False Discovery Rate Control Under General Dependence By Symmetrized Data Aggregation

Lilun Du[1], Xu Guo[2], Wenguang Sun[3] and Changliang Zou[4]

Abstract

We develop a new class of distribution–free multiple testing rules for false discovery rate (FDR) control under general dependence. A key element in our proposal is a symmetrized data aggregation (SDA) approach to incorporating the dependence structure via sample splitting, data screening and information pooling. The proposed SDA filter first constructs a sequence of ranking statistics that fulfill global symmetry properties, and then chooses a data–driven threshold along the ranking to control the FDR. The SDA filter substantially outperforms the knockoff method in power under moderate to strong dependence, and is more robust than existing methods based on asymptotic p-values. We first develop finite–sample theories to provide an upper bound for the actual FDR under general dependence, and then establish the asymptotic validity of SDA for both the FDR and false discovery proportion (FDP) control under mild regularity conditions. The procedure is implemented in the R package `sdafilter`. Numerical results confirm the effectiveness and robustness of SDA in FDR control and show that it achieves substantial power gain over existing methods in many settings.

Keywords: Empirical distribution; Integrative multiple testing; Moderate deviation theory; Sample-splitting; Uniform convergence.

[1]Hong Kong University of Science and Technology, Hong Kong
[2]Beijing Normal University, Beijing, China
[3]University of Southern California. Corresponding Email: wenguans@marshall.usc.edu.
[4]Nankai University, Tianjin, China

1 Introduction

Multiple testing provides a useful approach to identifying sparse signals from massive data. Recent developments on false discovery rate (FDR; Benjamini and Hochberg, 1995) methodologies have greatly influenced a wide range of scientific disciplines including genomics (Tusher et al., 2001; Roeder and Wasserman, 2009), neuroimaging (Pacifico et al., 2004; Schwartzman et al., 2008), geography (Caldas de Castro and Singer, 2006; Sun et al., 2015) and finance (Barras et al., 2010). Conventional FDR procedures, such as the Benjamini–Hochberg (BH) procedure, adaptive p-value procedure (Benjamini and Hochberg, 1997) and adaptive z-value procedure based on local FDR (Efron et al., 2001; Sun and Cai, 2007), are developed under the assumption that the test statistics are independent. However, data arising from large–scale testing problems are often dependent. FDR control under dependence is a critical problem that requires much research. Two key issues include (a) how the dependence may affect existing FDR methods, and (b) how to properly incorporate the dependence structure into inference.

1.1 FDR control under dependence

The impact of dependence on FDR analysis was first investigated by Benjamini and Yekutieli (2001), who showed that the BH procedure, when adjusted at level $\alpha/(\sum_{j=1}^{p} 1/j)$ with p being the number of tests, controls the FDR at level α under arbitrary dependence among the p-values. However, this adjustment is often too conservative in practice. Benjamini and Yekutieli (2001) further proved that applying BH without any adjustment is valid for FDR control for correlated tests satisfying the PRDS property. This result was strengthened by Sarkar (2002), who showed that the FDR control theory under positive dependence holds for a generalized class of step-wise methods. Storey et al. (2004), Wu (2008) and Clarke and Hall (2009) respectively showed that, in the asymptotic sense, BH is valid under weak dependence, Markovian dependence and linear process models. Although controlling the FDR does not always require independence, some key quantities in FDR analysis, such as the expectation and variance of the number of false positives, may possess substantially different properties under dependence (Owen, 2005; Finner et al., 2007). This implies that conventional FDR methods such as BH can suffer from low power and high variability under strong

dependence. Efron (2007) and Schwartzman and Lin (2011) showed that strong correlations degrade the accuracy in both estimation and testing. In particular, positive/negative correlations can make the empirical null distributions of z-values narrower/wider, which has substantial impact on subsequent FDR analyses. These insightful findings suggest that it is crucial to develop new FDR methods tailored to capture the structural information among dependent tests.

Intuitively high correlations can be exploited to aggregate weak signals from individuals to increase the signal to noise ratio (SNR). Hence informative dependence structures can become a bless for FDR analysis. For example, the works of Benjamini and Heller (2007), Sun and Cai (2009) and Sun and Wei (2011) showed that incorporating functional, spatial, and temporal correlations into inference can improve the power and interpretability of existing methods. However, these methods are not applicable to general dependence structures. Efron (2007), Efron (2010) and Fan et al. (2012) discussed how to obtain more accurate FDR estimates by taking into account arbitrary dependence. For a general class of dependence models, Leek and Storey (2008), Friguet et al. (2009), Fan et al. (2012) and Fan and Han (2017) showed that the overall dependence can be much weakened by subtracting the common factors out, and factor–adjusted p-values can be employed to construct more powerful FDR procedures. The works by Hall and Jin (2010), Jin (2012) and Li and Zhong (2017) showed that, under both the global testing and multiple testing contexts, the covariance structures can be utilized, via transformation, to construct test statistics with increased SNR, revealing the beneficial effects of dependence. However, the above methods, for example by Fan and Han (2017) and Li and Zhong (2017), rely heavily on the accuracy of estimated models and the asymptotic normality of the test statistics. Under the finite–sample setting, poor estimates of model parameters or violations of normality assumption may lead to less powerful and even invalid FDR procedures. This article aims to develop a robust and assumption–lean method that effectively controls the FDR under general dependence with much improved power.

1.2 Model and problem formulation

We consider a setup where p-dimensional vectors $\xi_i = (\xi_{i1}, \ldots, \xi_{ip})^\top$, $i = 1, \cdots, n$, follow a multivariate distribution with mean $\mu = (\mu_1, \ldots, \mu_p)^\top$ and covariance matrix Σ. The problem of interest

is to test p hypotheses simultaneously:

$$H_j^0 : \mu_j = 0 \quad \text{versus} \quad H_j^1 : \mu_j = 0, \quad \text{for } j = 1, \ldots, p.$$

The summary statistic $\bar{\xi} = n^{-1} \sum_{i=1}^{n} \xi_i$ obeys a multivariate normal (MVN) model asymptotically

$$\bar{\xi} \overset{d}{\approx} \text{MVN}(\boldsymbol{\mu}, n^{-1}\Sigma). \tag{1}$$

Denote $\Omega = \Sigma^{-1}$ the precision matrix. We first assume that Ω is known. For the case with unknown precision matrix, a data-driven methodology and its theoretical properties are discussed in Section 4. The problem of multiple testing under dependence can be recast as a variable selection problem in linear regression. Specifically, by taking a "whitening" transformation, Model (1) is equivalent to the following model:

$$Y = X\boldsymbol{\mu} + \varepsilon, \qquad \varepsilon \overset{d}{\approx} \text{MVN}(0, n^{-1} I_p), \tag{2}$$

where $Y = \Omega^{1/2}\bar{\xi} \in R^p$ is the pseudo response, $X = \Omega^{1/2} \in R^{p \times p}$ is the design matrix, I_p is a p-dimensional identity matrix and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)^{>}$ are noise terms that are approximately inde-pendent and normally distributed. The connection between model selection and FDR was discussed in Abramovich et al. (2006) and Bogdan et al. (2015), respectively under the normal means model and regression model with orthogonal designs.

Let $\theta_j = I\{\mu_j = 0\}$, $j = 1, \cdots, p$, where I is an indicator function, and $\theta_j = 0/1$ corresponds to a null/non-null variable. Let $\delta_j \in \{0, 1\}$ be a decision, where $\delta_j = 1$ indicates that H_j^0 is rejected and $\delta_j = 0$ otherwise. Let $A = \{j : \mu_j = 0\}$ denote the non–null set and $A^c = \{1, \cdots, p\} \setminus A$ the null set. The set of coordinates selected by a multiple testing procedure is denoted $A^b = \{j : \delta_j = 1\}$. Define the false discovery proportion (FDP) and true discovery proportion (TDP) as:

$$\text{FDP} = \frac{\sum_{j=1}^{p}(1-\theta_j)\delta_j}{(\sum_{j=1}^{p}\delta_j) \vee 1}, \quad \text{TDP} = \frac{\sum_{j=1}^{p}\theta_j\delta_j}{(\sum_{j=1}^{p}\theta_j) \vee 1}, \tag{3}$$

where $a \vee b = \max(a, b)$. The FDR is the expectation of the FDP: $\text{FDR} = E(\text{FDP})$. The average power is defined as $\text{AP} = E(\text{TDP})$.

1.3 FDR control by symmetrized data aggregation

This article introduces a new information pooling strategy, the symmetrized data aggregation (SDA), for handling the dependence issue in multiple testing. The SDA involves splitting and re-

assembling data to construct a sequence of statistics fulfilling symmetry properties. Our proposed SDA filter for FDR control consists of three steps:

- The first step splits the sample into two parts, both of which are utilized to construct statistics to assess the evidence against the null.
- The second step aggregates the two statistics to form a new ranking statistic fulfilling symmetry properties.
- The third step chooses a threshold along the ranking by exploiting the symmetry property between positive and negative null statistics to control the FDR.

To get intuitions on how the idea works, we start with the independent case [Zou et al. (2020)]. The more interesting but complicated dependent case will be described shortly, with detailed discussions, refinements and justifications deferred to later sections. Suppose the vectors $\xi_i = (\xi_{i1}, \ldots, \xi_{ip})^\top$ are i.i.d. obeying $\mathrm{MVN}(\boldsymbol{\mu}, I_p)$. The proposed SDA method first splits the full sample into two disjoint subsets D_1 and D_2, with sizes n_1 and n_2 and $n = n_1 + n_2$. A pair of statistics, both of which follow $N(0,1)$ under the null, are then calculated to test H_j^0:

$$(T_{1j}, T_{2j}) = \left(\frac{\sum_{i \in D_1} \xi_j}{\sqrt{n_1}}, \frac{\sum_{i \in D_2} \xi_j}{\sqrt{n_2}} \right).$$

The product $W_j = T_{1j} T_{2j}$ is used to aggregate the evidence across the two groups. If $|\mu_j|$ is large, then both T_{1j} and T_{2j} tend to have large absolute values with the same sign, thereby leading to a positive and large W_j. By contrast, W_j fulfills the symmetry property under H_j^0, i.e.

$$\Pr(W_j \geq t \mid H_j^0) = \Pr(W_j \leq -t \mid H_j^0), \text{ for any } t > 0. \tag{4}$$

This motivates one to consider the following selection procedure $A = \{j : W_j \geq L\}$, where L is the threshold chosen to control the FDR at level α:

$$L = \inf \left\{ t > 0 : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \alpha \right\}. \tag{5}$$

According to the symmetry property (4), the count of negative W_j's below $-t$ strongly resembles the count of false positives in the selected subset (i.e. the null W_j's above t). It follows that the fraction in Equation (5) provides a good estimate of the FDP.

The dependent case involves a more carefully designed SDA filter. After sample splitting, we apply variable selection techniques such as LASSO to D_1 to construct T_{1j}. T_{1j}, which is calculated based on linear model (2), can effectively capture the dependence structure. Before using D_2 to construct T_{2j}, we carry out a data screening step to narrow down the focus. We show that the screening step can significantly increase the SNR of T_{2j} under strong dependence, hence the correlations are exploited again to increase the power. The ranking statistic W_j is constructed by combining T_{1j} and T_{2j} with proven asymptotic symmetry properties. The theory of the proposed SDA filter is divided into two parts: the finite sample theory provides an upper bound for the FDR under general dependence, while the asymptotic theory shows that both the FDR and FDP can be controlled at $\alpha + o(1)$ under mild regularity conditions.

1.4 Connections to existing work and our contributions

The SDA is closely related to existing ideas of sample–splitting (Wasserman and Roeder, 2009; Meinshausen et al., 2009) and data carving (Fithian et al., 2014; Lei et al., 2021), both of which firstly divide the data into two independent parts, secondly use one part to narrow down the focus (or rank the hypotheses) and finally use the remainder to perform inference tasks such as variable selection, estimation or multiple testing. These ideas have a common theme with covariate–assisted multiple testing (Lei and Fithian, 2018; Cai et al., 2019; Li and Barber, 2019), where the primary statistic plays the key role to assess the significance while the side information plays an auxiliary role to assist inference [see also the discussion by Ramdas (2019)]. SDA provides a novel way of data aggregation where both parts of data, which are combined under the symmetry principle, play essential roles in both ranking and selection. This substantially reduces the information loss in conventional sample–splitting methods, while the symmetry principle, which is fulfilled by construction, enables the development of an effective and assumption-lean FDR filter.

The SDA is inspired by the elegant knockoff filter for FDR control (Barber and Candès, 2015), which creates knockoff features that emulate the correlation structure in original features, to form symmetrized ranking statistics for selecting important variables via the same mechanism (5). The knockoff method, which is originally developed under regression models, can be applied for FDR

control in Model (1) via the equivalent Model (2). The knockoff filter employs local pairwise contrasts: the ranking variable is constructed to capture the differential evidences against the null exhibited by the pair (i.e. the original feature vs. its knockoff copy). While it is desirable to make the pair as "independent" as possible, high correlations will greatly restrict the geometric space in which the knockoff can be constructed; see Appendix B.1 for detailed discussions and illustrations. This would significantly increase the difficulty for distinguishing the variable and its knockoff and hence lower the power. By contrast, the SDA filter, which does not rely on pairwise contrasts, will not suffer from high correlations.

To visualize the correlation effects, we consider a setup similar to Figure 5 in Barber and Candès (2015), where correlated normal, t, and exponential data are generated based on an autoregressive model $\Sigma = (\rho^{|j-i|})$ (see Section 5.2 for more details about the setup). We vary ρ from -0.9 to 0.9 and apply BH, knockoff and SDA at FDR level $\alpha = 0.2$. The actual FDRs and APs based on 500 replications are summarized in Figure 1. Our first column (normal data) shows that knockoff outperforms BH in some situations, but both the FDR and AP of the knockoff method decrease when correlations grow higher. By contrast, SDA controls the FDR near the nominal level consistently, and the power of SDA increases sharply with growing correlations. This pattern corroborates the insights by Benjamini and Heller (2007), Sun and Cai (2009) and Hall and Jin (2010) that high correlations, which can be exploited to increase the SNR, may become a bless in large–scale inference.

The proposed research improves the previous work by Zou et al. (2020) in several ways. First, Zou et al. (2020) has mainly focused on the independent and weak dependent case, with the major goal of deriving convergence rate of false discovery proportions when simultaneously performing thousands of t-tests. The methodology in Zou et al. (2020), which does not utilize LASSO and does not include the data screening step, becomes highly inefficient under strong dependence. See Appendix B.2 for an illustration. Second, our new theories for FDR and FDP control under dependence and the robustness of the SDA filter under model misspecification substantially depart from the theory in Zou et al. (2020).

The SDA filter provides a model–free framework that overcomes the limitations of many selective inference procedures, for example, the methods in Lockhart et al. (2014) and Javanmard and Javadi

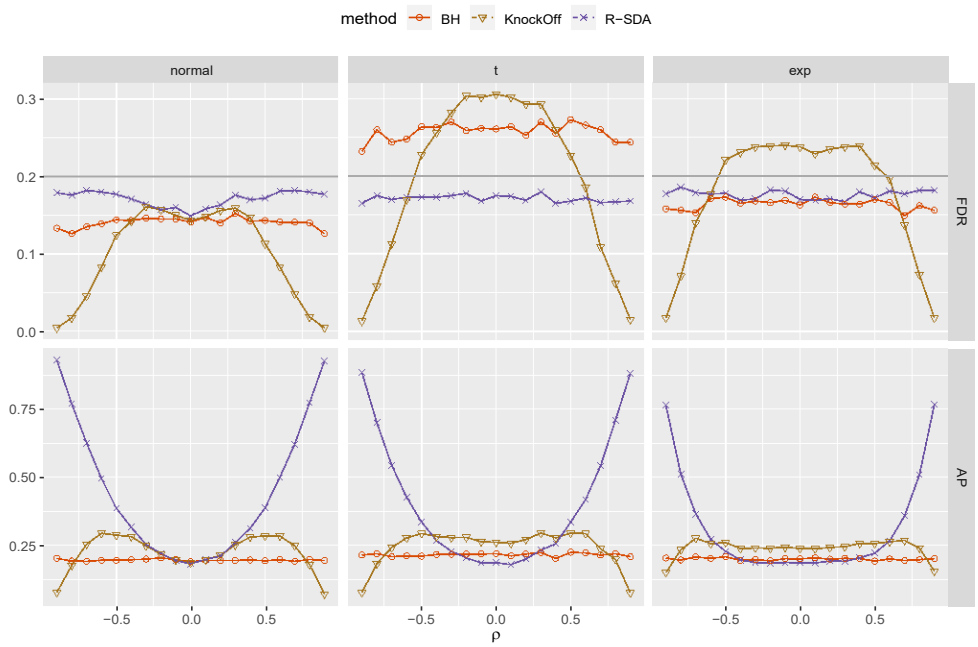


Figure 1: Impacts of correlation on different FDR procedures: "t" denotes the t distribution with 3 df and "exp" denotes the exponential distribution with scale parameter 2. In both cases the models have been mis-specified as normal when computing the p-values.

(2019), which require strong assumptions about the conditional distribution to construct asymptotic p-values. Our numerical results show that the methods in Fan and Han (2017) and Li and Zhong (2017), which require correctly specified models, accurate estimates of parameters and normality assumptions, are in general not robust for FDR control. The SDA filter, which employs empirical distributions instead of asymptotic distributions, only requires the global symmetry of the ranking statistics. It is more robust than its competitors for a wide range of scenarios since the asymptotic symmetry property is much easier to achieve in practice compared to asymptotic normality[1]. As illustrated by the second column (multivariate t data) of Figure 1, BH fails to control the FDR under heavy–tailed models. The failure in accounting for the deviations from normality may result in misleading empirical null and severe bias in FDR analysis (Efron, 2004; Delaigle et al., 2011; Liu and Shao, 2014). Finally, our Theorem 1, which develops a finite–sample upper bound of FDR

[1] For example, the average of several t-variables fulfills the symmetry property perfectly but violates the normality assumption. For asymmetric distributions such as exponential, we usually need a smaller sample size to achieve asymptotic symmetry compared to asymptotic normality – the latter is stronger than the former since it requires an additional accurate approximation in the tail areas.

under dependence, is closely connected to robust knockoffs theory and is established utilizing key arguments from Barber et al. (2020). More specifically, we employ the leave-one-out technique suggested in Barber et al. (2020) to analyze the effect on the SDA filter of possible deviations from normality and the sure screening property, similarly to the analysis of the effect on the Model-X knockoff filter of errors in estimating the true covariance structure. This important connection sheds lights on how the model uncertainty can affect the actual FDR level and how the error bound in FDR can be explicitly quantified using appropriate deviation measures; a detailed discussion is provided in Section B.3 of the Supplementary Material.

1.5 Organization

The remainder of our paper is structured as follows. In Section 2, we introduce the SDA filter for FDR control and discuss the effects of dependence on multiple testing. We develop finite sample and asymptotic theories for FDR control in Section 3. Methodology and theory for the unknown dependence case are discussed in Section 4. Simulation and real data analysis are presented in Sections 5 and 6, respectively. The extensions, proofs of theories and additional comparisons are provided in the Supplementary Material.

Notations. For $M \subseteq \{1, \cdots, p\}$, let X_M be the design matrix with columns $(X_j : j \in M)$ and $X_j = (X_{1j}, \ldots, X_{pj})^\top$ being the jth column. For a matrix or a vector $A = (a_{ij})$, A_M is similarly defined. Let $\|A\|$ be the L_2 norm, $\|A\|_1 = \max_j \sum_i |a_{ij}|$, $\|A\|_{max} = \max_{i,j} |a_{ij}|$ and $\|A\|_\infty = \max_i \sum_j |a_{ij}|$. Let $\lambda_{min}(B)$ and $\lambda_{max}(B)$ denote the smallest and largest eigenvalues of a square matrix B. The notation $A_n \asymp B_n$ means that A_n/B_n and B_n/A_n are both bounded in probability as $n \to \infty$. The "&" and "." are similarly defined. Let $A_n \approx B_n$ denote the two quantities are asymptotically equivalent, in the sense that $A_n/B_n \xrightarrow{p} 1$.

2 The SDA Filter for FDR Control

We start with the assumption that the covariance matrix Σ is known and then move to the case with unknown Σ in Section 4. Our discussion is mainly based on regression model (2); an equivalent

description of the methodology via model (1) follows similarly. We first outline in Section 2.1 the steps for constructing the ranking statistics, then provide intuitive explanations on how the SDA filter works in Sections 2.2 and 2.3. The detailed SDA algorithm is provided in Section A.4.

2.1 Construction of ranking statistics and the symmetry property

SDA first splits the data into two independent parts D_1 and D_2, which are respectively used to construct statistics T_{1j} and T_{2j}. The information in the two parts is then combined to form the ranking statistic $W_j = T_{1j}T_{2j}$. A wide class of pairs may be constructed from the sample. This section presents a specific pair (T_{1j}, T_{2j}), which is used in all numerical studies. Examples of other possible pairs are presented in Section A.2 in the Supplementary Material.

We propose to use LASSO (Tibshirani, 1996) to extract information from D_1 as it simultaneously takes into account the sparsity and dependency structures. Let $\bar{\xi}_1 = n_1^{-1} \sum_{i \in D_1} \xi_i$ and $y_1 = X\bar{\xi}_1$. The LASSO estimator is given by $\hat{\mu}_1 = (\hat{\mu}_{11}, \ldots, \hat{\mu}_{1p})^\top = \arg\min L(\mu)$, where

$$L(\mu) = (y_1 - X\mu)^\top (y_1 - X\mu) + \lambda \|\mu\|_1. \tag{6}$$

Let $S = \{j : \hat{\mu}_{1j} = 0\}$ denote the subset of coordinates selected by LASSO and $S^c = \{1, \cdots, p\} \setminus S$ its complement.

Remark 1 Following Wasserman and Roeder (2009), we suggest using $n_1 = \lfloor 2n/3 \rfloor$, which provides stable performance across a wide range of settings. To obtain asymptotically unbiased estimator in the next step, it is required that S contains all the signals with high probability. In practice, this can be achieved by deliberately choosing an overfitted model that includes most true signals and many false positives; see also Barber and Candès (2019) and Remark 2 in Section 3.2.

Next we use D_2 to obtain the least–squares estimates (LSEs). Let $\bar{\xi}_2 = n_2^{-1} \sum_{i \in D_2} \xi_i$, $y_2 = X\bar{\xi}_2$, $X_S = (X_j : j \in S)$ and $e_j = (0, \cdots, 0, 1, 0, \cdots, 0)^\top$[2]. The LSEs are only calculated for

[2]Specifically, e_j is an $|S|$-vector with 1 in the jth coordinate and 0 elsewhere.

coordinates on the narrowed subset S. Let $\hat{\beta}_2 = (\hat{\beta}_{21}, \ldots, \hat{\beta}_{2p})^\top$, where

$$\hat{\beta}_{2j} = \begin{cases} e_j^\top (X_S^\top X_S)^{-1} X_S^\top y_2, & j \in S; \\ 0, & j \in S^c. \end{cases} \tag{7}$$

Section 2.3 provides insights on why this data screening step can lead to increased SNR.

To aggregate information across both D_1 and D_2, let $W_j = T_{1j} T_{2j}$, where

$$(T_{1j}, T_{2j}) = \left(\frac{\sqrt{n_1}\hat{\beta}_{1j}}{\sigma_{S,j}}, \frac{\sqrt{n_2}\hat{\beta}_{2j}}{\sigma_{S,j}} \right) \tag{8}$$

and $\sigma_{S,j}^2$'s are the diagonal elements of $(X_S^\top X_S)^{-1}$. A multiple testing procedure consists of two steps: ranking and thresholding. Next we show that W_j's play key roles in both steps. Intuitively, the positive W_j's can be used for ranking because a large and positive W_j indicates strong evidence against the null. Meanwhile, the negative W_j's, which usually correspond to null cases, can be used for thresholding. The key idea is to exploit the following asymptotic symmetry property:

$$\sup_{0 \le t \le c \log p} \left| \frac{\sum_{j \in S \cap A^c} I(W_j \ge t)}{\sum_{j \in S \cap A^c} I(W_j \le -t)} - 1 \right| = o_p(1) \qquad \text{for some } c > 0, \tag{9}$$

which holds if $P(A \supseteq S) \to 1$[3]. Next we explain how the SDA filter works.

2.2 FDR thresholding

The asymptotic symmetry property (9) motivates us to choose the following data–driven threshold to control the FDR at level α:

$$L = \inf \left\{ t > 0 : \frac{\#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1} \le \alpha \right\}. \tag{10}$$

Our decision rule is given by $\delta = (\delta_j : 1 \le j \le p)^\top = \{I(W_j \ge L) : 1 \le j \le p\}^\top$. Denote $\hat{A} = \{j : \delta_j = 1\}$ the discovery set. To see why (10) makes sense, note that $\#\{j : W_j \le -t\}$ is an overestimation of $\#\{j : W_j \le -t, j \in A^c\}$, which is asymptotically equal to $\#\{j : W_j \ge t, j \in A^c\}$, the number of false positives, due to the asymptotic symmetry property (9). It follows that the

[3] We shall see that S contains all signals, then the LSEs of the null coordinates are symmetrically distributed around 0. Hence W_j's satisfy (4). It is easy to see that (9) is an asymptotic version of the symmetry property given by (4); see Lemmas S.1–S.2 in Section C of the Supplementary Material for a rigorous discussion.

fraction in (10) provides an overestimate of the FDP, which (desirably) leads to a conservative FDR control. Moreover, the empirical FDR level is typically very close to α because the gap between the fraction in (10) and the actual FDP is usually small in practice, where, for a suitably chosen L, most cases in $\{j : W_j \leq -L\}$ should come from the null.

The operation of the SDA filter can be visualized in Figure 2. We generate $\{\xi_i : i = 1, \ldots, 90\}$ from an MVN distribution with $\mu \in R^{p=1000}$ and $\Sigma = (0.8^{|i-j|})_{1\leq i,j\leq p}$. We randomly set 10% of the coordinates in $\boldsymbol{\mu}$ to be 0.2 and 0 elsewhere. Panel (a) presents the scatter plot of 288 nonzero W_j's with red triangles and black dots respectively denoting true signals and nulls. Panel (d) plots the normalized knockoff statistics that are constructed according to (1.7) in Barber and Candès (2015)[4]. We can see that both SDA and knockoff fulfill the symmetry property approximately for the null W_j's (black dots). However, SDA achieves a more clearcut separation of signals and noise. As explained in Section B.1 of the Supplement, the symmetrized knockoff statistics suffers from high correlations. By contrast, the construction of SDA statistic, which does not depend pairwise contrasts, eliminates the needs for creating fake variables. We can see from Panel (a) that the SDA ranking places most true signals above 0, and many true signals stay well above the majority of the null cases. However, in Panel (d) that illustrates the knockoff ranking, the true signals are not well separated from the nulls, and many true signals even fall below 0. Since the threshold must be positive, signals with negative W_j's will be missed, which leads to substantial power loss.

The impacts on the FDP processes are shown in the second column in Figure 2. We can see that the estimated FDP process [$\widehat{\text{FDP}}(t)$] of SDA approximates the true FDP process [FDP(t)] fairly accurately. However, the knockoff method yields overly conservative estimates of the true FDPs, which leads to overly conservative thresholds (marked by blue vertical lines). The last column in Figure 2 compares the TDP processes of SDA and knockoff. At the FDR level 0.2, the TDP of SDA is 0.87 (threshold $L = 0.62$), which is much higher than that of knockoff (TDP=0.03 with threshold $L = 6.80$). The low TDP of knockoff is due to the decreased power in distinguishing the signal from noise [Panel (d)] and an overly conservative threshold [Panel (e)].

[4] The normalization, which makes the plot easier to read, does not affect the results of the knockoff method. This is because only the relative magnitudes of W_j matter in the thresholding step of the knockoff method.

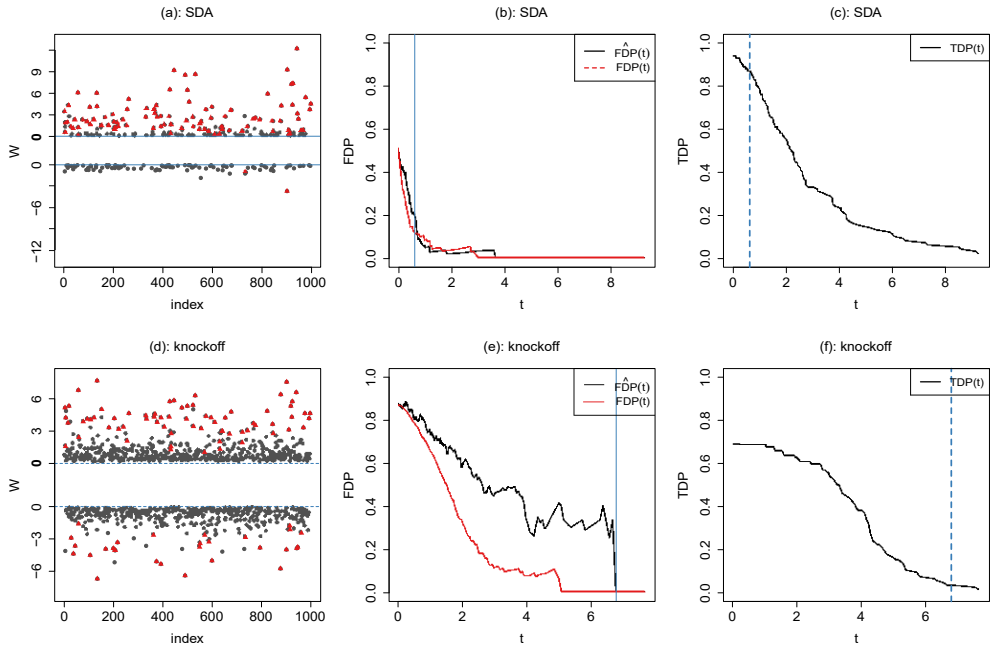


Figure 2: (a): Scatter plot of the 288 nonzero W_js from the SDA filter with red triangles and black dots denoting true signals and nulls respectively. A vertical space is added to the middle of the plot to better contrast positive and negative W_j's. (b): the corresponding estimate of FDP curve (against t) along with the true FDP for the SDA filter; (c): the true power curve (against t) for the SDA filter. (d)-(f): the scatter plot of p = 1000 W_js, the corresponding FDP estimate, and the true power for the knockoff method.

2.3 Power and effects of dependence

The impact of dependence on FDR analysis has been extensively studied but most discussions have focused on the validity issue. This section first discusses the impact of dependence on power, and then provides insights on the information loss of conventional data splitting methods.

Under the SDA framework, many possible pairs of (T_{1j}, T_{2j}) may be constructed. It is easy to show that W_j constructed via the pairs of sample averages

$$(T^0_{1j}, T^0_{2j}) = (\sqrt{n_1}\bar{\xi}_1, \sqrt{n_2}\bar{\xi}_2) \tag{11}$$

also fulfill the asymptotic symmetry property. However, the pair in (11), which falls into the class of marginal testing techniques, can be highly inefficient since it completely ignores the dependence structure. Next we provide intuitions on how the dependence structure is incorporated into the SDA filter to improve the efficiency of existing methods.

First, T_{1j} is superior to T_{1j}^0 by leveraging joint modeling techniques. The merit of joint modeling has been carefully illustrated by Barber and Candès (2015) through extensive simulations. Candès et al. (2018) further argued that the conditional testing techniques are in general more powerful in recovering sparse signals than marginal testing methods. T_{1j} is constructed based on LASSO (a conditional inference technique) and serves as a more suitable building block than T_{1j}^0 for constructing W_{1j}. Second, T_{2j} enjoys a higher SNR than T_{2j}^0 by exploiting the dependence between ξ_S and ξ_{S^c}. Clearly, the expectations of both $\hat{\mu}_{2S}$ and $\bar{\xi}_{2S}$ are μ_{2S}. The covariance of $\hat{\mu}_{2S}$ is $n_2^{-1}Q$, where $Q = (X_S^\top X_S)^{-1}$. By the inversion formula of a block matrix, we have $X_S^\top X_S = \Omega_{S,S} = \left(\Sigma_{S,S} - \Sigma_{S,S^c}\Sigma_{S^c,S^c}^{-1}\Sigma_{S^c,S}\right)^{-1}$. Hence, $Q = \Sigma_{S,S} - \Sigma_{S,S^c}\Sigma_{S^c,S^c}^{-1}\Sigma_{S^c,S}$, which is the conditional covariance of ξ_S given ξ_{S^c}. Let s_{jl} be the (j,l)-th element of Σ. Then $n_2\mathrm{Var}(\bar{\xi}_{2j}) = s_{jj}$. However, $n_2\mathrm{Var}(\hat{\mu}_{2j}) = s_{jj} - e_j^\top \Sigma_{S,S^c}\Sigma_{S^c,S^c}^{-1}\Sigma_{S^c,S}e_j < s_{jj}$. This provides the key insight on the effect of data screening. In regression terms, strong correlations indicate that a large fraction of variability in the variables in S can be explained by the variables in S^c. The higher the correlations, the more reductions in the uncertainties and hence the higher SNRs. This explains why SDA becomes more powerful as correlations increase (Figure 1).

Finally, both knockoff and SDA achieve the symmetry property at the expense of possibly reduced SNR: the former increases the dimension of the design matrix by adding noise variables while the latter involves sample splitting. In contrast with the sample splitting method in Wasserman and Roeder (2009), where D_1 is thrown away after model selection, SDA provides a new aggregation strategy: T_{1j} is kept and combined with T_{2j} to form the ranking statistic W_j. This substantially reduces the information loss in conventional sample splitting methods.

2.4 Effects of data screening

The data screening step is always beneficial as long as the tests are correlated. Intuitively, the smaller the set S, the larger amount of uncertainty can be explained by the variables in S^c. Hence a more effective dimension reduction implies increased SNR and higher power. Meanwhile, our theory on FDR control requires that $P(A \subseteq S)$ holds with high probability, indicating that an overly aggressive data screening step can hurt the FDR procedure. In practice, we recommend

deliberately choosing an overfitted model to ensure the validity in FDR control; this would slightly compromise the power. To illustrate the tradeoff, Figure 3 presents a numerical study to investigate how the size of S may affect both the FDR and power. We can see that the actual FDRs of SDA may deviate from the nominal level when S is too small. By contrast, a large S (overfitted model) has little impact on the FDR levels, but affects the power negatively.

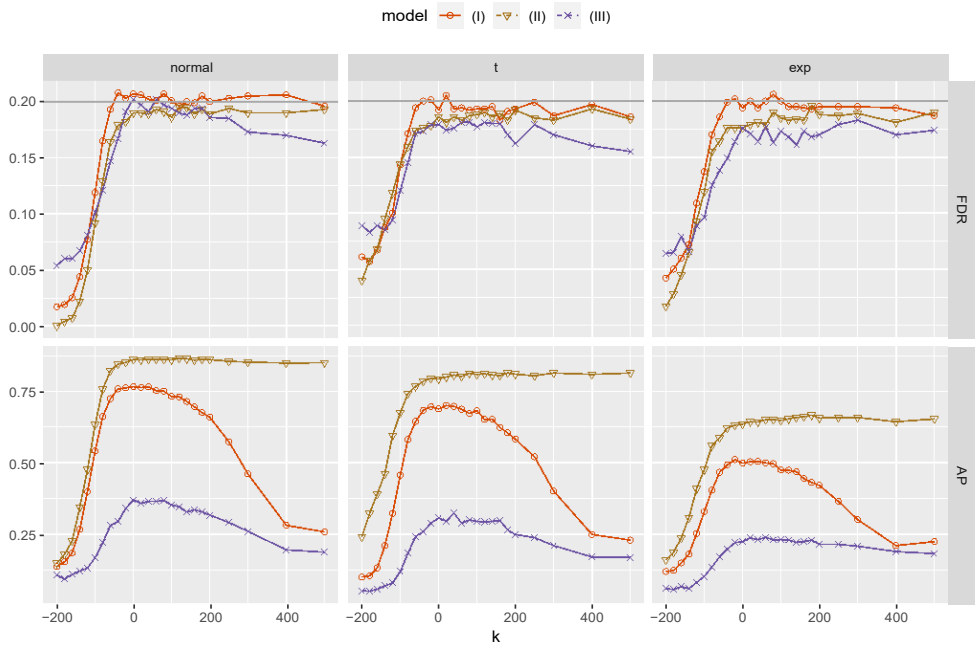


Figure 3: The effects of data screening. We choose n = 90, p = 500, and μ = ±0.2. The proportion of non-nulls is 10% and α = 0.2. We investigate the performance of SDA over 3 distributions and 3 covariance structures described in Section 5. Here k denotes the excess counts of |S| with λ selected by the AIC criterion (k can be negative).

3 Theoretical Properties of the SDA Filter

This section first establishes finite sample theory for FDR bounds (Section 3.1), and then develops asymptotic theories for FDR and FDP control.

3.1 Finite–sample theory on FDR control

Our finite–sample theory, which requires no model assumptions, establishes an upper bound for the FDR under general dependence. We emphasize that the upper bound holds for both known and estimated covariance matrices.

Our theory is developed for a modified SDA filter (SDA+) which chooses the threshold

$$L = \inf \left\{ t > 0 : \frac{1 + \#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1} \le \alpha \right\}.$$

SDA+ is slightly more conservative than SDA but their difference is negligible when the number of rejections is large. Recall $S = \{j : \mu_{1j} = 0\}$. Denote $W_S = (W_j : j \in S)^\top$ and $W_{-j} = W_S \setminus W_j$. The key quantity that controls the upper bound is

$$\Delta_j = |\Pr(W_j > 0 \,||W_j|, W_{-j}) - 1/2|, \tag{12}$$

which can be interpreted as a measure of the extent to which the "flip–sign" property of W_j is violated[5]. Our finite sample theory for FDR control is given by Theorem 1.

Theorem 1 For any $\alpha \in (0, 1)$, the FDR of the SDA+ method satisfies

$$\mathrm{FDR} \le \min_{\tau \ge 0} \left\{ \alpha(1 + \epsilon) + \Pr\Big(\max_{j \in A^c \cap S} \Delta_j > \tau \Big) \right\}. \tag{13}$$

Our theorem is closely connected to Theorem 1 in Barber et al. (2020). Both theorems involve assessing how the deviations from the "idealized situation" would affect the actual FDR level. However, the interpretations are very different. In model-X knockoff the deviation (from the assumption of a known X matrix) comes from the estimation errors of the X matrix whereas in SDA the deviation (from the perfect symmetry property) comes from the possible violations of the normality assumption and sure screening property. Our theorem shows that a tight control of Δ_j's leads to effective FDR control. Next we carefully interpret the bound and present several important settings in which the upper bound in (13) exactly achieves or is very close to the nominal level α.

[5] For a null variable (i.e. $j \in A^c$), the flip–sign property means that W_j is equally likely to be positive or negative conditioning on its magnitude and other W_k's in S.

Consider the ideal case where (a) the error distribution is symmetric, (b) S contains all signals and (c) W_j's are independent of each other for $j \in S$. We can show that $\Delta_j = 0$ for all $j \in A^c \cap S$. The upper bound achieves the nominal level α exactly since $\Pr(W_j > 0 \,||W_j|, W_{-j}) = \Pr(W_j > 0 \,| \,|W_j|) = 1/2$ and hence we can set $= 0$. Even when the error distribution is asymmetric, we expect that Δ_j's would become vanishingly small for moderate sample size n due to the convergence of μ_{2j} to a symmetric distribution (Lemma S.1). Hence the FDR bound would be close to α.

Next we turn to the dependent case. For simplicity, assume that ξ_i's come from a multivariate normal distribution. Let $Q = (X_S^\top X_S)^{-1} := (Q_{jk})_{q_n \times q_n}$ with $q_n = |S|$. The matrix $Q = \Sigma_{S,S} - \Sigma_{S,S^c}\Sigma_{S^c,S^c}^{-1}\Sigma_{S^c,S}$ is the conditional covariance matrix of ξ_S given ξ_{S^c}. The following lemma shows that the magnitude of Δ_j is controlled by the matrix Q.

Lemma 1 (Flip–sign property under Gaussian dependence). Assume that ξ_i's obey a multivariate normal distribution. Denote $Q_{-j,j}$ the jth column of Q excluding Q_{jj}. If $Q_{-j,j} = 0$, then $\Delta_j = 0$.

To provide some intuitions on how close the bound is to α in practice, consider the autoregressive (AR) structure $\Sigma = (\sigma_{j,l}) = (\rho^{|j-l|})$. Since the precision matrix of AR structure is tridiagonal, only consecutive coordinates are correlated with each other conditional on remaining variables. Suppose sparse signals are randomly distributed on the p coordinates and the dimension reduction via S is performed effectively, e.g. $q_n \ll p$. Let E be an event such that for any null variable $j \in S \cap A^c$, remaining variables in S are conditionally uncorrelated with it. We expect E to occur with high probability since for large tridiagonal precision matrices, there is a small chance that two consecutive coordinates are selected into a small set S simultaneously. On event E, we have $Q_{-j,j} = 0$ and it follows from Lemma 1 that $\Delta_j = 0$. Consequently the FDR bound would converge to α when $\Pr(E) \to 1$. In the same vein, we expect that the bound would be close to α for the class of power decay covariance matrices and the class of sparse precision matrices.

3.2 Asymptotic theory on FDP control

Under the asymptotic paradigm we can prove that the FDR can be controlled at $\alpha + o(1)$ under suitable conditions (asymptotic validity). Denote $\varepsilon_i = X(\xi_i - \mu)$. Let $d_n = |A|$, $q_n = |S|$, $q_{0n} =$

$|S \cap A^c|$, and $A(S) := (X_S^\top X_S)^{-1} X_S^\top = (a_{jk})_{q_n \times p}$. Assume that q_n is uniformly bounded above by some non-random sequence $\bar{q}_n$ that will be specified later. We start with some regularity conditions.

Condition 1 (Sure screening property) As $n \to \infty$, $\Pr(A \supseteq S) \to 1$.

Remark 2 Condition 1 ensures that $\hat{\mu}_{2j}$ is unbiased for $j \in S$. This pre–selection property, which has been commonly used (Wasserman and Roeder, 2009; Meinshausen et al., 2009; Barber and Candès, 2019), can be fulfilled with suitably chosen λ under the "zonal" assumption (Bühlmann and Mandozzi, 2014). In practice, we recommend applying AIC to deliberately choose an overfitted model. The sure screening property may not hold exactly but missing small μ_j's is inconsequential. For example, if we ignore "unimportant" signals, then Condition 1 is fulfilled by LASSO for large signals exceeding the rate of $d_n \sqrt{\log p/n}$. Asymptotically unbiased estimators are usually sufficient for effective FDR control. This has been corroborated by our empirical results in Section 5.

Condition 2 (Estimation accuracy) The estimator $\hat{\mu}_1$ fulfills $\|\hat{\mu}_1 - \mu\|_\infty = O_p(c_{np})$, where c_{np} is a sequence satisfying $c_{np} \to 0$ and $1/(\sqrt{n} c_{np}) = O(1)$.

Remark 3 Condition 2 assumes that $\hat{\mu}_1$ is a reasonable estimator of μ; this condition typically holds with $c_{np} = d_n \sqrt{\log p/n}$ for the LASSO solution (Van de Geer and Bühlmann, 2009).

The next two conditions are standard: Condition 3 imposes constraints on the diverging rates of $\bar{q}_n$ and p, both of which depend on the existence of certain moments; Condition 4 requires that the eigenvalues of the design matrix are doubly bounded by two constants.

Condition 3 (Moments) There exist two positive diverging sequences K_{n1} and K_{n2} such that $E(\|\xi_i - \mu\|_\infty^\theta) \le K_{n1}^\theta$ and $E(\|A(S)\varepsilon_i\|_\infty^\theta) \le K_{n2}^\theta$ uniformly in S and $i \in D_2$, where $\theta > 2$. Assume that as $n \to \infty$, $K_{n1} \sqrt{\log p}/n^{1/2-\gamma-\theta^{-1}} \to 0$, $\bar{q}_n^{2/\theta} K_{n2}/n^{1/2-\gamma-\theta^{-1}} \to 0$ for some small $\gamma > 0$.

Condition 4 (Covariance) There exist positive constants $\bar{\kappa}$ and $\underline{\kappa}$ such that with probability one,

$$\underline{\kappa} \le \liminf_{n\to\infty} \lambda_{\min}(X_S^\top X_S) < \limsup_{n\to\infty} \lambda_{\max}(X_S^\top X_S) \le \bar{\kappa}.$$

Condition 5 (Signals) As $n, p \to \infty$, $\eta_n = |C_\mu| \to \infty$, where

$$C_\mu = \{j \in A : \mu_j^2 / \{\max(c_{np}^2, \log \bar{q}_n / n)\} \to \infty\}.$$

Remark 4 Condition 5 implies that the number of identifiable effect sizes should not be too small as $p \to \infty$. This seems to be a necessary condition for FDP control. For example, Liu and Shao (2014) showed that if a multiple testing method controls the FDP with high probability, then its number of true alternatives must diverge when the number of tests goes to infinity.

Condition 6 (Dependence) Let $\rho_{jk} = Q_{jk} / \sqrt{Q_{jj} Q_{kk}}$. Assume that for each j, $\mathrm{Card}\{1 \le k \le q_n : |\rho_{jk}| \ge C(\log n)^{-2-v}\} \le r_p$, where $C > 0$, $v > 0$ is any small constant, and $r_p / \eta_n \to 0$ as $n, p \to \infty$.

Remark 5 Condition 6 allows ξ_j to be correlated with all others but requires that the number of large correlations cannot diverge too fast. The condition appears to be similar to the regularity conditions in Fan et al. (2012) and Xia et al. (2020) but in fact our condition is much weaker. For instance, the correlation between $\boldsymbol{\mu}_{2j_1}$ and $\boldsymbol{\mu}_{2j_2}$ is just the partial correlation of ξ_{j_1} and ξ_{j_2} given the rest variables. In particular, large correlations would be highly unlikely after data screening for a wide range of popular models, such as the class of power decay covariance matrices and the class of moderately sparse precision matrices. This reveals the advantage of SDA, which effectively de–correlates the strong dependence via data screening and conditioning.

Our main theoretical result on the asymptotic validity of the SDA method for both FDP and FDR control is given by the next theorem.

Theorem 2 Suppose Conditions 1–6 hold. For any $\alpha \in (0, 1)$, the FDP of the SDA method satisfies

$$\mathrm{FDP}_W(L) := \frac{\#\{j : W_j \ge L, j \in A^c\}}{\#\{j : W_j \ge L\} \vee 1} \le \alpha + o_p(1). \tag{14}$$

It follows that $\limsup_{(n,p)\to\infty} \mathrm{FDR} \le \alpha$.

4 Unknown dependence

Now we turn to the case where the covariance structure is unknown. When Ω is unknown, the SDA filter operates in the same way except that we substitute the estimate $\hat{\Omega}$ in place of Ω.

We propose to estimate Ω using only the first part of the sample D_1. Denote $\hat{\Omega}$ the corresponding estimator. Then the SDA filter can be readily constructed via the steps in Sections 2.1-2.2 with $X = \hat{\Omega}^{1/2}$. Various high-dimensional precision matrix estimation methods, such as the graphical LASSO (Friedman et al., 2008) and CLIME (Cai et al., 2011), can be used to obtain $\hat{\Omega}$. An attractive feature of the SDA filter under unknown dependence is its robustness for FDR control. We next show that the SDA filter is robust for FDR control if $\hat{\Omega}$ is constructed based only on D_1. We first state a modified version of Condition 6, which uses Q^0 in place of Q.

Condition 6' Let $Q^0 = (X_S^\top X_S)^{-1} X_S^\top X \Omega^{-1} X^\top X_S (X_S^\top X_S)^{-1} := (Q^0_{jk})_{q_n \times q_n}$ and $\rho^0_{jk} = Q^0_{jk} / \sqrt{Q^0_{jj} Q^0_{kk}}$. Assume that for each j, $\mathrm{Card}\{1 \le k \le q_n : |\rho^0_{jk}| \ge C(\log n)^{-2-v}\} \le r_p$, where $C > 0$, $v > 0$ is any small constant, and $r_p/\eta_n \to 0$ as $n, p \to \infty$.

The following theorem, which is in parallel with Theorem 2, establishes the asymptotic validity of the SDA filter for estimated covariance.

Theorem 3 Let $\hat{\Omega}$ denote an estimator based on D_1. Suppose Conditions 1-5 and 6' hold. Then the FDP of the SDA method utilizing $X = \hat{\Omega}^{1/2}$ satisfies $\mathrm{FDP} \le \alpha + o_p(1)$. It follows that $\limsup_{(n,p)\to\infty} \mathrm{FDR} \le \alpha$.

Remark 6 Our FDR theory does not require an accurate estimator for Ω. The accuracy of the estimator only affects the power but not the validity. Consider a working covariance structure that "estimates" Ω as the identity matrix. Then it can be shown that the FDP can still be controlled. This is more attractive than the FDR theories in, for example, Fan and Han (2017) and Li and Zhong (2017) that critically depend on the accuracy of the covariance estimators.

The key step in the proof is to verify the validity of (9). This amounts to addressing two major issues: the asymptotic symmetry of W_j under the null and the uniform convergence of

$q_{0n}^{-1} \sum_{j \in S \cap A^c} I(W_j \geq t)$. Because $\hat{\Omega}$ is obtained from D_1, then $\hat{\mu}_{2j}$ is unbiased conditional on D_1 and thus $\sum_{j \in S \cap A^c} P(W_j > t)$ is approximately equal to $\sum_{j \in S \cap A^c} P(W_j < -t)$, establishing the symmetry property. The dependence assumption on Q^0 ensures the convergence of $q_{0n}^{-1} \sum_{j \in S \cap A^c} I(W_j \geq t)$.

While sample–splitting ensures the independence between $\hat{\mu}_1$ and $\hat{\mu}_2$ and hence the robustness of the SDA filter, as one would expect, a more accurate estimate of Ω yields better power. Previously we have proposed to estimate Ω using D_1 and construct the LSE (7) using D_2. In practice one may consider using D_1 to construct T_{1j}, and then obtaining the LSE via the full sample estimator, denoted $\hat{\Omega}_F$, that is estimated using $\{D_1, D_2\}$. The caveat is that, although $X = \hat{\Omega}_F^{1/2}$ can potentially increase the power, stronger conditions will be needed to guarantee the asymptotic validity of the "full–sample" SDA method. As pointed out by an insightful referee, the asymptotic theory requires that $\hat{\Omega}_F$ must converge to Ω at a very fast rate, which can be impractical in applications. We recommend the robust SDA filter that estimates Ω using only D_1. Next we specify the requirements on the estimation accuracy of $\hat{\Omega}_F$.

Condition 7 The estimated precision matrix $\hat{\Omega}_F$ satisfies $\|\hat{\Omega}_F - \Omega\|_\infty = O_p(a_{np})$ with $a_{np} \to 0$.

The following theorem shows that the FDR and FDP can be controlled asymptotically when $\hat{\Omega}_F$ is sufficiently close to Ω. Let $s_n = \|\Omega\|_\infty$.

Theorem 4 Consider a modified SDA procedure where we use D_1 to construct T_{1j} and the full sample estimator $\hat{\Omega}_F$ to construct the LSE (7). Suppose Conditions 1–6 hold and $\hat{\Omega}_F$ satisfies Condition 7. Then, if

$$c_{np} a_{np} s_n \bar{q}_n \sqrt{n / \log p} (\log \bar{q}_n)^{1+\gamma} \to 0 \tag{15}$$

for a small $\gamma > 0$, the results in Theorem 2 hold for the procedure with $\hat{\Omega}_F$.

This theorem, which is a complementary result to Theorem 3, provides conditions that warrant the implementation of a more efficient version of SDA. It is worth further investigating the condition (15), which seems to be unavoidable because T_{1j} and T_{2j} are no longer independent when the whole sample is used to estimate Ω. To fix ideas, suppose that $\Omega = (\omega_{ij})_{p \times p}$ is k_n-sparse, i.e.

$\max_{1\le i\le p}\sum_{j=i} I(\omega_{ij} = 0) \le k_n$, and that all its elements ω_{ij}s are bounded. First, standard arguments in, for example, Yuan (2010) and Liu et al. (2012) indicate that $a_{np} = O_p(k_n \sqrt{\log p/n})$. Accordingly, with $c_{np} = d_n \sqrt{\log p/n}$, Equation (15) is equivalent to the condition $d_n k_n s_n \bar{q}_n/n^{1/2} \to 0$ if p is of a polynomial rate of n. The condition above imposes restrictions on the diverging rates of d_n, k_n, s_n and $\bar{q}_n$. Assume that d_n, k_n and s_n are all bounded. Then we must require that $\bar{q}_n = o(n^{1/2})$. Alternatively, if we only assume that k_n and s_n are bounded, then a sufficient condi-tion for (15) is $\bar{q}_n = o(n^{1/4})$ (since $d_n \le \bar{q}_n$). These rates are consistent with those in the literature; see, for example, Portnoy et al. (1984) and Fan and Peng (2004).

5 Simulation

This section first introduces the R package `sdafilter` (Section 5.1), followed by simulation designs (Section 5.2) and comparison results (Section 5.3). Additional results for comparisons with unknown covariance matrix and other correlation structures are provided in the Supplementary Material.

5.1 Implementation details

We describe the implementation details of the R package `sdafilter`. For sample–splitting, we follow the strategy in Wasserman and Roeder (2009), which uses $n_1 = [2/3n]$ for selecting variables, and the rest $n_2 = n - n_1$ for obtaining the LSEs. The AIC is used to select the tuning parameter in LASSO. If the number of the variables selected by AIC exceeds [p/3], then only the first [p/3] variables will be retained. For the case with unknown Ω, our default option is to apply the R package `glasso` to D_1, where the tuning parameter is set by the R package huge. If prior knowledge suggests a nonsparse $\mathbf{\Omega}$, the "nonsparse" option in our package can be used. This option first estimates the covariance matrix using the R package POET and then takes its inverse as the input. The `stable` option implements the R-SDA method described in Section A.1 of the Supplementary Material. The kwd option enables the usage of different estimators to summarizes the information in the first part of data, including the de-biased LASSO, innovated transformation of the sample means (Hall and Jin, 2010), and factor-adjusted sample means (Fan and Han, 2017).

5.2 Simulation settings

We consider three types of covariance structures: (I) Autoregressive (AR) structure: $\Sigma = (\rho^{|j-i|})$. (II) Compound symmetry structure: all off-diagonal elements of the Σ are ρ, which can be regarded as a factor model with one principal component. (III) Sparse covariance structure: $\Sigma = \Gamma\Gamma^{\top} + I_p$, where Γ is a $p \times p$ matrix and each row of Γ has only one position with nonzero value sampled from uniform distribution [1, 2].

The diagonal elements are normalized as unity for all three settings. To investigate the robustness of different methods, we consider three error distributions: (i) multivariate normal; (ii) t-distribution with df = 3 and (iii) exponential distribution with scale parameter 2. The observations are then standardized to have mean zero and standard deviation one. The correlation structure remains nearly unchanged after transformation. The following six methods will be compared:

(a) The Benjamini–Hochberg (BH) procedure with the p-values transformed from the t statistics.

(b) The principal factor approximation (PFA) procedure proposed by Fan et al. (2012) for known covariance and Fan and Han (2017) for estimated covariance. Two versions of the PFA procedure using the unadjusted p-values and adjusted p-values are implemented using the R package pfa, denoted as PFA_U and PFA_A respectively. We only report the results for PFA_A as it generally outperforms PFA_U.

(c) The sample-splitting method (SS; Wasserman and Roeder, 2009), which conducts data screening using LASSO and then applies BH to the p-values calculated based on $\hat{\mu}_2$.

(d) The knockoff method (Knockoff; Barber and Candès, 2015), which is implemented using function "create.fixed" in the R package knockoff.

(e) The DATE method (DATE; Li and Zhong, 2017), which we implemented by ourselves.

(f) The stability–refined SDA filter (R-SDA) implemented using our package sdafilter with the "stable" option. We only presented R-SDA, which we recommend to use in practice, to make the plots easier to read. SDA has similar performance to R-SDA.

Let n be the sample size, p the number tests, and π_1 the proportion of signals. For each combination (n, p, π_1), we generate data and apply the six methods at FDR level α. The FDR and AP are calculated by averaging the proportions from 500 replications.

5.3 Comparison results for known covariance structures

We fix $(n, p, \pi_1, \alpha) = (90, 500, 0.1, 0.2)$ and generate μ_j from the following random mixture model:

$$\mu_j \overset{iid}{\sim} (1 - \pi_1)\delta_0 + \pi_1 g(\cdot), \quad j = 1, \cdots, p,$$

where δ_0 is the dirac delta function (denoting a point mass at 0), and $g(\cdot)$ is the density of the non-null distribution, specified as a uniform distribution $[\mu_0 - 0.1, \mu_0 + 0.1]$. The signals μ_j's are then randomly multiplied by a flip-sign. To assess the effect of signal strength, we vary μ_0 from 0.1 to 0.3 and apply the six methods to simulated data. The results for Structures (I) and (III) are summarized in Figure 4, where in the top row we fix $\rho = 0.8$. The results for Structure (II) with $\rho = 0.8$ are shown in Figure S5 of the Supplementary Material. The following observations can be made.

(a) For the Gaussian error case, BH, knockoff, R-SDA and SS control the FDR at the nominal level. The FDR levels of PFA_A and DATE are inflated when signals are weak.

(b) For the non-Gaussian error case, BH, DATE, SS and PFA_A fail to control the FDR under various settings and the FDR levels can be much higher than the nominal level. Knockoff controls the FDR in all settings but can be very conservative. R-SDA has the most accurate and stable FDR levels among all methods.

(c) R-SDA vs SS and BH. As expected, SS and BH control the FDR under the Gaussian case but are not robust for non-Gaussian errors. R-SDA has much higher power than both methods (even when the FDR levels of R-SDA are much lower). It is interesting to note that although SS only uses the second part of the data, its power can be much higher than BH when the correlation structure is highly informative [Normal case under Structure (I) on top left]. This is because the data screening step can significantly increase the SNR (Section 2.3).

(d) R-SDA vs Knockoff. R-SDA and knockoff, both of which are distribution–free, are the only methods that can control the FDR at the nominal level across all scenarios. The knockoff method is overly conservative in Setting (I) due to the high correlation. The conservativeness become less severe under Setting (III). By contrast, R-SDA controls the FDR more accurately near the target level and has significantly higher power than knockoff.

(e) R-SDA vs DATE and PFA_A. In some scenarios, DATE and PFA_A can outperform SDA in power. However, the higher power may be attributed to the severely inflated FDRs. The numerical results reveal the promise of extending the SDA framework by employing other methods, such as factor–adjusted z-scores or innovated transformations, as alternatives to the LASSO estimates, to construct T_{1j}.

Next we turn to investigate how the six methods are affected by the strength of correlation. For covariance structures (I) and (II), we fix $\mu = 0.2$ under alternative and vary the magnitude of correlation ρ from independence ($\rho = 0$) to strong dependence ($\rho = 0.9$). The results are summarized in Figure 5. In addition to the observations that we have made based on the previous graph, the following additional patterns are worthy of mentioning.

(a) The knockoff method becomes more conservative when correlations become higher. Note that the average correlations in Structure (II) is much higher than that in Structure (I), the power of the knockoff method deteriorates faster for Structure (II) as ρ increases. For Structure (II), the FDR of BH also decreases as ρ increases.

(b) In contrast with BH and knockoff, both of which suffer from high correlations, the FDR of R-SDA remains at the nominal level consistently, and the power increases with the correlation. The power grows faster for Structure (II). This corroborates the insights that high correlations can be useful in FDR analysis (Benjamini and Heller, 2007; Sun and Cai, 2009).

(c) In Column 2 of Figure 5, knockoff fails to control the FDR for heavy tailed distributions when correlation is low. By contrast, SDA controls the FDR accurately under non-Gaussian errors.

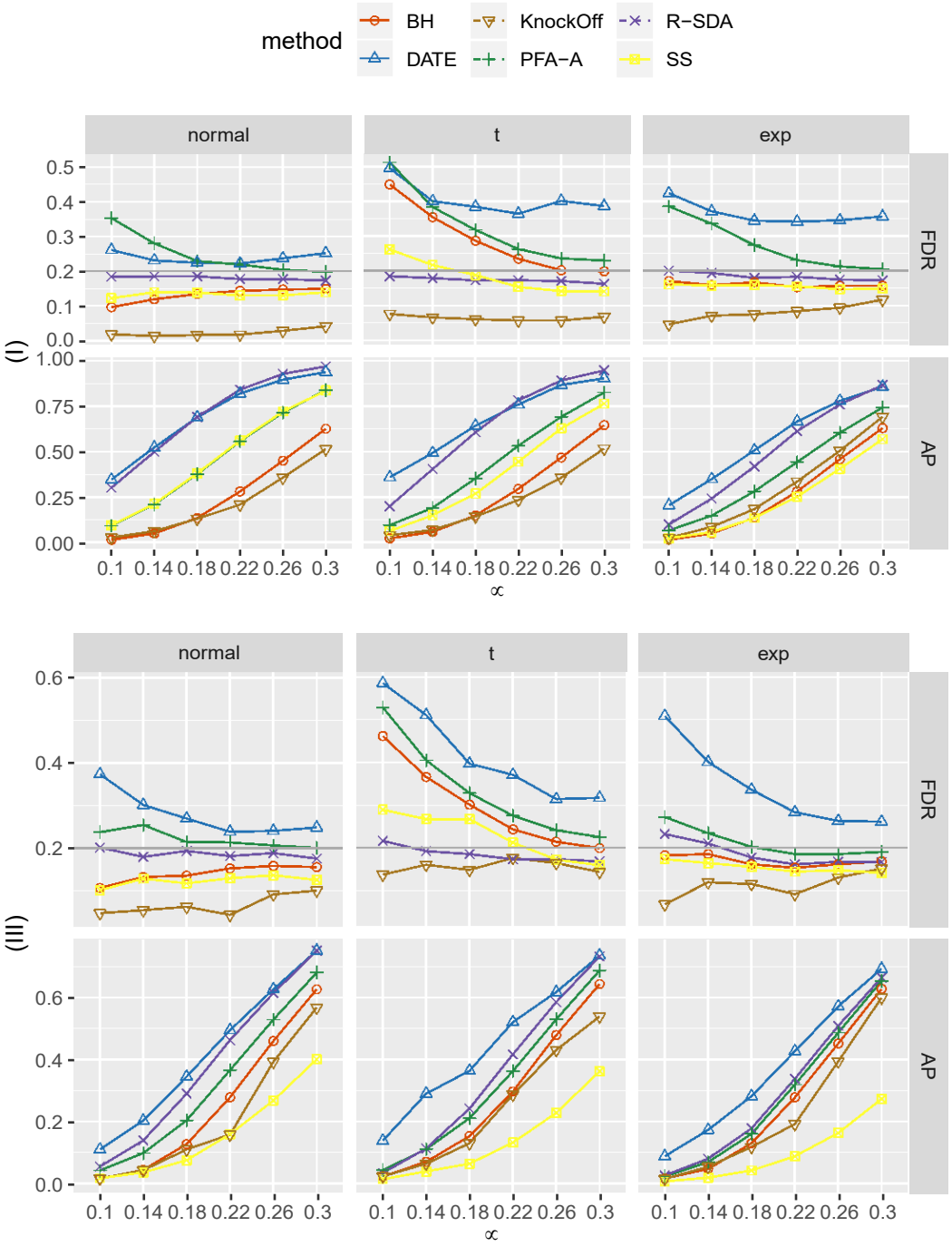


Figure 4: FDR and AP comparison for varying μ in Settings (I) and (III) with known variance.

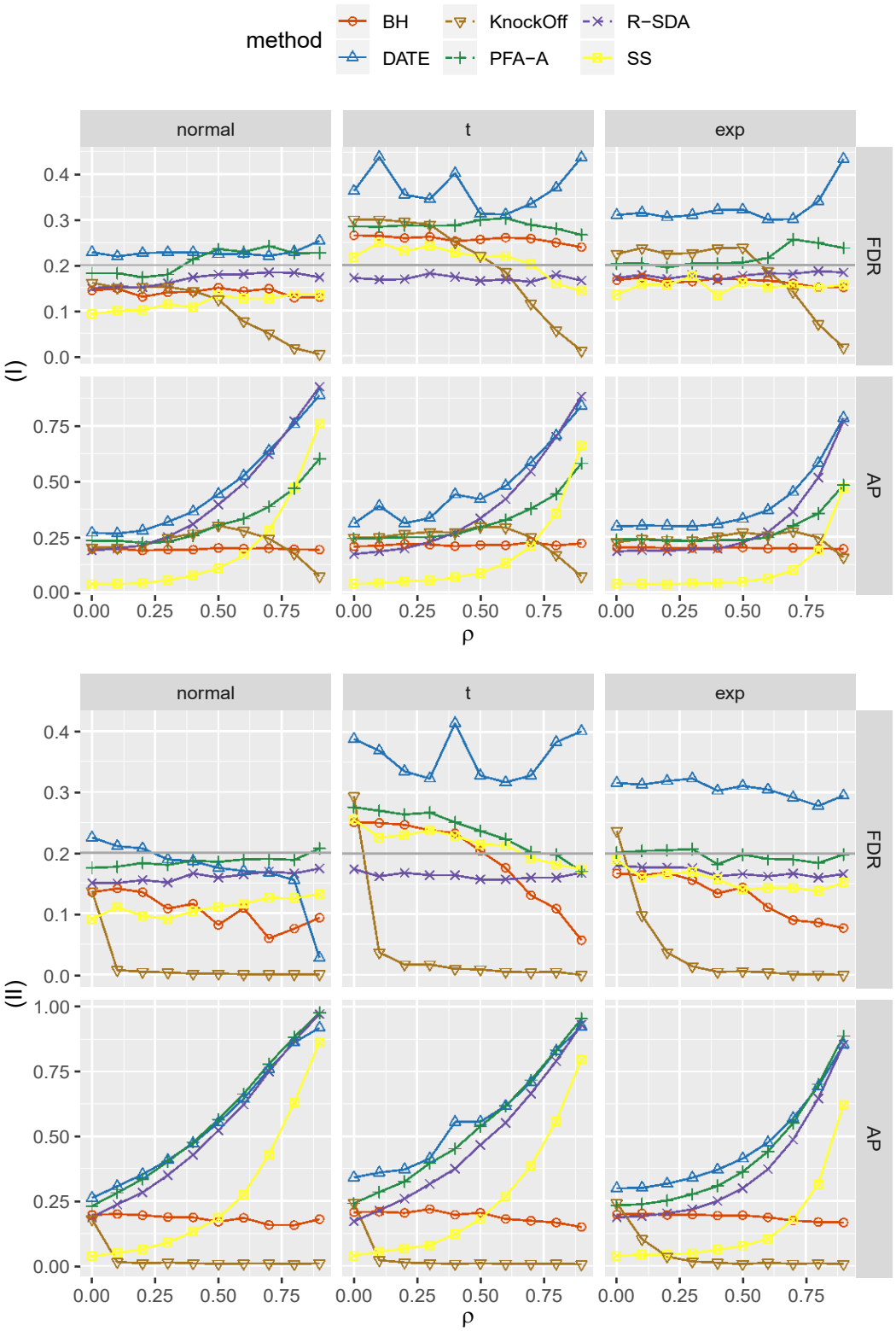


Figure 5: FDR and AP comparison for varying ρ in Settings (I)–(II) with known covariance matrix.

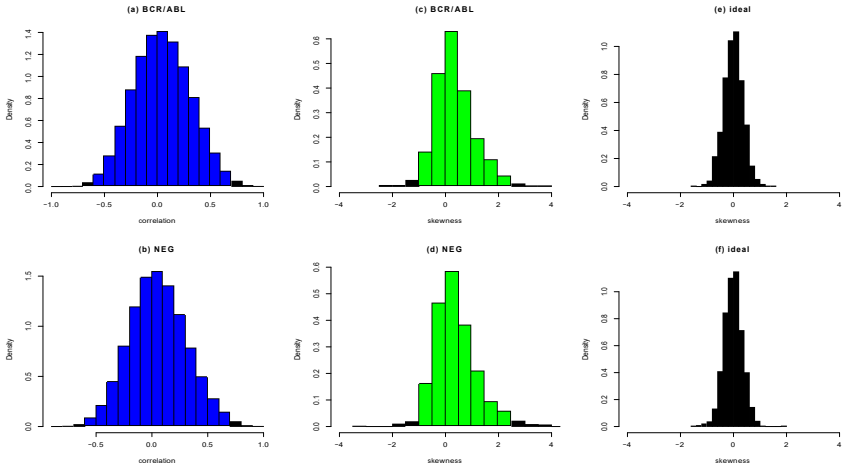


Figure 6: (a)-(b): Histograms of the off-diagonal elements of the sample correlation matrix for BCR/ABL and NEG; (c)-(d): Histogram of the skewness of the $p = 1263$ genes for BCR/ABL and NEG; (e)-(f): the ideal patterns of (c)-(d) when the data are normal.

6 A real-data example

This section illustrates the SDA filter for analysis of high-density oligonucleotide microarrays. The data set, which contains 12,625 probe sets from 128 adult patients enrolled in the Italian GIMEMA multi–center clinical trial, has been used in Chiaretti et al. (2005) and Bourgon et al. (2010) for identifying genetic factors that are associated with acute lymphoblastic leukemia (ALL). The ALL dataset is available at http://www.bioconductor.org.

We focus on a subset of 79 patients with B-cell differentiation because existing research reveals that malignant cells in B-lineage ALL are often associated with genetic abnormalities that have significant impacts on the clinical course of the disease. The patients are divided into two groups based on the molecular heterogeneity of the B-lineage ALL: 37 with the BCR/ABL mutation and 42 with NEG. We further narrow down the focus to 10% of the genes (i.e., $p = 1,263$) before carrying out the FDR analysis. Specifically, the uncorrelated screening method (Bourgon et al., 2010) has been used to remove probe sets with small overall sample variances since they are unlikely to be differentially expressed.

We apply a two–sample version of R-SDA (see Section A.3 for details), BH, SS, PFA_A, Knockoff

and DATE at several significance levels for identifying differentially expressed genes across the two groups. Table 1 summarizes the number of significant probe sets for each method. In Figure 6(a)-(b), we plot the pairwise correlations of the genes. We can see that a significant proportion of the correlations exceed 0.4. These correlations can jointly exhibit non-negligible dependence effect. This explains why the knockoff method is overly conservative. R-SDA is more powerful than SS by exploiting additional information from the second part of data. BH, PFA_{A} and DATE claims more significant genes than R-SDA. However, some caveats need to be given regarding the reliability of BH, PFA_{A} and DATE, which all require normality assumptions (and the latter two require accurate estimates of the unknown covariance matrices).

Next we conduct a preliminary analysis to investigate the normality assumption, which seems to have been severely violated in this data set. From Column 2 of Figure 6 we can see that the skewness scores of many genes exceed the conventional cutoff ± 1. As a comparison, we display in Column 3 of Figure 6 the "ideal" pattern where the normality assumption holds. The histograms in Column 2 are much wider than the histograms in Column 3, indicating a possibly highly skewed error distribution. One possible explanation for the difference in power is that BH, PFA-A and DATE may have inflated FDR levels under violation of normality. This has been observed in our simulation studies (e.g. last column in Figure S3). By contrast, SDA and knockoff are distribution–free methods, which tend to produce more reliable and replicable findings. The lists of 19 highest ranked probe sets by the six methods are presented in Table S1 of Appendix E.

Table 1: The number of rejections for six multiple testing procedures and various significance levels.

	R-SDA	SS	BH	PFA-A	Knockoff	DATE
$\alpha = 0.01$	19	7	29	98	2	364
$\alpha = 0.05$	33	15	146	182	2	452
$\alpha = 0.10$	56	37	229	252	2	501
$\alpha = 0.20$	139	68	350	339	7	546

Acknowledgments

The authors thank the Editor, Associate Editor and two anonymous referees for their many helpful comments that have resulted in significant improvements of the article.

References

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006), "Adapting to unknown sparsity by controlling the false discovery rate," The Annals of Statistics, 34, 584–653.

Barber, R. F. and Candès, E. J. (2015), "Controlling the false discovery rate via knockoffs," The Annals of Statistics, 43, 2055–2085.

— (2019), "A knockoff filter for high-dimensional selective inference," The Annals of Statistics, 47, 2504–2537.

Barber, R. F., Candès, E. J., and Samworth, R. J. (2020), "Robust inference with knockoffs," The Annals of Statistics, 48, 1409–1431.

Barras, L., Scaillet, O., and Wermers, R. (2010), "False discoveries in mutual fund performance: Measuring luck in estimated alphas," The journal of finance, 65, 179–216.

Benjamini, Y. and Heller, R. (2007), "False discovery rates for spatial signals," Journal of the American Statistical Association, 102, 1272–1281.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," Journal of the Royal Statistical Society: Series B (Methodological), 57, 289–300.

— (1997), "Multiple Hypotheses Testing with Weights," Scandinavian Journal of Statistics, 24, 407–418.

Benjamini, Y. and Yekutieli, D. (2001), "The control of the false discovery rate in multiple testing under dependency," The Annals of Statistics, 29, 1165–1188.

Bickel, P. J. and Levina, E. (2008), "Regularized estimation of large covariance matrices," The Annals of Statistics, 36, 199–227.

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015), "SLOPE-daptive variable selection via convex optimization," The Annals of Applied Statistics, 9, 1103.

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

Bühlmann, P. and Mandozzi, J. (2014), "High-dimensional variable screening and bias in subsequent inference, with an empirical comparison," Computational Statistics, 29, 407–430.

Cai, T. and Liu, W. (2016), "Large-scale multiple testing of correlations," Journal of the American Statistical Association, 111, 229–240.

Cai, T., Liu, W., and Luo, X. (2011), "A constrained l_1 minimization approach to sparse precision matrix estimation," Journal of the American Statistical Association, 106, 594–607.

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

Caldas de Castro, M. and Singer, B. H. (2006), "Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association," Geographical Analysis, 38, 180–208.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for gold: model-X knockoffs for high dimensional controlled variable selection," Journal of the Royal Statistical Society: Series B (Methodological), 80, 551–577.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Wang, K. S., Mandelli, F., Foa, R., and Ritz, J. (2005), "Gene expression profiles of B-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation," Clinical cancer research, 11, 7209–7219.

Clarke, S. and Hall, P. (2009), "Robustness of multiple testing procedures against dependence," The Annals of Statistics, 37, 332–358.

Delaigle, A., Hall, P., and Jin, J. (2011), "Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic," Journal of the Royal Statistical Society: Series B (Methodological), 73, 283–301.

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

— (2007), "Correlation and large-scale simultaneous significance testing," Journal of the American Statistical Association, 102, 93–103.

— (2010), "Correlated z-values and the accuracy of large-scale statistical estimates," Journal of the American Statistical Association, 105, 1042–1055.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes analysis of a microarray experiment," Journal of the American Statistical Association, 96, 1151–1160.

Fan, J. and Han, X. (2017), "Estimation of the false discovery proportion with unknown dependence," Journal of the Royal Statistical Society: Series B (Methodological), 79, 1143–1164.

Fan, J., Han, X., and Gu, W. (2012), "Estimating false discovery proportion under arbitrary covariance dependence," Journal of the American Statistical Association, 107, 1019–1035.

Fan, J., Liao, Y., and Mincheva, M. (2013), "Large covariance estimation by thresholding principal orthogonal complements," Journal of the Royal Statistical Society: Series B (Methodological), 75, 603–680.

Fan, J. and Peng, H. (2004), "Nonconcave penalized likelihood with a diverging number of parameters," The Annals of Statistics, 32, 928–961.

Finner, H., Dickhaus, T., and Roters, M. (2007), "Dependency and false discovery rate: asymptotics," The Annals of Statistics, 35, 1432–1455.

Fithian, W., Sun, D., and Taylor, J. (2014), "Optimal inference after model selection," arXiv preprint arXiv:1410.2597.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse inverse covariance estimation with the graphical lasso," Biostatistics, 9, 432–441.

Friguet, C., Kloareg, M., and Causeur, D. (2009), "A factor model approach to multiple testing under dependence," Journal of the American Statistical Association, 104, 1406–1415.

Hall, P. and Jin, J. (2010), "Innovated higher criticism for detecting sparse signals in correlated noise," The Annals of Statistics, 38, 1686–1732.

Javanmard, A. and Javadi, H. (2019), "False discovery rate control via debiased lasso," Electronic Journal of Statistics, 13, 1212–1253.

Jin, J. (2012), "Comment," Journal of the American Statistical Association, 107, 1042–1045.

Leek, J. T. and Storey, J. D. (2008), "A general framework for multiple testing dependence," Proceedings of the National Academy of Sciences, 105, 18718–18723.

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

Lei, L., Ramdas, A., and Fithian, W. (2021), "A general interactive framework for false discovery rate control under structural constraints," Biometrika, 108, 253–267.

Li, A. and Barber, R. F. (2019), "Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm," Journal of the Royal Statistical Society: Series B (Methodological), 81, 45–74.

Li, J. and Zhong, P.-S. (2017), "A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence," The Annals of Statistics, 45, 557–590.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012), "High-dimensional semiparametric Gaussian copula graphical models," The Annals of Statistics, 40, 2293–2326.

Liu, W. and Shao, Q.-M. (2014), "Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control," The Annals of Statistics, 42, 2003–2025.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014), "A significance test for the lasso," The Annals of Statistics, 42, 413–468.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "P-values for high-dimensional regression," Journal of the American Statistical Association, 104, 1671–1681.

Owen, A. B. (2005), "Variance of the number of false discoveries." Journal of the Royal Statistical Society: Series B (Methodological), 67, 411–426.

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

Petrov, V. (2002), "On probabilities of moderate deviations," Journal of Mathematical Sciences, 109, 2189–2191.

Portnoy, S. et al. (1984), "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency," The Annals of Statistics, 12, 1298–1309.

Ramdas, A. (2019), "Discussion of CARS: Covariate assisted ranking and screening for large-scale two-sample inference," Journal of the Royal Statistical Society: Series B (Methodological), 81, 228.

Roeder, K. and Wasserman, L. (2009), "Genome-wide significance levels and weighted hypothesis testing," Statistical science: a review journal of the Institute of Mathematical Statistics, 24, 398–413.

Sarkar, S. K. (2002), "Some results on false discovery rate in stepwise multiple testing procedures," The Annals of Statistics, 30, 239–257.

Schwartzman, A., Dougherty, R. F., and Taylor, J. E. (2008), "False discovery rate analysis of brain diffusion direction maps," The Annals of Applied Statistics, 2, 153–175.

Schwartzman, A. and Lin, X. (2011), "The effect of correlation in false discovery rate estimation," Biometrika, 98, 199–214.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach," Journal of the Royal Statistical Society: Series B (Methodological), 66, 187–205.

Sun, W. and Cai, T. (2009), "Large-scale multiple testing under dependence," Journal of the Royal Statistical Society: Series B (Methodological), 71, 393–424.

Sun, W. and Cai, T. T. (2007), "Oracle and adaptive compound decision rules for false discovery rate control," Journal of the American Statistical Association, 102, 901–912.

Sun, W., Reich, B. J., Cai, T. T., Guindani, M., and Schwartzman, A. (2015), "False discovery control in large-scale spatial multiple testing," Journal of the Royal Statistical Society: Series B (Methodological), 77, 59–83.

Sun, W. and Wei, Z. (2011), "Large-Scale multiple testing for pattern identification, with applications to time-course microarray experiments," Journal of the American Statistical Association, 106, 73–88.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), 58, 267–288.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001), "Significance analysis of microarrays applied to the ionizing radiation response," Proceedings of the National Academy of Sciences of the United States of America, 98, 5116–5121.

Van de Geer, S. A. and Bühlmann, P. (2009), "On the conditions used to prove oracle results for the Lasso," Electronic Journal of Statistics, 3, 1360–1392.

Wasserman, L. and Roeder, K. (2009), "High dimensional variable selection," The Annals of Statistics, 37, 2178–2201.

Wu, W. B. (2008), "On false discovery control under dependence," The Annals of Statistics, 36, 364–380.

Xia, Y., Cai, T. T., and Sun, W. (2020), "Gap: A general framework for information pooling in two-sample sparse inference," Journal of the American Statistical Association, 115, 1236–1250.

Yuan, M. (2010), "High dimensional inverse covariance matrix estimation via linear programming," The Journal of Machine Learning Research, 11, 2261–2286.

Zou, C., Ren, H., Guo, X., and Li, R. (2020), "A New Procedure for Controlling False Discovery Rate in Large-Scale t-tests," arXiv preprint arXiv:2002.12548.

Supplementary Material for "False Discovery Rate Control Under General Dependence By Symmetrized Data Aggregation"

This supplement contains some refinements and extensions of the SDA filter (Appendix A), comparisons of the SDA filter with related ideas in the literature (Appendix B), the proofs of main theorems (Appendix C), other theoretical results (Appendix D), and additional numerical results (Appendix E).

A Refinements and Extensions

SDA provides a general framework for constructing symmetrized statistics to aggregate structural information from dependent data. In this section, we discuss some extensions to illustrate how this framework can be implemented in different scenarios.

A.1 A stability refinement

To improve the stability in selection and avoid "p-value lottery" occurred in a single sample splitting (Meinshausen et al., 2009), we propose a modified SDA algorithm that employs the "bagging" technique to aggregate results from multiple sample–splitting procedures.

Denote $\hat{A}_k$, $k = 1, \ldots, B$, the discovery sets from repeatedly applying B times the SDA filter at level α via random sample splittings. The decisions are aggregated by $\hat{A}_v = \#\{j : \sum_{k=1}^{B} I(j \in \hat{A}_k) > \lfloor B/2 \rfloor\}$, the set of variables that are consistently selected in at least 50% of the replications. The stability refinement picks $\hat{A}_{\hat{k}}$ having the biggest overlap with $\hat{A}_v$:

$$\hat{k} = \arg\max_{1 \le k \le B} \sum_{j=1}^{p} \left\{ I(j \in \hat{A}_k \cap \hat{A}_v) + I(j \in \hat{A}_k \cap \hat{A}_v) \right\}. \tag{S.1}$$

The new method with stability refinement is denoted R-SDA. The asymptotic theory for the R-SDA filter is presented and proven in Section D. Our theory implies that the FDPs of $\hat{A}_k$ can be controlled uniformly for all k. Hence the discovery set $\hat{A}_{\hat{k}}$ produces more stable results with guaranteed FDR

control. Our numerical studies show that compared to SDA, R-SDA generally yields similar FDR and power but smaller variations in the FDP.

A.2 Other types of ranking statistics

The SDA filter utilizes $W_j = T_{1j}T_{2j}$ to rank the hypotheses. The asymptotic symmetry property (9) is fulfilled as long as T_{2j} are constructed as the LSEs on a subset S that includes all signals with high probability. This leaves much flexibility for constructing T_{1j}. We provide a few examples.

1) $T_{1j} = |\hat{\mu}_{1j}|$, where $\hat{\mu}_{1j}$ is the LASSO estimate. In contrast with the scaled version $|\hat{\mu}_{1j}|/\sigma_{S,j}$, using $|\hat{\mu}_{1j}|$ directly reflects the preference of selecting large effect sizes over significant ones. In our numerical studies the two methods seem to perform similarly.

2) If there is prior knowledge that the covariance structure can be well described by a factor model, then we can substitute the factor-adjusted statistics (Fan and Han, 2017) in place of T_{1j}.

3) T_{1j} is the de-biased estimate of μ_j (or its scaled version) based on inverse regression method (Xia et al., 2020).

4) T_{1j} is the innovated transformation of the sample means (Hall and Jin, 2010; Jin, 2012).

In our simulation studies, we found LASSO works well and stably in a wide range of settings but can be outperformed by other choices of T_{1j} in special situations. How to develop more powerful ranking statistics is an interesting and challenging problem that requires further research. The main message of this section is that in applications practitioners may develop new types of ranking statistics tailored to problem contexts and prior knowledge about the data structure.

Finally we stress that our theory requires that T_{2j} must be chosen so that the asymptotic symmetry property is fulfilled. For example, it is not allowed to use the LASSO estimate again to construct T_{2j} because this improper choice would lead to a violation of the symmetry property, which no longer guarantees that the FDR can be controlled at the nominal level.

A.3 Two–sample inference

Suppose we are interested in identifying features that exhibit differential levels across two conditions. Let $\xi^{(k)} = (\xi_1^{(k)}, \ldots, \xi_p^{(k)})^>$, $k = 1, 2$, be two p-dimensional random vectors. The population mean vectors and covariance matrices are $\mu^{(k)}$ and $\Sigma^{(k)}$, $k = 1, 2$, respectively. Consider the following two-sample multiple testing problem:

$$H_j^0 : \mu_j^{(1)} = \mu_j^{(2)} \text{ versus } H_j^1 : \mu_j^{(1)} = \mu_j^{(2)}, \text{ for } j = 1, \ldots, p.$$

The SDA filter can be easily generalized to handle the two-sample situation. Denote $D^{(k)} = \{\xi_i^{(k)} = (\xi_{i1}^{(k)}, \ldots, \xi_{ip}^{(k)})^>, i = 1, \cdots, n^{(k)}\}$. First, we split $D^{(k)}$ into two disjoint groups $D_1^{(k)} = (\xi_1^{(k)})$ and $D_2^{(k)} = (\xi_2^{(k)})$, with sizes $n_1^{(k)}$ and $n_2^{(k)}$, respectively. Denote $n_l = n_l^{(1)} + n_l^{(2)}$, $D_l = D_l^{(1)} \cup D_l^{(2)}$, $l = 1, 2$. Based on D_1, the LASSO estimator can be obtained via minimizing $(y_1 - X\omega)^>(y_1 - X\omega) + \lambda\|\omega\|_1$, where $y_1 = X(\bar{\xi}_1^{(1)} - \bar{\xi}_1^{(2)})$, $X = \Omega^{1/2}$, and $\Omega = (n_1/n_1^{(1)}\Sigma^{(1)} + n_1/n_1^{(2)}\Sigma^{(2)})^{-1}$. Denote S the selected subset by LASSO. Next we calculate the LSEs, using data D_2, for coordinates in S. The formula is identical to (7) except that now we take $y_2 = X(\bar{\xi}_2^{(1)} - \bar{\xi}_2^{(2)})$ and $X = \Omega^{1/2}$. Finally, we can calculate W_j and determine the threshold L using (10). This procedure is implemented in Section 6 in the main text to identify differentially expressed genes in microarray studies. Asymptotic theories for the two–sample SDA method, which are presented in Appendix D, can be established similarly as done for the standard SDA method.

A.4 The SDA algorithm: detailed steps

We summarize the operation of the SDA algorithm in this subsection.

- Step 1: Split the data set into two parts D_1 and D_2. If the precision matrix Ω is unknown, use D_1 to obtain its estimate $\hat{\Omega}$.
- Step 2: Let $X = \hat{\Omega}^{1/2}$. Compute $\hat{\omega}_1$ by (6) and find the narrowed subset S. Record the estimated coefficients $\hat{\omega}_{1j}$.
- Step 3: Compute $\hat{\omega}_2$ by (7) by restricting on the coordinates in the subset S.

- Step 4: Compute the ranking statistic W_j by (8).
- Step 5: Find the threshold L using (10) and output $\hat{A} = \{j : W_j \geq L\}$ as the selected features.

B Comparisons with Existing Literature

This section presents comparisons of SDA with existing literature. The goal is to provide insights on the limitations of existing works and highlight some key features of SDA.

B.1 SDA vs. Knockoff

We present some theoretical insights on why the knockoff method suffers from power loss under dependence. The whitening transformation from Model (1) to Model (2) implies that the fixed-design knockoff filter in Barber and Candès (2015) is directly applicable to our problem with the Gram matrix $X^\top X = \Omega$, where Ω is the precision matrix. The augmented design matrix can accordingly be constructed as $(\Omega^{1/2}, 0)^\top$ (c.f. Section 2.1.2 of Barber and Candès, 2015). The knockoffs $\tilde{X}$ must fulfill $\tilde{X}^\top \tilde{X} = \Omega$ and $X^\top \tilde{X} = \Omega - \mathrm{diag}\{s\}$, where $s = (s_1, \ldots, s_p)^\top$ is a p-dimensional nonnegative vector. Denote X_j the jth column of the design matrix and $\tilde{X}_j$ its knockoff copy. In a setting where the features are normalized, i.e. $\Omega_{jj} = 1$ for all j, the correlation between X_j and $\tilde{X}_j$ is $1 - s_j$, where $0 \leq s_j \leq 1$. Intuitively, it is desirable to make the entries of s as large as possible; this ensures that X_j would deviate from its knockoff copy as much as possible (hence we will hopefully have sufficient power to distinguish the true signals from faked ones).

Consider two settings where the correlation structures are respectively AR(1) $[\mathrm{Corr}(X_j, X_k) = \rho^{|j-k|}, j = k]$ and compound symmetric $[\mathrm{Corr}(X_j, X_k) = \rho, j = k]$. We consider two approaches, namely equi-correlated and SDP knockoffs, both of which were considered in Barber and Candès (2015) for optimizing s_j's. Figure S1 depicts the "average similarity score" $1 - s$ as a function of different correlation levels ρ, where $s = p^{-1} \sum_{j=1}^{p} s_j$ is calculated using both the equi-correlated (left column) and SDP (right column) optimizers. The plots for AR(1) and compound symmetric structures are shown in the top and bottom rows, respectively. We can see that the similarity score $1 - s$ increases rapidly in ρ. For example, $1 - s$ has already exceeded 95% when ρ is only 0.25 under

the compound symmetric structure. Consequently, it becomes extremely difficult to distinguish the original variables and their faked copies. This leads to substantial power loss of the knockoff filter. The relationship between the similarity scores and the correlation levels are consistent with the patterns in the power loss of the knockoff method as noted in Fig.5 of Barber and Candès (2015) and Figure 1 in the main text of this article.

In contrast with the knockoff filter, the operation of SDA does not rely on pairwise contrasts. It only utilizes the global symmetry property among all W_j's. The sample-splitting approach eliminates the needs for constructing fake variables under a possibly highly restricted geometric space. This explains why the SDA does not suffer from high correlations.

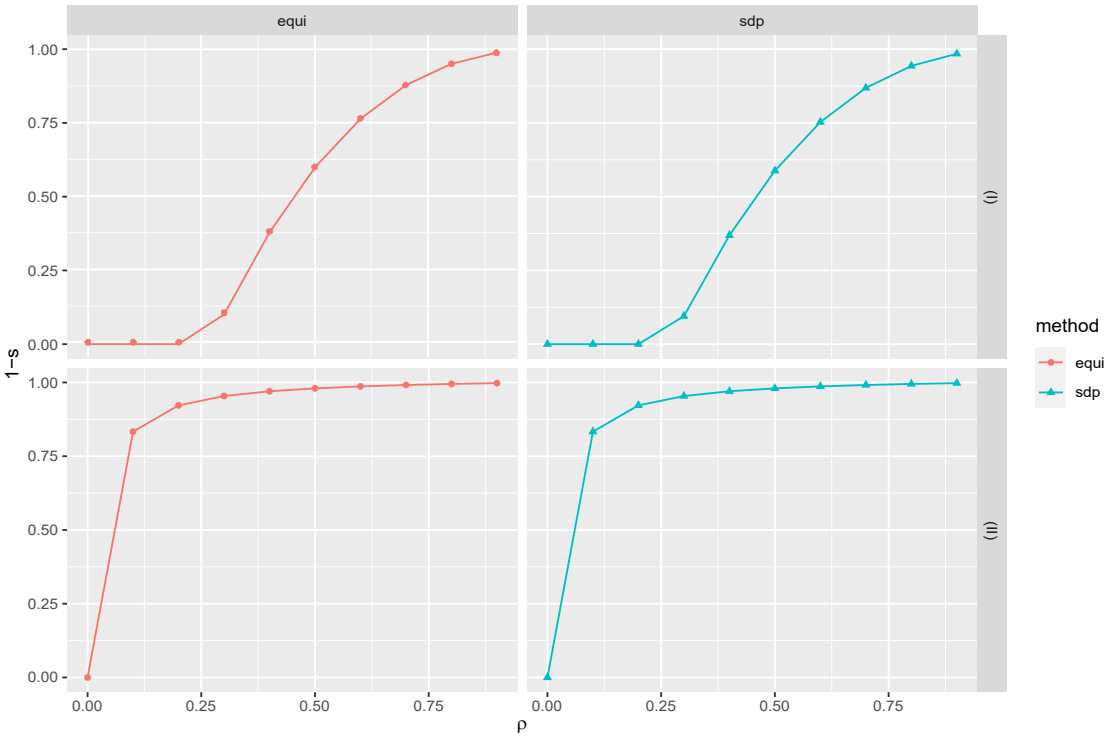


Figure S1: The knockoff filter suffers from power loss under moderate to strong dependence. The average similarity score (i.e., $1-s$) between the original variable and its knockoff as a function of ρ. Top row: AR(1) structure; bottom row: compound symmetric structure. Both equi-correlated knockoff (left) and SDP knockoff (right) have been considered. The number of tests is $p = 100$.

B.2 SDA vs. RESS

The reflection via sample-splitting (RESS) method in Zou et al. (2020) was developed for independent two-sample t-tests. It can be substantially improved by SDA that effectively exploits the informative dependence structure. For illustration, Figure S2 compares the FDR levels and average powers (AP) for SDA vs. BH and RESS in Zou et al. (2020) at different correlation levels. The simulation settings are the same as those in Figure 1 in the main text. We can see that the average powers of RESS and BH remain roughly the same across all correlation levels since the dependence structure has been ignored. In contrast, the power of SDA increases sharply with growing correlation levels. Section 2.3 in the main text provides high-level ideas on how the dependence is incorporated into the SDA filter to improve the power.

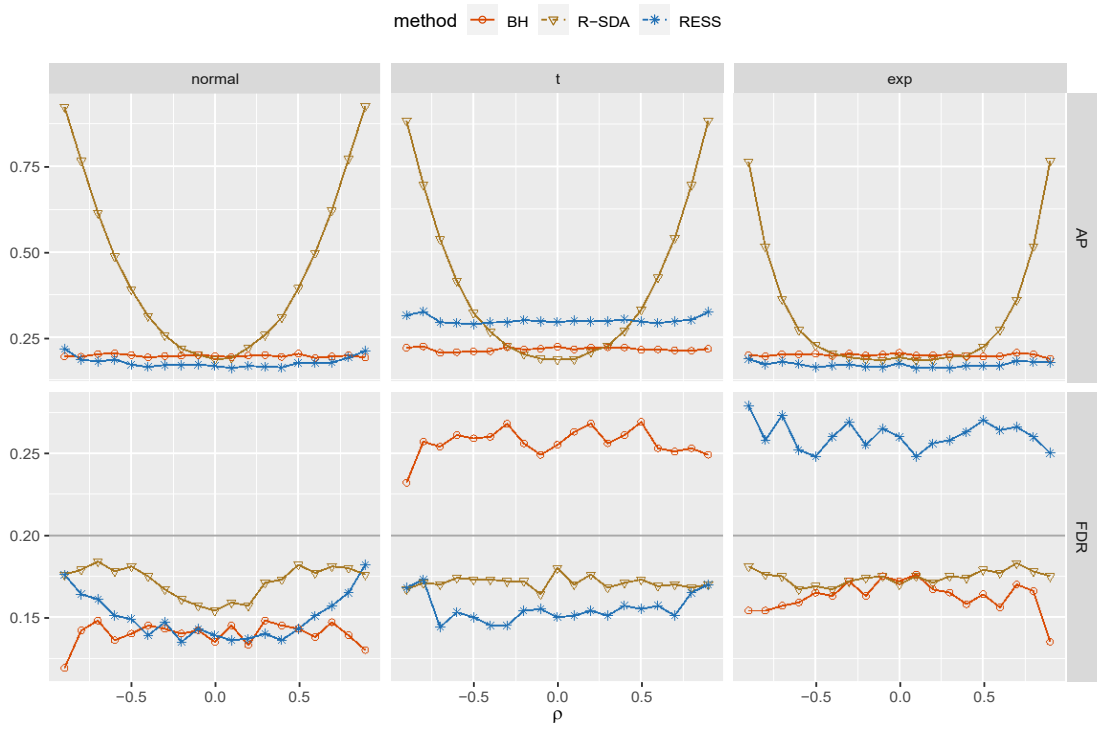


Figure S2: Impacts of correlation on different FDR procedures. Here RESS refers to the Refection via Sample Splitting procedure in Zou et al. (2020).

B.3 Model uncertainty and error bound for FDR analysis

This section highlights the important connection of our theory to the robust knockoff theory in Barber et al. (2020), as pointed out by an insightful referee.

The model-X knockoff assumes that the distribution of the feature vector X is known exactly. However, in practical situations the X distribution must be estimated. In Theorem 1 of Barber et al. (2020), the KL divergence between the true distribution and its estimate is employed to quantify the effect of estimation errors on FDR control. The KL divergence can be interpreted as a measure of the extent to which the pairwise exchangeability property of the model-X knockoff is violated.

Under the SDA inferential framework, the idealized setting corresponds to the case where the error distribution is perfectly symmetric about 0 and W_j's are independent of each other for $j \in S$. This idealized situation implies that $\Pr(W_j > 0 \,||W_j|, W_{-j}) = 1/2$. We call this, borrowing the term from Barber et al. (2020), the flip-sign property, which indicates that W_j is equally likely to be positive or negative conditional on its magnitude and other W_k's in S. However, in practical situations the flip-sign property only holds asymptotically. Therefore the actual FDR would unfortunately deviate from the nominal level. The amount of deviation is characterized by

$$\Delta_j = |\Pr(W_j > 0 \,||W_j|, W_{-j}) - 1/2|,$$

which can be interpreted as a measure of the extent to which the flip-sign property is violated. We subsequently use Δ_j's to quantify the effect of asymmetry (i.e. deviation from the perfect symmetry assumption) on FDR control.

Barber et al. (2020) introduced an elegant leave-one-out argument to establish the upper bound for the actual FDR level of the model-X knockoff where the X matrix must be estimated from data. The analysis of SDA in Section 3.1 reveals that the technique can be readily extended to other important settings where the issue on model uncertainty must be addressed[6]. In summary, the work of Barber et al. (2020) provides a set of useful technical tools for developing finite sample theory on (a) how the FDR control can be affected by the model uncertainty and (b) how the

[6] In model-X knockoff the model uncertainty comes from the estimation errors whereas in SDA the model uncertainty corresponds to the possible deviation from normality and sure screening property.

error bound can be explicitly quantified using appropriate deviation measures. The connection of our theory to the robust knockoff theory also provides insights on the impact of deviation from symmetry on the performance of the SDA filter.

C Proofs of Main Theorems

C.1 Finite Sample Theory

This section proves Theorem 1. The proof of this theorem has extensively used the techniques developed by Barber et al. (2020), which shows that the Model-X knockoff (Candès et al., 2018) incurs an inflation of the FDR that is proportional to the errors in estimating the distribution of each feature conditional on the remaining features.

Fix > 0 and for any $t > 0$, define

$$R(t) = \frac{\sum_{j \in A^c} I(W_j \geq t, \Delta_j \leq) 1}{+ \sum_{j \in A^c} I(W_j \leq -t)}.$$

Consider the event that $A = \{\Delta := \max_{j \in A^c} \Delta_j \leq \}$. Furthermore, consider a thresholding rule $L = T(W)$ that maps statistics W to a threshold $L \geq 0$. For each index $j = 1, \ldots, p$, by adopting the leave-one-out argument in Barber et al. (2020), define

$$L_j = T(W_1, \ldots, W_{j-1}, |W_j|, W_{j+1}, \ldots, W_p) \geq 0.$$

For the SDA filter with threshold L, we can write

$$\frac{\sum_{j \in A^c} I(W_j \geq L, \Delta_j \leq)}{\sum_j I(W_j \geq L)} = \frac{1 + \sum_j I(W_j \leq -L)}{1 \sum_j I(W_j \geq L)} \cdot \frac{\sum_{j \in A^c} I(W_j \geq L, \Delta_j \leq) 1}{1 + \sum_j I(W_j \leq -L)}$$

$$\leq \alpha R(L).$$

Next we derive an upper bound for $E\{R(L)\}$. Note that

$$
\begin{aligned}
E\{R(L)\} &= \sum_{j \in A^c} E\left(\frac{I(W_j \geq L, \Delta_j \leq)}{1 + \sum_{j} I(W_j \leq -L)} \right) \\
&= \sum_{j \in A^c} E\left(\frac{I(W_j \geq L_j, \Delta_j \leq)}{1 + \sum_{k \in A^c, k=j} I(W_k \leq -L_j)} \right) \\
&= \sum_{j \in A^c} E\left[E\left(\frac{I(W_j \geq L_j, \Delta_j \leq)}{1 + \sum_{k \in A^c, k=j} I(W_k \leq -L_j)} \,||W_j|, W_{-j} \right) \right] \\
&= \sum_{j \in A^c} E\left(\frac{\Pr(W_j > 0 \,||W_j|, W_{-j}) I(|W_j| \geq L_j, \Delta_j \leq)}{1 + \sum_{k \in A^c, k=j} I(W_k \leq -L_j)} \right). \qquad \text{(S.2)}
\end{aligned}
$$

The last step (S.2) holds since, after conditioning on $(|W_j|, W_{-j})$, the only unknown quantity is the sign of W_j. By the definition of Δ_j, we have $\Pr(W_j > 0 \,||W_j|, W_{-j}) \leq 1/2 + \Delta_j$. Hence,

$$
\begin{aligned}
& E\{R(L)\} \\
& \leq \sum_{j \in A} E\left(\frac{(\tfrac{1}{2} + \Delta_j) I(|W_j| \geq L_j, \Delta_j \leq)_c}{1 + \sum_{k \in A^c, k=j} I(W_k \leq -L_j)} \right) \\
& \leq (\tfrac{1}{2} +) \left[\sum_{j \in A^c} E\left(\frac{I(W_j \geq L_j, \Delta_j \leq)}{1 + \sum_{k \in A^c, k=j} I(W_k \leq -L_j)} \right) + \sum_{j \in A^c} E\left(\frac{I(W_j \leq -L_j)}{1 + \sum_{k \in A^c, k=j} I(W_k \leq -L_j)} \right) \right] \\
& = (\tfrac{1}{2} +) E\{R(L)\} + \sum_{j \in A^c} E\left(\frac{I(W_j \leq -L_j)}{1 + \sum_{k \in A^c, k=j} I(W_k \leq -L_j)} \right).
\end{aligned}
$$

The sum in the last expression can be simplified. If for all null j, $W_j > -L_j$, then the sum is equal to zero. Otherwise

$$
\sum_{j \in A^c} E\left(\frac{I(W_j \leq -L_j)}{1 + \sum_{k \in A^c, k=j} I(W_k \leq -L_j)} \right) = \sum_{j \in A^c} E\left(\frac{I(W_j \leq -L_j)}{1 + \sum_{k \in A^c, k=j} I(W_k \leq -L_k)} \right) = 1,
$$

where the first equality holds because for any j, k, if $W_j \leq -\min(L_j, L_k)$ and $W_k \leq -\min(L_j, L_k)$, then $L_j = L_k$. Accordingly, we have

$$
E\{R(L)\} \leq \frac{1/2 + }{1/2 -} \leq 1 + 5,
$$

which proves the theorem.

C.2 Asymptotic Theory with Known Ω

We present the proofs of Theorem 2 here along with two key lemmas. The lemmas play key roles in our technical arguments and may be of independent interest in their own rights. Other technical lemmas and proofs are provided in Appendix D.

For notational convenience, throughout this section, we consider variables that are included in the set S, and suppress "$j \in S$" in all the summations with respect to j. Let $\Phi(x) = 1 - \Phi(x)$, $G(t) = \bar{q}_{0n}^{-1} \sum_{j \in A^c} \Pr(W_j \geq t \mid D_1)$, $G_-(t) = \bar{q}_{0n}^{-1} \sum_{j \in A^c} \Pr(W_j \leq -t \mid D_1)$ and $G^{-1}(y) = \inf\{t \geq 0 : G(t) \leq y\}$ for $0 \leq y \leq 1$.

The first lemma characterizes the closeness between $G(t)$ and $G_-(t)$.

Lemma S.1 Suppose Conditions 1, 3, and 4 hold. We have

$$\frac{G(t)}{G_-(t)} - 1 \to 0.$$

uniformly for all $0 \leq t \leq G_-^{-1}(\alpha\eta_n/q_{0n})$.

Proof. Define $b_n = \sigma\sqrt{C \log \bar{q}_n}$ where $C > 4$. Denote $T_{kj} = \sqrt{n_k}\hat{\phi}_{kj}/\sigma_j$ for $j = 1, \ldots, q_n$ and $\sigma^2 = Q_{jj}/\sigma_j^2$. Observe that

$$\frac{G(t)}{G_-(t)} - 1 = \frac{\sum_{j \in A^c} \{\Pr(T_{1j}T_{2j} \geq t, |T_{2j}| \leq b_n \mid D_1) - \Pr(T_{1j}T_{2j} \leq -t, |T_{2j}| \leq b_n \mid D_1)\}}{q_{0n}G_-(t)}$$
$$+ \frac{\sum_{j \in A^c} \{\Pr(T_{1j}T_{2j} \geq t, |T_{2j}| > b_n \mid D_1) - \Pr(T_{1j}T_{2j} \leq -t, |T_{2j}| > b_n \mid D_1)\}}{q_{0n}G_-(t)}$$
$$:= \Delta_1 + \Delta_2.$$

Firstly, for the term Δ_2, by Lemma S.8 we obtain that

$$\frac{\sum_{j \in A^c} \Pr(T_{1j}T_{2j} \geq t, |T_{2j}| > b_n \mid D_1)}{q_{0n}G_-(t)} \leq \frac{\sum_{j \in A^c} \Pr(|T_{2j}| > b_n \mid D_1)}{\alpha\eta_n} \cdot \frac{\bar{q}_n \times o(1/\bar{q}_n)}{\eta_n}.$$

It follows that $\Delta_2 = o(1)$.

By Lemma S.7, it can be verified that

$$\frac{\Pr(T_{1j}T_{2j} \geq t, |T_{2j}| \leq b_n \mid D_1)}{\Pr(T_{1j}Z \geq t, |Z| \leq b_n \mid D_1)} \to 1,$$

where $Z \sim N(0, \sigma^2)$ which is independent of T_{1j}. Recall that

$$\mu_{2j}/\sigma_j = n_2^{-1} \sum_{i=1}^{n_2} e_j' X_2 S_{X_2}^{-1} X_2' \varepsilon_i/\sigma_j := n_2^{-1} \sum_{i=1}^{n_2} \eta_{ij}/\sigma_j.$$

Note that $B_n = n_2\sigma^2$ and $L_n = B_n^{-\theta/2} \sum_{i=1}^{n_2} E(|\eta_{ij}|^\theta) \le C n_2^{1-\theta/2} K_{n2}^\theta$. We have

$$\{2\log(1/L_n)\}^{1/2} \ge [2\log\{n_2^{\theta/2-1}/(K_{n2}^\theta)\}]^{1/2} \ge \sqrt{4\log q_n},$$

according to Condition 4. The result follows by applying Lemma S.7.

Similarly we get

$$\frac{\Pr(T_{1j}T_{2j} \le -t, |T_{2j}| \le b_n \mid D_1)}{\Pr(T_{1j}Z \le -t, |Z| \le b_n \mid D_1)} \to 1.$$

Note that

$$\Pr(T_{1j}Z \le -t, |Z| \le b_n \mid D_1) = \Pr(T_{1j}Z \ge t, |Z| \le b_n \mid D_1).$$

This implies that $\Delta_1 = o(1)$, which completes the proof.

The next lemma establishes the uniform convergence of $\sum_{j \in A^c} I(W_j \ge t)/(q_{0n}G(t))$.

Lemma S.2 Suppose Conditions 3, 4, and 6 hold. Then, conditional on D_1, we have

$$\sup_{0 \le t \le G^{-1}(\alpha\eta_n/q_{0n})} \left| \frac{\sum_{j \in A^c} I(W_j \ge t)}{q_{0n}G(t)} - 1 \right| = o_p(1), \tag{S.3}$$

$$\sup_{0 \le t \le G^{-1}(\alpha\eta_n/q_{0n})} \left| \frac{\sum_{j \in A^c} I(W_j \le -t)}{q_{0n}G(t)} - 1 \right| = o_p(1). \tag{S.4}$$

Proof. We only prove the first formula; the second can be proven similarly. In the proof of Lemma S.1, we show that

$$G(t) = q_{0n}^{-1} \sum_{j \in A^c} \Pr(T_{1j}T_{2j} \ge t, |T_{2j}| \le b_n \mid D_1)\{1 + o(1)\} := G^*(t)\{1 + o(1)\}.$$

Similarly we can show that

$$q_{0n}^{-1} \sum_{j \in A^c} I(W_j \ge t) = q_{0n}^{-1} \sum_{j \in A^c} I(W_j \ge t, |T_{2j}| \le b_n)\{1 + o_p(1)\}.$$

Hence, it suffices to show that

$$\sup_{0 \le t \le G^{-1}(\alpha\eta_n/q_{0n})} \left| \frac{\sum_{j \in A^c} I(W_j \ge t, |T_{2j}| \le b_n)}{q_{0n}G^*(t)} - 1 \right| = o_p(1).$$

Note that the $G(t)$ is a decreasing and continuous function. Let $a_p = \alpha\eta_n$, $z_0 < z_1 < \cdots < z_{h_n} \le 1$ and $t_i = G^{-1}(z_i)$, where $z_0 = a_p/q_{0n}$, $z_i = a_p/q_{0n} + b_p \exp(i\zeta)/q_{0n}$, $h_n = \{\log((q_{0n} - a_p)/b_p)\}^{1/\zeta}$ with $b_p/a_p \to 0$ and $0 < \zeta < 1$. Note that $G(t_i)/G(t_{i+1}) = 1 + o(1)$ uniformly in i. It is therefore enough to derive the convergence rate of

$$D_n = \sup_{j \in A^c} \frac{\{I(W_j > t_i, |T_{2j}| \le b_n) - \Pr(W_j > t_i, |T_{2j}| \le b_n \mid D_1)\}}{q_{0n} G(t_i)} \quad 0 \le i \le h_n .$$

Define $M_j = \{k \in A^c : |\rho_{jk}| \ge C(\log n)^{-2-\nu}\}$, $B = \{|T_{2j}| \le b_n, j \in A^c\}$ and

$$\begin{aligned} D(t) &= E\Big[\sum_{j \in A^c} \{I(W_j > t, |T_{2j}| \le b_n) - \Pr(W_j > t, |T_{2j}| \le b_n \mid D_1)\}^2 \mid D_1\Big] \\ &= \sum_{j \in A^c} \sum_{k \in A^c} \{\Pr(W_j > t, W_k > t \mid D_1, B) - \Pr(W_k > t \mid D_1, B) \Pr(W_j > t \mid D_1, B)\}\{1 + o(1)\}. \end{aligned}$$

Note that

$$D(t) \le r_p q_{0n} G(t) + \sum_{j \in A^c} \sum_{k \in M_j^c} \{\Pr(W_k > t, W_j > t \mid D_1, B) - \Pr(W_k > t \mid D_1, B) \Pr(W_j > t \mid D_1, B)\}.$$

However, for each $j \in A^c$ and $k \in M_j^c$, conditional on D_1, the Pearson correlation coefficient between W_j and W_k is ρ_{jk}. By Lemma 1 in Cai and Liu (2016),

$$\frac{\Pr(W_k > t, W_j > t \mid D_1, B) - \Pr(W_k > t \mid D_1, B) \Pr(W_j > t \mid D_1, B)}{\Pr(W_k > t \mid D_1, B) \Pr(W_j > t \mid D_1, B)} \le A_n,$$

uniformly holds, where $A_n = (\log n)^{-1-\nu_1}$ for $\nu_1 = \min(\nu, 1/2)$.

From the above results, we can get

$$\begin{aligned} \Pr(D_n \ge \eta \mid D_1) &\le \sum_{i=0}^{h_n} \Pr\Big(\frac{\sum_{j \in A^c} [I(W_j > t_i, |T_{2j}| \le b_n) - \Pr(W_j > t_i, |T_{2j}| \le b_n \mid D_1)]}{q_{0n} G(t_i)} \ge \eta \mid D_1\Big) \\ &\le \frac{1}{\eta^2} \sum_{i=0}^{h_n} \frac{1}{q_{0n} G^2(t_i)} D(t_i) \\ &\le \frac{1}{\eta^2} \Big(r_p \sum_{i=0}^{h_n} \frac{1}{q_{0n} G(t_i)} + h_n A_n \Big). \end{aligned}$$

Moreover, observe that

$$\sum_{i=0}^{h_n} \frac{1}{q_{0n} G(t_i)} = \frac{1}{a_p} + \sum_{i=1}^{h_n} \frac{1}{a_p + b_p e^{i\zeta}} \cdot b_p^{-1}.$$

Finally, note that (a) ζ can be arbitrarily close to 1 such that $h_n A_n \to 0$, and (b) b_p can be made arbitrarily large as long as $b_p/a_p \to 0$, we conclude that $D_n = o_p(1)$ when $r_p/\eta_n \to 0$. This completes the proof.

In Lemma S.1 and Lemma S.2, we have established the symmetry property and uniform consistency for W_j's. Now we are ready to present the proof of Theorem 2.

Proof of Theorem 2 By definition, SDA selects the jth variable if $W_j \geq L$, where

$$L = \inf\Big\{t \geq 0 : \sum_{j} I(W_j \leq -t) \leq \alpha \max\Big\{\sum_{j} I(W_j \geq t), 1\Big\}\Big\}.$$

We need to establish an asymptotic bound for L so that Lemmas S.1-S.2 can be applied.

Let $t^* = G_-^{-1}(\alpha\eta_n/q_{0n})$. It follows from Lemma S.2 that

$$\alpha\eta_n/q_{0n} = G_-(t^*) = \frac{1}{q_{0n}} \sum_{j \in A^c} I(W_j < -t^*)\{1 + o(1)\}.$$

On the other hand, for any $j \in C_\mu$, we can show that $\Pr(W_j < t^*, j \in C_\mu) \to 0$. In fact, it is straightforward to see that

$$\begin{aligned}
&\Pr(W_j < t^*, \text{ for some } j \in C_\mu) \\
&\leq \eta_n \Pr\Big(T_{1j}T_{2j} - \sqrt{n_1 n_2}\mu_j^2/\sigma_j^2 < t^* - \sqrt{n_1 n_2}\mu_j^2/\sigma_j^2\Big) \\
&\leq \eta_n \Pr\Big(|\mu_j|(|\hat{\mu}_{1j} - \mu_j| + |\hat{\mu}_{2j} - \mu_j|) + |\hat{\mu}_{1j} - \mu_j||\hat{\mu}_{2j} - \mu_j| > \mu_j^2 - t^*\sigma_j^2/\sqrt{n_1 n_2}\Big) \to 0.
\end{aligned}$$

To see the last equation, denote $d_j = \mu_j^2 - t^*\sigma_j^2/\sqrt{n_1 n_2}$. Under Condition 5, it follows that $d_j = \mu_j^2\{1 + o(1)\}$. We then get

$$\begin{aligned}
&\Pr(|\mu_j|(|\hat{\mu}_{1j} - \mu_j| + |\hat{\mu}_{2j} - \mu_j|) + |\hat{\mu}_{1j} - \mu_j||\hat{\mu}_{2j} - \mu_j| > d_j) \\
&\leq \Pr(|\mu_j|(|\hat{\mu}_{1j} - \mu_j| + |\hat{\mu}_{2j} - \mu_j|) > d_j/2) + \Pr(|\hat{\mu}_{1j} - \mu_j||\hat{\mu}_{2j} - \mu_j| > d_j/2) =: H_1 + H_2.
\end{aligned}$$

Note that $d_j/|\mu_j| = |\mu_j|\{1 + o(1)\}$. We observe that

$$\begin{aligned}
H_1 &\leq \Pr(|\hat{\mu}_{1j} - \mu_j| > d_j/(4|\mu_j|)) + \Pr(|\hat{\mu}_{2j} - \mu_j| > d_j/(4|\mu_j|)), \\
H_2 &\leq \Pr(|\hat{\mu}_{1j} - \mu_j| > c_{np}) + \Pr\Big(|\hat{\mu}_{2j} - \mu_j| > C\sqrt{\log \tilde{q}_n/n}\Big).
\end{aligned}$$

Then the result follows from Lemmas S.8 and Condition 2.

Consequently, we have $Pr(\sum_j I(W_j > t^*) \geq \eta_n) \to 1$. We conclude that $\sum_j I(W_j < -t^*) \cdot \alpha\eta_n \leq \alpha \sum_j I(W_j > t^*)$, and hence $L \cdot t^*$. By Lemmas S.1-S.2, we get

$$\frac{\sum_{j \in A^c} I(W_j \geq L)}{\sum_{j \in A^c} I(W_j \leq -L)} - 1 \to 0. \tag{S.5}$$

Next write

$$FDP = \frac{\sum_{j \in A^c} I(W_j \geq L)}{1 \vee \sum_j I(W_j \geq L)} = \frac{\sum_j I(W_j \leq -L)}{1 \vee \sum_j I(W_j \geq L)} \times \frac{\sum_{j \in A^c} I(W_j \geq L)}{\sum_j I(W_j \leq -L)}$$

$$\leq \alpha \times R(L).$$

Note that $R(L) \leq \sum_{j \in A^c} I(W_j \geq L) / \sum_{j \in A^c} I(W_j \leq -L)$, and thus $\limsup_{n \to \infty} FDP \leq \alpha$ by (S.5). Then, for any > 0,

$$FDR \leq (1 +)\alpha R(L) + Pr(FDP \geq (1 +)\alpha R(L)),$$

which proves the second part of this theorem.

C.3 Asymptotic Theory with unknown Ω: Proof of Theorems 3 and 4

Proof of Theorem 3 The proof follows similar lines as those of Theorem 2, except that we now establish Lemmas S.1 and S.2 under Conditions 1-5 and 6'. Note that Lemma S.8 still holds under Conditions 1, 3, and 4. With unknown Ω, conditional on D_1, the Pearson correlation coefficient between W_j and W_k is changed to ρ^c_{jk}. The rest of the proof is essentially the same as that of Theorem 2 and thus omitted.

Proof of Theorem 4 To establish this theorem, we consider another SDA procedure with the statistics $\hat{W}_j = \sqrt{n_1 n_2}(\hat{\mu}_{1j} - \hat{\mu}_{2j})/\sigma_j^2$, where $\hat{\mu}_2$ is the least–squares estimate that uses $\hat{X} = \Omega^{1/2}$ and $\hat{\mu}_2 = \hat{X}\hat{\xi}_2$. We choose a threshold $\hat{L} > 0$ by setting

$$\hat{L} = \inf \left\{ t > 0 : \frac{\#\{j : \hat{W}_j \leq -t\}}{\#\{j : \hat{W}_j \geq t\} \vee 1} \leq \alpha \right\}.$$

The proof of this theorem involves a careful investigation of the difference between W_j and $\hat{W}_j$. The main results are summarized by Lemmas S.3-S.5. Define $G = \{j : \mu_j = o(c_{np})\}$.

From Lemma S.3, we have, for any j,

$$W_j - \hat{W}_j = \sqrt{n_1 n_2}\hat{\beta}_{1j}(\hat{\beta}_{2j} - \beta_{2j})/\sigma_j^2 = O_p(n \times s_n\tilde{q}_n a_{np}\sqrt{\log p/n}) \times \{\mu_j + O_p(c_{np})\}.$$

Thus for any $j \in G$, under condition that $c_{np}a_{np}s_n\tilde{q}_n\sqrt{n \log p}(\log \tilde{q}_n)^{1+\gamma} \to 0$ for a small $\gamma > 0$, the absolute difference between W_j and $\hat{W}_j$ is negligible. While for $j \in G^c$, we need to consider the relative difference. That is,

$$\hat{W}_j = W_j\left(1 + \frac{\beta_{2j} - \hat{\beta}_{2j}}{\hat{\beta}_{2j}}\right) = W_j\left(1 + \frac{O_p(s_n\tilde{q}_n a_{np}\sqrt{\log p/n})}{\mu_j + O_p(\sqrt{\log \tilde{q}_n/n})}\right) = W_j\{1 + o_p(1)\}.$$

In fact, under conditions $c_{np}a_{np}s_n\tilde{q}_n\sqrt{n \log p}(\log \tilde{q}_n)^{1+\gamma} \to 0$ and $1/(\sqrt{n}c_{np}) = O(1)$, we have:

$$\frac{s_n\tilde{q}_n a_{np}\sqrt{\log p/n}}{c_{np}} = o(1), \quad \frac{s_n\tilde{q}_n a_{np}\sqrt{\log p/n}}{\sqrt{\log \tilde{q}_n/n}} = o(1).$$

From Lemma S.4 and Lemma S.5 given below, we conclude that

$$FDP_{\hat{W}}(t) := \frac{\#\{j : \hat{W}_j \geq t, j \in A^c\}}{\#\{j : \hat{W}_j \geq t\} \vee 1} = FDP_W(t)\{1 + o_p(1)\}.$$

Under Conditions 1-6, similar to the proof of Theorem 2, we can show that $FDP_{\hat{W}}(\hat{t})$ is controlled at the nominal level asymptotically. Thus the claimed result follows.

Lemma S.3 If Conditions 1, 3, 4 and 7 hold, then we have $\hat{\beta}_j = \beta_j + O_p(a_{np}s_n\tilde{q}_n\sqrt{\log p/n})$ uniformly in $j \in S$.

Proof. Note that

$$\begin{aligned}|\hat{\beta}_j - \beta_j| &= e_j^\top\left\{(X_S^\top X_S)^{-1}X_S^\top e - (X_S^\top X_S)^{-1}X_S^\top X_{\cdot}\right\}(\xi - \bar{\mu}) \\ &\leq \|(X_S^\top X_S)^{-1}X_S^\top e - (X_S^\top X_S)^{-1}X_S^\top X_{\cdot}\|_\infty \|\xi - \bar{\mu}\|_\infty =: \\ &\|\Delta\|_\infty \|\xi - \bar{\mu}\|_\infty.\end{aligned}$$

Similar to Lemma S.8, we get $\|\xi - \mu\|_\infty = O_p(\sqrt{\log p/n})$. For the analysis of Δ, we note the following fact

$$\|\hat{X}_S^\top \hat{X}_S - X_S^\top X_S\|_\infty \le \|\hat{\Omega}_F - \Omega\|_\infty = O_p(a_{np}),$$

$$\|\hat{X}_S^\top \hat{X}_{S^c} - X_S^\top X_{S^c}\|_\infty \le \|\hat{\Omega}_F - \Omega\|_\infty = O_p(a_{np}),$$

$$\|(\hat{X}_S^\top \hat{X}_S)^{-1} - (X_S^\top X_S)^{-1}\|_\infty \le \|(\hat{X}_S^\top \hat{X}_S)^{-1}\|_\infty \|(X_S^\top X_S)^{-1}\|_\infty \|\hat{X}_S^\top \hat{X}_S - X_S^\top X_S\|_\infty = O_p(\bar{q}_n a_{np}).$$

Thus, by triangle inequality, we can conclude that

$$\begin{aligned}
\|\Delta\|_\infty &= \|(\hat{X}_S^\top \hat{X}_S)^{-1}\hat{X}_S^\top \hat{X}_{S^c} - (X_S^\top X_S)^{-1} X_S^\top X_{S^c}\|_\infty \\
&\le \|(\hat{X}_S^\top \hat{X}_S)^{-1} - (X_S^\top X_S)^{-1}\|_\infty \|X_S^\top X_{S^c}\|_\infty + \|\hat{X}_S^\top \hat{X}_{S^c} - X_S^\top X_{S^c}\|_\infty \|(\hat{X}_S^\top \hat{X}_S)^{-1}\|_\infty \\
&= O_p(\bar{q}_n s_n a_{np})
\end{aligned}$$

and accordingly $\max_j |\hat{\mu}_j - \mu_j| = O_p(\bar{q}_n s_n a_{np}\sqrt{\log p/n})$.

The next lemma establishes the approximation result of W_j to $\hat{W}_j$ for those $j \in G$.

Lemma S.4 Suppose Conditions 1, 2, 3, 4 and 7 hold and $c_{np} a_{np} s_n \bar{q}_n \sqrt{n \log p} (\log \bar{q}_n)^{1+\gamma} \to 0$ for a small $\gamma > 0$. Then, for any $M > 0$,

$$\sup_{M \le t \le G^{-1}(\alpha\eta_n/q_{0n})} \frac{\sum_{j \in G} I(\hat{W}_j \ge t)}{\sum_{j \in G} I(W_j \ge t)} - 1 = o_p(1),$$

$$\sup_{M \le t \le G^{-1}(\alpha\eta_n/q_{0n})} \frac{\sum_{j \in G} I(-\hat{W}_j \le -t)}{\sum_{j \in G} I(-W_j \le -t)} - 1 = o_p(1).$$

Proof. By Lemma S.3, with probability tending to one,

$$\begin{aligned}
&\sum_{j \in G} I(W_j \ge t) - \sum_{j \in G} I(\hat{W}_j \ge t) \\
&\le \sum_{j \in G} \{I(W_j \ge t + l_n) - I(W_j \ge t)\} + \sum_{j \in G} \{I(W_j \ge t - l_n) - I(W_j \ge t)\} \\
&:= \Delta_1 + \Delta_2,
\end{aligned}$$

where $l_n/(c_{np} a_{np} s_n \bar{q}_n \sqrt{n \log p}) \to \infty$ as $n, p \to \infty$. We will deal with Δ_1 only and the part of Δ_2 is

similar. Define the events $C_t = \{|T_{1j}| > t/(C\sqrt{\log q_n}), |T_{2j}| > t/(\sqrt{n c_{np}}), j \in G\}$.

$$
\begin{aligned}
E(\Delta_1) &= E\left[\sum_{j\in G} I(t \le W_j \le t + l_n)\right] \\
&\le \sum_{j\in G} Pr(t \le W_j \le t + l_n \,|\, C_t) + \sum_{j\in G} Pr(t \le W_j \le t + l_n, C_t^c) \\
&\le \sum_{j\in G} Pr(t \le W_j \le t + l_n \,|\, C_t) + o(1),
\end{aligned}
$$

where we use Lemmas S.8 and Condition 2 to get $\sum_{j\in G} Pr(t \le W_j \le t + l_n, C_t) = o(1)$. Further note that under the event, $\{t \le W_j \le t + l_n, C_t\}$, we have

$$
|T_{2j}| \le \frac{t + l_n}{|T_{1j}|} \le \frac{C(t + l_n)\sqrt{\log q_n}}{t} = C\sqrt{\log q_n} + \frac{l_n\sqrt{\log q_n}}{M} \le C\sqrt{\log q_n} = b_n,
$$

under condition that $l_n \to 0$. Let $T_{2j}^* = \sqrt{n_2}(\hat{\mu}_{2j} - \mu_j)/\sigma_j$ and $U_j = \sqrt{n_2}\hat{\mu}_j/\sigma_j$. Thus from Lemma S.1, we conclude that

$$
\begin{aligned}
&\sum_{j\in G} Pr(t - T_{1j}U_j \le T_{1j}Z \le t + l_n - T_{1j}U_j \,|\, C_t) \\
&= \sum_{j\in G} E\{\Phi((t + l_n)/|T_{1j}| - U_j) - \Phi(t/|T_{1j}| - U_j) \,|\, C_t\} \\
&\le \sum_{j\in G} l_n E|T_{1j}|^{-1}\phi(t/|T_{1j}| - U_j) \,|\, C_t \\
&\le l_n \sum_{j\in G} E\left[(t/T_{1j}^2 - U_j/|T_{1j}|) + \frac{1}{t - U_j|T_{1j}|}\right]\bar{\Phi}(t/|T_{1j}| - U_j) \,|\, C_t \\
&\doteq l_n M^{-1}\log q_n \sum_{j\in G} E\left[\bar{\Phi}(t/|T_{1j}| - U_j) \,|\, C_t\right],
\end{aligned}
$$

where $\bar{\Phi}(x) = 1 - \Phi(x)$. The second to last inequality is due to

$$
\frac{x}{x^2 + 1}\phi(x) < \bar{\Phi}(x), \quad \text{for all } x > 0.
$$

On the other hand,

$$
\sum_{j\in G} Pr(W_j > t) = \sum_{j\in G} E\left[\bar{\Phi}(t/|T_{1j}| - U_j) \,|\, C_t\right]\{1 + o(1)\}.
$$

Therefore, by Markov inequality and similar arguments in the proof of Lemma S.2, the assertion holds if $c_{np}a_{np}s_n\sqrt{n\log p}\log q_n h_n \to 0$. Note that h_n can be made arbitrarily small as long as $h_n \to \infty$ as $n \to \infty$, from which we completes the proof.

In the next lemma, we obtain the approximation result for those j with relatively large μ_j.

Lemma S.5 Suppose Conditions 1, 2, 3, 4 and 7 hold and $c_{np} a_{np} s_n \bar{q}_n \sqrt{n \log p} (\log \bar{q}_n)^{1+\gamma} \to 0$. Then, for any $M > 0$,

$$\sup_{M \le t \le G^{-1}(\alpha \eta_n / q_{0n})} \frac{\sum_{j \in G^c} I(\hat{W}_j \ge t)}{\sum_{j \in G^c} I(W_j \ge t)} - 1 = o_p(1),$$

$$\sup_{M \le t \le G^{-1}(\alpha \eta_n / q_{0n})} \frac{\sum_{j \in G^c} I(\hat{W}_j \le -t)}{\sum_{j \in G^c} I(W_j \le -t)} - 1 = o_p(1).$$

Proof. Under the designed conditions, we have $W_j = \hat{W}_j \{1 + o_p(1)\}$ for any $j \in G^c$ uniformly. Then the results follow.

D Proofs of Additional Theoretical Results

D.1 Proof of Lemma 1 (the coin-flip property under dependence)

Observe that $W_j = \sqrt{n_1 n_2 / (n_1 + n_2)} (\hat{\mu}_{1j} - \hat{\mu}_{2j}) / \sigma_j^2 =: c_j \times (\hat{\mu}_{1j} - \hat{\mu}_{2j})$. Conditional on D_1, we have $W_j \mid W_{-j} \sim N(\mu_{j|-j}, \sigma^2_{j|-j})$ with

$$\mu_{j|-j} = \mathrm{Cov}(W_j, W_{-j}) \mathrm{Var}(W_{-j})^{-1} (W_{-j} - E W_{-j}) \text{ and}$$

$$\sigma^2_{j|-j} = \mathrm{Var}(W_j) - \mathrm{Cov}(W_j, W_{-j}) \{\mathrm{Var}(W_{-j})\}^{-1} \mathrm{Cov}(W_j, W_{-j})^{\top}.$$

For any $k, l \in S$, we have $\mathrm{Cov}(W_k, W_l) = c_k c_l Q_{kl}$. Let $C = \mathrm{diag}\{c_1, \ldots, c_{q_n}\}$ and $D = CQC$.

Then $Cov(W_j, W_{-j}) = D_{j,-j}$ and $Var(W_{-j}) = D_{-j,-j}$. So, we obtain that

$$
\begin{aligned}
&Pr(W_j > 0 \,|\, |W_j|, W_{-j}, D_1) \\
&= \frac{\phi\left(\frac{|W_j|-\mu_{j|-j}}{\sigma_{j|-j}}\right)}{\phi\left(\frac{|W_j|-\mu_{j|-j}}{\sigma_{j|-j}}\right) + \phi\left(\frac{|W_j|+\mu_{j|-j}}{\sigma_{j|-j}}\right)} \\
&= \frac{\phi\left(\frac{|W_j| - D_{j,-j}D^{-1}_{j,-j}(W_{-j}-EW_{-j})}{\sqrt{D_{jj} - D_{j,-j}D^{-1}_{-j,-j}D_{-j,j}}}\right)}{\phi\left(\frac{|W_j| - D_{j,-j}D^{-1}_{j,-j}(W_{-j}-EW_{-j})}{\sqrt{D_{jj} - D_{j,-j}D^{-1}_{-j,-j}D_{-j,j}}}\right) + \phi\left(\frac{|W_j| + D_{j,-j}D^{-1}_{j,-j}(W_{-j}-EW_{-j})}{\sqrt{D_{jj} - D_{j,-j}D^{-1}_{-j,-j}D_{-j,j}}}\right)} \\
&:= \Delta_j(|W_j|, W_{-j}, D_1)
\end{aligned}
$$

Denote $Q_{-j,j} = 0$ the jth column of Q excluding Q_{jj}. Finally we have

$$
\begin{aligned}
Pr(W_j > 0 \,|\, |W_j|, W_{-j}) &= E\{Pr(W_j > 0 \,|\, |W_j|, W_{-j}, D_1) \,|\, |W_j|, W_{-j}\} \\
&= E\{\Delta_j(|W_j|, W_{-j}, D_1) \,|\, |W_j|, W_{-j}\} - 1/2.
\end{aligned}
$$

It can be easily verified that if $Q_{-j,j} = 0$, $\Delta_j(|W_j|, W_{-j}, D_1) = 1/2$ and consequently $\Delta_j = 0$.

D.2 Asymptotic results for R-SDA and two–sample SDA

The next result is a direct corollary of Theorem 2 which establishes the FDR control of the multi-splitting procedure R-SDA.

Corollary 1 Suppose Conditions 1–6 hold. For any $\alpha \in (0,1)$ and a given B, the FDR of the R-SDA method satisfies $\limsup_{(n,p)\to\infty} FDR \le \alpha$.

As in (14), the FDP is controlled for each replication so is the FDP of R-SDA, resulting in the FDR control.

To establish the FDR control result of SDA procedure for the two-sample problem, we introduce a new sequence of independent random variables $\{\xi_i\}$ defined as follows:

$$
\xi_i - \omega = \begin{cases} n_2/n_2^{(1)}(\xi^{(1)}_{2i} - \mu^{(1)}); & 1 \le i \le n_2^{(1)}; \\ -n_2/n_2^{(2)}(\xi^{(2)}_{2i+n_2^{(1)}} - \mu^{(2)}); & n_2^{(1)} + 1 \le i \le n_2. \end{cases}
$$

Note that

$$\xi_2^{(1)} - \xi_2^{(2)} - \omega = \frac{1}{n_2^{(1)}} \sum_{i=1}^{n_2^{(1)}} (\xi_{2i}^{(1)} - \mu^{(1)}) - \frac{1}{n_2^{(2)}} \sum_{i=1}^{n_2^{(2)}} (\xi_{2i}^{(2)} - \mu^{(2)}) = \frac{1}{n_2} \sum_{i=1}^{n_2} (\xi_i - \omega).$$

By the proofs for Theorem 2, if we replace μ as ω and set $\Omega^{-1} = \Sigma^{(1)}/\pi + \Sigma^{(2)}/(1-\pi)$ with $\pi = \lim n_l^{(1)}/n_l$, Theorem 2 holds also for the two-sample problem.

Corollary 2 Suppose Conditions 1–6 hold. For any $\alpha \in (0,1)$ and $0 < \pi < 1$, the FDR of the SDA for the two-sample problem satisfies $\limsup_{(n,p)\to\infty} \mathrm{FDR} \le \alpha$.

We want to emphasize that as long as Condition 2 is satisfied, the above results hold for other choices of T_{1j} as discussed in Appendix A.2. For example, consider a hard-thresholding estimator $\hat{\mu}_{1j} = \xi_{1j} I(|\xi_{1j}| > c\sqrt{\log p/n})$ for some $c > 0$. We know that $c_{np} = \sqrt{\log p/n}$ if ξ_{ij}'s have uniformly bounded fourth moments.

D.3 Additional lemmas

The first one is the standard Bernstein's inequality.

Lemma S.6 (Bernstein's inequality) Let $X_1, \ldots, X_n$ be independent centered random variables a.s. bounded by $A < \infty$ in absolute value. Let $\sigma^2 = n^{-1} \sum_{i=1}^n E(X_i^2)$. Then for all $x > 0$,

$$\Pr\left(\sum_{i=1}^n X_i \ge x \right) \le \exp\left(- \frac{x^2}{2n\sigma^2 + 2Ax/3} \right).$$

The second one is a moderate deviation result for the mean; See Petrov (2002).

Lemma S.7 (Moderate deviation for the independent sum) Suppose that $X_1, \ldots, X_n$ are independent random variables with mean zero, satisfying $E(|X_j|^{2+\delta}) < \infty$ $(j = 1, 2, \ldots)$. Let $B_n = \sum_{i=1}^n E(X_i^2)$. Then,

$$\frac{\Pr(\sum_{i=1}^n X_i > x\sqrt{B_n})}{1 - \Phi(x)} \to 1,$$

as $n \to \infty$ uniformly in x in the domain $0 \le x \le C\{2\log(1/L_n)\}^{1/2}$, where $L_n = B_n^{-1-\delta/2} \sum_{i=1}^n E|X_i|^{2+\delta}$ and C is a positive constant satisfying the condition $C < 1$.

The next lemma establishes uniform bounds for $\hat{\mu}_{2j}$.

Lemma S.8 Suppose Conditions 1, 3, and 4 hold. Then, as $n \to \infty$, $\Pr\sigma^{-1}$

$$|\hat{\mu}_{2j} - \mu_j| > \sigma_j C \log \bar{q}_n \sqrt{n_2 / |D_1|} = o(1/\bar{q}_n),$$

holds uniformly in S, where $C > 4$.

Proof. Write

$$\hat{\mu}_{2j} - \mu_j = n_2^{-1} \sum_{i=1}^{n_2} e_j' \hat{\Sigma}_{2S}^{-1} X_{2S}' \varepsilon_i := n_2^{-1} \sum_{i=1}^{n_2} \xi_{ij}.$$

Let $m_n = (n_2 \bar{q}_n)^{1/\theta+\gamma} K_{n2}$ and note that

$$\begin{aligned}\xi_{ij} &= \xi_{ij} I(|\xi_{ij}| \le m_n) - E\{\xi_j I(|\xi_j| \le m_n)\} + \xi_{ij} I(|\xi_{ij}| > m_n) - E\{\xi_j I(|\xi_j| > m_n)\} \\ &=: \xi_{ij,1} + \xi_{ij,2}.\end{aligned}$$

Conditioned on the first split D_1,

$$\begin{aligned}&\Pr\left(|\sqrt{n_2}(\hat{\mu}_{2j} - \mu_j)| > \sigma_j x \text{ for some } j \mid D_1\right) \\ &= \Pr\left(\sum_{i=1}^{n_2} \xi_{ij,1} + \sum_{i=1}^{n_2} \xi_{ij,2} > \sqrt{n_2}\sigma_j x \text{ for some } j \mid D_1\right) \\ &\le \Pr\left(\sum_{i=1}^{n_2} \xi_{ij,1} + \sum_{i=1}^{n_2} \xi_{ij,2} > \sqrt{n_2}\sigma_j x \text{ for some } j \mid D_1\right) \\ &\le \Pr\left(\sum_{i=1}^{n_2} \xi_{ij,1} > \sqrt{n_2}\sigma_j x(1-a) \text{ for some } j \mid D_1\right) \\ &\quad + \Pr\left(\sum_{i=1}^{n_2} \xi_{ij,2} > \sqrt{n_2}\sigma_j x a \text{ for some } j \mid D_1\right) =: P_1 + P_2. \qquad \text{(S.6)}\end{aligned}$$

Here a is a small positive value.

Firstly consider the term P_1. Note that $\xi_{1j,1}, \ldots, \xi_{n_2 j,1}$ are independent centered random variables a.s. bounded by $2m_n$ in absolute value. Then the Bernstein inequality in Lemma S.6 yields that

$$P_1 \le 2q_n \max_j \exp\left(-\frac{n_2 \sigma_j^2 x^2 (1-a)^2}{2n_2 E(\xi_{j,1}^2) + 2 \cdot 2 m_n \cdot \sqrt{n_2}\sigma_j x(1-a)/3}\right).$$

Recall that $\epsilon_{ij,1} = \epsilon_{ij} I(|\epsilon_{ij}| \le m_n) - E[\epsilon_j I(|\epsilon_j| \le m_n)]$. Thus

$$E(\epsilon_{j,1}^2) = \mathrm{Var}\{\epsilon_j I(|\epsilon_j| \le m_n)\} \le E\{\epsilon_j I(|\epsilon_j|^2 \le m_n)\} \le E(\epsilon_j^2) = Q_{jj}.$$

We then have:

$$\begin{aligned} P_1 &\le 2\bar{q}_n \max_j \exp\left(-\frac{n_2\sigma_j^2 x^2(1-a)^2}{2n_2 Q_{jj} + 2\cdot 2m_n \cdot \sqrt{n_2}\sigma_j x(1-a)/3}\right) \\ &\le 2\bar{q}_n \max_j \exp\left(-\frac{x^2(1-a)^2}{2\sigma^2 + 4(1-a)\sigma_j^{-1} x m_n/(3\sqrt{n_2})}\right). \end{aligned} \tag{S.7}$$

Next we turn to consider P_2. First note that

$$P_2 \le \Pr\left(\sum_{i=1}^{n_2} \max_j |\epsilon_{ij}| I(|\epsilon_{ij}| > m_n) + \max_j n_2 E\{|\epsilon_j| I(|\epsilon_j| > m_n)\} > \sqrt{n_2}\sigma_j \sqrt{x} a \mid D_1\right)$$

Further note that

$$E^2\{|\epsilon_j| I(|\epsilon_j| > m_n)\} \le E(\epsilon_j^2) \Pr(|\epsilon_j| > m_n) \le E(\epsilon_j^2) \frac{E(|\epsilon_j|^\theta)}{m_n^\theta}$$

We then conclude that

$$\max_j n_2 E\{|\epsilon_j| I(|\epsilon_j| > m_n)\} \le \max_j n_2 \frac{\sqrt{E(\epsilon_j^2) E(|\epsilon_j|^\theta)}}{m_n^{\theta/2}} = o(\sqrt{n_2}).$$

From this, we then have

$$\begin{aligned} P_2 &\le \Pr\left(\sum_{i=1}^{n_2} \max_j |\epsilon_{ij}| I(|\epsilon_{ij}| > m_n) > \sqrt{n_2}\sigma_j \sqrt{x} a/2 \mid D_1\right) \\ &\le \Pr\left(\max_j |\epsilon_{ij}| > m_n \text{ for some } i \mid D_1\right) \\ &\le n_2 \frac{E(kA(S)\varepsilon_i k_\infty^\theta)}{m_n^\theta} = o(\bar{q}_n^{-1}). \end{aligned} \tag{S.8}$$

Let $x = \sigma \sqrt{C \log \bar{q}_n}$. From the inequalities (S.6), (S.7), and (S.8), we conclude that

$$\begin{aligned} &\Pr(|\sqrt{n_2}(\hat{\mu}_{2j} - \mu_j)| > \sigma_j x \text{ for some } j \mid D_1) \\ &\le 2\bar{q}_n \max_j \exp\left(-\frac{x^2(1-a)^2}{2\sigma^2 + 4(1-a)\sigma_j^{-1} x m_n/(3\sqrt{n_2})}\right) + o(\bar{q}_n^{-1}) = o(\bar{q}_n^{-1}). \end{aligned}$$

holds uniformly in S, where we use the condition $m_n/\sqrt{n/\log \bar{q}_n} = o(1)$ which is implied by Condition 3.

E Additional Numerical Results

E.1 Estimated covariance structures

This section compares the methods mentioned in Section 5 for the unknown covariance case. In practice, one should adopt the most appropriate estimator tailored to specific correlation structures. Specifically, we have used the method based on Cholesky decomposition in Bickel and Levina (2008), the POET method proposed by Fan et al. (2013), and the graphical lasso (Friedman et al., 2008) to estimate the unknown Structures (I)–(III), respectively.

Figure S3 follows the settings in Figure 4 (except that the covariance matrix or its inverse is estimated). Figures S4 uses the same settings as those in Figures 5 with estimated covariance matrix. We omit a detailed discussion as the observed patterns seem to be very similar to those in the known covariance case (except that the FDR control sometimes becomes less accurate due to the additional estimation errors). Our conclusions based on Figure S3 and Figures S4 remain essentially the same as before. Knockoff and R-SDA seem to be the only methods that can control the FDR reasonably well in all scenarios, with the R-SDA method having much higher power in most scenarios.

E.2 Additional comparisons

Figure S5 demonstrates the FDR and AP for various signal magnitude μ under the compound symmetry error structure (II) and three error distributions, for known and unknown covariance structures, respectively.

E.3 Boxplots of FDPs

When the noises are sampled from the multivariate normal distribution, Figure S6 shows the boxplot of the FDP and AP of the testing procedures for $\pi_1 = 0.05$ and 0.2, while fixing $(n, p, \alpha) = (90, 500, 0.2)$. The signal magnitude μ is adjusted according to the covariance structures so that the APs are in a similar range. While the BH is conservative with little power, the R-SDA outperforms

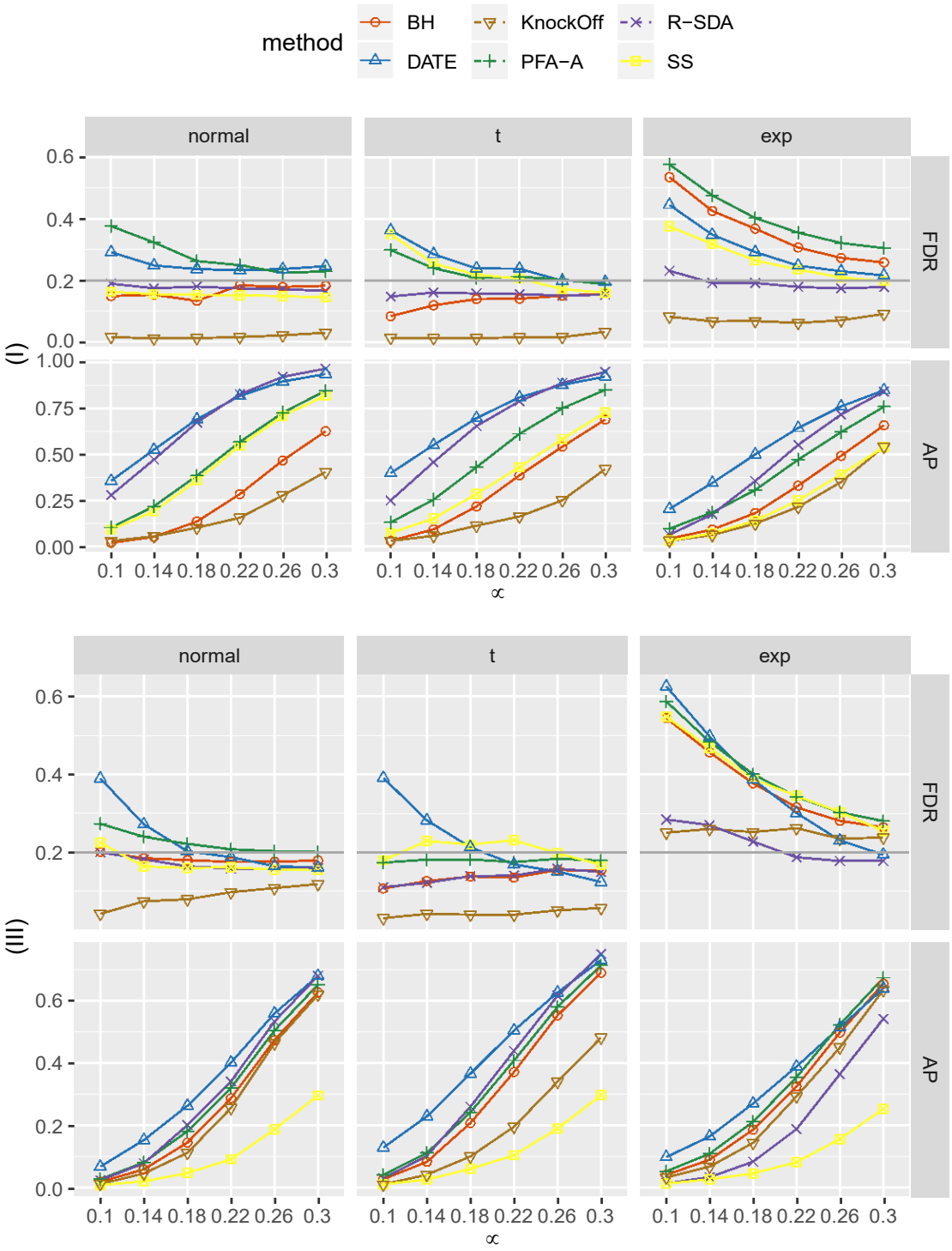


Figure S3: FDR and AP comparison for varying μ in Settings (I) and (III) with estimated covariance matrix.

the DATE and PFA in the sense that it provides more accurate estimate of FDP and generally higher power. The conclusions are consistent for different choices of π_1, with narrower interquartile range of FDP and AP for larger π_1. As we can expect, the R-SDA has smaller variation than the single-splitting SDA.

E.4 The impact of the number of tests and sample sizes

We also conduct experiments by altering the number of tests p, while keeping $(n, \pi_1, \alpha) = (90, 0.1, 0.2)$. To make the AP comparable across p, the signal μ is adjusted via $\mu = C\sqrt{\log p/n}$ with C depending on the covariance structures. The results are summarized in the top half of Figure S7. We can see that all methods have more accurate control of FDR as p increases, but the PFA_A and DATE fail to control the FDR when p is small. To investigate the impact on sample sizes with unknown covariance, we set $\mu = C\sqrt{\log(p)/n}$, fix $(p, \pi_1, \alpha) = (500, 0.1, 0.2)$, and consider the normal error. The results are summarized in the bottom half of Figure S7. We can see that that our R-SDA method is able to control the FDR and close to the nominal level regardless of the choice of n. Its superior performance relative to the other three methods is significant in some cases. Though all the methods exhibit steady AP pattern, the BH, PFA and DATE appear to need larger sample to achieve satisfactory FDR control than the R-SDA does. This again concurs with our theoretical result in Theorem 2 and demonstrates the advantage of using the nonparametric estimation of FDP in the SDA procedure.

E.5 List of selected genes by different methods

Table S1 reports the list of 19 most differentially expressed probe sets obtained by the methods R-SDA, BH, SS, PFA-A and DATE in the real-data example.

Table S1: Differentially expressed probe sets in the B lineage ALL with BCR/ABL versus NEG molecular rearrangement, for five different multiple testing adjustment methods

R-SDA	BH	SS	PFA	DATE
1635_at	1636_g_at	39730_at	1636_g_at	36502_at
39730_at	39730_at	39317_at	39730_at	38385_at
1636_g_at	1635_at	37027_at	1635_at	40202_at
36502_at	1674_at	38052_at	1674_at	37403_at
37403_at	40504_at	1635_at	40202_at	38052_at
32134_at	40202_at	1636_g_at	37403_at	33690_at
38052_at	37015_at	40202_at	32434_at	39317_at
36821_at	37027_at	34850_at	37014_at	40876_at
38385_at	32434_at	37403_at	32979_at	33440_at
37027_at	40167_s_at	37024_at	1249_at	1674_at
1674_at	40480_s_at	1249_at	38111_at	36908_at
41872_at	36591_at	36802_at	37015_at	33774_at
33440_at	33774_at	37025_at	37147_at	39730_at
32434_at	37403_at	32979_at	40504_at	41592_at
40876_at	37014_at	34870_at	33440_at	32134_at
40202_at	37363_at	36502_at	38112_g_at	39070_at
39317_at	34472_at	33891_at	36502_at	37558_at
32562_at	32542_at	34800_at	31786_at	33304_at
34990_at	39329_at	36543_at	34850_at	34180_at

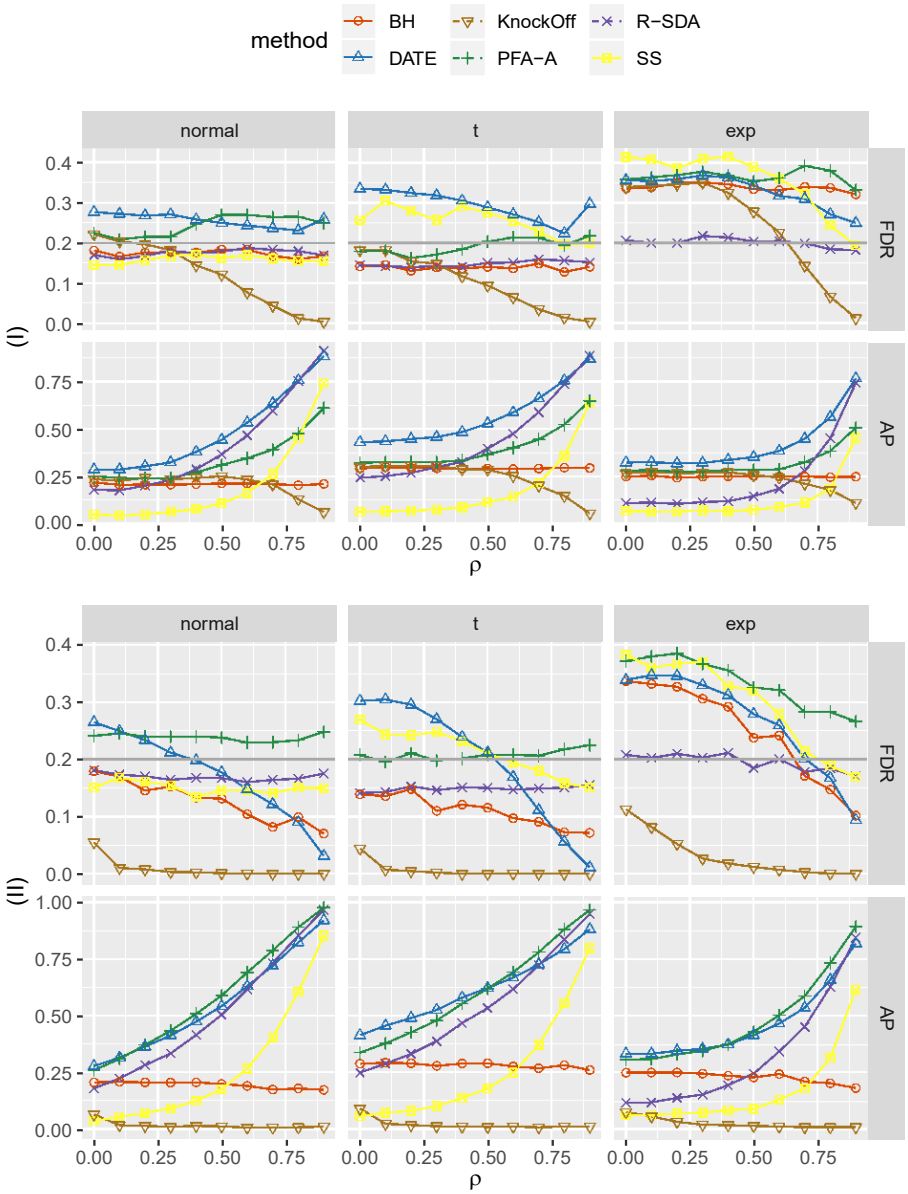


Figure S4: FDR and AP comparison for varying ρ in Settings (I)–(II) with estimated covariance matrix.

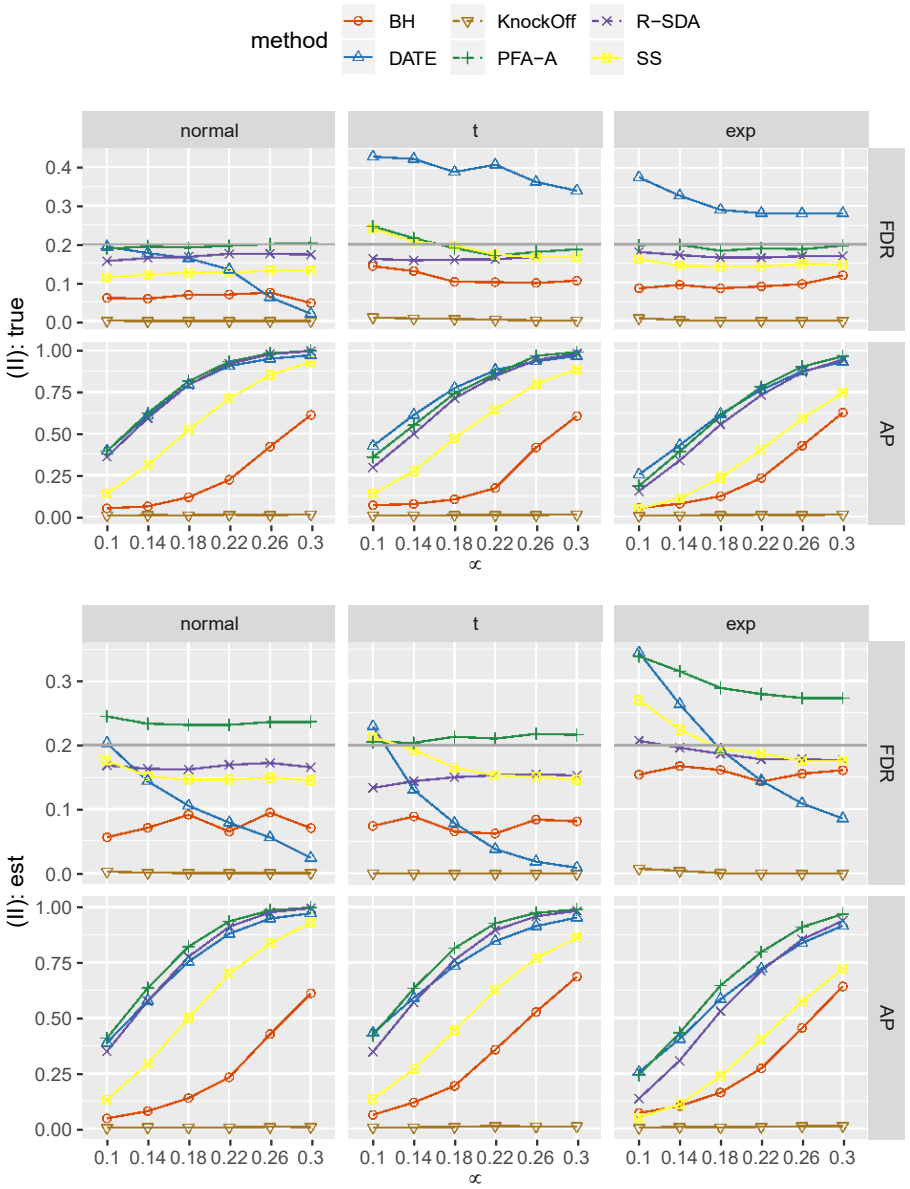


Figure S5: FDR and AP comparison for varying μ in Setting (II) with known (top half) and unknown variances (bottom half).

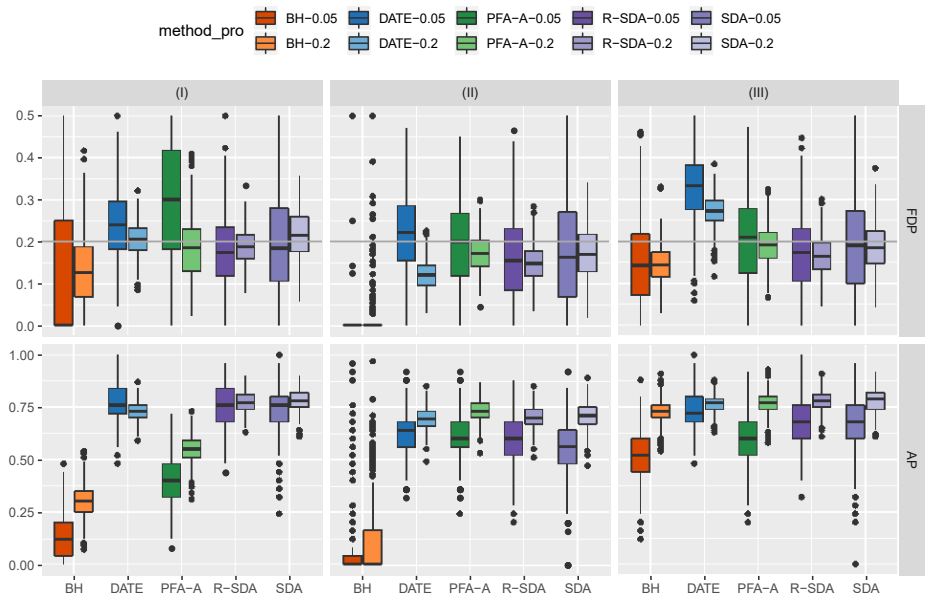


Figure S6: The boxplot of FDP and AP when the proportions of alternative are 0.05 and 0.2. The normal error is considered and $(n, p, \alpha) = (90, 500, 0.2)$. The signal strength μ is set as 0.2, 0.15, 0.3 for the covariance structures (I)-(III), respectively.

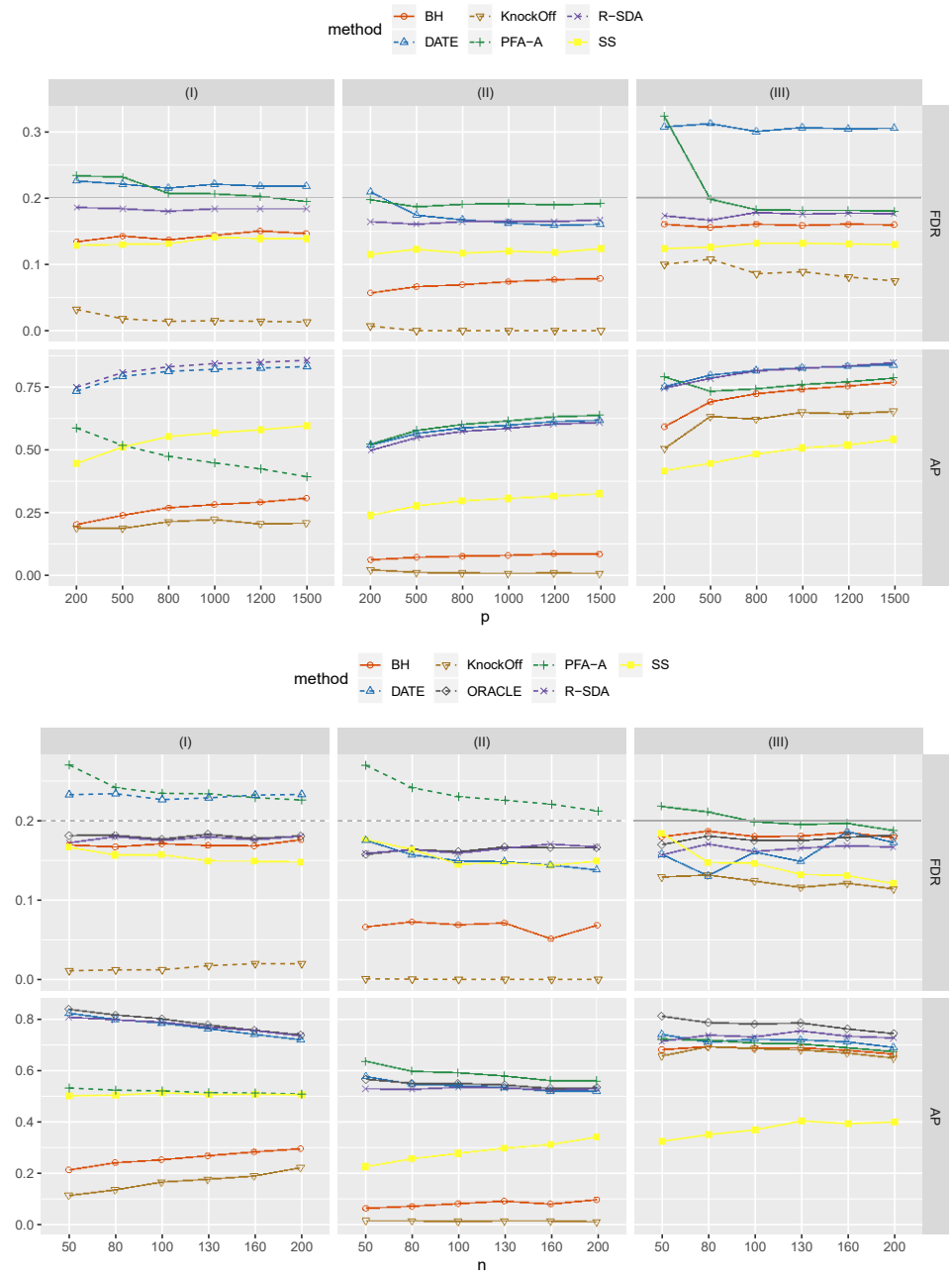
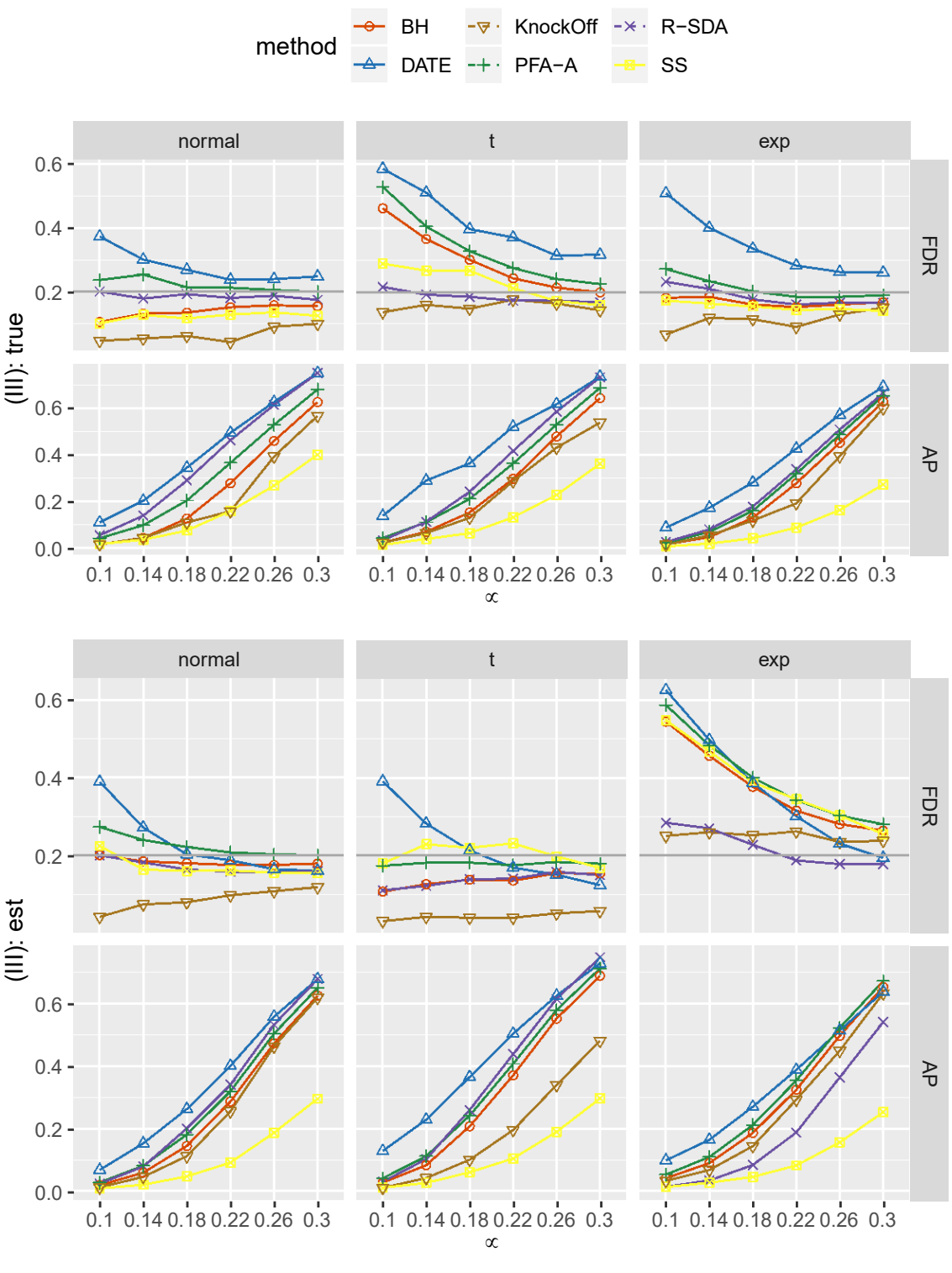


Figure S7: Top half: The empirical FDR and AP for varying p. $(n, \pi_1, \alpha) = (90, 0.1, 0.2)$ and $\mu_n = C\sqrt{p/\log(p)/n}$ with $C = 0.8, 0.5, 1.2$. Bottom half: The FDR and AP for varying n when the covariances are estimated. $(p, \pi_1, \alpha) = (500, 0.1, 0.2)$ and $\mu_n = C\sqrt{p/\log p/n}$ with $C = 0.8, 0.5, 1.2$.


method
BH
DATE
KnockOff
PFA−A
R−SDA
SS
normal
t
exp
FDR
AP
(III): true
(III): est
0.0
0.2
0.4
0.6
0.1
0.14
0.18
0.22
0.26
0.3
∝