ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)
81, *Part* 2, *pp.* 187–234

Covariate-assisted ranking and screening for large-scale two-sample inference

T. Tony Cai,

University of Pennsylvania, Philadelphia, USA

and Wenguang Sun 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.

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

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

 1369–7412/19/81187

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 $\boldsymbol{\mu}_x = \mathbb{E}(\mathbf{X}) = (\mu_{x1}, \ldots, \mu_{xm})^{\mathrm{T}}$ and $\boldsymbol{\mu}_y = \mathbb{E}(\mathbf{Y}) = (\mu_{y1}, \ldots, \mu_{ym})^{\mathrm{T}}$. A classical formulation for identifying differential features is to carry out m two-sample tests:

$$H_{i,0} : \mu_{xi} = \mu_{yi} \quad \textit{versus} \quad H_{i,1} : \mu_{xi} \neq \mu_{yi}, \qquad 1 \leqslant i \leqslant m. \tag{1.1}$$

Suppose that we have collected two random samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}\}$ 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 $\{X_{ij}: 1 \leqslant j \leqslant n_x\}$ and $\{Y_{ik}: 1 \leqslant k \leqslant n_y\}$, $i = 1, \ldots, m$, are repeated measurements of generic independent random variables X_i and Y_i respectively. Let $\boldsymbol{\beta}_0 = (\beta_{0i}: 1 \leqslant i \leqslant m)$ be a latent baseline vector which itself is sparse (including the special case where $\beta_{0i} = 0$ for all i). Consider the hierarchical model

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

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

Remark 1. Model (2.1) can be applied to scenarios with non-sparse $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ when some baseline measurements are available. See section A.8 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 μ_{xi} and μ_{yi} are estimated by $\bar{X}_i = n_x^{-1}\Sigma_{j=1}^{n_x} X_{ij}$ and $\bar{Y}_i = n_y^{-1}\Sigma_{k=1}^{n_y} Y_{ik}$ respectively.

The two-sample inference problem is concerned with the simultaneous testing of m hypotheses $H_{i,0} : \mu_{xi} = \mu_{yi}$ *versus* $H_{i,1} : \mu_{xi} \neq \mu_{yi}$, $i = 1, \ldots, m$. Let $\mathbb{I}(\cdot)$ be an indicator function. Let T_{1i} and T_{2i} be summary statistics that contain the information about $\theta_{1i} = \mathbb{I}(\mu_{xi} \neq \mu_{yi})$ (support of mean difference) and $\theta_{2i} = \mathbb{I}(\mu_{xi} \neq 0 \text{ or } \mu_{yi} \neq 0)$ (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 θ_{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 $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are sparse, exploiting this structural information would improve the efficiency of tests. To see this, note that the continuity of μ_{xi} and μ_{yi} implies that, with probability 1, θ_{1i} and θ_{2i} obey the logical relationship

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

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

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

2.1. Constructing the primary and auxiliary statistics

A key step in our formulation is to construct a pair of statistics (T_{1i}, T_{2i}) such that

(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 θ_{1i}; hence $\bar{X}_i - \bar{Y}_i$ should be incorporated in its expression. Second, to capture the information on the union support θ_{2i}, we propose to use the weighted sum $\bar{X}_i + \kappa_i \bar{Y}_i$, where $\kappa_i > 0$ is the weight to be specified later. Under the normality assumption, the covariance of $\bar{X}_i - \bar{Y}_i$ and $\bar{X}_i + \kappa_i \bar{Y}_i$ is given by $\sigma_{xi}^2/n_x - \kappa_i \sigma_{yi}^2/n_y$. This motivates us to choose the weight $\kappa_i^* = \gamma_y \sigma_{xi}^2/(\gamma_x \sigma_{yi}^2)$, which leads to zero correlation, 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:

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

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

2.2. A bivariate random-mixture model

We develop a bivariate model to describe the joint distribution of T_{1i} and T_{2i}. Let $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i})$.

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

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

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

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:

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

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

Proposition 1. The conditional independence (2.5) implies that

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

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

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

2.3. Problem formulation

Our goal is to make inference on $\theta_{1i}=\mathbb{I}(\mu_{xi}\neq\mu_{yi})$, $1\leqslant i\leqslant m$, 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 $\{(T_{1i},T_{2i}):1\leqslant i\leqslant m\}$ instead of a single vector $\{T_{1i}:1\leqslant i\leqslant m\}$. This new problem can be recast and solved in the framework of multiple testing with a covariate: T_{1i} is viewed as the primary statistic for assessing significance, and T_{2i} is viewed as a covariate to assist inference by providing supporting information.

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

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

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

In large-scale testing problems, the false discovery rate FDR (Benjamini and Hochberg, 1995) has been widely used to control the inflation of type I errors. For a given decision rule

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

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

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

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

Genovese and Wasserman (2002) showed that $\mathrm{mFDR} = \mathrm{FDR} + O(m^{-1/2})$ 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 $\mathrm{ETP}_{\boldsymbol{\delta}} = \mathbb{E}(\sum_{i=1}^{m} \theta_{1i}\delta_i)$. 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 $f_{j\cdot} = (1-\pi_j) f_{j0} + \pi_j f_{j1}$, where $\pi_j = \mathbb{P}(\theta_{ji} = 1)$ and $f_{j0} = f(t_{ji}|\theta_{ji} = 0)$ and $f_{j1} = f(t_{ji}|\theta_{ji} = 1)$ are the conditional densities for T_{ji} 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

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

where subscript '1' indicates a quantity 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

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

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

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

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

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

Define the oracle statistic

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

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

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

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

$$\delta^i_{\mathrm{OR}} = \mathbb{I}(T^i_{\mathrm{OR}} < \lambda_{\mathrm{OR}}). \tag{3.5}$$

Then $\boldsymbol{\delta}_{\mathrm{OR}}$ is optimal in the sense that $\mathrm{ETP}_{\boldsymbol{\delta}} \leqslant \mathrm{ETP}_{\boldsymbol{\delta}_{\mathrm{OR}}}$ for any $\boldsymbol{\delta}$ 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 $\mathrm{mFDR}_{\boldsymbol{\delta}} \leqslant \alpha$.

Remark 2. The oracle statistic T^i_{OR} 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_{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: $n_x = n_y$, $\sigma^2_{xi} = \sigma^2_{yi}$ and $\mu_{xi} = -\mu_{yi}$. In section A.9 of the on-line supplementary material, we show that under the above conditions T^i_{OR} 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 that the power will unlikely be decreased by pooling the auxiliary information in T_{2i}. Further numerical evidence is provided in Section 5.4.

The oracle rule (3.5) motivates us to consider a stepwise procedure that operates in two steps: ranking and thresholding. The ranking step orders all hypotheses from the most significant to the least significant according to T_{OR}, and the thresholding step identifies the largest threshold along the ranking subject to the constraint on FDR. Specifically, denote $T^{(1)}_{\mathrm{OR}} \leqslant \ldots \leqslant T^{(m)}_{\mathrm{OR}}$

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

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

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_{OR} via screening

The oracle statistic T_{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 $T_{\mathrm{OR}}^{\tau,i}$ that involves only quantities that can be well estimated from data. The new statistic provides a good approximation to T_{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 $q^*(t_2) = \lim_{h \to 0} \mathbb{E}\{Q^*(t_2, h)\}/h$. The counting task is difficult as we do not know the value of θ_{1i}. Our idea is 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 τ, say $\tau = 0.9$, we would reasonably predict that $\theta_{1i} = 0$ if $P_i > \tau$, as most likely large p-values should be from the null. Hence we may count those T_{2i} with large p-values:

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

The adjustment $1 - \tau$ in the denominator comes from the fact that we have 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^{τ} to replace Q^*, a sensible approximation of $q^*(t_2)$ would be

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

Intuitively, a large τ 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 τ (proposition 4). However, a larger τ would increase the variability of our proposed estimator (as we have fewer samples to construct the estimator), affecting the testing procedure adversely. The bias–variance 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^i_{OR}:

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

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

Proposition 2.

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

3.3. Estimation of the test statistic

We now turn to the estimation of $T^{\tau,i}_{\mathrm{OR}}$. 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):

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

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:

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

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

Proposition 3. Consider $\hat q^\tau$ and q^τ respectively defined in 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 $\int K(t)=1$, $\int tK(t)\mathrm{d}t=0$ and $\int t^2K(t)\mathrm{d}t<\infty$, and
(c) $f^{(2)}_{2\cdot}(t_2|\tau)=\int_{t_1\in\mathcal{A}_\tau}f^{(2)}_{2\cdot}(t_2|t_1)\,f_{1\cdot}(t_1)\mathrm{d}t_1$ 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

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

Combining these results, we propose to estimate T^τ_{OR} by the statistic

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

where $\hat q^\tau(t_2)$ and $\hat f(t_1,t_2)$ are respectively given in 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^τ_{OR} defined in equation (3.9).

(a) Denote $B_q(\tau) = \int |q^\tau(t_2) - q^*(t_2)| \mathrm{d}t_2$ the total approximation error. If $G_1(\cdot)$ is concave, then $B_q(\tau)$ decreases in τ.
(b) If $\lim_{x\uparrow 1} g_1(x) = 0$, then $\lim_{\tau\uparrow 1} q^\tau(t_2) = q^*(t_2)$.

Remark 5. The concavity assumption (or the more general monotone likelihood ratio condition) 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 (Neuvial, 2013). It would be desirable to develop a more general condition in future work.

It follows from proposition 4 that a large τ is helpful to reduce the bias and the bias converges to 0 when $\tau \to 1$. However, a large τ 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 τs, say $\{\tau_1, \ldots, \tau_k\}$, and then to use a smoothing method to obtain the limiting value of $\hat{q}^\tau$ when $\tau \to 1$. This approach aims to borrow strength from the entire sample to minimize the bias without blowing up the variance. Specifically, let $\tau_0 < \tau_1 < \ldots < \tau_k$ be ordered and equally spaced points in the interval $(0, 1)$. Denote $\hat{q}^{\tau_j}(t_2)$ the estimates from equation (3.11), $j = 1, \ldots, k$. We propose to obtain the local linear kernel estimator $\hat{q}^*(t_2) \equiv \hat{q}\{\tau = 1; \hat{q}^{\tau_1}(t_2), \ldots, \hat{q}^{\tau_k}(t_2)\}$ as the height of the fit $\hat{\beta}_0$, where $(\hat{\beta}_0, \hat{\beta}_1)$ minimizes the weighted kernel least squares $\Sigma_{j=1}^k (\hat{q}^{\tau_j} - \beta_0 - \beta_1 \tau_j)^2 K_{h_\tau}(\tau_j - \tau_k)$. For a given integer r, denote $\hat{s}_r = k^{-1} \Sigma_{j=1}^k (\tau_j - 1)^r K_{h_\tau}(\tau_j - \tau_k)$. It can be shown that (e.g. Wand and Jones (1995), page 119)

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

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

Proposition 5. Consider $\hat{q}^\tau$ and $\hat{q}^*$ that are respectively defined in equations (3.11) and (3.13). Denote $q^{\tau_k,(2)}(t_2) = (\mathrm{d}/\mathrm{d}\tau)^2 q^{\tau,(2)}(t_2)|_{\tau=\tau_k}$. 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_τ is a sequence satisfying $h_\tau \to 0$ and $kh_\tau \to \infty$ as $k \to \infty$.

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

$$\mathbb{E}\|\hat{q}^* - q^*\|^2 = \mathbb{E}\left[\int \{\hat{q}^*(x) - q^*(x)\}^2 \mathrm{d}x\right] \to 0 \qquad \text{when } (m,k) \to 0. \tag{3.14}$$

3.5. The covariate-assisted ranking and screening procedure

The estimated statistics $\hat{T}^{\tau,i}_{\mathrm{OR}}$ will be used as a significance index to rank the relative importance of all hypotheses. Motivated by the stepwise algorithm (3.6), we propose the following CARS procedure (procedure 1).

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

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

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

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

Condition 2. $\mathbb{E}\|\hat{f} - f\|^2 = \mathbb{E}\big[\int\int \{\hat{f}(t_1,t_2) - f(t_1,t_2)\}^2 \mathrm{d}t_1 \mathrm{d}t_2\big] \to 0$.

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 $\hat{T}^{\tau}_{\mathrm{OR}}$, then the FDR-level of the CARS procedure is $\alpha + o(1)$. Moreover, denote $\mathrm{ETP}_{\mathrm{CARS}}$ and $\mathrm{ETP}_{\mathrm{OR}}$ the ETP-levels of CARS and the oracle procedure respectively. Then we have $\mathrm{ETP}_{\mathrm{CARS}}/\mathrm{ETP}_{\mathrm{OR}} = 1 + o(1)$.

3.6. Case with unknown variances and non-Gaussian errors

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

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

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

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

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

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

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

4. Extensions and connections with existing work

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

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

4.1. A general bivariate model

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

Theorem 3. Define the oracle statistic under the general model

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

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

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

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

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

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

for $k=1,\ldots,K$, where π_1^k is the proportion of non-null cases in group k, f_{10}^k and f_{11}^k are the null and non-null densities of T_i and $f_1^k=(1-\pi_1^k)f_{10}^k+\pi_1^k f_{11}^k$ is the mixture density for all observations in group k. The model allows the conditional distributions in 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) $\mathrm{CLfdr}_i=(1-\pi_1^k)f_{10}^k(t_i)/f_1^k(t_i)$ for $S_i=k$, $k=1,\ldots,K$. 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:

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

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

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

4.3. Discretization and uncorrelated screening

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

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

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

4.4. The 'pooling-within' strategy for information integration

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

multiple-testing fields, has led to some impactful theories and methodologies exemplified by the James–Stein 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.

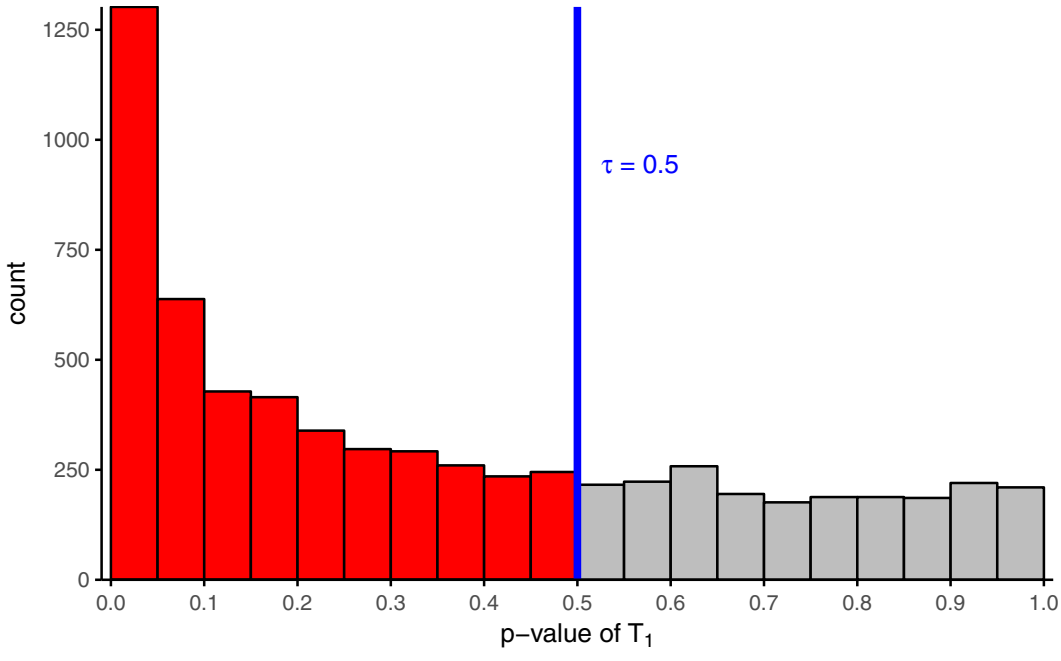
Our work differs in several ways from existing works on multiple testing with covariates (Ferkingstad *et al.*, 2008; Zablocki *et al.*, 2014; Scott *et al.*, 2015). First, the covariate in other works is collected *externally* from other data sources, whereas the auxiliary information in our work is gathered *internally* within the primary data set. Second, in other works it has been assumed that the null density would not be affected by the external covariate. However, the assumption should be scrutinized in practice as it may not always hold. Under our testing framework, the requirement of a fixed null density is formalized as the conditional independence between the primary and auxiliary statistics. The conditional independence is proposed as a principle for information extraction and is automatically fulfilled by our approach to constructing the auxiliary sequence.

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

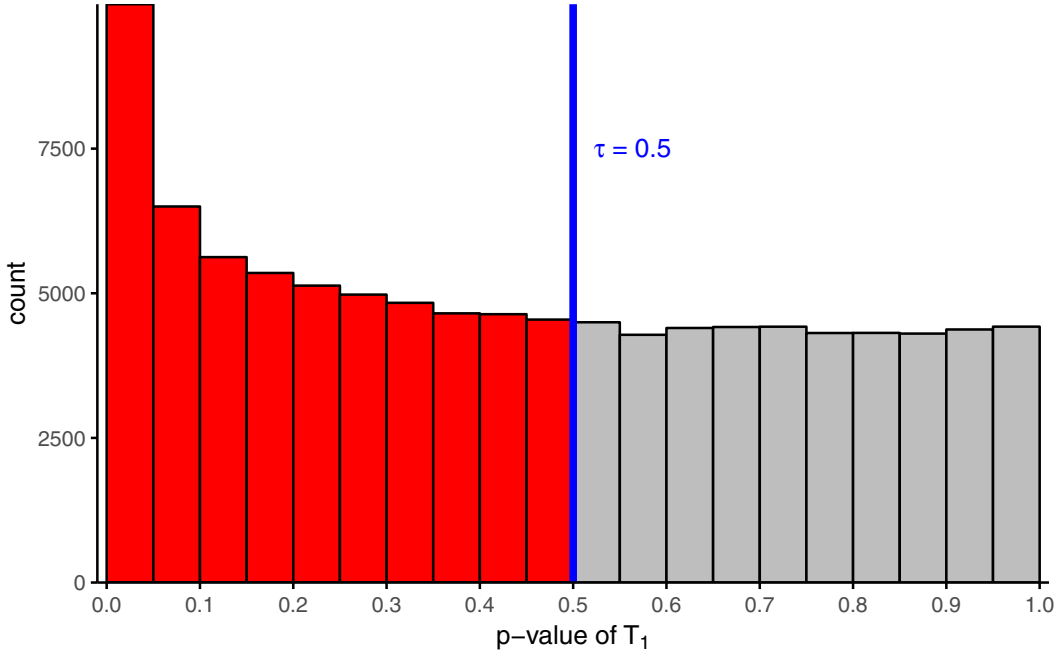
4.5. Capturing sparsity information via screening

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 (Ferkingstad *et al.*, 2008; Cai and Sun, 2009; Hu *et al.*, 2010). Estimating $\pi_1(s)$ with a continuous covariate is a challenging problem. Most existing works (Zablocki *et al.*, 2014; Scott *et al.*, 2015) employ Bayesian computational algorithms that do not provide theoretical guarantees. Notable progress has been made by Boca and Leek (2017). However, their theory requires a correct specification of the underlying regression model, which cannot be checked in practice. Next we discuss how the screening idea in Sections 3.3 and 3.4 can be extended to derive a simple and elegant non-parametric estimator of $\pi_1(s)$.

Fig. 1 gives a graphical illustration of the 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 τ 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 π_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 τ can be approximated as $\mathbb{E}\{Q(\tau)\} = m(1-\pi_1)(1-\tau)$. Setting the expected and actual counts equal, we obtain $\hat{\pi}_1^\tau = 1 - Q(\tau)/\{m(1-\tau)\}$.

Next we consider the conditional proportion $\pi_1(s) = P(\theta_i = 1 | S_i = s)$. Assume that $\pi_1(s)$ and $f(s)$, the density of S_i, are constants in a small 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 τ. Taking the limit, we obtain $\pi_1(s) = 1 - f(s)^{-1} \lim_{h \to 0} Q^\tau(s, h)/h = 1 - q^\tau(s)/f(s)$. Finally, utilizing the screening approach (3.11), we propose the following non-parametric smoothing estimator

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

Choosing tuning parameter τ is an important issue but has gone beyond the scope of the current work; see Storey (2002) 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 and 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 data. We compare the oracle and data-driven CARS procedures, respectively denoted by OR and DD, with existing methods, including the Benjamini–Hochberg procedure BH (Benjamini and Hochberg, 1995), the adaptive z-value procedure AZ (Sun and Cai, 2007) and the US procedure (Liu, 2014). We first describe the implementation of CARS in Section 5.1. Sections 5.2 and 5.3 respectively consider

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

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

5.1. The implementation and R package CARS

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

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

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

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

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

The R package `np` is used to choose the bandwidths h_1 and h_2 in equation (3.10). We have adopted two strategies to improve the performance. First, the bandwidths h_1 and h_2 are chosen based on the *normal reference rule* restricted to the samples with $\mathrm{Lfdr}_1 < 0.5$. The restriction

leads to more informative bandwidths as this subset is the more relevant part of the sample for the multiple-testing problem. Second, the same h_2 has been used in 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 I: known variances

Consider model (2.1). Denote $\boldsymbol{\mu}_{x,i_1:i_2} = (\mu_{x,i_1}, \ldots, \mu_{x,i_2})$ and $\boldsymbol{\mu}_{y,i_1:i_2} = (\mu_{y,i_1}, \ldots, \mu_{y,i_2})$ the vectors of consecutive observations from i_1 to i_2. The two original samples are denoted $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n_x}\}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_{n_y}\}$, 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 μ_0 from 3.5 to 5.

We can see that the CARS procedure is more powerful than conventional univariate methods such BH and AZ, and is superior 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 Fig. 2). 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 μ_0 corresponds to a larger difference in effect sizes).

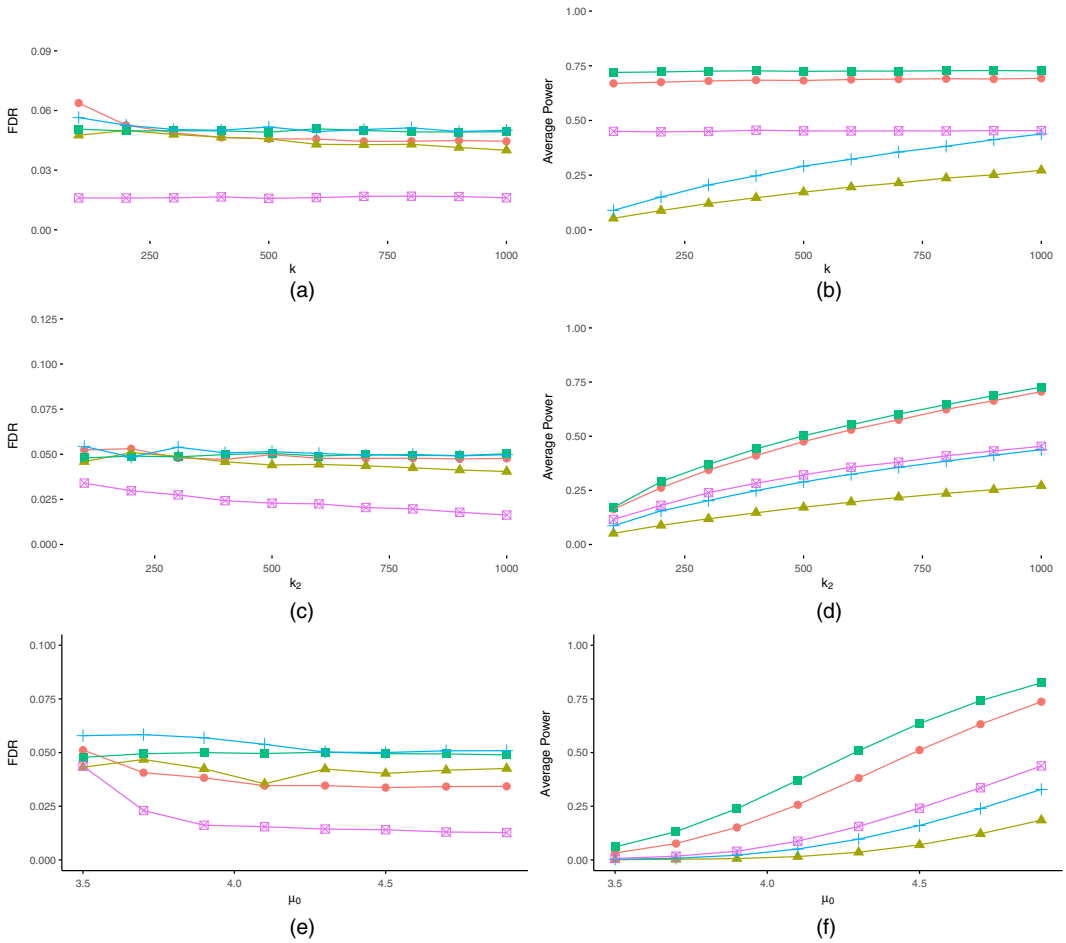


Fig. 2. Two-sample tests with known variances: FDR and average power (ETP divided by the number of non-nulls) are plotted against (a), (b) varied non-null proportions (setting 1), (c), (d) conditional proportions (setting 2) and (e), (f) effect sizes (setting 3) (●, DD; ▲, BH; ■, OR; +, AZ; ⊠, US)

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 ϵ_{xij} and ϵ_{yik} from a t-distribution with degrees of freedom $\mathrm{df}=4$ and $\mathrm{df}=5$ respectively. Finally, we vary k from 1 to 200 to investigate the performance of CARS under various sparsity regimes. The modified density estimator $\hat{f}^{\mathcal{U}}(t_1,t_2)$, 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.

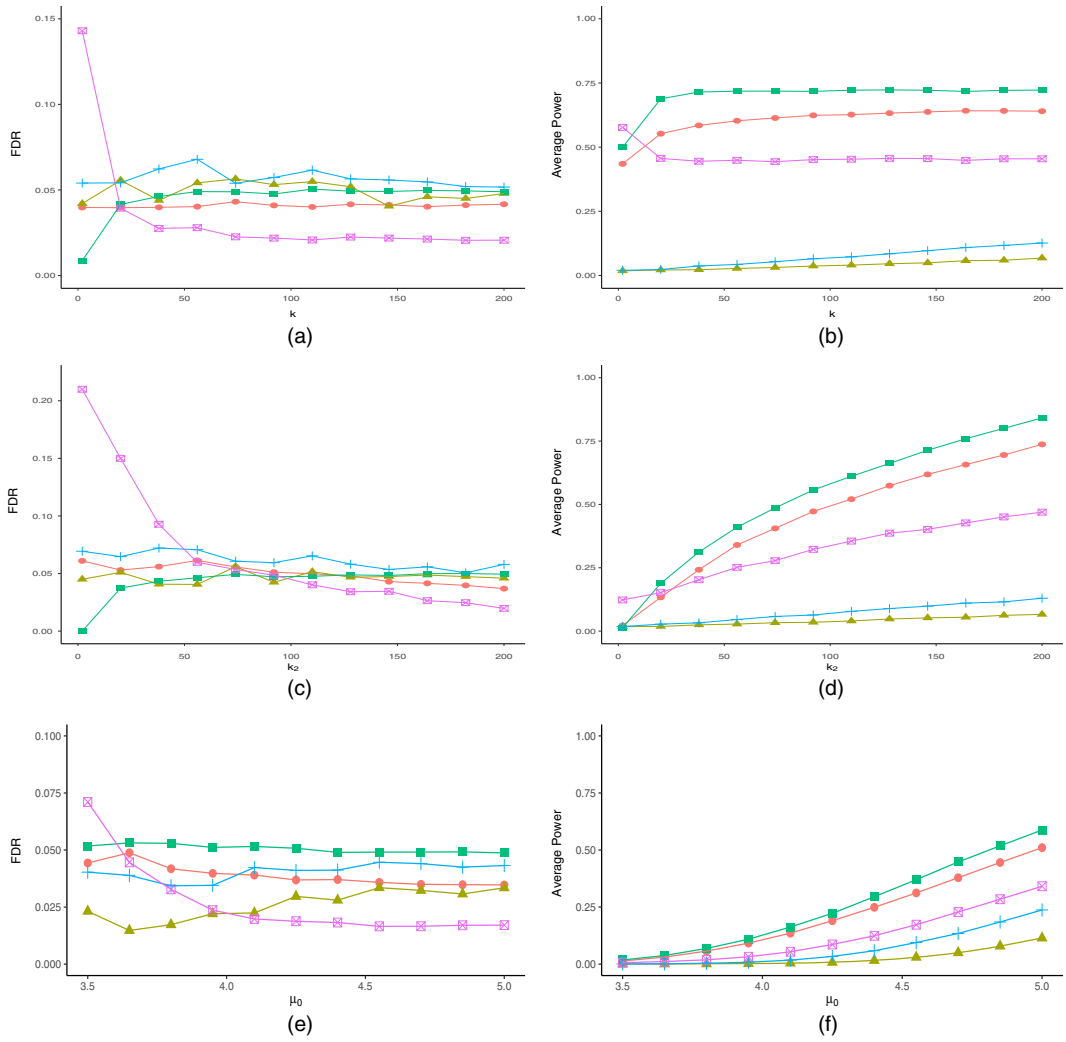


Fig. 3. Two-sample tests with unknown variances and non-Gaussian errors: FDR and MDR are plotted against (a), (b) varied non-null proportions (setting 1), (c), (d) conditional proportions and (e), (f) effect sizes (setting 3) (●, DD; ▲, BH; ■, OR; +, AZ; ⊠, US)

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

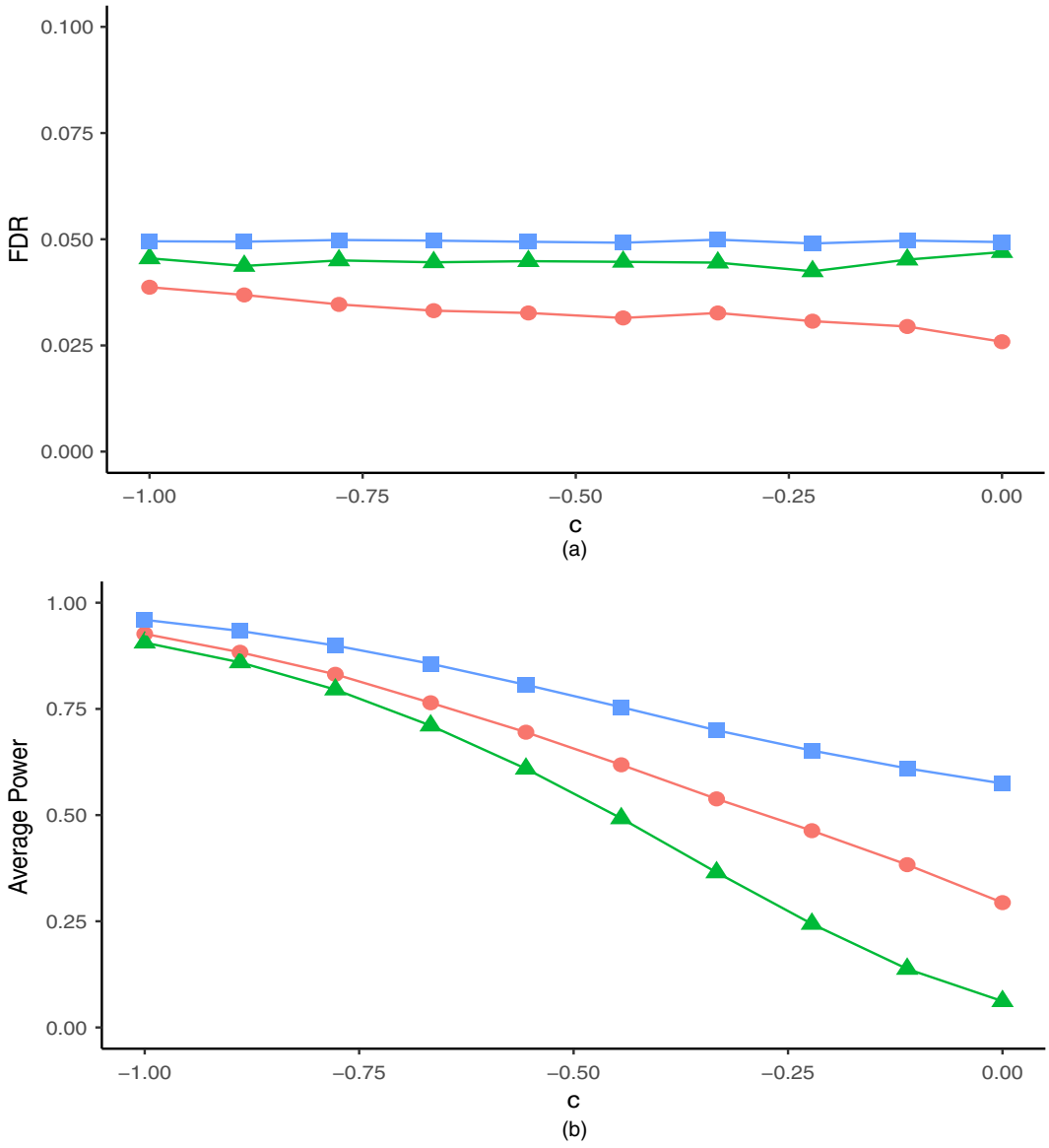


Fig. 4. Comparison of BH (▲), DD (●) and OR (■) when the non-zero means have opposite signs

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

and

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

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.

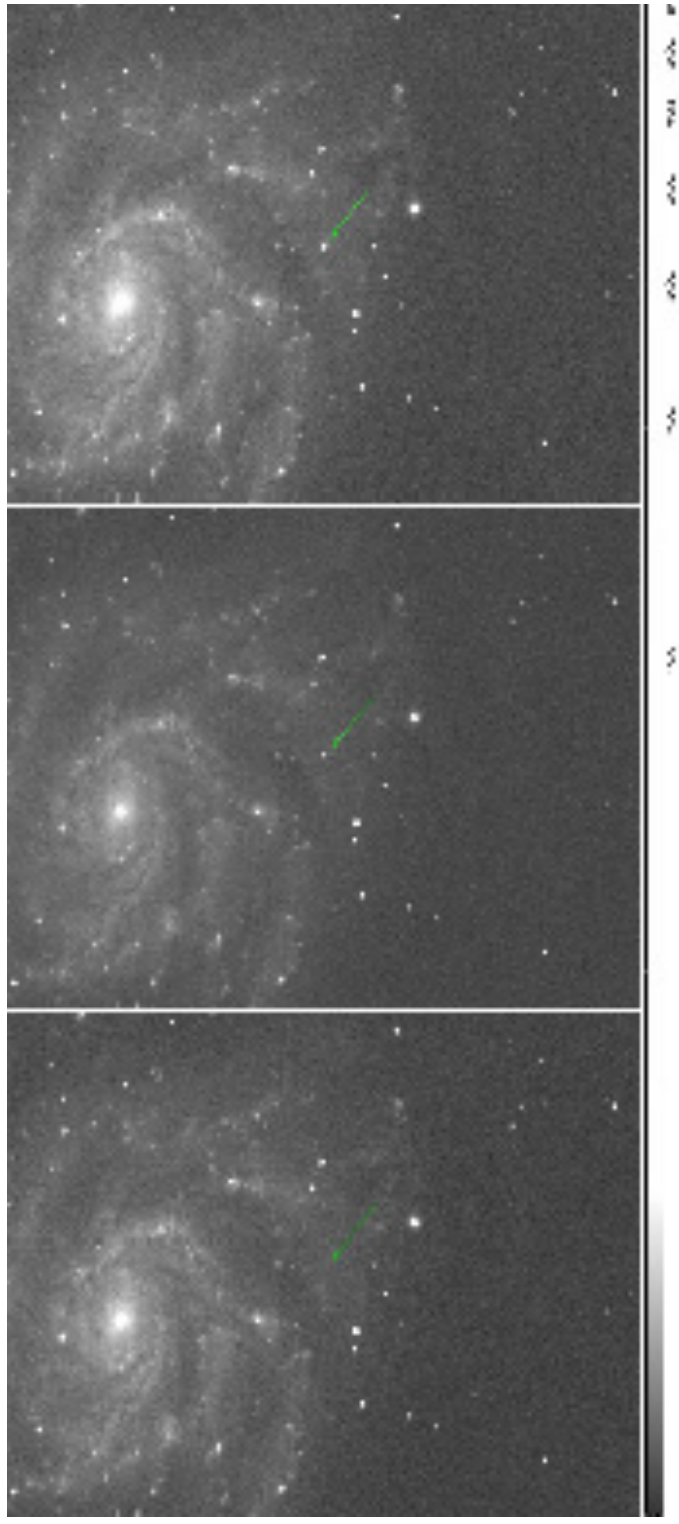


Fig. 5. Images taken on (a) August 23rd, (b) August 24th and (c) August 25th, 2011: the arrows clearly indicate the explosion of supernova SN 2011fe

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 516×831 (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 $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_m)$ (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 $\mathbf{x}^{\mathrm{st}}$ and $\mathbf{y}^{\mathrm{st}}$, 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 $\mathbf{x}^{\mathrm{st}}$ and $\mathbf{y}^{\mathrm{st}}$ from the two images seem to be comparable as most of the pairs $(x_i^{\mathrm{st}}, y_i^{\mathrm{st}})$ have similar values.

Next we carry out $m = 428796$ two-sample tests with known variances. For standardized observations $\mathbf{x}^{\mathrm{st}}$ and $\mathbf{y}^{\mathrm{st}}$ the variances are known to be 1. Then $t_{1i} = (x_i^{\mathrm{st}} - y_i^{\mathrm{st}})/\sqrt{2}$ and $t_{2i} = (x_i^{\mathrm{st}} + y_i^{\mathrm{st}})/\sqrt{2}$. 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 516×831 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}	3.66×10^{-10} (4)	2.75×10^{-4} (5)	3.24, 5.46 (22)	4.38×10^{-4} (35)
0.01	9.91×10^{-7} (22)	3.37×10^{-2} (24)	2.51, 4.87 (38)	9.25×10^{-2} (58)
0.05	7.38×10^{-6} (64)	0.39 (69)	1.92, 4.42 (80)	0.26 (109)

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*., 2018), 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 to 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^i_{OR} 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

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

A.1.1. Proof of part (a)

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

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

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

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

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

$$\mathbb{E}_{\mathbf{T}_1,\mathbf{T}_2}\left\{\sum_{i=1}^m (T^i_{\mathrm{OR}}-\alpha_2)\mathbb{I}(T^i_{\mathrm{OR}}<t_2)\right\}>0. \tag{A.2}$$

This inequality holds since

(a) $\mathbb{E}_{\mathbf{T}_1,\mathbf{T}_2}\{\sum_{i=1}^m (T^i_{\mathrm{OR}}-\alpha_1)\mathbb{I}(T^i_{\mathrm{OR}}<t_1)\}=0$,
(b) $\mathbb{E}_{\mathbf{T}_1,\mathbf{T}_2}\{\sum_{i=1}^m (\alpha_1-\alpha_2)\mathbb{I}(T^i_{\mathrm{OR}}<t_1)\}>0$ and
(c) $\mathbb{E}_{\mathbf{T}_1,\mathbf{T}_2}\{\sum_{i=1}^m (T^i_{\mathrm{OR}}-\alpha_1)\mathbb{I}(t_1\leqslant T^i_{\mathrm{OR}}<t_2)\}>0$,

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 α_2, which implies that $\mathbb{E}_{\mathbf{T}_1,\mathbf{T}_2}\{\sum_{i=1}^m (T^i_{\mathrm{OR}}-\alpha_2)\mathbb{I}(T^i_{\mathrm{OR}}<t_2)\}=0$. Hence we must have $\alpha_1\leqslant\alpha_2$.

A.1.2. Proof of part (b)

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

$$\begin{aligned}\mathbb{E}_{\mathbf{T}_1,\mathbf{T}_2}\left\{\sum_{i=1}^m (T^i_{\mathrm{OR}}-\alpha)\delta^i_{\mathrm{OR}}\right\}&=0,\\ \mathbb{E}_{\mathbf{T}_1,\mathbf{T}_2}\left\{\sum_{i=1}^m (T^i_{\mathrm{OR}}-\alpha)\delta^i_{*}\right\}&\leqslant 0.\end{aligned} \tag{A.3}$$

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

$$\mathbb{E}_{\mathbf{T}_1,\mathbf{T}_2}\left\{\sum_{i=1}^m (\delta^i_{\mathrm{OR}}-\delta^i_*)(T^i_{\mathrm{OR}}-\alpha)\right\}\geqslant 0. \tag{A.4}$$

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

$$\delta^i_{\mathrm{OR}}=I\left(\frac{T^i_{\mathrm{OR}}-\alpha}{1-T^i_{\mathrm{OR}}}<\lambda_{\mathrm{OR}}\right),$$

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

Note that

(a) $T^i_{\mathrm{OR}}-\alpha-\lambda_{\mathrm{OR}}(1-T^i_{\mathrm{OR}})<0$ if $\delta^i_{\mathrm{OR}}>\delta^i_*$ and
(b) $T^i_{\mathrm{OR}}-\alpha-\lambda_{\mathrm{OR}}(1-T^i_{\mathrm{OR}})>0$ if $\delta^i_{\mathrm{OR}}<\delta^i_*$.

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

$$\mathbb{E}_{\mathbf{T}_1,\mathbf{T}_2}\left[\sum_{i=1}^m (\delta^i_{\mathrm{OR}}-\delta^i_*)\{T^i_{\mathrm{OR}}-\alpha-\lambda_{\mathrm{OR}}(1-T^i_{\mathrm{OR}})\}\right]\leqslant 0. \tag{A.5}$$

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

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

Finally, noting that $\lambda_{\mathrm{OR}} > 0$ and that ETP for a decision rule $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$ is given by $\mathbb{E}_{\mathbf{T}_1, \mathbf{T}_2}\{\Sigma_{i=1}^m \delta_i (1 - T_{\mathrm{OR}}^i)\}$, we conclude that $\mathrm{ETP}(\boldsymbol{\delta}_{\mathrm{OR}}) \geqslant \mathrm{ETP}(\boldsymbol{\delta}_*)$.

A.2. Proof of theorem 2

We first provide a summary of useful notation.

(a) $Q^\tau(t) = m^{-1}\Sigma_{i=1}^m (T_{\mathrm{OR}}^{\tau,i} - \alpha) I(T_{\mathrm{OR}}^{\tau,i} < t)$.
(b) $\hat{Q}^\tau(t) = m^{-1}\Sigma_{i=1}^m (\hat{T}_{\mathrm{OR}}^{\tau,i} - \alpha) I(\hat{T}_{\mathrm{OR}}^{\tau,i} < t)$.
(c) $Q_\infty^\tau(t) = \mathbb{E}\{(T_{\mathrm{OR}}^\tau - \alpha)\mathbb{I}(T_{\mathrm{OR}}^\tau < t)\}$, where T_{OR}^τ is a generic member from $\{T_{\mathrm{OR}}^i : 1 \leqslant i \leqslant m\}$.

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

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

A.2.1. Proof of part (a)

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

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

$$Q_{\mathrm{C}}^\tau(t) = \frac{t - T_{\mathrm{OR}}^{\tau,(k)}}{T_{\mathrm{OR}}^{\tau,(k+1)} - T_{\mathrm{OR}}^{\tau,(k)}} Q_k^\tau + \frac{T_{\mathrm{OR}}^{\tau,(k+1)} - t}{T_{\mathrm{OR}}^{\tau,(k+1)} - T_{\mathrm{OR}}^{\tau,(k)}} Q_{k+1}^\tau, \tag{A.6}$$

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

$$Q_{\mathrm{C}}^{\tau,-1}(0) \xrightarrow{\mathrm{p}} t_\infty^\tau, \tag{A.7a}$$

$$\hat{Q}_{\mathrm{C}}^{\tau,-1}(0) \xrightarrow{\mathrm{p}} t_\infty^\tau. \tag{A.7b}$$

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

$$\mathbb{P}\{|Q_{\mathrm{C}}^\tau(t_\infty^\tau) - \alpha| > \eta\} \geqslant \mathbb{P}[|Q_{\mathrm{C}}^{\tau,-1}(\alpha) - Q_{\mathrm{C}}^{\tau,-1}\{Q_{\mathrm{C}}^\tau(t_\infty^\tau)\}| > \epsilon].$$

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

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

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

$$\begin{aligned}(\hat{U}_i - U_i)^2 &= (\hat{T}_{\mathrm{OR}}^\tau - T_{\mathrm{OR}}^\tau)^2\mathbb{I}(\hat{T}_{\mathrm{OR}}^\tau \leqslant t, T_{\mathrm{OR}}^\tau \leqslant t) + (\hat{T}_{\mathrm{OR}}^\tau)^2\mathbb{I}(\hat{T}_{\mathrm{OR}}^\tau \leqslant t, T_{\mathrm{OR}}^\tau > t) + (T_{\mathrm{OR}}^\tau)^2\mathbb{I}(\hat{T}_{\mathrm{OR}}^\tau > t, T_{\mathrm{OR}}^\tau \leqslant t) \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

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

$$\mathbb{P}(\hat{T}_{\mathrm{OR}}^\tau \leqslant t, T_{\mathrm{OR}}^\tau > t) \leqslant \mathbb{P}\{T_{\mathrm{OR}}^\tau \in (t, t + \eta)\} + \mathbb{P}(|\hat{T}_{\mathrm{OR}}^\tau - T_{\mathrm{OR}}^\tau| > \eta) \to 0.$$

Since $\hat{T}_{\mathrm{OR}}^\tau$ 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 $\hat{Q}^{\tau}(t) \to^{\mathrm{p}} Q^{\tau}_{\infty}(t)$. Noting that $Q^{\tau}(t) \to^{\mathrm{p}} Q^{\tau}_{\infty}(t)$, we need to show only that $\hat{Q}^{\tau}(t) \to^{\mathrm{p}} Q^{\tau}(t)$. The dependence among $\hat{U}_i$ in the expression $\hat{Q}^{\tau}(t) = m^{-1}\Sigma_i \hat{U}_i$ 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 $S_n = \Sigma_{i=1}^{m}(\hat{U}_i - U_i)$. 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

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

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

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

Therefore $\hat{Q}^{\tau}(t) \to^{\mathrm{p}} Q^{\tau}_{\infty}(t)$. By definition, $|\hat{Q}^{\tau}_{\mathrm{C}}(t) - \hat{Q}^{\tau}(t)| \leqslant m^{-1}$. We claim that $\hat{Q}^{\tau}_{\mathrm{C}}(t) \to^{\mathrm{p}} Q^{\tau}_{\infty}(t)$ and result (A.7b) follows.

According to results (A.7a) and (A.7b), $\hat{Q}^{\tau,-1}_{\mathrm{C}}(0) = Q^{\tau,-1}_{\mathrm{C}}(0) + o_{\mathbb{P}}(1)$. The mFDR-levels of the testing procedures are

$$\mathrm{mFDR}(\boldsymbol{\delta}^{\tau}_{\mathrm{OR}}) = \frac{\mathbb{P}_{H_0}\{T^{\tau,i}_{\mathrm{OR}} < Q^{\tau,-1}_{\mathrm{C}}(\alpha)\}}{\mathbb{P}\{T^{\tau,i}_{\mathrm{OR}} < Q^{\tau,-1}_{\mathrm{C}}(\alpha)\}}$$

and

$$\mathrm{mFDR}(\boldsymbol{\delta}_{\mathrm{DD}}) = \frac{\mathbb{P}_{H_0}\{\hat{T}^{\tau,i}_{\mathrm{OR}} < \hat{Q}^{\tau,-1}_{\mathrm{C}}(\alpha)\}}{\mathbb{P}\{\hat{T}^{\tau,i}_{\mathrm{OR}} < \hat{Q}^{\tau,-1}_{\mathrm{C}}(\alpha)\}}.$$

The operation of our testing procedure implies that $Q^{\tau,-1}_{\mathrm{C}}(\alpha) \geqslant \alpha$. It follows that $\mathbb{P}\{T^{\tau,i}_{\mathrm{OR}} < Q^{\tau,-1}_{\mathrm{C}}(\alpha)\}$ is bounded away from zero. We conclude that $\mathrm{mFDR}(\boldsymbol{\delta}_{\mathrm{DD}}) = \mathrm{mFDR}(\boldsymbol{\delta}^{\tau}_{\mathrm{OR}}) + o(1)$.

The result on mFDR-control can be extended to FDR-control. The next proposition, which is proved in the on-line supplementary 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 $\boldsymbol{\delta}^{\tau}_{\mathrm{DD}}$. It follows from proposition 7 that CARS controls FDR at level $\alpha + o(1)$.

Proposition 7.

(a) Consider a general decision rule $\boldsymbol{\delta}$. Let $\mathcal{Y} = m^{-1}\Sigma_{i=1}^{m}\delta_i$. Then $\mathrm{mFDR}(\boldsymbol{\delta}) = \mathrm{FDR}(\boldsymbol{\delta}) + o(1)$ if
   (i) $\mathbb{E}(\mathcal{Y}) \geqslant \underline{\eta}$ for some $\underline{\eta} > 0$ and
   (ii) $\mathrm{var}(\mathcal{Y}) = o(1)$.

(b) Conditions (i) and (ii) are fulfilled by the CARS procedure $\boldsymbol{\delta}^{\tau}_{\mathrm{DD}}$.

A.2.2. Proof of part (b)

The CARS procedure utilizes $\hat{q}*$, and the corresponding test statistic is $\hat{T}^{*,i}_{\mathrm{OR}}$. It follows from conditions 1′ and 2, and the continuous mapping theorem that $\hat{T}^{*}_{\mathrm{OR}} \to^{\mathrm{p}} T_{\mathrm{OR}}$. Denote $Q_{\mathrm{OR}}(t)$ the oracle mFDR-function and t_{OR} the oracle threshold. Then

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

Define $\hat{Q}^{*}(t) = m^{-1}\Sigma_{i=1}^{m}\hat{T}^{*,i}_{\mathrm{OR}}\mathbb{I}(\hat{T}^{*,i}_{\mathrm{OR}} < t)$. Similarly to equation (A.6) we define a continuous version of $\hat{Q}^{*}(t)$ and denote it by $\hat{Q}^{*}_{\mathrm{C}}(t)$. It can be shown that $\hat{Q}^{*}_{\mathrm{C}}(t)$ is continuous and monotone; so is its inverse $\hat{Q}^{*,-1}_{\mathrm{C}}(t)$. The CARS procedure is given by $\boldsymbol{\delta}^{*}_{\mathrm{DD}} = [\mathbb{I}\{\hat{T}^{*,i}_{\mathrm{OR}} \leqslant \hat{Q}^{*,-1}_{\mathrm{C}}(\alpha)\} : 1 \leqslant i \leqslant m]$. Following the steps in the proof of part (a) we can show that

$$\begin{aligned} \hat{Q}^{*}_{\mathrm{C}}(t) &\xrightarrow{\mathrm{p}} Q_{\mathrm{OR}}(t), \\ \hat{Q}^{*,-1}_{\mathrm{C}}(t) &\xrightarrow{\mathrm{p}} t_{\mathrm{OR}}. \end{aligned} \tag{A.8}$$

The operation of CARS implies that $Q^{*,-1}_{\mathrm{C}}(0) \geqslant \alpha$ (thus the denominator of mFDR is bounded away from zero). Note that $\mathrm{mFDR}(\boldsymbol{\delta}_{\mathrm{OR}}) = \alpha$; we have $\mathrm{mFDR}(\boldsymbol{\delta}^{*}_{\mathrm{DD}}) = \alpha + o(1)$. Next, we consider ETP. It follows from $\hat{T}^{*}_{\mathrm{OR}} \to^{\mathrm{p}} T_{\mathrm{OR}}$ and expression (A.8) that

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

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. (2018) Weighted false discovery control in large-scale multiple testing. *J. Am. Statist. Ass.*, **113**, 1172–1183.

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*. Johns Hopkins University, Baltimore.

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., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., Lowry, S. F. and Inflamm and Host Response to Injury Large Scale Collab. Res. Program (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*. Laboratoire de Probabilités et Modèles Aléatoires, Université Pierre et Marie Curie, Paris.

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.

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 (ed. J. Neyman), pp. 361–379. Berkeley: University of California Press.

Jin, J. and Cai, T. T. (2007) Estimating the null and the proportion 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.

Law, N. M., Kulkarni, S. R., Dekany, R. G., Ofek, E. O., Quimby, R. M., Nugent, P. E., Surace, J., Grillmair, C. C., Bloom, J. S., Kasliwal, M. M., Bildsten, L., Brown, T., Cenko, S. B., Ciardi, D., Croner, E., Djorgovski, S. G., van Eyken, J. C., Filippenko, A. V., Fox, D. B., Gal-Yam, A., Hale, D., Hamam, N., Helou, G., Henning, J. R., Howell, D. A., Jacobsen, J., Laher, R., Mattingly, S., McKenna, D., Pickles, A., Poznanski, D., Rahmer, G., Rau, A., Rosing, W., Shara, M., Smith, R., Starr, D., Sullivan, M., Velur, V., Walters, R. S. and Zolkower, J. (2009) The Palomar Transient Factory: system overview, performance, and first results. *Publ. Astron. Soc. Pacif.*, **121**, 1395.

Lehmann, E. L. and Casella, G. (2006) *Theory of Point Estimation*. New York: Springer Science and Business Media.

Li, A. and Barber, R. F. (2016) Multiple testing with the structure adaptive Benjamini-Hochberg algorithm. *Preprint arXiv:1606.07926*.

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

Liu, Y., Sarkar, S. K. and Zhao, Z. (2016) A new approach to multiple testing of grouped hypotheses. *J. Statist. Planng Inf.*, **179**, 1–14.

Neuvial, P. (2013) Asymptotic results on adaptive false discovery rate controlling procedures based on kernel estimators. *J. Mach. Learn. Res.*, **14**, 1423–1459.

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

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

Roeder, K. and Wasserman, L. (2009) Genome-wide significance levels and weighted hypothesis testing. *Statist. Sci.*, **24**, 398–413.

Roquain, E. and Van De Wiel, M. A. (2009) Optimal weighting for false discovery rate control. *Electron. J. Statist.*, **3**, 678–711.

Rubin, D., Dudoit, S. and van der Laan, M. (2006) A method to increase the power of multiple testing procedures through sample splitting. *Statist. Appl. Genet. Molec. Biol.*, **5**, article 19.

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

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

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

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P. and Kass, R. E. (2015) False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *J. Am. Statist. Ass.*, **110**, 459–471.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. Boca Raton: CRC Press.

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

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

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

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

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

Tukey, J. W. (1994) *The Collected Works of John W. Tukey*, vol. 3. New York: Taylor and Francis.

Wand, M. and Jones, M. (1995) *Kernel Smoothing*. London: Chapman and Hall.

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

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

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

Zehetmayer, S., Bauer, P. and Posch, M. (2008) Optimized multi-stage designs controlling the false discovery or 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"'.

Discussion on the Paper by Cai, Sun and Wang

Etienne Roquain (*Sorbonne Université, Paris*)

Introduction

I congratulate Cai, Sun and Wang for this excellent contribution to multiple-testing methodology. This work shows, in a two-sided setting, that reducing a data set to a column of test statistics is suboptimal and that the multiple-testing decision can gain much from incorporating side information. In a nutshell the authors proposed

(a) to find the procedure, called oracle covariate-assisted ranking and screening (CARS), that solves the problem of maximizing the power while controlling the (marginal) false discovery rate (FDR),
(b) to approximate this optimal procedure by a data-driven version, called CARS, by way of kernel estimators and
(c) to prove theoretical consistency of CARS when all the parameters are kept fixed and the number m of nulls grows to ∞.

Many extensions are discussed and a package is implementing CARS, which make this method available for practitioners.

The purpose of this discussion contribution is to underline the Bayesian flavour of CARS and to discuss a Cauchy slab version of it. In particular, since the theoretical framework of the paper seems to exclude the case where the model parameters depend on m, and so exclude sparse signals, we consider the problem of obtaining a uniform FDR control under sparsity, which can be formulated as

$$\sup_{\Delta:\|\Delta\|_0 \leqslant s_m} \mathrm{FDR}(\Delta) \leqslant \alpha, \qquad s_m/m \to 0, \tag{1}$$

where $\Delta \in \mathbb{R}^m$ is the true mean difference and $\|z\|_0$ denotes the number of non-zeros in z. Since the original Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) provides such a guarantee, we may ask whether procedures improving on the Benjamini–Hochberg procedure also provide expression (1).

Stylized covariate-assisted ranking and screening setting

Briefly, consider the following simplified version of the CARS setting. Let $X \sim \mathcal{N}(\mu_x, I_m)$ and $Y \sim \mathcal{N}(\mu_y, I_m)$ be two independent random vectors, for two $\mathbb{R}^m$ mean vectors μ_x and μ_y. We consider the problem of testing simultaneously the nulls $H_{0,i}$: '$\mu_{x,i} = \mu_{y,i}$' against $H_{1,i}$: '$\mu_{x,i} \neq \mu_{y,i}$'. We let

$$\theta_i = \mathbf{1}\{\mu_{x,i} \neq \mu_{y,i}\}, \qquad 1 \leqslant i \leqslant m,$$

the true–false status of the hypotheses. Testing can be done by using the standard test statistics

$$T_1 = (X - Y)/\sqrt{2} \sim \mathcal{N}(\Delta, I_m), \qquad \Delta = (\mu_x - \mu_y)/\sqrt{2}.$$

The idea of CARS is to keep the information of the (independent) covariate $T_2 = (X + Y)/\sqrt{2} \sim \mathcal{N}\{(\mu_x + \mu_y)/\sqrt{2}, I_m\}$ in the analysis to help in making the decision.

The authors' model additionally uses some random effects for the mean couple $(\mu_x, \mu_y) \in \mathbb{R}^{2m}$, which can be interpreted as choosing a particular prior distribution on the true parameters. Let $\phi_i = \mathbf{1}\{\mu_{x,i} \neq 0 \text{ or } \mu_{y,i} \neq 0\}$, and generate (θ, ϕ) as follows: $\theta_i \sim^{\mathrm{IID}} \mathcal{B}(1 - \pi_0)$; $\phi_i | \theta_i = 0 \sim^{\mathrm{IID}} \mathcal{B}(1 - \pi_{0|0})$; $\phi_i = 1$ if $\theta_i = 1$. Also let $\pi_{00} = \pi_0 \pi_{0|0}$ the probability that both means are equal to 0. To complete the description of the prior, the components of (μ_x, μ_y) are generated independently as

$$(\mu_{x,i} - \mu_{y,i})/\sqrt{2} \sim \begin{cases} \delta_0 & \text{if } \theta_i = 0, \\ \gamma & \text{if } \theta_i = 1, \end{cases}$$

$$(\mu_{x,i} - \mu_{y,i})/\sqrt{2} \sim \begin{cases} \delta_0 & \text{if } \phi_i = 0, \\ \gamma & \text{if } \phi_i = 1. \end{cases}$$

At this point, the strategy of CARS is to make the decision by estimating the posterior probability $\mathbb{P}(\theta_i = 0 | T_{1,i}, T_{2,i})$, which can be seen as implicity using a non-parametric estimator of the slab γ.

Cauchy slab covariate-assisted ranking and screening version

We discuss now the possibility of avoiding non-parametric estimation, by fixing γ equal to $\gamma(x) = (2\pi)^{-1/2}\{1 - |x|\bar{\Phi}(x)/\phi(x)\}$, which is the so-called quasi-Cauchy distribution (Johnstone and Silverman, 2004). Recent studies suggest that such a prior is particularly suitable to obtain posterior distributions with good frequentist properties; see for example Castillo and Mismer (2008) and references therein. An

intuitive explanation is that this density has heavy tails, which thus puts mass 'everywhere' and can thus account for any true alternative distribution of the test statistics.

We introduce the Cauchy slab version of the CARS procedure, which rejects $H_{0,i}$ as soon as $q(T_{1,i}, T_{2,i}) \leqslant \alpha$, for which

$$q(t_1, t_2) = \mathbb{P}(\theta_i = 0 | |T_{1,i}| \geqslant t_1, T_{2,i} = t_2) = \frac{\pi_{00}\bar{\Phi}(t_1)\phi(t_2) + (\pi_0 - \pi_{00})\bar{\Phi}(t_1)g(t_2)}{\pi_{00}\bar{\Phi}(t_1)\phi(t_2) + (\pi_0 - \pi_{00})\bar{\Phi}(t_1)g(t_2) + (1 - \pi_0)\bar{G}(t_1)g(t_2)},$$

where $g(x) = (\phi * \gamma)(x) = (2\pi)^{-1/2}x^{-2}\{1 - \exp(-x^2/2)\}$ and $\bar{G}(s) = \int_s^\infty g(x)\mathrm{d}x$. In the equation above, the only unknown quantities are the hyperparameters π_0 and π_{00}. The parameter π_0 and π_{00} can be easily estimated by marginal maximum likelihood from respectively the sample T_1 and T_2 (see Johnstone and Silverman (2004) and the devoted package Johnstone and Silverman (2005)), because the $T_{1,i}$s are independent and identically distributed $\pi_0\phi + (1 - \pi_0)g$ and the $T_{2,i}$s are independent and identically distributed $\pi_{00}\phi + (1 - \pi_{00})g$.

Now, two questions are

(a) does this Cauchy slab version enjoy the uniform FDR control (1) and
(b) does it still improve on the Benjamini–Hochberg procedure in terms of power?

I tend to believe that both answers are positive. First it seems reasonable to think that the uniform FDR control can be proved by extending the methodology of Castillo and Roquain (2018) to the bivariate case. Second, even with the uninformative Cauchy slab, the covariate T_2 can still help T_1 to make the correct decision, as illustrated in Fig. 6. These two facts have been confirmed by unreported numerical computations.

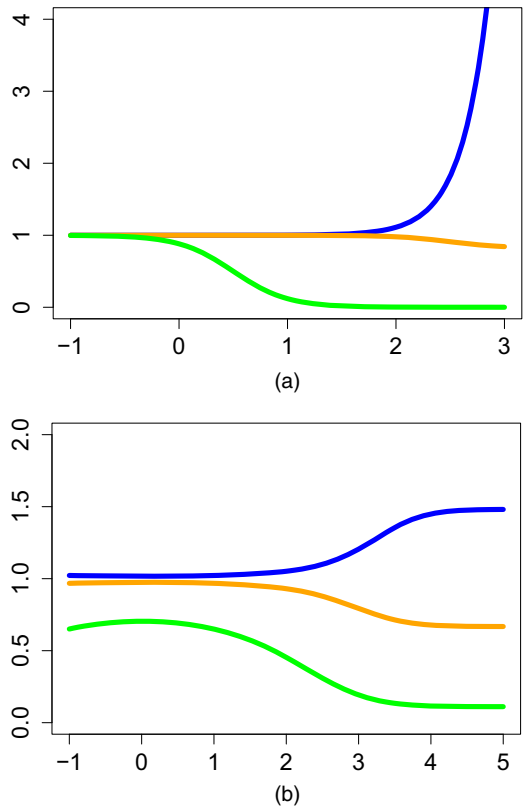


Fig. 6. Ratio of the posteriors $\mathbb{P}(\theta_i = 0|T_{1,i}, T_{2,i})/\mathbb{P}(\theta_i = 0|T_{1,i})$ as a function of $T_{1,i}$, for various values of $T_{2,i}$: the prior is computed either from (a) the true mixture density (——, $T_{2,i} = 1$; ——, $T_{2,i} = 6$; ——, $T_{2,i} = 10$) or from (b) the Cauchy slab (——, $T_{2,i} = 1$; ——, $T_{2,i} = 2$; ——, $T_{2,i} = 5$); in both cases, small or large values of $T_{2,i}$ help $T_{1,i}$ to accept or reject the null respectively

Conclusion

Overall, this discussion puts forward the issue of approaching the oracle version of the CARS procedure under sparsity. There may be a trade-off in the choice of the complexity of the slab estimator: although using a non-parametric kernel estimator is ambitious from a power point of view, it might be too unstable for reliable control of the FDR. Instead, using a simpler Cauchy slab introduces a bias that will stabilize the FDR, but it does reduce power. Finding a principled trade-off between these two extremes is certainly an interesting avenue for future research.

To conclude, it is a pleasure for me to propose the vote of thanks.

I also warmly acknowledge Ismael Castillo, Sebastian Dölher and Mark van de Wiel for discussions that helped while preparing these comments.

Thomas E. Nichols (*University of Oxford*)

Cai, Sun and Wang (CSW) are to be congratulated on a relevant work that brings modern statistical methods to bear on the age-old problem of comparing two groups. Their 'nested sparsity' setting, where a sparse difference of two populations is informed by the sparse average of the population, is highly relevant to functional magnetic resonance imaging (MRI). In this short comment I give a brief background on functional MRI, provide a reformulation of their method that I think provides even more intuition and finally provide an illustrative numerical simulation and a functional MRI data analysis to explore the value of the method.

Just as the colour of blood changes from blue to red as it is exposed to oxygen, the magnetic properties change, giving rise to the blood oxygenation level dependent (BOLD) effect. The BOLD effect allows an MRI scanner to be used to track changes in brain activity in response to an experimental task. In the scanner, the subject alternates between control and active tasks, and the per-subject outcome is an image of change in the BOLD signal between control and active states. For example, in a working memory experiment the control task is viewing a sequence of letters, pressing a button when an 'X' appears, whereas in the active task the subject must remember successive letters and detect when a letter is repeated in a given interval. Subtracting functional MRI data in the control state from the active state should cancel out irrelevant activity—related to vision and button clicking—and leave only BOLD changes related to short-term storage and retrieval of letters.

Like genetics, functional MRI comprises a massive multiple-testing problem, where there are $m = 10^5$–10^6 voxels (volume elements) tested. The basic question, detecting voxels with a non-zero change in the BOLD signal, is typically less interesting than finding regions that differ between groups (e.g. young *versus* old) or that vary with a covariate (e.g. age). The essential feature of functional MRI group comparisons is exactly the nested sparsity setting of the method of CSW—group differences in BOLD activation are expected to be found in voxels where there is an overall, average BOLD effect. This aspect of functional MRI data is so well understood that an informal, two-step method is often used for group inference: first conduct a test for the average BOLD effect, identify significant voxels by some method and then conduct a test of a differential effect *only* at the average significant voxels, dramatically reducing the severity of the multiple-testing problem for the group difference inference (Kriegeskorte *et al*. (2009), supplementary text, page 13).

The method of CSW can be seen as a much more nuanced version of this informal functional MRI method: instead of completely ignoring tests that lack significance for the average effect, inferences are adjusted according to evidence for the average effect.

What most draws me to the method of CSW is that it can be seen as a direct extension of Efron's local false discovery rate (FDR). Retaining CSW's notation, where the test of the group difference for element i is T_{1i} and the test of the average is T_{2i} (based on a weighted average constructed to be independent of the difference), the local FDR for a group difference is

$$\mathrm{Lfdr}_i(t_1) = P(\theta_{1i} = 0 | T_{1i} = t) \tag{2}$$

and the covariate-assisted ranking and screening (CARS) statistic is

$$T^i_{\mathrm{OR}}(t_1, t_1) = P(\theta_{1i} = 0 | T_{1i} = t_1, T_{2i} = t_2); \tag{3}$$

both are for inference on the group difference effect θ_{1i}, but with the CARS statistic information on average is now considered via additional conditioning on $T_{2i} = t_2$. Moreover, the two definitions can be combined to show that the CARS statistic is simply the local FDR weighted according to a measure of dependence of T_2 on T_1, expressed as a ratio:

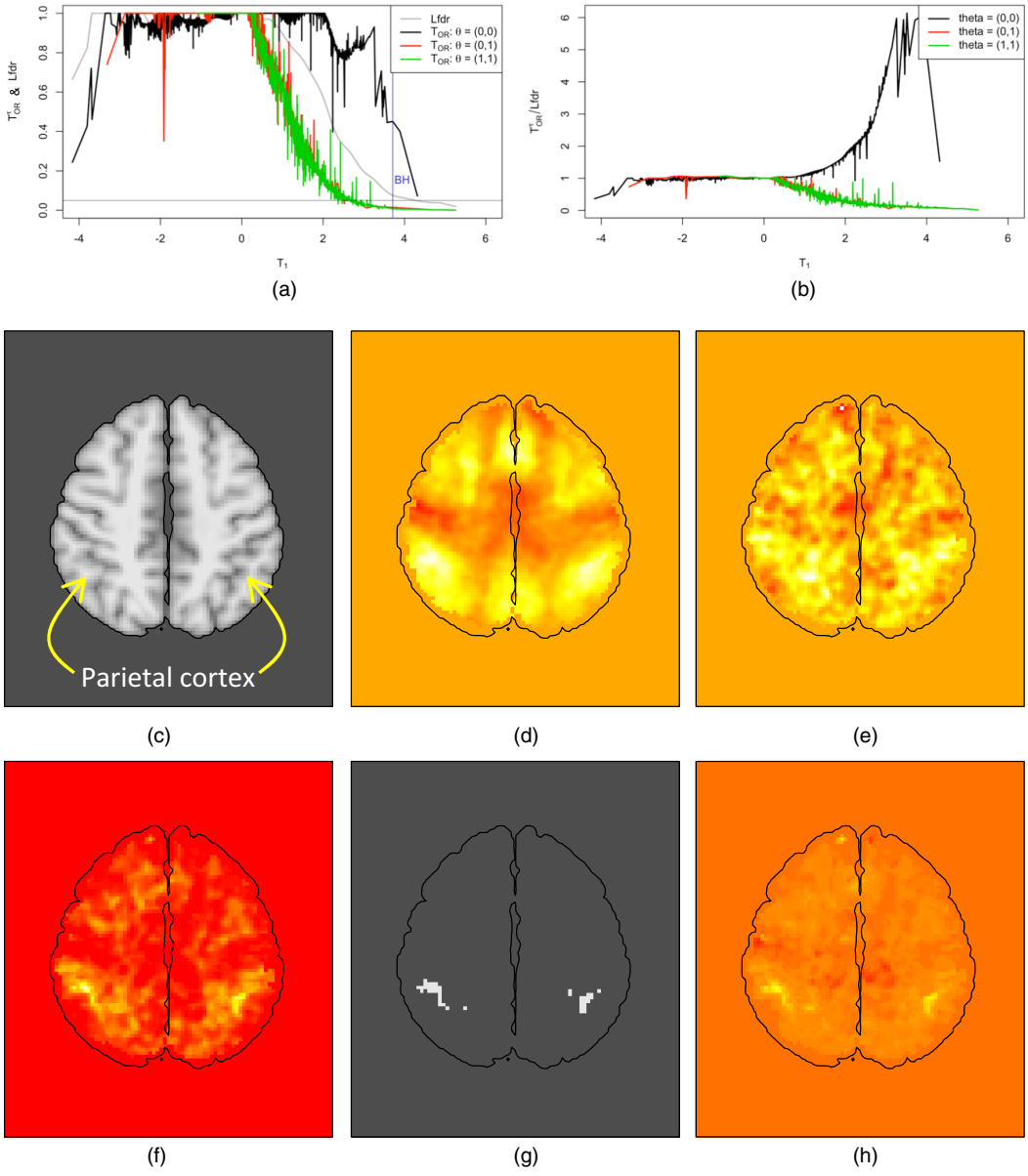


Fig. 7. Illustration of CARS methods on simulated and real functional MRI data: (a), (b) local FDR and CARS are closely related, here shown as T_{OR} diverging above or below Lfdr depending on the dependence between T_1 and T_2; (c)–(h) functional MRI results showing how a group difference effect receives a 'power boost' from the average effect (see the text for a full description); (a) Lfdr- and CARS statistic; (b) CARS/Lfdr ratio—effect of T_2; (c) anatomical reference; (d) T_2—average; (e) T_1—group difference; (f) $-\log_{10}$(CARS)—group difference; (g) CARS detections; (h) $-\log_{10}$(CARS/Lfdr)—T_2 boost

$$\frac{T^i_{\mathrm{OR}}(t_1,t_2)}{\mathrm{Lfdr}_i(t_1)} = \frac{f(t_2|\theta_{1i}=0)}{f(t_2|T_{1i}=t_1)}. \tag{4}$$

This provides a clear intuition for CARS: if t_2 is more likely given $T_1 = t_1$ than under the state of no

difference, evidence for a difference is amplified. Although this ratio of null marginal and conditional densities is not a practical approach to estimating the CARS statistic, once T^i_{OR} has been estimated the ratio can be computed as $T^i_{\mathrm{OR}}(t_1, t_2)/\mathrm{Lfdr}_i(t_1)$ and used as a diagnostic.

We illustrate this with a variant of CSW's simulation setting 1. I applied the CARS R package to one realization of $n_x = 50$, $n_y = 60$ and $m = 10000$ data, with 2000 elements with a positive mean, of which 1000 elements had group difference, with common means $\mu_x = \mu_y = 4/\sqrt{30}$, differential means $\mu_x = 5/\sqrt{30}$ and $\mu_y = 2/\sqrt{30}$, and $\sigma_x = 1$ and $\sigma_y = 2$. This produces three types of effect: no signal at all $\theta = (0, 0)$, an average but no group difference effect $\theta = (0, 1)$ and both average and group difference effects $\theta = (1, 1)$. Fig. 7(a) shows that Lfdr (grey) detects some tests (critical $T_1 = 4.283$) but fewer than the Benjamini–Hochberg (BH) FDR (the vertical line, $T_1 = 3.709$); the CARS statistic, plotted by signal type, shows dramatically distinct behaviour between complete null tests ($\theta = (0, 0)$; the black curve) and those with a mean effect, both those without ($\theta = (0, 1)$; red curve) and with a group difference effect ($\theta = (1, 1)$; green curve); the smallest CARS-detected statistic was $T_1 = 2.369$. The diagnostic ratio, T_{OR}/Lfdr (Fig. 7(b)), clearly shows the amplifying and attenuating effect that information from T_2 brings to the inference on T_1.

Finally, I examined the effect of intelligence quotient (IQ) on a working memory task functional MRI data set on 78 unrelated subjects in the Human Connectome Project (Van Essen *et al.*, 2013). I created low and high IQ groups by using a median split of the fluid intelligence score PMAT_CR, testing high − low. Figs 7(c)–7(h) show results for one slice with 2930 voxels, with anatomical reference in Fig. 7(c). The average activation (one-sample t-test; Fig. 7(d)) shows prominent effects in the left and right parietal areas of the brain, which are thought to be involved in storage (as opposed to retrieval) of information. The group difference (Fig. 7(e)) is a much weaker effect and neither the Benjamini–Hochberg FDR nor Lfdr finds any significance at level 0.05. The CARS statistic (Fig. 7(f)) shows a different pattern of effects and finds 42 significant voxels (Fig. 7(g)); the CARS–Lfdr ratio (Fig. 7(h)) specifically records where dependence between T_1 and T_2 exists and boosts the effect that would be otherwise missed.

Although I am enthusiastic for the potential of this procedure, a significant limitation is that it can only work with the two-sample case. As motivated by the functional MRI IQ example, we would ideally be testing for a linear effect of a covariate while using information on the overall mean. As the sample mean and centred covariate effects are independent in a Gaussian linear model, it would seem to be a direct extension of this work. Also additional work needs to be done on the implementation to scale up to truly large data. The paper's examples produced an error when scaled up to 100000 tests, and simulations under increased sparsity also failed (though the sparsity mode that is referenced in the paper, but was unavailable at the time of the presentation, should address this). Over all, this is an important and outstanding work and it is my pleasure to offer to second the vote of thanks.

The vote of thanks was passed by acclamation.

Felipe A. Medina (*University of Chile, Santiago, and University of Valparaíso*) **and Milan Stehlík** (*Johannes Kepler University in Linz, University of Valparaíso, and Arizona State University, Tempe*)

Our congratulations go to Cai, Sun and Wang for their work on ranking and screening in large-scale two-sample inference. Here, we point out a potential issue that arises in the context of differential expression (DE) studies where biological samples are genetically heterogeneous. In particular, consider a data-generating process corresponding to the following modification of the authors' model:

$$X_{ij} = \beta_{0i} + \mu^*_{xi} + \eta_{xij} + \epsilon_{xij},$$

$$Y_{ij} = \beta_{0i} + \mu^*_{yi} + \eta_{yik} + \epsilon_{yik}$$

where η_{xij} and η_{yik} are fixed effects accounting for the effect of the genetic background of individuals j and k on the expression level of gene i respectively. These effects vary depending on each gene's degree of genomic-dependent variation and every individual's genetic background variability. This model is then similar to those used in the analysis of expression quantitative trait loci (see Michaelson *et al.* (2009)).

It can be shown that under this model the primary test statistic T_{1i} is biased under the null, i.e.

$$\mathbb{E}[T_{1i} | H_{i,0} : \mu^*_{xi} = \mu^*_{yi}] = \sqrt{\left(\frac{n_x n_y}{n}\right)} \sigma^{-1}_{pi} \left(n_x^{-1} \sum_{j=1}^{n_x} \eta_{xij} - n_y^{-1} \sum_{k=1}^{n_y} n_{xik} \right) = \Delta_{xyi}.$$

This bias is unobserved and can have a substantial effect on the inference of DE studies. First, it shifts the centre of the empirical null distribution of $T_{1,i}$ from 0 to Δ_{xyi}, which in a univariate analysis setting increases

the empirical type I error rate. Second, this shift can go against the direction of the unstandardized effect size, $\mu^*_{xi} - \mu^*_{yi}$, which in a univariate setting reduces empirical power.

Given the increasing need to study multiethnic cohorts and admixed populations, methods for performing DE analyses in this type of studies need to be able to handle sources of variability like those found in expression quantitative trait locus studies. We mention some (Michaelson *et al.*, 2009): batch effects, population (sub)structure and cryptic relatedness. Because of this, we think that adaptation of the method of Cai, Sun and Wang to consider this issue is an interesting direction for further research.

The following contributions were received in writing after the meeting.

Trambak Banerjee and Gourab Mukherjee (*University of Southern California, Los Angeles*)

We congratulate Cai, Sun and Wang for developing the highly potent covariate-assisted ranking and screening (CARS) procedure for two-sample multiple testing of means. In CARS, an auxiliary covariate sequence that contains additional structural information on the support where the true mean differences are 0, is combined with the primary test statistic to construct an improved multiple-testing procedure. The improvement achieved by CARS in multiple-testing frameworks motivated us to study the efficient use of side information for improving mean-squared error in the estimation of several contrasts.

As in CARS, we observe two samples $X_i \sim N(\mu_{i,1}, \sigma^2_{i,1})$ and $Y_i \sim N(\mu_{i,2}, \sigma^2_{i,2})$ with $\sigma_{i,1}$ and $\sigma_{i,2}$ known and $\mu_{i,1}$ and $\mu_{i,2}$ unknown for all $i = 1, \ldots, m$, and we consider estimating the contrast vector $\boldsymbol{\nu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ when $\boldsymbol{\nu}$ is sparse. Applying SureShrink (Donoho and Johnstone, 1995) on the observed differences $\mathbf{X} - \mathbf{Y}$ is a popular procedure that incorporates sparsity information of $\boldsymbol{\nu}$ data-adaptive soft thresholding. Consider augmenting an auxiliary sequence (AS) $S_i = |X_i + \kappa_i Y_i|$ to the primary statistic (PS) $\mathbf{X} - \mathbf{Y}$ with $\kappa_i = \sigma_{i,1}/\sigma_{i,2}$, which makes them conditionally independent. Note that the AS contains additional sparsity information on $\boldsymbol{\nu}$ that can be leveraged by combining it with the PS. Perhaps, the easiest combined procedure is groupwise adaptive thresholding where groups are based on the AS and estimators within each group is based on the PS (Banerjee *et al.*, 2018). We call it the adaptive sparse estimator using side information 'ASUS' and study its performance compared with competing methods that do not leverage side information. Fig. 8 and Table 2 demonstrate the risk performance of ASUS where $\mu_{i,1}$ and $\mu_{i,2}$ are generated from a sparse mixture model (details are given in Fig. 8). We see that ASUS uses the side information in $\mathbf{S}$ and exhibits superior performance across both scenarios. This corroborates the importance of the auxiliary covariate sequence in constructing groups with disparate sparsity levels and thereby improving the overall estimation accuracy.

Over the last decade, tremendous advances in data gathering and sharing facilities have led to the accumulation of huge digital repositories that can be cheaply and readily accessed for collection of supplementary

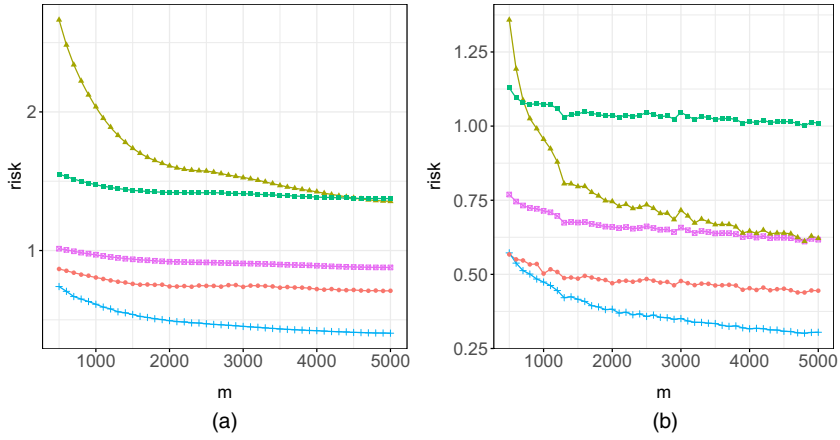


Fig. 8. Average risks of various estimators; ASUS (●), empirical Bayes thresholding (▲) of Johnstone and Silverman (2004), extended James–Stein estimator (■) discussed in Brown (2008) and the SureShrink estimator (■) in Donoho and Johnstone (1995) (here for $i = 1, \ldots, m$, $\mu_{i,1} \sim^{\text{IID}} (1 - m^{-0.6})\delta_0 + m^{-0.6}\text{Unif}(4, 8) + 0.1Z_{i,1}$, $\mu_{i,2} \sim^{\text{IID}} (1 - m^{-0.2})\delta_0 + m^{-0.2}\delta_{\{5\}} + 0.1Z_{i,2}$ and $Z_{i,1}, Z_{i,2} \sim^{\text{IID}} N(0, 1)$; the oracle estimator (+) used here is the loss oracle defined in Xie *et al.* (2012)): (a) scenario 1 wherein $\sigma_{i,1} = \sigma_{i,2} = 1$; (b) scenario 2 wherein $(\sigma^2_{1,i}, \sigma^2_{2,i}) \sim^{\text{IID}} \text{Unif}(0.1, 1.5)$

Table 2. Risk estimates for ASUS at $m = 5000$†

Method	*Results for scenario 1*	*Results for scenario 2*
Oracle	0.404	0.305
ASUS	0.710	0.445
SureShrink	0.878	0.617
EBT	1.355	0.622
EJS	1.374	1.009

†Scenario 1, $\sigma_{i,1} = \sigma_{i,2} = 1$; scenario 2, $(\sigma^2_{1,i}, \sigma^2_{2,i}) \overset{\text{IID}}{\sim} \text{Unif}(0.1, 1.5)$.

information that is pertinent to several inferential problems. In this context, the general idea behind the CARS or the ASUS procedure of adding auxiliary information to the PS for improved inference is very useful. However, in situations where there are multiple covariate sequences, it is unclear how to modify the ASUS or the CARS framework to construct an effective ranking strategy using an auxiliary matrix. An easily implementable procedure would be to construct a new sequence that represents the 'optimal use' of all available side information and thereafter use ASUS-type estimators.

Marina Bogomolov (*Technion—Israel Institute of Technology, Haifa*) **and Ruth Heller and Daniel Yekutieli** (*Tel-Aviv University*)

Cai, Sun and Wang provide us with new methodology for testing multiple two-sample problems, which makes use of a carefully constructed auxiliary statistic for more powerful identification of differential signal. Thresholding the local false discovery rate (FDR) Lfdr has been shown to be an optimal rejection policy for mFDR control when the test statistics are assumed to be generated independently from the two-group model (Cai and Sun, 2017). The authors show that, for the two-sample problem, the performance of the optimal univariate procedure can be greatly improved by exploiting the information of the standardized weighted sum of means, in addition to the standardized mean difference. The authors suggest a clever estimation method of their oracle test statistic and demonstrate its usefulness.

In Heller and Yekutieli (2014) we generalized the two-group model for inference using $N > 1$ independent studies. Let $T_{1i}, \ldots, T_{Ni}$ be the test statistics for feature i in the N studies. For θ_{ji}, the indicator of whether the association of feature i in study j is non-null, we explicitly compute the conditional probability of each configuration of $\theta_{1i}, \ldots, \theta_{Ni}$ given $T_{1i}, \ldots, T_{Ni}$. We can thus compute Lfdr, which is the marginal conditional distribution that $\theta_{1i} = 0$. In Heller and Yekutieli (2014) the conditional independence property of the statistics given the parameters applies since the statistics are from different studies. In this work the authors cleverly construct conditionally independent statistics from a single study. We think that the estimation method suggested in covariate-assisted ranking and screening is very promising and can aid, for example, the identification of genotypes associated with a psychiatric disorder by using the available studies on associations of genotypes with other (genetically related) psychiatric disorders (Andreassen *et al.*, 2013).

CARS highlights the potential in using an auxiliary statistic to aid discovery. When the auxiliary statistic comes from another study examining a similar problem, it can be of interest to infer on the replicability of signal across studies for each feature. In Bogomolov and Heller (2018) we suggest an approach that provides FDR control in each study separately as well as on replicability findings across both studies. When the method is applied at level 2α, the FDR is controlled in each study separately at level α. The discoveries may differ from those obtained if each study was analysed separately with the Benjamini–Hochberg procedure at level α, for the reason highlighted by the authors: auxiliary test statistics that tend to come from non-null hypotheses whenever $\theta_{1i} \neq 0$ aid in ranking the evidence against the null.

Edgar Dobriban (*University of Pennsylvania, Philadelphia*)

I congratulate Cai, Sun and Wang for this important contribution. They address the problem of large-scale two-sample testing, when the effect sizes (i.e. the differences in the effects in the two groups) are sparse. Often there is additional prior or side information about the effects that can improve power. The authors propose a general methodology and a concrete algorithm termed covariate-assisted ranking and screening (CARS) to exploit such information.

I would like to summarize the key idea as follows (following the notation from Section 4.1, roughly speaking): given a primary test statistic T (for instance the difference of the means of the two groups) and an auxiliary (or secondary) test statistic S (for instance a weighted combination of the means of the two groups), the CARS method estimates the probability $P_0(T, S)$ of the null hypothesis that the effects are equal, given specific values of S and T. Then an oracle method rejects those hypotheses for which this probability is small: $P_0(T, S) \leqslant c$. Under a specific mixture model, the authors show how to estimate everything to make this a practical procedure that controls the false discovery rate and has asymptotically optimal mean number of true discoveries.

The broad question here is how to improve power in multiple-hypothesis testing. We have contributed to this area by developing *weighted* multiple-testing procedures, where the weights are chosen *a priori* (Dobriban *et al*., 2015; Fortney *et al*., 2015; Dobriban, 2017). The authors mention in their discussion that 'The connection of CARS to theories on optimal weights is still an open issue'. We believe that this direction deserves further investigation. Our non-rigorous intuition is that weighting can sometimes be viewed as a special case of the class of procedures discussed here. Suppose that $w(S)$ is a weight depending on the auxiliary variable. If the auxiliary statistic S is independent of T under the null, then a rule of the form $G(T) \leqslant w(S)$, where G is a transform of the test statistic, can be viewed as a valid weighting rule. This is a special case of rules depending on T and S. Hence, heuristically, weights are a special case of the procedures of the form considered in the paper.

However, in contrast with most existing weighting methods, the optimal rules here must be estimated from the data at hand, which makes the problem different, and poses significant challenges. The connections to weighting may deserve more thought.

Jianqing Fan (*Princeton University*)

Professor Tony Cai, Professor Wenguang Sun and Dr Weinan Wang are wholeheartedly congratulated for important and stimulating contributions to the beautiful theory and elegant methods for large-scale two-sample inference with optimality guarantee.

An important contribution of the paper is to recognize that the sample difference $T_{1i} = \bar{X}_i - \bar{Y}_i$ is inadequate for testing $H_i : \mu^*_{xi} = \mu^*_{yi}$. This is understandable from the point of view of sufficient statistics. When we consider the power of the test, we have two mean parameters and sufficient statistics are $(\bar{X}_i, \bar{Y}_i)$ when the data are from normal distributions with known variances. This is equivalent to the authors' (T_{1i}, T_{2i}) where $T_{2i} = \bar{X}_i + \kappa/\bar{Y}_i$ $(\kappa \neq -1)$ and lends further support to their claim that the power can be enhanced by using the auxiliary variables. However, developing an optimal procedure like covariate-assisted ranking and screening (CARS) requires creative ideas and a substantial amount of work.

CARS utilizes the very intuitive oracle statistics $T^i_{\mathrm{OR}} = P(\theta_{1i} = 0 | T_{1i} = t_1, T_{2i} = t_2)$. This requires independent assumptions among m-variate vectors. This assumption can be too strong for high dimensional applications including genomics and finance. Dependent adjustments of data are necessary before the applications of CARS. Examples of dependent adjustments are given in Zhou *et al*. (2018) and Fan *et al*. (2018). After dependence adjustments, we can assume that adjusted data are weakly correlated. Can CARS be applied to weakly dependent data, rather than independent data? How robust is CARS to the weak dependence among m-variate vectors?

CARS relies on an estimate of $q^*(t)$. The authors provide a natural estimator (3.11). This estimator aggregates test statistics T_{2i} under true nulls. Under the true nulls, the test statistic T_{2i} depends on $\beta_{0i} + \mu^*_{xi}$. When these common means are very different across i, can the CARS procedure still be effective?

The authors use proposition 6 to indicate the applicability of CARS to non-normal data. The finite fourth-moment assumption is adequate for the asymptotic normality of (T_{1i}, T_{i2}) but not enough for uniform convergence of sample means when m is much larger than n. Robustifications such as those in Zhou *et al*. (2018) are needed. It will be very interesting to see the full development of CARS in more realistic settings with non-Gaussian, dependent and possibly heavy tail errors.

Jelle Goeman (*Leiden University Medical Center*) **and Aldo Solari** (*University of Milano-Bicocca, Milan*)

Auxiliary information, independent of all the null p-values but not of the non-null p-values, may be used to improve the power of multiple-testing procedures. This principle has been used in familywise error control and false discovery rate (FDR) control, leading to data-driven filtering or weighting (Kropf and Läuter, 2002; Westfall *et al*., 2004; Roeder and Wasserman, 2009; Bourgon *et al*., 2010; Ignatiadis *et al*., 2016; Pecanka *et al*., 2017). The main contribution of Cai, Sun and Wang to this literature is to present a data-driven procedure, similar to a weighted procedure, that is asymptotically optimal for the doubly sparse two-sample problem.

The classical choice of auxiliary information is the total variance. Cai, Sun and Wang chose to use a weighted sum of group means instead. Surprisingly, the total variance was not explicitly considered as a competitor, and we wonder why. Perhaps the sparsity assumption or the assumed heterogeneity of variances makes the use of the total variance unattractive. Heterogeneity, however, may be countered by per-group standardization as in the data example. An alternative for the total variance in the sparse case might be total variance around zero, i.e. simply the sum of all squared observations. Such auxiliary information would fit the general framework, and we are curious to see how it would perform.

The main comparison is with the Benjamini–Hochberg (BH) procedure and we see that CARS rejects more than that classical method. However, the superior performance comes at a high price. In the first place, it depends on two strong assumptions: sparsity and independence. Secondly, it replaces the exact FDR control of the BH method by only asymptotic control, requiring the number of tests to approach ∞. That CARS does not provide exact control can be seen in the toy example provided with the `CARS` function. With 200 tests and nominal 5% FDR, we obtain a disappointing 11% FDR. Asymptotic FDR control follows from consistent estimation of the FDR. To obtain consistent FDR estimation, (near) independence of the p-values is crucial. If stronger dependence is present, consistency, and therefore asymptotic control, may be lost since the variance of the estimator may not disappear when the testing problem grows. Independence or near independence could be appropriate for the astronomy example but is rare in biological data. The BH method remains one of very few methods that provides exact, rather than asymptotic, FDR control under realistic dependence and is still powerful.

Joshua Habiger (*Oklahoma State University, Stillwater*)

The main idea in the covariate-assisted ranking and screening method is that single statistics $T_1, T_2, \ldots, T_m$ for testing null hypotheses $H_{1,0}, H_{2,0}, \ldots, H_{m,0}$ may not be *sufficient*. The optimal oracle bivariate procedure is based on the local false discovery rate (FDR) or posterior probability

$$\text{lFDR}(t_i, s_i) = \Pr(H_{i,0} \text{ true } | T_i = t_i, S_i = s_i) = \frac{\Pr(H_{i,0} \text{ true }) f(t_i, s_i | H_{i,0} \text{ true})}{f(t_i, s_i)},$$

where S_i is an auxiliary covariate that provides further information regarding the distribution of the data when the null hypotheses are false through $f(t_i, s_i)$. It is show that this bivariate procedure dominates its univariate counterpart that utilizes $\text{lFDR}(t_i)$ (Sun and Cai, 2007) instead of $\text{lFDR}(t_i, s_i)$ for each test.

A discrete data example is given in Habiger *et al.* (2017) for the analysis of next generation sequencing data. Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$ be a collection of independent Poisson(μ_{ij}) random variables, with $\log(\mu_{ij}) = \beta_{0i} + \beta_{1i} x_j$ for x_j a covariate. The goal is to test $H_{i,0} : \beta_{1i} = 0$ for $i = 1, 2, \ldots, m$. Motivated by Sun and Cai (2007) and McCullagh and Nelder (1989), Habiger *et al.* (2017) proposed a conditional lFDR-procedure based on

$$\text{clFDR}(y_i, s_i) = \Pr(\beta_{1i} = 0 | Y_i = y_i, S_i = s_i) = \frac{\pi_0 \Pr(Y_i = y_i | S_i = s_i, \beta_{1i} = 0)}{\Pr(Y_i = y_i | S_i = s_i)},$$

where

$$\Pr(y_i = y_i | S_i = s_i) = \pi_0 \Pr(Y_i = y_i | s_i, \beta_{1i} = 0) + \sum_{k=1}^{K} \pi_k \Pr(Y_i = y_i | S_i = s_i, \beta_{1i} = \gamma_k).$$

Here, $\Pr(Y_i = y_i | S_i = s_i, \beta_{1i} = \gamma_k)$ is a multinominal probability mass function with parameters s_i and probability vector $p(\gamma_k) \propto (\exp(\gamma_k x_1), \exp(\gamma_k x_2), \ldots, \exp(\gamma_k x_n))$. Mixing proportions $\pi_0, \pi_1, \ldots, \pi_K$ and parameters $\gamma_1, \gamma_2, \ldots, \gamma_K$ are consistently estimated with the expectation–maximization algorithm, thereby facilitating a data-driven procedure.

In this example, the sufficient statistics for (β_{0i}, β_{1i}) are $(S_i, T_i) = (\Sigma_j Y_{ij}, \Sigma_j x_j Y_{ij})$, and T_i and S_i are discrete and dependent (even under the null hypothesis). Further discretizing the S_is or ignoring them would result in efficiency loss, but is the clFDR-statistic above based on (Y_i, S_i) or even (T_i, S_i) efficient? This seems to be related to the ancillarity paradox in remark 7 and comment 10.

Regarding the discussion in Section 4.6, the above finite mixture model may also facilitate the empirical null hypothesis (Efron, 2004) by simply relaxing the condition that $\beta_{1i} = 0$ under the null hypothesis, and allowing it to be estimated via maximum likelihood. However, it is also not clear how to select K above, or what to do if $\hat{\gamma}_k$ is approximately 0 for some k when utilizing the theoretical null.

The weighted p-value FDR literature (see Genovese *et al.* (2006), Roquain and Van De Wiel (2009), Peña *et al.* (2011) and Habiger (2017) among others) is similar in spirit to the current paper. For example,

the Benjamini–Hochberg or adaptive Benjamini–Hochberg procedure (Storey *et al.*, 2004) can be applied to weighted p-values $Q_i = P_i/w_i$. The main disadvantage of weighted p-value procedures is that optimal weights can be complex. However, an advantage is that a p-value is readily available for most hypothesis tests, and weights are robust (Habiger, 2017; Genovese *et al.*, 2006) in that, even if optimal oracle weights are not utilized or well estimated, some gain in power is still expected and FDR control is maintained. For example, Habiger (2017) showed that this robustness property enables a simple closed form expression for approximating optimal weights with $s_1, s_2, \ldots, s_m$, and demonstrated that some improvement over regular p-value procedures is still expected.

Jialiang Li (*National University of Singapore*) **and Weng Kee Wong** (*University of California at Los Angeles*)

We congratulate Cai, Sun and Wang for their interesting contributions on multiple-testing problems. We wish to comment on three practical issues related to the proposed development. First, the solution to the two-sample test problem may be generalized to multiple-group comparisons. In fact, multiclass comparison problems can usually be decomposed into a series of two-class comparison problems (one *versus* another, or one *versus* the rest) and then the two-sample Z-test or t-test that is discussed in the paper can be directly applied to each pair of groups. However, the so-constructed hypotheses may not be all independent and modelling the data will need to incorporate such latent dependence. This issue is slightly different from what was addressed in Section 4.2 where the m hypotheses were artificially divided into K heterogeneous groups. Second, it seems attractive to replace the parametric tests based on the Gaussian assumption by non-parametric tests for the location shift. In fact, assuming that a massive number of variables have a symmetric normal distribution can be very restrictive and unrealistic. For low dimensional problems, Wilcoxon test and other rank-based non-parametric tests are usually preferred as they offer robust results. We downloaded the microarray time course data which were analysed in section B.7 of the on-line supplementary file of the paper, and we ran Shapiro–Wilk tests on the $m = 22283$ genes. We noted that 9940 genes (44.6%) could not satisfy the normality assumption at the significance level 0.05. In the genetic literature more popular tests include those based on the empirical Bayes methods and moderated t-test with variance shrinkage. If one insists on using a Z-test or t-test as recommended in this paper, then our third point is that some normalization transformation may be helpful. In each comparison, only two distributions are involved and it is fairly easy to find a suitable monotone transformation such as the Box–Cox power transformation that is introduced in standard textbooks. However, an appropriate transformation for one hypothesis may not be appropriate for another. One therefore may need to find optimal transformations for the variables before testing all the m hypotheses. How to interpret the transformed data might also be an open issue.

Nicholas T. Longford (*Imperial College London*)

While applauding the authors' innovation in an area that is crowded with attention stimulated by emerging research themes, I want to highlight a profound weakness of the building block in their set-up, namely, the hypothesis test. After providing sterling service for many decades, the hypothesis test should be condemned to a statistical museum because it is hopelessly deficient for the needs of modern scientific endeavour. It is disqualified from purposeful inference by having no means of incorporating the consequences of the two kinds of inappropriate choices that the analyst may make: false discovery and failure to discover. Focus on one kind of error at the expense of the other is a gross scientific error; in addition to their minimization, the balance of the two kinds of error that reflects their ramifications is a more appropriate target.

Associated with this is the false dichotomy of the null hypothesis, with zero (or another 'special' value) pitted against the rest of the real axis. Without the indoctrination by the mechanics of the hypothesis test, the null hypothesis (1.1) should always be rejected because it is irrational to bet on any specific value of a parameter against an uncountable set of alternatives, uncountably many of which are arbitrarily close to the hypothesized value.

The consequences of the two kinds of error are difficult to establish, but the suggestion that they might be unimportant, and therefore the analysis could be oblivious to them, is difficult to sustain. If we continue with the practice of ignoring the consequences, our outputs will by definition remain inconsequential or will require considerable improvisation, or verbal massage, politely referred to as interpretation, to make them consequential.

The premise of this paper is the difficult calculus of p-values generated by multiple tests. The difficulty would evaporate if we switched to a framework in which the counterparts of the p-values are additive.

In decision theory, they are the expected gains or losses. These comments summarize the conclusions of Longford (2014).

However, I agree with Cai, Sun and Wang that hypothesis tests (and problems of making elementary decisions) are unequal both *a priori* and after drawing on auxiliary information.

Aaditya Ramdas (*Carnegie Mellon University, Pittsburgh*)

I discuss some recent work in the literature that is morally related to covariate-assisted ranking and screening (CARS). One common thread has been

(a) separate the available information about each hypothesis into two parts that are independent under the null and
(b) use one part to estimate and control the false discovery rate (FDR), and use the 'auxiliary information' to rank hypotheses, like a data-dependent prior, to gain power.

Through the lens of selective inference, these procedures can be seen as instances of 'data carving' (Fithian *et al.*, 2014). For example, approaches (a) and (b) can be seen as the spirit behind the knockoffs procedure (Barber and Candès, 2015), where only one bit of information per hypothesis (signs of knockoff statistics) suffice for FDR control, and the remaining bits (magnitudes of knockoff statistics) are used to sort hypotheses. The same idea was central in the design of the interactive AdaPT procedure (Lei and Fithian, 2018), where p-values are split into a single hidden bit $h(p) = I(p > \frac{1}{2})$ and a *masked* p-value $g(p) = \min\{p, 1-p\}$. Again, the former is used for estimating and controlling the FDR, and, along with additional covariate information, the latter is used to gain power by guiding the interactions. General constructions of data carving functions (h, g) were employed by the interactive STAR procedure (Lei *et al.*, 2017) which can additionally maintain any (possibly data-dependent, interactively discovered) structural constraints on the rejected set while controlling the FDR.

There are naturally numerous differences between CARS and the aforementioned work. CARS applies to high dimensional two-sample testing, knockoffs to high dimensional regression, and AdaPT and STAR to structured multiple testing with covariates. STAR and AdaPT carve the p-values themselves and are agnostic about how they were constructed, whereas CARS and knockoffs work with the raw data from scratch. (Technically, AdaPT and STAR can both work with knockoff statistics instead of p-values, which would equalize them all from this perspective.) Another difference is that AdaPT and STAR may use an assumed generative model, such as a covariate-based two-groups model, to relate the covariates to the p-values; however, they are robust to misspecification of the said model—if the model is completely wrong, FDR control still provably holds and only power is hurt. It is possible that CARS cannot (even asymptotically) control the FDR if the models assumed, like the bivariate random-mixture model, are wrong. A last technical difference lies in the proof techniques: knockoffs, AdaPT and STAR use martingale techniques to guarantee *non-asymptotic* FDR control, whereas the current paper does not use martingales and guarantees *asymptotic* FDR control. In fact, unlike CARS, many martingale-based methods like the above also satisfy strong *uniform post hoc* false discovery proportion guarantees (Katsevich and Ramdas, 2018).

There are other methods like SABHA (Li and Barber, 2019) that carve the p-values, but for brevity we do not discuss them. We suspect that we shall see additional use of the (a) plus (b) strategy in the coming years by using novel data carving techniques and other clever uses of ancillarity or sufficiency like Basu's theorem.

Qing Yang and Guang Cheng (*Purdue University, West Lafayette*)

We congratulate Cai, Sun and Wang for this inspiring work. We would like to evaluate the classification performance of the proposed covariate-adjusted ranking and screening (CARS) approach. This is in contrast with the recent studies on using classifiers to do statistical testing especially for high dimensional data, e.g. Friedman (2004), Ramdas *et al.* (2016) and Rosenblatt *et al.* (2016).

Consider two classes $N_m(\boldsymbol{\mu}_i, \mathbf{I}_m)$ and $N_m(\boldsymbol{\mu}_2, \mathbf{I}_m)$ with training data $\mathbf{X}_{m\times n_1}$ and $\mathbf{Y}_{m\times n_2}$. The well-known Fisher discriminant can perform as poorly as random guessing under high dimensionality (see Bickel and Levina (2004) and Fan and Fan (2008) among others). Hence, we first apply CARS to do marginal screening according to procedure 1; then we work only on those locations that reject H_0 (say d such positions). A new observation $\mathbf{z}$ is classified to class 1 if and only if

$$\left(\mathbf{z}_d - \frac{\bar{\mathbf{x}}_d + \bar{\mathbf{y}}_d}{2}\right)^{\mathrm{T}} \mathrm{diag}(\mathbf{S}_d)^{-1}(\bar{\mathbf{x}}_d - \bar{\mathbf{Y}}_d) > 0, \tag{5}$$

where the subscript d indicates the reduced parameters. Here $\text{diag}(\mathbf{S}_d)$ is used because this paper considers independent multiple tests. If we apply $\mathbf{S}_d$ in a general situation, an additive term

$$\frac{n_1+n_2-2}{n_1+n_2-d-3}\left(\frac{d}{2n_1}-\frac{d}{2n_2}\right)$$

is suggested to be added in condition (5) for offsetting the dimensionality effect, similarly to the rescaled terms in Yang and Cheng (2017).

Empirically, we also adopt the Benjamini and Hochberg (1995) procedure to select reduced locations. The population means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are generated as in simulation setting 1 of this paper.

Fig. 9 shows that method (5) enjoys comparable performance with the Bayes method. Moreover,

(a) the more powerful testing method ('CARS') leads to better classification performance and
(b) the gap between the two testings' powers roughly keeps stable, whereas that between the misclassification rates decreases rapidly to zero.

The positions are selected by controlling both methods at the same false discovery rate level. More positions are selected by the proposed CARS because of its larger power. Meanwhile, this increases the data dimension. So, it is not easy to tell which testing method leads to better classification outcomes before running simulations. To have a further comparison, in Fig. 10, the same number of locations is chosen by setting different false discovery rate values. It demonstrates a similar pattern, while the decreasing speed of the Benjamini–Hochberg method is a little faster. All these observations motivate one intriguing research direction—how do we quantify the relationship between testing power and misclassification rate?; what kind of statistical testing leads to optimal classification performance? These questions are more subtle for high dimensional sparse data.

Guo Yu (*University of Washington, Seattle*), **Jacob Bien** (*University of Southern California, Los Angeles*) **and Daniela Witten** (*University of Washington, Seattle*)

In this discussion contribution, we connect the elegant proposal of Cai, Sun and Wang to *multiview data*, in which multiple sets of variables (or 'views') are measured on the same observations. Using ideas from Section 4, we show that we can exploit a secondary view to improve power for testing on the first view.

Consider independent and identically distributed observations of m random variables under two conditions. In condition $l \in \{1, 2\}$, observation $i \in \{1, \ldots, n_l\}$ of variable $j \in \{1, \ldots, m\}$ is given by (view 1)

$$X_{ij}(l) = \mu_j(l) + \varepsilon_{ij}(l),$$

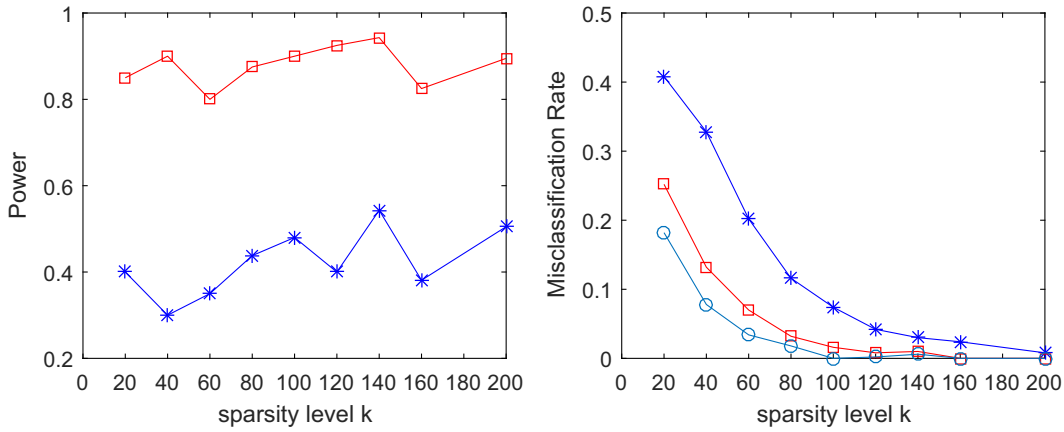


Fig. 9. (a) Power comparison and (b) empirical misclassification rates for two classes $N_m(\mu_1, \mathbf{I}_m)$ and $N_m(\mu_2, \mathbf{I}_m)$ based on 500 replications (FDR level $\alpha = 0.05$; $n_1 = 50$; $n_2 = 60$; $m = 1000$; $\mu_{1,1:k} = 5/\sqrt{30}$; $\mu_{1,(k+1):(2k)} = 4/\sqrt{30}$; $\mu_{1,(2k+1):m} = 0$; $\mu_{21:k} = 2/\sqrt{30}$; $\mu_{2,(k+1):(2k)} = 4/\sqrt{30}$; $\mu_{2,(2k+1):m} = 0$): ✳, method (5) based on Benjamini and Hochberg (1995); □, method (5) based on CARS; ○, Bayes rule

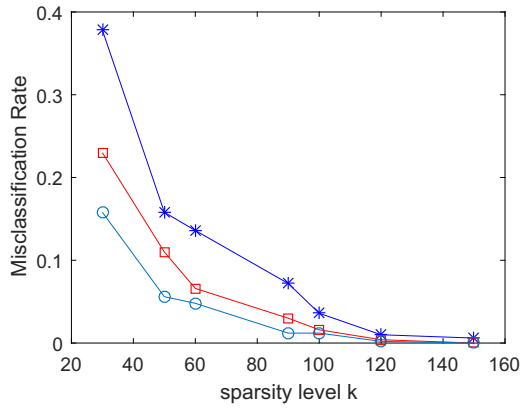


Fig. 10. Empirical misclassification rates when the same amount of locations are chosen for both methods: ✳, Benjamini and Hochberg (1995); □, CARS; ○, Bayes rule

where $\varepsilon_{ij}(l)$ is zero mean, and we suppress the common intercept. The random-mean vectors $\boldsymbol{\mu}(1)$ and $\boldsymbol{\mu}(2)$ are sparse. Furthermore, for the same individuals, we also observe a second view of $\tilde{m}$ variables (view 2):

$$Z_{ik}(l) = \tilde{\mu}_k(l) + \tilde{\varepsilon}_{ik}(l) \qquad \text{for } k \in \{1, \ldots, \tilde{m}\}.$$

The mean vectors $\tilde{\boldsymbol{\mu}}(l)$ are sparse, $\tilde{\varepsilon}_{ik}(l)$ is zero mean and again we suppress the intercept. Suppose that the two views satisfy a hierarchical sparsity constraint: for $j \in \{1, \ldots, m\}$ and $l \in \{1, 2\}$,

$$\tilde{\mu}_{\sigma(j)}(l) = 0 \Rightarrow \mu_j(l) = 0, \tag{6}$$

where $\sigma(j)$ maps the jth entry of $\boldsymbol{\mu}(l)$ to its parent in $\tilde{\mu}(l)$: Fig. 11.

Concretely, suppose that $X(l)$ and $Z(l)$ contain protein and gene expression measurements respectively. If transcripts that encode the jth protein are absent (i.e. $\tilde{\mu}_{\sigma(j)}(l) = 0$), then the jth protein cannot be present (i.e. $\mu_j(l) = 0$).

Suppose that $(\mu_j(1), \tilde{\mu}_{\sigma(j)}(1))$ is independent of $(\mu_j(2), \tilde{\mu}_{\sigma(j)}(2))$. Further assume that the random errors $(\varepsilon_{ij}(l), \tilde{\varepsilon}_{i\sigma(j)}(l))$ are bivariate normal and independent across j, l and i, and independent of $\boldsymbol{\mu}(l)$ and $\tilde{\boldsymbol{\mu}}(l)$.

Using the terminology of Cai, Sun and Wang the 'primary statistic' for testing $H_{0j} : \mu_j(1) = \mu_j(2)$ is

$$T_j = C_j\{\bar{X}_j(1) - \bar{X}_j(2)\}$$

for some constant C_j. We consider a pair of 'auxiliary statistics',

$$R_j = D_j\left[\bar{X}_j(1) + \frac{n_2 \mathrm{var}\{\varepsilon_{ij}(1)\}}{n_1 \mathrm{var}\{\varepsilon_{ij}(2)\}}\bar{X}_j(2)\right],$$

$$S_j = E_j\left[\bar{Z}_{\sigma(j)}(1) + \frac{n_2 \mathrm{cov}\{\varepsilon_{ij}(1), \tilde{\varepsilon}_{i\sigma(j)}(1)\}}{n_1 \mathrm{cov}\{\varepsilon_{ij}(1), \tilde{\varepsilon}_{i\sigma(j)}(2)\}}\bar{Z}_{\sigma(j)}(2)\right],$$

for some constants D_j and E_j. The statistic R_j is the same as T_{2j} in the paper, whereas S_j is constructed by using the second data view. A small value of $|S_j|$ provides evidence for $\tilde{\mu}_{\sigma(j)}(1) = \tilde{\mu}_{\sigma(j)}(2) = 0$, which by constraint (6) suggests that $\mu_j(1) = \mu_j(2)$. By analogy with proposition 1 in the paper, the oracle statistic is

$$T^{(j)}_{\mathrm{OR}}(t_j, r_j, s_j) \equiv \Pr(\theta_{1j} = 0 | T_j = t_j, R_j = r_j, S_j = s_j) = \frac{f(t_j, r_j, s_j | \theta_{1j} = 0)\Pr(\theta_{1j} = 0)}{f(t_j, r_j, s_j)}$$
$$= \frac{f(t_j | \theta_{1j} = 0) f(r_j, s_j | \theta_{1j} = 0)\Pr(\theta_{1j} = 0)}{f(t_j, r_j, s_j)}.$$

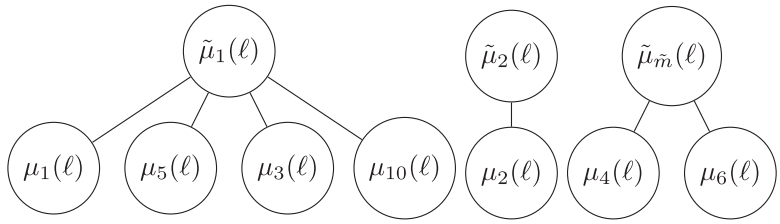


Fig. 11. Schematic diagram of constraint (6) with $\sigma(3) = 1$

Moreover, $T^{(j)}_{\mathrm{OR}}(t_j, r_j, s_j)$ enjoys the properties in theorem 3 of the paper. Detailed proofs are available from `https://hugogogo.github.io/paper/cars_discussion_supplement.pdf`. If there is not a one-to-one mapping between $\sigma(j)$ and j, then $T^{(j)}_{\mathrm{OR}}(t_j, r_j, s_j)$ must be estimated carefully.

The **authors** replied later, in writing, as follows.

We thank the discussants for their insightful comments and excellent contributions. It is our great delight to meet some discussants in London, and we are pleased to participate in further discussions in writing. The discussions are wide ranging. For brevity, we focus only on some key topics.

Key message: data reduction, information loss and optimality
Data reduction via constructing linear contrasts has long been used as an essential tool for statistical analyses. Examining the process at a high level, the conventional practice involves first dividing raw data into 'relevant' and 'irrelevant' parts (or data carving, per Professor Ramdas), and then developing inference procedures based solely on the summary of the relevant part. This practice is widespread in statistical analysis. A major surprise is that such standard practices in data processing could lead to significant information loss in large-scale inference. Our work marks a clear departure from the existing work where auxiliary information is gleaned from external data. We propose new strategies to extract structural information *within the same sample* by using auxiliary covariates. We thank Professor Fan for the comments on our contributions to the optimality theory in false discovery rate FDR control, which has been lacking in the literature. This is an important direction in large-scale inference, considering that optimality has been the goal in the development of many fundamental results in statistics including Fisher's theory on the asymptotic efficiency of maximum likelihood estimation and the Neyman–Pearson lemma on the optimality of the likelihood ratio test.

*Structural information in high dimensional inferenc*e
We appreciate the illuminating comments from Roquain and Nichols on the role of auxiliary data in amplifying the signals. From a decision theoretic view, classical ideas such as Robbins's compound decision theory and the James–Stein shrinkage estimator show that the joint structure of primary statistics can be exploited to construct more efficient estimation or testing rules. A key message conveyed through this work is that *extra* valuable structural information can be extracted from the seemingly irrelevant part of the data. This point is particularly crucial in high dimensional settings. When the number of parameters is small, the information loss is inconsequential (since the joint structure cannot be estimated well). However, in the case with thousands of parameters structural information can be recovered with good precision from auxiliary statistics, which can play a key role in improving the power.

The sufficiency principle and broad applicability of covariate-assisted ranking and screening
We concur with Fan and Habiger in their insightful comments on *sufficient statistics*: a fundamental principle that seems to have been largely ignored in the FDR-literature. The comments also shed new lights on the applicability of covariate-assisted ranking and screening (CARS) beyond the case that requires doubly sparse means. The general idea in CARS works for a broad class of bivariate models (see Section 4.1 and our remark 7) and the doubly sparse assumption should be viewed as a special setting to explain intuitively why CARS works. Moreover, violations of the sufficiency principle are common in data processing and CARS can benefit from a non-sparse auxiliary sequence as long as the covariate encodes useful structural information. These points have been nicely corroborated by Professor Habiger's several interesting papers on heterogeneous discrete data. Professor Habiger's inspiring discussion also points the way forward for developing effective data combination strategies across qualitative and quantitative variables.

Using covariate-assisted ranking and screening in other high dimensional problems

CARS provides a generic tool for inferring sparsity structure by integrating evidence from multiple sources. The interesting discussions by Bogomolov, Heller and Yekutieli, Yu, Bien and Witten, Yang and Cheng, Li and Wong, and Banerjee and Mukherjee, among others, show that CARS has considerable potential for providing better solutions to a wide range of high dimensional problems including large-scale analysis-of-variance tests, high dimensional replicability analysis, sparse linear discriminant analysis, multiview analysis, hierarchical inference and sparse compound estimation. It is encouraging to see some preliminary successes reported by the discussants. We feel that revisiting the fundamental sufficiency principle in large-scale inference and carefully investigating possible information loss in data reduction would be an important and fruitful direction for future research. We appreciate the creative ideas and stimulating comments from the discussants on applying CARS to various high dimensional inference problems. We very much look forward to further explorations along these lines.

The dependent case and the 'grouping, adjusting and pooling' procedure

Fan, Goeman and Solari expressed legitimate concerns on the independence assumption. Although the robustness of CARS under dependence has been investigated numerically in the on-line appendix B.5, we take this opportunity to describe briefly our recent work aiming to address the important dependence issue. Xia *et al*. (2018) developed a general information pooling framework that involves grouping, adjusting and pooling (GAP) to leverage the structural information from an auxiliary sequence. GAP is built on the Benjamini–Hochberg (BH) procedure and utilizes weighted p-values to capture the heterogeneity among hypotheses. We generalize the weighted multiple-testing theory in Genovese *et al*. (2006) to show that GAP controls FDR under a range of dependence structures, including weakly dependent tests arising from high dimensional linear regression and Gaussian graphical models. However, the optimal choice of weights is still an open issue that deserves more research; inspiring discussions can be found in the comments by Dobriban, Ramdas and Habiger on use of weighted p-values and interactive use of masked p-values.

Asymptotic false discovery rate control and variations of the Benjamini–Hochberg procedure

CARS and Lfdr-methods offer asymptotic FDR-control and work better for large-scale testing problems where the density functions can be well estimated. By contrast, the BH procedure offers guaranteed FDR-control under a range of dependence structures. For smaller-scale problems with a few dozen or several hundred tests as considered by Goeman and Solari, we recommend GAP and other variations of the BH procedure (see the discussions by Dobriban, Ramdas and Habiger) to incorporate useful side information. It would be of great interest to investigate the performance of Bayesian CARS (see Professor Roquain's comments) to increase the stability in small sample settings.

The sparse case and updated covariate-assisted ranking and screening package

We thank Roquain and Nichols for noticing the issues of our CARS package under the very sparse case. We have uploaded to the Comprehensive R Archive Network the updated package that includes the 'sparse option' described in Section 5.1 and a new section on the vignette illustrating that CARS, using the sparse option, controls FDR when $m = 10000$ and $k = 10$: a setting considered by Professor Nichols. The key idea for the sparsity adjustment is to use the known densities to stabilize the bivariate density estimate in regions with few observations. Through communication with Professor Mark van de Wiel, we recognize that, for methods based on CARS and Lfdr, the instability of the non-parametric density estimator (in the denominator) seems to be a common issue. In the sparse regime, Professor Roquain's proposal of employing Cauchy slab priors is a promising direction with the potential of having the best of both worlds: the method avoids non-parametric modelling of a bivariate density, while the choice of priors has great promise of leading to good frequentist properties.

On choosing the auxiliary sequence

We briefly address the interesting question from Goeman and Solari whether the total variance could be a good competitor as an auxiliary variable. First, it can be shown that, with known and homoscedastic variances, the pair (T_1, T_2) is a sufficient statistic (per Professor Fan); hence T_2 is optimal in the sense that it has no information loss. Although the sufficiency principle may be satisfied by other pairs, our choice of T_2 not only is intuitively appealing but also simplifies the development of both methodology and theory; see the discussion in Section 2.1. Second, an important consideration in choosing the auxiliary variable is to avoid selection bias. As noted in a post by Professor Ryan Tibshirani on the 'Normal deviate' blog, screening based on between-group variance leads to severe selection bias. The total variance is not promising either, at least under the CARS framework, because under heteroscedasticity it is correlated with

the primary statistic and cannot capture the sparsity structure effectively. Moreover, the total variances are not suitable as useful structures to inform BH algorithms. The p-value null distribution is likely to be distorted when screening, grouping or weighting is carried out via total variance.

Open issues and concluding remarks

Large-scale multiple testing is a fundamental building block in contemporary statistics and developing efficient procedures that control the FDR, a celebrated innovation in the past two decades, has been a prominent and impactful research area. Although the hypothesis testing framework is not omnipotent as pointed out by Professor Longford, we believe that some concerns may be possibly addressed by tailoring the general FDR-concept to the needs of specific applications; notable ideas include weighted FDR (Benjamini and Hochberg, 1997), directional FDR (Guo *et al.*, 2010) and the false important discovery rate (Sun and McLain, 2012). As pointed out by Medina and Stehlík, the null hypothesis should be carefully formulated, and existing methods should be properly modified for specific applications. Much more research is still needed in this area.

References in the discussion

Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., Sklar, P., Roddey, J. C., Chen, C.-H., McEvoy, L., Desikan, R. S., Djurovic, S., Dale, A. M., Psychiatric Genomics Consortium and Bipolar Disorder and Schizophrenia Working Groups (2013) Improved detection of common variants associated with schizophrenia and polar disorder using pleiotropy-informed conditional false discovery rate. *PLOS Genet.*, **9**, article e1003455.

Banerjee, T., Mukherjee, G. and Sun, W. (2018) Adaptive sparse estimation with side information. *Preprint arXiv: 1811.11930*.

Barber, R. F. and Candès, E. J. (2015) Controlling the false discovery rate via knockoffs. *Ann. Statist.*, **43**, 2055–2085.

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 hypothesis testing with weights. *Scand. J. Statist.*, **24**, 407–418.

Bickel, P. J. and Levina, E. (2004) Some theory for Fisher's linear discriminant function, naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989–1010.

Bogomolov, M. and Heller, R. (2018) Assessing replicability of findings across two studies of multiple features. *Biometrika*, **105**, 505–516.

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. (2008) In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Statist.*, **2**, 113–152.

Cai, T. and Sun, W. (2017) Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. *J. R. Statist. Soc.* B, **79**, 197–223.

Castillo, I. and Mismer, R. (2018) Empirical Bayes analysis of spike and slab posterior distributions. *Electron. J. Statist.*, **12**, 3953–4001.

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

Dobriban, E. (2017) Weighted mining of massive collections of p-values by convex optimization. *Informn Inf.*, **7**, 251–275.

Dobriban, E., Fortney, K., Kim, S. K. and Owen, A. B. (2015) Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika*, **102**, 753–766.

Donoho, D. L. and Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.

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

Fan, J. and Fan, Y. (2008) High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.

Fan, J., Ke, Y., Sun, Q. and Zhou, W.-X. (2018) FarmTest: factor-adjusted robust multiple testing with false discovery control. *J. Am. Statist. Ass.*, to be published.

Fithian, W., Sun, D. and Taylor, J. (2014) Optimal inference after model selection. *Preprint arXiv:1410.2597*. University of California at Berkeley, Berkeley.

Fortney, K., Dobriban, E., Garagnani, P., Pirazzini, C., Monti, D., Mari, D., Atzmon, G., Barzilai, N., Franceschi, C., Owen, A. B. and Kim, S. K. (2015) Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLOS Genet.*, **11**, no. 12, article e1005728.

Friedman, J. (2004) On multivariate goodness-of-fit and two-sample testing. *Report SLAC-PUB-10325*. Stanford Linear Accelerator Center, Menlo Park.

Genovese, C. R., Roeder, K. and Wasserman, L. (2006) False discovery control with p-value weighting. *Biometrika*, **93**, 509–524.
Guo, W., Sarkar, S. K. and Peddada, S. D. (2010) Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics*, **66**, 485–492.
Habiger, J. D. (2017) Adaptive false discovery rate control for heterogeneous data. *Statist. Sin.*, **27**, 1731–1756.
Habiger, J., Watts, D. and Anderson, M. (2017) Multiple testing with heterogeneous multinomial distributions. *Biometrics*, **73**, 562–570.
Heller, R. and Yekutieli, D. (2014) Replicability analysis for genome-wide association studies. *Ann. Appl. Statist.*, **8**, 481–498.
Ignatiadis, N., Klaus, B., Zaugg, J. B. and Huber, W. (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Meth.*, **13**, 577.
Johnstone, I. M. and Silverman, B. W. (2004) Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, **32**, 1594–1649.
Johnstone, I. M. and Silverman, B. W. (2005) Ebayes Thresh: R programs for empirical Bayes thresholding. *J. Statist. Softwr.*, **12**, no. 8.
Katsevich, E. and Ramdas, A. (2018) Towards 'simultaneous selective inference': post-hoc bounds on the false discovery proportion. *Preprint arXiv:1803.06790*. Stanford University, Stanford.
Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. and Baker, C. I. (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neursci.*, **12**, 535–540.
Kropf, S. and Läuter, J. (2002) Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. *Biometr. J.*, **44**, 789–800.
Lei, L. and Fithian, W. (2018) AdaPT: an interactive procedure for multiple testing with side information. *J. R. Statist. Soc.* B, **80**, 649–679.
Lei, L., Ramdas, A. and Fithian, W. (2017) STAR: a general interactive framework for FDR control under structural constraints. *Preprint arXiv:1710.02776*. University of California at Berkeley, Berkeley.
Li, A. and Barber, R. F. (2019) Multiple testing with the structure adaptive Benjamini–Hochberg algorithm. *J. R. Statist. Soc.* B, **81**, 45–74.
Longford, N. T. (2014) A decision-theoretical alternative to testing many hypotheses. *Biostatistics*, **15**, 154–169.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Michaelson, J. J., Loguercio, S. and Beyer, A. (2009) Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, **48**, 265–276.
Pecanka, J., Jonker, M. A., IPDGC, Bochdanovits, Z. and Van Der Vaart, A. W. (2017) A powerful and efficient two-stage method for detecting gene-to-gene interactions in GWAS. *Biostatistics*, **18**, 477–494.
Peña, E., Habiger, J. and Wu, W. (2011) Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *Ann. Statist.*, **39**, 556–583.
Ramdas, A., Singh, A. and Wasserman, L. (2016) Classification accuracy as a proxy for two sample testing. *Preprint arXiv:1602.02210*.
Roeder, K. and Wasserman, L. (2009) Genome-wide significance levels and weighted hypothesis testing. *Statist. Sci.*, **24**, 398–413.
Roquain, E. and van de Wiel, M. A. (2009) Optimal weighting for false discovery rate control. *Electron. J. Statist.*, **3**, 678–711.
Rosenblatt, J. D., Benjamini, Y., Gilron, R., Mukamel, R. and Goeman, J. J. (2016) Better-than-chance classification for signal detection. *Preprint arXiv:1608.08873*.
Storey, J. D., Taylor, J. E. and Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc.* B, **66**, 187–205.
Sun, W. and Cai, T. T. (2007) Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Statist. Ass.*, **102**, 901–912.
Sun, W. and McLain, A. C. (2012) Multiple testing of composite null hypotheses in heteroscedastic models. *J. Am. Statist. Ass.*, **107**, 673–687.
Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E. and Ugurbil, K. (2013) The WU-Minn Human Connectome Project: an overview. *NeuroImage*, **80**, 62–79.
Westfall, P. H., Kropf, S. and Finos, L. (2004) *Weighted FWE-controlling Methods in High-dimensional Situations*, pp. 143–154. Beachwood: Institute of Mathematical Statistics.
Xia, Y., Cai, T. T. and Sun, W. (2018) GAP: a general framework for information pooling in two-sample sparse inference. *Technical Report*. Fudan University, Shanghai. (Available from `http://www.bcf.usc.edu/~wenguans/Papers/GAP.pdf`.)
Xie, X., Kou, S. and Brown, L. D. (2012) Sure estimates for a heteroscedastic hierarchical model. *J. Am. Statist. Ass.*, **107**, 1465–1479.
Yang, Q. and Cheng, G. (2018) Quadratic discriminant analysis under moderate dimension. *Preprint arXiv:1808.10065*. Purdue University, West Lafayette.
Zhou, W.-X., Bose, K., Fan, J. and Liu, H. (2018) A new perspective on robust M-estimation: finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.*, **46**, 1904–1931.