

# 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*

and Weinan Wang

*University of Southern California, Los Angeles, USA*

[Read before The Royal Statistical Society at a meeting organized by the Research Section on Tuesday, December 4th, 2018, Professor A. Doucet in the Chair]

**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 cut-off to adjust for multiplicity. However, this data reduction step could cause significant loss of information and thus lead to suboptimal testing procedures. 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 the auxiliary variables. The proposed CARS procedure is shown to be asymptotically valid and optimal for false discovery rate control. The procedure is implemented in the R package CARS. Numerical results confirm the effectiveness of CARS in false discovery rate control and show that it achieves substantial power gain over existing methods. CARS is also illustrated through an application to the analysis of a 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 first to reduce the original samples to a vector of  $p$ -values and then to choose a cut-off 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 by using an auxiliary covariate sequence and develops optimal covariate-assisted inference procedures for large-scale two-sample multiple-testing problems.

*Address for correspondence:* T. Tony Cai, Department of Statistics, Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, USA  
E-mail: [tcai@wharton.upenn.edu](mailto:tcai@wharton.upenn.edu)

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, which is considered in section B.7 in the on-line 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 that is 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 have been converted into greyscale 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  $\mathbf{X}$  and  $\mathbf{Y}$  be two random vectors recording the measurement levels of the same  $m$  features under two experimental conditions. The population mean vectors are given by  $\boldsymbol{\mu}_x = \mathbb{E}(\mathbf{X}) = (\mu_{x1}, \dots, \mu_{xm})^\top$  and  $\boldsymbol{\mu}_y = \mathbb{E}(\mathbf{Y}) = (\mu_{y1}, \dots, \mu_{ym})^\top$ . A classical formulation for identifying differential features is to carry out  $m$  two-sample tests:

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

Suppose that we have collected two random samples  $\{\mathbf{X}_1, \dots, \mathbf{X}_{n_1}\}$  and  $\{\mathbf{Y}_1, \dots, \mathbf{Y}_{n_2}\}$  as independent copies of  $\mathbf{X}$  and  $\mathbf{Y}$  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 utilizes only a vector of  $p$ -values, may suffer from substantial loss of information.

This paper 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 method proposed 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, 2008; Reiner-Benaim *et al.*, 2007; Wasserman and Roeder, 2009; Bourgon *et al.*, 2010). Following that strategy, we 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 eliminate further 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 with a naive joint analysis that combines the data from both stages. The second strategy (Liu, 2014) can be described as *testing following grouping*, i.e. the hypotheses are analysed 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 the divide-and-test strategies involve dichotomizing a continuous variable, which fails to utilize the auxiliary information fully. 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 the 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 analyse a time course satellite image data set 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 Appendix A and the on-line appendix A. Additional numerical results are given in the on-line appendix B.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

<https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets>

## 2. Extracting structural information by using an auxiliary sequence

Suppose that  $\{X_{ij} : 1 \leq j \leq n_x\}$  and  $\{Y_{ik} : 1 \leq k \leq n_y\}$ ,  $i = 1, \dots, m$ , are repeated measurements of generic independent random variables  $X_i$  and  $Y_i$  respectively. Let  $\beta_0 = (\beta_{0i} : 1 \leq i \leq m)$  be a latent baseline vector which itself is sparse (including the special case where  $\beta_{0i} = 0$  for all  $i$ ). Consider the hierarchical model

$$\begin{aligned} X_{ij} &= \beta_{0i} + \mu_{xi}^* + \epsilon_{xij}, \\ Y_{ik} &= \beta_{0i} + \mu_{yi}^* + \epsilon_{yik}, \end{aligned} \quad (2.1)$$

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  $\epsilon_{xij} \sim^{\text{IID}} N(0, \sigma_{xi}^2)$  and  $\epsilon_{yij} \sim^{\text{IID}} N(0, \sigma_{yi}^2)$ . More general settings will be discussed in Sections 3.6. We assume that  $\mu_{xi}^*$  and  $\mu_{yi}^*$ , 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  $\delta_0$  is the Dirac delta function, and  $g_{\mu_x}$  and  $g_{\mu_y}$  are unspecified densities of non-zero effects.

*Remark 1.* Model (2.1) can be applied to scenarios with non-sparse  $\mu_x$  and  $\mu_y$  when some baseline measurements are available. See section A.8 in the on-line appendix for further details.

The methodology proposed 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  $\mu_{xi}$  and  $\mu_{yi}$  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  $H_{i,0} : \mu_{xi} = \mu_{yi}$  versus  $H_{i,1} : \mu_{xi} \neq \mu_{yi}$ ,  $i = 1, \dots, m$ . Let  $\mathbb{1}(\cdot)$  be an indicator function. Let  $T_{1i}$  and  $T_{2i}$  be summary statistics that contain the information about  $\theta_{1i} = \mathbb{1}(\mu_{xi} \neq \mu_{yi})$  (support of mean difference) and  $\theta_{2i} = \mathbb{1}(\mu_{xi} \neq 0 \text{ or } \mu_{yi} \neq 0)$  (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  $\theta_{1i}$  directly. Instead, we aim to incorporate  $T_{2i}$  in inference to support (indirectly) the evidence that is provided in the primary variable  $T_{1i}$ . The intuition is that, since the union support is sparse if both  $\mu_x$  and  $\mu_y$  are sparse, exploiting this structural information would improve the efficiency of tests. To see this, note that the continuity of  $\mu_{xi}$  and  $\mu_{yi}$  implies that, with probability 1,  $\theta_{1i}$  and  $\theta_{2i}$  obey the logical relationship

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

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

- (a) the pair extracts information from the data effectively and
- (b) 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  $\theta_{1i}$ ; hence  $\bar{X}_i - \bar{Y}_i$  should be incorporated in its expression. Second, to capture the information on the union support  $\theta_{2i}$ , 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, which is 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 these considerations, we propose to use the following pair of statistics to summarize the information in the data:

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

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

## 2.2. A bivariate random-mixture model

We develop a bivariate model to describe the joint distribution of  $T_{1i}$  and  $T_{2i}$ . Let  $\theta_i = (\theta_{1i}, \theta_{2i})$ . Assume that  $\theta_i$  are independent and identically distributed 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  $\theta_i = (j, k)$ ,  $(T_{1i}, T_{2i})$  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 observe only  $(T_{1i}, T_{2i})$  from a mixture model

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

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

The mixture model (2.4) is difficult to analyse. 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 have only three terms in equation (2.4). Second, according to our construction (2.3),  $T_{1i}$  and  $T_{2i}$  are conditionally independent:

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

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

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

$$\left. \begin{aligned} f(t_{1i}, t_{2i} | \theta_{1i} = 0, \theta_{2i} = 0) &= f(t_{1i} | \theta_{1i} = 0) f(t_{2i} | \theta_{2i} = 0), \\ f(t_{1i}, t_{2i} | \theta_{1i} = 0, \theta_{2i} = 1) &= f(t_{1i} | \theta_{1i} = 0) f(t_{2i} | \theta_{1i} = 0, \theta_{2i} = 1), \\ f(t_{1i}, t_{2i} | \theta_{1i} = 0) &= f(t_{1i} | \theta_{1i} = 0) f(t_{2i} | \theta_{1i} = 0). \end{aligned} \right\} \quad (2.6)$$

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

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

## 2.3. Problem formulation

Our goal is to make inference on  $\theta_{1i} = \mathbb{1}(\mu_{xi} \neq \mu_{yi})$ ,  $1 \leq i \leq m$ , by simultaneously testing  $m$  hypotheses  $H_{i,0} : \theta_{1i} = 0$  versus  $H_{i,1} : \theta_{1i} = 1$ . Compared with conventional approaches, we aim to develop methods utilizing  $m$  pairs  $\{(T_{1i}, T_{2i}) : 1 \leq i \leq m\}$  instead of a single vector  $\{T_{1i} : 1 \leq i \leq m\}$ . 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

$$\delta = \{\delta_i = \mathbb{1}(S_i < t) : i = 1, \dots, m\} \in \{0, 1\}^m, \quad (2.8)$$

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

$\delta = (\delta_i : 1 \leq i \leq m)$  of the form (2.8), FDR is defined as

$$\text{FDR}_\delta = \mathbb{E} \left\{ \frac{\sum_{i=1}^m (1 - \theta_{1i}) \delta_i}{\left( \sum_{i=1}^m \delta_i \right) \vee 1} \right\}, \quad (2.9)$$

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

$$\text{mFDR}_\delta = \frac{\mathbb{E} \left\{ \sum_{i=1}^m (1 - \theta_{1i}) \delta_i \right\}}{\mathbb{E} \left( \sum_{i=1}^m \delta_i \right)}. \quad (2.10)$$

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

Define the expected number of true positive results  $\text{ETP}_\delta = \mathbb{E}(\sum_{i=1}^m \theta_{1i} \delta_i)$ . Other related power measures include the missed discovery rate (Taylor *et al.*, 2005), the average power (Efron, 2007) and false non-discovery or false negative rate (Genovese and Wasserman, 2002; Sarkar, 2002). Our optimality result is developed based on mFDR and ETP. We call a multiple-testing procedure *valid* if it controls 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  $f_j = (1 - \pi_j) f_{j0} + \pi_j f_{j1}$ , 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  $T_{ji}$  respectively. Conventional FDR-procedures, which are developed based on a vector of  $p$ -values or  $z$ -values, are essentially *univariate inference procedures* that utilize only the information of  $T_{1i}$ . Define the local false discovery rate Lfdr (Efron *et al.*, 2001) as

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

where subscript '1' indicates a quantity that is 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

$$\delta(\text{Lfdr}_1, c) = [\{ \text{Lfdr}_1(t_i) < c \} : 1 \leq i \leq m], \quad (3.2)$$

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

The following theorem derives an oracle procedure for mFDR-control when the pairs  $(T_{1i}, T_{2i})$  are given. We shall see that the performance of  $\delta^*$ , 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  $T_{\text{OR}}^i$  that optimally pools information from both  $T_{1i}$  and  $T_{2i}$ , and an oracle threshold  $t_{\text{OR}}$  that controls the mFDR with the largest ETP.

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

$$q^*(t_2) = (1 - \pi_1)f(t_2 | \theta_{1i} = 0). \quad (3.3)$$

Define the oracle statistic

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

where  $f(t_1, t_2)$  is the joint density given by equation (2.7). Then we have the following results.

- (a) For  $0 < \lambda \leq 1$ , let  $Q_{\text{OR}}(\lambda)$  be the mFDR-level of testing rule  $\{\mathbb{1}(T_{\text{OR}}^i < \lambda) : 1 \leq i \leq m\}$ . Then  $Q_{\text{OR}}(\lambda) < \lambda$  and  $Q_{\text{OR}}(\lambda)$  is non-decreasing in  $\lambda$ .
- (b) Suppose that we choose  $\alpha < \bar{\alpha} \equiv Q_{\text{OR}}(1)$ . Then the oracle threshold  $\lambda_{\text{OR}} = \sup\{\lambda : Q_{\text{OR}}(\lambda) \leq \alpha\}$  exists uniquely and  $Q_{\text{OR}}(\lambda_{\text{OR}}) = \alpha$ . Furthermore, define oracle rule  $\delta_{\text{OR}} = (\delta_{\text{OR}}^i : i = 1, \dots, m)$ , where

$$\delta_{\text{OR}}^i = \mathbb{1}(T_{\text{OR}}^i < \lambda_{\text{OR}}). \quad (3.5)$$

Then  $\delta_{\text{OR}}$  is optimal in the sense that  $\text{ETP}_{\delta} \leq \text{ETP}_{\delta_{\text{OR}}}$  for any  $\delta$  in  $\mathcal{D}_{\alpha}$ , where  $\mathcal{D}_{\alpha}$  is the collection of all testing rules based on  $\mathbf{T}_1$  and  $\mathbf{T}_2$  such that  $\text{mFDR}_{\delta} \leq \alpha$ .

*Remark 2.* The oracle statistic  $T_{\text{OR}}^i$  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 neighbourhood of  $t_2$ . The estimation of  $T_{\text{OR}}$  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 that were 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 on-line supplementary material, we show that under the above conditions  $T_{\text{OR}}^i$  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  $T_{\text{OR}}$ , and the thresholding step identifies the largest threshold along the ranking subject to the constraint on FDR. Specifically, denote  $T_{\text{OR}}^{(1)} \leq \dots \leq T_{\text{OR}}^{(m)}$

the ordered oracle statistics and  $H_{(1)}, \dots, H_{(m)}$  the corresponding hypotheses. The stepwise procedure operates as follows:

$$\text{let } k = \max \left\{ j : j^{-1} \sum_{i=1}^j T_{\text{OR}}^{(i)} \leq \alpha \right\}; \text{ reject } H_{(1)}, \dots, H_{(k)}. \quad (3.6)$$

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

### 3.2. Approximating $T_{\text{OR}}$ via screening

The oracle statistic  $T_{\text{OR}}^i$  is unknown and needs to be estimated. However, standard methods do not work well for the bivariate model. For example, the popular expectation–maximization 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  $T_{\text{OR}}^{\tau, i}$  that involves only quantities that can be well estimated from data. The new statistic provides a good approximation to  $T_{\text{OR}}^i$  and guarantees the FDR-control.

In definition (3.4), the null density  $f_{10}$  is known by construction. The bivariate density  $f(t_1, t_2)$  can be well estimated by using a standard kernel method (Silverman, 1986; Wand and Jones, 1995). Hence we shall focus on the quantity  $q^*(t_2)$ . Suppose that 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 neighbourhood 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  $q^*(t_2) = \lim_{h \rightarrow 0} \mathbb{E}\{Q^*(t_2, h)\} / h$ . The counting task is difficult as we do not know the value of  $\theta_{1i}$ . Our idea is first to apply a screening method to select the nulls (i.e.  $\theta_{1i} = 0$ ), and then to construct an estimator based on selected cases.

Denote  $P_i$  the  $p$ -value that is associated with  $T_{1i} = t_{1i}$ . For a large  $\tau$ , 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:

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

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

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

Intuitively, a large  $\tau$  would yield a sample that is close to a sample that is 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  $\tau$  (proposition 4). However, a larger  $\tau$  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 trade-off is further discussed in Section 3.4.



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

$$T_{\text{OR}}^{\tau,i}(t_1, t_2) = \frac{q^\tau(t_2) f_{10}(t_1)}{f(t_1, t_2)}. \quad (3.9)$$

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

*Proposition 2.*

- (a)  $T_{\text{OR}}^i(t_1, t_2) \leq T_{\text{OR}}^{\tau,i}(t_1, t_2)$  for all  $\tau$ .
- (b) Let  $\delta_{\text{OR}}^\tau$  be a decision rule that substitutes  $T_{\text{OR}}^{\tau,i}$  in place of  $T_{\text{OR}}^i$  in procedure (3.6). Then both the FDR- and the mFDR-levels of  $\delta_{\text{OR}}^\tau$  are controlled below level  $\alpha$ .

### 3.3. Estimation of the test statistic

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

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

where  $K(t)$  is a kernel function, and  $h_1$  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:

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

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  $q^\tau$  respectively defined in equations (3.8) and (3.11). Assume that

- (a)  $q^\tau(\cdot)$  is bounded and has continuous first and second derivatives;
- (b) 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
- (c)  $f_{2 \cdot}^{(2)}(t_2|\tau) = \int_{t_1 \in \mathcal{A}_\tau} \int f_{2 \cdot}^{(2)}(t_2|t_1) f_{1 \cdot}(t_1) dt_1$  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

$$\mathbb{E} \|\hat{q}^\tau - q^\tau\|^2 = \mathbb{E} \left[ \int \{\hat{q}^\tau(x) - q^\tau(x)\}^2 dx \right] \rightarrow 0.$$

Combining these above results, we propose to estimate  $T_{\text{OR}}^\tau$  by the statistic

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

where  $\hat{q}^\tau(t_2)$  and  $\hat{f}(t_1, t_2)$  are respectively given in equations (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 equation (3.10) and equation (3.11). Utilizing the same bandwidth across the numerator

and denominator of equation (3.12) has no effect 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 estimator proposed 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 to prove its consistency. Let the cumulative distribution function of the  $p$ -value that is associated with  $T_{1i}$  be  $G(\tau) = (1 - \pi_1)\tau + \pi_1 G_1(\tau)$ , where  $G_1$  is the alternative cumulative distribution function. Denote  $g$  and  $g_1$  the corresponding density functions.

*Proposition 4.* Consider  $T_{OR}^\tau$  defined in equation (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  $\tau$ .
- (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) has been commonly used in the literature (Storey, 2002; Genovese and Wasserman, 2002; Sun and Cai, 2007); the monotone likelihood ratio condition 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 (Neuivial, 2013). It would be desirable to develop a more general condition in future work.

It follows from proposition 4 that a large  $\tau$  is helpful to reduce the bias and the bias converges to 0 when  $\tau \rightarrow 1$ . However, a large  $\tau$  would increase the variance of our estimator (3.11), which is constructed by using the sample  $\mathcal{T}(\tau) = \{i : P_{1i} > \tau\}$ . To address the bias–variance trade-off, we propose first to obtain  $\hat{q}^\tau$  for a range of  $\tau$ s, say  $\{\tau_1, \dots, \tau_k\}$ , and then to use a smoothing method to obtain the limiting value of  $\hat{q}^\tau$  when  $\tau \rightarrow 1$ . This approach aims to borrow strength from the entire sample to minimize the bias without blowing up the variance. Specifically, let  $\tau_0 < \tau_1 < \dots < \tau_k$  be ordered and equally spaced points in the interval  $(0, 1)$ . Denote  $\hat{q}^{\tau_j}(t_2)$  the estimates from equation (3.11),  $j = 1, \dots, k$ . We propose to obtain the local linear kernel estimator  $\hat{q}^*(t_2) \equiv \hat{q}\{\tau = 1; \hat{q}^{\tau_1}(t_2), \dots, \hat{q}^{\tau_k}(t_2)\}$  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), page 119)

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

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 equations (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 that the following conditions hold:

- (a)  $q^{\tau k, (2)}(t_2)$  is square integrable,  
 (b)  $K(\cdot)$  is symmetric about zero and is supported on  $[-1, 1]$  and  
 (c) the bandwidth  $h_\tau$  is a sequence satisfying  $h_\tau \rightarrow 0$  and  $kh_\tau \rightarrow \infty$  as  $k \rightarrow \infty$ .

Moreover, assume that conditions (a)–(c) in proposition 3 and condition (b) in proposition 4 hold. We have

$$\mathbb{E}\|\hat{q}^* - q^*\|^2 = \mathbb{E}\left[\int\{\hat{q}^*(x) - q^*(x)\}^2 dx\right] \rightarrow 0 \quad \text{when } (m, k) \rightarrow 0. \quad (3.14)$$

### 3.5. The covariate-assisted ranking and screening procedure

The estimated statistics  $\hat{T}_{\text{OR}}^{\tau, i}$  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 CARS procedure.

(Procedure 1). Consider model (2.7) and estimated statistics  $\hat{T}_{\text{OR}}^{\tau, i}$  (3.12). Denote  $\hat{T}_{\text{OR}}^{\tau, (1)} \leq \dots \leq \hat{T}_{\text{OR}}^{\tau, (m)}$  the ordered statistics and  $H_{(1)}, \dots, H_{(m)}$  the corresponding hypotheses. Let  $k = \max\{j: j^{-1}\sum_{i=1}^j \hat{T}_{\text{OR}}^{\tau, (i)} \leq \alpha\}$ . Then reject  $H_{(1)}, \dots, H_{(k)}$ .

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

*Condition 1.*  $\mathbb{E}\|\hat{q}^\tau - q^\tau\|^2 \rightarrow 0$ .

*Condition 1'.*  $\mathbb{E}\|\hat{q}^* - q^*\|^2 \rightarrow 0$ .

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

*Remark 6.* Proposition 3 shows that condition 1 is satisfied by the proposed estimator (3.11). Proposition 5 shows that condition 1' is satisfied by the smoothing estimator (3.14) under stronger assumptions. Finally, condition 2 is satisfied by the standard choice of bandwidth  $h_{t1} \sim m^{-1/6}$  and  $h_{t2} \sim m^{-1/6}$ ; 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 1 and 2 hold, then both the mFDR and the FDR of the CARS procedure are controlled at level  $\alpha + o(1)$ .  
 (b) If conditions 1' and 2 hold, and we substitute  $\hat{q}^*$  (3.14) in place of  $\hat{q}^\tau$  in equation (3.12) to compute  $\hat{T}_{\text{OR}}^\tau$ , then the FDR-level of the CARS procedure is  $\alpha + o(1)$ . Moreover, denote  $\text{ETP}_{\text{CARS}}$  and  $\text{ETP}_{\text{OR}}$  the ETP-levels of CARS and the oracle procedure respectively. Then we have  $\text{ETP}_{\text{CARS}}/\text{ETP}_{\text{OR}} = 1 + o(1)$ .

### 3.6. Case with unknown variances and non-Gaussian errors

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

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

For the case with unknown but equal variances (e.g.  $\sigma_{xi}^2 = \sigma_{yi}^2$ ), we modify equation (3.15) as follows. First,  $\hat{\kappa}_i^*$  is replaced by  $\kappa^* = \gamma_y / \gamma_x$ . Second,  $S_{pi}^2$  is instead estimated by  $S_{pi}^2 = \gamma_x S_{xi}^2 +$

1  $\gamma_y S_{y_i}^2$ . Finally  $T_{1i}$  and  $T_{2i}$  are plugged into equation (3.12) to compute the CARS statistic, which  
 2 is further employed to implement procedure 1.

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

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

8  
 9 The expression for the asymptotic variance–covariance matrix, which is given in section A.6  
 10 of the on-line supplementary material, reveals that the asymptotic independence holds for non-  
 11 Gaussian errors as long as the error distributions are symmetric. Our simulation results in  
 12 Section 5.3 confirm that unknown variances and non-Gaussian errors have almost no effect  
 13 on the performance of CARS. Therefore the plug-in methods are recommended for practical  
 14 applications. The case with a skewed distribution requires further research, and a full theoretical  
 15 justification of CARS methodology is still an open question.

## 17 4. Extensions and connections with existing work

18  
 19 This section considers the extension of our theory to a general bivariate model. The results in  
 20 the general model provide a unified theoretical framework for understanding different testing  
 21 strategies, which helps to gain insights on the connections between existing methods.

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

### 32 4.1. A general bivariate model

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

35  
 36 *Theorem 3.* Define the oracle statistic under the general model

$$37 T_{\text{OR}}^G(t, s) = \mathbb{P}(\theta_i = 0 | T_i = t, S_i = s). \quad (4.1)$$

38 Denote  $Q_{\text{OR}}^G(\lambda)$  the mFDR-level of  $\delta(T_{\text{OR}}^G, \lambda)$ , where  $\delta(T_{\text{OR}}^G, \lambda) = \{\mathbb{1}\{T_{\text{OR}}^G(t, s) < \lambda\} : 1 \leq i$   
 39  $\leq m\}$ . Let  $\lambda_{\text{OR}} = \sup\{\lambda \in (0, 1) : Q_{\text{OR}}^G(\lambda) \leq \alpha\}$ . Define the oracle mFDR-procedure under  
 40 the general model as  $\delta_{\text{OR}}^G = \delta(T_{\text{OR}}^G, \lambda_{\text{OR}})$ . Then  $\delta_{\text{OR}}^G$  is optimal in the sense that, for any  $\delta$   
 41 such that  $\text{mFDR}_\delta \leq \alpha$ , we always have  $\text{ETP}_\delta \leq \text{ETP}_{\delta_{\text{OR}}^G}$ .

42  
 43 Theorem 1 can be viewed as a special case of theorem 3. However, theorem 3 is of less practical  
 44 importance as the ‘best’ data-driven solution to theorem 3 may depend on various factors such as

- 45 (a) whether the auxiliary statistic is categorical or continuous,
- 46 (b) whether the null distribution of  $T_i$  is fixed and known and
- 47 (c) whether  $S_i$  and  $T_i$  are independent etc.

The key issue is that estimating  $T_{\text{OR}}^{G,i}$  is very difficult in a general bivariate model. Under some special cases,  $T_{\text{OR}}^G$  can be approximated well. For example,  $T_{\text{OR}}^T$  provides a good approximation to  $T_{\text{OR}}^G$  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 *multigroup random-mixture model* that was 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  $\{1, \dots, K\}$  with prior probabilities  $\{\pi_1, \dots, \pi_K\}$ . Consider the conditional distributions

$$(T_i | S_i = k) \sim f_1^k = (1 - \pi_1^k) f_{10}^k + \pi_1^k f_{11}^k, \quad (4.2)$$

for  $k = 1, \dots, K$ , where  $\pi_1^k$  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 expression (4.2) to vary across groups; this is desirable in practice when groups are heterogeneous.

*Remark 7.* In section A.10 in the on-line 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 (Lehmann and Casella (2006) page 420),

‘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  $S_i = k$ ,  $k = 1, \dots, K$ . 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. 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:

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

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).

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

### 7 4.3. *Discretization and uncorrelated screening*

8 The idea in Liu (2014) involves discretizing a continuous  $S_i$  as a binary variable. Define index sets  
 9  $\mathcal{G}_1 = \{1 \leq i \leq m : S_i > \lambda\}$  and  $\mathcal{G}_2 = \{1 \leq i \leq m : S_i \leq \lambda\}$ , where the tuning parameter  $\lambda$  divides  
 10  $t_{1i}$ 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 US method (Liu, 2014)  
 11 operates in two steps. First, the Benjamini and Hochberg (1995) method is applied at level  $\alpha$  to  
 12 the two groups separately, and then the rejected hypotheses from two groups are combined as  
 13 the final output. The tuning parameter  $\lambda$  is chosen in a way such that it yields the largest number  
 14 of rejections (two groups combined). US is closely related to the *separate analysis* strategy that  
 15 was proposed in Efron (2008). The key difference is that the groups were known *a priori* in  
 16 Efron (2008), whereas the groups were chosen adaptively in Liu (2014). The main merit of US is  
 17 that the screening statistic is constructed to be uncorrelated with the test statistic, which ensures  
 18 that the selection bias issue can be avoided. Moreover, the divide-and-test strategy combines  
 19 the results in both groups; this is different from conventional independent filtering approaches  
 20 (Bourgon *et al.*, 2010), in which one group is completely filtered out.

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

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

### 45 4.4. *The ‘pooling-within’ strategy for information integration*

46 Tukey (1994) coined two witted terms to advocate some of his favourite information integration  
 47 strategies: *borrowing strength* and *pooling within*. The idea of borrowing strength, which was  
 48 investigated extensively and systematically by researchers in both simultaneous estimation and

multiple-testing fields, has led to some 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), page 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 evidence that is exhibited from both statistics.

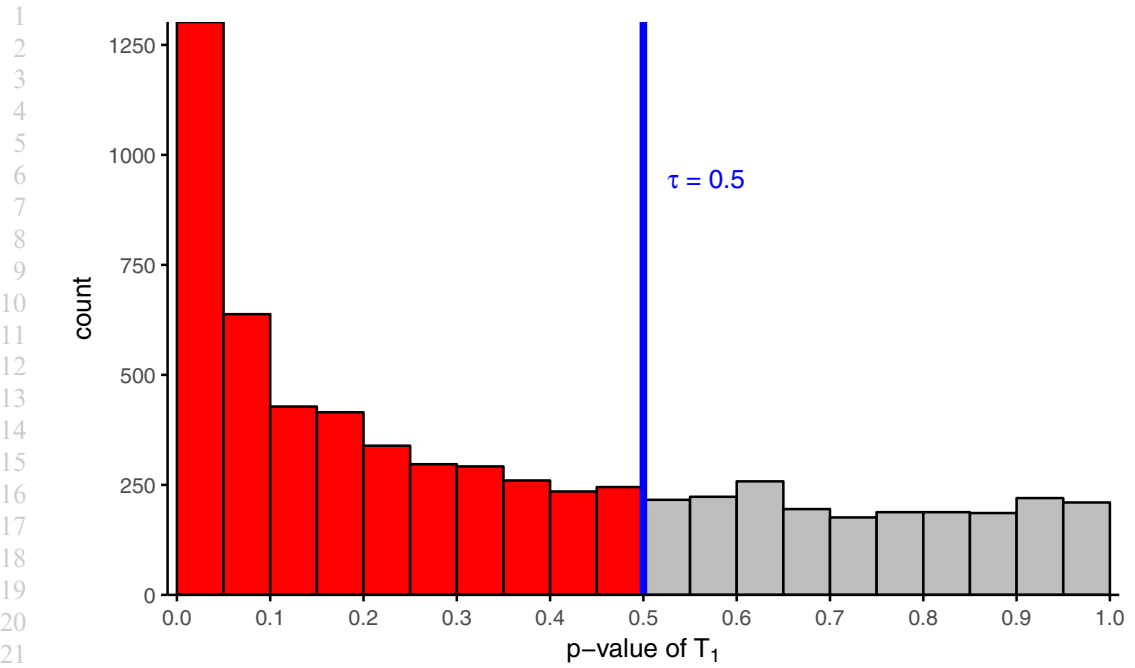
Our work differs in several ways from existing works on multiple testing with covariates (Feringstad *et al.*, 2008; Zablocki *et al.*, 2014; Scott *et al.*, 2015). First, the covariate in other works is collected *externally* from other data sources, whereas the auxiliary information in our work is 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 that is employed by CARS reveals interesting connections between sparsity estimation and multiple testing with covariates, which is elaborated next.

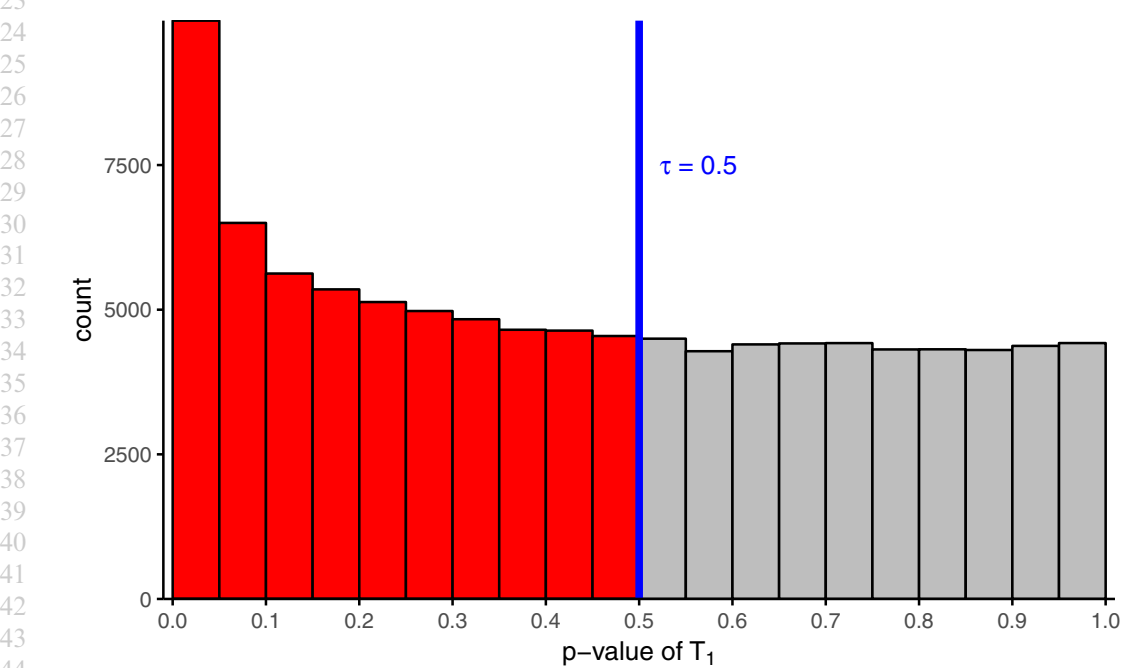
#### 4.5. Capturing sparsity information via screening

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 of existing methods (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 (Feringstad *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 estimator proposed. 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 that 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 in Fig. 1(a) for each  $p$ -value interval. As shown in proposition 1,  $T_i$  and  $S_i$  are independent under the null



(a)



(b)

**Fig. 1.** 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  $\tau$  increases; (b) histogram of all  $p$ -values—Similarly, the  $p$ -values are approximately uniformly distributed on the right



(see 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 (Fig. 1 (b)).

We start with a description of a classical estimator (Schweder and Spjøtvoll, 1982; Storey, 2002) for  $\pi_1$ ; see Langaas *et al.* (2005) for an 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  $\tau$  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 neighbourhood  $[s - h/2, s + h/2]$ . 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 counting can be done by using equation (3.7) in Section 3.2. In obtaining equation (3.7), we exploit the fact that the counts  $S_i$  are roughly uniformly distributed to the right of the threshold  $\tau$ . Taking the limit, we obtain  $\pi_1(s) = 1 - f(s)^{-1} \lim_{h \rightarrow 0} Q^\tau(s, h)/h = 1 - q^\tau(s)/f(s)$ . Finally, utilizing the screening approach (3.11), we propose the following non-parametric smoothing estimator

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

Choosing tuning parameter  $\tau$  is an important issue but has gone beyond the scope of the current work; see Storey (2002) and 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 procedure BH 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 FDR. However, as we previously mentioned, discretizing a continuous  $S_i$  fails to utilize the auxiliary information fully. The estimation of the empirical null with a continuous  $S_i$  is an interesting problem for future research. The non-parametric smoothing idea in estimator (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 by using both simulated and real

1 data. We compare the oracle and data-driven CARS procedures, respectively denoted by OR and  
 2 DD, with existing methods, including the Benjamini–Hochberg procedure BH (Benjamin and  
 3 Hochberg, 1995), the adaptive  $z$ -value procedure AZ (Sun and Cai 2007) and the US procedure  
 4 (Liu 2014). We first describe the implementation of CARS in Section 5.1. Sections 5.2 and 5.3  
 5 respectively consider

- 6 (a) the case with known and unequal variances and
- 7 (b) the case with estimated variances and non-Gaussian errors.

9 Section 5.4 provides numerical evidence to show the merit of CARS when the two means have  
 10 opposite signs. An application to supernova detection is discussed in Section 5.5. Additional nu-  
 11 merical results including the non-informative case, completely informative case and dependent  
 12 tests are provided in sections B.2, B.3 and B.6 respectively in the on-line supplementary material.

### 14 5.1. The implementation and R-package CARS

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

17 The bivariate density estimator  $\hat{f}(t_1, t_2)$  can become unstable in very sparse settings, which may  
 18 lead to slightly elevated FDR-levels (see Fig. 2(a)). To increase the stability of CARS in the ex-  
 19 tremely sparse setting where the non-null proportion is vanishingly small, the CARS package has  
 20 included a ‘sparse’ option, which implements a conservative but more stable density estimator

$$22 \hat{f}^v(t_1, t_2) = (1 - \hat{\pi}_2) f_{10}(t_1) f_{20}(t_2) + \mathbb{G}_L(v) \{1 - \widehat{\text{FDR}}_2(v)\} \hat{f}(t_1, t_2 | \widehat{\text{Lfd}}_2 < v). \quad (5.1)$$

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

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

46 The R package `np` is used to choose the bandwidths  $h_1$  and  $h_2$  in equation (3.10). We have  
 47 adopted two strategies to improve the performance. First, the bandwidths  $h_1$  and  $h_2$  are chosen  
 48 based on the *normal reference rule* restricted to the samples with `Lfd`<sub>1</sub> < 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 expression (3.11) to obtain the numerator of expression (3.12). This strategy is helpful to increase the stability of the estimator (see remark 4). Finally we note that these strategies are only practical guidelines in finite samples; the asymptotic theories are not affected.

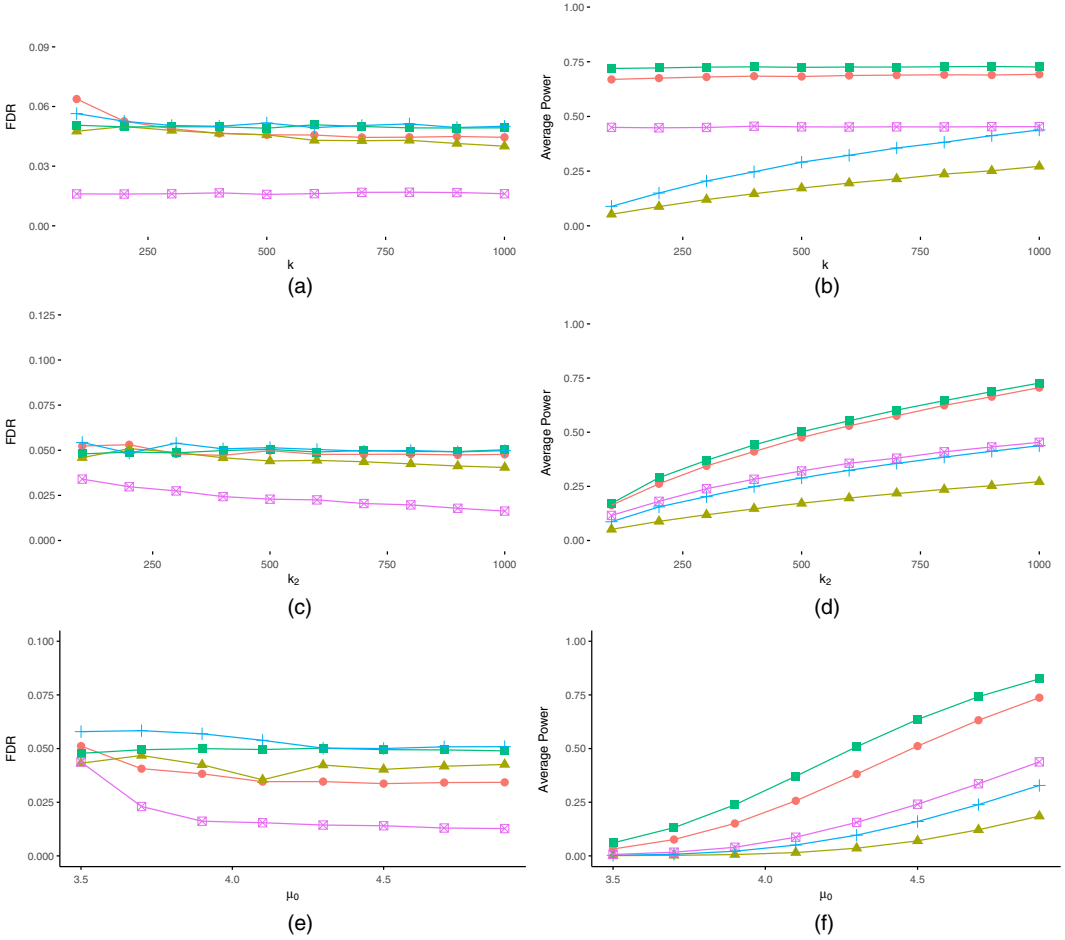
## 5.2. Simulation 1: known variances

Consider model (2.1). Denote  $\boldsymbol{\mu}_{x,i_1:i_2} = (\mu_{x,i_1}, \dots, \mu_{x,i_2})$  and  $\boldsymbol{\mu}_{y,i_1:i_2} = (\mu_{y,i_1}, \dots, \mu_{y,i_2})$  the vectors of consecutive observations from  $i_1$  to  $i_2$ . The two original samples are denoted  $\{\mathbf{x}_1, \dots, \mathbf{x}_{n_x}\}$  and  $\{\mathbf{y}_1, \dots, \mathbf{y}_{n_y}\}$ , 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 three 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 various parameter values and displayed in Fig. 2.

- (a) 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 non-zero locations is  $2k/m$ . We vary  $k$  from 100 to 1000 to investigate the effect of sparsity.
- (b) Setting 2: we use  $k_1$  and  $k_2$  to denote the number of non-zero 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.
- (c) 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  $\boldsymbol{\mu}_{y,(k+1):(2k)} = 3/\sqrt{30}$  and  $\boldsymbol{\mu}_{y,(2k+1):m} = 0$ . To investigate the effect of the effect sizes, we vary  $\mu_0$  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 to US which only partially utilizes the auxiliary information. A more detailed description of simulation results is given below.

- (a) All methods control 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 models  $T_{2i}$  as only a binary variable whereas CARS fully utilizes the information in  $T_{2i}$ .
- (d) DD has a similar performance to that of OR in most settings. However, DD can be conservative in FDR control in some settings and hence has less power compared with OR (see 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 the gain in efficiency of CARS increases when  $k_2$  increases. Note that  $k_2$  is proportional to  $\pi_{1|1}$  (the informativeness of the auxiliary covariate).
- (g) Setting 3 shows that the gain in efficiency of CARS increases as the signal strength decreases (note that a smaller  $\mu_0$  corresponds to a larger difference in effect sizes).



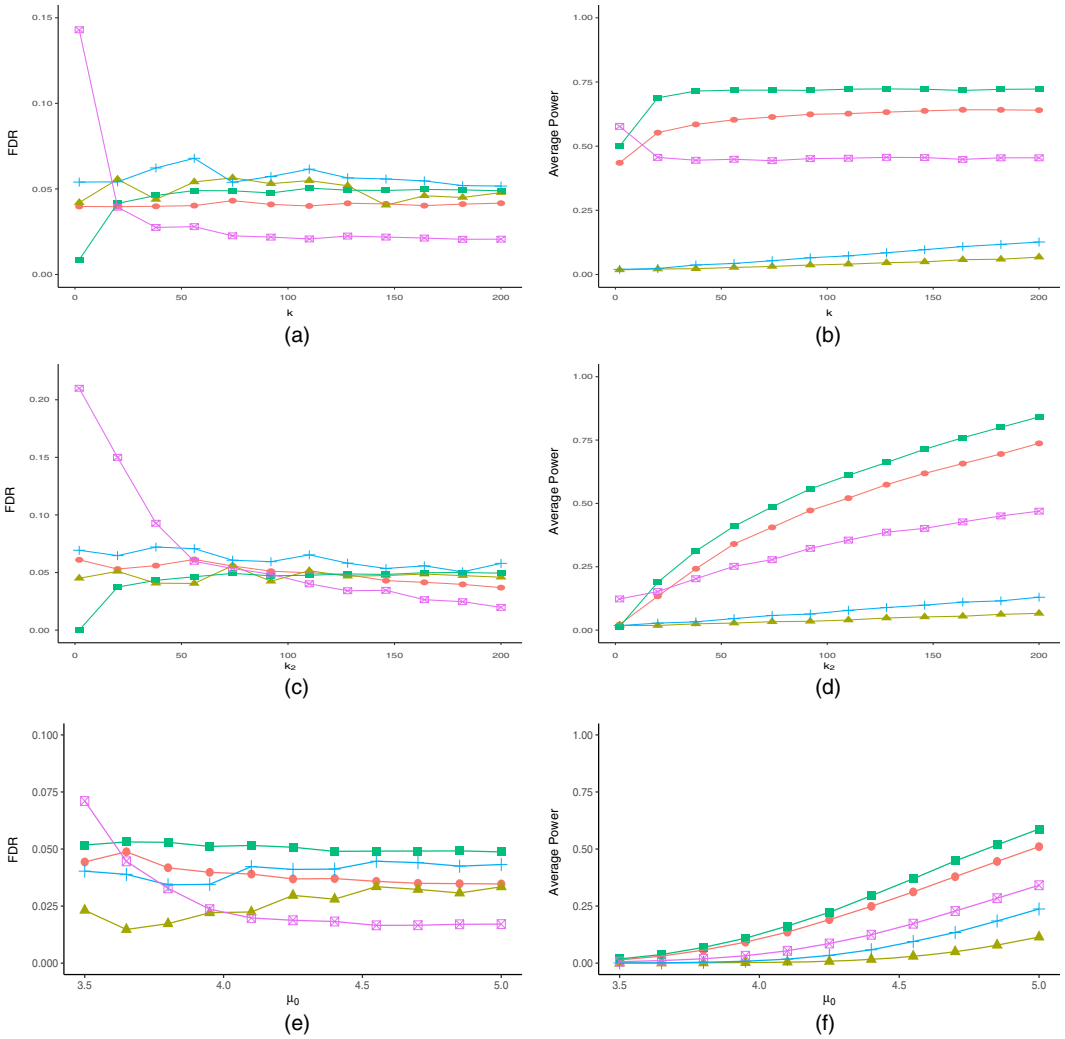
**Fig. 2.** Two-sample tests with known variances: FDR and average power (ETP divided by the number of non-nulls) are plotted against (a), (b) varied non-null proportions (setting 1), (c), (d) conditional proportions (setting 2) and (e), (f) effect sizes (setting 3) (●, DD; ▲, BH; ■, OR; +, AZ; ✕, US)

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

We consider similar simulation settings to those in 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  $\epsilon_{xij}$  and  $\epsilon_{yik}$  from a  $t$ -distribution with degrees of freedom  $df = 4$  and  $df = 5$  respectively. Finally, we vary  $k$  from 1 to 200 to investigate the performance of CARS under various sparsity regimes. The modified density estimator  $\hat{f}^v(t_1, t_2)$ , which is defined in equation (5.1), has been used in all settings. The simulation results are summarized in Fig. 3.

The patterns are very similar to those in simulation I; our conclusions on the comparison of various methods remain the same. We mention the following points.

- (a) Settings 1 and 2 show that 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.



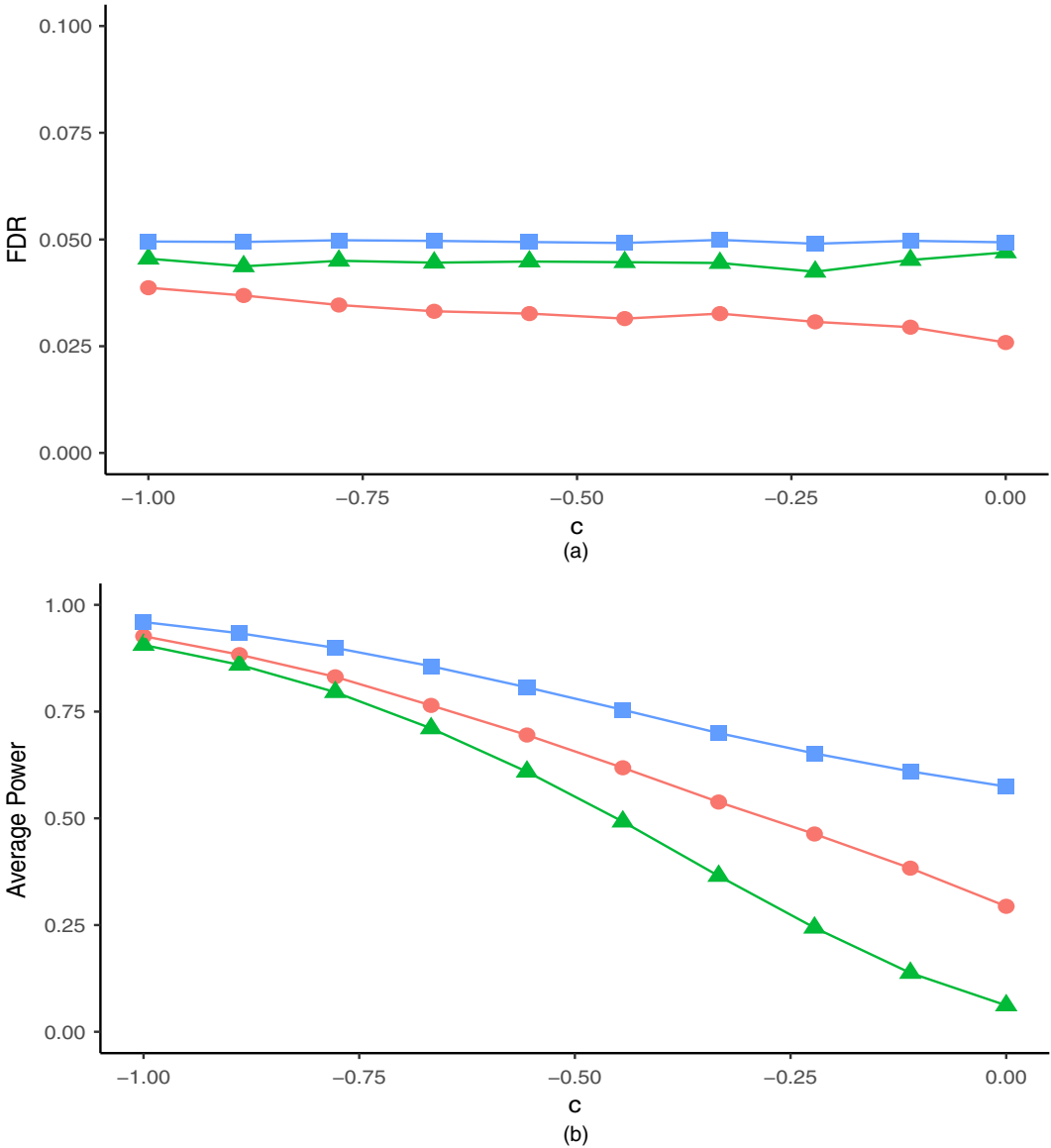
**Fig. 3.** Two-sample tests with unknown variances and non-Gaussian errors: FDR and MDR are plotted against (a), (b) varied non-null proportions (setting 1), (c), (d) conditional proportions and (e), (f) effect sizes (setting 3) (●, DD; ▲, BH; ■, OR; +, AZ; ✕, US)

### 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 on-line 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 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 independent and identically distributed noise. Set  $n_x = n_y = 50$ . In our simulation, the number of tests is  $m = 10000$ . The two mean vectors are



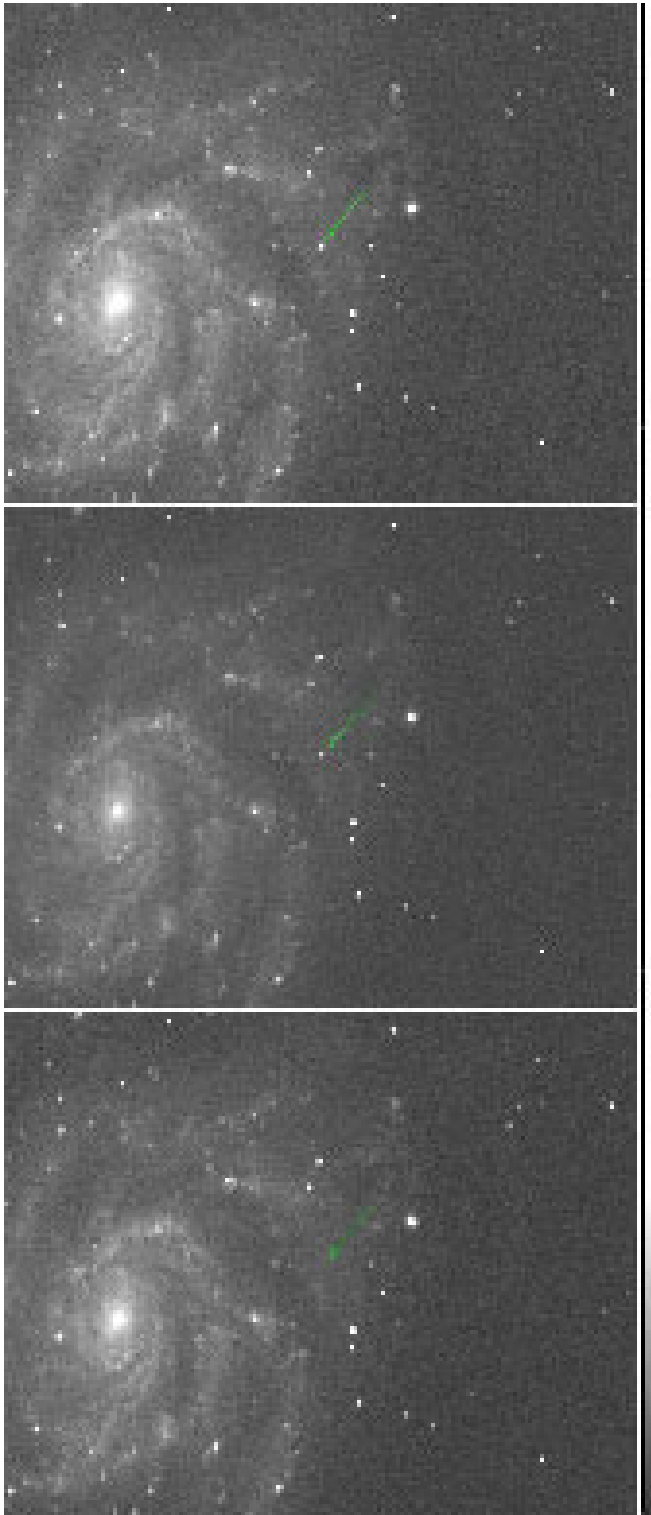
**Fig. 4.** Comparison of BH ( $\blacktriangle$ ), DD ( $\bullet$ ) and OR ( $+$ ) when the non-zero means have opposite signs

$$\mu_{x,1:500} = 3/\sqrt{50}, \quad \mu_{x,501:1000} = 4/\sqrt{50}, \quad \mu_{x,1001:m} = 0$$

and

$$\mu_{y,1:500} = 3c/\sqrt{50}, \quad \mu_{y,501:1000} = 4/\sqrt{50}, \quad \mu_{y,1001:m} = 0.$$

We vary  $c$  from  $-1$  to  $0$ , where  $c = -1$  corresponds to the least favourable situation where the two means cancel out precisely. We apply the Benjamini–Hochberg 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 Fig. 4.



**Fig. 5.** Images are taken on (a) August 23rd, (b) August 24th, and (c) August 25th, 2011: the arrows clearly indicate the explosion of supernova SN 2011fe

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48

We can see that, when  $c = -1$ , all methods have similar power. As  $c$  increases to 0, 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. Fig. 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, which is one of the brightest supernovas known 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 the potential locations for supernova explosions. More accurate measurements and further investigations will then be carried out on the narrowed subset of potential locations.

The satellite data are recorded and converted into greyscale images of size  $516 \times 831$  (or  $m = 428796$  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 greyscale pixel values are subtracted from those in images 2 and 3. These differences are then vectorized as  $\mathbf{x} = (x_1, \dots, x_m)$  and  $\mathbf{y} = (y_1, \dots, y_m)$  (respectively representing ‘image 2 – image 1’ and ‘image 3 – image 1’).

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

Next we carry out  $m = 428796$  two-sample tests with known variances. For standardized observations  $\mathbf{x}^{\text{st}}$  and  $\mathbf{y}^{\text{st}}$  the variances are known to be 1. Then  $t_{1i} = (x_i^{\text{st}} - y_i^{\text{st}})/\sqrt{2}$  and  $t_{2i} = (x_i^{\text{st}} + y_i^{\text{st}})/\sqrt{2}$ . We apply BH, AZ, US and CARS at FDR-levels 0.01%, 1% and 5%. Fig. 12 in the on-line supplementary material shows the rejected pixels in the  $516 \times 831$  layout for each method under various FDR-levels. The estimated sparsity levels for  $\mathbf{t}_1$  and  $\mathbf{t}_2$  are respectively 1.47% and 49.5%. The corresponding estimated support size at  $\tau = 0.5$  is 6285. We report the thresholds of various 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 detected by

**Table 1.** Thresholds and total number of rejections (in parentheses) of various testing procedures

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



CARS. To quantify our procedure's superiority further, we count the total number of rejections in Table 1 (the numbers in parentheses). We can see that CARS consistently detects more signals from the satellite images than do the 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 setwise 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 are yet to be developed for handling various types of problem 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 whereas others are secondary.

We conclude our discussion with a few more open issues.

- (a) *Are there better ways to construct the auxiliary sequence?* Our theory shows only 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.
- (b) *How can we deal with multiple-testing dependence?* The CARS method cannot be applied to dependent tests as it assumes that  $T_i$  are independent. Our simulation studies show that CARS controls FDR under weak dependence. However, the result is based on very limited empirical studies, which lack theoretical support. An important direction is to develop new theory and methodology for the dependent case.
- (c) *How can we 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 distribution is calibrated with permutations. We conjecture that the CARS procedure, which requires an 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.
- (d) *How can we construct the auxiliary sequence in more general settings?* This paper focuses on the two-sample tests. It would be of interest to extend the methodology to simultaneous analysis-of-variance 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 dependence, 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.

- (e) *How can we summarize the auxiliary information in high dimensional settings?* The proposed CARS methodology requires the joint modelling 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.
- (f) *How can we 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 statistics*, in partial conjunction tests and replicability analysis all sequences are of equal importance, which poses new challenges for problem formulation and methodological development.

## Acknowledgements

We thank the referees and Research Section Committee for the thorough and useful comments which have greatly helped for improve the presentation of the paper. In particular, we are grateful for several excellent suggestions from the referees that have inspired our discussions on the conditional independence assumption, Tukey's procedures for combination, Brown's ancillarity paradox and the robustness of CARS when pooling non-informative auxiliary data. W. Sun thanks Ms Pallavi Basu from Tel Aviv University for helpful suggestions on theory. Tony Cai was supported in part by National Science Foundation grant DMS-1712735 and National Institutes of Health grant R01 CA127334. W. Sun was supported in part by National Science Foundation grants DMS-CAREER-1255406 and DMS-1712983.

## Appendix A: Proofs of main theorems

This section proves the main theorems. The proofs of other propositions are provided in the on-line supplementary material.

### A.1. Proof of theorem 1

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

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

#### A.1.1. Proof of part (a)

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

$$\mathbb{E}_{(\mathbf{T}_1, \mathbf{T}_2)} \left\{ \sum_{i=1}^m (T_{\text{OR}}^i - \alpha_t) \mathbb{1}(T_{\text{OR}}^i < t) \right\} = 0, \quad (\text{A.1})$$

where the subscript  $(\mathbf{T}_1, \mathbf{T}_2)$  indicates that the expectation is taken over the joint distribution of  $(\mathbf{T}_1, \mathbf{T}_2)$ . Equation (A.1) implies that  $\alpha_t < t$ ; otherwise all terms in the summation on its left-hand side would be either 0 or negative.

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

$$\begin{aligned} (T_{\text{OR}}^i - \alpha_2) \mathbb{1}(T_{\text{OR}}^i < t_2) &= (T_{\text{OR}}^i - \alpha_2) \mathbb{1}(T_{\text{OR}}^i < t_1) + (T_{\text{OR}}^i - \alpha_2) \mathbb{1}(t_1 \leq T_{\text{OR}}^i < t_2) \\ &\geq (T_{\text{OR}}^i - \alpha_1) \mathbb{1}(T_{\text{OR}}^i < t_1) + (\alpha_1 - \alpha_2) \mathbb{1}(T_{\text{OR}}^i < t_1) + (T_{\text{OR}}^i - \alpha_1) \mathbb{1}(t_1 \leq T_{\text{OR}}^i < t_2). \end{aligned}$$

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

$$\mathbb{E}_{T_1, T_2} \left\{ \sum_{i=1}^m (T_{\text{OR}}^i - \alpha_2) \mathbb{1}(T_{\text{OR}}^i < t_2) \right\} > 0. \quad (\text{A.2})$$

This inequality holds since

- (a)  $\mathbb{E}_{T_1, T_2} \left\{ \sum_{i=1}^m (T_{\text{OR}}^i - \alpha_1) \mathbb{1}(T_{\text{OR}}^i < t_1) \right\} = 0$
- (b)  $\mathbb{E}_{T_1, T_2} \left\{ \sum_{i=1}^m (\alpha_1 - \alpha_2) \mathbb{1}(T_{\text{OR}}^i < t_1) \right\} > 0$  and
- (c)  $\mathbb{E}_{T_1, T_2} \left\{ \sum_{i=1}^m (T_{\text{OR}}^i - \alpha_1) \mathbb{1}(t_1 \leq T_{\text{OR}}^i < t_2) \right\} > 0$ ,

which are respectively due to equation (A.1), the assumption that  $\alpha_1 > \alpha_2$  and the fact that  $\alpha_1 < t_1$ . However, inequality (A.2) is a contradiction to our definition of  $\alpha_2$ , which implies that  $\mathbb{E}_{T_1, T_2} \left\{ \sum_{i=1}^m (T_{\text{OR}}^i - \alpha_2) \mathbb{1}(T_{\text{OR}}^i < t_2) \right\} = 0$ . Hence we must have  $\alpha_1 \leq \alpha_2$ .

### A.1.2. Proof of part (b)

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

$$\begin{aligned} \mathbb{E}_{T_1, T_2} \left\{ \sum_{i=1}^m (T_{\text{OR}}^i - \alpha) \delta_{\text{OR}}^i \right\} &= 0, \\ \mathbb{E}_{T_1, T_2} \left\{ \sum_{i=1}^m (T_{\text{OR}}^i - \alpha) \delta_*^i \right\} &\leq 0. \end{aligned} \quad (\text{A.3})$$

Combining the two results in expression (A.3) we conclude that

$$\mathbb{E}_{T_1, T_2} \left\{ \sum_{i=1}^m (\delta_{\text{OR}}^i - \delta_*^i) (T_{\text{OR}}^i - \alpha) \right\} \geq 0. \quad (\text{A.4})$$

Next, consider a monotonic transformation of the oracle decision rule  $\delta_{\text{OR}}^i = \mathbb{1}(T_{\text{OR}}^i < t_{\text{OR}})$  via  $f(x) = (x - \alpha)/(1 - x)$  (note that the derivative  $f'(x) = (1 - \alpha)/(1 - x)^2 > 0$ ). The oracle decision rule is equivalent to

$$\delta_{\text{OR}}^i = I \left( \frac{T_{\text{OR}}^i - \alpha}{1 - T_{\text{OR}}^i} < \lambda_{\text{OR}} \right),$$

where  $\lambda_{\text{OR}} = (t_{\text{OR}} - \alpha)/(1 - t_{\text{OR}})$ . A useful fact is that  $\alpha < t_{\text{OR}} < 1$ . Hence  $\lambda_{\text{OR}} > 0$ .

Note that

- (a)  $(T_{\text{OR}}^i - \alpha) - \lambda_{\text{OR}}(1 - T_{\text{OR}}^i) < 0$  if  $\delta_{\text{OR}}^i > \delta_*^i$  and
- (b)  $(T_{\text{OR}}^i - \alpha) - \lambda_{\text{OR}}(1 - T_{\text{OR}}^i) > 0$  if  $\delta_{\text{OR}}^i < \delta_*^i$ .

Combining (a) and (b), we conclude that the following inequality holds for all  $i$ :  $(\delta_{\text{OR}}^i - \delta_*^i) \{(T_{\text{OR}}^i - \alpha) - \lambda_{\text{OR}}(1 - T_{\text{OR}}^i)\} \leq 0$ . Summing over  $i$  and taking expectations, we have

$$\mathbb{E}_{T_1, T_2} \left[ \sum_{i=1}^m (\delta_{\text{OR}}^i - \delta_*^i) \{(T_{\text{OR}}^i - \alpha) - \lambda_{\text{OR}}(1 - T_{\text{OR}}^i)\} \right] \leq 0. \quad (\text{A.5})$$

Combining inequalities (A.4) and (A.5) we have

$$\lambda_{\text{OR}} \mathbb{E}_{T_1, T_2} \left\{ \sum_{i=1}^m (\delta_{\text{OR}}^i - \delta_*^i) (1 - T_{\text{OR}}^i) \right\} \geq \mathbb{E}_{T_1, T_2} \left\{ \sum_{i=1}^m (\delta_{\text{OR}}^i - \delta_*^i) (T_{\text{OR}}^i - \alpha) \right\} \geq 0.$$

1 Finally, noting that  $\lambda_{\text{OR}} > 0$  and that ETP for a decision rule  $\delta = (\delta_1, \dots, \delta_m)$  is given by  $\mathbb{E}_{T_1, T_2} \{\sum_{i=1}^m \delta_i (1 - T_{\text{OR}}^i)\}$ , we conclude that  $\text{ETP}(\delta_{\text{OR}}) \geq \text{ETP}(\delta_*)$ .

4 **A.2. Proof of theorem 2**

5 We first provide a summary of useful notation.

- 6  
 7 (a)  $Q^\tau(t) = m^{-1} \sum_{i=1}^m (T_{\text{OR}}^{\tau, i} - \alpha) I(T_{\text{OR}}^{\tau, i} < t)$ .  
 8 (b)  $\hat{Q}^\tau(t) = m^{-1} \sum_{i=1}^m (\hat{T}_{\text{OR}}^{\tau, i} - \alpha) I(\hat{T}_{\text{OR}}^{\tau, i} < t)$ .  
 9 (c)  $Q_\infty^\tau(t) = \mathbb{E}\{(T_{\text{OR}}^\tau - \alpha) \mathbb{1}(T_{\text{OR}}^\tau < t)\}$ , where  $T_{\text{OR}}^\tau$  is a generic member from  $\{T_{\text{OR}}^i : 1 \leq i \leq m\}$ .

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

11 
$$12 \quad t_\infty^\tau = \sup\{t \in (0, 1) : Q_\infty^\tau(t) \leq 0\}.$$

13  
 14 **A.2.1. Proof of part (a)**

15 In proposition 5, we show that  $\delta_{\text{OR}}^\tau$  is conservative in mFDR-control. To establish the desired property in mFDR-control, we need to show only that  $\text{mFDR}(\delta_{\text{DD}}) = \text{mFDR}(\delta_{\text{OR}}^\tau) + o(1)$ .

16 Define a continuous version of  $Q^\tau(t)$  as follows. For  $T_{\text{OR}}^{\tau, (k)} < t \leq T_{\text{OR}}^{\tau, (k+1)}$ , let

17  
 18  
 19 
$$20 \quad Q_C^\tau(t) = \frac{t - T_{\text{OR}}^{\tau, (k)}}{T_{\text{OR}}^{\tau, (k+1)} - T_{\text{OR}}^{\tau, (k)}} Q_k^\tau + \frac{T_{\text{OR}}^{\tau, (k+1)} - t}{T_{\text{OR}}^{\tau, (k+1)} - T_{\text{OR}}^{\tau, (k)}} Q_{k+1}^\tau, \quad (\text{A.6})$$

21 where  $Q_k^\tau = Q^\tau(T_{\text{OR}}^{\tau, (k)})$ . It is easy to see that  $Q_C^\tau(t)$  is continuous and monotone. Hence the inverse of  $Q_C^\tau(t)$ , which is denoted  $Q_C^{\tau, -1}$ , is well defined. Moreover,  $Q_C^{\tau, -1}$  is continuous and monotone. We can similarly define a continuous version of  $\hat{Q}^\tau(t)$ , which is denoted by  $\hat{Q}_C^\tau(t)$ .  $\hat{Q}_C^\tau(t)$  is continuous and monotone; so is its inverse  $\hat{Q}_C^{\tau, -1}(\cdot)$ . By construction, we have  $\delta_{\text{OR}}^\tau = [I\{T_{\text{OR}}^{\tau, i} \leq Q_C^{\tau, -1}(0)\} : 1 \leq i \leq m]$  and  $\delta_{\text{DD}}^\tau = [I\{\hat{T}_{\text{OR}}^{\tau, i} \leq \hat{Q}_C^{\tau, -1}(0)\} : 1 \leq i \leq m]$ . We shall show that

22  
 23  
 24  
 25  
 26 
$$27 \quad Q_C^{\tau, -1}(0) \xrightarrow{P} t_\infty^\tau, \quad (\text{A.7a})$$

28 
$$29 \quad \hat{Q}_C^{\tau, -1}(0) \xrightarrow{P} t_\infty^\tau. \quad (\text{A.7b})$$

30 To show result (A.7a), note that the continuity of  $Q_C^{\tau, -1}(\cdot)$  implies that, for any  $\epsilon > 0$ , we can find  $\eta > 0$  such that  $|Q_C^{\tau, -1}(0) - Q_C^{\tau, -1}\{Q_C^\tau(t_\infty^\tau)\}| < \epsilon$  if  $|Q_C^\tau(t_\infty^\tau)| < \eta$ . Hence

31  
 32 
$$\mathbb{P}\{|Q_C^\tau(t_\infty^\tau) - \alpha| > \eta\} \geq \mathbb{P}[|Q_C^{\tau, -1}(\alpha) - Q_C^{\tau, -1}\{Q_C^\tau(t_\infty^\tau)\}| > \epsilon].$$

33 Next, by the weak law of large numbers  $Q_C^\tau(t) \xrightarrow{P} Q_\infty^\tau(t)$ . Noting that  $Q_C^\tau(t_\infty^\tau) = \alpha$ , we have  $\mathbb{P}\{|Q^\tau(t_\infty^\tau) - \alpha| > \eta\} \rightarrow 0$ . By the Markov inequality, we conclude that  $Q_C^{\tau, -1}(\alpha) \xrightarrow{P} Q_C^{\tau, -1}\{Q_C^\tau(t_\infty^\tau)\} = t_\infty^\tau$ .

34 Next we show result (A.7b). By inspecting the proof of result (A.7a), we need to show only that  $\hat{Q}_C^\tau(t) \xrightarrow{P} Q_\infty^\tau(t)$ . Denote a variable without index  $i$  (e.g.  $\hat{T}_{\text{OR}}^\tau$  and  $T_{\text{OR}}^\tau$ ) as a generic member from the sample. It follows from condition and the continuous mapping theorem that  $\hat{T}_{\text{OR}}^\tau \xrightarrow{P} T_{\text{OR}}^\tau$ . Note that both  $T_{\text{OR}}^\tau$  and  $\hat{T}_{\text{OR}}^\tau$  are bounded above by 1. It follows that  $\mathbb{E}(\hat{T}_{\text{OR}}^\tau - T_{\text{OR}}^\tau)^2 \rightarrow 0$ .

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

36  
 37  
 38  
 39  
 40 
$$41 \quad (\hat{U}_i - U_i)^2 = (\hat{T}_{\text{OR}}^\tau - T_{\text{OR}}^\tau)^2 \mathbb{1}(\hat{T}_{\text{OR}}^\tau \leq t, T_{\text{OR}}^\tau \leq t) + (\hat{T}_{\text{OR}}^\tau)^2 \mathbb{1}(\hat{T}_{\text{OR}}^\tau \leq t, T_{\text{OR}}^\tau > t) + (T_{\text{OR}}^\tau)^2 \mathbb{1}(\hat{T}_{\text{OR}}^\tau > t, T_{\text{OR}}^\tau \leq t) \\ 42 \quad = \text{I} + \text{II} + \text{III}.$$

43 The first term I equals  $o(1)$  because  $\mathbb{E}(\hat{T}_{\text{OR}}^\tau - T_{\text{OR}}^\tau)^2 \rightarrow 0$ . Let  $\eta > 0$ . Noting that  $T_{\text{OR}}^\tau$  is continuous and that  $\hat{T}_{\text{OR}}^\tau \xrightarrow{P} T_{\text{OR}}^\tau$ , we have

44  
 45 
$$46 \quad \mathbb{P}(\hat{T}_{\text{OR}}^\tau \leq t, T_{\text{OR}}^\tau > t) \leq \mathbb{P}\{T_{\text{OR}}^\tau \in (t, t + \eta)\} + \mathbb{P}(|\hat{T}_{\text{OR}}^\tau - T_{\text{OR}}^\tau| > \eta) \rightarrow 0.$$

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

Next we show that  $\hat{Q}^\tau(t) \rightarrow^P Q_\infty^\tau(t)$ . Noting that  $Q^\tau(t) \rightarrow^P Q_\infty^\tau(t)$ , we need to show only that  $\hat{Q}^\tau(t) \rightarrow^P Q^\tau(t)$ . 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 the Cauchy–Schwarz inequality, we have  $\mathbb{E}(\hat{U}_i - U_i)(\hat{U}_j - U_j) = o(1)$ . It follows that

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

Therefore  $E\{S_n - \mathbb{E}(S_n)/n\}^2 \rightarrow 0$ . 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  $\hat{Q}^\tau(t) \rightarrow^P Q_\infty^\tau(t)$ . By definition,  $|\hat{Q}_C^\tau(t) - \hat{Q}^\tau(t)| \leq m^{-1}$ . We claim that  $\hat{Q}_C^\tau(t) \rightarrow^P Q_\infty^\tau(t)$  and result (A.7b) follows.

According to results (A.7a) and (A.7b),  $\hat{Q}_C^{\tau,-1}(0) = Q_C^{\tau,-1}(0) + o_P(1)$ . The mFDR-levels of the testing procedures are

$$\text{mFDR}(\delta_{\text{OR}}^\tau) = \frac{\mathbb{P}_{H_0}\{T_{\text{OR}}^{\tau,i} < Q_C^{\tau,-1}(\alpha)\}}{\mathbb{P}\{T_{\text{OR}}^{\tau,i} < Q_C^{\tau,-1}(\alpha)\}}$$

and

$$\text{mFDR}(\delta_{\text{DD}}^\tau) = \frac{\mathbb{P}_{H_0}\{\hat{T}_{\text{OR}}^{\tau,i} < \hat{Q}_C^{\tau,-1}(\alpha)\}}{\mathbb{P}\{\hat{T}_{\text{OR}}^{\tau,i} < \hat{Q}_C^{\tau,-1}(\alpha)\}}.$$

The operation of our testing procedure implies that  $Q_C^{\tau,-1}(\alpha) \geq \alpha$ . It follows that  $\mathbb{P}\{T_{\text{OR}}^{\tau,i} < Q_C^{\tau,-1}(\alpha)\}$  is bounded away from zero. We conclude that  $\text{mFDR}(\delta_{\text{DD}}^\tau) = \text{mFDR}(\delta_{\text{OR}}^\tau) + o(1)$ .

The result on mFDR-control can be extended to FDR-control. The next proposition, which is proved in the on-line supplement material, first gives sufficient conditions under which the definitions of mFDR and FDR are asymptotically equivalent and then verifies that these conditions are fulfilled by the CARS procedure  $\delta_{\text{DD}}^\tau$ . It follows from proposition 7 that CARS controls FDR at level  $\alpha + o(1)$ .

*Proposition 7.*

- (a) Consider a general decision rule  $\delta$ . Let  $\mathcal{Y} = m^{-1} \sum_{i=1}^m \delta_i$ . Then  $\text{mFDR}(\delta) = \text{FDR}(\delta) + o(1)$  if
- (i)  $\mathbb{E}(\mathcal{Y}) \geq \eta$  for some  $\eta > 0$  and
  - (ii)  $\text{var}(\mathcal{Y}) = o(1)$ .
- (b) Conditions (i) and (ii) are fulfilled by the CARS procedure  $\delta_{\text{DD}}^\tau$ .

### A.2.2. Proof of part (b)

The CARS procedure utilizes  $\hat{q}^*$ , and the corresponding test statistic is  $\hat{T}_{\text{OR}}^{*,i}$ . It follows from conditions 1' and 2, and the continuous mapping theorem that  $\hat{T}_{\text{OR}}^* \rightarrow^P T_{\text{OR}}$ . Denote  $Q_{\text{OR}}(t)$  the oracle mFDR-function and  $t_{\text{OR}}$  the oracle threshold. Then

$$Q_{\text{OR}}(t) = \mathbb{E}\{(T_{\text{OR}} - \alpha) \mathbb{1}(T_{\text{OR}} < t)\},$$

$$t_{\text{OR}} = \sup\{t \in (0, 1) : Q_{\text{OR}}(t) \leq 0\}.$$

Define  $\hat{Q}^*(t) = m^{-1} \sum_{i=1}^m \hat{T}_{\text{OR}}^{*,i} \mathbb{1}(\hat{T}_{\text{OR}}^{*,i} < t)$ . Similarly to equation (A.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 is its inverse  $\hat{Q}_C^{*,,-1}(t)$ . The CARS procedure is given by  $\delta_{\text{DD}}^* = [\mathbb{1}\{\hat{T}_{\text{OR}}^{*,i} \leq \hat{Q}_C^{*,,-1}(\alpha)\} : 1 \leq i \leq m]$ . Following the steps in the proof of part (a) we can show that

$$\hat{Q}_C^*(t) \xrightarrow{P} Q_{\text{OR}}(t), \tag{A.8}$$

$$\hat{Q}_C^{*,,-1}(t) \xrightarrow{P} t_{\text{OR}}.$$

The operation of CARS implies that  $Q_C^{*,,-1}(0) \geq \alpha$  (thus the denominator of mFDR is bounded away from zero). Note that  $\text{mFDR}(\delta_{\text{OR}}^*) = \alpha$ , we have  $\text{mFDR}(\delta_{\text{DD}}^*) = \alpha + o(1)$ . Next, we consider ETP. It follows from  $\hat{T}_{\text{OR}}^* \rightarrow^P T_{\text{OR}}$  and expression (A.8) that

$$\frac{\text{ETP}(\delta_{\text{DD}}^*)}{\text{ETP}(\delta_{\text{OR}})} = \frac{\mathbb{P}_{H_1}\{\hat{T}_{\text{OR}}^* < \hat{Q}_{\text{C}}^{*-1}(\alpha)\}}{\mathbb{P}_{H_1}(T_{\text{OR}} < t_{\text{OR}})} = 1 + o(1).$$

## References

- Barber, R. F. and Ramdas, A. (2017) The  $p$ -filter: multilayer false discovery rate control for grouped hypotheses. *J. R. Statist. Soc. B*, **79**, 1247–1268.
- Basu, P., Cai, T. T., Das, K. and Sun, W. (2017) Weighted false discovery control in large-scale multiple testing. *J. Am. Statist. Ass.*, to be published.
- Benjamini, Y. and Heller, R. (2008) Screening for partial conjunction hypotheses. *Biometrics*, **64**, 1215–1222.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.
- Benjamini, Y. and Hochberg, Y. (1997) Multiple hypotheses testing with weights. *Scand. J. Statist.*, **24**, 407–418.
- Benjamini, Y. and Hochberg, Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, **25**, 60–83.
- Boca, S. M. and Leek, J. T. (2017) A regression framework for the proportion of true null hypotheses. *Preprint bioRxiv 035675*.
- Bourgon, R., Gentleman, R. and Huber, W. (2010) Independent filtering increases detection power for high-throughput experiments. *Proc. Natn. Acad. Sci USA*, **107**, 9546–9551.
- Brown, L. D. (1990) An ancillarity paradox which appears in multiple linear regression. *Ann. Statist.*, **18**, 471–493.
- Cai, T. T. and Jin, J. (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 Sun, W. (2009) Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. *J. Am. Statist. Ass.*, **104**, 1467–1481.
- Cai, T. T. and Wu, Y. (2014) Optimal detection of sparse mixtures against a given null distribution. *IEEE Trans. Inform. Theory*, **60**, 2217–2232.
- Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K. et al. (2005) A network-based analysis of systemic inflammation in humans. *Nature*, **437**, 1032–1037.
- Cao, H., Sun, W. and Kosorok, M. R. (2013) The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing. *Biometrika*, **100**, 495–502.
- Donoho, D. and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 962–994.
- Du, L. and Zhang, C. (2014) Single-index modulated multiple testing. *Ann. Statist.*, **42**, 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. *J. Am. Statist. Ass.*, **99**, 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. Statist.*, **2**, 197–223.
- Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Ass.*, **96**, 1151–1160.
- Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G. and Kong, A. (2008) Unsupervised empirical bayesian multiple testing with external covariates. *Ann. Appl. Statist.*, **2**, 714–735.
- Foster, D. P. and George, E. I. (1996) A simple ancillarity paradox. *Scand. J. Statist.*, **23**, 233–242.
- Genovese, C. and Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc. B*, **64**, 499–517.
- Genovese, C. and Wasserman, L. (2004) A stochastic process approach to false discovery control. *Ann. Statist.*, **32**, 1035–1061.
- Haupt, J., Castro, R. M. and Nowak, R. (2011) Distilled sensing: adaptive sampling for sparse detection and estimation. *IEEE Trans. Inform. Theory*, **57**, 6222–6235.
- Heller, R., Bogomolov, M. and Benjamini, Y. (2014) Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proc. Natn. Acad. Sci. USA*, **111**, 16262–16267.
- Heller, R. and Yekutieli, D. (2014) Replicability analysis for genome-wide association studies. *Ann. Appl. Statist.*, **8**, 481–498.
- Hu, J. X., Zhao, H. and Zhou, H. H. (2010) False discovery rate control with groups. *J. Am. Statist. Ass.*, **105**, 1215–1227.
- James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 361–379. Berkeley: University of California Press.
- Jin, J. and Cai, T. T. (2007) Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Am. Statist. Ass.*, **102**, 495–506.
- Langaas, M., Lindqvist, B. H. and Ferkingstad, E. (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Statist. Soc. B*, **67**, 555–572.

- 1 Law, N. M., Kulkarni, S. R., Dekany, R. G., Ofek, E. O., Quimby, R. M., Nugent, P. E., Surace, J., Grillmair, C. C.,  
 2 Bloom, J. S., Kasliwal, M. M. *et al.* (2009) The Palomar Transient Factory: system overview, performance, and  
 3 first results. *Publ. Astron. Soc. Pacif.*, **121**, 1395.
- 4 Lehmann, E. L. and Casella, G. (2006) *Theory of Point Estimation*. New York: Springer Science and Business  
 5 Media.
- 6 Li, A. and Barber, R. F. (2016) Multiple testing with the structure adaptive benjamini-hochberg algorithm.  
 7 *Preprint arXiv:1606.07926*.
- 8 Liu, W. (2014) Incorporation of sparsity information in large-scale multiple two-sample  $t$  tests. *Preprint*  
 9 *arXiv:1410.4282*.
- 10 Liu, Y., Sarkar, S. K. and Zhao, Z. (2016) A new approach to multiple testing of grouped hypotheses. *J. Statist.*  
 11 *Planng Inf.*, **179**, 1–14.
- 12 Meinshausen, N. and Rice, J. (2006) Estimating the proportion of false null hypotheses among a large number of  
 13 independently tested hypotheses. *Ann. Statist.*, **34**, 373–393.
- 14 Neuvial, P. (2013) Asymptotic results on adaptive false discovery rate controlling procedures based on kernel  
 15 estimators. *J. Mach. Learn. Res.*, **14**, 1423–1459.
- 16 Nugent, P. E., Sullivan, M., Cenko, S. B., Thomas, R. C., Kasen, D., Howell, D. A., Bersier, D., Bloom, J. S.,  
 17 Kulkarni, S. R., Kandrashoff, M. T., Filippenko, A. V., Silverman, J. M., Marcy, J. M., Howard, A. W., Isaacson,  
 18 H. T., Maguire, K., Suzuki, N., Tarlton, J. E., Pan, Y.-C., Bildsten, L., Fulton, B. J., Parrent, J. T., Sand, D.,  
 19 Podsiadlowski, P., Bianco, F. B., Dilday, B., Graham, M. L., Lyman, J., James, P., Kasliwal, M. M., Law, N.  
 20 M., Quimby, R. M., Hook, I. M., Walker, E. S., Mazzali, P., Pian, E., Ofek, E. O., Gal-Yam, A. and Poznanski,  
 21 D. (2011) Supernova SN 2011fe from an exploding carbon-oxygen white dwarf star. *Nature*, **480**, 344–347.
- 22 Reiner-Benaim, A., Yekutieli, D., Letwin, N. E., Elmer, G. I., Lee, N. H., Kafkafi, N. and Benjamini, Y. (2007)  
 23 Associating quantitative behavioral traits with gene expression in the brain: searching for diamonds in the hay.  
 24 *Bioinformatics*, **23**, 2239–2246.
- 25 Roeder, K. and Wasserman, L. (2009) Genome-wide significance levels and weighted hypothesis testing. *Statist.*  
 26 *Sci.*, **24**, 398.
- 27 Roquain, E. and Van De Wiel, M. A. (2009) Optimal weighting for false discovery rate control. *Electron. J. Statist.*,  
 28 **3**, 678–711.
- 29 Rubin, D., Dudoit, S. and van der Laan, M. (2006) A method to increase the power of multiple testing procedures  
 30 through sample splitting. *Statist. Appl. Genet. Molec. Biol.*, **5**, article 19.
- 31 Sarkar, S. K. (2002) Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, **30**,  
 32 239–257.
- 33 Sarkar, S. K. and Zhao, Z. (2017) Local false discovery rate based methods for multiple testing of one-way  
 34 classified hypotheses. *Preprint arXiv:1712.05014*.
- 35 Schweder, T. and Spjøtvoll, E. (1982) Plots of  $p$ -values to evaluate many tests simultaneously. *Biometrika*, **69**,  
 36 493–502.
- 37 Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P. and Kass, R. E. (2015) False discovery rate regression: an  
 38 application to neural synchrony detection in primary visual cortex. *J. Am. Statist. Ass.*, **110**, 459–471.
- 39 Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. Boca Raton: CRC Press.
- 40 Skol, A. D., Scott, L. J., Abecasis, G. R. and Boehnke, M. (2006) Joint analysis is more efficient than replication-  
 41 based analysis for two-stage genome-wide association studies. *Nat. Genet.*, **38**, 209–213.
- 42 Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. B*, **64**, 479–498.
- 43 Sun, W. and Cai, T. T. (2007) Oracle and adaptive compound decision rules for false discovery rate control. *J. Am.*  
 44 *Statist. Ass.*, **102**, 901–912.
- 45 Sun, W. and Wei, Z. (2011) Large-scale multiple testing for pattern identification, with applications to time-course  
 46 microarray experiments. *J. Am. Statist. Ass.*, **106**, 73–88.
- 47 Taylor, J., Tibshirani, R. and Efron, B. (2005) The “miss rate” for the analysis of gene expression data. *Biostatistics*,  
 48 **6**, 111–117.
- 49 Tukey, J. W. (1994) *The Collected Works of John W. Tukey*, vol. 3. New York: Taylor and Francis.
- 50 Wand, M. and Jones, M. (1995) *Kernel Smoothing*. London: Chapman and Hall.
- 51 Wasserman, L. and Roeder, K. (2009) High-dimensional variable selection. *Ann. Statist.*, **37**, 2178–2201.
- 52 Zablocki, R. W., Schork, A. J., Levine, R. A., Andreassen, O. A., Dale, A. M. and Thompson, W. K. (2014)  
 53 Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics*, **30**, 2098–  
 54 2104.
- 55 Zehetmayer, S., Bauer, P. and Posch, M. (2005) Two-stage designs for experiments with a large number of hy-  
 56 potheses. *Bioinformatics*, **21**, 3771–3777.
- 57 Zehetmayer, S., Bauer, P. and Posch, M. (2008) Optimized multi-stage designs controlling the false discovery or  
 58 the family-wise error rate. *Statist Med.*, **27**, 4145–4160.

#### Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary material for “CARS: covariate assisted ranking and screening for large-scale two-sample inference”’.