,2),$$

$$\sigma_i \overset{iid}{\sim} U(0.5,4), \quad X_i|\mu_i,\sigma_i \sim N(\mu_i,\sigma_i^2), \quad i \in [5000].$$

We aim to test the hypotheses $H_{0,i} : \mu_i \leq \mu_0$ versus $H_{a,i} : \mu_i > \mu_0$, with $\mu_0 = 0$. The nominal FDR level varies from 0.01 to 0.1 for different settings. The final results are obtained by averaging the results in 100 replications and are presented in Figure 18.

In Figure 19(a) we look at one particular run with nominal FDR= 0.01. We can see that the rejection pattern is similar as before, compare to Clfdr, DD still prefers hypotheses

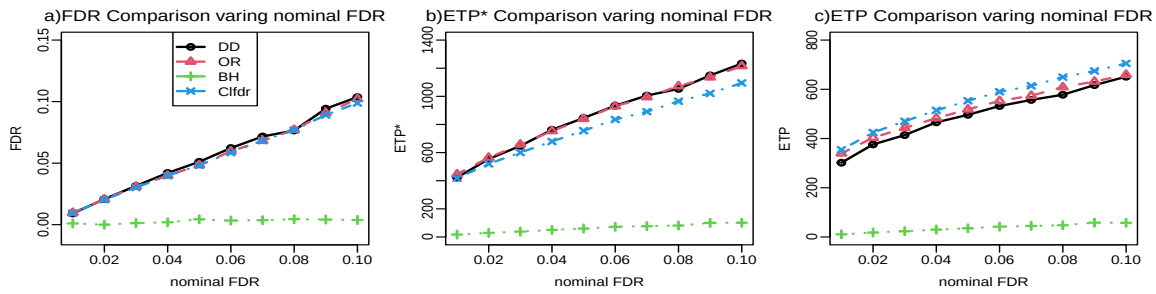

Figure 18: We vary nominal FDR from 0.01 to 0.1. The FDRs for DD are are close to the target level.

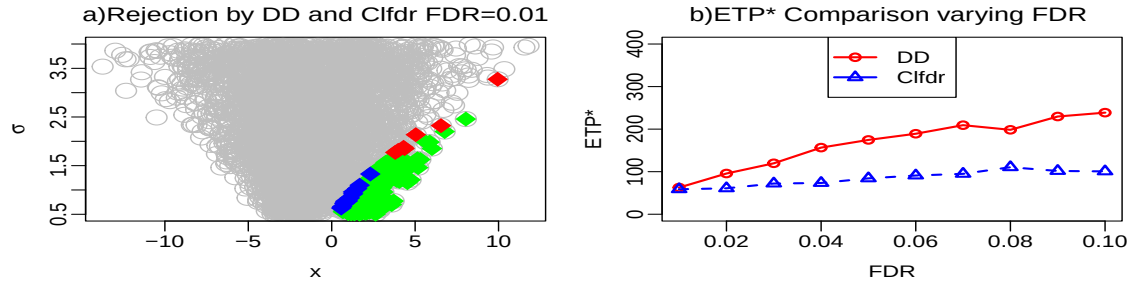

Figure 19: (a): A scatter plot of the hypotheses when nominal FDR is 0.01. The gray circles are hypotheses rejected by neither DD or Clfdr, green dots are hypotheses rejected by both DD and Clfdr, red dots are hypotheses rejected by DD but not Clfdr, blue dots are hypotheses rejected by Clfdr but not DD. (b): ETP* comparison of hypotheses rejected by either DD or Clfdr, but not both.

with larger X_i. In Figure 19 (b) we see that for the hypotheses that are rejected by only one method, DD has a superior ETP* in comparison to Clfdr across all nominal FDR levels.

F.7 The analysis of the AYP data

The unprocessed data sets for the AYP study are available at `https://www.cde.ca.gov/re/pr/api-Bdatarecordlayouts.asp`. We begin by defining Y and Y' as the passing rates of students from socially-economically advantaged backgrounds (SEA) and socially-economically disadvantaged backgrounds (SED), respectively Our objective is to identify significant differences in the passing rates $X_i = Y_i - Y_i'$ for each school i, where $i \in [m]$ and

$m = 6,398$. The standard error of X_i is calculated as

$$s_i = \sqrt{Y_i(1 - Y_i)/n_i + Y'(1 - Y_i')/n_{i'}},$$

where n_i and $n_{i'}$ are the number of SEA and SED students tested, respectively. To ensure numerical stability, we remove observations that have a standard error below the 1% percentile or above the 99% percentile. Figure 20 presents the scatter plot and histograms of the observed data.

We aim to test the following hypotheses: $H_{0,i} : \mu_i \leq \mu_0$ vs $H_{a,i} : \mu_i > \mu_0$, with $\mu_0 = 0.2$ being the cutoff of the indifference region and FDR level set at $\alpha = 0.01$. We calculate the z-values as $z_i = (x_i - \mu_0)/s_i$, and the p-values as $p_i = 1 - \Phi(z_i)$, where Φ is the standard normal cumulative distribution function.

Our primary focus is to compare our approach against analyses that solely rely on statistical significance indices (Clfdr and BH). In this context, the Clfdr method refers to the data-driven HART procedure (Fu et al., 2022). We summarize the results in Table 2, which reports the total number of rejections (a proxy for traditional power) and weighted number of rejections based on Equation (2.5) (a proxy for modified power).

We can see that DD and Clfdr outperform BH in terms of both the traditional and modified powers. Although DD and Clfdr exhibit similar performances, the hypotheses rejected by the two methods exhibit different patterns. Figure 21 (a) displays a scatter plot of hypotheses rejected by Clfdr and DD. The gray circles represent hypotheses that were not rejected by either method, the green dots represent hypotheses rejected by both Clfdr and DD, the red dots represent hypotheses rejected by DD but not Clfdr, and the blue dots represent hypotheses rejected by Clfdr but not DD. It is clear that DD displays a predilection for rejecting hypotheses with larger effect sizes, while Clfdr has a preference

for rejecting hypotheses with low standard error. Figure 21 (b) presents the top 20 schools ranked according to both p-values (indicated by blue dots) and r_{μ_0}-values (represented by red dots). Notably, the r_{μ_0}-value demonstrates a distinct inclination towards schools with larger effect sizes as compared to the p-value.

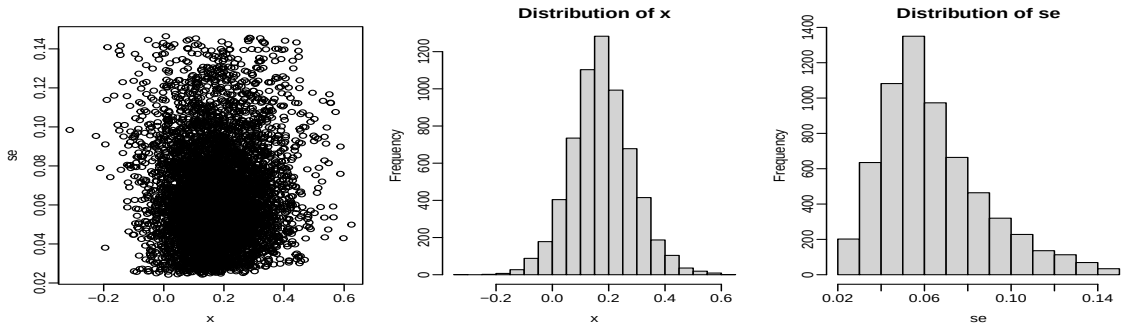

Figure 20: Left: scatter plot of the AYP data: x-axis is the observation, y-axis is the standard error. Middle: histogram of the observations. Right: histogram of the standard errors.

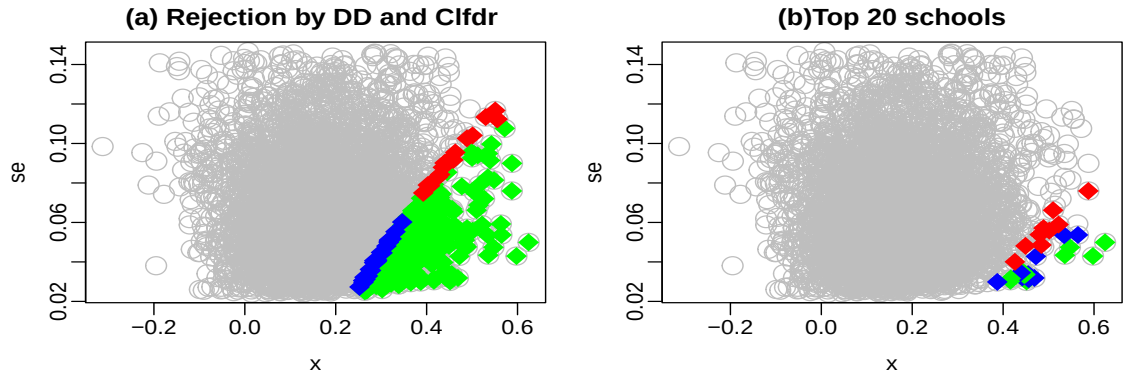

Figure 21: (a) The scatter plot displays the school data. The x-axis represents the raw observations (x), while the y-axis corresponds to the SEs. The gray circles represent schools that were not selected by either DD or Clfdr. The green dots denote schools selected by both Clfdr and DD, while the blue dots represent schools selected by Clfdr but not DD. The hypotheses selected by DD but not Clfdr are shown as red dots. (b) The scatter plot displays the top 20 schools ranked according to p-value and r_{μ_0}-value (Definition 2 with $\alpha = 0.01$). The top 20 schools ranked according to the r_{μ_0}-value are depicted as red dots, while the top 20 schools ranked according to the p-value are shown as blue dots. The schools ranked as top 20 by both p-value and r_{μ_0}-value are represented by green dots.

Table 2: Summary of power by each method on the AYP data.

	DD	Clfdr	BH
Number of hypotheses rejected	388	398	158
Modified Power	69.5	67.7	34.3

F.8 Comparison of r-values

We have introduced two notions of r-values. r_α is obtained by fixing μ_0 and vary α, r_{μ_0} is obtained by fixing α and vary μ_0. In Figure 22 we illustrate the difference between the two r-values on the CRSP and AYP data. As shown by the blue dots in Figure 22, we can see that the selection process using r_α, favors units with small standard errors. This issue can be ameliorated using r_{μ_0}, as shown by the red dots in Figure 22. In both the AYP and CRSP datasets, which are marked by considerable heteroscedasticity, we recommend using r_{μ_0} to identify units with the greatest effect sizes.

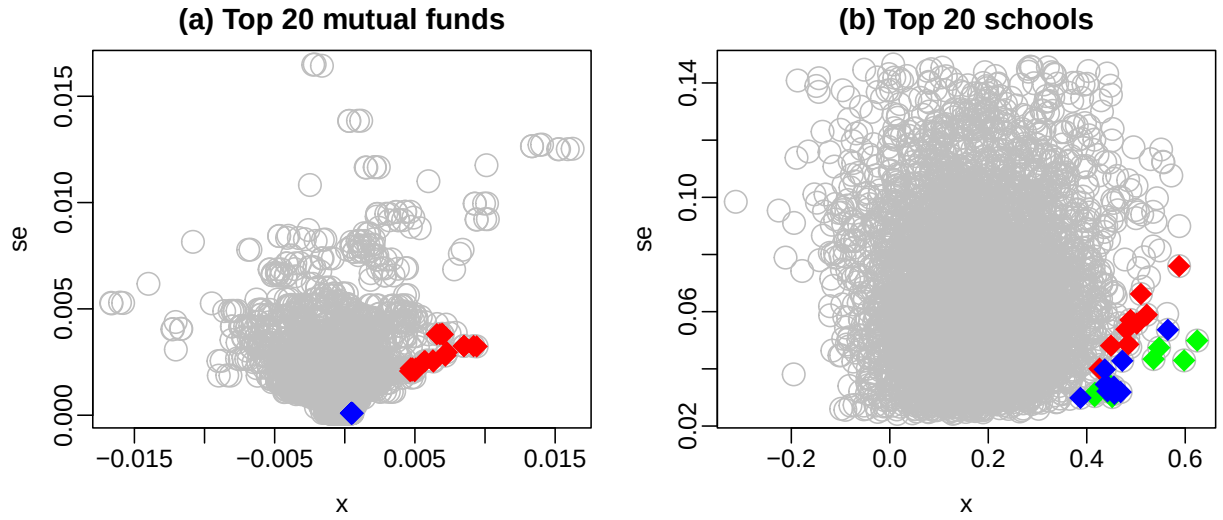

Figure 22: (a) The scatter plot displays the top 20 mutual funds ranked according to the two definitions of r-value. The top 20 mutual funds ranked according to r_α (fix $\mu_0 = 0$ and vary α) are depicted as blue dots (the dots represent the top 20 mutual funds are on top of each other so they appear as one blue dot in the graph), while the top 20 mutual funds ranked according to r_{μ_0} (fix $\alpha = 0.1$ and vary μ_0) are shown as red dots. (b) The scatter plot displays the top 20 schools ranked according to the two definitions of r-value. The top 20 schools ranked according to r_α (fix $\mu_0 = 0.2$ and vary α) are depicted as blue dots, while the top 20 schools ranked according to r_{μ_0} (fix $\alpha = 0.01$ and vary μ_0) are shown as red dots. The schools ranked as top 20 by both methods are represented by green dots.