Heterocedasticity-Adjusted Ranking and Thresholding for Large-Scale Multiple Testing

Luella Fu[1], Bowen Gang[2], Gareth M. James[3], and Wenguang Sun[3]

Abstract

Standardization has been a widely adopted practice in multiple testing, for it takes into account the variability in sampling and makes the test statistics comparable across different study units. However, despite conventional wisdom to the contrary, we show that there can be a significant loss in information from basing hypothesis tests on standardized statistics rather than the full data. We develop a new class of heteroscedasticity–adjusted ranking and thresholding (HART) rules that aim to improve existing methods by simultaneously exploiting commonalities and adjusting heterogeneities among the study units. The main idea of HART is to bypass standardization by directly incorporating both the summary statistic and its variance into the testing procedure. A key message is that the variance structure of the alternative distribution, which is subsumed under standardized statistics, is highly informative and can be exploited to achieve higher power. The proposed HART procedure is shown to be asymptotically valid and optimal for false discovery rate (FDR) control. Our simulation results demonstrate that HART achieves substantial power gain over existing methods at the same FDR level. We illustrate the implementation through a microarray analysis of myeloma.

Keywords: covariate-assisted inference; data processing and information loss; false discovery rate; heteroscedasticity; multiple testing with side information; structured multiple testing

[1]Department of Mathematics, San Francisco State University.
[2]Department of Statistics, Fudan University.
[3]Department of Data Sciences and Operations, University of Southern California. The research of Wenguang Sun was supported in part by NSF grant DMS-1712983.

1 Introduction

In a wide range of modern scientific studies, multiple testing frameworks have been routinely employed by scientists and researchers to identify interesting cases among thousands or even millions of features. A representative sampling of settings where multiple testing has been used includes: genetics, for the analysis of gene expression levels (Tusher et al., 2001; Dudoit et al., 2003; Sun and Wei, 2011); astronomy, for the detection of galaxies (Miller et al., 2001); neuro-imaging, for the discovery of di↵erential brain activity (Pacifico et al., 2004; Schwartzman et al., 2008); education, to identify student achievement gaps (Efron, 2008a); data visualization, to find potentially interesting patterns (Zhao et al., 2017); and finance, to evaluate trading strategies (Harvey and Liu, 2015).

The standard practice involves three steps: reduce the data in di↵erent study units to a vector of summary statistics X_i, with associated standard deviation $_i$; standardize the summary statistics to obtain z-values, $Z_i = X_i / _i$; and finally, order the z-values, or associated p-values, and apply a threshold to keep the rate of Type I error below a pre-specified level. Classical approaches concentrated on setting a threshold that controls the family-wise error rate (FWER), using methods such as the Bonferroni correction or Holm's procedure (Holm, 1979). However, the FWER criterion becomes infeasible once the number of hypotheses under consideration grows to thousands. The seminal contribution of Benjamini and Hochberg (1995) proposed replacing the FWER by the false discovery rate (FDR) and provided the BH algorithm for choosing a threshold on the ordered p-values which, under certain assumptions, is guaranteed to control the FDR.

While the BH procedure o↵ers a significant improvement over classical approaches, it only controls the FDR at level $(1 \quad \uparrow)\downarrow$, where $\uparrow$ is the proportion of non-nulls, suggesting that its power can be improved by incorporating an adjustement for $\uparrow$ into the procedure. Benjamini and Hochberg (2000), Storey (2002) and Genovese and Wasserman (2002) proposed to first estimate the non-null proportion by $\uparrow$ and then run BH at level $\downarrow/(1 \quad \uparrow)$. Efron et al. (2001) proposed the local false discovery rate (Lfdr), which incorporates, in addition to the sparsity parameter $\uparrow$, information about the alternative distribution. Sun and Cai (2007) proved that the z-value optimal procedure is an Lfdr thresholding rule and that this rule uniformly dominates the p-value optimal procedure in Genovese and Wasserman

(2002). The key idea is that the shape of the alternative could potentially affect the rejection region but the important structural information is lost when converting the z-value to p-value. For example, when the means of non-null effects are more likely to be positive than negative, then taking this asymmetry of the alternative into account increases the power. However, the sign information is not captured by conventional p-value methods, which only consider information about the null.

Although a wide variety of multiple testing approaches have been proposed, they almost all begin with the standardized data Z_i (or its associated p-value, P_i). In fact, in large-scale studies where the data are collected from intrinsically diverse sources, the standardization step has been upheld as conventional wisdom, for it takes into account the variability of the summary statistics and suppresses the heterogeneity – enabling one to compare multiple study units on an equal footing. For example, in microarray studies, Efron et al. (2001) first compute standardized two-sample t-statistics for comparing the gene expression levels across two biological conditions and then convert the t-statistics to z–scores, which are further employed to carry out FDR analyses. Binomial data is also routinely standardized by rescaling the number of successes X_i by the number of trials n_i to obtain success probabilities $\hat{p}_i = X_i/n_i$ and then converting the probabilities to z-scores (Efron, 2008a,b). However, while standardization is an intuitive, and widely adopted, approach, we argue in this paper that there can be a significant loss in information from basing hypothesis tests on Z_i rather than the full data $(X_i, \ _i)$[1]. This observation, which we formalize later in the paper, is based on the fact that the power of tests can vary significantly as changes, but this difference in power is suppressed when the data is standardized and treated as equivalent. In the illustrative example in Section 2.2, we show that by accounting for differences in an alternative ordering of rejections can be obtained, allowing one to identify more true positives at the same FDR level.

This article develops a new class of heteroscedasticity-adjusted ranking and thresholding (HART) rules for large-scale multiple testing that aim to improve existing methods by simultaneously exploiting commonalities and adjusting heterogeneities among the study

[1]Unless otherwise stated, the term "full data" specifically refers to the pair $(X_i, \ _i)$ in this article. In practice, the process of deriving the pair $(X_i, \ _i)$ from the original (full) data could also suffer from information loss, but this point is beyond the scope of this work; see the rejoinder of Cai et al. (2019) for related discussions.

units. The main strategy of HART is to bypass standardization by directly incorporating $(X_i, \ _i)$ into the testing procedure. We adopt a two-step approach. In the first step a new significance index is developed by taking into account the alternative distribution of each X_i conditioned on $\ _i$; hence HART avoids power distortion. Then, in the second step the significance indices are ordered and the smallest ones are rejected up to a given cuto↵. We develop theories to show that HART is optimal for integrating the information from both X_i and $\ _i$. Numerical results are provided to confirm that in finite sample settings HART controls the FDR and uniformly dominates existing methods in terms of power.

We are not the first to consider adjusting for heterogeneity. Ignatiadis et al. (2016) and Lei and Fithian (2018) mentioned the possibility of using the p-value as a primary significance index while employing $\ _i$ as side-information to weight or pre-order hypotheses. Earlier works by Efron (2008a) and Cai and Sun (2009) also suggest grouping methods to adjust for heterogeneous variances in data. However, the variance issue is only briefly mentioned in these works and it is unclear how a proper pre–ordering or grouping can be created based on $\ _i$. It is important to note that the ordering or grouping based on the magnitudes of $\ _i$ will not always be informative. Our numerical results show that ordering by $\ _i$ is suboptimal, even potentially leading to power loss compared to methods that utilize no side information. In contrast with existing works, we explicitly demonstrate the key role that $\ _i$ plays in characterizing the shape of the alternative in simultaneous testing (Section 2.2). Moreover, HART provides a principled and optimal approach for incorporating the structural information encoded in $\ _i$. We prove that HART guarantees FDR control and uniformly improves upon all existing methods in terms of asymptotic power.

The findings are impactful for three reasons. First, the observation that standardization can be ine cient has broad implications since, due to inherent variabilities or di↵ering sample sizes between study units, standardized tests are commonly applied to large-scale heterogeneous data to make di↵erent study units comparable. Second, our finding enriches the recent line of research on multiple testing with side and structural information (e.g. Cai et al., 2019; Li and Barber, 2019; Xia et al., 2019, among others). In contrast with these works that have focused on the usefulness of sparsity structure, our characterization of the impact of heteroscedasticity, or more concretely the shape of alternative distribution, is new. Finally, HART convincingly demonstrates the benefits of leveraging structural information

in high-dimensional settings when the number of tests is in the thousands or more.

The rest of the paper is organized as follows. Section 2 reviews the standard multiple testing model and provides a motivating example that clearly illustrates the potential power loss from standardization. Section 3 describes our HART procedure and its theoretical properties. Section 4 contains simulations, and Section 5 demonstrates the method on a microarray study. We conclude the article with a discussion of connections to existing work and open problems. Technical materials, proofs and additional numerical results are provided in the Appendix.

2 Problem Formulation and the Issue of Standardizing

This section first describes the problem formulation and then discusses an example to illustrate the key issue.

2.1 Problem formulation

Let $\checkmark_i$ denote a Bernoulli($\Uparrow$) variable, where $\checkmark_i = 0/1$ indicates a null/alternative hypothesis, and $\Uparrow = P(\checkmark_i = 1)$ is the proportion of nonzero signals coming from the alternative distribution. Suppose the summary statistics $X_1, \ldots, X_m$ are normal variables obeying distribution

$$X_i \,|\, \mu_i, \ {}_i^2 \overset{ind}{\longleftarrow} N(\mu_i, \ {}_i^2), \tag{2.1}$$

where μ_i follows a mixture model with a point mass at zero and $_i$ is drawn from an unspecified prior:

$$\mu_i \overset{iid}{\longleftarrow} (1 \quad \Uparrow) \ {}_0(\cdot) + \Uparrow g_\mu(\cdot), \qquad {}_i^2 \overset{iid}{\longleftarrow} g \ (\cdot). \tag{2.2}$$

In (2.2), $_0(\cdot)$ is a Dirac delta function indicating a point mass at 0 under the null hypothesis, while $g_\mu(\cdot)$ signifies that μ_i under the alternative is drawn from an unspecified distribution which is allowed to vary across i. In this work, we focus on a model where μ_i and $_i$ are not linked by a specific function. The more challenging situation where $_i$ may be informative for predicting μ_i is briefly discussed in Section 6.2.

Following tradition in dealing with heteroscedasticity problems (e.g. Xie et al., 2012; Weinstein et al., 2018), we assume that $_i$ are known. This simplifies the discussion and

enables us to focus on key ideas. For practical applications, we use a consistent estimator of $_i$. The goal is to simultaneously test m hypotheses:

$$H_{0,i} : \mu_i = 0 \quad \text{vs.} \quad H_{1,i} : \mu_i = 0; \quad i = 1, \ldots, m. \tag{2.3}$$

The multiple testing problem (2.3) is concerned with the simultaneous inference of $\checkmark = \{\checkmark_i = I(\mu_i = 0) : i = 1, \ldots, m\}$, where $I(\cdot)$ is an indicator function. The decision rule is represented by a binary vector $= (_i : 1 \leq i \leq m) \in \{0, 1\}^m$, where $_i = 1$ means that we reject $H_{0,i}$, and $_i = 0$ means we do not reject $H_{0,i}$. The false discovery rate (FDR) (Benjamini and Hochberg, 1995), defined as

$$FDR = E\left[\frac{\sum_i (1 - \checkmark_i)\, _i}{\max\{\sum_i _i, 1\}}\right], \tag{2.4}$$

is a widely used error criterion in large-scale testing problems. A closely related criterion is the marginal false discovery rate

$$mFDR = \frac{E\{\sum_i (1 - \checkmark_i)\, _i\}}{E(\sum_i _i)}. \tag{2.5}$$

The mFDR is asymptotically equivalent to the FDR for a general set of decision rules satisfying certain first- and second-order conditions on the number of rejections (Basu et al., 2018), including p–value based tests for independent hypotheses (Genovese and Wasserman, 2002) and weakly dependent hypotheses (Storey et al., 2004). We shall show that our proposed data-driven procedure controls both the FDR and mFDR asymptotically; the main consideration of using the mFDR criterion is to derive optimality theory and facilitate methodological developments.

We use the expected number of true positives $ETP = E(\sum_{i=1}^m \checkmark_i\, _i)$ to evaluate the power of an FDR procedure. Other power measures include the missed discovery rate (MDR, Taylor et al., 2005), average power (Benjamini and Hochberg, 1995; Efron, 2007) and false negative rate or false non-discovery rate (FNR, Genovese and Wasserman, 2002; Sarkar, 2002). Cao et al. (2013) showed that under the monotone likelihood ratio condition (MLRC), maximizing the ETP is equivalent to minimizing the MDR and FNR. The ETP is used in this article because it is intuitive and simplifies the theory. We call a multiple

testing procedure valid if it controls the FDR at the nominal level and optimal if it has the largest ETP among all valid FDR procedures.

The building blocks for conventional multiple testing procedures are standardized statistics such as Z_i or P_i. Let $\mu_i^{\star} = \mu_i / \sigma_i$. The tacit rationale in conventional practice is that the simultaneous inference problem

$$H_{0,i} : \mu_i^{\star} = 0 \quad \text{vs.} \quad H_{1,i} : \mu_i^{\star} = 0; \quad i = 1, \ldots, m, \tag{2.6}$$

is equivalent to the formulation (2.3); hence the standardization step has no impact on multiple testing. However, this seemingly plausible argument, which only takes into account the null distribution, fails to consider the change in the structure of the alternative distribution. Next we present an example to illustrate the information loss and power distortion from standardizing.

2.2 Data processing and power loss: an illustrative example

The following diagram describes a data processing approach that is often adopted when performing hypothesis tests:

$$(X_i, \sigma_i) \quad \Rightarrow \quad Z_i = \frac{X_i}{\sigma_i} \quad \Rightarrow \quad P_i = 2\,(1 - |Z_i|). \tag{2.7}$$

We start with the full data consisting of X_i and $\sigma_i^2 = \mathrm{Var}(X_i \mid \mu_i)$. The data is then standardized, $Z_i = X_i / \sigma_i$, and finally converted to a two-sided p-value, P_i. Typically these p-values are ordered from smallest to largest, a threshold is chosen to control the FDR, and hypotheses with p-values below the threshold are rejected.

Here we present a simple example to illustrate the information loss that can occur at each of these data compression steps. Consider a hypothesis testing setting with $H_{0,i} : \checkmark_i = 0$ and the data coming from a normal mixture model, where

$$\mu_i \overset{iid}{\sim} (1 - \pi_0)\,\delta_0 + \pi_0\,\delta_{\mu_a}, \qquad \sigma_i \overset{iid}{\sim} U[0.5, 4]. \tag{2.8}$$

This is a special case of (2.2), where μ_i are specifically drawn from a mixture of two point masses, and where we set $\mu_a = 2$.

We examine three possible approaches to controlling the FDR at ↵ = 0.1. In the p-value approach we reject for all p-values below a given threshold. Note that, when the FDR is exhausted, this is the uniformly most powerful p-value based method (Genovese and Wasserman, 2002), so is superior to, for example, the BH procedure. Alternatively, in the z-value approach we reject for all suitably small $P(H_0|Z_i)$, which is in turn the most powerful z-value based method (Sun and Cai, 2007). Finally, in the full data approach we reject when $P(H_0|X_i, \ _i)$ is below a certain threshold, which we show later is optimal given X_i and $\ _i$. In computing the thresholds, we assume that there is an oracle knowing the alternative distribution; the formulas for our theoretical calculations are provided in Section A of the Appendix. For the model given by (2.8) these rules correspond to:

$$\begin{aligned} {}^{p} &= \{I(P_i \leq 0.0006) : 1 \leq i \leq m\} = \{I(|Z_i| \quad 3.43) : 1 \leq i \leq m\}, \\ {}^{z} &= \{I(P(H_0|Z_i) \leq 0.24) : 1 \leq i \leq m\} = \{I(Z_i \quad 3.13) : 1 \leq i \leq m\}, \\ {}^{full} &= \{I(P(H_0|X_i, \ _i) \leq 0.28) : 1 \leq i \leq m\}, \end{aligned}$$

with the thresholds chosen such that the FDR is exactly 10% for all three approaches. However, while the FDRs of these three methods are identical, the average powers, $AP(\ ) = \frac{1}{m\uparrow} E(\sum_{i=1}^{m} \checkmark_i \ _i)$, di↵er significantly:

$$AP(\ ^{p}) = 5.0\%, \quad AP(\ ^{z}) = 7.2\%, \quad AP(\ ^{full}) = 10.5\%. \tag{2.9}$$

To better understand these di↵erences consider the left hand plot in Figure 1, which illustrates the rejection regions for each approach as a function of Z and ^{2}. In the blue region all methods fail to reject the null hypothesis, while all methods reject in the black region. The green region corresponds to the space where the full data approach rejects the null while the other two methods do not. Alternatively, in the red region both the z-value and full data methods reject while the p-value approach fails to do so. Finally, in the white region the full data approach fails to reject while the z-value method does reject.

We first compare ^{z} and ^{p}. Let $\uparrow^{+}$ and $\uparrow^{-}$ denote the proportions of positive e↵ects and negative e↵ects, respectively. Then $\uparrow^{+} = 0.1$ and $\uparrow^{-} = 0$. This asymmetry of the al-

[2]The p-value method will also reject for large negative values of Z but, to keep the figure readable, we have not plotted that region.

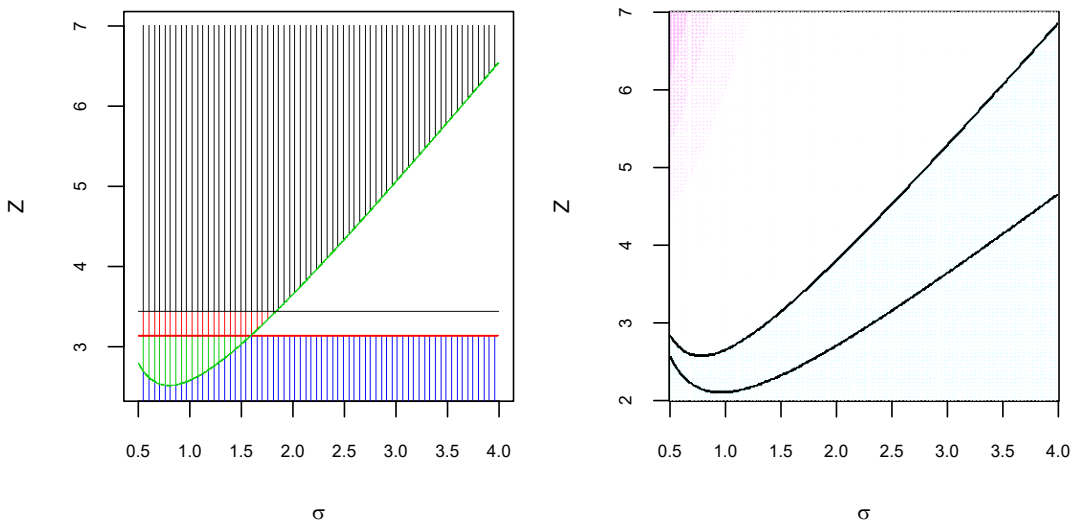

Figure 1: Left: Rejection regions for the p-value approach (black line), z-value approach (red line) and full data approach (green line) as a function of Z and . Approaches reject for all points above their corresponding line. Right: Heat map of relative proportions (on log scale) of alternative vs null hypotheses for di↵erent Z and . Blue corresponds to lower ratios and purple to higher ratios. The solid black line represents equal fractions of null and alternative, while the dashed line corresponds to three times as many alternative as null.

ternative distribution can be captured by ^{z}, which uses a one-sided rejection region. (Note that this asymmetric rejection region is not pre-specified but a consequence of theoretical derivation. In practice ^{z} can be emulated by an adaptive z-value approach that is fully data-driven (Sun and Cai, 2007).) By contrast, ^{p} enforces a two-sided rejection region that is symmetrical about 0, trading o↵ extra rejections in the region Z_i ⇡ 3.43 for fewer rejections in the region where 3.13 ⇡ Z_i ⇡ 3.43. As all nonzero e↵ects are positive, negative z-values are highly unlikely to come from the alternative; this accounts for the 2.2% loss in AP for the p-value method. Next consider ^{full} vs ^{z}. The full data approach trades o↵ extra rejections in the green space for fewer rejections in the white space. This may seem like a sub-optimal trade-o↵ given that the green space is smaller. However, the green space actually contains many more true alternative hypotheses. Approximately 3.8% of the true alternatives occur in the green region as opposed to only 0.5% in the white region, which accounts for the 3.3% higher AP for the full data approach.

At first Figure 1 may appear counterintuitive. Why should we reject for low z-values in the green region but fail to reject for high z-values in the white region? The key observation

here is that not all z-values are created equal. In the green region the observed data is far more consistent with the alternative hypothesis than the null hypothesis. For example, with Z = 4 and = 0.5 our observed X is four standard deviations from the null mean but exactly equal to the alternative mean. Alternatively, while it is true that in the white region the high z-values suggest that the data are inconsistent with the null hypothesis, they are also highly inconsistent with the alternative hypothesis. For example, with Z = 4 and = 2 our observed X is 8, which is four standard deviations from the null mean, but also three standard deviations from the alternative mean. Given that 90% of observations come from the null hypothesis, we do not have conclusive evidence as to whether this data is from the null or alternative. A z-value of 4 with = 0.5 is far more likely to come from the alternative hypothesis than is a z-value of 4 with = 2.

The right hand plot of Figure 1 makes this clear. Here we have plotted (on a log scale) the relative proportions of alternative vs null hypotheses for di↵erent Z and . Blue corresponds to lower ratios and purple to higher ratios. The solid black line represents equal fractions of null and alternative, while the dashed line corresponds to three times as many alternative as null. Clearly, for the same z-value, alternative hypotheses are relatively more common for low values. Notice how closely the shape of the dashed line maps the green rejection boundary in the left hand plot, which indicates that the full data method is correctly capturing the regions with most alternative hypotheses. By contrast, the p-value and z-value methods fail to correctly adjust for di↵erent values of .

Figure 2 provides one further way to understand the e↵ect of standardizing the data. Here we have plotted the density functions of Z under the null hypothesis (black solid) and alternative hypothesis (red dashed) for di↵erent values of . The densities have been multiplied by the relative probability of each hypothesis occurring so points where the densities cross correspond to an equal likelihood for either hypothesis. The blue line represents an observation, which is fixed at Z = 2 in each plot. The alternative density is centered at Z = 2/ so when is large the standardized null and alternative are very similar, making it hard to know which distribution Z = 2 belongs to. As decreases the standardized alternative distribution moves away from the null and becomes more consistent with Z = 2. However, eventually the alternative moves past Z = 2 and it again becomes unclear which distribution our data belongs to. Standardizing means that the null hypothesis is consistent

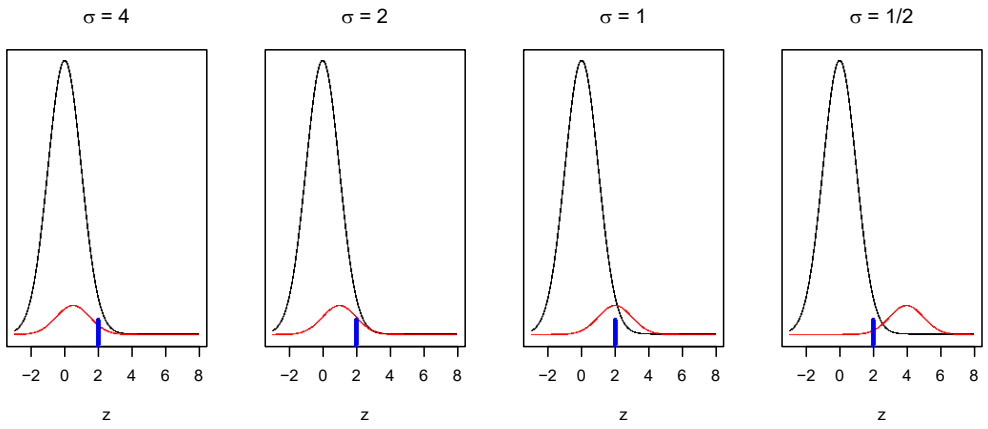

Figure 2: Plots of the density functions of Z under the null hypothesis (black solid) and alternative hypothesis (red dashed) for di↵erent values of . The blue line represents an observation at Z = 2.

for all values of , but the alternative hypothesis can change dramatically as a function of the standard deviation.

To summarize, the information loss incurred in both steps of data processing (2.7) reveals the essential role of the alternative distribution in simultaneous testing. This structure of the alternative is not captured by the p-value, which is calculated only based on the null. Our result (2.9) in the toy example shows that by exploiting (i) the overall asymmetry of the alternative via the z-value and (ii) the heterogeneity among individual alternatives via the full data, the average power of conventional p-value based methods can be doubled.

2.3 Heteroscadasticity and empirical null distribution

In the context of simultaneous testing with composite null hypotheses, Sun and McLain (2012) argued that the conventional testing framework, which involves rescaling or standardization, can become problematic:

"In multiple testing problems where the null is simple ($H_{0,i} : \mu_i = 0$), the heteroscedasticity in errors can be removed by rescaling all $_i$ to 1. However, when the null is composite, such a rescaling step would distort the scientific question."

Sun and McLain (2012) further proposed the concept of empirical composite null as an extension of Efron's empirical null (Efron, 2004a) for testing composite nulls $H_{0,i} : \mu_i$ 2

$[-a_0, a_0]$ under heteroscedastic models. It is important to note that the main message of this article, which focuses on the impact of heteroscedastiticy on the alternative instead of the null, is fundamentally different from that in Sun and McLain (2012). In fact, we show that even when the null is simple, the heteroscedasticity still matters. Our finding, which somehow contradicts the above quotes, is more striking and even counter-intuitive. Moreover, we shall see that our data-driven HART procedure, which is based on Tweedie's formula (or the f-modeling approach, Efron, 2011), is very different from the deconvoluting kernel method (or g-modeling approach) in Sun and McLain (2012)[3]. The new two-step bivariate estimator in Section 3.2 is novel and highly nontrivial; the techniques employed in the proofs of theory are also very different.

3 HART: Heteroscedasticity Adjusted Ranking and Thresholding

The example in the previous section presents a setting where hypothesis tests based on the full data (X_i, σ_i) can produce higher power than that from only using the standardized data Z_i. In this section we formalize this idea and show that the result holds in general for heteroscedasticity problems. In Section 3.1 we first assume that the distributional information is known and derive an oracle rule based on the full data. Then Section 3.2 develops data-driven schemes and computational algorithms to implement the oracle rule. Finally theoretical properties of the proposed method are established in Section 3.3.

3.1 The oracle rule under heteroscedasity

Note that the models given by (2.1) and (2.2) imply that

$$X_i \mid \sigma_i \overset{ind}{\sim} f_{\pi,\sigma_i}(x) = (1-\pi) f_{0,\sigma_i}(x) + \pi f_{1,\sigma_i}(x), \tag{3.1}$$

[3] The deconvoluting kernel method has an extremely slow convergence rate. Our numerical studies show that the method in Sun and McLain (2012) only works for composite nulls where the uncertainties in estimation can be smoothed out over an interval $[-a_0, a_0]$. However, the deconvoluting method is highly unstable and does not work well when testing simple nulls $H_{0,i} : \mu_i = 0$. Our numerical results show that the two-step method in Section 3.2 works much better.

where $f_{0,\ }(x) = \frac{1}{\ }(x/\ )$ is the null density, $f_{1,\ }(x) = \frac{1}{\ }\int_{\ }^{R} {}^{x\ \mu} g_{\mu}(\mu)d\mu$ is the alter-native density, (x) is the density of a standard normal variable, and $f\ (x)$ is the mixture density. In conventional practice, the data are standardized as $Z_i = X_i/\ _i$, and the follow-ing mixture model is used

$$Z_i \overset{iid}{\sim} f(z) = (1 \quad \uparrow)f_0(z) + \uparrow f_1(z), \tag{3.2}$$

where $f_0(z) = \ (z)$, $f_1(z)$ is the non-null density, and $f(z)$ is the mixture density of the z-values. As discussed previously, a standard approach involves converting the z-value to a two-sided p–value $P_i = 2\ (\ |Z_i|)$, where $(\cdot)$ is the standard normal cdf. The mixture model based on p-values is

$$P_i \overset{iid}{\sim} g(p) = (1 \quad \uparrow)I_{[0,1]}(p) + \uparrow g_1(p), \quad \text{for } p\ 2\ [0,1], \tag{3.3}$$

where $I(\cdot)$ is an indicator function, and $g(\cdot)$ and $g_1(\cdot)$ are the mixture density and non-null density of the p-values, respectively. Models 3.2 and 3.3 provide a powerful and flexible framework for large-scale inference and have been used in a range of related problems such as signal detection, sparsity estimation and multiple testing [e.g. Efron et al. (2001); Storey (2002); Genovese and Wasserman (2002); Donoho and Jin (2004); Newton et al. (2004); Jin and Cai (2007)].

The oracle FDR procedures for Models 3.2 and 3.3 are both known. We first review the oracle z-value procedure (Sun and Cai, 2007). Define the local FDR (Efron et al., 2001)

$$\text{Lfdr}_i = P(H_0|z_i) = P(\checkmark_i = 0|z_i) = \frac{(1 \quad \uparrow)f_0(z_i)}{f(z_i)}. \tag{3.4}$$

Then Sun and Cai (2007) showed that the optimal z-value FDR procedure is given by

$$^{z} = [I\{\text{Lfdr}(z_i) < c^{\leftarrow}\} : 1 \leq i \leq m], \tag{3.5}$$

where $c^{\leftarrow}$ is the largest Lfdr threshold such that mFDR $\leq$ ↵. Similarly, Genovese and

Wasserman (2002) showed that the optimal p-value based FDR procedure is given by

$$^{p} = [I\{P_i < c^{\leftarrow}\} : 1 \leq i \leq m], \tag{3.6}$$

where $c^{\leftarrow}$ is the largest p-value threshold such that mFDR $\leq \alpha$.

Next we derive the oracle rule based on m pairs $\{(x_i, \ _i) : i = 1, \ldots, m\}$. This new problem can be recast and solved in the framework of multiple testing with a covariate sequence. Consider Model 3.1 and define the heterogeneity–adjusted significance index[4]

$$T_i \equiv T(x_i, \ _i) = P(\checkmark_i = 0 | x_i, \ _i) = \frac{(1 \quad \hat{\cdot})f_{0, \ _i}(x_i)}{f \ _i(x_i)}. \tag{3.7}$$

Let Q(t) denote the mFDR level of the testing rule $[I\{T_i < t\} : 1 \leq i \leq m]$. Then the oracle full data procedure is denoted

$$^{full} = [I\{T_i < t^{\leftarrow}\} : 1 \leq i \leq m], \tag{3.8}$$

where $t^{\leftarrow} = \sup\{t : Q(t) \leq \alpha\}$.

The next theorem provides the key result showing that ^{full} has highest power amongst all α–level FDR rules based on $\{(x_i, \ _i) : i = 1, \cdots, m\}$.

Theorem 1 Let D_α be the collection of all testing rules based on $\{(x_i, \ _i) : i = 1, \ldots, m\}$ such that mFDR $\leq \alpha$. Then ETP $\geq$ ETP $_{full}$ for any $2 D_\alpha$. In particular we have

$$ETP \ _p \geq ETP \ _z \geq ETP \ _{full}.$$

Based on Theorem 1, our proposed methodology employs a heteroscedasticity–adjusted ranking and thresholding (HART) rule that operates in two steps: first rank all hypotheses according to T_i and then reject all hypotheses with $T_i \leq t^{\leftarrow}$. We discuss in Section 3.2 our finite sample approach for implementing HART using estimates for T_i and $t^{\leftarrow}$.

[4]Note that the oracle statistic $P(\checkmark_i = 0 | X_i, \ _i)$ is equivalent to $P(\checkmark_i = 0 | Z_i, \ _i)$ since the pairs $(X_i, \ _i)$ and $(Z_i, \ _i)$ contain the same amount of information. We use the pairs $(X_i, \ _i)$ in the next formula just to facilitate the development of estimation procedures.

3.2 Data-driven procedure and computational algorithms

We first discuss how to estimate T_i and then turn to $t^{\leftarrow}$. Inspecting T_i's formula (3.7), the null density $f_{0,\ i}(x_i)$ is known and the non-null proportion $\uparrow$ can be estimated by $\hat{\uparrow}$ using existing methods such as Storey's estimator (Storey, 2002) or Jin-Cai's estimator (Jin and Cai, 2007). Hence we focus on the problem of estimating $f_{\ i}(x_i)$.

There are two possible approaches for implementing this step. The first involves directly estimating $f_{\ i}(x_i)$ while the second is implemented by first estimating $f_{1,\ i}(x_i)$ and then computing the marginal distribution via

$$\hat{f}_{\ i}(x_i) = (1 \quad \hat{\uparrow})f_{0,\ i}(x_i) + \hat{\uparrow}\hat{f}_{1,\ i}(x_i). \tag{3.9}$$

Our theoretical and empirical results strongly suggest that this latter approach provides superior results so we adopt this method.

Remark 1 The main concern about the direct estimation of $f_{\ i}(x_i)$ is that the tail areas of the mixture density are of the greatest interest in multiple testing but unfortunately the hardest parts to accurately estimate due to the small number of observations in the tails. The fact that $f_{\ i}(x_i)$ appears in the denominator exacerbates the situation. The decomposition in (3.9) increases the stability of the density by incorporating the known null density.

Standard bivariate kernel methods (Silverman, 1986; Wand and Jones, 1994) are not suitable for estimating $f_{1,\ i}(x_i)$ because, unlike a typical variable, $_i$ plays a special role in a density function and needs to be modeled carefully. Fu et al. (2020) recently addressed a closely related problem using the following weighted bivariate kernel estimator:

$$\hat{f}^{\leftarrow}(x) := \sum_{j=1}^{m} \frac{h(\ \ _j)}{\sum_{j=1}^{m} h(\ \ _j)} h_{xj}(x \quad x_j), \tag{3.10}$$

where $h = (h_x, h\ )$ is a pair of bandwidths, $h(\ \ _j)/\{\sum_{j=1}^{m} h(\ \ _j)\}$ determines the contribution of $(x_j, \ _j)$ based on $_j$, $h_{xj} = h_{x\ j}$ is a bandwidth that varies across j, and $h(z) = \frac{1}{\sqrt{2\uparrow}h}\exp\left\{\frac{z^2}{2h^2}\right\}$ is a Gaussian kernel. The variable bandwidth h_{xj} up-weights/down-weights observations corresponding to small/large $_j$; this suitably adjusts

for the heteroscedasticity in the data.

Let $M_1 = \{i : \checkmark_i = 1\}$. In the ideal setting where $\checkmark_j$ is observed one could extend (3.10) to estimate $f_{1,\ i}(x_i)$ via

$$\tilde{f}_{1,\ i}(x) = \sum_{j \in M_1} \frac{K_h(\ \ _j)}{\sum_{k \in M_1} K_h(\ \ _k)} h_{x_j}(x \quad x_j). \tag{3.11}$$

Given that $\checkmark_j$ is unknown, we cannot directly implement (3.11). Instead we apply a weighted version of (3.11),

$$\hat{f}_{1,\ i}(x_i) = \sum_{j=1}^{m} \frac{\hat{w}_j K_h(\ _i \ \ _j)}{\sum_{k=1}^{m} \hat{w}_k K_h(\ _i \ \ _k)} h_{x_j}(x_i \quad x_j) \tag{3.12}$$

with weights $\hat{w}_j$ equal to an estimate of $P(\checkmark_j = 1 | x_j, \ _j)$. In particular we adopt a two step approach:

1. Compute $\hat{f}^{(0)}_{1,\ i}(x_i)$ via (3.12) with initial weights $\hat{w}^{(0)}_j = (1 \quad \hat{T}^{(0)}_j)$ for all j, where $\hat{T}^{(0)}_j = \min\left\{\frac{(1 \quad \hat{\pi})f_{0,\ j}(x_j)}{\hat{f}_j(x_j)}, 1\right\}$, $\hat{\pi}$ is the estimated non-null proportion, and $\hat{f}_j(x_j)$ is computed using (3.10).

2. Compute $\hat{f}^{(1)}_{1,\ i}(x_i)$ via (3.12) with updated weights $\hat{w}^{(1)}_j = (1 \quad \hat{T}^{(1)}_j)$ where

$$\hat{T}^{(1)}_j = \frac{(1 \quad \hat{\pi})f_{0,\ j}(x_j)}{(1 \quad \hat{\pi})f_{0,\ j}(x_j) + \hat{\pi}\hat{f}^{(0)}_{1,\ j}(x_j)}.$$

This leads to our final estimate for $T_i = P(H_0 | x_i, \ _i)$:

$$\hat{T}_i = \hat{T}^{(2)}_i = \frac{(1 \quad \hat{\pi})f_{0,\ i}(x_i)}{(1 \quad \hat{\pi})f_{0,\ i}(x_i) + \hat{\pi}\hat{f}^{(1)}_{1,\ i}(x_i)}.$$

In the next section, we carry out a detailed theoretical analysis to show that both $\hat{f}_{\ i}(x_i)$ and $\hat{T}_i$ are consistent estimators with $E\|\hat{f}_{\ i} \quad f_{\ i}\|^2 = E\int\{\hat{f}_{\ i}(x) \quad f_{\ i}(x)\}^2 dx \to 0$ and $\hat{T}_i \xrightarrow{P} T_i$, uniformly for all i.

To implement the oracle rule (3.8), we need to estimate the optimal threshold t^*, which can be found by carrying out the following simple stepwise procedure.

Procedure 1 (data-driven HART procedure) Rank hypotheses by increasing order of $\hat{T}_i$. Denote the sorted ranking statistics $\hat{T}_{(1)} \leq \cdots \leq \hat{T}_{(m)}$ and $H_{(1)}, \ldots, H_{(m)}$ the corresponding hypotheses. Let

$$k = \max\left\{ j : \frac{1}{j} \sum_{i=1}^{j} \hat{T}_{(i)} \leq \alpha \right\}.$$

Then reject the corresponding ordered hypotheses, $H_{(1)}, \ldots, H_{(k)}$.

The idea of the above procedure is that if the first j hypotheses are rejected, then the moving average $\frac{1}{j}\sum_{i=1}^{j} \hat{T}_{(i)}$ provides a good estimate of the false discovery proportion, which is required to fulfill the FDR constraint. Comparing with the oracle rule (3.8), Procedure 1 can be viewed as its plug-in version:

$$\text{dd} = \{I(\hat{T}_i \leq t^{\text{dd}}) : 1 \leq i \leq m\}, \quad \text{where } t^{\text{dd}} = \hat{T}_{(k)}. \tag{3.13}$$

The theoretical properties of Procedure 1 are studied in the next section.

3.3 Theoretical properties of Data-Driven HART

In Section 3.1, we have shown that the (full data) oracle rule $^{\text{full}}$ (3.8) is valid and optimal for FDR analysis. This section discusses the key theoretical result, Theorem 2, which shows that the performance of $^{\text{full}}$ can be achieved by its finite sample version $^{\text{dd}}$ (3.13) when $m \to \infty$. Inspecting (3.13), the main steps involve showing that both $\hat{T}_i$ and t^{dd} are "close" to their oracle counterparts. To ensure good performance of the proposed procedure, we require the following conditions.

(C1) $\text{supp}(g_0) \subset (M_1, M_2)$ and $\text{supp}(g_\mu) \subset (-M, M)$ for some $M_1 > 0$, $M_2 < 1$, $M < \infty$.

(C2) The kernel function K is a positive, bounded and symmetric function satisfying $\int K(t)\,dt = 1$, $\int tK(t)\,dt = 0$ and $\int t^2 K(t)\,dt < \infty$. The density function $f(t)$ has bounded and continuous second derivative and is square integrable.

(C3) The bandwidths satisfy $h_x = o\{(\log m)^{-1}\}$, $\lim_{m\to\infty} m h_x h^2 = \infty$, $\lim_{m\to\infty} m^{-1} h\, h_x^2 = \infty$ and $\lim_{m\to\infty} m^{\epsilon/2} h^2 h_x^{-1} \neq 0$ for some $\epsilon > 0$.

(C4) $\pi_0 \to^p \pi_0$.

Remark 2 For Condition (C2), the requirement on f is standard in density estimation theory, and the requirements on the kernel K is satisfied by our choice of a Gaussian kernel. Condition (C3) is satisfied by standard choices of bandwidths in Wand and Jones (1994) and Silverman (1986). The Jin-Cai estimator (Jin and Cai, 2007) fulfills Condition (C4) in a wide class of mixture models.

Our theory is divided into two parts. The next proposition establishes the theoretical properties of the proposed density estimator $\hat{f}$ and the plug-in statistic $\hat{T}_i$. The convergence of $\hat{t}^{\leftarrow}$ to $t^{\leftarrow}$ and the asymptotic properties of ^{dd} are established in Theorem 2.

Proposition 1 Suppose Conditions (C1) to (C4) hold. Then

$$E\|\hat{f} - f\|^2 = E\int \{\hat{f}(x) - f(x)\}^2 dx \to 0,$$

where the expectation E is taken over $(\mathbf{X}, \ , \boldsymbol{\mu})$. Further, we have $\hat{T}_i \xrightarrow{P} T_i$.

Next we turn to the performance of our data-driven procedure ^{dd} when $m \to 1$. A key step in the theoretical development is to show that $t^{\leftarrow} \to \hat{t}^{\leftarrow}$, where $t^{\leftarrow}$ and $\hat{t}^{\leftarrow}$ are defined in (3.13) and (3.8), respectively.

Theorem 2 Under the conditions in Proposition 1, we have $\hat{t}^{\leftarrow} \xrightarrow{P} t^{\leftarrow}$. Further, both the mFDR and FDR of ^{dd} are controlled at level $\alpha + o(1)$, and $ETP_{dd}/ETP_{full} = 1 + o(1)$.

In combination with Theorem 1, these results demonstrate that the proposed finite sample HART procedure (Procedure 1) is asymptotically valid and optimal.

4 Simulation

We first describe the implementation of HART in Section 4.1. Section 4.2 presents results for the general setting where $_i$ comes from a continuous density function. In Section 4.3, we further investigate the effect of heterogeneity under a mixture model where $_i$ takes on one of two distinct values. Simulation results for additional settings, including a non-Guassian alternative, unknown $_i$, weak dependence structure, non-Gaussian noise, estimated empirical null, correlated μ_i and $_i$, and the global null, are provided in Section E of the Supplementary Material.

4.1 Implementation of HART

The accurate estimation of $\hat{T}_i$ is crucial for ensuring good performance of the HART procedure. The key quantity is the bivariate kernel density estimator $\hat{f}_{1,}(x)$, which depends on the choice of tuning parameters $h = (h_x, h)$. Note that the ranking and selection process in Procedure 1 only involves small $\hat{T}_i$. To improve accuracy, the bandwidth should be chosen based on the pairs $(x_i, _i)$ that are less likely to come from the null. We first implement Jin and Cai's method (Jin and Cai, 2007) to estimate the overall proportion of non-nulls in the data, denoted $\hat{}$. We then compute h_x and h by applying Silverman's rule of thumb (Silverman, 1986) to the subset of the observations $\{x_i : P_i < \hat{}\}$. When implementing HART, we first estimate $f(x)$ using the data without $(X_i, _i)$, and then plug-in the unused data $(X_i, _i)$ to calculate $\hat{T}_i$. This method can increase the stability of the density estimator. The asymptotic property of this approach is established in Proposition 1.

4.2 Comparison in general settings

We consider simulation settings according to Models 2.1 and 2.2, where $_i$ are uniformly generated from $U[0, _{max}]$. We then generate X_i from a two-component normal mixture model

$$X_i | _i \overset{iid}{\sim} (1 \quad \hat{})N(0, _i^2) + \hat{}N(2, _i^2).$$

In the first setting, we fix $_{max} = 4$ and vary $\hat{}$ from 0.05 to 0.15. In the second setting, we fix $\hat{} = 0.1$ and vary $_{max}$ from 3.5 to 4.5. Five methods are compared: the ideal full data oracle procedure (OR), the z-value oracle procedure of Sun and Cai (2007) (ZOR)[5], the Benjamini-Hochberg procedure (BH), AdaPT (Lei and Fithian, 2018), and the proposed data–driven HART procedure (DD). Note that we do not include methods that explore the usefulness of sparsity structure (Scott et al., 2015; Boca and Leek, 2018; Li and Barber, 2019; Cai et al., 2019) since the primary objective here is to incorporate structural information encoded in $_i$. Also, although Ignatiadis et al. (2016) mention the idea of using $_i$ as a covariate to construct weighted p-values, no guidance is given on how to do so, and since the way in which $_i$ are incorporated is particularly important, we exclude it.

[5]We omit the comparison with the adaptive z-value (AZ) method in Sun and Cai (2007), the data-driven version of ZOR, as AZ is dominated by ZOR.

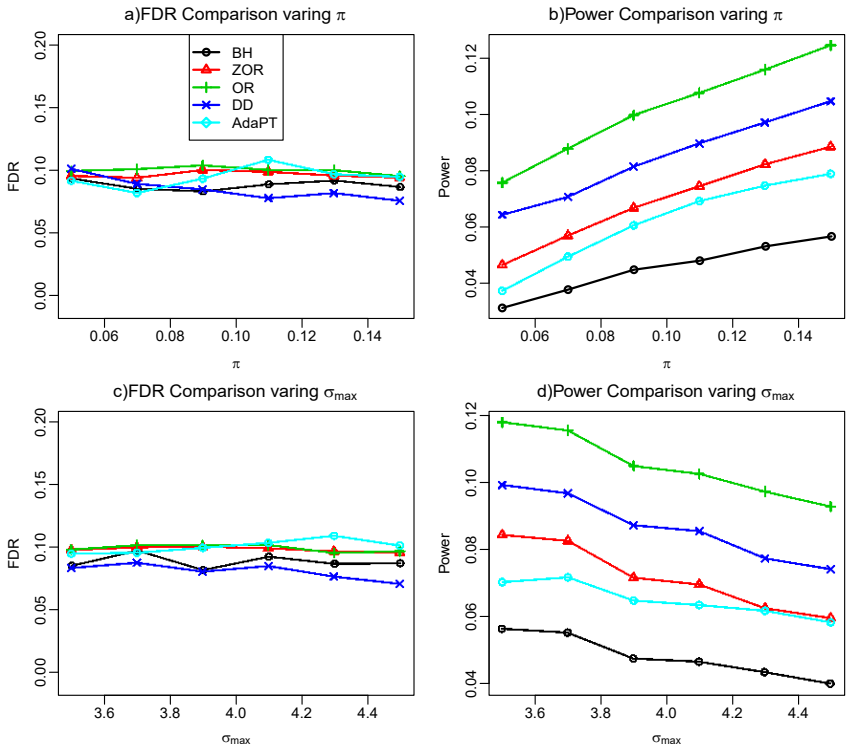

Figure 3: Comparison when $_i$ is generated from a uniform distribution. We vary ⇡ in the top row and $_{max}$ in the bottom row. All methods control the FDR at the nominal level. DD has roughly the same FDR but higher power compared to ZOR in all settings.

The nominal FDR level is set to ↵ = 0.1. For each setting, the number of tests is m = 20,000. Each simulation is also run over 100 repetitions. Then, the FDR is estimated as the average of the false discovery proportion FDP( ) = $\sum_{i=1}^{m}\{(1 \quad \surd_i)\ _i\}/(\sum_{i=1}^{m} \ _i \_ 1)$ and the average power is estimated as the average proportion of true positives that are correctly identified, $\sum_{i=1}^{m}(\surd_i \ _i)/(mp)$, both over the number of repetitions. The results for di↵ering values of ⇡ and $_{max}$ are respectively displayed in the first and second rows of Figure 3.

Next we discuss some important patterns of the plots and provide interpretations.

(a) Panel (a) of Figure 3 shows that all methods appropriately control FDR at the nominal level, with DD being slightly conservative.

(b) Panel (b) illustrates the advantage of the proposed HART procedure over existing methods. When ⇡ is small, the power of OR can be 60% higher than ZOR. This

shows that exploiting the structural information of the variance can be extremely beneficial. DD has lower power compared to OR due to the inaccuracy in estimation. However, DD still dominates ZOR and BH in all settings.

(c) ZOR dominates BH and the e ciency gain increases as ⇡ increases. To explain the power gain of ZOR over BH, let ⇡+ and ⇡ denote the proportion of true positive signals and true negative signals, respectively. Then ⇡+ = ⇡ and ⇡ = 0. This asymmetry can be captured by ZOR, which uses a one-sided rejection region. By contrast, BH adopts a two-sided symmetric rejection region. Under the setting being considered, the power loss due to the conservativeness of BH is essentially negligible, whereas the failure of capturing important structural information in the alternative accounts for most power loss.

(d) From the second row of Figure 3, we can again see that all methods control the FDR at the nominal level. OR dominates the other three methods in all settings. DD is less powerful than OR but has a clear advantage over ZOR with slightly lower FDR and higher power.

(e) In most cases, AdaPT outperforms BH. However, it is important to note that pre-ordering based on $_i$ is a suboptimal way for using side information. Moreover, the dominance of AdaPT over BH is not uniform (Section E.1 in the supplement). This shows that pre-ordering based on $_i$ can be anti-informative and lead to possible power loss for AdaPT. By contrast, HART utilizes the side information in a principled and systematic way. It uniformly improves competitive methods.

Finally, it should be noted that incorporating side information comes with computational costs: conventional methods including BH and ZOR both run considerably faster than DD. However, DD runs faster than AdaPT.

4.3 Comparison under a two-group model

To illustrate the heteroscedasticity e↵ect more clearly, we conduct a simulation using a simpler model where $_i$ takes on one of two distinct values. The example illustrates that

the heterogeneity adjustment is more useful when there is greater variation in the standard deviations among the testing units.

Consider the setup in Models 2.1 and 2.2. We first draw σ_i randomly from two possible values $\{\sigma_a, \sigma_b\}$ with equal probability, and then generate X_i from a two-point normal mixture model $X_i \mid \sigma_i \overset{iid}{\sim} (1-\pi)N(0, \sigma_i^2) + \pi N(\mu, \sigma_i^2)$. In this simpler setting, it is easy to show that HART reduces to the CLfdr method in Cai and Sun (2009), where the conditional Lfdr statistics are calculated for separate groups defined by σ_a and σ_b. As previously, we apply BH, ZOR, OR and DD to data with $m = 20,000$ tests and the experiment is repeated on 100 data sets. We fix $\pi = 0.1$, $\mu = 2.5$, $\sigma_a = 1$ and vary σ_b from 1.5 to 3. The FDRs and powers of different methods are plotted as functions of σ_b, with results summarized in the first row of Figure 4. In the second row, we plot the group-wise z-value cutoffs and group-wise powers as functions of σ_b for the DD method.

We can see that DD has almost identical performance to OR, and the power gain over ZOR becomes more pronounced as σ_b increases. This is intuitive, because more variation in σ tends to lead to more information loss in standardization. The bottom row shows the z-value cutoffs for ZOR and DD for each group. We can see that in comparison to ZOR, which uses a single z-value cutoff, HART uses different cutoffs for each group. The z-value cutoff is bigger for the group with larger variance, and the gap between the two cutoffs increases as the degree of heterogeneity increases. In Panel d), we can see that the power of Group b decreases as σ_b increases. These interesting patterns corroborate those we observed in our toy example in Section 2.2. □

5 Data Analysis

This section compares the adaptive z-value procedure [AZ, the data-driven implementation of ZOR in Sun and Cai (2007)], BH, and HART on a microarray data set. The data set measures expression levels of 12,625 genes for patients with multiple myeloma, 36 for whom magnetic resonance imaging (MRI) detected focal lesions of bone (lesions), and 137 for whom MRI scans could not detect focal lesions (without lesions) of bone (Tian et al., 2003). For each gene, we calculate the differential gene expression levels (X_i) and standard errors (S_i). The FDR level is set at $\alpha = 0.1$.

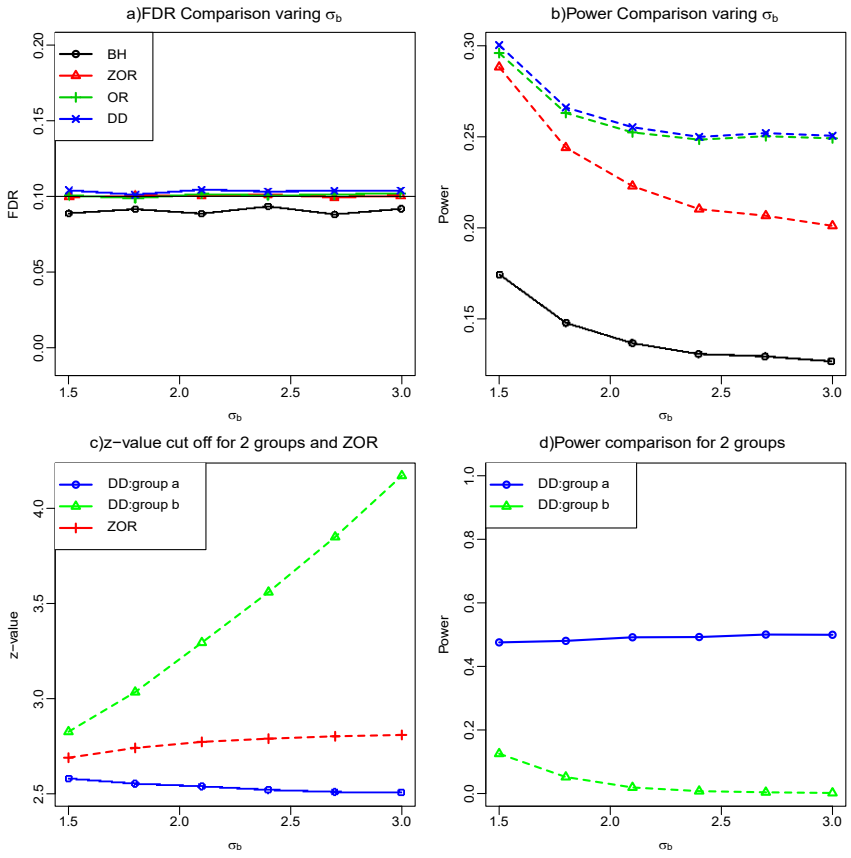

Figure 4: Two groups with varying $_b$ from 1.5 to 3. As $_b$ increases, the cut-o↵ for group a decreases whereas the cut-o↵ for group b increases. The power for tests in group b drops quickly as $_b$ increases. This corroborates our calculations in the toy example in Section 2.2 and the patterns revealed by Figure 1.

We first address two important practical issues. The first issue is that the theoretical null $N(0, 1)$ (red curve on the left panel of Figure 5) is much narrower compared to the histogram of z-values. Efron (2004b) argued that a seemingly small deviation from the theoretical z-curve can lead to severely distorted FDR analysis. For this data set, the analysis based on the theoretical null would inappropriately reject too many hypotheses, resulting in a very high FDR. To address the distortion of the null, we adopted the empirical null approach (Efron, 2004b) in our analysis. Specifically, we first used the middle part of the histogram, which contains 99% of the data, to estimate the null distribution as $N(0, 1.3^2)$ [see Efron (2004b) for more details]. The new p-values are then converted from the z-values based on the estimated empirical null: $P_i^{\leftarrow} = 2\ ^{\leftarrow}(\ 2|Z_i|)$, where $^{\leftarrow}$ is the

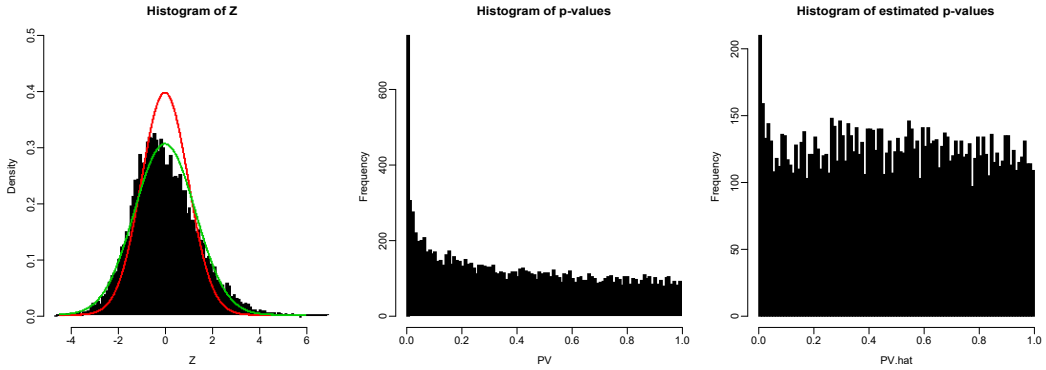

Figure 5: Left: histogram of z-values: the estimated empirical null $N(0, 1.3^2)$ (green line) seems to provide a better fit to the data compared to the theoretical null $N(0, 1)$ (red line). Middle: histogram of original p-values. Right: histogram of estimated p-values based on the empirical null. The z-value histogram suggests that the theoretical null is inappropriate (too narrow, leading to too many rejections). The use of an empirical null corrects the non-uniformity of the histogram of the p-values.

CDF of a $N(0, 1.3^2)$ variable. We can see from Figure 5 that the empirical null (green curve) provides a better fit to the histogram of z-values. Another piece of evidence for the suitability of the empirical null approach is that the histogram of the estimated p-values (right panel) looks closer to uniform compared to that of original p-values (middle panel). The uniformity assumption is crucial for ensuring the validity of p-value based procedures.

The second issue is the estimation of $f(x)$, which usually requires a relatively large sample size to ensure good precision. Figure 6 presents the histogram of S_i and scatter plot of S_i vs Z_i. Based on the histogram, we propose to only focus on data points with S_i less than 1 (12172 out of 12625 genes are kept in the analysis) to ensure the estimation accuracy of $\hat{T}_i$. Compared to conventional approaches, there is no e ciency loss because no hypothesis with $S_i > 1$ is rejected by BH at $\leftarrow = 0.1$ – note that the BH p-value cuto↵ is $6 \rightarrow 10^{-5}$, which corresponds to a z-value cuto↵ of 5.22; see also Figure 7. (If BH rejects hypotheses with large S_i, we recommend to carry out a group-wise FDR analysis, which first tests hypotheses at ↵ in separate groups and then combines the testing results, as suggested by Efron (2008a).)

Finally we apply BH, AZ and HART to the data points with $S_i < 1$. BH uses the new p-values $P_i^{\leftarrow}$ based on the estimated empirical null $N(0, 1.3^2)$. Similarly AZ uses Lfdr statistics where the null is taken as the density of a $N(0, 1.3^2)$ variable. When implementing HART,

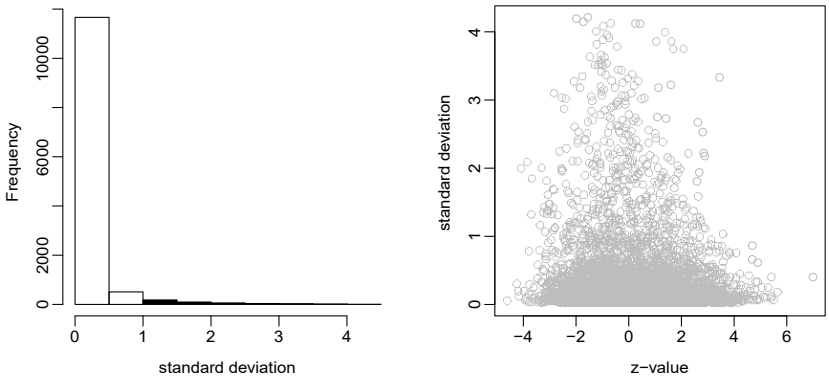

Figure 6: Histogram of S_i (left), scatter plot of (Z_i, S_i) (right)

Table 1: Numbers of genes (% of total) that are selected by each method.

↵-level	BH	AZ	HART
0.1	8 (0.07%)	25 (0.2%)	122 (1%)

we estimate the non-null proportion ⇡ using Jin-Cai's method with the empirical null taken as $N(0, 1.3^2)$. We further employ the jacknifed method to estimate $f(x)$ by following the steps in Section 4.1. We summarize the number of rejections by each method in Table 1 and display the testing results in Figure 7, where we have marked rejected hypotheses by each method using di↵erent colors.

HART rejects more hypotheses than BH and AZ. The numbers should be interpreted with caution as BH and AZ have employed the empirical null $N(0, 1.3^2)$ whereas HART has utilized null density $N(0, \ ^2_i)$ conditioned on individual $\ _i$ – it remains an open question how to extend the empirical null approach to the heteroscedastic case. Since we do not know the ground truth, it is di cult to assess the power gains. However, the key point of this analysis, and the focus of our paper, is to compare the shapes of rejection regions to gain some insights on the di↵erences between the methods. It can be seen that for this data set, the rejection rules of BH and AZ only depend on Z_i. By contrast, the rejection region for HART depends on both Z_i and S_i. HART rejects more z-values when S_i is small compared to BH and AZ. Moreover, HART does not reject any hypothesis when S_i is large. This pattern is consistent with the intuitions we gleaned from the illustrative example (Figure 1) and the results we observed in simulation studies (Figure 4, Panel c).

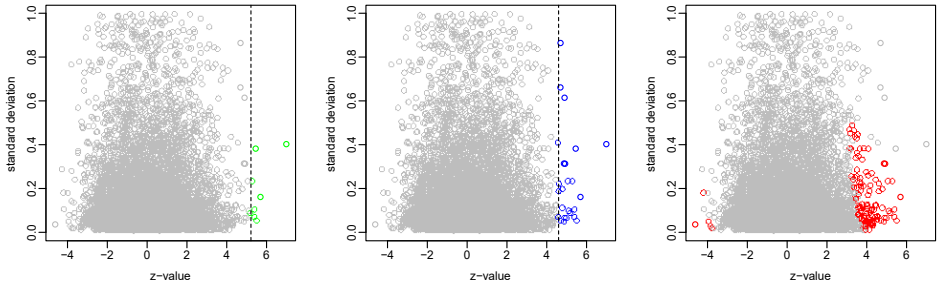

Figure 7: Scatter plot of rejected hypotheses by each method. Green: BH, blue: AZ, red: HART. AZ and BH reject every hypothesis to the right of the dashed line. The rejection region for HART depends on both z and .

6 Discussion

6.1 Multiple testing with side information

Multiple testing with side or auxiliary information is an important topic that has received much attention recently. The research directions are wide-ranging as there are various types of side information, which may be either extracted from the same data using carefully constructed auxiliary sequences or gleaned from secondary data sources such as prior studies, domain-specific knowledge, structural constraints and external covariates. The recent works by Xia et al. (2019), Li and Barber (2019) and Cai et al. (2019) have focused on utilizing side information that encodes the sparsity structure. By contrast, our work investigates the impact of the alternative distribution, showing that incorporating ¡ can be extremely useful for improving the ranking and hence the power in multiple testing [6].

In the context of FDR analysis, the key issue is that the hypotheses become unequal in light of side information. Efron (2008b) argued that ignoring the heterogeneity among study units may lead to FDR rules that are ine cient, noninterpretable and even invalid. We discuss two lines of work to further put our main contributions in context and to guide future research developments.

Grouping, pioneered by Efron (2008b), provides an e↵ective strategy for capturing the heterogeneity in the data. Cai and Sun (2009) showed that the power of FDR procedures

[6] A method is said to have better ranking if it rejects more true positives than its competitor at the same FDR level. Theorem 1 in Section 3.1 shows that the oracle HART procedure has the optimal ranking in the sense that it has the largest power among all FDR procedures at level ↵.

can be much improved by utilizing new ranking statistics adjusted for grouping. Recent works along this direction, including Liu et al. (2016), Barber and Ramdas (2017) and Sarkar and Zhao (2017), develop general frameworks for dealing with a class of hierarchical and grouping structures. However, the groups can be characterized in many ways and the optimal grouping strategy still remains unknown. Moreover, discretizing a continuous covariate by grouping leads to loss of information. HART directly incorporates $_i$ into the ranking statistic and hence eliminates the need to define groups.

Weighting is another widely used strategy for incorporating side information into FDR analyses (Benjamini and Hochberg, 1997; Genovese et al., 2006; Roquain and Van De Wiel, 2009; Basu et al., 2018). For example, when the sparsity structure is encoded by a covariate sequence, weighted p-values can be constructed to up-weight the tests at coordinates where signals appear to be more frequent (Hu et al., 2010; Xia et al., 2019; Li and Barber, 2019). However, the derivation of weighting functions for directly incorporating heteroscedasticity seems to be rather complicated (Peña et al., 2011; Habiger et al., 2017). Notably, Habiger (2017) developed novel weights for p-values as functions of a class of auxiliary parameters, including $_i$ as a special case, for a generic two-group mixture model. However, the formulation is complicated and the weights are hard to compute – the methodology requires handling the derivative of the power function, estimating several unknown quantities and tuning a host of parameters.

6.2 Open issues and future directions

We conclude the article by discussing several open issues. First, HART works better for large-scale problems where the density with heteroscedastic errors can be well estimated. For problems with several hundred tests or fewer, p-value based algorithms such as BH or the WAMDF approach (Habiger, 2017) are more suitable. The other promising direction for dealing with smaller-scale problems, suggested by Castillo and Roquain (2018), is to employ spike and slab priors to produce more stable empirical Bayes estimates (with frequentist guarantees under certain conditions). Second, in practice the model given by (2.2) can be extended to

$$\mu_i \mid {}_i \overset{\text{ind}}{\dashleftarrow} (1 - \uparrow_i)\ {}_0(\cdot) + \uparrow_i g_\mu(\cdot \mid {}_i), \qquad {}_i^2 \overset{\text{iid}}{\dashleftarrow} g\ (\cdot), \tag{6.1}$$

where both the sparsity level and distribution of non-null e↵ects depend on $_i$; this setting has been considered in a related work by Weinstein et al. (2018). The heterogeneity-adjusted statistic is then given by

$$T_i = P(\checkmark_i = 0|x_i, \ _i) = \frac{(1 \ \ \Updownarrow \ _i)f_{0,\ i}(x_i)}{f \ _i(x_i)}, \tag{6.2}$$

where the varying proportion $\Updownarrow \ _i$ indicates that $_i$ also captures the sparsity structure. This is possible, for example, in applications where observations from the alternative have larger variances compared to those from the null. An interesting, but challenging, direction for future research is to develop methodologies that can simultaneously incorporate both the sparsity and heterocedasticity structures into inference. Third, the HART-type methodology can only handle one covariate sequence $\{ \ _i : 1 \leq i \leq m\}$. It would be of great interest to develop new methodologies and principles for information pooling for multiple testing with several covariate sequences. Finally, our work has assumed that $_i$ are known in order to illustrate the key message (i.e. the impact of the alternative distribution on the power of FDR analyses). Although this is a common practice, it is desirable to carefully investigate the impact of estimating $_i$ on the accuracy and stability of large-scale inference, and to develop more accurate simultaneous estimation procedures for unknown $_i$. We have provided some empirical results in Appendix E.2 but a rigorous theoretical study, along the lines of Fan et al. (2007) and Kosorok and Ma (2007), will be of much interest for future research.

References

Barber, R. F. and Ramdas, A. (2017). The p-filter: multilayer false discovery rate control for grouped hypotheses. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(4):1247–1268.

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018). Weighted false discovery rate control in large-scale multiple testing. Journal of the American Statistical Association, 113:1172–1183.

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

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. Scandinavian Journal of Statistics, 24:407–418.

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

Boca, S. M. and Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariate. Journal of the American Statistical Association, 6:e6035.

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

Cai, T. T., Sun, W., and Wang, W. (2019). Cars: covariate assisted ranking and screening for large-scale two-sample inference (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81:187–234.

Cao, H., Sun, W., and Kosorok, M. R. (2013). The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing. Biometrika, 100(2):495–502.

Castillo, I. and Roquain, E. (2018). On spike and slab empirical bayes multiple testing. arXiv:1808.09748.

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

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. Statist. Sci., 18(1):71–103.

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

Efron, B. (2004b). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. Journal of the American Statistical Association, 99(465):96–104.

Efron, B. (2007). Size, power and false discovery rates. Ann. Statist., 35:1351–1377.

Efron, B. (2008a). Microarrays, empirical Bayes and the two-groups model. Statist. Sci., 23:1–22.

Efron, B. (2008b). Simultaneous inference: When should hypothesis testing problems be combined? Ann. Appl. Stat., 2:197–223.

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

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

Fan, J., Hall, P., and Yao, Q. (2007). To how many simultaneous hypothesis tests can normal, student's t or bootstrap calibration be applied? Journal of the American Statistical Association, 102(480):1282–1288.

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

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

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. Biometrika, 93(3):509–524.

Habiger, J., Watts, D., and Anderson, M. (2017). Multiple testing with heterogeneous multinomial distributions. Biometrics, 73(2):562–570.

Habiger, J. D. (2017). Adaptive false discovery rate control for heterogeneous data. Statistica Sinica, pages 1731–1756.

Harvey, C. R. and Liu, Y. (2015). Backtesting. The Journal of Portfolio Management, 42(1):13–28.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, 6:65–70.

Hu, J. X., Zhao, H., and Zhou, H. H. (2010). False discovery rate control with groups. Journal of the American Statistical Association, 105:1215–1227.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nature methods, 13(7):577.

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

Kosorok, M. R. and Ma, S. (2007). Marginal asymptotics for the large p, small n paradigm: with applications to microarray data. The Annals of Statistics, 35(4):1456–1486.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(4):649–679.

Li, A. and Barber, R. F. (2019). Multiple testing with the structure-adaptive benjamini–hochberg algorithm. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81(1):45–74.

Liu, Y., Sarkar, S. K., and Zhao, Z. (2016). A new approach to multiple testing of grouped hypotheses. Journal of Statistical Planning and Inference, 179:1–14.

Miller, C., Genovese, C., Nichol, R., Wasserman, L., Connolly, A., Reichart, D., Hopkins, D., Schneider, J., and Moore, A. (2001). Controlling the false-discovery rate in astrophysical data analysis. Astronomical Journal, 122:3492–3505.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting di↵erential gene expression with a semiparametric hierarchical mixture method. Biostatistics, 5(2):155–176.

Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. Journal of the American Statistical Association, 99(468):1002–1014.

Peña, E. A., Habiger, J. D., and Wu, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. Annals of statistics, 39(1):556.

Roquain, E. and Van De Wiel, M. A. (2009). Optimal weighting for false discovery rate control. Electronic journal of statistics, 3:678–711.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. Ann. Statist., 30:239–257.

Sarkar, S. K. and Zhao, Z. (2017). Local false discovery rate based methods for multiple testing of one-way classified hypotheses. arXiv:1712.05014.

Schwartzman, A., Dougherty, R. F., and Taylor, J. E. (2008). False discovery rate analysis of brain di↵usion direction maps. Ann. Appl. Stat., 2(1):153–175.

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015). False discovery rate regression: An application to neural synchrony detection in primary visual cortex. Journal of the American Statistical Association, 110(510):459–471.

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

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

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

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

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

Sun, W. and Wei, Z. (2011). Large-scale multiple testing for pattern identification, with applications to time-course microarray experiments. J. Amer. Statist. Assoc., 106:73–88.

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

Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., and Shaughnessy, J. D. (2003). The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma. New England Journal of Medicine, 349(26):2483–2494. PMID: 14695408.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U. S. A., 98(9):5116–5121.

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

Weinstein, A., Ma, Z., Brown, L. D., and Zhang, C.-H. (2018). Group-linear empirical Bayes estimates for a heteroscedastic normal mean. Journal of the American Statistical Association, 113:698–710.

Xia, Y., Cai, T. T., and Sun, W. (2019). Gap: A general framework for information pooling in two-sample sparse inference. Journal of the American Statistical Association, 0(just-accepted):to appear.

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

Zhao, Z., De Stefani, L., Zgraggen, E., Binnig, C., Upfal, E., and Kraska, T. (2017). Controlling false discoveries during interactive data exploration. In Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD '17, pages 527–540, New York, NY, USA. ACM.

Supplementary Material for "Heterocedasticity-Adjusted Ranking and Thresholding for Large-Scale Multiple Testing"

This supplement contains the formulas for the illustrative example (Section A), proofs of main theorems (Section B), propositions (Section C), and lemmas (Section D), as well as additional numerical results (Section E).

A Formulas for the Illustrative Example

Consider Model 2.8 in Section 2.2. We derive the formulas for the oracle p-value, oracle z-value and oracle full data procedures.

- ^{p} corresponds to the thresholding rule $I(|Z_i| > t_p)$, where

$$t_p = \inf\left\{ t > 0 : \frac{2(1-\pi)\tilde{\Phi}(t)}{2(1-\pi)\tilde{\Phi}(t) + \pi\int(\tilde{\Phi}(t+\mu_a) + \tilde{\Phi}(t-\mu_a))\,dG(\cdot)} \le \right\},$$

  with $\tilde{\Phi}$ being the survival function of the $N(0,1)$ variable.

- ^{z} is a one-sided thresholding rule of the form $I(Z_i > t_z)$, where

$$t_z = \inf\left\{ t > 0 : \frac{(1-\pi)\tilde{\Phi}(t)}{(1-\pi)\tilde{\Phi}(t) + \pi\int\tilde{\Phi}(t-\mu_a)dG(\cdot)} \le \right\}.$$

- ^{full} is of the form $I\{P(\checkmark_i = 0|x_i, \ _i) < \ \}$. It can be written as $I\{Z_i > t_{z,\ (\ )}\}$, where

$$t_{z,\ (\ )} = \frac{\mu_a^2 \quad 2 \ ^2\log\left\{\frac{\pi}{(1-\ )(1-\pi)}\right\}}{2\mu_a}.$$

  Denote $^{\leftarrow}$ the optimal threshold. Hence ^{full} is given by $I\{P(\checkmark_i = 0|x_i, \ _i) < {}^{\leftarrow}\}$, where

$${}^{\leftarrow} = \sup\left\{ \ \in [0,1] : \frac{(1-\pi)\int\tilde{\Phi}\{t_{z,\ (\ )}\}dG(\ )}{(1-\pi)\int\tilde{\Phi}\{t_{z,\ (\ )}\}dG(\ ) + \pi\int\tilde{\Phi}\{t_{z,\ (\ )} - \mu_a\}dG(\ )} \right\}.$$

The optimal cuto↵s can be solved numerically from the above. The powers are given

by

$$AP(\ ^{p}) = \int \left\{ \tilde{\ }\downarrow t_p + \underline{\mu_a}\ \right\} + \left\{ \tilde{\ }\downarrow t_p \quad \underline{\mu_a}\ \right\} o \quad dG(\ ),$$

$$AP(\ ^{z}) = \int \tilde{\ }\downarrow t \quad \underline{\mu_a}\ \ dG(\ ),$$

$$AP(\ ^{full}) = \int \tilde{\ } n \quad t_{z,}(\ ) \quad \underline{\mu_a} o \quad dG(\ ).$$

B Proofs of Theorems

B.1 Proof of Theorem 1

We divide the proof into two parts. In Part (a), we establish two properties of the testing rule $^{full}(t) = \{I(T_i < t) : 1 \leq i \leq m\}$ for an arbitrary $0 < t < 1$. In Part (b) we show that the oracle rule $^{full}(t^{\leftarrow})$ attains the mFDR level exactly and is optimal amongst all FDR procedures at level $\leftarrow$.

Part (a). Denote $\leftarrow(t)$ the mFDR level of $^{full}(t)$. We shall show that (i) $\leftarrow(t) < t$ for all $0 < t < 1$ and that (ii) $\leftarrow(t)$ is nondecreasing in t. Note that $E\{\sum_{i=1}^{m}(1 \quad \surd_i)\ _i\} = E_{X,}\ (\sum_{i=1}^{m} T_{i\ i})$. According to the definition of $\leftarrow(t)$, we have

$$E_{X,} \left(\sum_{i=1}^{m} \{T_i \qquad \leftarrow(t)\} I(T_i \leq t) \right) = 0. \tag{B.1}$$

We claim that $\leftarrow(t) < t$. Otherwise if $\leftarrow(t) \quad t$, then we must have $T_i < t \leq \leftarrow(t)$. It follows that the LHS must be negative, contradicting (B.1).

Next we show (ii). Let $\leftarrow(t_j) = \leftarrow_j$. We claim that if $t_1 < t_2$, then we must have $\leftarrow_1 \leq \leftarrow_2$. We argue by contradiction. Suppose that $t_1 < t_2$ but $\leftarrow_1 > \leftarrow_2$. Then

$$\begin{aligned}(T_i \quad \leftarrow_2)I(T_i < t_2) &= (T_i \quad \leftarrow_1)I(T_i < t_1) + (\leftarrow_1 \quad \leftarrow_2)I(T_i < t_1) + (T_i \quad \leftarrow_2)I(t_1 \leq T_i < t_2) \\ &\quad (T_i \quad \leftarrow_1)I(T_i < t_1) + (\leftarrow_1 \quad \leftarrow_2)I(T_i < t_1) + (T_i \quad \leftarrow_1)I(t_1 \leq T_i < t_2).\end{aligned}$$

It follows that $E\{\sum_{i=1}^{m}(T_i \quad \leftarrow_2)I(T_i < t_2)\} > 0$ since $E\{\sum_{i=1}^{m}(T_i \quad \leftarrow_1)I(T_i < t_1)\} = 0$ according to (B.1), $\leftarrow_1 > \leftarrow_2$ and $T_i \quad t_1 > \leftarrow_1$, contradicting (B.1). Hence we must have $\leftarrow_1 < \leftarrow_2$.

Part (b). Let $\bar{\lambda} = \lambda(1)$. In Part (a), we show that $\lambda(t)$ is non–decreasing in t. It follows that for all $\lambda < \bar{\lambda}$, there exists a $t^{\leftarrow}$ such that $t^{\leftarrow} = \sup\{t : \lambda(t^{\leftarrow}) = \lambda\}$. By definition, $t^{\leftarrow}$ is the oracle threshold. Consider an arbitrary decision rule $d = (d_1, \ldots, d_m) \in \{0,1\}^m$ such that $mFDR(d) \leq \lambda$. We have $E\left\{\sum_{i=1}^m (T_i - \lambda)\, \delta_i^{full}\right\} = 0$ and $E\{\sum_{i=1}^m (T_i - \lambda) d_i\} \geq 0$. Hence

$$E\left(\sum_{i=1}^m (\delta_i^{full} - d_i)(T_i - \lambda)\right) \leq 0. \tag{B.2}$$

Consider transformation $f(x) = (x - \lambda)/(1 - x)$. Note that $f(x)$ is monotone, we rewrite $\delta_i^{full} = I[\{(T_i - \lambda)/(1 - T_i)\} < \gamma]$, where $\gamma = (t_{\leftarrow} - \lambda)/(1 - t^{\leftarrow})$. In Part (a) we have shown that $\lambda < t_{OR} < 1$, which implies that $\gamma > 0$. Hence

$$E\left[\sum_{i=1}^m (\delta_i^{full} - d_i)\{(T_i - \lambda) - \gamma(1 - T_i)\}\right] \geq 0. \tag{B.3}$$

To see this, consider the terms where $\delta_i^{full} - d_i = 0$. Then we have two situations: (i) $\delta_i^{full} > d_i$ or (ii) $\delta_i^{full} < d_i$. In situation (i), $\delta_i^{full} = 1$, implying that $\{(T_i - \lambda)/(1 - T_i)\} < \gamma$. In situation (ii), $\delta_i^{full} = 0$, implying that $\{(T_i - \lambda)/(1 - T_i)\} \geq \gamma$. Therefore we always have $(\delta_i^{full} - d_i)\{(T_i - \lambda) - \gamma(1 - T_i)\} \geq 0$. Summing over the m terms and taking the expectation yield (B.3). Combining (B.2) and (B.3), we obtain

$$0 \geq E\left(\sum_{i=1}^m (\delta_i^{full} - d_i)(T_i - \lambda)\right) \geq E\left(\sum_{i=1}^m (\delta_i^{full} - d_i)(T_i - \lambda)\right).$$

Finally, since $\gamma > 0$, it follows that $E\left\{\sum_{i=1}^m (\delta_i^{full} - d_i)(T_i - \lambda)\right\} > 0$. Finally, we apply the definition of ETP to conclude that $ETP(\delta^{full}) \geq ETP(d)$ for all $d \in D_\lambda$.

B.2 Proof of Theorem 2

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

- $Q(t) = m^{-1} \sum_{i=1}^m (T_i - \lambda) I\{T_i < t\}$.
- $\hat{Q}(t) = m^{-1} \sum_{i=1}^m (\hat{T}_i - \lambda) I\{\hat{T}_i < t\}$.
- $Q_1(t) = E\{(T_{OR} - \lambda) I\{T_{OR} < t\}\}$.
- $t_1 = \sup\{t \in (0,1) : Q_1(t) \geq 0\}$: the "ideal" threshold.

For $T_{OR}^{(k)} < t < T_{OR}^{(k+1)}$, define a continuous version of $\hat{Q}(t)$ as

$$\hat{Q}_C(t) = \frac{t - \hat{p}_{OR}^{(k)}}{\hat{p}_{OR}^{(k+1)} - \hat{p}_{OR}^{(k)}} \hat{Q}_k + \frac{\hat{p}_{OR}^{(k+1)} - t}{\hat{p}_{OR}^{(k+1)} - \hat{p}_{OR}^{(k)}} \hat{Q}_{k+1},$$

where $\hat{Q}_k = \hat{Q}\{\hat{p}_{OR}^{(k)}\}$. Since $\hat{Q}_C(t)$ is continuous and monotone, its inverse $\hat{Q}_C^{-1}$ is well–defined, continuous and monotone. Next we show the following two results in turn: (i) $\hat{Q}(t) \to^p Q_1(t)$ and (ii) $\hat{Q}_C^{-1}(0) \to^p t_1$.

To show (i), note that $Q(t) \to^p Q_1(t)$ by the WLLN, so that we only need to establish that $\hat{Q}(t) - Q(t) \to^p 0$. We need the following lemma, which is proven in Section D.

Lemma 1 Let $U_i = (T_i - t)I(T_i < t)$ and $\hat{U}_i = (\hat{T}_i - t)I\{\hat{T}_i < t\}$. Then $E\{\hat{U}_i - U_i\}^2 = o(1)$.

By Lemma 1 and Cauchy-Schwartz inequality, $E\{\hat{U}_i - U_i\}\{\hat{U}_j - U_j\} = o(1)$. Let $S_m = \sum_{i=1}^m \{\hat{U}_i - U_i\}$. It follows that

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

By Proposition 1, $E(m^{-1}S_m) \to 0$, applying Chebyshev's inequality, we obtain $m^{-1}S_m = \hat{Q}(t) - Q(t) \to^p 0$. Hence (i) is proved. Notice that $Q_1(t)$ is continuous by construction, we also have $\hat{Q}(t) \to^p \hat{Q}_C(t)$.

Next we show (ii). Since $\hat{Q}_C(t)$ is continuous, for any $\varepsilon > 0$, we can find $\delta > 0$ such that $\hat{Q}_C^{-1}(0) - \hat{Q}_C^{-1}\{\hat{Q}_C(t_1)\} < \varepsilon$ if $\hat{Q}_C(t_1) < \delta$. It follows that

$$P\{\hat{Q}_C(t_1) > \delta\} \le P\{\hat{Q}_C^{-1}(0) - \hat{Q}_C^{-1}\{\hat{Q}_C(t_1)\} > \varepsilon\}.$$

Proposition 1 and the WLLN imply that $\hat{Q}_C(t) \to^p Q_1(t)$. Note that $Q_1(t_1) = 0$. Then $P\{\hat{Q}_C(t_1) > \delta\} \to 0$. Hence we have $\hat{Q}_C^{-1}(0) \to^p \hat{Q}_C^{-1}\{\hat{Q}_C(t_1)\} = t_1$, completing the proof of (ii).

To show $FDR(^{dd}) = FDR(^{full}) + o(1) = \alpha + o(1)$, we only need to show $mFDR(^{dd}) = mFDR(^{full}) + o(1)$. The result then follows from the asymptotic equivalence of FDR and $mFDR$, which was proven in Cai et al. (2019).

Define the continuous version of $Q(t)$ as $Q_C(t)$ and the corresponding threshold as $Q_C^{-1}(0)$. Then by construction, we have

$$\hat{R}^{dd} = \{ i : \hat{T}_i \le \hat{Q}_C^{-1}(0), 1 \le i \le m \} \quad \text{and} \quad R^{full} = \{ i : T_i \le Q_C^{-1}(0), 1 \le i \le m \}.$$

Following the previous arguments, we can show that $\hat{Q}_C^{-1}(0) \to_p t_1$. It follows that $\hat{Q}_C^{-1}(0) = Q_C^{-1}(0) + o_p(1)$. By construction $mFDR(R^{full}) = \alpha$. The mFDR level of R^{dd} is

$$mFDR(R^{dd}) = \frac{P_{H_0}\{ \hat{T}_i \le \hat{Q}_C^{-1}(0) \}}{P\{ \hat{T}_i \le \hat{Q}_C^{-1}(0) \}}.$$

From Proposition 2, $\hat{T}_i \to_p T_i$. Using the continuous mapping theorem, $mFDR^{dd} = mFDR^{full} + o(1) = \alpha + o(1)$. The desired result follows.

Finally, using the fact that $\hat{T}_i \to_p T_i$ and $\hat{Q}_C^{-1}(0) \to_p Q_C^{-1}(0)$, we can similarly show that

$$ETP(R^{dd})/ETP(R^{full}) = 1 + o(1).$$

C Proof of Proposition 1

Summary of notation

The following notation will be used throughout the proofs:

- $\hat{f}^{(k)}(x) = \sum_{j=1}^{m} \frac{\vec{h}(\hat{\pi}_j)}{\sum_{i=1}^{m} h(\hat{\pi}_i)} h_{x_j}(x - x_j)$.
- $\hat{f}^{(k)}_{1,-j}(x) = \sum_{j=1}^{m} \frac{\vec{h}(\hat{\pi}_j) I(\hat{\sigma}_j = 1)}{\sum_{i=1}^{m} h(\hat{\pi}_i) I(\hat{\sigma}_i = 1)} h_{x_j}(x - x_j)$.
- $\tilde{f}_{1,-j}(x) = \sum_{j=1}^{m} \frac{\vec{h}(\pi_j) P(\sigma_j = 1 | x_j, \pi_j)}{\sum_{i=1}^{m} h(\pi_i) P(\sigma_i = 1 | x_i, \pi_i)} h_{x_j}(x - x_j)$.
- $\hat{f}_{1,-j}(x) = \sum_{j=1}^{m} \frac{\vec{h}(\hat{\pi}_j) \hat{P}(\sigma_j = 1 | x_j, \hat{\pi}_j)}{\sum_{i=1}^{m} h(\hat{\pi}_i) \hat{P}(\sigma_i = 1 | x_i, \hat{\pi}_i)} h_{x_j}(x - x_j)$.
- $\hat{f}(x) = (1 - \hat{\pi}) f_{0,-j}(x) + \hat{\pi} \hat{f}_{1,-j}(x)$.

The basic idea is that a consistent one–step estimator constructed via $\hat{f}^{(k)}(x)$ leads to a consistent two–step estimator via $\hat{f}(x)$. By Condition (C4) and the triangle inequality, it

is sufficient to show that

$$E\int \left\{\hat{f}_{1,\lambda}(x) - f_{1,\lambda}(x)\right\}^2 dx \to 0. \tag{C.4}$$

Let $u_j = \frac{\sum_{i=1}^{m} h(\theta_j)}{\sum_{i=1}^{m} h(\theta_i)}$. A direct consequence of condition (C1) is $0 < \frac{C_1}{m} \le Eu_j \le \frac{C_2}{m} < 1$ for some positive constants C_1 and C_2. Let $C' = \min(1, C_1)$.Consider event

$$A = \left\{ \sum_{i=1}^{m} \sqrt{\lambda_j}\, m^{-\frac{1}{2}} < C\, m^{-\frac{C'}{2}} \right\}. \tag{C.5}$$

By Hoeffding's inequality and Condition (C2), $P(A^C) = O(h_x^2) \le \exp(-C'^2 m/2) = O(h_x^2) \to 0$. Therefore it suffices to prove (C.4) under A. We establish the result in three steps:

1. $E\int \{\hat{f}^{*}_{1,\lambda}(x) - f_{1,\lambda}(x)\}^2 dx \to 0.$
2. $E\int \{\tilde{f}_{1,\lambda}(x) - \hat{f}^{*}_{1,\lambda}(x)\}^2 dx \to 0.$
3. $E\int \{\hat{f}_{1,\lambda}(x) - \tilde{f}_{1,\lambda}(x)\}^2 dx \to 0.$

The proposition then follows from the triangle inequality.

C.1 Proof of Step (a)

Let $b_j^{*} = \frac{\sum_{i=1}^{m} h(\theta_j) I(\sqrt{\lambda_j} \neq \sqrt{\lambda_i})}{\sum_{i=1}^{m} h(\theta_i) I(\lambda_i = 1)}$. It is easy to show that

$$\left\{\hat{f}^{*}_{1,\lambda}(x) - f_{1,\lambda}(x)\right\}^2 = \sum_{j=1}^{m}\sum_{k=1}^{m} b_j^{*} b_k^{*} \left\{h_{x,k}(x - x_k) - f_{1,\lambda}(x)\right\}\left\{h_{x,j}(x - x_j) - f_{1,\lambda}(x)\right\}.$$

Under condition (C1) and event A, we have $E(b_j^{*} b_k^{*}) = O(m^{-2})$. Using standard arguments in density estimation theory (e.g. Wand and Jones (1994) page 21), and the fact that $E\sum_{j=1}^{m}(b_j^{*})^2 = O(m^{-1}h^{-1})$, we have $E\int \{\hat{f}^{*}_{1,\lambda}(x) - f_{1,\lambda}(x)\}^2 dx = O\left((mh\, h_x)^{-1} + h^4\right)$.

Under condition (C2) and (C3) the RHS $\to 0$, establishing Step (a).

C.2 Proof of Step (b)

Let $b_j = \frac{h(\quad_j)P(\surd_j = 1|x_j, \quad_j)}{\sum_{i=1}^m h(\quad_i)P(\surd_i = 1|x_i, \quad_i)}$. Then

$$\left\{\tilde{f}_{1,}(x) - \hat{f}_{1,}(x)\right\}^2 = \sum_{j=1}^m (b_j^{\leftarrow} - b_j)^2 \, {}^2_{h_{x\,j}}(x - x_j) + \sum_{(j,k):j\neq k} (b_j^{\leftarrow} - b_j)(b_k^{\leftarrow} - b_k) \, {}_{h_{x\,j}}(x - x_j) \, {}_{h_{x\,k}}(x - x_k). \quad (C.6)$$

We first bound $E(b_j^{\leftarrow} - b_j)^2$. Write $E(b_j^{\leftarrow} - b_j)^2 = \{E(b_j^{\leftarrow} - b_j)\}^2 + Var(b_j^{\leftarrow} - b_j)$. It is clear that $E(b_j^{\leftarrow})$ and $E(b_j)$ are both $O(m^{-1})$. Hence $\{E(b_j^{\leftarrow} - b_j)\}^2 = O(m^{-2})$. Next consider $Var(b_j^{\leftarrow} - b_j) = Var(b_j^{\leftarrow}) + Var(b_j) - 2Cov(b_j^{\leftarrow}, b_j)$. We have by condition (C3)

$$Var(b_j^{\leftarrow}) = Var\left(\frac{I(\surd_j = 1) \, h(\quad_j)}{\sum_{i=1}^m h(\quad_i) I(\surd_i = 1)}\right) \leq E(b_j^{\leftarrow})^2 = O(m^{-2}h^{-1}).$$

Similarly $Var(b_j) = O(m^{-2}h^{-1})$. It follows that $Cov(b_j^{\leftarrow}, b_j) = O(m^{-2}h^{-1})$. Therefore $Var(b_j^{\leftarrow} - b_j) = O(m^{-2}h^{-1})$ and $E(b_j^{\leftarrow} - b_j)^2 = O(m^{-2}h^{-1})$. Using the fact that $\int {}^2_{h_{x\,j}}(x - x_j)dx = O(h_x^{-1})$, we have

$$E\left[\sum_{j=1}^m (b_j^{\leftarrow} - b_j)^2 \, {}^2_{h_{x\,j}}(x - x_j)dx\right] = O\{(mh_xh)^{-1}\} \to 0. \quad (C.7)$$

Next we bound $E\{(b_j^{\leftarrow} - b_j)(b_k^{\leftarrow} - b_k)\}$ for $j \neq k$. Consider the decomposition

$$E\{(b_j^{\leftarrow} - b_j)(b_k^{\leftarrow} - b_k)\} = E(b_j^{\leftarrow} - b_j)E(b_k^{\leftarrow} - b_k) + Cov(b_j^{\leftarrow} - b_j, b_k - b_k). \quad (C.8)$$

Our goal is to show that $E\{(b_j^{\leftarrow} - b_j)(b_k^{\leftarrow} - b_k)\} = O(m^{-3}h^{-2}) + O(m^{-4}h^{-4})$. It suffices to show

$$E_{\surd|\,,\mathbf{x}}\{(b_j^{\leftarrow} - b_j)(b_k^{\leftarrow} - b_k)\} = O(m^{-3}h^{-2}) + O(m^{-4}h^{-4}). \quad (C.9)$$

Observe that $Var\left\{\frac{1}{mh^{-1}}\sum_{i=1}^m h(\quad_j)I(\surd_i = 1)\Big|\,,\mathbf{x}\right\} = O(m^{-1})$ and

$$E_{\surd|\,,\mathbf{x}}\left(\frac{1}{mh^{-1}}\sum_{i=1}^m h(\quad_j)I(\surd_i = 1)\right) = \frac{1}{mh^{-1}}\sum_{i=1}^m h(\quad_j)P(\surd_i = 1|\quad_i, x_i).$$

Applying Chebyshev's inequality,

$$\frac{1}{mh^{1}}\sum_{i=1}^{m} h(\quad_j)I(\checkmark_i = 1) \quad \frac{1}{mh^{1}}\sum_{i=1}^{m} h(\quad_j)P(\checkmark_i = 1|\ _i, x_i) \;\overset{p}{\to}\; 0.$$

It follows that for any $\varepsilon > 0$,

$$P\left(\sum_{i=1}^{m} h(\quad_j)I(\checkmark_i = 1) \sum_{i=1}^{m} h(\quad_j)P(\checkmark_i = 1|\ _i, x_i) < \varepsilon mh^{1}\right) \to 1.$$

Under A defined in (C.5), we have $\sum_{i=1}^{m} h(\quad_j)I(\checkmark_i = 1) > h^{1}C_3 m$ for some C_3, and

$$P\left(\sum_{i=1}^{m} h(\quad_j)P(\checkmark_i = 1|\ _i, x_i) < h^{1}C_3 m\right) \to 0. \qquad (C.10)$$

The boundedness of $b_j^{\leftarrow}$ and b_j and (C.10) imply that we only need to prove (C.9) on the event $A^{\leftarrow} = \{(\mathbf{x}, \quad) : \sum_{i=1}^{m} h(\quad_j)P(\checkmark_i = 1|\ _i, x_i) \quad h^{1}C_3 m\}$. We shall consider $E_{\checkmark|\ ,\mathbf{x}}(b_j^{\leftarrow} \quad b_j)$ and $\mathrm{Cov}(b_j^{\leftarrow} \quad b_j, b_k^{\leftarrow} \quad b_k|\ , \mathbf{x})$ in turn. Write

$$\begin{aligned} E_{\checkmark|\ ,\mathbf{x}}(b_j^{\leftarrow}) &= E_{\checkmark|\ ,\mathbf{x}}\left(\frac{I(\checkmark_j = 1)\ h(\quad_j)}{\sum_{i=1}^{m} h(\quad_{\vec{i}})I(\checkmark_i = 1)}\right) \\ &= P(\checkmark_j = 1|x_j, \ _j)E_{\checkmark|\ ,\mathbf{x}}\left(\frac{h(\quad)_j)\checkmark}{\sum_{i=1}^{m} h(\quad_i I(\ _i =\right. \\ &\quad \left. 1) + \mathrm{Cov}\left(\checkmark_j, \frac{h(\quad_j)}{\sum_{i=1}^{m} h(\quad}\right) \checkmark_i I(\ _i = 1)\right) \ ,\mathbf{x} \ . \end{aligned}$$

Let $Y = \left(\frac{h(\quad)_j)\checkmark}{\sum_{i=1}^{m} h(\quad_i I(\ _i = 1}\right)$. We state three lemmas that are proven in Section D.

Lemma 2 Under event $A^{\leftarrow}$, we have $E_{\checkmark|\ ,\mathbf{x}}(Y) \quad E_{\checkmark|\ ,\mathbf{x}}(Y|\checkmark_j = 1) = O(m^{2})$ and $E_{\checkmark|\ ,\mathbf{x}}(Y) \quad E_{\checkmark|\ ,\mathbf{x}}(Y|\checkmark_j = 0) = O(m^{2}h^{1})$.

Lemma 3 Under event $A^{\leftarrow}$, we have

$$E_{\checkmark|\ ,\mathbf{x}}\left(\frac{h(\quad_j)}{\sum_{i=1}^{m} h(\quad_{\vec{i}})I(\checkmark_i = 1)}\right) = \left(\frac{h(\quad_j)}{\sum_{i=1}^{m} h(\quad_i)P(\checkmark_i = 1|x_i, \ _i)}\right) + O\left(m^{2}h^{2}\right).$$

Lemma 4 Under event $A^{\leftarrow}$, we have $\mathrm{Cov}(b_j^{\leftarrow} \quad b_j, b_k^{\leftarrow} \quad b_k|\ , \mathbf{x}) = O(m^{3}h^{2})$.

According to Lemma 2, we have

$$\begin{aligned} & \mathrm{Cov}\left(\checkmark_j, \frac{\sum_{i=1}^m K_h(\quad_j)\checkmark_i}{\sum_{i=1}^m K_h(\quad_i) I(\quad_i = 1)} \,\middle|\, \mathbf{x}, \mathbf{Z}\right) \\ = & \{P(\checkmark_j = 0|x_j, \quad_j) P(\checkmark_j = 1|x_j, \quad_j)\}\{y \quad E_{\checkmark|\,\cdot,\mathbf{x}}(Y)\}^2 f_{Y|\checkmark=1,\cdot,\mathbf{x}}(y)dy \\ & \{1 \quad P(\checkmark_j = 1|x_j, \quad_j)\}\{y \quad E_{\checkmark|\,\cdot,\mathbf{x}}(Y)\}^2 f_{Y|\checkmark_j=0,\cdot,\mathbf{x}}(y)dy = O(m^{-2}h^{-1}). \end{aligned} \tag{C.11}$$

Together with Lemma 3, we have

$$E_{\checkmark|\,\cdot,\mathbf{x}} \frac{I(\checkmark_j = 1) K_h(\quad_j)}{\sum_{i=1}^m K_h(\quad_i) I(\checkmark_i = 1)} = \frac{P(\checkmark_j = 1|x_j, \quad_j) K_h(\quad_j)}{\sum_{i=1}^m K_h(\quad_i) P(\checkmark_i = 1|x_i, \quad_i)} + O(m^{-2}h^{-2}). \tag{C.12}$$

It follows that $E(b_j^{\leftarrow} \quad b_j) = O(m^{-2}h^{-2})$ and $E(b_j^{\leftarrow} \quad b_j)E(b_k^{\leftarrow} \quad b_k) = O(m^{-4}h^{-4})$. The decomposition (C.8) and Lemma 4 together imply $E\{(b_j^{\leftarrow} \quad b_j)(b_k^{\leftarrow} \quad b_k)\} = O(m^{-3}h^{-2}) + O(m^{-4}h^{-4})$. It follows that

$$E \int \sum_{(j,k):j\neq k} (b_j^{\leftarrow} \quad b_j)(b_k^{\leftarrow} \quad b_k) K_{h_x}(x \quad x_j) K_{h_x}(x \quad x_k)dx = O\{(mh_xh^2)^{-1} + O(mh)^{-2}\} \to 0. \tag{C.13}$$

Combing (C.6), (C.7) and (C.13), we conclude that $E\int \{\tilde{f}_{1,\cdot}(x) \quad \hat{f}_{1,\cdot}^{\leftarrow}(x)\}^2 dx \to 0$.

C.3 Proof of Step (c)

Let $q_j = P(\checkmark_j = 1|\quad_j, x_j)$, $\hat{q}_j = \hat{P}(\checkmark_j = 1|\quad_j, x_j) = \min\left\{\frac{(1 \quad \hat{\pi})\hat{f}_{0,\cdot}(x_j)}{\hat{f}_j^{\leftarrow}(x_j)}, 1\right\}$ and $\hat{f}_{1,\cdot}(x) = \sum_{j=1}^m \frac{K_h(\quad_j)\hat{q}_j}{\sum_{i=1}^m K_h(\quad_i)\hat{q}_i} K_{h_x}(x \quad x_j)$.

Write $\hat{q}_j = q_j + a_j$, then $|a_j| \leq 1$ and $a_j = o_P(1)$. We have

$$\begin{aligned} E\int \{\tilde{f}_{1,\cdot}(x) \quad \hat{f}_{1,\cdot}(x)\}^2 dx &= O\left(h_x^{-1} m^{-2} E\left\{\frac{K_h(\quad_j)\hat{q}_j}{\sum_{i=1}^m K_h(\quad_i)\hat{q}_i} \quad \frac{K_h(\quad_j)q_j}{\sum_{i=1}^m K_h(\quad_i)q_i}\right\}^2\right) \\ &= O\left(h_x^{-1}h^{-2}Ea_j^2\right). \end{aligned}$$

Next we explain why the last equality holds. Let $c_i = K_h(X_i - x)h$. Then

$$E\left[\left(\frac{\sum_{i=1}^m K_h(X_i - x)\hat{q}_j}{\sum_{i=1}^m K_h(X_i - x)\hat{q}_i} - \frac{\sum_{i=1}^m K_h(X_i - x)q_j}{\sum_{i=1}^m K_h(X_i - x)q_i}\right)^2\right]$$
$$= E\left[\left(\frac{a_j \sum_{i=1}^m K_h(X_i - x)q_i - q_j \sum_{i=1}^m K_h(X_i - x)a_i}{\{\sum_{i=1}^m K_h(X_i - x)\hat{q}_i\}\{\sum_{i=1}^m K_h(X_i - x)q_i\}}\right)^2\right]$$
$$= h^2 \frac{1}{m^4} O\left[4E\left(a_j \sum_{i=1}^m c_i q_i - q_j \sum_{i=1}^m c_i a_i\right)^2 5\right]$$
$$= \frac{h^2}{m^4} O\left[4E\left\{m^2 a_j^2 - 2m a_j \left(\sum_{i=1}^m a_i + \sum_{i=1}^m a_i\right)\right\}^2 5\right] = \frac{h^2}{m^2} O\left[E\left(a_j^2\right)\right].$$

The last line holds by noting that $E(a_j a_i) \le \sqrt{E(a_j^2)E(a_i^2)} = O\{E(a_j^2)\}$.

The next step is to bound $E(a_j^2)$. Note that $a_j = O\left(\frac{f_{0,j}(x_j)\{\hat{f}^{\leftarrow}_j(x_j) - f_j(x_j)\}}{f_j(x_j)\hat{f}^{\leftarrow}_j(x_j)}\right)$.

By the construction of $\hat{q}_j$, we have $\hat{f}_j(x_j) \ge (1 - \delta)f_{0,j}(x_j)$. Hence

$$a_j = O\left(1 - \frac{\hat{f}^{\leftarrow}_j(x_j)}{f_j(x_j)}\right) \quad \text{and} \quad E(a_j^2) = O\left[4E\left(1 - \frac{\hat{f}^{\leftarrow}_j(x_j)}{f_j(x_j)}\right)^2 5\right].$$

Let $K_j = \left[\mu_j - \sqrt{\log m}\,M, \mu_j + \sqrt{\log m}\,M\right]$. By the Gaussian tail bound, $P\{x_j \in K_j\} = O(m^{-M/2})$. By the boundedness of a_j^2 and the fact that $h_x^{-1}h^2 m^{-M/2} \to 0$ (Condition (C3)), we only need to consider $E\left[4\left(1 - 2\frac{\hat{f}^{\leftarrow}_j(x_j)}{f^{\leftarrow}_j(x_j)} + \left(\frac{\hat{f}_j(x_j)}{f_j(x_j)}\right)^2\right)^2 3 x_j 5\right]$ for $x_j \in K_j$.

Let $f(x_j) = \int K_h(x)\{(1-\delta)\phi_0(x_j - x) + \delta g_\mu(x_j - x)\}dx$. Define a jacknifed version of $\hat{f}^{\leftarrow,(j)}_j$ that is formed without the pair (μ_j, x_j). It follows that

$$E\{\hat{f}^{\leftarrow,(j)}_j(x_j) \mid x_j\} = \int\int \frac{1}{\sqrt{2\pi h_x^2}} K_h(x)\{(1-\delta)\phi_0(x_j - x) + \delta g_\mu(x_j - x)\} g(\mu_j)d\mu_j dx.$$

By the intermediate value theorem and Condition (C1),

$$E\{\hat{f}^{\leftarrow,(j)}_j(x_j) \mid x_j\} = \int \frac{1}{\sqrt{2\pi h_x^2 c}} K_h(x)\{(1-\delta)\phi_0(x_j - x) + \delta g_\mu(x_j - x)\}dx$$

for some constant c. Next consider the ratio

$$\frac{E\{\hat{f}_j^{(-j)}(x_j)\,|\,x_j\}}{f_j(x_j)} = \frac{\int \frac{\phi_{\sigma^2+h_x^2c}(x)}{\phi_{\sigma^2}(x)}\{(1-\pi_0)\phi_0(x_j-x)+\pi g_\mu(x_j-x)\}\,dx}{\int \{(1-\pi_0)\phi_0(x_j-x)+\pi g_\mu(x_j-x)\}\,dx}.$$

By Condition (C1), $\mathrm{supp}(g_\mu) \subset (-M, M)$ with $M < 1$, we have

$$\inf_{x_j-M<x<x_j+M} \frac{\phi_{\sigma^2+h_x^2c}(x)}{\phi_{\sigma^2}(x)} \le \frac{E\{\hat{f}_j^{(-j)}(x_j)\,|\,x_j\}}{f_j(x_j)} \le \sup_{x_j-M<x<x_j+M} \frac{\phi_{\sigma^2+h_x^2c}(x)}{\phi_{\sigma^2}(x)}.$$

Note that $x_j \in (\sqrt{2\log m} - M, \sqrt{2\log m} + M)$. The above infimum and supremum are taken over $x \in K = (\sqrt{2\log m} - 2M, \sqrt{2\log m} + 2M)$. Using Taylor expansion

$$\frac{\phi_{\sigma^2+h_x^2c}(x)}{\phi_{\sigma^2}(x)} = \frac{1}{\sqrt{1+h_x^2c/\sigma^2}}\left\{1+\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{h_x^2cx^2}{2(\sigma^2+h_x^2c)}\right)^k\right\},$$

we have $\inf_{x\in K} \frac{\phi_{\sigma^2+h_x^2c}(x)}{\phi_{\sigma^2}(x)} = \frac{1}{\sqrt{1+h_x^2c/\sigma^2}} = 1+O(h_x^2)$. Similarly,

$$\sup_{x\in K} \frac{1}{\sqrt{1+h_x^2c/\sigma^2}}\left\{1+\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{h_x^2cx^2}{2(\sigma^2+h_x^2c)}\right)^k\right\} = 1+O(h_x^2)+O\left\{\sup_{x\in K}\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{h_x^2cx^2}{2(\sigma^2+h_x^2c)}\right)^k\right\}.$$

It follows that

$$\frac{E\{\hat{f}_j^{(-j)}(x_j)\,|\,x_j\}}{f_j(x_j)} = 1+O(h_x^2)+O\left\{\sup_{x\in K}\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{h_x^2cx^2}{2(\sigma^2+h_x^2c)}\right)^k\right\}, \quad \text{(C.14)}$$

$$1-\frac{E\{\hat{f}_j^{(-j)}(x_j)\,|\,x_j\}}{f_j(x_j)} = 1+O(h_x)+O\left\{\sup_{x\in K}\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{h_x^2cx^2}{2(\sigma^2+h_x^2c)}\right)^k\right\}. \quad \text{(C.15)}$$

Next consider $E\left(\frac{\hat{f}_j^{(-j)}(x_j)}{f_j(x_j)}\,\middle|\,x_j\right) = \left\{\frac{E\{\hat{f}_j^{(-j)}(x_j)\,|\,x_j\}}{f_j(x_j)}\right\}^2 + \mathrm{Var}\left(\frac{\hat{f}_j^{(-j)}(x_j)}{f_j(x_j)}\,\middle|\,x_j\right)$. By the same computation on page 21 of Wand and Jones (1994),

$$\mathrm{Var}\left(\frac{\hat{f}_j^{(-j)}(x_j)}{f_j(x_j)}\,\middle|\,x_j\right) = O\{(mh_x)^{-1}f_j(x_j)^{-1}\} + o\{(mh_x)^{-1}f_j(x_j)^{-2}\}.$$

Since $x_j \in K_j$, $f_j(x_j) \ge C_3 m^{-/2}$ for some constant C_3, together with Condition (C3), we

have $h_x^{-1} Var\left(\frac{\hat{f}_j^{\leftarrow,(j)}(x_j)}{\hat{f}_j(x_j)} \mid x_j\right) = o(1)$. It follows from (C.14) and (C.15) that

$$h_x^{-1} \left[2h_x^{-1} E\left(\frac{\hat{f}_j^{\leftarrow,(j)}(x_j)}{\hat{f}_j(x_j)} \mid x_j\right) + h_x^{-1} E\left(\frac{\hat{f}_j^{\leftarrow,(j)}(x_j)}{\hat{f}_j(x_j)} \mid x_j\right)^2 \right]$$
$$= O\left(h_x + \sup_{k=1}^{\infty} \frac{1}{k!} \frac{h_x^{2k-1} c^k x^{2k}}{2^k (\sigma^2 + h_x^2 c)^k}\right) + o(1). \qquad (C.16)$$

By condition (C2) and the range of x, the RHS goes to 0.

Let $S_j = \sum_{i=1}^{m} h(x_j - x_i)$ and $S_j = \sum_{i=j}^{m} h(x_j - x_i)$. Some algebra shows $\hat{f}^{\leftarrow}_j(x_j) = \frac{S_j}{\hat{S}_j} \hat{f}^{\leftarrow,(j)}_j(x_j) + \frac{1}{2\hat{S}_j h h_x}$. We use the fact that $\hat{f}_j(x_j) \geq C_3 m^{-1/2}$ for some constant C_3 and Condition (C3) to claim that on A_κ,

$$h_x^{-1} \frac{E\{\hat{f}^{\leftarrow}_j(x_j) \mid x_j\}}{\hat{f}_j(x_j)} = h_x^{-1} \frac{E\{\hat{f}^{\leftarrow(j)}_j(x_j) \mid x_j\}}{\hat{f}_j(x_j)} + o(1). \qquad (C.17)$$

Similar computation shows that

$$h_x^{-1} E\left(\frac{\hat{f}^{\leftarrow}_j(x_j)}{\hat{f}_j(x_j)} \mid x_j\right)^2 = h_x^{-1} E\left(\frac{\hat{f}^{\leftarrow,(j)}_j(x_j)}{\hat{f}_j(x_j)} \mid x_j\right)^2 + o(1). \qquad (C.18)$$

(C.16), (C.17) and (C.18) together implies $h_x^{-1} h^2 E\{a_j^2 \mid x_j\} \to 0$. Hence $E \int \left(\hat{f}_{1,\kappa}(x) - \tilde{f}_{1,\kappa}(x)\right)^2 dx \to 0$ and Step (c) is established. $\hat{T}_i \to^p T_i$ then follows from Lemma A.1 and Lemma A.2 in Sun and Cai (2007).

D Proof of Lemmas

D.1 Proof of lemma 1

Using the definitions of $\hat{U}_i$ and U_i, we can show that $\left(\hat{U}_i - U_i\right)^2 = \left(\hat{T}_i - T_i\right)^2 I\left(\hat{T}_i \leq t, T_i \leq t\right) + \left(\hat{T}_i - \tau\right)^2 I\left(\hat{T}_i \leq t, T_i > t\right) + (T_i - \tau)^2 I\left(\hat{T}_i > t, T_i \leq t\right)$. Denote the three sums on the

RHS as I, II, and III respectively. By Proposition 2, $E(I) = o(1)$. Let $\varepsilon > 0$. Consider

$$P\left(\hat{T}_i \le t, T_i > t\right) \le P\left(\hat{T}_i \le t, T_i \in (t, t+\varepsilon)\right) + P\left(\hat{T}_i \le t, T_i \ge t+\varepsilon\right)$$

$$\le P\left\{T_i \in (t, t+\varepsilon)\right\} + P\left(|\hat{T}_i - T_i| > \varepsilon\right)$$

The first term on the right hand is vanishingly small as $\varepsilon \to 0$ because T_{OR} is a continuous random variable. The second term converges to 0 by Proposition 2. we conclude that $II = o(1)$. In a similar fashion, we can show that $III = o(1)$, thus proving the lemma.

D.2 Proof of lemma 2

Note that $E_{\sqrt{|},x}(Y|\sqrt{_j} = 0) \le E_{\sqrt{|},x}Y \le E_{\sqrt{|},x}(Y|\sqrt{_j} = 1)$. It is sufficient to bound $E_{\sqrt{|},x}(Y|\sqrt{_j} = 0)$ and $E_{\sqrt{|},x}(Y|\sqrt{_j} = 1)$. The lemma follows by noting that

$$E_{\sqrt{|},x}(Y|\sqrt{_j} = 0) - E_{\sqrt{|},x}(Y|\sqrt{_j} = 1)$$

$$= E_{\sqrt{|},x}\left(\frac{\sum_{i=j} h(\quad_j)}{\sum_{i=j} h(\quad_i)\sqrt{_i}}\right) - E_{\sqrt{|},x}\left(\frac{\sum_{i=j} h(\quad_j)}{\sum_{i=j} h(\quad_i)\sqrt{_i} + h(\quad_j)}\right)$$

$$= E_{\sqrt{|},x}\left(\frac{\sum_{i=j} h^2(\quad_j)}{\{\sum_{i=j} h(\quad_i)\sqrt{_i}\}\{\sum_{i=j} h(\quad_i)\sqrt{_i} + h(\quad_j)\}}\right)$$

$$\le E_{\sqrt{|},x}\left(\frac{\sum h^2(\quad_j))}{(\sum_{i=j} h(\quad_i\sqrt{_i})^2}\right) = O(m^{-2}h^{-1}).$$

D.3 Proof of lemma 3

Let $Z = \sum_{i=1}^{m} h(\quad_i)I(\sqrt{_i} = 1)$, We expand $\frac{1}{Z}$ around $E_{\sqrt{|},x}(Z)$ and take expected value:

$$E_{\sqrt{|},x}\left(\frac{1}{Z}\right) = E_{\sqrt{|},x}\left[\frac{1}{E_{\sqrt{|},x}(Z)} - \frac{E\ 1}{\{\ _{\sqrt{|},x}(Z)\}^2}(Z - E_{\sqrt{|},x}Z) + \sum_{k=3}^{1}\frac{(-1)^{k-1}}{\{E_{\sqrt{|},x}(Z)\}^k}(Z - E_{\sqrt{|},x}Z)^{k-1}\right].$$

The series converges on A. Moreover, using normal approximation of binomial distribution, it can be shown that $E_{\sqrt{|},x}(Z - E_{\sqrt{|},x}Z)^k = O((mh^{-1})^{k/2})$. The lemma follows by noting that $E_{\sqrt{|},x}\ Z^{-1} = \{E_{\sqrt{|},x}(Z)\}^{-1} + O(m^{-2}h^2)$.

D.4 Proof of lemma 4

Consider $\tilde{b}_j = \frac{h(\quad j)I_j(\checkmark = 1)}{\sum_{i=} h(\quad i)(\checkmark_i 1)}$ defined in Section C.1. Let $\tilde{b}_j = \frac{j\checkmark}{\sum_{i=} {}_1\checkmark_i}$. By Condition (C1), $Cov(\tilde{b}_j, b_k | \ , \mathbf{x}) = O\{h \ Cov(\tilde{b}_j, b_k | \ , \mathbf{x})\}$. Note that $Cov(\tilde{b}_j, b_k | \ , \mathbf{x}) =$ $E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_j \tilde{b}_k) \quad E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_j) E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_k)$. Using similar argument as in the proof for (C.12), we have

$$E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_j) = \frac{P\ \checkmark_j = 1|}{\sum_{i=1}(\ ,\mathbf{x})\checkmark_i = 1|\ ,\mathbf{x})} + O(m^{\ 2}) \text{ and } E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_k) = \frac{P(\checkmark_k = 1|\ |}{\sum_{i=1}(\ ,\mathbf{x})\checkmark_i = 1\ \ ,\mathbf{x})} + O(m^{\ 2}).$$

It follows that $E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_j) E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_k) = \frac{P\ \checkmark_j = 1|\ \ |}{\sum_{i=1}(\mathbf{x})\ P(\checkmark_i = 1\ \ ,\mathbf{x})} \frac{\dashrightarrow}{\sum_{i=1}(\mathbf{x})P(\checkmark_i = 1\ \ ,\mathbf{x})} + O(m^{\ 3})$.

Next we compute $E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_j \tilde{b}_k)$. Note that $E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_j \tilde{b}_k) = P(\checkmark_j = 1|\ ,\mathbf{x}) E_{\checkmark|\ ,\mathbf{x}} \frac{\sum_m \checkmark_k}{(\ .\ i=1\checkmark_i{}^2} \checkmark_j = 1$

Using similar arguments as the proof for (C.11), we have $Cov\ \checkmark_k, \frac{1}{(\sum_{i=1}\checkmark_i{}^2} \checkmark_j = 1, \ ,\mathbf{x} = O(m^{\ 3})$. Let $\checkmark_{\ k} = (\checkmark_j : 1 \leq j \leq m, j = k)$. We have

$$E_{\checkmark|\ ,\mathbf{x}} \frac{\checkmark_k}{(\sum_{i=1}^m \checkmark_i{}^2)} \checkmark_j = 1 \ = P(\checkmark_k = 1|\ ,\mathbf{x}) E_{\checkmark_k} \frac{1}{(\sum_{i=1}^m \checkmark_i{}^2)} \checkmark_j = 1, \ ,\mathbf{x} + O(m^{\ 3}).$$

Using similar arguments in Lemmas 3 and 2, we have

$$E_{\checkmark\ _k} \frac{1\ \ )}{(\sum_{i=1}^m \checkmark_i{}^2} \checkmark_j = 1, \ ,\mathbf{x} = \frac{1}{E_{\checkmark|\ ,\mathbf{x}}(\sum_{i=1}^m \checkmark_i)^2} + O(m^{\ 3}).$$

In the previous equation, the conditional expectation $E_{\checkmark\ _k}$ can be replaced by $E_{\checkmark}$ because the term $\checkmark_k$ only a↵ects the ratio by a term of order $O(m^{\ 3})$. Note that $E_{\checkmark|\ ,\mathbf{x}}(\sum_{i=1}^m \checkmark_i)^2$ $= E_{\checkmark|\ ,\mathbf{x}}(\sum^m \quad \sum_{i=1}^m \checkmark_i)^{\ 2} + Var(\quad \sum_{i=1}^m \checkmark_i|\ ,\mathbf{x})$ and $Var(\quad _{i=1}\checkmark_i|\ ,\mathbf{x}) \leq m$. We have

$$E_{\checkmark|\ ,\mathbf{x}} \frac{\checkmark_k}{(\sum_{i=\ 1}^m \checkmark_i{}^2} \checkmark_j = 1 \ = \frac{P\checkmark_k = 1|\ \ |}{\{\sum_{i=1}^m (\mathbf{x})P(\checkmark_i = 1\ \ ,\mathbf{x})\}^2} + O(m^{\ 3}).$$

Finally, the lemma follows from the fact that

$$Cov(\tilde{b}_j, \tilde{b}_k | \ ,\mathbf{x}) = E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_j \tilde{b}_k) \quad E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_j) E_{\checkmark|\ ,\mathbf{x}}(\tilde{b}_k) = O(m^{\ 3}).$$

E Supplementary Numerical Results

E.1 Non-Gaussian alternative

We generate σ_i uniformly from $[0.5, \sigma_{max}]$, and generate X_i according to the following model:

$$X_i \mid \sigma_i \overset{ind}{\sim} (1 - \pi)N(0, \sigma_i^2) + \pi N(\mu_i, \sigma_i^2), \quad \mu_i \overset{iid}{\sim} 0.5N(-1.5, 0.1^2) + 0.5N(2, 0.1^2).$$

In the first setting, we fix $\sigma_{max} = 2$ and vary π from 0.05 to 0.15. In the second setting, we fix $\pi = 0.1$ and vary σ_{max} from 1.5 to 2.5. Five methods are compared: the ideal full data oracle procedure (OR), the z-value oracle procedure of (Sun and Cai, 2007) (ZOR), the Benjamini-Hochberg procedure (BH), AdaPT (Lei and Fithian, 2018) (AdaPT), and the proposed data–driven HART procedure (DD). The nominal FDR level is set to $\alpha = 0.1$. For each setting, the number of tests is $m = 20,000$. Each simulation is also run over 100 repetitions. The results are summarized in Figure 8. All methods can control the FDR at the nominal level with BH slightly conservative. DD performs almost as well as OR. The ordering information from σ_i seems to help AdaPT in some cases, but in other cases, it causes AdaPT to underperform BH. There is a clear power gap between DD and ZOR.

E.2 Unknown σ_i

This section investigates the robustness of our method when σ_i is unknown. In some applications, the exact value of σ_i is unknown but can be estimated. For this simulation, we independently generate 200 copies of X_i using the following model:

$$X_i \mid \sigma_i \overset{ind}{\sim} (1 - \pi)N(0, \sigma_i^2) + \pi N(2/\sqrt{200}, \sigma_i^2), \quad \sigma_i \overset{iid}{\sim} U[0.5, \sigma_{max}].$$

For fair comparison we replace ZOR by AZ, the data driven version of ZOR described in (Sun and Cai, 2007). We use the sample standard deviation of x_i (denoted s_i) as an estimate of σ_i for DD, AZ, BH and AdaPT. OR has access to the exact value of σ_i. We then apply the testing procedures to the pairs $(\sqrt{200}\bar{x}_i, s_i)$. The z-value is computed as $\frac{\sqrt{200}\bar{x}_i}{s_i}$ and the p-value is computed using $\frac{1}{2}\{1 - \Phi(|z_i|)\}$ where Φ is the CDF for standard normal distribution. Strictly speaking the z-values should follow a t distribution, but since

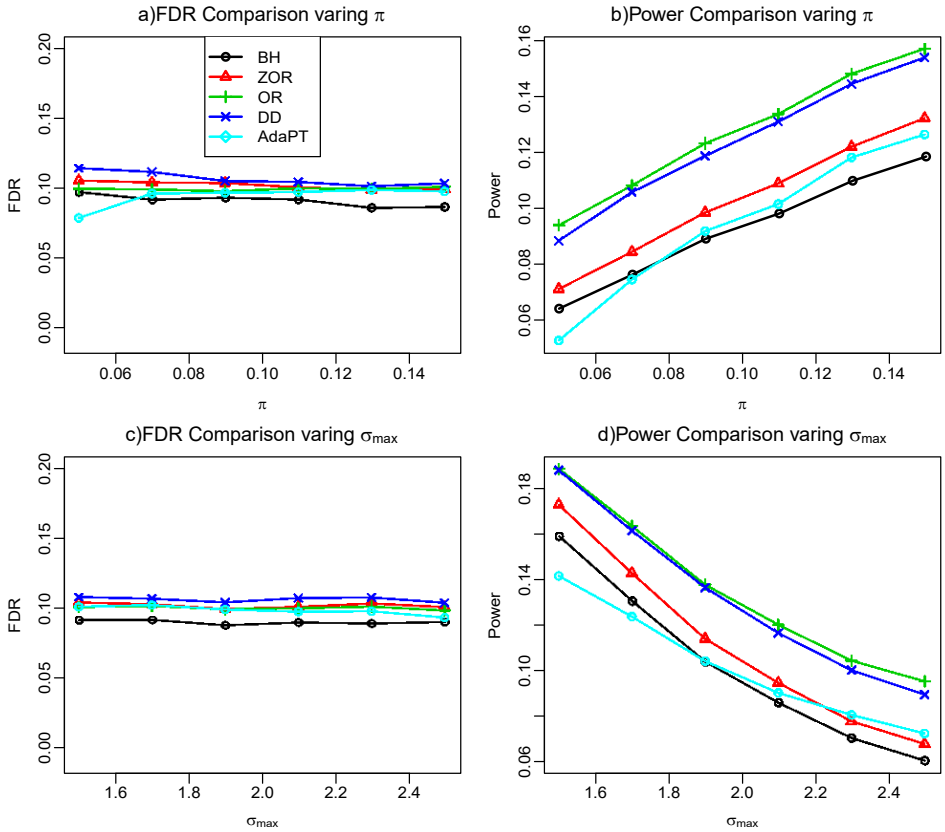

Figure 8: Comparison when $_i$ is generated from a uniform distribution and μ_i is generated from a Gaussian mixture. We vary ⇡ in the top row and $_{max}$ in the bottom row. All methods control the FDR at the nominal level. DD performs almost as well as OR and has a significant power advantage over ZOR, BH and AdaPT.

the sample size is relatively large, normal distribution serves as a good approximation. The number of tests is $m = 20,000$. Each simulation is run over 100 repetitions. In the first setting, we fix $_{max} = 4$ and vary ⇡ from 0.05 to 0.15. In the second setting, we fix $⇡ = 0.1$ and vary $_{max}$ from 3.5 to 4.5. The results are summarized in Figure 9.

We can see that all data-driven methods have been adversely a↵ected. However, the inflation of the FDR for all methods are mild and the overflow of FDR for DD is less severe compared to the other three data-driven methods. The gap in power between OR and DD becomes larger but the power advantage of DD over other methods is maintained.

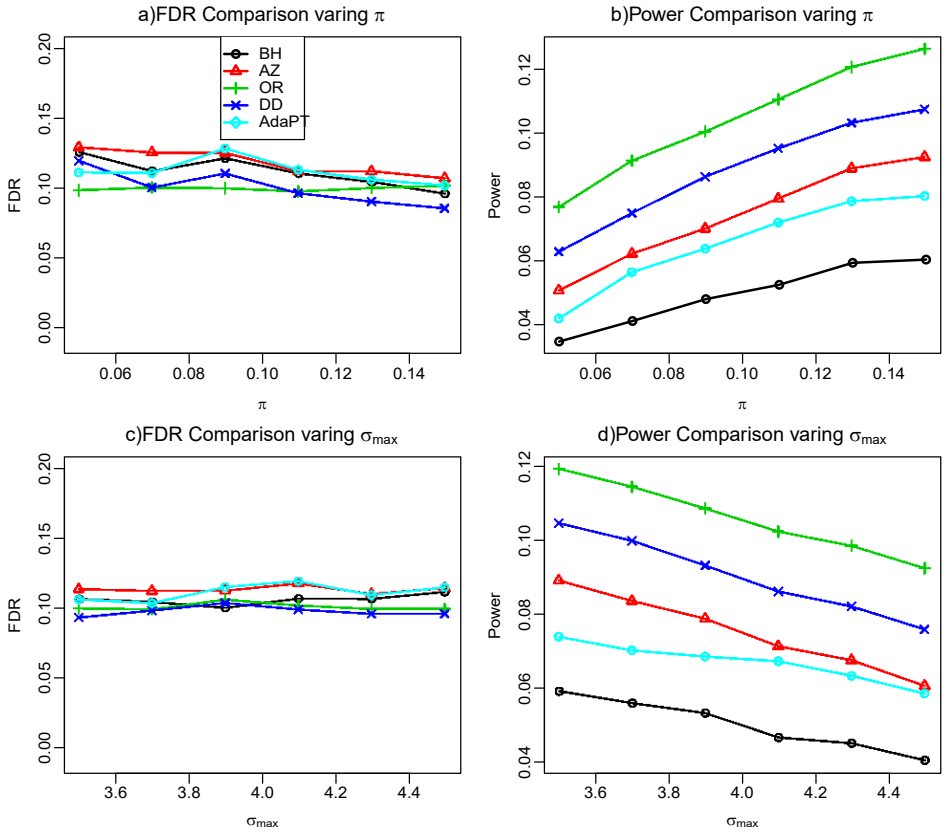

Figure 9: Comparison when $_i$ is unknown. We vary ⇡ in the top row and $_{max}$ in the bottom row. All data-driven methods are adversely a↵ected. But the e↵ect on DD is the smallest.

E.3 Weak dependence

We investigate the robustness of our method under weak dependence. We consider two weak dependence models:

Case 1: We use the following model:

$$\mu_i \overset{iid}{\leftarrow\cdots} (1 \quad ⇡)\ _0(\cdot) + ⇡\ _2(\cdot), \qquad _i \overset{iid}{\leftarrow\cdots} U[0, \ _{max}], \quad \mathbf{x} \leftarrow\cdots N(\boldsymbol{\mu}, ⍰),$$

where the covariance matrix is a block matrix:

$$⍰ = \begin{matrix} 0 & & 1 \\ @ M_{11} & 0 & A \\ 0 & M_{22} & \end{matrix}.$$

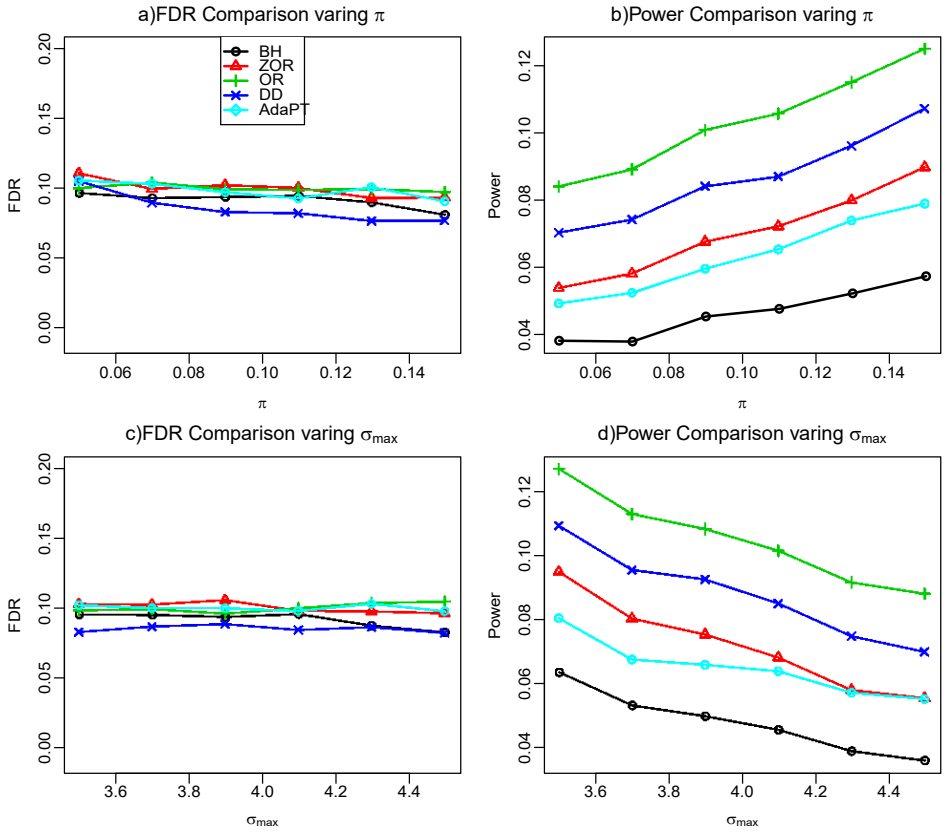

Figure 10: Comparison under weak dependence case 1. The dependence structure has little effect on all the methods. The pattern is almost the same as in the independent case.

Here M_{11} is a 4000×4000 matrix, the (i,i) entry is $\frac{2}{i}$, the $(i,i+1)$ and $(i+1,i)$ entries are $0.5_{\ i\ i+1}$, the $(i,i+2)$ and $(i+2,i)$ entries are $0.4_{\ i\ i+2}$. The rest of the entries are 0. M_{22} is a 16000×16000 diagonal matrix, with the (i,i) entry being $\frac{2}{i+4000}$. In the first setting, we fix $_{max} = 4$ and vary $\uparrow$ from 0.05 to 0.15. In the second setting, we fix $\uparrow = 0.1$ and vary $_{max}$ from 3.5 to 4.5. The results are summarized in Figure 10. The number of tests is $m = 20,000$. We can see that the weak dependence has little effect on the pattern. This demonstrates the robustness of DD under dependence.

Case 2: We use the following model:

$$\mu_i \overset{iid}{\leftarrow\cdots} (1 \quad \uparrow)\ _0(\cdot) + \uparrow\ _2(\cdot), \qquad _i \overset{iid}{\leftarrow\cdots} U[0,\ _{max}], \quad \mathbf{x} \leftarrow\cdots N(\boldsymbol{\mu}, \square),$$

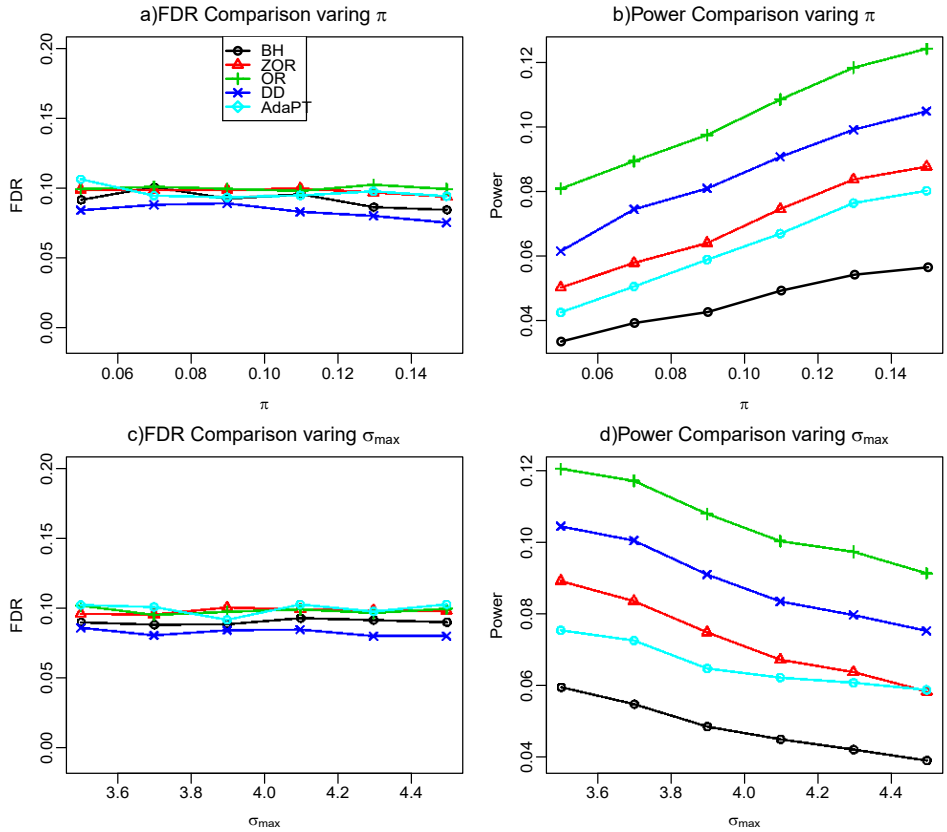

Figure 11: Comparison under weak dependence case 2. Again, the dependence structure has little e↵ect on all the methods. The pattern is almost the same as in the independent case.

where $\Sigma = \begin{pmatrix} M_{11} & 0 \\ 0 & M_{22} \end{pmatrix}$. Here M_{11} is a 4000×4000 matrix, the (i, j) entry of M_{11} is $0.5^{|i-j|}$ $_{i\ j}$. M_{22} is a 16000×16000 diagonal matrix with diagonal being $(\ _{4001}^2, \ldots, \ _{20000}^2)$. In the first setting, we fix $\ _{max} = 4$ and vary $\Uparrow$ from 0.05 to 0.15. In the second setting, we fix $\Uparrow = 0.1$ and vary $\ _{max}$ from 3.5 to 4.5. The results are summarized in Figure 11. We can see again, there is no noticeable di↵erence in pattern from the independent setting.

This shows the robustness of DD under positive dependence. This seems to be consistent with existing results in the literature.

E.4 Non-Gaussian noise

We study the performance of our method when the noise follows a heavy-tailed distribution. For this simulation, we independently generate 200 copies of X_i using the following model:

$$X_i = \mu_i + \cdot_i, \quad \mu_i \overset{iid}{\sim} (1 - \dagger) \, _0(\cdot) + \dagger \, _{2/}{}^{p}{}_{200}(\cdot), \qquad _i \overset{iid}{\sim} U[0, \ _{max}], \quad \cdot_i \overset{iid}{\sim} t_5.$$

We use the sample standard deviation of x_i (denoted $\hat{}_i$) as an estimate of $_i$ for all five methods. We then apply the testing procedures to the pairs $(\sqrt{200}\bar{x}_i, \ \hat{}_i)$. The number of tests is $m = 20,000$. Each simulation is run over 100 repetitions. We compare BH, the adaptive z-value procedure (AZ, Sun and Cai, 2007), DD, OR and AdaPT. Note that in this case, the model is mis-specified even for OR. But OR has access to the distribution of μ_i and $\dagger$ while other data-driven methods do not. In the first setting, we fix $_{max} = 4$ and vary $\dagger$ from 0.05 to 0.15. In the second setting, we fix $\dagger = 0.1$ and vary $_{max}$ from 3.5 to 4.5. The results are summarized in Figure 12. We can see that the pattern is similar to the Gaussian-noise case.

E.5 Unknown null distribution

We study the performance of our method when the z-values do not follow a standard normal distribution under the null hypothesis. For this simulation, we use the following model:

$$Z_i \overset{iid}{\sim} N(0, 0.8^2), \qquad _i \overset{iid}{\sim} U[0, \ _{max}], \quad X_i = Z_i \ _i + \mu_i, \quad \mu_i \overset{iid}{\sim} (1 - \dagger) \, _0(\cdot) + \dagger \, _2(\cdot).$$

For the data-driven methods, we estimate the null distribution of z_i's using the method described in Jin and Cai (2007). Let $_0$ be the estimated variance of Z_i. For DD, $f_{0, \ j}$ is now the density function of $N(0, \ ^2 \ ^2_0)._j$ The p-values are obtained as the two-sided tail probabilities with respect to the estimated empirical null. In the first setting, we fix $_{max} = 4$ and vary $\dagger$ from 0.05 to 0.15. In the second setting, we fix $\dagger = 0.1$ and vary $_{max}$ from 3.5 to 4.5. The results are summarized in Figure 13.

We can see that the variance of the estimator of null variance has a noticeable effect on all data-driven methods. The FDR control are not as precise as before for all data-driven methods. In particular, the gap between DD and ZOR has become smaller. However, DD

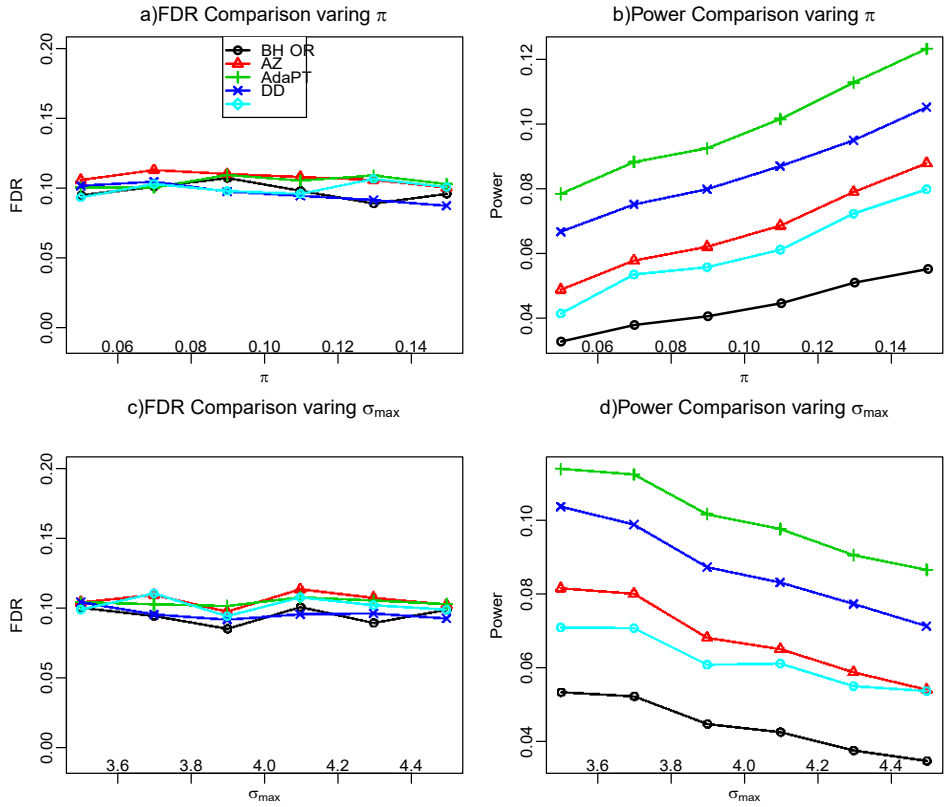

Figure 12: Comparison when the noise is non-Gaussian. The non-Gaussian noise has little effect on the overall pattern.

still performs the best among all data-driven methods.

E.6 Correlated $\mu_{a,i}$, σ_i

E.6.1 z-value is a lossless summary

We study the performance of our method under the following model:

$$\sigma_i \overset{iid}{\sim} U[0.5, \sigma_{max}], \quad \mu_{a,i} | \sigma_i \sim U[\sigma_i, 2\sigma_i], \quad X_i | \mu_i, \sigma_i \sim (1-\pi)N(0, \sigma_i^2) + \pi N(\mu_{a,i}, \sigma_i^2).$$

In this situation standardization, which leads to the following model

$$Z_i \overset{iid}{\sim} (1-\pi)N(0,1) + \pi N(u,1), \quad u \sim U[1,2],$$

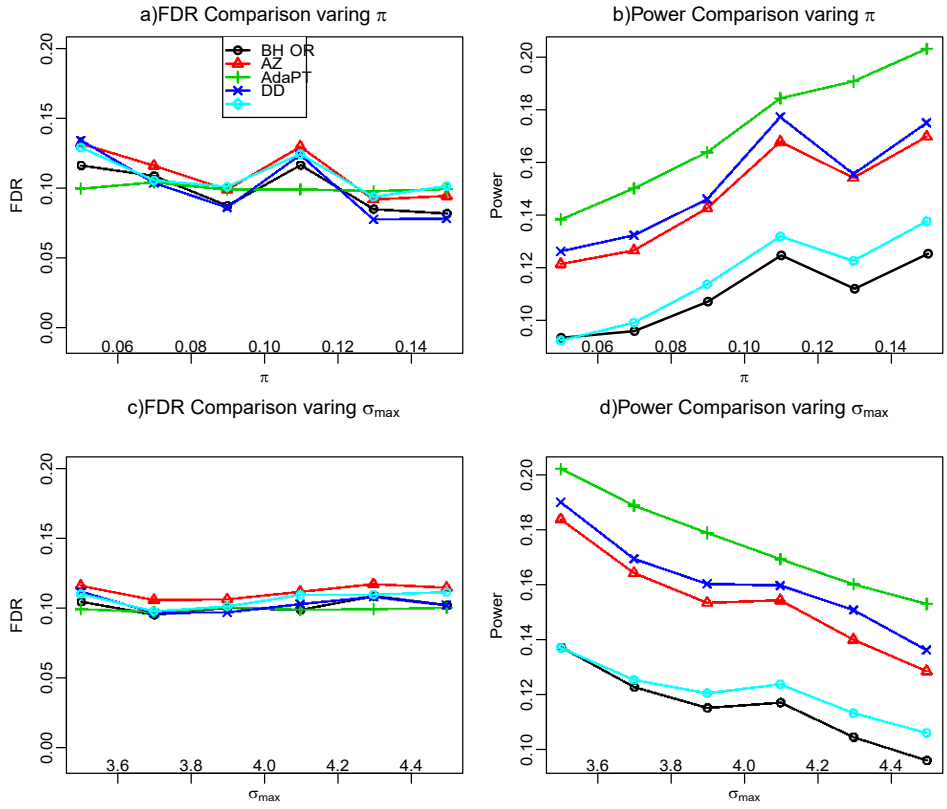

Figure 13: Comparison when the null distribution is estimated. We can see that all the data-driven methods show some instability: this is due to the variance of the estimator of null variance. But DD still out-performs other data-driven methods

would make a lot of sense.

It can be shown that in this case the oracle method based on Z_i coincides with the oracle method based on $(x, \ _i)$. So there is no information loss from standardization.

We compare di↵erent methods and the simulation result is summarized in Figure 14. The number of tests is $m = 20,000$. Each simulation is run over 100 repetitions. In the first setting, we fix $\ _{max} = 4$ and vary ⇡ from 0.05 to 0.15. In the second setting, we fix ⇡ = 0.1 and vary $\ _{max}$ from 3.5 to 4.5. One can see that performance of the data-driven HART is almost identical to the two oracle procedures.

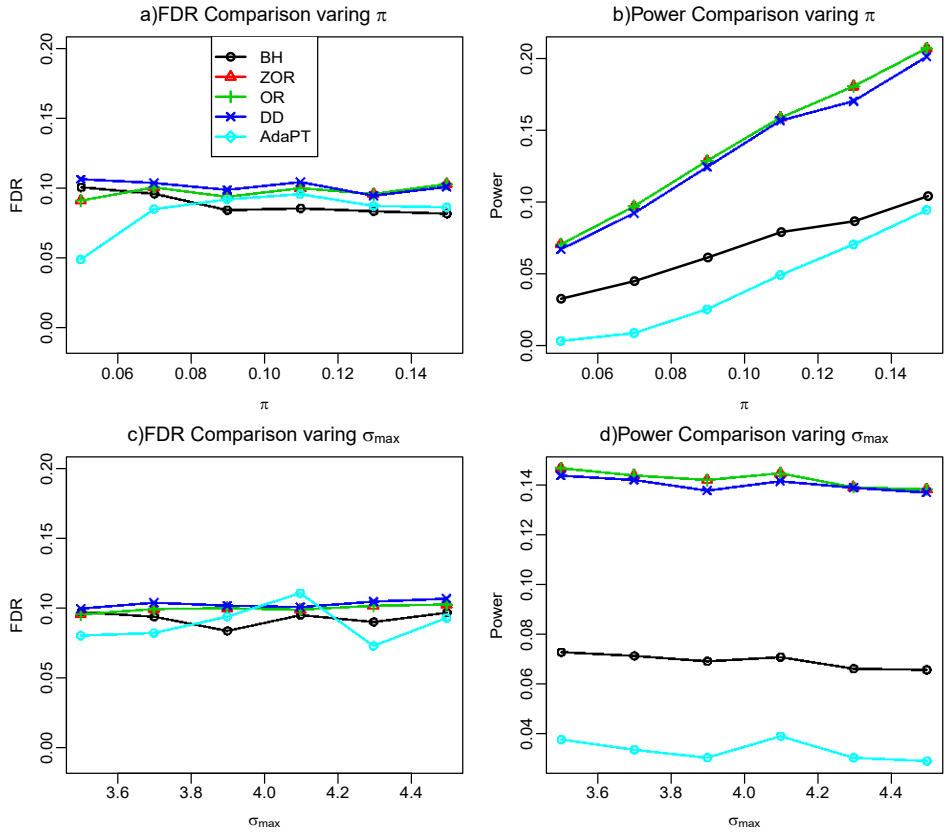

Figure 14: Comparison when $\mu_{a,i}$ and $_i$ are correlated but z-value is a lossless summary. We can see that in this case oracle method based on Z_i coincides with the oracle method based on $(x, _i)$. But the data-driven method still performs well.

E.6.2 z-value is not a lossless summary

We study the performance of our method under the following model:

$$_i \overset{iid}{\leftarrow\cdots} U[0.5, _{max}], \quad \mu_{a,i} | _i \leftarrow\cdots U[0.5\ ^2_i, \ ^2_i], \quad X_i | \mu_i, _i \leftarrow\cdots (1 \quad \Updownarrow) N(0, \ ^2_i) + \Updownarrow N(\mu_{a,i}, \ ^2_i).$$

Under this setting, the oracle method based on z-value only is not equivalent to the full oracle. In this case both the oracle and the data-driven HART procedure would outperform the oracle z-value procedure. The simulation result is summarized in Figure 15. The number of tests is $m = 20,000$. Each simulation is run over 100 repetitions. In the first setting, we fix $_{max} = 4$ and vary $\Updownarrow$ from 0.05 to 0.15. In the second setting, we fix $\Updownarrow = 0.1$ and vary $_{max}$ from 3.5 to 4.5.

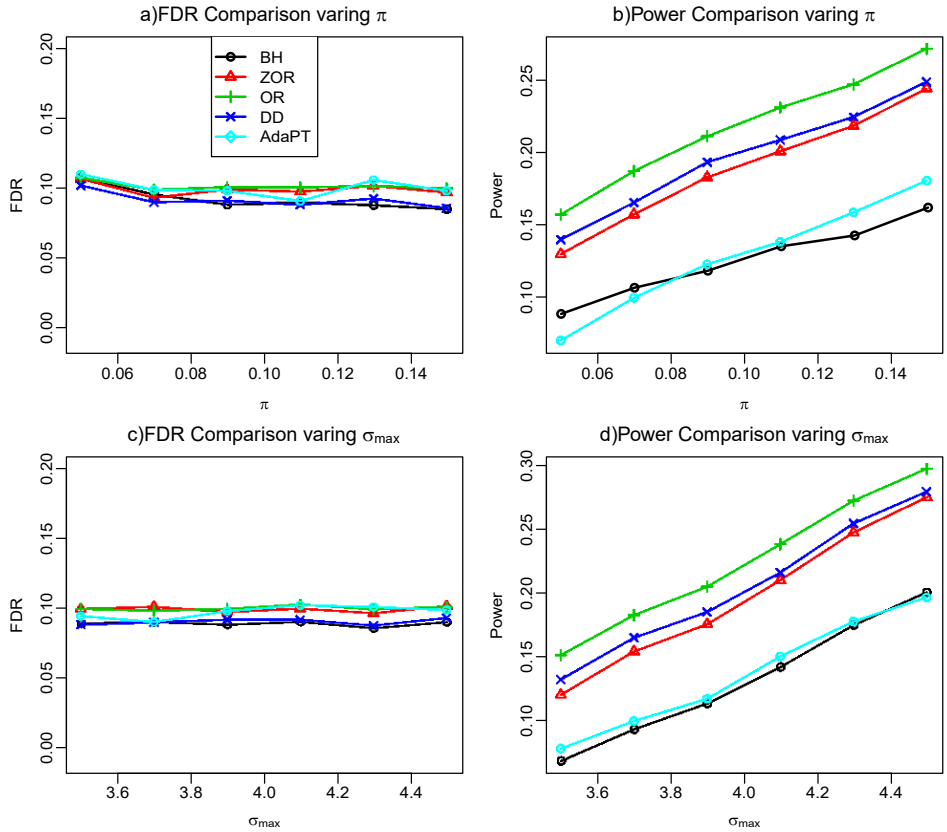

Figure 15: Comparison when $\mu_{a,i}$ and σ_i^2 are correlated and z-value is not a lossless summary. We can see that in this case the data-driven HART procedure outperforms methods based on z-value.

We can see from the simulation that OR has much higher power than ZOR. Moreover, the data-driven HART procedure (DD) has higher power than ZOR as well.

E.7 Global null

We study the stability of our data-driven procedure under the global null (no signals). We use the following model:

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

We vary σ_{max} from 3.5 to 4.5. The number of tests is $m = 20,000$. Each simulation is run over 100 repetitions. The simulation result is summarized in Figure 16. Recall that FDR is

defined as $E\left[\frac{\sum_i (1-\sqrt{\ }_i)\ _i}{\max\{\ _i\ _i, 1\}}\right]$. In the case of global null, the FDP $\frac{\sum_i (1-\sqrt{\ }_i)\ _i}{\max\{\ _i\ _i, 1\}}$ is either 0 or 1. We report the FDR as the average of FDPs. For each $_{max}$, we also record the number of repetitions with n rejections for BH and DD in Table 2. Note that under the global null the oracle knows $\uparrow = 0$. The oracle statistics $T_i = 1$ for all i. Hence the number of rejections is zero in all replications.

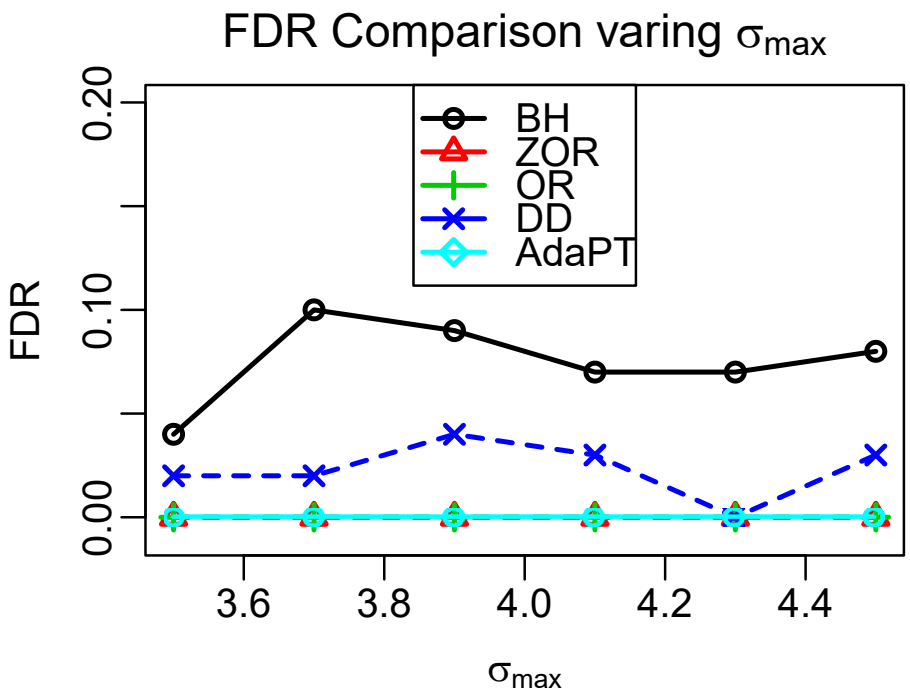

Figure 16: The data driven procedure is stable under global null.

$_{max}$	3.5		3.7		3.9		4.1		4.3		4.5	
n	BH	DD	BH	DD	BH	DD	BH	DD	BH	DD	BH	DD
0	96	98	90	98	92	96	93	97	93	100	92	97
1	4	2	7	1	6	4	5	3	4	0	6	2
2	0	0	1	0	2	0	2	0	3	0	1	1
3	0	0	2	1	0	0	0	0	0	0	1	0

Table 2: Distribution of the numbers of rejections by BH and DD. n is the number of rejections.

We can see that our data-driven HART procedure seems to be comparable to BH in term of the stability in FDR control. The reason for the robustness of data-driven HART under the global null is that when estimating the marginal distribution $f(x_i, \ _i)$ we have

already employed the known (null) density. This step has stabilized the bivariate density estimate in regions with few observations.