CARS: Covariate Assisted Ranking and Screening for Large-Scale Two-Sample Inference

T. Tony Cai
University of Pennsylvania, Philadelphia, USA
Wenguang Sun
University of Southern California, Los Angeles, USA
Weinan Wang
University of Southern California, Los Angeles, USA

Summary. Two-sample multiple testing has a wide range of applications. The conventional practice first reduces the original observations to a vector of p-values and then chooses a cutoff to adjust for multiplicity. However, this data reduction step could cause significant loss of information and thus lead to suboptimal testing procedures. In this paper, we introduce a new framework for two-sample multiple testing by incorporating a carefully constructed auxiliary variable in inference to improve the power. A data-driven multiple testing procedure is developed by employing a covariate-assisted ranking and screening (CARS) approach that optimally combines the information from both the primary and auxiliary variables.
The proposed CARS procedure is shown to be asymptotic valid and optimal for false discovery rate (FDR) control. The procedure is implemented in the R-package `CARS`. Numerical results confirm the effectiveness of CARS in FDR control and show that it achieves substantial power gain over existing methods. CARS is also illustrated through an application to the analysis of satellite imaging data set for supernova detection.

Keywords: Compound decision theory; False discovery rate; Logically correlated tests; Multiple testing with covariates; Uncorrelated screening.

1. Introduction

A common goal in modern scientific studies is to identify features that exhibit differential levels across two or more conditions. The task becomes difficult in large-scale comparative experiments, where differential features are sparse among thousands or even millions of features being investigated. The conventional practice is to first reduce the original samples to a vector of p-values and then choose a cutoff to adjust for multiplicity. However, the first step of data reduction could cause significant loss of information and thus lead to suboptimal testing procedures. This paper proposes new strategies to extract structural information in the sample using an auxiliary covariate sequence and develops optimal covariate-assisted inference procedures for large-scale two-sample multiple testing problems.

We focus on a setting where both mean vectors are individually sparse. Such a setting arises naturally in many modern scientific applications. For example, the detection of sequentially activated genes in time-course microarray experiments, considered in Section B.7 in the Supplementary Material, involves identifying varied effect sizes across different time points [Calvano et al. (2005); Sun and Wei (2011)]. Since only a small fraction of genes are differentially expressed from the baseline, the problem of identifying varied levels over time essentially reduces to a multiple testing problem with several high-dimensional sparse

vectors (after removing the baseline effects). The second example arises from the detection of supernova explosions considered in Section 5.5. The potential locations can be identified by testing sudden changes in brightness in satellite images taken over a period of time. After the measurements are converted into grey-scale images and vectorized, multiple tests are conducted to compare the intensity levels between two sparse vectors. Another case in point is the analysis of differential networks, where the goal is to detect discrepancies between two or more networks with possibly sparse edges.

We first describe the conventional framework for two-sample inference and then discuss its limitations. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random vectors recording the measurement levels of the same m features under two experimental conditions, respectively. The population mean vectors are given by $\boldsymbol{\mu}_x = \mathbb{E}(\boldsymbol{X}) = (\mu_{x1}, \ldots, \mu_{xm})^\intercal$ and $\boldsymbol{\mu}_y = \mathbb{E}(\boldsymbol{Y}) = (\mu_{y1}, \ldots, \mu_{ym})^\intercal$. A classical formulation for identifying differential features is to carry out m two-sample tests:

$$H_{i,0} : \mu_{xi} = \mu_{yi} \quad \text{vs.} \quad H_{i,1} : \mu_{xi} \neq \mu_{yi}, \quad 1 \leq i \leq m. \tag{1.1}$$

Suppose we have collected two random samples $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}\}$ and $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}\}$ as independent copies of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The standard practice starts with a data reduction step: a two-sample t-statistic T_i is computed to compare the two conditions for feature i, then T_i is converted to a p-value or z-value. Finally a significance threshold is chosen to control the multiplicity. However, this conventional practice, which only utilizes a vector of p-values, may suffer from substantial information loss.

This article proposes a new testing framework that involves two steps. In the first step, besides the usual primary test statistics, an auxiliary covariate sequence is constructed from the original data to capture important structural information that is discarded by conventional practice. In the second step, the auxiliary covariates are combined with the primary test statistics to construct a multiple testing procedure that improves the accuracy in inference. Our idea is that the hypotheses become "unequal" in light of the auxiliary sequence. A key step in our methodological development is to incorporate the heterogeneity by recasting the problem in the framework of multiple testing with a covariate sequence. This requires a carefully constructed pair of statistics that lead to a simple bivariate model and an easily implementable methodology. Section 2 discusses strategies for constructing the pair of primary and auxiliary variables. Then we develop oracle and data-driven multiple testing procedures for the consequent bivariate model in Section 3. The proposed method employs a covariate-assisted ranking and screening (CARS) approach that simultaneously incorporates the primary and auxiliary information in multiple testing. We show that the CARS procedure controls the false discovery rate at the nominal level and outperforms existing methods in power.

We mention two related strategies in the literature: testing following screening and testing following grouping. In the first strategy, the hypotheses are formed and tested hierarchically via a screen-and-clean method [Zehetmayer et al. (2005); Reiner-Benaim et al. (2007); Zehetmayer et al. (2008); Wasserman and Roeder (2009); Bourgon et al. (2010)]. Following that strategy, one can first inspect the sample to identify the union support of $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$, and then conduct two-sample tests on the narrowed subset to further eliminate the null locations with no differential levels. The screen-and-clean approach requires sample splitting to ensure the independence between the screening and testing stages to avoid selection bias [Rubin et al. (2006)]. However, even the screening stage can significantly narrow down the focus, sample splitting often leads to loss of power. For example, the empirical studies in Skol et al. (2006) concluded that a two-stage analysis is in general inferior compared to a naive joint analysis that combines the data from both stages. The

second strategy (Liu, 2014) can be described as *testing following grouping*, that is, the hypotheses are analyzed in groups via a divide-and-test method. Liu (2014) developed an uncorrelated screening (US) method, which first divides the hypotheses into two groups according to a screening statistic, and then applies multiple testing procedures to the groups separately to identify non-null cases. It was shown in Liu (2014) that US controls the error rate at the nominal level and outperforms competitive methods in power.

Our approach marks a clear departure from existing methods. Both the screen-and-clean and divide-and-test strategies involve dichotomizing a continuous variable, which fails to fully utilize the auxiliary information. By contrast, our proposed CARS procedure models the screening covariate as a continuous variable and employs a novel ranking and selection procedure that optimally integrates the information from both the primary and auxiliary variables. In Section 4, we develop further results on a general bivariate model; our study reveals the connections between existing methods and provides insights on the advantage of the proposed CARS procedure. Simulation results in Section 5 demonstrate that CARS controls the false discovery rate in finite samples and uniformly dominates all existing methods. The gain in power is substantial in many settings. We illustrate our method to analyze a time-course satellite image dataset in Section 5.5. The application shows improved sensitivity of the proposed method in identifying changes between images taken over time. Section 6 further discusses related issues and open problems. The proofs are provided in Section 7 and Appendix A. Additional numerical results are given in Appendix B.

2. Extracting Structural Information Using An Auxiliary Sequence

Suppose $\{X_{ij} : 1 \le j \le n_x\}$ and $\{Y_{ik} : 1 \le k \le n_y\}$, $i = 1, \cdots, m$, are repeated measurements of generic independent random variables X_i and Y_i, respectively. Let $\boldsymbol{\beta}_0 = (\beta_{0i} : 1 \le i \le m)$ be a latent baseline vector which itself is sparse [including the special case where $\beta_{0i} = 0$ for all i]. Consider the following hierarchical model

$$X_{ij} = \beta_{0i} + \mu^*_{xi} + \epsilon_{xij}, \quad Y_{ik} = \beta_{0i} + \mu^*_{yi} + \epsilon_{yik}, \tag{2.1}$$

with corresponding population means given by $\mu_{xi} = \beta_{0i} + \mu^*_{xi}$ and $\mu_{yi} = \beta_{0i} + \mu^*_{yi}$. For ease of presentation, we focus on the Gaussian model for the error terms $\epsilon_{xij} \overset{iid}{\sim} N(0, \sigma^2_{xi})$ and $\epsilon_{yij} \overset{iid}{\sim} N(0, \sigma^2_{yi})$. More general settings will be discussed in Sections 3.6. We assume that μ^*_{xi} and μ^*_{yi}, which can be viewed as random perturbations from the baseline, satisfy $\mu^*_{xi} \sim (1 - \pi_x)\delta_0 + \pi_x g_{\mu_x}(\cdot)$ and $\mu^*_{yi} \sim (1 - \pi_y)\delta_0 + \pi_y g_{\mu_y}(\cdot)$, where δ_0 is the Dirac delta function, and g_{μ_x} and g_{μ_y} are unspecified densities of nonzero effects.

Remark 1. Model (2.1) can be applied to scenarios with non-sparse $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ when some baseline measurements are available. See Section A.8 for further details. The proposed methodology only requires $\bar{X}_i$ and $\bar{Y}_i$ to be normal. In practical situations where n_x and n_y are large, our method works well without the normality assumption. Numerical results with non-Gaussian errors are provided in Section 5.3.

Let $n = n_x + n_y$. Denote $\gamma_x = n_x/n$ and $\gamma_y = n_y/n$. The population means μ_{xi} and μ_{yi} are estimated by $\bar{X}_i = (n_x)^{-1} \sum_{j=1}^{n_x} X_{ij}$ and $\bar{Y}_i = (n_y)^{-1} \sum_{k=1}^{n_y} Y_{ik}$, respectively.

The two-sample inference problem is concerned with the simultaneous testing of m hypotheses $H_{i,0} : \mu_{xi} = \mu_{yi}$ vs $H_{i,1} : \mu_{xi} \ne \mu_{yi}$, $i = 1, \cdots, m$. Let $\mathbb{I}(\cdot)$ be an indicator function. Let T_{1i} and T_{2i} be summary statistics that contain the information about $\theta_{1i} = \mathbb{I}(\mu_{xi} \ne \mu_{yi})$ (support of mean difference) and $\theta_{2i} = \mathbb{I}(\mu_{xi} \ne 0 \text{ or } \mu_{yi} \ne 0)$ (union support),

respectively. T_{1i} is the primary statistic in inference and T_{2i} is an *auxiliary covariate*. The term "auxiliary" indicates that we do not use T_{2i} to make inference on θ_{1i} directly. Instead, we aim to incorporate T_{2i} in inference to (indirectly) support the evidence provided in the primary variable T_{1i}. The intuition is that, since the union support is sparse if both $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are sparse, exploiting this structural information would improve the efficiency of tests. To see this, note that the continuity of μ_{xi} and μ_{yi} implies that, with probability one, θ_{1i} and θ_{2i} obey the following logical relationship:

$$\theta_{1i} = 0 \quad \text{if} \quad \theta_{2i} = 0. \tag{2.2}$$

Hence the auxiliary sequence can be utilized to assist inference by providing supplementary evidence on whether a hypothesis is promising.

We first discuss how to construct the primary and auxiliary statistics from the original data, and then introduce a bivariate random mixture model to describe their joint distribution. Finally, we formulate a decision-theoretic framework for two-sample simultaneous inference with an auxiliary covariate.

2.1. Constructing the primary and auxiliary statistics

A key step in our formulation is to construct a pair of statistics (T_{1i}, T_{2i}) such that (i) the pair extracts information from the data effectively; (ii) the pair leads to a simple bivariate model via which the logical relationship (2.2) can be exploited. To focus on the main ideas, we first discuss the Gaussian case with known variances. Extensions to two-sample tests with non-Gaussian errors and unknown variances are discussed in Section 3.6.

The general strategies for constructing the pair (T_{1i}, T_{2i}) can be described as follows. First, T_{1i} is used to capture the information on θ_{1i}; hence $\bar{X}_i - \bar{Y}_i$ should be incorporated in its expression. Second, to capture the information on the union support θ_{2i}, we propose to use the weighted sum $\bar{X}_i + \kappa_i \bar{Y}_i$, where $\kappa_i > 0$ is the weight to be specified later. Under the normality assumption, the covariance of $\bar{X}_i - \bar{Y}_i$ and $\bar{X}_i + \kappa_i \bar{Y}_i$ is given by $\sigma_{xi}^2/n_x - \kappa_i \sigma_{yi}^2/n_y$. This motivates us to choose the weight $\kappa_i^* = (\gamma_y \sigma_{xi}^2)/(\gamma_x \sigma_{yi}^2)$, which leads to zero correlation, a crucial property for simplifying the model and facilitating the methodological development. Finally, the difference and weighted sum are standardized to make the statistics comparable across tests. Combining the above considerations, we propose to use the following pair of statistics to summarize the information in the data:

$$(T_{1i}, T_{2i}) = \sqrt{\frac{n_x n_y}{n}} \left(\frac{\bar{X}_i - \bar{Y}_i}{\sigma_{pi}}, \frac{\bar{X}_i + \kappa_i^* \bar{Y}_i}{\sqrt{\kappa_i^*}\sigma_{pi}} \right), \tag{2.3}$$

where $\sigma_{pi}^2 = \gamma_y \sigma_{xi}^2 + \gamma_x \sigma_{yi}^2$. Denote $\boldsymbol{T}_1 = (T_{1i} : 1 \le i \le m)$ and $\boldsymbol{T}_2 = (T_{2i} : 1 \le i \le m)$.

2.2. A bivariate random mixture model

We develop a bivariate model to describe the joint distribution of T_{1i} and T_{2i}. Let $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i})$. Assume that $\boldsymbol{\theta}_i$ are independent and identically distributed (i.i.d.) bivariate random vectors that take values in the Cartesian product space $\{0,1\}^2 = \{(0,0),(0,1),(1,0),(1,1)\}$. For each combination $\boldsymbol{\theta}_i = (j,k)$, (T_{1i}, T_{2i}) are jointly distributed with conditional density $f(t_{1i}, t_{2i} | \theta_{1i} = j, \theta_{2i} = k)$. Denote $\pi_{jk} = \mathbb{P}(\theta_{1i} = j, \theta_{2i} = k)$. In practice, we do not know $(\theta_{1i}, \theta_{2i})$ but only observe (T_{1i}, T_{2i}) from a mixture model

$$f(t_{1i}, t_{2i}) = \textstyle\sum_{(j,k)\in\{0,1\}^2} \pi_{jk} f(t_{1i}, t_{2i} | \theta_{1i} = j, \theta_{2i} = k). \tag{2.4}$$

Denote $\pi_j = \mathbb{P}(\theta_{ji} = 1)$, $j = 1, 2$. Assume that $\pi_1 > 0$. The goal is to determine the value of θ_{1i} based on pairs $\{(T_{1i}, T_{2i}) : 1 \le i \le m\}$.

The mixture model (2.4) is difficult to analyze. However, if T_{1i} and T_{2i} are carefully constructed as done in Section 2.1, then several simplifications can be made. First, the logical relationship (2.2) implies that $\pi_{10} = 0$; thus we only have three terms in (2.4). Second, according to our construction (2.3), T_{1i} and T_{2i} are conditionally independent:

$$f(t_{1i}, t_{2i}|\mu_{xi}, \mu_{yi}) = f(t_{1i}|\mu_{xi}, \mu_{yi}) f(t_{2i}|\mu_{xi}, \mu_{yi}). \tag{2.5}$$

The next proposition utilizes (2.5) to further simplify the model.

PROPOSITION 1. *The conditional independence* (2.5) *implies that*

$$\begin{aligned} f(t_{1i}, t_{2i}|\theta_{1i} = 0, \theta_{2i} = 0) &= f(t_{1i}|\theta_{1i} = 0) f(t_{2i}|\theta_{2i} = 0); \\ f(t_{1i}, t_{2i}|\theta_{1i} = 0, \theta_{2i} = 1) &= f(t_{1i}|\theta_{1i} = 0) f(t_{2i}|\theta_{1i} = 0, \theta_{2i} = 1); \quad \textit{and} \\ f(t_{1i}, t_{2i}|\theta_{1i} = 0) &= f(t_{1i}|\theta_{1i} = 0) f(t_{2i}|\theta_{1i} = 0). \end{aligned} \tag{2.6}$$

The last equation shows that T_{1i} and T_{2i} are independent under the null hypothesis $H_{i0} : \theta_{1i} = 0$. This is a critical result for our later methodological and theoretical developments. The joint density is given by

$$\begin{aligned} f(t_{1i}, t_{2i}) = {} & \pi_{00} f(t_{1i}|\theta_{1i} = 0) f(t_{2i}|\theta_{2i} = 0) + \pi_{01} f(t_{1i}|\theta_{1i} = 0) f(t_{2i}|\theta_{1i} = 0, \theta_{2i} = 1) \\ & + \pi_{11} f(t_{1i}, t_{2i}|\theta_{1i} = 1, \theta_{2i} = 1). \end{aligned} \tag{2.7}$$

2.3. Problem formulation

Our goal is to make inference on $\theta_{1i} = \mathbb{I}(\mu_{xi} \neq \mu_{yi})$, $1 \le i \le m$, by simultaneously testing m hypotheses $H_{i,0} : \theta_{1i} = 0$ vs $H_{i,1} : \theta_{1i} = 1$. Compared to conventional approaches, we aim to develop methods utilizing m pairs $\{(T_{1i}, T_{2i}) : 1 \le i \le m\}$ instead of a single vector $\{T_{1i} : 1 \le i \le m\}$. This new problem can be recast and solved in the framework of multiple testing with a covariate: T_{1i} is viewed as the primary statistic for assessing significance, and T_{2i} is viewed as a covariate to assist inference by providing supporting information.

The concepts of error rate and power are similar to those in the conventional settings. A multiple testing procedure is represented by a thresholding rule of the form

$$\boldsymbol{\delta} = \{\delta_i = \mathbb{I}(S_i < t) : i = 1, \cdots, m\} \in \{0, 1\}^m, \tag{2.8}$$

where $\delta_i = 1$ if we reject hypothesis i and $\delta_i = 0$ otherwise. Here S_i is a significance index that ranks the hypotheses from the most significant to least significant, and t is a threshold.

In large-scale testing problems, the false discovery rate (FDR, Benjamini and Hochberg, 1995) has been widely used to control the inflation of Type I errors. For a given decision rule $\boldsymbol{\delta} = (\delta_i : 1 \le i \le m)$ of the form (2.8), the FDR is defined as

$$\mathrm{FDR}_{\boldsymbol{\delta}} = \mathbb{E}\left[\frac{\sum_{i=1}^m (1 - \theta_{1i})\delta_i}{(\sum_{i=1}^m \delta_i) \vee 1}\right], \tag{2.9}$$

where $x \vee y = \max(x, y)$. A closely related concept is the marginal false discovery rate (mFDR), which is defined by

$$\mathrm{mFDR}_{\boldsymbol{\delta}} = \frac{\mathbb{E}\{\sum_{i=1}^m (1 - \theta_{1i})\delta_i\}}{\mathbb{E}(\sum_{i=1}^m \delta_i)}. \tag{2.10}$$

Genovese and Wasserman (2002) showed that mFDR = FDR + $O(m^{-1/2})$ when the BH procedure (Benjamini and Hochberg, 1995) is applied to m independent tests. We use mFDR mainly for technical considerations to obtain optimality result. Proposition 7 in Section 7.2 gives sufficient conditions under which the mFDR and FDR are asymptotically equivalent and shows that the conditions are fulfilled by our proposed method.

Define the expected number of true positives $\mathrm{ETP}_{\boldsymbol{\delta}} = \mathbb{E}\left(\sum_{i=1}^{m} \theta_{1i}\delta_i\right)$. Other related power measures include the missed discovery rate (MDR, Taylor et al., 2005), average power (Efron, 2007) and false non-discovery/negative rate (FNR, Genovese and Wasserman, 2002; Sarkar, 2002). Our optimality result is developed based on the mFDR and ETP. We call a multiple testing procedure *valid* if it controls the mFDR at the nominal level and *optimal* if it has the largest ETP among all valid mFDR procedures.

3. Oracle and Data-driven Procedures

The basic framework of our methodological developments is explained as follows. We first consider an ideal situation where an oracle knows all parameters in model (2.7). Section 3.1 derives an oracle procedure. Sections 3.2 and 3.3 discuss an approximation strategy and related estimation methods, with a refinement given in Section 3.4. The data-driven procedure and extensions are presented in Sections 3.5 and 3.6.

3.1. Oracle procedure with pairs of observations

The marginal density function for T_{ji} is defined as $f_{j\cdot} = (1-\pi_j)f_{j0} + \pi_j f_{j1}$, where $\pi_j = \mathbb{P}(\theta_{ji} = 1)$ and $f_{j0} = f(t_{ji}|\theta_{ji} = 0)$ and $f_{j1} = f(t_{ji}|\theta_{ji} = 1)$ are the conditional densities for T_{ji}. Conventional FDR procedures, which are developed based on a vector of p-values or z-values, are essentially *univariate inference procedures* that only utilize the information of T_{1i}. Define the local false discovery rate (Lfdr, Efron et al., 2001)

$$\mathrm{Lfdr}_1(t_1) = \frac{(1-\pi_1)f_{10}(t_1)}{f_{1\cdot}(t_1)}, \tag{3.1}$$

where subscript "1" indicates a quantity associated with T_{1i}. It was shown in Sun and Cai (2007) that the optimal univariate mFDR procedure is a thresholding rule of the form

$$\boldsymbol{\delta}(\mathrm{Lfdr}_1, c) = [\mathbb{I}\{\mathrm{Lfdr}_1(t_{1i}) < c\} : 1 \le i \le m], \tag{3.2}$$

where $0 \le c \le 1$ is a cutoff. Denote $Q_{\mathrm{LF}}(c)$ the mFDR level of $\boldsymbol{\delta}(\mathrm{Lfdr}_1, c)$. Let $c^* = \sup\{c : Q_{\mathrm{LF}}(c) \le \alpha\}$ be the largest cutoff under the mFDR constraint. Then $\boldsymbol{\delta}^* = \boldsymbol{\delta}(\mathrm{Lfdr}_1, c^*)$ is optimal among all univariate mFDR procedures in the sense that it has the largest ETP subject to mFDR $\le \alpha$.

The next theorem derives an oracle procedure for mFDR control when the pairs (T_{1i}, T_{2i}) are given. We shall see that the performance of $\boldsymbol{\delta}^*$, the optimal univariate procedure, can be greatly improved by exploiting the information in T_{2i}. The oracle procedure under the bivariate model (2.7) has two important components: an oracle statistic T^i_{OR} that optimally pools information from both T_{1i} and T_{2i}, and an oracle threshold t_{OR} that controls the mFDR with the largest ETP.

THEOREM 1. *Suppose* (T_{1i}, T_{2i}) *follow model* (2.7). *Let*

$$q^*(t_2) = (1-\pi_1) f\left(t_2 \mid \theta_{1i} = 0\right). \tag{3.3}$$

Define the oracle statistic

$$T_{OR}^i(t_1,t_2)=\mathbb{P}(\theta_{1i}=0|T_{1i}=t_1,T_{2i}=t_2)=\frac{q^*(t_2)f_{10}(t_1)}{f(t_1,t_2)}, \tag{3.4}$$

where $f(t_1,t_2)$ is the joint density given by (2.7)*. Then*

(a) For $0<\lambda\le 1$, let $Q_{OR}(\lambda)$ be the mFDR level of testing rule $\{\mathbb{I}(T_{OR}^i<\lambda):1\le i\le m\}$. Then $Q_{OR}(\lambda)<\lambda$ and $Q_{OR}(\lambda)$ is non-decreasing in λ.

(b) Suppose we choose $\alpha<\bar{\alpha}\equiv Q_{OR}(1)$. Then the oracle threshold $\lambda_{OR}=\sup\{\lambda: Q_{OR}(\lambda)\le\alpha\}$ exists uniquely and $Q_{OR}(\lambda_{OR})=\alpha$. Furthermore, define oracle rule $\boldsymbol{\delta}_{OR}=(\delta_{OR}^i:i=1,\cdots,m)$, where

$$\delta_{OR}^i=\mathbb{I}(T_{OR}^i<\lambda_{OR}). \tag{3.5}$$

Then $\boldsymbol{\delta}_{OR}$ is optimal in the sense that $ETP_{\boldsymbol{\delta}}\le ETP_{\boldsymbol{\delta}_{OR}}$ for any $\boldsymbol{\delta}$ in $\mathcal{D}_\alpha$, where $\mathcal{D}_\alpha$ is the collection of all testing rules based on $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ such that $mFDR_{\boldsymbol{\delta}}\le\alpha$.

REMARK 2. The oracle statistic T_{OR}^i is the posterior probability that $H_{i,0}$ is true given the pair of primary and auxiliary statistics. It serves as a significance index providing evidence against the null. Section 3.2 gives a detailed discussion of $q^*(t_2)$ and explains that it roughly describes how frequently T_{2i} from the null distribution would fall into the neighborhood of t_2. The estimation of T_{OR} and $q^*(t_2)$ is discussed in Section 3.3.

REMARK 3. Theorem 1 indicates that pooling auxiliary information would not result in efficiency loss, provided that T_{2i} are carefully constructed according to the principles described in Section 2.1. Consider the "worst case scenario" where T_{2i} is completely non-informative: $n_x=n_y$, $\sigma_{xi}^2=\sigma_{yi}^2$ and $\mu_{xi}=-\mu_{yi}$. In Section A.9 of the Supplementary Material, we show that under the above conditions T_{OR}^i reduces to the Lfdr statistic (3.1), and the oracle (bivariate) procedure would coincide with the optimal univariate rule (3.2). Contrary to the intuition that incorporating T_{2i} might negatively affect the performance, Theorem 1 indicates the power will unlikely be decreased by pooling the auxiliary information in T_{2i}. Further numerical evidence is provided in Section 5.4.

The oracle rule (3.5) motivates us to consider a stepwise procedure that operates in two steps: ranking and thresholding. The ranking step orders all hypotheses from the most significant to the least significant according to T_{OR}, and the thresholding step identifies the largest threshold along the ranking subject to the constraint on FDR. Specifically, denote $T_{OR}^{(1)}\le\ldots\le T_{OR}^{(m)}$ the ordered oracle statistics and $H_{(1)},\cdots,H_{(m)}$ the corresponding hypotheses. The step-wise procedure operates as follows.

$$\text{Let } k=\max\Big\{j: j^{-1}\textstyle\sum_{i=1}^j T_{OR}^{(i)}\le\alpha\Big\}.\ \text{Reject } H_{(1)},\cdots,H_{(k)}. \tag{3.6}$$

The moving average of the top j ordered statistics, gives an estimate of the FDR [cf. Sun and Cai (2007)]. Thus the stepwise algorithm (3.6) identifies the largest threshold subject to the FDR constraint.

3.2. Approximating T_{OR} via screening

The oracle statistic T^i_{OR} is unknown and needs to be estimated. However, standard methods do not work well for the bivariate model. For example, the popular EM algorithm usually requires the specification of a parametric form of the non-null distribution; this is often impractical in large-scale studies where little is known about the alternative. Moreover, existing estimators often suffer from low accuracy and convergence issues when signals are sparse. To overcome the difficulties in estimation, we propose a new test statistic $T^{\tau,i}_{OR}$ that only involves quantities that can be well estimated from data. The new statistic provides a good approximation to T^i_{OR} and guarantees the FDR control.

In (3.4), the null density f_{10} is known by construction. The bivariate density $f(t_1, t_2)$ can be well estimated using a standard kernel method [Silverman (1986); Wand and Jones (1995)]. Hence we shall focus on the quantity $q^*(t_2)$. Suppose we are interested in counting how frequently T_{2i} from the null distribution (i.e. $\theta_{1i} = 0$) would fall into an interval in the neighborhood of t_2: $Q^*(t_2, h) = \#\{i : T_{2i} \in [t_2 - h/2, t_2 + h/2] \text{ and } \theta_{1i} = 0\}/m$. The quantity is relevant because $q^*(t_2) = \lim_{h\to 0} \mathbb{E}\{Q^*(t_2, h)\}/h$. The counting task is difficult as we do not know the value of θ_{1i}. Our idea is to first apply a screening method to select the nulls (i.e. $\theta_{1i} = 0$), and then construct an estimator based on selected cases.

Denote P_i the p-value associated with $T_{1i} = t_{1i}$. For a large τ, say $\tau = 0.9$, we would reasonably predict that $\theta_{1i} = 0$ if $P_i > \tau$, as most likely large p-values should be from the null. Hence we may count those T_{2i} with large p-values:

$$Q^\tau(t_2, h) = \frac{\#\{i : T_{2i} \in [t_2 - h/2, t_2 + h/2] \text{ and } P_i > \tau\}}{m(1-\tau)}. \tag{3.7}$$

The adjustment $(1-\tau)$ in the denominator comes from the fact that we have only utilized $100(1-\tau)\%$ of the data while counting the frequency. Let $\mathcal{A}_\tau$ denote the set of possible t_{1i} such that $P_i > \tau$. Using Q^τ to replace Q^*, a sensible approximation of $q^*(t_2)$ would be

$$q^\tau(t_2) = \lim_{h\to 0} \frac{\mathbb{E}\{Q^\tau(t_2, h)\}}{h} = \frac{\int_{\mathcal{A}_\tau} f(t_1, t_2) dt_1}{1-\tau}. \tag{3.8}$$

Intuitively, a large τ would yield a sample that is close to one generated from a "pure" null distribution and thus reduce the bias $q^\tau(t_2) - q^*(t_2)$. Our theory reveals that the bias is always positive (Proposition 2), and would decrease in τ (Proposition 4). However, a larger τ would increase the variability of our proposed estimator (as we have fewer samples to construct the estimator), affecting the testing procedure adversely. The bias-variance tradeoff is further discussed in Section 3.4.

Substituting $q^\tau(t_2)$ in place of $q^*(t_2)$, we obtain the following approximation of T^i_{OR}:

$$T^{\tau,i}_{OR}(t_1, t_2) = \frac{q^\tau(t_2) f_{10}(t_1)}{f(t_1, t_2)}. \tag{3.9}$$

Some important properties of the approximation in (3.9) are summarized in the next proposition, which shows that $T^{\tau,i}_{OR}$ overestimates T^i_{OR}. Hence if we substitute $T^{\tau,i}_{OR}$ in place of T^i_{OR} in (3.6), then fewer rejections will be made, leading to a conservative FDR level.

PROPOSITION 2. *(a) $T^i_{OR}(t_1, t_2) \le T^{\tau,i}_{OR}(t_1, t_2)$ for all τ. (b) Let $\boldsymbol{\delta}^\tau_{OR}$ be a decision rule that substitutes $T^{\tau,i}_{OR}$ in place of T^i_{OR} in (3.6). Then both the FDR and mFDR levels of $\boldsymbol{\delta}^\tau_{OR}$ are controlled below level α.*

3.3. Estimation of the test statistic

We now turn to the estimation of $T_{OR}^{\tau,i}$. By our construction, the null density $f_{10}(t_1)$ is known. The bivariate density $f(t_1, t_2)$ can be estimated using a kernel method [Silverman (1986); Wand and Jones (1995)]:

$$\hat{f}(t_1, t_2) = m^{-1} \sum_{i=1}^{m} K_{h_1}(t_1 - t_{1i}) K_{h_2}(t_2 - t_{2i}), \tag{3.10}$$

where $K(t)$ is a kernel function, h_1 and h_2 are the bandwidths, with $K_h(t) = h^{-1}K(t/h)$. To estimate $q^\tau(t_2)$, we first carry out a screening procedure to obtain sample $\mathcal{T}(\tau) = \{i : P_{1i} > \tau\}$, and then apply kernel smoothing to the selected observations:

$$\hat{q}^\tau(t_2) = \frac{\sum_{i \in \mathcal{T}(\tau)} K_{h_2}(t_2 - t_{2i})}{m(1-\tau)}. \tag{3.11}$$

The next proposition shows that $\hat{q}^\tau(\cdot)$ converges to $q^\tau(\cdot)$ in L_2 norm.

PROPOSITION 3. *Consider $\hat{q}^\tau$ and q^τ respectively defined in* (3.8) *and* (3.11). *Assume that (i) $q^\tau(\cdot)$ is bounded and have continuous first and second derivatives; (ii) the kernel K is a positive, bounded and symmetric function satisfying $\int K(t) = 1$, $\int tK(t)dt = 0$ and $\int t^2 K(t)dt < \infty$; and (iii) $f_{2\cdot}^{(2)}(t_2|\tau) = \int_{t_1 \in \mathcal{A}_\tau} \int f_{2\cdot}^{(2)}(t_2|t_1) f_{1\cdot}(t_1) dt_1$ is square integrable, where $f_{2\cdot}(t_2|t_1)$ is the conditional density of T_2 given T_1. Then with the common choice of bandwidth $h \sim m^{-1/6}$, we have*

$$\mathbb{E} \|\hat{q}^\tau - q^\tau\|^2 = \mathbb{E} \int \{\hat{q}^\tau(x) - q^\tau(x)\}^2 dx \to 0.$$

Combining the above results, we propose to estimate T_{OR}^τ by the following statistic

$$\hat{T}_{OR}^\tau(t_1, t_2) = \frac{\hat{q}^\tau(t_2) f_{10}(t_1)}{\hat{f}(t_1, t_2)} \wedge 1, \tag{3.12}$$

where $\hat{q}^\tau(t_2)$ and $\hat{f}(t_1, t_2)$ are respectively given in (3.11) and (3.10), and $x \wedge y = \min(x, y)$.

REMARK 4. In our proposed estimator, the same bandwidth h_2 has been used for the kernels in both (3.10) and (3.11). Utilizing the same bandwidth across the numerator and denominator of (3.12) has no impact on the theory, but is beneficial for increasing the stability of our estimator. More practical guidelines are provided in Section 5.1.

3.4. A refined estimator

This section develops a consistent estimator of $q^*(t_2)$. The proposed estimator is important for the optimality theory in Section 3.5. However, it is computationally intensive and requires much stronger assumptions which should be scrutinized in practice. The power gain tends to be limited. In practice we still recommend the simple estimator (3.11). This section may be skipped for readers who are mainly interested in methodology.

We state in the next proposition some theoretical properties for the approximation error $q^\tau(t_2) - q^*(t_2)$; these properties are helpful to motivate the new estimator and prove its consistency. Let the CDF of the p-value associated with T_{1i} be $G(\tau) = (1-\pi_1)\tau + \pi_1 G_1(\tau)$, where G_1 is the alternative CDF. Denote g and g_1 the corresponding density functions.

PROPOSITION 4. *Consider* T^{τ}_{OR} *defined in* (3.9).

(a). Denote $B_q(\tau) = \int |q^{\tau}(t_2) - q^*(t_2)|dt_2$ *the total approximation error. If* $G_1(\cdot)$ *is concave, then* $B_q(\tau)$ *decreases in* τ.

(b). If $\lim_{x\uparrow 1} g_1(x) = 0$, *then* $\lim_{\tau\uparrow 1} q^{\tau}(t_2) = q^*(t_2)$.

REMARK 5. The concavity assumption (or the more general monotone likelihood ratio condition, MLRC) has been commonly used in the literature (Storey, 2002; Genovese and Wasserman, 2002; Sun and Cai, 2007); the MLRC should be treated with caution (Cao et al., 2013). Assumption (b) is also a typical condition (Genovese and Wasserman, 2004), which requires that the null cases are dominant on the right of the p-value histogram. The condition holds for one-sided p-values but can be violated by two-sided p-values (Neuvial, 2013). It would be desirable to develop a more general condition in future work.

It follows from Proposition 4 that a large τ is helpful to reduce the bias and the bias converges to zero when $\tau \to 1$. However, a large τ would increase the variance of our estimator (3.11), which is constructed using the sample $\mathcal{T}(\tau) = \{i : P_{1i} > \tau\}$. To address the bias-variance tradeoff, we propose to first obtain $\hat{q}^{\tau}$ for a range of τ's, say $\{\tau_1, \cdots, \tau_k\}$, and then use a smoothing method to obtain the limiting value of $\hat{q}^{\tau}$ when $\tau \to 1$. This approach aims to borrow strength from the entire sample to minimize the bias without blowing up the variance. Specifically, let $\tau_0 < \tau_1 < \cdots < \tau_k$ be ordered and equally spaced points in the inteval $(0, 1)$. Denote $\hat{q}^{\tau_j}(t_2)$ the estimates from (3.11), $j = 1, \cdots, k$. We propose to obtain the local linear kernel estimator $\hat{q}^*(t_2) \equiv \hat{q}\{\tau = 1; \hat{q}^{\tau_1}(t_2), \cdots, \hat{q}^{\tau_k}(t_2)\}$ as the height of the fit $\hat{\beta}_0$, where $(\hat{\beta}_0, \hat{\beta}_1)$ minimizes the weighted kernel least squares $\sum_{j=1}^{k}(\hat{q}^{\tau_j} - \beta_0 - \beta_1\tau_j)^2 K_{h_\tau}(\tau_j - \tau_k)$. For a given integer r, denote $\hat{s}_r = k^{-1}\sum_{j=1}^{k}(\tau_j - 1)^r K_{h_\tau}(\tau_j - \tau_k)$. It can be shown that (e.g. Wand and Jones, 1995, pp. 119)

$$\hat{q}^*(t_2) = k^{-1}\textstyle\sum_{j=1}^{k} \dfrac{\{\hat{s}_2 - \hat{s}_1(\tau_j - \tau_k)\}K_{h_\tau}(\tau_j - \tau_k)\hat{q}^{\tau_j}}{\hat{s}_2\hat{s}_0 - \hat{s}_1^2}. \tag{3.13}$$

The next proposition shows that $\hat{q}^*(t_2)$ is a consistent estimator for $q^*(t_2)$.

PROPOSITION 5. *Consider* $\hat{q}^{\tau}$ *and* $\hat{q}^*$ *that are respectively defined in* (3.11) *and* (3.13). *Denote* $q^{\tau_k,(2)}(t_2) = (d/d\tau)^2 q^{\tau,(2)}(t_2)|_{\tau=\tau_k}$. *Assume the following conditions hold: (i)* $q^{\tau_k,(2)}(t_2)$ *is square integrable; (ii)* $K(\cdot)$ *is symmetric about zero and is supported on* $[-1, 1]$*; and (iii) the bandwidth* h_τ *is a sequence satisfying* $h_\tau \to 0$ *and* $kh_\tau \to \infty$ *as* $k \to \infty$. *Moreover, assume Condition (i) to (iii) in Proposition 3 and Condition (b) in Proposition 4 hold. We have*

$$\mathbb{E}\,\|\hat{q}^* - q^*\|^2 = \mathbb{E}\int\{\hat{q}^*(x) - q^*(x)\}^2 dx \to 0 \quad \textit{when } (m, k) \to 0. \tag{3.14}$$

3.5. The CARS procedure

The estimated statistics $\hat{T}^{\tau,i}_{OR}$ will be used as a significance index to rank the relative importance of all hypotheses. Motivated by the stepwise algorithm (3.6), we propose the following covariate-assisted ranking and screening (CARS) procedure.

PROCEDURE 1 (CARS PROCEDURE). *Consider model* (2.7) *and estimated statistics* $\hat{T}^{\tau,i}_{OR}$ *(3.12). Denote* $\hat{T}^{\tau,(1)}_{OR} \le \cdots \le \hat{T}^{\tau,(m)}_{OR}$ *the ordered statistics and* $H_{(1)}, \cdots, H_{(m)}$ *the corresponding hypotheses. Let* $k = \max\left\{j : j^{-1}\sum_{i=1}^{j}\hat{T}^{\tau,(i)}_{OR} \le \alpha\right\}$. *Then reject* $H_{(1)}, \cdots, H_{(k)}$.

To ensure good performance of the data-driven procedure, we require the following conditions for estimated quantities:

(C1). $\mathbb{E}\|\hat{q}^\tau - q^\tau\|^2 \to 0$. (C1′). $\mathbb{E}\|\hat{q}^* - q^*\|^2 \to 0$.

(C2). $\mathbb{E}\left\|\hat{f} - f\right\|^2 = \mathbb{E}\left[\int\int\{\hat{f}(t_1,t_2) - f(t_1,t_2)\}^2 dt_1 dt_2\right] \to 0$.

REMARK 6. Proposition 3 shows that (C1) is satisfied by the proposed estimator (3.11). Proposition 5 shows that (C1′) is satisfied by the smoothing estimator (3.14) under stronger assumptions. Finally, (C2) is satisfied by the standard choice of bandwidth $h_{t1} \sim m^{-1/6}$, $h_{t2} \sim m^{-1/6}$; see, for example, page 111 in Wand and Jones (1995) for a proof.

The asymptotic properties of the CARS procedure are established by the next theorem.

THEOREM 2. Asymptotic validity and optimality of CARS.

(a). If Conditions (C1) and (C2) hold, then both the mFDR and FDR of the CARS procedure are controlled at level $\alpha + o(1)$.

(b). If Conditions (C1′) and (C2) hold, and we substitute $\hat{q}^*$ (3.14) *in place of* $\hat{q}^\tau$ *in* (3.12) *to compute* $\hat{T}^\tau_{OR}$, *then the FDR level of the CARS procedure is* $\alpha + o(1)$. *Moreover, denote* ETP_{CARS} *and* ETP_{OR} *the ETP levels of CARS and the oracle procedure, respectively. Then we have* $ETP_{CARS}/ETP_{OR} = 1 + o(1)$.

3.6. The case with unknown variances and non-Gaussian errors

For two-sample tests with unknown and unequal variances, we can estimate σ^2_{xi} and σ^2_{yi} by $S^2_{xi} = (n_x)^{-1}\sum_{j=1}^{n_x}(X_{ij} - \bar{X}_i)^2$ and $S^2_{yi} = (n_y)^{-1}\sum_{j=1}^{n_y}(Y_{ij} - \bar{Y}_i)^2$, respectively. Let $\hat{\kappa}^*_i = (\gamma_y S^2_{xi})/(\gamma_x S^2_{yi})$ and $S^2_{pi} = \gamma_y S^2_{xi} + \gamma_x S^2_{yi}$. The following pair will be used to summarize the information in the sample:

$$(T_{1i}, T_{2i}) = \sqrt{\frac{n_x n_y}{n}}\left(\frac{\bar{X}_i - \bar{Y}_i}{S_{pi}}, \frac{\bar{X}_i + \hat{\kappa}^*_i \bar{Y}_i}{\sqrt{\hat{\kappa}^*_i} S_{pi}}\right). \tag{3.15}$$

For the case with unknown but equal variances (e.g. $\sigma^2_{xi} = \sigma^2_{yi}$), we modify (3.15) as follows. First, $\hat{\kappa}^*_i$ is replaced by $\kappa^* = \gamma_y/\gamma_x$. Second, S^2_{pi} is instead estimated by $S^2_{pi} = \gamma_x S^2_{xi} + \gamma_y S^2_{yi}$. Finally T_{1i} and T_{2i} are plugged into (3.12) to compute the CARS statistic, which is further employed to implement Procedure 1.

T_{1i} and T_{2i} are not strictly independent when estimated variances are used. The following proposition shows that T_{1i} and T_{2i} are asymptotically independent under the null.

PROPOSITION 6. *Consider model* (2.1). *Assume that the error terms (possibly non-Gaussian) of* X_{ij} *and* Y_{ij} *have symmetric distributions and finite fourth moments. Then* (T_{1i}, T_{2i}) *defined in* (3.15) *are asymptotically independent when* $H_{i,0} : \mu_{xi} = \mu_{yi}$ *is true.*

The expression for asymptotic variance-covariance matrix, which is given in Section A.6 of the Supplementary Material, reveals that the asymptotic independence holds for non-Gaussian errors as long as the error distributions are symmetric. Our simulation results in Section 5.3 confirm that unknown variances and non-Gaussian errors have almost no impact on the performance of CARS. Therefore the plug-in methods are recommended for practical applications. The case with skewed distribution requires further research, and a full theoretical justification of CARS methodology is still an open question.

4. Extensions and Connections with Existing Work

This section considers the extension of our theory to a general bivariate model. The results in the general model provide a unified theoretical framework for understanding different testing strategies, which helps gain insights on the connections between existing methods.

We substitute (T_i, S_i) in place of (T_{1i}, T_{2i}) in this section. This change reflects a more flexible view of the auxiliary covariate: S_i can be either continuous or discrete, from either internal or external data, and we do not explicitly estimate the joint density of T_i and S_i as done in previous sections. Sections 4.1 to 4.5 assume that T_i follow a continuous distribution with a known density under the null; the case with unknown null density is discussed in Section 4.6. We allow S_i to be either continuous or categorical and hence eliminate the notation θ_{2i}. (Previously θ_{2i} denotes the union support, which is only needed when T_{2i} has a density with a point mass at 0.) As a result, the subscript "1" in θ_{1i} is suppressed for notational convenience, where $\theta_i = 0/1$ stands for a null/non-null case.

4.1. A general bivariate model

Suppose T_i and S_i follow a joint distribution $F_i(t, s)$. The optimal (oracle) testing rule is given by the next theorem, which can be proven similarly as Theorem 1.

THEOREM 3. *Define the oracle statistic under the general model*

$$T^G_{OR}(t, s) = \mathbb{P}(\theta_i = 0 | T_i = t, S_i = s). \tag{4.1}$$

Denote $Q^G_{OR}(\lambda)$ the mFDR level of $\boldsymbol{\delta}(T^G_{OR}, \lambda)$, where $\boldsymbol{\delta}(T^G_{OR}, \lambda) = \{\mathbb{I}(T^G_{OR}(t, s) < \lambda) : 1 \le i \le m\}$. Let $\lambda_{OR} = \sup\{\lambda \in (0, 1) : Q^G_{OR}(\lambda) \le \alpha\}$. Define the oracle mFDR procedure under the general model as $\boldsymbol{\delta}^G_{OR} = \boldsymbol{\delta}(T^G_{OR}, \lambda_{OR})$. Then $\boldsymbol{\delta}^G_{OR}$ is optimal in the sense that for any $\boldsymbol{\delta}$ such that $mFDR_{\boldsymbol{\delta}} \le \alpha$, we always have $ETP_{\boldsymbol{\delta}} \le ETP_{\boldsymbol{\delta}^G_{OR}}$.

Theorem 1 can be viewed as a special case of Theorem 3. However, Theorem 3 is of less practical importance as the "best" data-driven solution to Theorem 3 may depend on a number of factors such as (i) whether the auxiliary statistic is categorical or continuous; (ii) whether the null distribution of T_i is fixed and known; (iii) whether S_i and T_i are independent, and etc. The key issue is that estimating $T^{G,i}_{OR}$ is very difficult in a general bivariate model. Under some special cases, T^G_{OR} can be approximated well. For example, T^τ_{OR} provides a good approximation to T^G_{OR} under bivariate model (2.4) and the conditional independence assumption (2.6). When S_i is categorical, the oracle procedure can also be approximated well. This important special case is discussed next.

4.2. Discrete case: multiple testing with groups

We now consider a special case where the auxiliary covariate S_i is categorical. A concrete scenario is the *multi-group random mixture model* first introduced in Efron (2008). See also Cai and Sun (2009). The model is useful to handle large-scale testing problems where data are collected from heterogeneous sources. Correspondingly, the m hypotheses may be divided into, say, K groups that exhibit different characteristics. Let S_i denote the group membership. Assume that S_i takes values in $\{1, \cdots, K\}$ with prior probabilities $\{\pi_1, \cdots, \pi_K\}$. Consider the following conditional distributions:

$$(T_i | S_i = k) \sim f^k_1 = (1 - \pi^k_1) f^k_{10} + \pi^k_1 f^k_{11}, \tag{4.2}$$

for $k = 1, \cdots, K$, where π_1^k is the proportion of non-null cases in group k, f_{10}^k and f_{11}^k are the null and non-null densities of T_i, and $f_1^k = (1-\pi_1^k)f_{10}^k + \pi_1^k f_{11}^k$ is the mixture density for all observations in group k. The model allows the conditional distributions in (4.2) to vary across groups; this is desirable in practice when groups are heterogeneous.

REMARK 7. In Section A.10 in the supplement, we present a simple example to show that T_i is not sufficient (as insightfully pointed out by a reviewer), whereas (T_i, S_i) is sufficient. S_i is *ancillary* in the sense that its value is determined by an external process independent from the main parameter. Contrary to the common intuition that S_i is "useless" for inference, our analysis reveals that S_i can be informative. The phenomenon is referred to as the *ancillarity paradox* because, to quote Lehmann (page 420, Lehmann and Casella, 2006), "the distribution of the ancillary, which should not affect the estimation of the main parameter, has an enormous effect on the properties of the standard estimator." A related phenomenon in the estimation context was discussed by Foster and George (1996). See also the seminal work by Brown (1990) for a paradox in multiple regression.

The problem of multiple testing with groups and related problems have received substantial attention in the literature (Efron, 2008; Ferkingstad et al., 2008; Cai and Sun, 2009; Hu et al., 2010; Liu et al., 2016; Barber and Ramdas, 2017). It can be shown that under model (4.2), the oracle statistic (4.1) is reduced to the conditional local false discovery rate (CLfdr, Cai and Sun 2009) $\text{CLfdr}_i = (1-\pi_1^k)f_{10}^k(t_i)/f_1^k(t_i)$ for $S_i = k$, $k = 1, \cdots, K$. The CLfdr statistic can be accurately estimated from data when the number of tests is large in separate groups. However, the CLfdr statistic cannot be well estimated when the number of groups is large, or when S_i becomes a continuous variable. An important recent progress for exploiting the grouping and hierarchical structures among hypotheses under more generic settings has been made in Liu et al. (2016), wherein an interesting decomposition of the oracle statistic was derived:

$$T_{OR}^i(t,s) = 1 - \{1 - P(\theta_{2i} = 0|t,s)\}\{1 - P(\theta_{1i} = 0|t,s,\theta_{2i} = 1)\}.$$

The decomposition explicitly shows how the auxiliary statistic can be used to adjust the Lfdr statistic, and provides insights on how the grouping effects S_i and individual effects T_i interplay in simultaneous testing. The logical correlation (2.2) can be conceptualized as a hierarchical constraint and exploited more efficiently (Sarkar and Zhao, 2017).

The result on multiple testing with groups motivates an interesting strategy to approximate the oracle rule. For a continuous auxiliary covariate S_i, we can first discretize S_i, then divide the hypotheses into groups according to the discrete variable, and finally apply group-wise multiple testing procedures. This idea is closely related to the uncorrelated screening (US) method in Liu (2014); the connection is discussed next.

4.3. Discretization and uncorrelated screening

The idea in Liu (2014) involves discretizing a continuous S_i as a binary variable. Define index sets $\mathcal{G}_1 = \{1 \le i \le m : S_i > \lambda\}$ and $\mathcal{G}_2 = \{1 \le i \le m : S_i \le \lambda\}$, where the tuning parameter λ divides t_{1i}'s into two groups: $\mathcal{T}_1(\mathcal{G}_1) = \{t_i : i \in \mathcal{G}_1\}$ and $\mathcal{T}_1(\mathcal{G}_2) = \{t_i : i \in \mathcal{G}_2\}$. The uncorrelated screening (US, Liu 2014) method operates in two steps. First, the BH method (Benjamini and Hochberg, 1995) is applied at level α to the two groups separately, and then the rejected hypotheses from two groups are combined as the final output. The tuning parameter λ is chosen in a way such that it yields the largest number of rejections (two groups combined). US is closely related to the *separate analysis* strategy

proposed in Efron (2008). The key difference is that the groups are known *a priori* in Efron (2008), whereas the groups are chosen adaptively in Liu (2014). The main merit of US is that the screening statistic is constructed to be uncorrelated with the test statistic, which ensures that the selection bias issue can be avoided. Moreover, the divide-and-test strategy combines the results in both groups; this is different from conventional independent filtering approaches [Bourgon et al. (2010)], in which one group is completely filtered out.

We now compare different methods under a unified framework. Both US and CARS can be viewed as approximations of the oracle rule (4.1). The goal is to borrow information from the external covariate S_i to improve the efficiency of simultaneous inference. US adopts the divide-and-test strategy and only models S_i as a binary variable. It suffers from information loss in the discretization step. Specifically, the auxiliary variables S_i can be used to reveal other useful data structures in addition to sparsity. Consider a toy example where the cases on the union support can be divided into two types, respectively characterized by low and high baseline activities; and among the more active ones, a larger proportion would exhibit differential levels between the two conditions. Intuitively the auxiliary statistics can be used to identify three groups, respectively with no, low and high activities. Hence the two-group strategy utilized by US can be potentially outperformed by a three-group strategy that captures the underlying data structure more effectively. In practice, the data structure can be complex and finding the "best" grouping is tricky; this sheds lights on the superiority of CARS, for it fully utilizes the auxiliary data by modeling S_i as a continuous variable.

The general framework suggests several directions in which US may be improved. First, the information of S_i may be better exploited, e.g. by creating more groups. However, it remains unknown how to choose the optimal number of groups. Second, US tests the hypotheses at FDR level α for both groups. However, Cai and Sun (2009) showed that the choice of identical FDR levels across groups can be suboptimal. In order to maximize the overall power, different group-wise FDR levels should be chosen. However, no matter how smart a divide-and-test strategy may be, discretizing a continuous covariate would inevitably result in information loss and hence will be outperformed by CARS.

4.4. The "pooling within" strategy for information integration

In *Philosophy and Principles of Data Analysis* (Tukey, 1994), Tukey coined two witted terms to advocate some of his favorite information integration strategies: *borrowing strength* and *pooling within*. The idea of borrowing strength, which was investigated extensively and systematically by researchers in both simultaneous estimation and multiple testing fields, has led to a number of impactful theories and methodologies exemplified by the James-Stein's estimator (James and Stein, 1961) and local false discovery rate methodology (Efron et al., 2001). By contrast, the direction of "pooling within" has been less explored. Tukey described it, in a very different scenario from ours, as a two step strategy that involves first gathering quantitative indications from "within" different parts of the data, and then "pooling" these indications into a single overall index (Tukey, 1994, pp. 278). Our work formalizes a theoretical framework for the "pooling within" idea in the context of two-sample multiple testing: first constructing multiple indications from within the data (i.e. independent and comparable pairs), and second deriving an overall index (i.e. the oracle statistic) that optimally combines the evidences exhibited from both statistics.

Our work differs in several ways from existing works on multiple testing with covariate (Ferkingstad et al., 2008; Zablocki et al., 2014; Scott et al., 2015). First, the covariate in other works is collected *externally* from other data sources, whereas the auxiliary information in our work is gathered *internally* within the primary data set. Second, in other

works it has been assumed that the null density would not be affected by the external covariate. However, the assumption should be scrutinized in practice as it may not always hold. Under our testing framework, the requirement of a fixed null density is formalized as the conditional independence between the primary and auxiliary statistics. The conditional independence is proposed as a principle for information extraction, and is automatically fulfilled by our approach to constructing the auxiliary sequence.

CARS makes several new methodological and theoretical contributions. First, under a decision-theoretic framework, the oracle CARS procedure is shown to be optimal for information pooling. Second, existing methodologies on testing with covariate are mostly developed under the Bayesian computational framework and lack theoretical justifications. By contrast, our data-driven CARS procedure is a non-parametric method and enjoys nice asymptotic properties. Such FDR theories, as far as we know, are new in the literature. Third, the screening approach employed by CARS reveals interesting connections between sparsity estimation and multiple testing with covariate, which is elaborated next.

4.5. Capturing sparsity information via screening

A celebrated finding in the FDR literature is that incorporating the estimated proportion of non-nulls $[\pi_1 = P(\theta_i = 1)]$ can improve the power (Benjamini and Hochberg, 2000; Storey, 2002; Genovese and Wasserman, 2002). In light of S_i, the proportion becomes heterogeneous; hence it is desirable to utilize the *conditional proportions* $\pi_1(s) = P(\theta_i = 1|S_i = s)$ to improve the power (Zablocki et al., 2014; Scott et al., 2015; Li and Barber, 2016). In a similar vein, earlier works on multiple testing with groups (or discrete S_i) reveal that varied sparsity levels across groups can be exploited to construct more powerful methods (Ferkingstad et al., 2008; Cai and Sun, 2009; Hu et al., 2010). Estimating $\pi_1(s)$ with a continuous covariate is a challenging problem. Most existing works (Zablocki et al., 2014; Scott et al., 2015) employ Bayesian computational algorithms that do not provide theoretical guarantees. Notable progress has been made by Boca and Leek (2017). However, their theory requires a correct specification of the underlying regression model, which cannot be checked in practice. Next we discuss how the screening idea in Sections 3.3 and 3.4 can be extended to derive a simple and elegant non-parametric estimator of $\pi_1(s)$.

Fig. 1 gives a graphical illustration of the proposed estimator. We generate $m = 10^5$ tests with $\bar{X}_i \sim N(0,1)$ and $\bar{Y}_i \sim 0.8N(0,1) + 0.2N(2,1)$; hence $T_i = (1/\sqrt{2})(\bar{X}_i - \bar{Y}_i)$ and $S_i = (1/\sqrt{2})(\bar{X}_i + \bar{Y}_i)$. Suppose we are interested in counting how many S_i would fall into the interval $[t_2 - h, t_2 + h]$ with $t_2 = 2$ and $h = 0.3$. The counts are represented by vertical bars on Panel (a) for each p-value interval. As shown in Proposition 1, T_i and S_i are independent under the null [cf. Equation (2.6)]. Therefore we can see that the counts of S_i are roughly uniformly distributed when the p-value of T_i is large. Expanding the interval $[t_2 - h, t_2 + h]$ to the entire real line (which actually corresponds to discarding the information in S_i), we obtain the histogram of all p-values [Panel (b)].

We start with a description of a classical estimator [Schweder and Spjøtvoll (1982); Storey (2002)] for π_1; see Langaas et al. (2005) for a detailed discussion of various extensions. Let $Q(\tau) = \#\{P_i > \tau\}$, then by Fig. 1. (b), the expected counts covered by light grey bars to the right of the threshold τ can be approximated as $\mathbb{E}\{Q(\tau)\} = m(1 - \pi_1)(1 - \tau)$. Setting the expected and actual counts equal, we obtain $\hat{\pi}_1^\tau = 1 - Q(\tau)/\{m(1 - \tau)\}$.

Next we consider the conditional proportion $\pi_1(s) = P(\theta_i = 1|S_i = s)$. Assume that $\pi_1(s)$ and $f(s)$, the density of S_i, are constants in a small neighborhood $[s - h/2, s + h/2]$. Then the expected counts of the p-values from the null distribution in the interval $[s - h/2, s + h/2]$ can be approximated by $Q^\tau(s,h) \approx \{1 - \pi_1(s)\}f(s)h$. The other way of

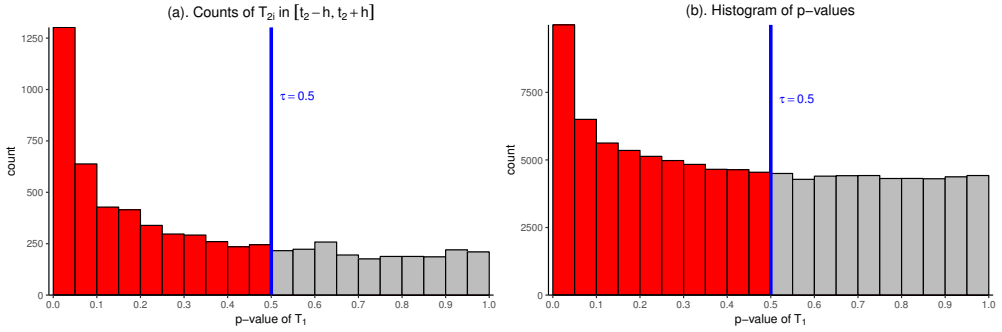

Fig. 1. A graphical illustration of the smoothing estimator (4.3). (a). The counts of S_i are uniformly distributed on the right. The bias decreases and the variability increases when τ increases. (b). The histogram of all p-values. Similarly, the p-values are approximately uniformly distributed on the right.

counting can be done using (3.7) in Section 3.2. In obtaining (3.7), we exploit the fact that the counts S_i are roughly uniformly distributed to the right of the threshold τ. Taking the limit, we obtain $\pi_1(s) = 1 - \{f(s)\}^{-1} \lim_{h\to 0} Q^\tau(s,h)/h = 1 - q^\tau(s)/f(s)$. Finally, utilizing the screening approach (3.11), we propose the following nonparametric smoothing estimator

$$\hat{\pi}_1^\tau(s) = 1 - \frac{\sum_{i\in\mathcal{T}(\tau)} K_h(s - s_i)}{(1-\tau)\sum_{i=1}^m K_h(s - s_i)}. \tag{4.3}$$

Choosing tuning parameter τ is an important issue but has gone beyond the scope of the current work; see Storey (2002); Langaas et al. (2005) for further discussions.

4.6. Heterogeneity, correlation and empirical null

Conventional FDR analyses treat all hypotheses exchangeably. However, the hypotheses become "unequal" in light of S_i, and it is desirable to incorporate, for example, the varied conditional proportions in a testing procedure to improve the efficiency. This section further discusses the case where the heterogeneity is reflected by disparate null densities.

A key principle in our construction is that the primary and auxiliary statistics are conditionally independent under the null. However, in many applications where the auxiliary information is collected from external data, S_i may be correlated with T_i. Then the FDR procedure may become invalid if S_i is incorporated inappropriately. For example, if the grouping variable S_i is correlated with the p-value, then applying Benjamini-Hochberg's (BH) procedure to hypotheses in separate groups would be problematic because the null distributions of the p-values in some groups may no longer be uniform. A partial solution to resolve the issue is to estimate the *empirical null* distributions (Efron, 2004; Jin and Cai, 2007) for different groups, instead of using the theoretical null directly. The theory and methodology in Efron (2008) and Cai and Sun (2009), which allow the use of varied empirical nulls across different groups, can be applied to control the FDR. However, as we previously mentioned, discretizing a continuous S_i fails to fully utilize the auxiliary information. The estimation of the empirical null with a continuous S_i is an interesting problem for future research. The nonparametric smoothing idea in (4.3) might be helpful but additional difficulties may arise. The limitations of the current methodology and open questions will be discussed in Section 6.

5. Numerical Results

This section investigates the numerical performance of CARS using both simulated and real data. We compare the oracle and data-driven CARS procedures, respectively denoted by OR and DD, with existing methods, including the Benjamini-Hochberg procedure (BH, Benjamin and Hochberg, 1995), the adaptive z-value procedure (AZ, Sun and Cai 2007), and the uncorrelated screening procedure (US, Liu 2014). We first describe the implementation of CARS in Section 5.1. Sections 5.2 and 5.3 respectively consider (i) the case with known and unequal variances and (ii) the case with estimated variances and non-Gaussian errors. Section 5.4 provides numerical evidence to show the merit of CARS when the two means have opposite signs. An application to Supernova Detection is discussed in Section 5.5. Additional numerical results, including the non-informative case (Section B.2), completely informative case (Section B.3) and dependent tests (Section B.6), are provided in Section B of the Supplementary Material.

5.1. The implementation and R-package `CARS`

The R-package `CARS` has been developed to implement the proposed method. This section describes implementation details and some practical guidelines.

The bivariate density estimator $\hat{f}(t_1, t_2)$ can become unstable in very sparse settings, which may lead to slightly elevated FDR level (cf. top left panel in Figure 2). To increase the stability of CARS in the extremely sparse setting where the non-null proportion is vanishingly small, the `CARS` package has included a "sparse" option, which implements a conservative but more stable density estimator

$$\hat{f}^{\upsilon}(t_1, t_2) = (1 - \hat{\pi}_2) f_{10}(t_1) f_{20}(t_2) + \mathbb{G}_L(\upsilon) \left\{ 1 - \widehat{\mathrm{FDR}_2}(\upsilon) \right\} \hat{f}(t_1, t_2 | \widehat{\mathrm{Lfdr}}_2 < \upsilon). \tag{5.1}$$

Here $\mathbb{G}_L(\upsilon) = m^{-1} \sum_{i=1}^{m} \mathbb{I}\{\mathrm{Lfdr}_2(t_{2i} < \upsilon)\}$ is an empirical CDF, $\widehat{\mathrm{FDR}_2}(\upsilon)$ is the estimated FDR level, and υ is the screening level. The first term on the right hand side of (5.1) is based on known densities, which stabilizes the bivariate density estimate in regions with few observations. Our numerical studies in Section B.3 show that the choice of υ has little impact in the range of 0.1 to 0.3; the default choice in the `CARS` package is $\upsilon = 0.1$. To estimate the bivariate density $f(t_1, t_2 | \mathrm{Lfdr}_2 < \upsilon)$, we apply the R package `ash` to the sample $\mathcal{T} = \{t_{2i} : \widehat{\mathrm{Lfdr}}(t_{2i}) < \upsilon\}$. We explain in Section A.11 that the screening step would underestimate $f(t_1, t_2)$ and hence lead to a *conservative* FDR control. The performance of the modified density estimator is investigated in Section 5.3 for extremely sparse case (including $k = 0$). For the global null case, one may consider a hybrid strategy as done in Durand (2017) that include a global testing step (Donoho and Jin, 2004; Cai and Wu, 2014) to test the hypothesis that all effects are zero; run CARS if the global null is rejected.

In estimating (3.11), the `CARS` package uses the Lfdr as the screening statistic (as opposed to the p-values). A correction factor similar to $1 - \tau$ is needed and can be easily computed from data. Related formulae and computational details are described in Section A.12 in the Supplement. Although the p-values lead to simpler and more intuitive descriptions of the methodology, we found that screening via Lfdr leads to improved stability in tuning for finite samples. The intuition is that the Lfdr, which contains information about the sparsity and non-null density, provides a testing rule that is more adaptive to the data. Section B.3 in the Supplement investigates the choice of the tuning parameter τ when the Lfdr is used. In general as τ increases, the FDR is closer to the nominal level but the stability decreases. The default choice in our package is $\tau = 0.9$, which has been used in all our simulations.

The R package `np` is used to choose the bandwidths h_1 and h_2 in equation (3.10). We have adopted two strategies to improve the performance. First, the bandwidth h_1 and h_2 are chosen based on the *normal reference rule* restricted to the samples with $\text{Lfdr}_1 < 0.5$. The restriction leads to more informative bandwidths as this subset is the more relevant part of the sample for the multiple testing problem. Second, the same h_2 has been used in (3.11) to obtain the numerator of (3.12). This strategy is helpful to increase the stability of the estimator [cf. Remark 4]. Finally we note that these strategies are only practical guidelines in finite samples; the asymptotic theories are not affected.

5.2. Simulation I: known variances

Consider model (2.1). Denote $\boldsymbol{\mu}_{x,i_1:i_2} = (\mu_{x,i_1}, \cdots, \mu_{x,i_2})$ and $\boldsymbol{\mu}_{y,i_1:i_2} = (\mu_{y,i_1}, \cdots, \mu_{y,i_2})$ the vectors of consecutive observations from i_1 to i_2. The two original samples are denoted $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n_x}\}$ and $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{n_y}\}$, with corresponding means $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$. Let $\sigma_{xi} = 1$, $\sigma_{yi} = 2$, $n_x = 50$ and $n_y = 60$. Our simulations use $m = 5000$ and FDR level $\alpha = 0.05$. We consider the following 3 settings, where different methods are applied to simulated data and the results are averaged over 500 replications. The FDR and average power (proportion of differential effects that are correctly identified) are plotted as functions of varied parameter values and displayed in Figure 2.

Setting 1: we set $\boldsymbol{\mu}_{x,1:k} = 5/\sqrt{30}$, $\boldsymbol{\mu}_{x,(k+1):(2k)} = 4/\sqrt{30}$, $\boldsymbol{\mu}_{x,(2k+1):m} = 0$, $\boldsymbol{\mu}_{y,1:k} = 2/\sqrt{30}$, $\boldsymbol{\mu}_{y,(k+1):(2k)} = 4/\sqrt{30}$ and $\boldsymbol{\mu}_{y,(2k+1):m} = 0$. Here k denotes the sparsity level: the proportion of locations with differential effects is k/m, and the proportion of nonzero locations is $(2k)/m$. We vary k from 100 to 1000 to investigate the effect of sparsity.

Setting 2: we use k_1 and k_2 to denote the number of nonzero locations and the number of locations with differential effects, respectively. Let $\boldsymbol{\mu}_{x,1:k_2} = 5/\sqrt{30}$, $\boldsymbol{\mu}_{x,(k_2+1):k_1} = 4/\sqrt{30}$, $\boldsymbol{\mu}_{x,(k_1+1):m} = 0$, $\boldsymbol{\mu}_{y,1:k_2} = 2/\sqrt{30}$, $\boldsymbol{\mu}_{y,(k_2+1):k_1} = 4/\sqrt{30}$ and $\boldsymbol{\mu}_{y,(k_1+1):m} = 0$. We fix $k_1 = 2000$ and vary k_2 from 100 to 1000. This setting investigates how the informativeness of the auxiliary covariate would affect the performance of different methods. Note that as k_2 increases, the conditional probability $\pi_{1|1} = \mathbb{P}(\theta_{1i} = 1 | \theta_{2i} = 1)$ also increases, and the auxiliary covariate becomes more informative.

Setting 3: We fix $k = 750$ and set $\boldsymbol{\mu}_{x,1:k} = \mu_0/\sqrt{30}$, $\boldsymbol{\mu}_{x,(k+1):(2k)} = 3/\sqrt{30}$, $\boldsymbol{\mu}_{x,(2k+1):m} = 0$, $\boldsymbol{\mu}_{y,1:k} = 2/\sqrt{30}$, and $\boldsymbol{\mu}_{y,(k+1):(2k)} = 3/\sqrt{30}$, $\boldsymbol{\mu}_{y,(2k+1):m} = 0$. To investigate the impact of the effect sizes, we vary μ_0 from 3.5 to 5.

We can see that the CARS procedure is more powerful than conventional univariate methods such BH and AZ, and is superior than US which only partially utilizes the auxiliary information. A more detailed description of simulation results is given below. (a). All methods control the FDR at the nominal level 0.05 approximately. BH is slightly conservative and US is very conservative. (b). Univariate methods (BH and AZ) are improved by bivariate methods (US and CARS) in most settings. This shows that exploiting the auxiliary information is helpful. (c). US is uniformly dominated by CARS. This is expected because US only models T_{2i} as a binary variable whereas CARS fully utilizes the information in T_{2i}. (d). DD has similar performance as OR in most settings. However, DD can be conservative in FDR control in some settings and hence has less power compared to OR (cf. Setting 3, bottom row of Figure 1). This has been predicted by our theory (Proposition 5). (e). Setting 1 shows that the gain in efficiency (of bivariate methods over univariate methods) decreases as k (or the sparsity level) increases. (f) Setting 2 shows that

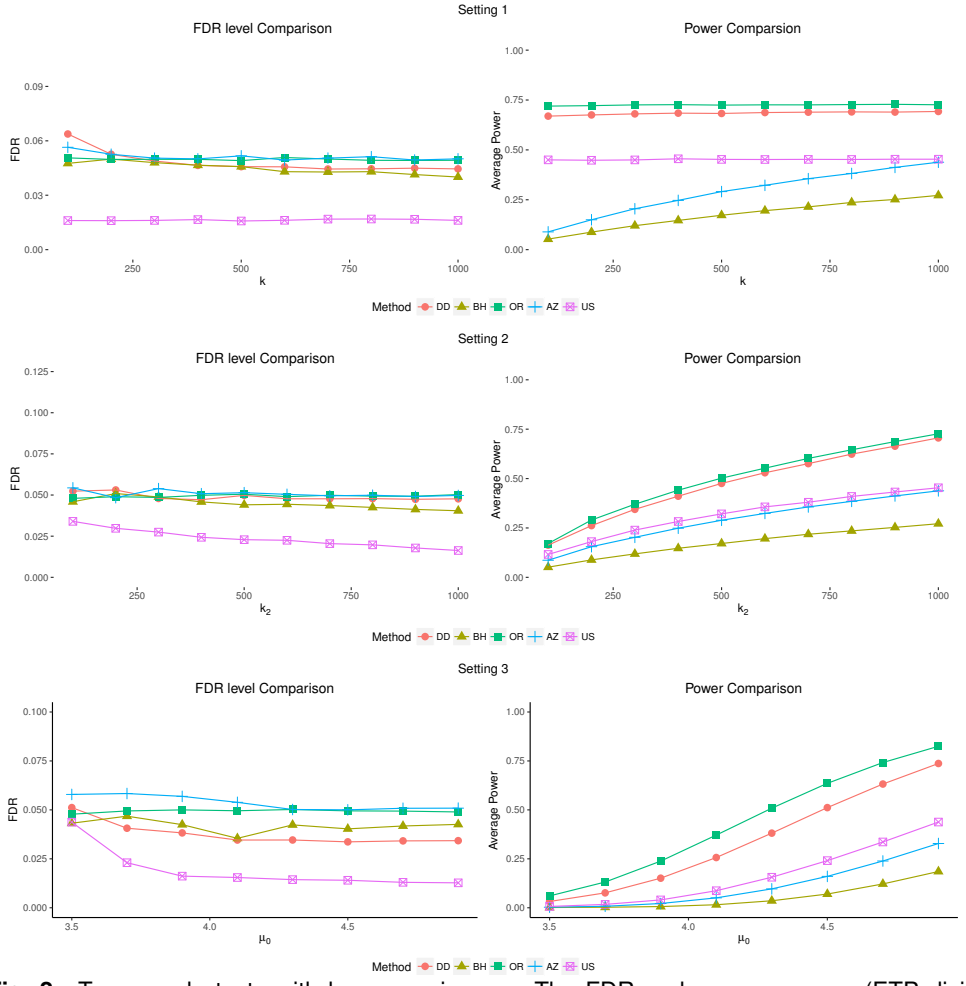

Fig. 2. Two-sample tests with known variances. The FDR and average power (ETP divided by the number of non-nulls) are plotted against varied non-null proportions (top row) and conditional proportions (middle row) and effect sizes (bottom row). DD (○), BH (△), OR (■), AZ (+) and US (□).

the efficiency gain of CARS increases when k_2 increases. Note k_2 is proportional to $\pi_{1|1}$ (the informativeness of the auxiliary covariate). (g). Setting 3 shows that the efficiency gain of CARS increases as the signal strength decreases (note that a smaller μ_0 corresponds a larger difference in effect sizes).

5.3. Simulation II: estimated variances and non-Gaussian errors

We consider similar simulation settings as the previous section with three modifications. First, we substitute the estimated variances in place of known variances. Second, to investigate the performance of our method with non-Gaussian errors, we modify Setting 3 slightly by generating ϵ_{xij} and ϵ_{yik} from t-distribution with $\mathrm{df} = 4$ and $\mathrm{df} = 5$, respectively. Finally, we vary k from 1 to 200 to to investigate the performance of CARS under different sparsity regimes. The modified density estimator $\hat{f}^v(t_1, t_2)$, defined in (5.1), has been used in all settings. The simulation results are summarized in Figure 3.

The patterns are very similar to those in Simulation I; our conclusions on the comparison of different methods remain the same. We mention the following points. (a). Settings 1-2 show that the CARS works well with estimated variances. (b). Setting 3 shows that CARS is robust to the Gaussian assumption. (c). Under the very sparse setting, the modified CARS procedure is conservative for FDR control but still outperforms competitive methods.

5.4. Simulation III: means with opposite signs

Our testing framework utilizes T_{2i} as auxiliary statistics to assist inference. It is possible that T_{2i} may be non-informative but *this auxiliary information cannot hurt*. This important point has been explained by Remark 3; see also Section A.9 in the Supplementary Material. Here we provide numerical evidence to support the claim.

Consider a setting in which the two means have opposite signs. We shall see that CARS outperforms univariate methods as long as the two means do not cancel out with each other precisely. This confirms our claim that CARS, which benefits from an enhanced signal to noise ratio by exploiting the auxiliary data, always dominates the univariate methods.

Let $\epsilon_{xij} \sim N(0,1)$ and $\epsilon_{yik} \sim N(0,1)$ be iid noise. Set $n_x = n_y = 50$. In our simulation, the number of tests is $m = 10,000$. The two mean vectors are given below:

$$\boldsymbol{\mu}_{x,1:500} = 3/\sqrt{50},\ \boldsymbol{\mu}_{x,501:1000} = 4/\sqrt{50},\ \boldsymbol{\mu}_{x,1001:m} = 0$$
$$\boldsymbol{\mu}_{y,1:500} = (3c)/\sqrt{50},\ \boldsymbol{\mu}_{y,501:1000} = 4/\sqrt{50},\ \boldsymbol{\mu}_{y,1001:m} = 0.$$

We vary c from -1 to 0, where $c = -1$ corresponds to the least favorable situation where the two means cancel out precisely. We apply the BH procedure (BH), oracle CARS procedure (OR) and data-driven CARS procedure (DD) to the simulated data sets. The FDR and power are obtained based on 200 replications. The simulation results are summarized by Figure 4. We can see that when $c = -1$, all methods have similar power. As c increases to zero, the power gain of CARS become more pronounced.

5.5. Application in Supernova Detection

This section applies CARS for analysis of time course satellite imaging data. Figure 5 shows the time course g-band images of galaxy M101 collected by the Palomar Transient Factory survey (Law et al., 2009). The images indicate the appearance of SN 2011fe, one of the brightest supernovas known up to date (Nugent et al., 2011). A major goal of our analysis is to detect the discrepancies between images taken over time so that we can narrow down

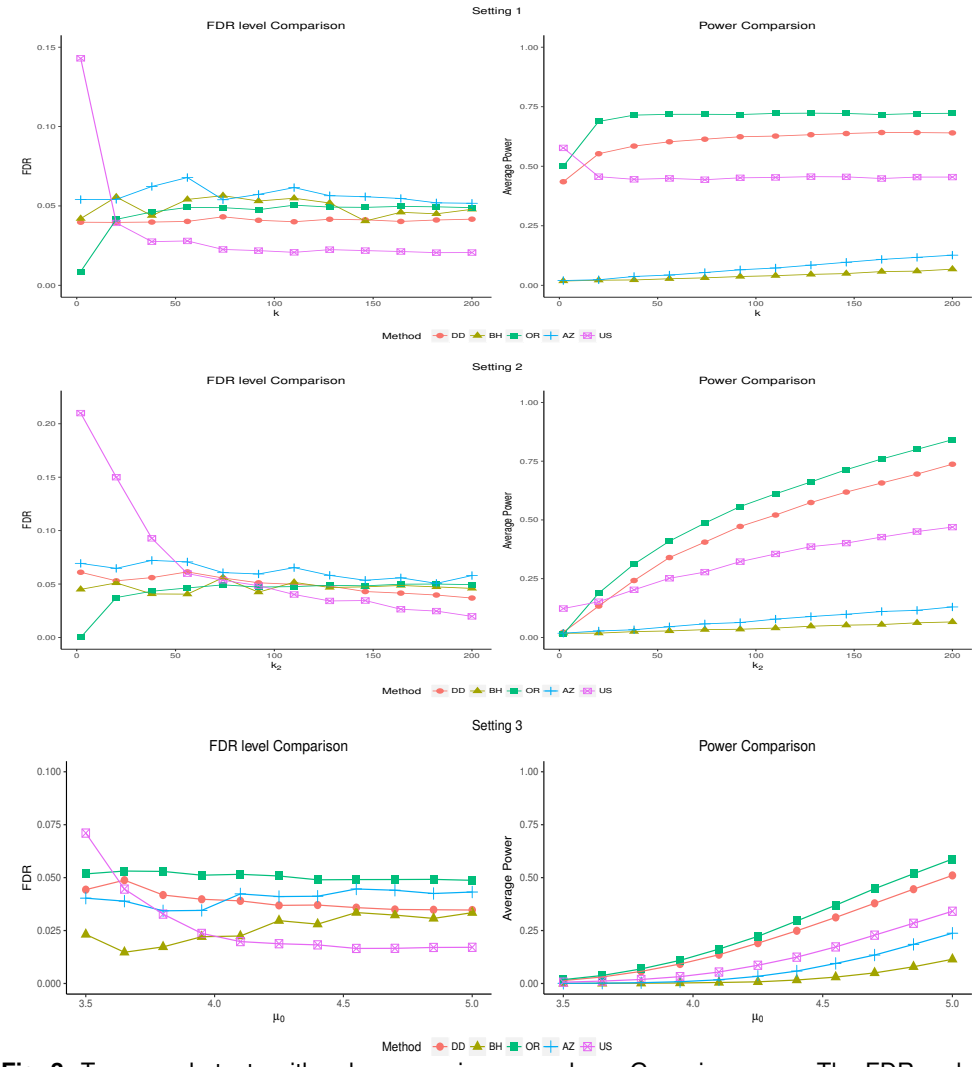

Fig. 3. Two-sample tests with unknown variances and non-Gaussian errors. The FDR and average power are plotted against varied non-null proportions (top row), conditional proportions (bottom row). The displayed procedures are DD (○), BH (△), OR (■), AZ (+) and US (□).

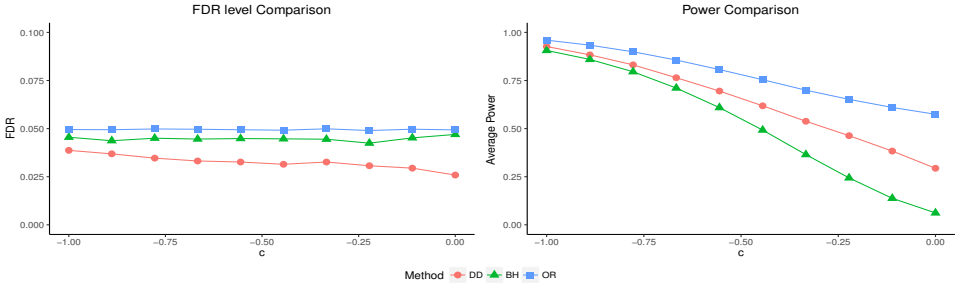

Fig. 4. Comparison of BH, DD and OR when the nonzero means have opposite signs.

the potential locations for Supernova explosions. More accurate measurements and further investigations will then be carried out on the narrowed subset of potential locations.

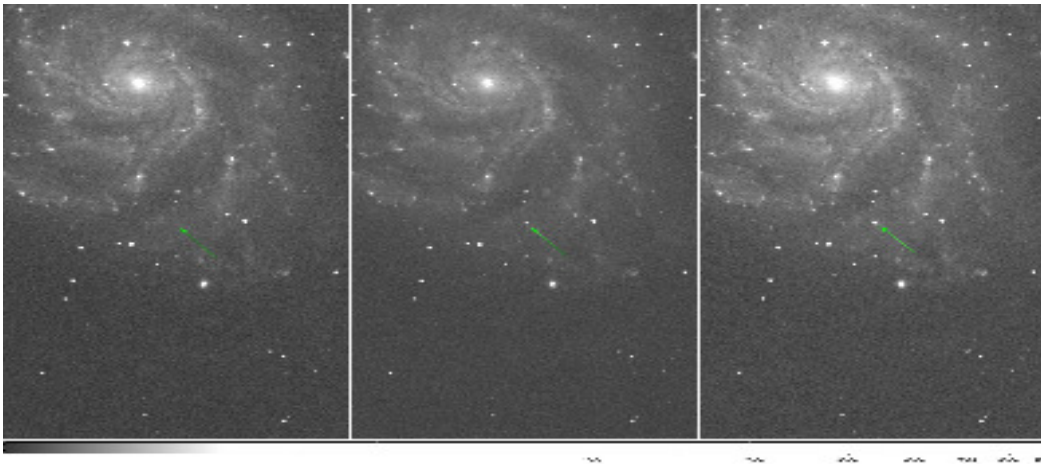

Fig. 5. From left to right, images are taken respectively on August 23, 24, and 25, 2011. The arrows clearly indicate the explosion of SN 2011fe.

The satellite data are recorded and converted into gray-scale images of size 516×831 (or $m = 428,796$ pixels). Each pixel corresponds to a value ranging from 0 to 1 that indicates the intensity level of the influx from stars. We use image 1 as the baseline. Its grey-scale pixel values are subtracted from those in images 2 and 3. These differences are then vectorized as $\boldsymbol{x} = (x_1, \cdots, x_m)$ and $\boldsymbol{y} = (y_1, \cdots, y_m)$ (respectively representing "image 2 − image 1" and "image 3 − image 1").

We plot the histograms and find that the null distributions of $\boldsymbol{x}$ and $\boldsymbol{y}$ are different. This can be explained by the lapse in times at which these images are taken (the brightness of these g-band images changes gradually over time). The supernova data have significant amount of background noise in each image, and the average magnitudes of the background noise vary a lot from image to image. To remove the image-specific background noise, we first estimate the empirical null distributions based on the center part of the histograms. The variances of observations are assumed to be homoscedastic and are estimated using all pixels. For $\boldsymbol{x}$ and $\boldsymbol{y}$, we have $N(0.0028, 0.0023)$ and $N(0.044, 0.0027)$, respectively. We then standardize the observations as $\boldsymbol{x}^{\text{st}}$ and $\boldsymbol{y}^{\text{st}}$, which are used in our analysis. We do not take the difference $\boldsymbol{x} - \boldsymbol{y}$ directly as it would create many false signals due to the varied magnitudes of the background noise. The standardized measurements $\boldsymbol{x}^{\text{st}}$ and $\boldsymbol{y}^{\text{st}}$ from the two images seem to be comparable as most of pairs $(x_i^{\text{st}}, y_i^{\text{st}})$ have similar values.

Next we carry out $m = 428,796$ two-sample tests with known variances. For standardized observations $\boldsymbol{x}^{\text{st}}$ and $\boldsymbol{y}^{\text{st}}$ the variances are known to be 1. Then $t_{1i} = \frac{x_i^{st} - y_i^{st}}{\sqrt{2}}$ and $t_{2i} = \frac{x_i^{st} + y_i^{st}}{\sqrt{2}}$. We apply BH, AZ, US and CARS procedures at FDR levels 0.01%, 1% and 5%. Figure 12 in the Supplementary Material shows the rejected pixels in the 516×831 layout for each method under different FDR levels. The estimated sparsity levels for $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$ are respectively 1.47% and 49.5%. The corresponding estimated support size at $\tau = 0.5$ is 6285. We report the thresholds of different testing procedures in Table 1.

We can see that more information can be harvested from the data by using auxiliary information. In particular, at FDR level 0.01%, the supernova is missed by BH and AZ but

Table 1. Thresholds and Total Number of Rejections (in parentheses) of Different Testing Procedures

FDR level	BH procedure	Adaptive Z procedure	US (two thresholds)	CARS
10^{-4}	3.66×10^{-10} (4)	2.75×10^{-4} (5)	3.24, 5.46 (22)	4.38×10^{-4} (35)
0.01	9.91×10^{-7} (22)	3.37×10^{-2} (24)	2.51, 4.87 (38)	9.25×10^{-2} (58)
0.05	7.38×10^{-6} (64)	0.39 (69)	1.92, 4.42 (80)	0.26 (109)

detected by CARS. To further quantify our procedure's superiority, we count the total number of rejections in Table 1 (numbers in parentheses). We can see that CARS consistently detects more signals from the satellite images than competing methods.

6. Discussion

Covariate-assisted multiple testing can be viewed as a special case of a much broader class of problems under the theoretical framework of *integrative simultaneous inference*, which covers a range of topics including multiple testing with external or prior domain knowledge (Benjamini and Hochberg, 1997; Basu et al., 2017), partial conjunction tests and set-wise inference (Benjamini and Heller, 2008; Sun and Wei, 2011; Du and Zhang, 2014), and replicability analysis (Heller et al., 2014; Heller and Yekutieli, 2014). A coherent theme in these problems is to combine the information from multiple sources to make more informative decisions. Tukey's "pooling within" strategy provides a promising approach in such scenarios where quantitative indications might be hidden in various parts of massive data sets.

The current formulations and methodologies in integrative data analysis differ substantially. A general theory and methodology is yet to be developed for handling various types of problems in a unified framework. For instance, in weighted FDR analysis (Benjamini and Hochberg, 1997), the external domain knowledge is incorporated as the weights in modified FDR and power definitions to reflect the varied gains and losses in decisions. By contrast, covariate-assisted multiple testing still utilizes unweighted FDR and power definitions. The connection of CARS to theories on optimal weights is still an open issue (Roeder and Wasserman, 2009; Roquain and Van De Wiel, 2009). Moreover, in partial conjunction tests and replicability analysis, the summary statistics from different studies are of equal importance, which marks a key difference from covariate assisted inference where some statistics are primary while others are secondary.

We conclude our discussion with a few more open issues.

- *Are there better ways to construct the auxiliary sequence?* Our theory only shows that CARS is optimal *when the pairs are given*. How to construct an optimal pair from data is still an open question. For instance, in situations where two means have opposite signs, the sum of absolute values may better capture the sparsity information but would give rise to a correlated pair, which cannot be handled by the current testing framework.

- *How to deal with multiple testing dependency?* The CARS method cannot be applied to dependent tests as it assumes that T_i are independent. Our simulation studies show that CARS controls the FDR under weak dependence. However, the result is based on very limited empirical studies, which lack theoretical supports. An important direction is to develop new theory and methodology for the dependent case.

- *How to generalize the idea to settings where the null distribution is unknown?* This important situation may arise from the classical two-sample tests where the null dis-

tribution is calibrated with permutations. We conjecture that the CARS procedure, which requires explicit form of the null density, may be tailored by using a different, probably more ad hoc, approximation. For example, informative weights may be derived from the auxiliary data and incorporated into the permutation based p-values via some grouping and weighting strategy.

- *How to construct the auxiliary sequence in more general settings?* This article focuses on the two-sample tests. It would be of interest to extend the methodology to simultaneous ANOVA tests. Moreover, CARS provides a useful strategy for extracting the sparsity structure from data. There are other important structures in the data such as heteroscedasticity, hierarchy and dependency, which may also be captured by an auxiliary sequence. It remains an open question on how to extract and incorporate such structural information effectively to improve the power of a testing procedure.

- *How to summarize the auxiliary information in high-dimensional settings?* The proposed CARS methodology requires the joint modeling of the primary and auxiliary statistics, which cannot handle many covariate sequences because the joint density estimator would greatly suffer from high-dimensionality. A fundamental issue is to develop new principles for information pooling, or optimal dimension reduction, in multiple testing with high-dimensional covariates.

- *How to make inference with multiple sequences?* In partial conjunction tests and replicability analysis, an important feature is that the means (of summary statistics) from separate studies are individually sparse. We expect that similar strategies for extracting and sharing sparsity information among multiple sequences would improve the accuracy of simultaneous inference. However, as opposed to covariate-assisted inference where there is a sequence of *primary statistic*, in partial conjunction tests and replicability analysis all sequences are of equal importance, which poses new challenges for problem formulation and methodological development.

7. Proofs of Main Theorems

This section proves the main theorems. The proofs of other propositions are provided in the Supplementary Material.

7.1. Proof of Theorem 1

PROOF. We first show that the two expressions of T^i_{OR} in (3.4) are equivalent. Recall $q^*(t_2) = (1-\pi_1)f(t_2 \mid \theta_{1i}=0)$. Applying Bayes theorem and using the conditional independence between T_{1i} and T_{2i} under the null $\theta_{1i}=0$ (Proposition 1), we obtain

$$T^i_{OR}(t_1,t_2) = \frac{\mathbb{P}(\theta_{1i}=0)f(t_1,t_2|\theta_{1i}=0)}{f(t_1,t_2)} = \frac{q^*(t_2)f_{10}(t_1)}{f(t_1,t_2)}.$$

Part (a). Let $Q_{OR}(t)=\alpha_t$. We first show that $\alpha_t < t$. According to the mFDR definition,

$$\mathbb{E}_{(\boldsymbol{T}_1,\boldsymbol{T}_2)}\left\{\sum_{i=1}^m (T^i_{OR}-\alpha_t)\mathbb{I}(T^i_{OR}<t)\right\} = 0, \tag{7.1}$$

where the subscript $(\boldsymbol{T}_1,\boldsymbol{T}_2)$ indicates that the expectation is taken over the joint distribution of $(\boldsymbol{T}_1,\boldsymbol{T}_2)$. The above equation implies that $\alpha_t < t$; otherwise all terms in the summation on its LHS would be either zero or negative.

Next we show that $Q_{OR}(t)$ is monotone in t. Let $Q_{OR}(t_j) = \alpha_j$ for $j = 1, 2$. We only need to show that if $t_1 < t_2$, then $\alpha_1 \leq \alpha_2$. We argue by contradiction. If $\alpha_1 > \alpha_2$, then

$$\begin{aligned}(T^i_{OR} - \alpha_2)\mathbb{I}(T^i_{OR} < t_2) &= (T^i_{OR} - \alpha_2)\mathbb{I}(T^i_{OR} < t_1) + (T^i_{OR} - \alpha_2)\mathbb{I}(t_1 \leq T^i_{OR} < t_2) \\ &\geq (T^i_{OR} - \alpha_1)\mathbb{I}(T^i_{OR} < t_1) + (\alpha_1 - \alpha_2)\mathbb{I}(T^i_{OR} < t_1) \\ &\quad + (T^i_{OR} - \alpha_1)\mathbb{I}(t_1 \leq T^i_{OR} < t_2).\end{aligned}$$

Next take expectations on both sides and sum over all i. We claim that

$$\mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left\{\sum_{i=1}^m (T^i_{OR} - \alpha_2)\mathbb{I}(T^i_{OR} < t_2)\right\} > 0. \tag{7.2}$$

The above inequality holds since (i) $\mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left\{\sum_{i=1}^m (T^i_{OR} - \alpha_1)\mathbb{I}(T^i_{OR} < t_1)\right\} = 0$, (ii) $\mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}$ $\left\{\sum_{i=1}^m (\alpha_1 - \alpha_2)\mathbb{I}(T^i_{OR} < t_1)\right\} > 0$, and (iii) $\mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left\{\sum_{i=1}^m (T^i_{OR} - \alpha_1)\mathbb{I}(t_1 \leq T^i_{OR} < t_2)\right\} > 0$, which are respectively due to (7.1), the assumption that $\alpha_1 > \alpha_2$ and the fact $\alpha_1 < t_1$. However, (7.2) is a contradiction to our definition of α_2, which implies that $\mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left\{\sum_{i=1}^m (T^i_{OR} - \alpha_2)\mathbb{I}(T^i_{OR} < t_2)\right\} = 0$. Hence we must have $\alpha_1 \leq \alpha_2$.

Part (b). The oracle threshold is defined as $t_{OR} = \sup_t\{t \in (0, 1) : Q_{OR}(t) \leq \alpha\}$. We want to show that at t_{OR}, the mFDR level is attained precisely. Let $\bar{\alpha} = Q_{OR}(1)$. Part (a) shows that the continuous function $Q_{OR}(t)$ is non-decreasing. Then we always have $Q_{OR}(t_{OR}) = \alpha$ if $\alpha < \bar{\alpha}$. Define $\boldsymbol{\delta}_{OR} = \{\mathbb{I}(T^i_{OR} < t_{OR}) : 1 \leq i \leq m\}$. Let $\boldsymbol{\delta}_* = (\delta^1_*, \cdots, \delta^m_*)$ be an arbitrary decision rule such that $\text{mFDR}(\boldsymbol{\delta}_*) \leq \alpha$. It follows that

$$\mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left\{\sum_{i=1}^m (T^i_{OR} - \alpha)\delta^i_{OR}\right\} = 0 \quad \text{and} \quad \mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left\{\sum_{i=1}^m (T^i_{OR} - \alpha)\delta^i_*\right\} \leq 0. \tag{7.3}$$

Combining the two results in (7.3), we conclude that

$$\mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left\{\sum_{i=1}^m (\delta^i_{OR} - \delta^i_*)(T^i_{OR} - \alpha)\right\} \geq 0. \tag{7.4}$$

Next, consider a monotonic transformation of the oracle decision rule $\delta^i_{OR} = \mathbb{I}(T^i_{OR} < t_{OR})$ via $f(x) = (x - \alpha)/(1 - x)$ (note that the derivative $\dot{f}(x) = (1 - \alpha)/(1 - x)^2 > 0$). The oracle decision rule is equivalent to $\delta^i_{OR} = I\left(\frac{T^i_{OR} - \alpha}{1 - T^i_{OR}} < \lambda_{OR}\right)$, where $\lambda_{OR} = \frac{t_{OR} - \alpha}{1 - t_{OR}}$. A useful fact is that $\alpha < t_{OR} < 1$. Hence $\lambda_{OR} > 0$.

Note that (i) $(T^i_{OR} - \alpha) - \lambda_{OR}(1 - T^i_{OR}) < 0$ if $\delta^i_{OR} > \delta^i_*$, and (ii) $(T^i_{OR} - \alpha) - \lambda_{OR}(1 - T^i_{OR}) > 0$ if $\delta^i_{OR} < \delta^i_*$. Combining (i) and (ii), we conclude that the following inequality holds for all i: $(\delta^i_{OR} - \delta^i_*)\left\{(T^i_{OR} - \alpha) - \lambda_{OR}(1 - T^i_{OR})\right\} \leq 0$. Summing over i and taking expectations, we have

$$\mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left[\sum_{i=1}^m (\delta^i_{OR} - \delta^i_*)\left\{(T^i_{OR} - \alpha) - \lambda_{OR}(1 - T^i_{OR})\right\}\right] \leq 0. \tag{7.5}$$

Combining (7.4) & (7.5), we have

$$\lambda_{OR} \cdot \mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left\{\sum_{i=1}^m (\delta^i_{OR} - \delta^i_*)(1 - T^i_{OR})\right\} \geq \mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left\{\sum_{i=1}^m (\delta^i_{OR} - \delta^i_*)(T^i_{OR} - \alpha)\right\} \geq 0.$$

Finally, note that $\lambda_{OR} > 0$ and that the ETP of a decision rule $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_m)$ is given by $\mathbb{E}_{\boldsymbol{T}_1, \boldsymbol{T}_2}\left\{\sum_{i=1}^m \delta_i(1 - T^i_{OR})\right\}$, we conclude that $\text{ETP}(\boldsymbol{\delta}_{OR}) \geq \text{ETP}(\boldsymbol{\delta}_*)$.

7.2. Proof of Theorem 2

Proof. We first provide a summary of useful notations.

- $Q^\tau(t) = m^{-1}\sum_{i=1}^m (T^{\tau,i}_{OR} - \alpha)I(T^{\tau,i}_{OR} < t)$.
- $\hat{Q}^\tau(t) = m^{-1}\sum_{i=1}^m (\hat{T}^{\tau,i}_{OR} - \alpha)I(\hat{T}^{\tau,i}_{OR} < t)$.
- $Q^\tau_\infty(t) = \mathbb{E}\{(T^\tau_{OR} - \alpha)\mathbb{I}(T^\tau_{OR} < t)\}$, where T^τ_{OR} is a generic member from $\{T^i_{OR} : 1 \leq i \leq m\}$.

Note that $Q^\tau(t)$ and $\hat{Q}^\tau(t)$ are non-decreasing and right-continuous. We can further define

$$t^\tau_\infty = \sup\{t \in (0,1) : Q^\tau_\infty(t) \le 0\}.$$

Part (a). In Proposition 5, we show that $\boldsymbol{\delta}^\tau_{OR}$ is conservative in mFDR control. To establish the desired property in mFDR control, we only need to show that $\mathrm{mFDR}(\boldsymbol{\delta}_{DD}) = \mathrm{mFDR}(\boldsymbol{\delta}^\tau_{OR}) + o(1)$.

Define a continuous version of $Q^\tau(t)$ as follows. For $T^{\tau,(k)}_{OR} < t \le T^{\tau,(k+1)}_{OR}$, let

$$Q^\tau_C(t) = \frac{t - T^{\tau,(k)}_{OR}}{T^{\tau,(k+1)}_{OR} - T^{\tau,(k)}_{OR}} Q^\tau_k + \frac{T^{\tau,(k+1)}_{OR} - t}{T^{\tau,(k+1)}_{OR} - T^{\tau,(k)}_{OR}} Q^\tau_{k+1}, \tag{7.6}$$

where $Q^\tau_k = Q^\tau\left(T^{\tau,(k)}_{OR}\right)$. It is easy to see that $Q^\tau_C(t)$ is continuous and monotone. Hence the inverse of $Q^\tau_C(t)$, denoted $Q^{\tau,-1}_C$, is well defined. Moreover, $Q^{\tau,-1}_C$ is continuous and monotone. We can similarly define a continuous version of $\hat{Q}^\tau(t)$, denoted by $\hat{Q}^\tau_C(t)$. $\hat{Q}^\tau_C(t)$ is continuous and monotone; so does its inverse $\hat{Q}^{\tau,-1}_C(\cdot)$. By construction, we have $\boldsymbol{\delta}^\tau_{OR} = \left[I\left\{T^{\tau,i}_{OR} \le Q^{\tau,-1}_C(0)\right\} : 1 \le i \le m\right]$ and $\boldsymbol{\delta}^\tau_{DD} = \left[I\left\{\hat{T}^{\tau,i}_{OR} \le \hat{Q}^{\tau,-1}_C(0)\right\} : 1 \le i \le m\right]$. We will show that

$$\text{(i) } Q^{\tau,-1}_C(0) \xrightarrow{p} t^\tau_\infty \quad \text{and} \quad \text{(ii) } \hat{Q}^{\tau,-1}_C(0) \xrightarrow{p} t^\tau_\infty. \tag{7.7}$$

To show (i), note that the continuity of $Q^{\tau,-1}_C(\cdot)$ implies that for any $\epsilon > 0$, we can find $\eta > 0$ such that $\left|Q^{\tau,-1}_C(0) - Q^{\tau,-1}_C\left\{Q^\tau_C(t^\tau_\infty)\right\}\right| < \epsilon$ if $|Q^\tau_C(t^\tau_\infty)| < \eta$. Hence

$$\mathbb{P}\left(|Q^\tau_C(t^\tau_\infty)| > \eta\right) \ge \mathbb{P}\left[\left|Q^{\tau,-1}_C(0) - Q^{\tau,-1}_C\left\{Q^\tau_C(t^\tau_\infty)\right\}\right| > \epsilon\right].$$

Next, by the WLLN $Q^\tau_C(t) \xrightarrow{p} Q^\tau_\infty(t)$. Note that $Q^\tau_\infty(t^\tau_\infty) = 0$, we have $\mathbb{P}\left(|Q^\tau(t^\tau_\infty)| > \eta\right) \to 0$. By Markov inequality, we conclude that $Q^{\tau,-1}_C(0) \xrightarrow{p} Q^{\tau,-1}_C\left\{Q^\tau_C(t^\tau_\infty)\right\} = t^\tau_\infty$.

Next we show (ii). By inspecting the proof of (i), we only need to show that $\hat{Q}^\tau_C(t) \xrightarrow{p} Q^\tau_\infty(t)$. Denote a variable without index i (e.g. $\hat{T}^\tau_{OR}$ and T^τ_{OR}) as a generic member from the sample. It follows from Condition (C2) and the continuous mapping theorem that $\hat{T}^\tau_{OR} \xrightarrow{p} T^\tau_{OR}$. Note that both T^τ_{OR} and $\hat{T}^\tau_{OR}$ are bounded above by 1. It follows that $\mathbb{E}(\hat{T}^\tau_{OR} - T^\tau_{OR})^2 \to 0$.

Let $U_i = (T^{\tau,i}_{OR} - \alpha)\mathbb{I}(T^{\tau,i}_{OR} < t)$ and $\hat{U}_i = (\hat{T}^{\tau,i}_{OR} - \alpha)\mathbb{I}(\hat{T}^{\tau,i}_{OR} < t)$. We will show that $\mathbb{E}(\hat{U}_i - U_i)^2 = o(1)$. To see this, consider the following decomposition

$$\begin{aligned}(\hat{U}_i - U_i)^2 &= (\hat{T}^\tau_{OR} - T^\tau_{OR})^2\mathbb{I}(\hat{T}^\tau_{OR} \le t, T^\tau_{OR} \le t) + (\hat{T}^\tau_{OR} - \alpha)^2\mathbb{I}(\hat{T}^\tau_{OR} \le t, T^\tau_{OR} > t) \\ &\quad + (T^\tau_{OR} - \alpha)^2\mathbb{I}(\hat{T}^\tau_{OR} > t, T^\tau_{OR} \le t) = I + II + III.\end{aligned}$$

The first term $I = o(1)$ because $\mathbb{E}(\hat{T}^\tau_{OR} - T^\tau_{OR})^2 \to 0$. Let $\eta > 0$. Note that T^τ_{OR} is continuous and that $\hat{T}^\tau_{OR} \xrightarrow{p} T^\tau_{OR}$, we have

$$\mathbb{P}(\hat{T}^\tau_{OR} \le t, T^\tau_{OR} > t) \le \mathbb{P}\left\{T^\tau_{OR} \in (t, t+\eta)\right\} + \mathbb{P}\left(|\hat{T}^\tau_{OR} - T^\tau_{OR}| > \eta\right) \to 0.$$

Since $\hat{T}^\tau_{OR}$ is bounded, we conclude that the second term $II = o(1)$. Similarly we can show that $III = o(1)$. Therefore $\mathbb{E}(\hat{U}_i - U_i)^2 = o(1)$.

Next we show that $\hat{Q}^\tau(t) \xrightarrow{p} Q^\tau_\infty(t)$. Note that $Q^\tau(t) \xrightarrow{p} Q^\tau_\infty(t)$, we only need to show that $\hat{Q}^\tau(t) \xrightarrow{p} Q^\tau(t)$. The dependence among $\hat{U}_i$ in the expression $\hat{Q}^\tau(t) = m^{-1}\sum_i \hat{U}_i$ creates some complications. The idea is to apply some standard techniques for the limit of triangular arrays that do not require independence between variables. Consider $S_n = \sum_{i=1}^m (\hat{U}_i - U_i)$. Then $\mathbb{E}(S_n) = m\{\mathbb{E}(\hat{U}_i) - \mathbb{E}(U_i)\}$. Applying standard inequalities such as Cauchy-Schwartz, we have $\mathbb{E}(\hat{U}_i - U_i)(\hat{U}_j - U_j) = o(1)$. It follows that

$$m^{-2}\mathrm{var}(S_n) \le m^{-1}\mathbb{E}(\hat{U}_i - U_i)^2 + (1 + o(1))\mathbb{E}\left\{(\hat{U}_i - U_i)(\hat{U}_j - U_j)\right\} = o(1).$$

Therefore $E\{S_n - \mathbb{E}(S_n)/n\}^2 \to 0$. Applying Chebyshev's inequality, we obtain

$$m^{-1}\{S_n - \mathbb{E}(S_n)\} = \hat{Q}^\tau(t) - Q^\tau(t) \xrightarrow{p} 0.$$

Therefore $\hat{Q}^\tau(t) \xrightarrow{p} Q^\tau_\infty(t)$. By definition, $|\hat{Q}^\tau_C(t) - \hat{Q}^\tau(t)| \le m^{-1}$. We claim that $\hat{Q}^\tau_C(t) \xrightarrow{p} Q^\tau_\infty(t)$ and (ii) follows.

According to (i) and (ii) in (7.7), $\hat{Q}^{\tau,-1}_C(0) = Q^{\tau,-1}_C(0) + o_{\mathbb{P}}(1)$. The mFDR levels of the testing procedures are:

$$\mathrm{mFDR}(\boldsymbol{\delta}^\tau_{OR}) = \frac{\mathbb{P}_{H_0}\left(T^{\tau,i}_{OR} < Q^{\tau,-1}_C(0)\right)}{\mathbb{P}\left(T^{\tau,i}_{OR} < Q^{\tau,-1}_C(0)\right)}, \quad \text{and} \quad \mathrm{mFDR}(\boldsymbol{\delta}_{dd}) = \frac{\mathbb{P}_{H_0}\left(\hat{T}^{\tau,i}_{OR} < \hat{Q}^{\tau,-1}_C(0)\right)}{\mathbb{P}\left(\hat{T}^{\tau,i}_{OR} < \hat{Q}^{\tau,-1}_C(0)\right)}.$$

The operation of our testing procedure implies that $Q^{\tau,-1}_C(0) \ge \alpha$. It follows that $\mathbb{P}\left(T^{\tau,i}_{OR} < Q^{\tau,-1}_C(0)\right)$ is bounded away from zero. We conclude that $\mathrm{mFDR}(\boldsymbol{\delta}_{dd}) = \mathrm{mFDR}(\boldsymbol{\delta}^\tau_{OR}) + o(1)$ and the result on mFDR control follows.

The result on mFDR control can be extended to FDR control. The next proposition, which is proved in the Supplement, first gives sufficient conditions under which the mFDR and FDR definitions are asymptotically equivalent, and then verifies that these conditions are fulfilled by the CARS procedure. It follows from the proposition that CARS controls the FDR at level $\alpha + o(1)$.

PROPOSITION 7. *(a) Consider a general decision rule $\boldsymbol{\delta}$. Let $\mathcal{Y} = m^{-1}\sum_{i=1}^m \delta_i$. Then $mFDR(\boldsymbol{\delta}) = FDR(\boldsymbol{\delta}) + o(1)$ if (i) $\mathbb{E}(\mathcal{Y}) \ge \underline{\eta}$ for some $\underline{\eta} > 0$, and (ii) $\mathsf{Var}(\mathcal{Y}) = o(1)$.*

(b) Conditions (i) and (ii) are fulfilled by the CARS procedure $\boldsymbol{\delta}^\tau_{dd}$.

Part (b). The CARS procedure utilizes $\hat{q}^*$, and the corresponding test statistic is $\hat{T}^{*,i}_{OR}$. It follows from Conditions (C1′) and (C2), and the continuous mapping theorem that $\hat{T}^*_{OR} \xrightarrow{p} T_{OR}$. Denote $Q_{OR}(t)$ the mFDR capacity function and t_{OR} the oracle threshold. Then

$$Q_{OR}(t) = \mathbb{E}\{(T_{OR} - \alpha)\mathbb{I}(T_{OR} < t)\}, \quad t_{OR} = \sup\{t \in (0,1) : Q_{OR}(t) \le 0\}.$$

Define $\hat{Q}^*(t) = m^{-1}\sum_{i=1}^m (\hat{T}^{*,i}_{OR} - \alpha)\mathbb{I}(\hat{T}^{*,i}_{OR} < t)$. Similar to (7.6), we define a continuous version of $\hat{Q}^*(t)$ and denote it by $\hat{Q}^*_C(t)$. It can be shown that $\hat{Q}^*_C(t)$ is continuous and monotone; so does its inverse $\hat{Q}^{*,-1}_C(t)$. The CARS procedure is given by $\boldsymbol{\delta}^*_{DD} = \left[\mathbb{I}\left\{\hat{T}^{*,i}_{OR} \le \hat{Q}^{*,-1}_C(0)\right\} : 1 \le i \le m\right]$. Following the steps in Part (a) we can show that

$$\hat{Q}^*_C(t) \xrightarrow{p} Q_{OR}(t), \quad \hat{Q}^{*,-1}_C(t) \xrightarrow{p} t_{OR}. \tag{7.8}$$

The operation of CARS implies that $Q^{*,-1}_C(0) \ge \alpha$ (thus the denominator of the mFDR is bounded away from zero). Note that $\mathrm{mFDR}(\boldsymbol{\delta}_{OR}) = \alpha$, we have $\mathrm{mFDR}(\boldsymbol{\delta}^*_{DD}) = \alpha + o(1)$. Next, we consider the ETP. It follows from $\hat{T}^*_{OR} \xrightarrow{p} T_{OR}$ and (7.8) that

$$\frac{\mathrm{ETP}(\boldsymbol{\delta}^*_{DD})}{\mathrm{ETP}(\boldsymbol{\delta}_{OR})} = \frac{\mathbb{P}_{H_1}\left\{\hat{T}^*_{OR} < \hat{Q}^{*,-1}_C(0)\right\}}{\mathbb{P}_{H_1}\left(T_{OR} < t_{OR}\right)} = 1 + o(1).$$

□

Acknowledgments

We thank the Editor, referees and RSC for the thorough and useful comments which have greatly helped to improve the presentation of the paper. In particular, we are grateful to several excellent suggestions from the referees that have inspired our discussions on the conditional independence assumption, Tukey's procedures for combination, Brown's ancilarity paradox and the robustness of CARS when pooling non-informative auxiliary data. W. Sun would like to thank Ms. Pallavi Basu from Tel Aviv University for helpful suggestions on theory. Tony Cai was supported in part by NSF Grant DMS-1712735 and NIH Grant R01 CA127334. W. Sun was supported in part by NSF Grants DMS-CAREER-1255406 and DMS-1712983.

References

Barber, RF. and A. Ramdas (2017+). The p-filter: multi-layer fdr control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) to appear.*

Basu, P., T. T. Cai, K. Das, and W. Sun (2017+). Weighted false discovery control in large-scale multiple testing. *Journal of the American Statistical Association. To appear.*

Benjamini, Y. and R. Heller (2008). Screening for partial conjunction hypotheses. *Biometrics* 64: 1215–1222.

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

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

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

Boca, S. M. and J. T. Leek (2017). A regression framework for the proportion of true null hypotheses. Preprint. *bioRxiv*: 035675.

Bourgon, R., R. Gentleman, and W. Huber (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences 107*(21), 9546–9551.

Brown, L. D. (1990). An ancillarity paradox which appears in multiple linear regression. *The Annals of Statistics 18*(2), 471–493.

Cai, T. T. and J. Jin (2010). Optimal rates of convergence for estimating the null density and proportion of non-null effects in large-scale multiple testing. *Ann. Statist. 38*, 100 – 145.

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

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

Calvano, S. E., W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, et al. (2005). A network-based analysis of systemic inflammation in humans. *Nature 437*(7061), 1032–1037.

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

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

Du, L. and C. Zhang (2014). Single-index modulated multiple testing. *The Annals of Statistics 42*(4), 1262–1311.

Durand, G. (2017). Adaptive p-value weighting with power optimality. *Preprint.* arXiv:1710.01094.

Efron, B. (2004). 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. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Stat. 2*, 197–223.

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

Ferkingstad, E., A. Frigessi, H. Rue, G. Thorleifsson, and A. Kong (2008). Unsupervised empirical bayesian multiple testing with external covariates. *The Annals of Applied Statistics 2*, 714–735.

Foster, D. P. and E. I. George (1996). A simple ancillarity paradox. *Scandinavian journal of statistics 23*(2), 233–242.

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

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

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

Heller, R., M. Bogomolov, and Y. Benjamini (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences 111*(46), 16262–16267.

Heller, R. and D. Yekutieli (2014). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics 8*(1), 481–498.

Hu, J. X., H. Zhao, and H. H. Zhou (2010). False discovery rate control with groups. *Journal of the American Statistical Association 105*(491), 1215–1227.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 361–379.

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

Langaas, M., B. H. Lindqvist, and E. Ferkingstad (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *J. Roy. Statist. Soc. B* ***67***, 555–572.

Law, N. M., S. R. Kulkarni, R. G. Dekany, E. O. Ofek, R. M. Quimby, P. E. Nugent, J. Surace, C. C. Grillmair, J. S. Bloom, M. M. Kasliwal, et al. (2009). The palomar transient factory: system overview, performance, and first results. *Publications of the Astronomical Society of the Pacific 121*(886), 1395.

Lehmann, E. L. and G. Casella (2006). *Theory of point estimation*. Springer Science & Business Media.

Li, A. and R. F. Barber (2016). Multiple testing with the structure adaptive benjamini-hochberg algorithm. *arXiv preprint arXiv:1606.07926*.

Liu, W. (2014). Incorporation of sparsity information in large-scale multiple two-sample t tests. *arXiv preprint arXiv:1410.4282*.

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

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

Neuvial, P. (2013). Asymptotic results on adaptive false discovery rate controlling procedures based on kernel estimators. *The Journal of Machine Learning Research 14*(1), 1423–1459.

Nugent, P. E., M. Sullivan, S. B. Cenko, R. C. Thomas, D. Kasen, D. A. Howell, D. Bersier, J. S. Bloom, S. R. Kulkarni, M. T. Kandrashoff, A. V. Filippenko, J. M. Silverman, G. W. Marcy, A. W. Howard, H. T. Isaacson, K. Maguire, N. Suzuki, J. E. Tarlton, Y.-C. Pan, L. Bildsten, B. J. Fulton, J. T. Parrent, D. Sand, P. Podsiadlowski, F. B. Bianco, B. Dilday, M. L. Graham, J. Lyman, P. James, M. M. Kasliwal, N. M. Law, R. M. Quimby, I. M. Hook, E. S. Walker, P. Mazzali, E. Pian, E. O. Ofek, A. Gal-Yam, and D. Poznanski (2011, 12). Supernova sn 2011fe from an exploding carbon-oxygen white dwarf star. *Nature 480*(7377), 344–347.

Reiner-Benaim, A., D. Yekutieli, N. E. Letwin, G. I. Elmer, N. H. Lee, N. Kafkafi, and Y. Benjamini (2007). Associating quantitative behavioral traits with gene expression in the brain: searching for diamonds in the hay. *Bioinformatics 23*(17), 2239–2246.

Roeder, K. and L. Wasserman (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics 24*(4), 398.

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

Rubin, D., S. Dudoit, and M. Van der Laan (2006). A method to increase the power of multiple testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology 5*(1), Article 19.

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

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

Schweder, T. and E. Spjøtvoll (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* 69: 493–502.

Scott, J. G., R. C. Kelly, M. A. Smith, P. Zhou, and R. E. Kass (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*, CRC press.

Wand, M. and M. Jones (1995). *Kernel smoothing*, Chapman & Hall, London.

Skol, A. D., L. J. Scott, G. R. Abecasis, and M. Boehnke (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet. 38*, 209–213.

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

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

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

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

Tukey, J. W. (1994). *The collected works of John W. Tukey*, Volume 3. Taylor & Francis.

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

Zehetmayer, S., P. Bauer, and M. Posch (2005). Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics 21*(19), 3771–3777.

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

Zablocki, R. W., A. J. Schork, R. A. Levine, O. A. Andreassen, A. M. Dale, and W. K. Thompson (2014). Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics 30*(15), 2098–2104.