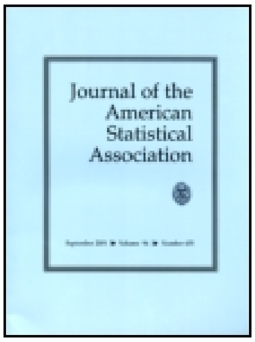


Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

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

Luella Fu , Bowen Gang , Gareth M. James & Wenguang Sun

To cite this article: Luella Fu , Bowen Gang , Gareth M. James & Wenguang Sun (2020): Heteroscedasticity-Adjusted Ranking and Thresholding for Large-Scale Multiple Testing, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1840992

To link to this article: https://doi.org/10.1080/01621459.2020.1840992

Taylor & Francis
Taylor & Francis Group

Check for updates

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

Luella Fu[a], Bowen Gang[b], Gareth M. James[c], and Wenguang Sun[c]

[a]Department of Mathematics, San Francisco State University, San Francisco, CA; [b]Department of Statistics, Fudan University, Shanghai, China; [c]Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA

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.

ARTICLE HISTORY
Received October 2019
Accepted October 2020

KEYWORDS
Covariate-assisted inference; Data processing and information loss; False discovery rate; Heteroscedasticity; Multiple testing with side information; Structured multiple testing

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, Tibshirani, and Chu 2001; Dudoit, Shaffer, and Boldrick 2003; Sun and Wei 2011); astronomy, for the detection of galaxies (Miller et al. 2001); neuro-imaging, for the discovery of differential brain activity (Pacifico et al. 2004; Schwartzman, Dougherty, and Taylor 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 different study units to a vector of summary statistics X_i, with associated standard deviation σ_i; standardize the summary statistics to obtain z-values, $Z_i = X_i/\sigma_i$; 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 offers a significant improvement over classical approaches, it only controls the FDR at level $(1 - \pi)\alpha$, where π is the proportion of non-nulls, suggesting that its power can be improved by incorporating an adjustement for π into the procedure. Benjamini and Hochberg (2000), Storey (2002), and Genovese and Wasserman (2002) proposed to first estimate the nonnull proportion by $\hat{\pi}$ and then run BH at level $\alpha/(1 - \hat{\pi})$. Efron et al. (2001) proposed the local false discovery rate (Lfdr), which incorporates, in addition to the sparsity parameter π, 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.

CONTACT Wenguang Sun ✉ wenguans@marshall.usc.edu Department of Data Sciences and Operations (DSO), University of Southern California (USC), Los Angeles, CA 90089.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

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, σ_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 σ 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 σ 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 units. The main strategy of HART is to bypass standardization by directly incorporating (X_i, σ_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 σ_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 cutoff. We develop theories to show that HART is optimal for integrating the information from both X_i and σ_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 σ_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 σ_i. It is important to note that the ordering or grouping based on the magnitudes of σ_i will not always be informative. Our numerical results show that ordering by σ_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 σ_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 σ_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 inefficient has broad implications since, due to inherent variabilities or differing sample sizes between study units, standardized tests are commonly applied to large-scale heterogeneous data to make different study units comparable. Second, our finding enriches the recent line of research on multiple testing with side and structural information (e.g. Cai, Sun, and Wang 2019; Li and Barber 2019; Xia, Cai, and Sun 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 article 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 θ_i denote a Bernoulli(π) variable, where $\theta_i = 0/1$ indicates a null/alternative hypothesis, and $\pi = P(\theta_i = 1)$ 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

$$X_i | \mu_i, \sigma_i^2 \overset{\text{ind}}{\sim} N(\mu_i, \sigma_i^2), \tag{1}$$

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

$$\mu_i \overset{\text{iid}}{\sim} (1-\pi)\delta_0(\cdot) + \pi g_\mu(\cdot), \quad \sigma_i^2 \overset{\text{iid}}{\sim} g_\sigma(\cdot). \tag{2}$$

In Equation (2), $\delta_0(\cdot)$ 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

[1] Unless otherwise stated, the term "full data" specifically refers to the pair (X_i, σ_i) in this article. In practice, the process of deriving the pair (X_i, σ_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.

which is allowed to vary across i. In this work, we focus on a model where μ_i and σ_i are not linked by a specific function. The more challenging situation where σ_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, Kou, and Brown 2012; Weinstein et al. 2018), we assume that σ_i are known. This simplifies the discussion and enables us to focus on key ideas. For practical applications, we use a consistent estimator of σ_i. The goal is to simultaneously test m hypotheses:

$$H_{0,i} : \mu_i = 0 \quad \text{vs.} \quad H_{1,i} : \mu_i \neq 0; \quad i = 1, \ldots, m. \tag{3}$$

The multiple testing problem (3) is concerned with the simultaneous inference of $\boldsymbol{\theta} = \{\theta_i = \mathbb{I}(\mu_i \neq 0) : i = 1, \ldots, m\}$, where $\mathbb{I}(\cdot)$ is an indicator function. The decision rule is represented by a binary vector $\boldsymbol{\delta} = (\delta_i : 1 \le i \le m) \in \{0,1\}^m$, where $\delta_i = 1$ means that we reject $H_{0,i}$, and $\delta_i = 0$ means we do not reject $H_{0,i}$. The false discovery rate (FDR) (Benjamini and Hochberg 1995), defined as

$$\text{FDR} = E\left[\frac{\sum_i (1-\theta_i)\delta_i}{\max\{\sum_i \delta_i, 1\}}\right], \tag{4}$$

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

$$\text{mFDR} = \frac{E\left\{\sum_i (1-\theta_i)\delta_i\right\}}{E\left(\sum_i \delta_i\right)}. \tag{5}$$

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, Taylor, and Siegmund 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 $\text{ETP} = E\left(\sum_{i=1}^m \theta_i\delta_i\right)$ to evaluate the power of an FDR procedure. Other power measures include the missed discovery rate (MDR, Taylor, Tibshirani, and Efron 2005), average power (Benjamini and Hochberg 1995; Efron 2007) and false negative rate or false nondiscovery rate (FNR, Genovese and Wasserman 2002; Sarkar 2002). Cao, Sun, and Kosorok (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 $\mu_i^* = \mu_i/\sigma_i$. The tacit rationale in conventional practice is that the simultaneous inference problem

$$H_{0,i} : \mu_i^* = 0 \quad \text{vs.} \quad H_{1,i} : \mu_i^* \neq 0; \quad i = 1, \ldots, m, \tag{6}$$

is equivalent to the formulation (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:

$$(X_i, \sigma_i) \quad \longrightarrow \quad Z_i = \frac{X_i}{\sigma_i} \quad \longrightarrow \quad P_i = 2\Phi(-|Z_i|). \tag{7}$$

We start with the full data consisting of X_i and $\sigma_i^2 = Var(X_i|\mu_i)$. 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 $H_{0,i} : \theta_i = 0$ and the data coming from a normal mixture model, where

$$\mu_i \overset{\text{iid}}{\sim} (1-\pi)\delta_0 + \pi\delta_{\mu_a}, \quad \sigma_i \overset{\text{iid}}{\sim} U[0.5, 4]. \tag{8}$$

This is a special case of Equation (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 $\alpha = 0.1$. 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 $\mathbb{P}(H_0|Z_i)$, which is in turn the most powerful z-value based method (Sun and Cai 2007). Finally, in the full data approach we reject when $\mathbb{P}(H_0|X_i, \sigma_i)$ is below a certain threshold, which we show later is optimal given X_i and σ_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 Equation (8) these rules correspond to

$$\begin{aligned}
\boldsymbol{\delta}^p &= \{\mathbb{I}(P_i \le 0.0006) : 1 \le i \le m\} \\
&= \{\mathbb{I}(|Z_i| \ge 3.43) : 1 \le i \le m\}, \\
\boldsymbol{\delta}^z &= \{\mathbb{I}(\mathbb{P}(H_0|Z_i) \le 0.24) : 1 \le i \le m\} \\
&= \{\mathbb{I}(Z_i \ge 3.13) : 1 \le i \le m\}, \\
\boldsymbol{\delta}^{\text{full}} &= \{\mathbb{I}(\mathbb{P}(H_0|X_i, \sigma_i) \le 0.28) : 1 \le i \le m\},
\end{aligned}$$

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, $\text{AP}(\boldsymbol{\delta}) = \frac{1}{m\pi}\mathbb{E}\left(\sum_{i=1}^m \theta_i\delta_i\right)$, differ significantly

$$\text{AP}(\boldsymbol{\delta}^p) = 5.0\%, \quad \text{AP}(\boldsymbol{\delta}^z) = 7.2\%, \quad \text{AP}(\boldsymbol{\delta}^{\text{full}}) = 10.5\%. \tag{9}$$

To better understand these differences consider the left-hand plot in Figure 1, which illustrates the rejection regions for each

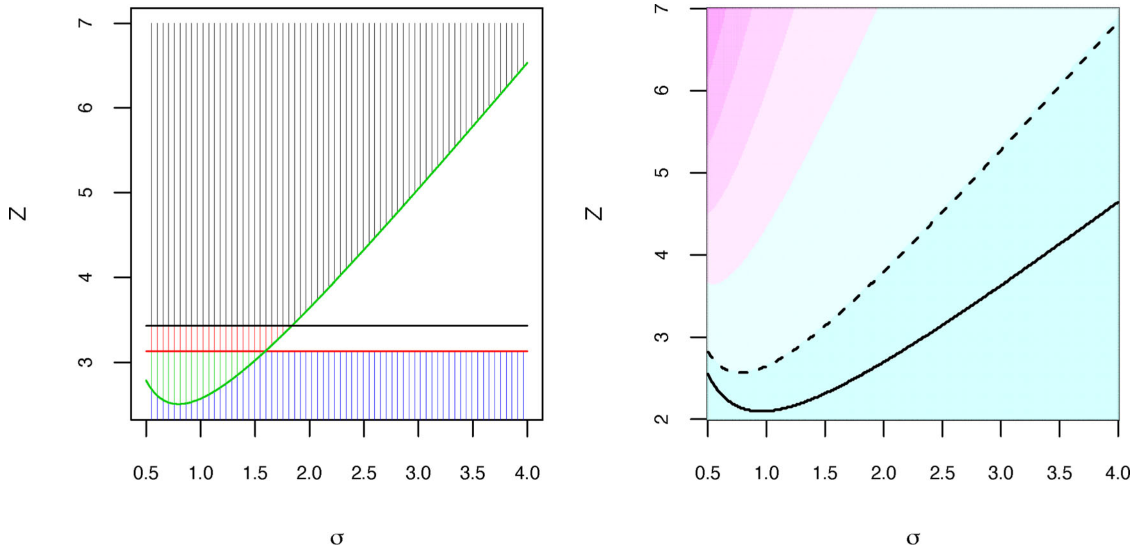


Figure 1. Left: Rejection regions for the p-value approach (black line), z-value approach (red line) and full data approach (green line) as a function of Z and σ. Approaches reject for all points above their corresponding line. Right: Heat map of relative proportions (on log scale) of alternative vs null hypotheses for different Z and σ. 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.

approach as a function of Z and σ[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 $\boldsymbol{\delta}^z$ and $\boldsymbol{\delta}^p$. Let π^+ and π^- denote the proportions of positive effects and negative effects, respectively. Then $\pi^+ = 0.1$ and $\pi^- = 0$. This asymmetry of the alternative distribution can be captured by $\boldsymbol{\delta}^z$, 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 $\boldsymbol{\delta}^z$ can be emulated by an adaptive z-value approach that is fully data-driven (Sun and Cai 2007).) By contrast, $\boldsymbol{\delta}^p$ enforces a two-sided rejection region that is symmetrical about 0, trading off extra rejections in the region $Z_i \leq -3.43$ for fewer rejections in the region where $3.13 \leq Z_i \leq 3.43$. As all nonzero effects 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 $\boldsymbol{\delta}^{\text{full}}$ vs $\boldsymbol{\delta}^z$. The full data approach trades off extra rejections in the green space for fewer rejections in the white space. This may seem like a suboptimal trade-off 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 $\sigma = 0.5$ 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 $\sigma = 2$ 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 $\sigma = 0.5$ is far more likely to come from the alternative hypothesis than is a z-value of 4 with $\sigma = 2$.

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 different Z and σ. 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 σ 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 different values of σ.

Figure 2 provides one further way to understand the effect 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 different values of σ. 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 $Z = 2/\sigma$ so when σ is large the standardized null and alternative are very similar,

[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.

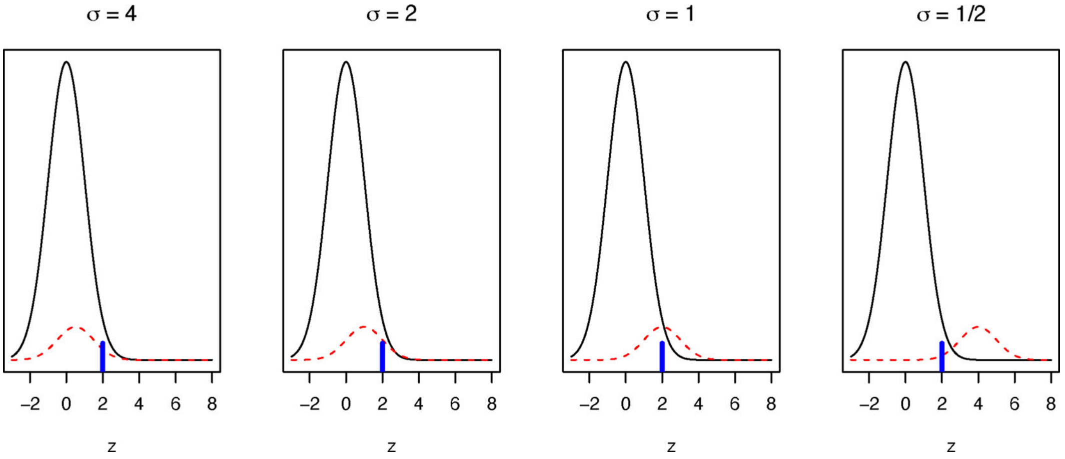


Figure 2. Plots of the density functions of Z under the null hypothesis (black solid) and alternative hypothesis (red dashed) for different values of σ. The blue line represents an observation at $Z = 2$.

making it hard to know which distribution $Z = 2$ belongs to. As σ 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 for all values of σ, 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 (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 (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 σ_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 $H_{0,i} : \mu_i \in [-a_0, a_0]$ 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 Heterocedasity*

Note that the models given by Equations (1) and (2) imply that

$$X_i|\sigma_i \overset{\text{ind}}{\sim} f_{\sigma_i}(x) = (1-\pi)f_{0,\sigma_i}(x) + \pi f_{1,\sigma_i}(x), \qquad (10)$$

where $f_{0,\sigma}(x) = \frac{1}{\sigma}\phi(x/\sigma)$ is the null density, $f_{1,\sigma}(x) = \frac{1}{\sigma}\int \phi_\sigma\left(\frac{x-\mu}{\sigma}\right) g_\mu(\mu) d\mu$ is the alternative density, $\phi(x)$ is the density of a standard normal variable, and $f_\sigma(x)$ is the mixture density. In conventional practice, the data are standardized as $Z_i = X_i/\sigma_i$, and the following mixture model is used

$$Z_i \overset{\text{iid}}{\sim} f(z) = (1-\pi)f_0(z) + \pi f_1(z), \qquad (11)$$

[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 $f_0(z) = \phi(z)$, $f_1(z)$ 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 $P_i = 2\Phi(-|Z_i|)$, where $\Phi(\cdot)$ is the standard normal cdf. The mixture model based on p-values is

$$P_i \overset{iid}{\sim} g(p) = (1-\pi)\mathbb{I}_{[0,1]}(p) + \pi g_1(p), \quad \text{for } p \in [0,1], \quad (12)$$

where $\mathbb{I}(\cdot)$ 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 11 and 12 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 11 and 12 are both known. We first review the oracle z-value procedure (Sun and Cai 2007). Define the local FDR (Efron et al. 2001)

$$\mathrm{Lfdr}_i = \mathbb{P}(H_0|z_i) = \mathbb{P}(\theta_i = 0|z_i) = \frac{(1-\pi)f_0(z_i)}{f(z_i)}. \quad (13)$$

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

$$\boldsymbol{\delta}^z = [\mathbb{I}\{\mathrm{Lfdr}(z_i) < c^*\} : 1 \le i \le m], \quad (14)$$

where c^* is the largest Lfdr threshold such that mFDR $\le \alpha$. Similarly, Genovese and Wasserman (2002) showed that the optimal p-value based FDR procedure is given by

$$\boldsymbol{\delta}^p = [\mathbb{I}\{P_i < c^*\} : 1 \le i \le m], \quad (15)$$

where c^* is the largest p-value threshold such that mFDR $\le \alpha$.

Next we derive the oracle rule based on m pairs $\{(x_i, \sigma_i) : i = 1, \ldots, m\}$. This new problem can be recast and solved in the framework of multiple testing with a covariate sequence. Consider Model 10 and define the heterogeneity-adjusted significance index[4]

$$T_i \equiv T(x_i, \sigma_i) = \mathbb{P}(\theta_i = 0|x_i, \sigma_i) = \frac{(1-\pi)f_{0,\sigma_i}(x_i)}{f_{\sigma_i}(x_i)}. \quad (16)$$

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

$$\boldsymbol{\delta}^{\mathrm{full}} = [\mathbb{I}\{T_i < t^*\} : 1 \le i \le m], \quad (17)$$

where $t^* = \sup\{t : Q(t) \le \alpha\}$.

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

Theorem 1. Let $\mathcal{D}_\alpha$ be the collection of all testing rules based on $\{(x_i, \sigma_i) : i = 1, \ldots, m\}$ such that $\mathrm{mFDR}_{\boldsymbol{\delta}} \le \alpha$. Then $\mathrm{ETP}_{\boldsymbol{\delta}} \le \mathrm{ETP}_{\boldsymbol{\delta}^{\mathrm{full}}}$ for any $\boldsymbol{\delta} \in \mathcal{D}_\alpha$. In particular we have

$$\mathrm{ETP}_{\boldsymbol{\delta}^p} \le \mathrm{ETP}_{\boldsymbol{\delta}^z} \le \mathrm{ETP}_{\boldsymbol{\delta}^{\mathrm{full}}}.$$

Based on Theorem 1, our proposed methodology employs a *HART* rule that operates in two steps: first rank all hypotheses according to T_i and then reject all hypotheses with $T_i \le t^*$. We discuss in Section 3.2 our finite sample approach for implementing HART using estimates for T_i and t^*.

3.2. Data-driven Procedure and Computational Algorithms

We first discuss how to estimate T_i and then turn to t^*. Inspecting T_i's formula (16), the null density $f_{0,\sigma_i}(x_i)$ is known and the non-null proportion π can be estimated by $\hat\pi$ 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 $f_{\sigma_i}(x_i)$.

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

$$\hat f_{\sigma_i}(x_i) = (1-\hat\pi)f_{0,\sigma_i}(x_i) + \hat\pi \hat f_{1,\sigma_i}(x_i). \quad (18)$$

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 $f_{\sigma_i}(x_i)$ 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 $f_{\sigma_i}(x_i)$ appears in the denominator exacerbates the situation. The decomposition in Equation (18) 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 $f_{1,\sigma_i}(x_i)$ because, unlike a typical variable, σ_i plays a special role in a density function and needs to be modeled carefully. Fu, James, and Sun (2020) recently addressed a closely related problem using the following weighted bivariate kernel estimator:

$$\hat f^*_\sigma(x) := \sum_{j=1}^{m} \frac{\phi_{h_\sigma}(\sigma - \sigma_j)}{\sum_{j=1}^{m} \phi_{h_\sigma}(\sigma - \sigma_j)} \phi_{h_{xj}}(x - x_j), \quad (19)$$

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

Let $\mathcal{M}_1 = \{i : \theta_i = 1\}$. In the ideal setting where θ_j is observed one could extend Equation (19) to estimate $f_{1,\sigma_i}(x_i)$ via

$$\tilde f_{1,\sigma}(x) = \sum_{j \in \mathcal{M}_1} \frac{\phi_{h_\sigma}(\sigma - \sigma_j)}{\sum_{k \in \mathcal{M}_1} \phi_{h_\sigma}(\sigma - \sigma_k)} \phi_{h_{xj}}(x - x_j). \quad (20)$$

Given that θ_j is unknown, we cannot directly implement Equation (20). Instead we apply a weighted version of Equation (20),

$$\hat f_{1,\sigma_i}(x_i) = \sum_{j=1}^{m} \frac{\hat w_j \phi_{h_\sigma}(\sigma_i - \sigma_j)}{\sum_{k=1}^{m} \hat w_k \phi_{h_\sigma}(\sigma_i - \sigma_k)} \phi_{h_{xj}}(x_i - x_j) \quad (21)$$

with weights $\hat w_j$ equal to an estimate of $P(\theta_j = 1|x_j, \sigma_j)$. In particular we adopt a two-step approach:

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

1. Compute $\hat{f}^{(0)}_{1,\sigma_i}(x_i)$ via Equation (21) with initial weights $\hat{w}_j^{(0)} = (1 - \hat{T}_j^{(0)})$ for all j, where $\hat{T}_j^{(0)} = \min\left\{\frac{(1-\hat{\pi})f_{0,\sigma_j}(x_j)}{\hat{f}^*_{\sigma_j}(x_j)}, 1\right\}$, $\hat{\pi}$ is the estimated non-null proportion, and $\hat{f}^*_{\sigma_j}(x_j)$ is computed using Equation (19).
2. Compute $\hat{f}^{(1)}_{1,\sigma_i}(x_i)$ via Equation (21) with updated weights $\hat{w}_j^{(1)} = (1 - \hat{T}_j^{(1)})$ where

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

This leads to our final estimate for $T_i = \mathbb{P}(H_0|x_i, \sigma_i)$:

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

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

To implement the oracle rule (17), 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 $\hat{T}_{(1)} \le \cdots \le \hat{T}_{(m)}$ and $H_{(1)}, \ldots, H_{(m)}$ the corresponding hypotheses. Let

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

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 (17), Procedure 1 can be viewed as its plug-in version:

$$\boldsymbol{\delta}^{dd} = \{\mathbb{I}(\hat{T}_i \le \hat{t}^*) : 1 \le i \le m\}, \text{ where } \hat{t}^* = \hat{T}_{(k)}. \tag{22}$$

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 $\boldsymbol{\delta}^{\text{full}}$ (17) is valid and optimal for FDR analysis. This section discusses the key theoretical result, Theorem 2, which shows that the performance of $\boldsymbol{\delta}^{\text{full}}$ can be achieved by its finite sample version $\boldsymbol{\delta}^{\text{dd}}$ (22) when $m \to \infty$. Inspecting (22), the main steps involve showing that both $\hat{T}_i$ and $\hat{t}^*$ are "close" to their oracle counterparts. To ensure good performance of the proposed procedure, we require the following conditions.

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

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

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

(C4) $\hat{\pi} \xrightarrow{P} \pi$.

Remark 2. For Condition (C2), the requirement on f_σ 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 $\hat{f}_\sigma$ and the plug-in statistic $\hat{T}_i$. The convergence of $\hat{t}^*$ to t^* and the asymptotic properties of $\boldsymbol{\delta}^{\text{dd}}$ are established in Theorem 2.

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

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

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

Next we turn to the performance of our data-driven procedure $\boldsymbol{\delta}^{\text{dd}}$ when $m \to \infty$. A key step in the theoretical development is to show that $\hat{t}^* \xrightarrow{P} t^*$, where $\hat{t}^*$ and t^* are defined in Equations (22) and (17), respectively.

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

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 σ_i comes from a continuous density function. In Section 4.3, we further investigate the effect of heterogeneity under a mixture model where σ_i takes on one of the two distinct values. Simulation results for additional settings, including a non-Guassian alternative, unknown σ_i, weak dependence structure, non-Gaussian noise, estimated empirical null, correlated μ_i and σ_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 $\hat{f}_{1,\sigma}(x)$, which depends on the choice of tuning parameters $\boldsymbol{h} = (h_x, h_\sigma)$. 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, σ_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 nonnulls in the data, denoted $\hat{\pi}$. We then compute h_x and h_σ by applying Silverman's rule of thumb (Silverman 1986) to the subset of the observations $\{x_i : P_i < \hat{\pi}\}$. When implementing HART, we first estimate $f_\sigma(x)$ using the data without (X_i, σ_i), and then plug-in the unused data (X_i, σ_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 1 and 2, where σ_i are uniformly generated from $U[0, \sigma_{\max}]$. We then generate X_i from a two-component normal mixture model

$$X_i|\sigma_i \overset{\text{iid}}{\sim} (1-\pi)N(0, \sigma_i^2) + \pi N(2, \sigma_i^2).$$

In the first setting, we fix $\sigma_{\max} = 4$ and vary π from 0.05 to 0.15. In the second setting, we fix $\pi = 0.1$ and vary $\sigma_{\max}$ 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, Sun, and Wang 2019) since the primary objective here is to incorporate structural information encoded in σ_i. Also, although Ignatiadis et al. (2016) mention the idea of using σ_i as a covariate to construct weighted p-values, no guidance is given on how to do so, and since the way in which σ_i are incorporated is particularly important, we exclude it.

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. Then, the FDR is estimated as the average of the false discovery proportion $\text{FDP}(\boldsymbol{\delta}) = \sum_{i=1}^m \{(1-\theta_i)\delta_i\}/(\sum_{i=1}^m \delta_i \vee 1)$ and the average power is estimated as the average proportion of true positives that are correctly identified, $\sum_{i=1}^m (\theta_i \delta_i)/(mp)$, both over the number of repetitions. The results for differing values of π and $\sigma_{\max}$ 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 π 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 efficiency gain increases as π increases. To explain the power gain of ZOR over BH, let π^+ and π^- denote the proportion of true positive signals and true negative signals, respectively. Then $\pi^+ = \pi$ and $\pi^- = 0$. 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 σ_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 σ_i can be anti-informative and lead to possible power loss for AdaPT. By contrast, HART uses the side information in a principled and systematic way. It uniformly improves competitive methods.

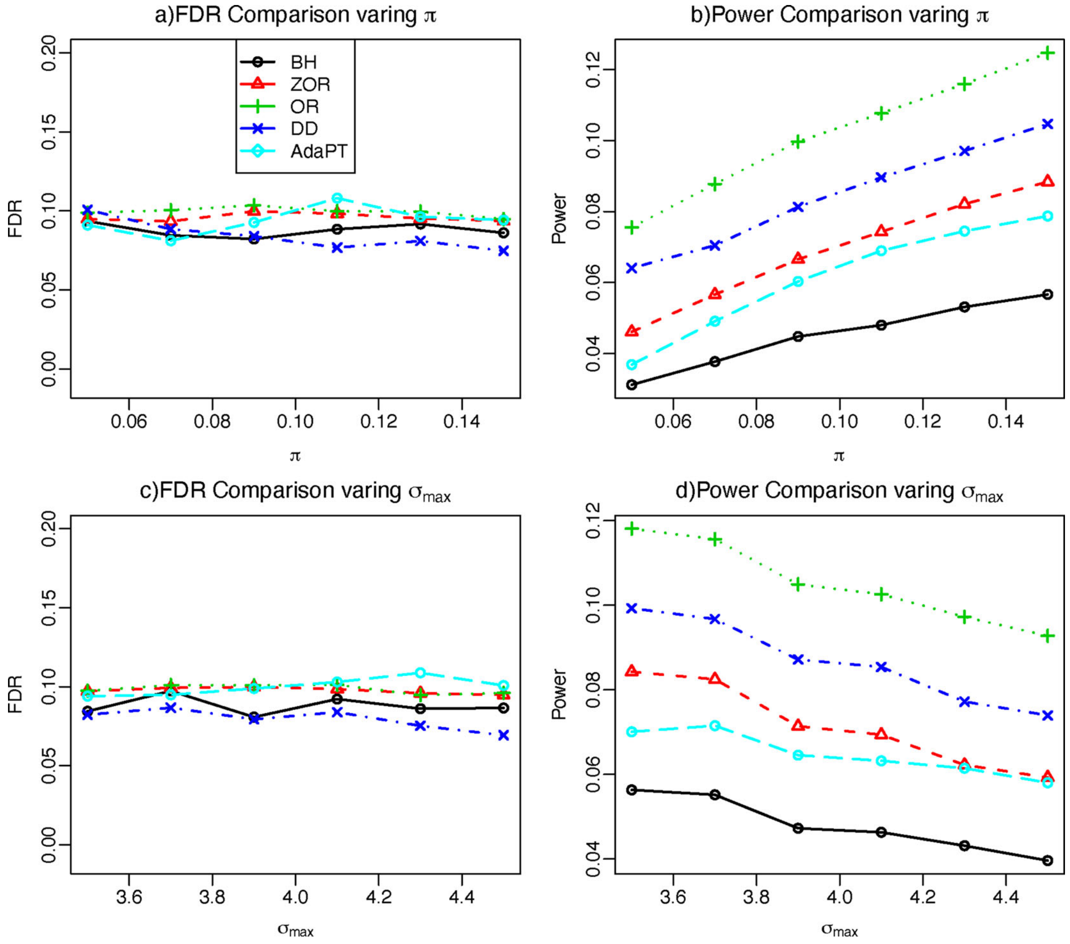
Finally, it should be noted that incorporating side information comes with computational costs: conventional methods including BH and ZOR both run considerably faster than DD. However, DD runs faster than AdaPT.

4.3. Comparison Under a Two-group Model

To illustrate the heteroscedasticity effect more clearly, we conduct a simulation using a simpler model where σ_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 1 and 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 $X_i|\sigma_i \overset{\text{iid}}{\sim} (1-\pi)N(0, \sigma_i^2) + \pi N(\mu, \sigma_i^2)$. 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 datasets. 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.

[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.


a)FDR Comparison varing π

b)Power Comparison varing π

c)FDR Comparison varing σ_{max}

d)Power Comparison varing σ_{max}

Figure 3. Comparison when σ_i is generated from a uniform distribution. We vary π in the top row and $\sigma_{\max}$ 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.

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 dataset. The dataset 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 $\alpha = 0.1$.

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: $P_i^* = 2\Phi^*(-2|Z_i|)$, where Φ^* is the 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 $f_\sigma(x)$, which usually requires a relatively large sample size to ensure good precision. Figure 6 presents the histogram of S_i and scatterplot of S_i vs Z_i. Based on the histogram, we propose to only focus on data points

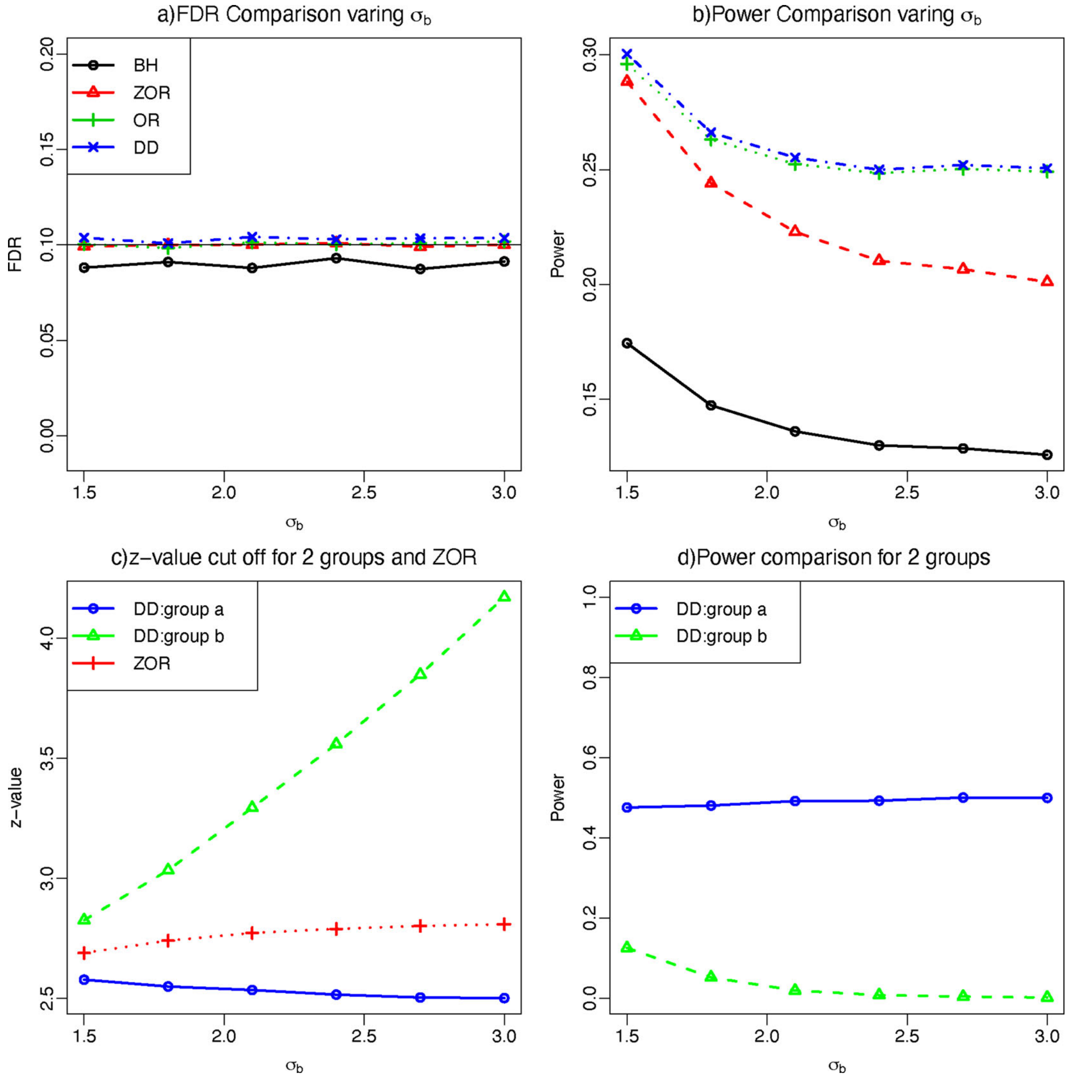


Figure 4. Two groups with varying σ_b from 1.5 to 3. As σ_b increases, the cut-off for group a decreases whereas the cut-off for group b increases. The power for tests in group b drops quickly as σ_b increases. This corroborates our calculations in the toy example in Section 2.2 and the patterns revealed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

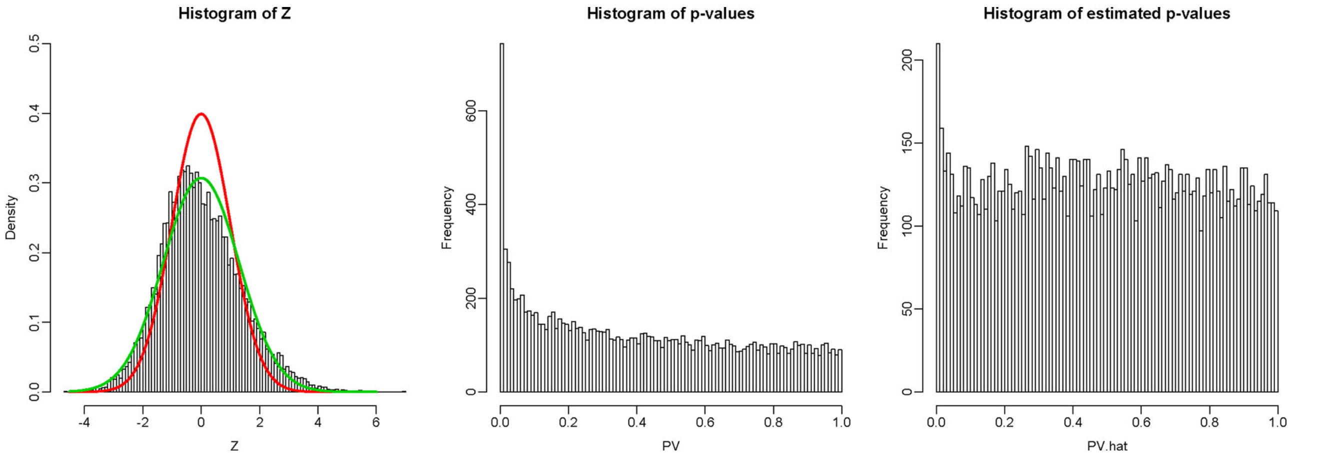


Figure 5. Left: histogram of z-values: the estimated empirical null $N(0, 1.3^2)$ (green line) seems to provide a better fit to the data compared to the theoretical null $N(0, 1)$ (red line). Middle: histogram of original p-values. Right: histogram of estimated p-values based on the empirical null. The 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.

with S_i less than 1 (12,172 out of 12,625 genes are kept in the analysis) to ensure the estimation accuracy of $\hat{T}_i$. Compared to conventional approaches, there is no efficiency loss because no hypothesis with $S_i > 1$ is rejected by BH at $\alpha = 0.1$ – note that the BH p-value cutoff is 6×10^{-5}, which corresponds to a z-value cutoff of 5.22; see also Figure 7. (If BH rejects hypotheses with

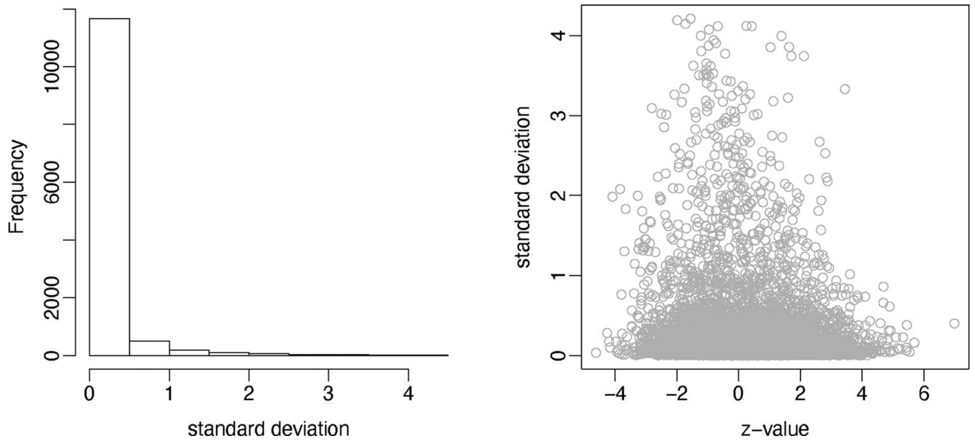


Figure 6. Histogram of S_i (left), scatterplot of (Z_i, S_i) (right)

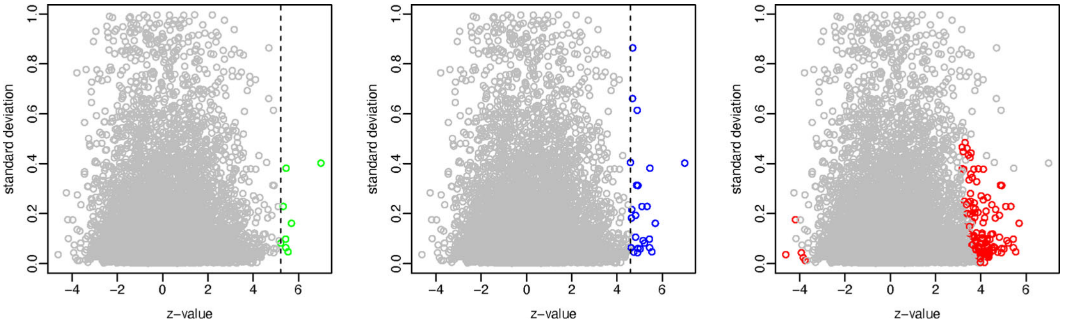


Figure 7. Scatterplot of rejected hypotheses by each method. Green: BH, blue: AZ, red: HART. AZ and BH reject every hypothesis to the right of the dashed line. The rejection region for HART depends on both z and σ.

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

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

large S_i, we recommend to carry out a group-wise FDR analysis, which first tests hypotheses at α 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^* 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, we estimate the nonnull proportion π using Jin-Cai's method with the empirical null taken as $N(0, 1.3^2)$. We further employ the jackknifed method to estimate $f_\sigma(x)$ 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 different 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 $N(0, \sigma_i^2)$ conditioned on individual σ_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 difficult to assess the power gains. However, the key point of this analysis, and the focus of our article, is to compare the *shapes of rejection regions* to gain some insights on the differences 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).

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, Cai, and Sun (2019), Li and Barber (2019) and Cai, Sun, and Wang (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 σ_i can be extremely useful for improving the ranking and hence the power in multiple testing.[6]

[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 α.

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 inefficient, 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 effective strategy for capturing the heterogeneity in the data. Cai and Sun (2009) showed that the power of FDR procedures can be much improved by using new ranking statistics adjusted for grouping. Recent works along this direction, including Liu, Sarkar, and Zhao (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 σ_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, Roeder, and Wasserman 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, Zhao, and Zhou 2010; Li and Barber 2019; Xia, Cai, and Sun 2019). However, the derivation of weighting functions for directly incorporating heteroscedasticity seems to be rather complicated (Peña, Habiger, and Wu 2011; Habiger, Watts, and Anderson 2017). Notably, Habiger (2017) developed novel weights for p-values as functions of a class of auxiliary parameters, including σ_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) can be extended to

$$\mu_i|\sigma_i \overset{\text{ind}}{\sim} (1-\pi_{\sigma_i})\delta_0(\cdot) + \pi_{\sigma_i} g_\mu(\cdot|\sigma_i), \quad \sigma_i^2 \overset{\text{iid}}{\sim} g_\sigma(\cdot), \tag{23}$$

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

$$T_i = \mathbb{P}(\theta_i = 0|x_i, \sigma_i) = \frac{(1-\pi_{\sigma_i})f_{0,\sigma_i}(x_i)}{f_{\sigma_i}(x_i)}, \tag{24}$$

where the varying proportion π_{σ_i} indicates that σ_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 $\{\sigma_i : 1 \le i \le m\}$. 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 σ_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 σ_i on the accuracy and stability of large-scale inference, and to develop more accurate simultaneous estimation procedures for unknown σ_i. We have provided some empirical results in Appendix E.2 but a rigorous theoretical study, along the lines of Fan, Hall, and Yao (2007) and Kosorok and Ma (2007), will be of much interest for future research.

Supplementary Materials

The supplementary material contains the formulas for the illustrative example; proofs of main theorems, propositions, and lemma;s and additional numerical results.

Funding

The research of Wenguang Sun was supported in part by NSF grant DMS-1712983.

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, 79, 1247–1268. [12]

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. [3,12]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of Royal Statistical Society*, Series B, 57, 289–300. [1,3]

——— (1997), "Multiple Hypotheses Testing With Weights," *Scandinavian Journal of Statistics*, 24, 407–418. [12]

——— (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. [1]

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. [8]

Cai, T. T., and Sun, W. (2009), "Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks," *Journal of American Statistical Association*, 104, 1467–1481. [2,8,12]

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, 81, 187–234. [2,8,11]

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, 495–502. [3]

Castillo, I., and Roquain, E. (2018), "On Spike and Slab Empirical Bayes Multiple Testing," arXiv:1808.09748. [12]

Donoho, D., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *Annals of Statistics*, 32, 962–994. [6]

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103. [1]

Efron, B. (2004a), "Large-scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99(465):96–104. [5]

——— (2004b), "Large-scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104. [9]

——— (2007), "Size, Power and False Discovery Rates," *Annals of Statistics*, 35, 1351–1377. [3]

——— (2008a), "Microarrays, Empirical Bayes and the Two-groups Model," *Statistical Science*, **23**:1–22. [1,2,11]

——— (2008b), "Simultaneous Inference: When Should Hypothesis Testing Problems Be Combined?" *Annals of Applied Statics*, **2**:197–223. [2,12]

——— (2011), "Tweedie's Formula and Selection Bias," *Journal of the American Statistical Association*, 106, 1602–1614. [5]

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of American Statistical Association*, 96:1151–1160. [1,2,6]

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, 1282–1288. [12]

Fu, L. J., James, G. M., and Sun, W. (2020), "Nonparametric Empirical Bayes Estimation on Heterogeneous Data," arXiv:2002.12586. [6]

Genovese, C. and Wasserman, L. (2002), "Operating Characteristics and Extensions of the False Discovery Rate Procedure," *Journal of Royal Statistical Society*, Series B, 64, 499–517. [1,3,6]

Genovese, C. R., Roeder, K., and Wasserman, L. (2006), "False Discovery Control With p-value Weighting," *Biometrika*, 93, 509–524. [12]

Habiger, J., Watts, D., and Anderson, M. (2017), "Multiple Testing With Heterogeneous Multinomial Distributions," *Biometrics*, 73, 562–570. [12]

Habiger, J. D. (2017), "Adaptive False Discovery Rate Control for Heterogeneous Data," *Statistica Sinica*, 27, 1731–1756. [12]

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

Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70. [1]

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. [12]

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, 556–583. [2,8]

Jin, J., and Cai, T. T. (2007), "Estimating the Null and the Proportional of Nonnull Effects in Large-scale Multiple Comparisons," *Journal of American Statistical Association*, 102, 495–506. [6,7,8]

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, 1456–1486. [12]

Lei, L., and Fithian, W. (2018), "Adapt: An Interactive Procedure for Multiple Testing With Side Information," *Journal of the Royal Statistical Society*, Series B, 80, 649–679. [2,8]

Li, A., and Barber, R. F. (2019), "Multiple Testing With the Structure-Adaptive Benjamini–Hochberg Algorithm," *Journal of the Royal Statistical Society*, Series B, 81, 45–74. [2,8,11,12]

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. [12]

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. [1]

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting Differential Gene Expression With a Semiparametric Hierarchical Mixture Method," *Biostatistics*, 5, 155–176. [6]

Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. (2004), "False Discovery Control for Random Fields," *Journal of the American Statistical Association*, 99, 1002–1014. [1]

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, 556–583. [12]

Roquain, E. and Van De Wiel, M. A. (2009), "Optimal Weighting for False Discovery Rate Control," *Electronic Journal of Statistics*, 3, 678–711. [12]

Sarkar, S. K. (2002), "Some Results on False Discovery Rate in Stepwise Multiple Testing Procedures," *Annals of Statistics*, 30, 239–257. [3]

Sarkar, S. K., and Zhao, Z. (2017), "Local False Discovery Rate Based Methods for Multiple Testing of One-way Classified Hypotheses," arXiv:1712.05014. [12]

Schwartzman, A., Dougherty, R. F., and Taylor, J. E. (2008), "False Discovery Rate Analysis of Brain Diffusion Direction Maps," *Annals of Applied Statistics*, 2, 153–175. [1]

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, 459–471. [8]

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Vol. 26, Boca Raton, FL: CRC Press. [6,7,8]

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of Royal Statistical Society*, Series B, 64, 479–498. [1,6]

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," *Journal of Royal Statistical Society*, Series B, 66, 187–205. [3]

Sun, W., and Cai, T. T. (2007), "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control," *Journal of American Statistical Association*, 102, 901–912. [1,3,4,6,8,9]

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. [5]

Sun, W., and Wei, Z. (2011), "Large-scale Multiple Testing for Pattern Identification, With Applications to Time-course Microarray Experiments," *Journal of American Statistical Association*, 106, 73–88. [1]

Taylor, J., Tibshirani, R., and Efron, B. (2005), The 'Miss Rate' for the Analysis of Gene Expression Data," *Biostatistics*, 6, 111–117. [3]

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, 2483–2494. PMID: 14695408. [9]

Tusher, V. G., Tibshirani, R., and Chu, G. (2001), "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proceedings of the National Academy of Science, USA*, 98, 5116–5121. [1]

Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing*, volume 60 of *Chapman and Hall CRC Monographs on Statistics and Applied Probability*. New York: Chapman and Hall CRC. [6,7]

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. [3,12]

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*, 115, 1236–1250. [2,11,12]

Xie, X., Kou, S., and Brown, L. D. (2012), "Sure Estimates for a Heteroscedastic Hierarchical Model," *Journal of the American Statistical Association*, 107, 1465–1479. [3]

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), pp. 527–540. [1]