Statistica Sinica Preprint No: SS-2024-0022	
Title	Multiple Testing of One-Sided Hypotheses under Unknown Dependence
Manuscript ID	SS-2024-0022
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0022
Complete List of Authors	Seonghun Cho, Youngrae Kim, Johan Lim, Hyungwon Choi, DoHwan Park and Woncheol Jang
Corresponding Authors	Woncheol Jang
E-mails	wcjang@snu.ac.kr

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Multiple Testing of One-Sided Hypotheses under Unknown Dependence

Seonghun Cho[1], Youngrae Kim[2], Johan Lim[2], Hyungwon Choi[3],
DoHwan Park[4] and Woncheol Jang*[2]

[1]*Inha University,* [2]*Seoul National University,*
[3]*National University of Singapore, and* [4]*University of Maryland at Baltimore County*

Abstract: The one-sided hypotheses in a multiple testing problem make the empirical null distribution (or p-values) conservative. Furthermore, it introduces a significant loss of power if not appropriately considered. We propose a multiple testing procedure named discarding adaptively with bounding on principal factor approximation (DAB-PFA) to simultaneously test a number of one-sided hypotheses under the general dependency of test statistics. Specifically, we use the principal factor approximation (PFA) by Fan and Han (2017) to account for the dependence structure among test statistics and adaptively discard small or large p-values when estimating the realized false discovery proportion (FDP). We derive the convergence rate of the proposed estimator and numerically compare the false discovery rate (FDR) and the true positive rate (TPR) of our method to many existing procedures, including those from Benjamini and Hochberg (1995), Efron (2004), and Wang and Fan (2017). We demonstrate our method through simulation studies and analysis of protein phosphorylation levels for serous ovarian adenocarcinoma samples.

Key words and phrases: Discarding adaptively with bounding (DAB), Principal Factor Approximation, Conservative null, False discovery rate, Multiple testing, One-sided hypothesis

*Corresponding author

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

1. Introduction

1.1 Multiple testing

In the last two decades, multiple testing, in which many hypotheses are tested simultaneously, has been one of the few central topics of statistics. The multiple testing problem arises in various applications, which include microarray analysis in genetics, functional magnetic resonance imaging (fMRI) studies of the brain, clinical trials with multiple endpoints, and tens of thousands of A/B tests performed by major internet companies. Early works on this topic introduced various type I errors, including the family-wise error rate (FWER), the generalized FWER, the false discovery rate (FDR), the positive FDR, and proposed procedures to control the aforementioned errors at the nominal level. Researchers mostly assumed that test statistics used for each hypothesis test are independent or weakly dependent. However, the independence assumption of test statistics is easily broken in practice, which makes the control of the multiple testing error rate inaccurate. Great efforts have been made to construct procedures that consider the dependency of test statistics. Some early works in this area focused on studying the validity of the proposed procedures, meaning that the FDR is controlled at a nominal level under some classes of dependence structure among the tests(Benjamini and Yekutieli, 2001; Finner and Roters, 2002; Efron, 2004, 2007; Owen, 2005; Sarkar, 2006; Romano et al., 2008; Wu, 2008). However, it was suggested that efficiency in terms of the false negative rate (FNR) should be considered in multiple testing (Genovese and Wasserman, 2002; Sarkar, 2004), and some works showed that efficiency improvements could be made by taking into account the dependence structure (Sun and Cai, 2009; Wei et al., 2009; Xiao et al., 2013; Liu et al., 2016; Fan et al., 2012; Fan and Han, 2017). The procedure proposed by Sun and Cai (2009) and its extensions (Wei et al., 2009; Xiao

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

et al., 2013) were based on a hidden Markov model (HMM), which only allows sequential dependence. Liu et al. (2016) replaced the HMM with a Markov-random-field-coupled mixture model, which can be applied to more general dependence structures. Fan et al. (2012) proposed the principal factor approximation (PFA) to estimate the realized false discovery proportion (FDP) under an arbitrary but known dependence structure, and Fan and Han (2017) extended the PFA method under an unknown dependence structure.

1.2 One-sided hypothesis

The main goal of this study is to develop a multiple testing procedure of one-sided hypotheses for the FDR control under unknown dependence. There are many applications where testing a number of one-sided hypotheses is of primary interest. For instance, researchers aim to identify protein modification levels that are uniquely elevated in a specific group of subjects compared to other groups in protein phosphorylation analysis. In clinical trials, noninferiority and superiority tests are commonly required to assess the benefit of new drugs. Another example of multiple one-sided hypotheses testing is the tens of thousands of A/B tests that major internet companies perform. For more detailed examples, refer to Cohen and Sackrowitz (2005); Tian and Ramdas (Tian and Ramdas) and the references therein.

The main difficulty in multiple testing of one-sided hypotheses results from the conservative null p-value. When performing hypothesis testing, one of the standard assumptions is that the p-value P is *valid*, which means that if the null hypothesis is true, then we have $\Pr(P \leq u) \leq u$ for all $u \in [0,1]$. If the inequality is strict for some $u \in [0,1]$, then the null p-value P is said to be *conservative*. In a one-sided hypothesis test, the null p-value is typically conservative since the true parameter of interest is rarely exactly at the boundary

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

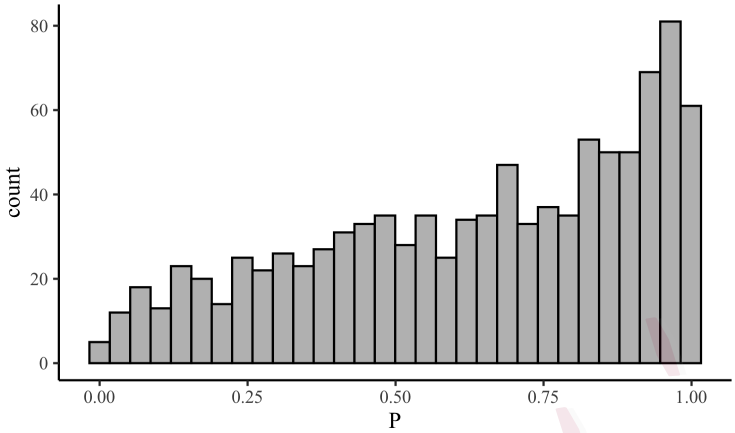

Figure 1: Histogram of p-values $\{P_j = \Phi(-Z_j)\}_{j=1}^{1000}$, where Z_js are independently generated from $N(\mu_j, 1)$ and $\mu_j \sim U[-1, 0]$.

of the null set. Figure 1 presents a histogram of p-values $\{P_j = \Phi(-Z_j)\}_{j=1}^{1000}$, where Z_js are independently generated from $N(\mu_j, 1)$ with $\mu_j \sim U[-1, 0]$. If we apply existing multiple testing procedures such as Fan et al. (2012); Fan and Han (2017) for the FDR control without considering the conservativeness of the null p-values, we will overestimate the realized false discovery proportion (FDP). This leads to a significant loss of power despite those procedures remaining valid for FDR control.

One general solution for the conservative null p-values is to discard p-values close to 1 (Zhao et al., 2019). This approach was also used in the online FDR control setting, which assumes an infinite sequence of p-values. Tian and Ramdas (Tian and Ramdas) proposed an adaptive algorithm that discards conservative nulls. By discarding overly conservative p-values, a power increase was achieved. However, these current procedures for conservative nulls are based on the independence assumption among test statistics. In this paper, we propose a procedure to control the FDR in testing multiple one-sided hypotheses under general unknown dependence.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Our main interest is to test p one-sided hypotheses

$$\mathcal{H}_{0j} : \mu_j \leq 0 \quad \text{vs.} \quad \mathcal{H}_{1j} : \mu_j > 0 \tag{1.1}$$

for each $j = 1, 2, \ldots, p$ based on a test statistic Z_j, where the vector $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$ of test statistics follows a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ and covariance matrix $\boldsymbol{\Sigma}$. Here, we assume that $\boldsymbol{\Sigma}$ is an unknown correlation matrix, which implies that the marginal variance of each test statistic is known, while the dependence structure among the test statistics is unknown. Then, for each $j = 1, \ldots, p$, the p-value for the j-th hypothesis is calculated as $P_j = \Phi(-Z_j)$, where Φ is the cumulative density function of the standard normal distribution. We use a common threshold value $t \in (0, 1)$. We reject the j-th hypothesis $\mathcal{H}_{0j}$ if and only if the corresponding p-value P_j does not exceed the threshold value t. Define $R(t) = \#\{j : P_j \leq t\}$ as the number of rejections (or discoveries), and $V(t) = \#\{j \in \mathcal{H}_0 : P_j \leq t\}$ as the number of false rejections, where $\mathcal{H}_0$ is the set of indices of true nulls. Under these definitions, we aim to control the FDR, which is defined by

$$\mathrm{FDR}(t) = \mathbb{E}\left\{\mathrm{FDP}(t)\right\} = \mathbb{E}\left\{\frac{V(t)}{R(t) \vee 1}\right\}, \tag{1.2}$$

under a predetermined level $\alpha \in (0, 1)$.

To control the FDR, we need to estimate the realized false discovery proportion $\mathrm{FDP}(t)$ for a given threshold level $t \in (0, 1)$ and find an optimal level $\widehat{t}$ such that $\widehat{\mathrm{FDP}}(\widehat{t}) \leq \alpha$. Since the number of rejections $R(t)$ is observable, we only need to approxomate the number of false rejections $V(t)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

1.3 Principal factor approximation (PFA)

To approximate $V(t)$ under the unknown dependence of test statistics $\{Z_j\}_{j=1}^p$, we use an approximated factor model as in Fan and Han (2017). Let $\{(\lambda_j, \boldsymbol{\gamma}_j)\}_{j=1}^p$ be the eigenvalues and the corresponding eigenvectors of $\boldsymbol{\Sigma}$ with the ordering $\lambda_1 \geq \cdots \geq \lambda_p$. For a fixed integer K satisfying

$$\text{(C1) } p^{-1}\sqrt{\lambda_{K+1}^2 + \cdots + \lambda_p^2} = O(p^{-\delta}) \text{ for some } \delta > 0,$$

the correlation matrix $\boldsymbol{\Sigma}$ is decomposed by

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^\top + \boldsymbol{\Sigma}_u, \tag{1.3}$$

where $\mathbf{B} = (\sqrt{\lambda_1}\boldsymbol{\gamma}_1, \ldots, \sqrt{\lambda_K}\boldsymbol{\gamma}_K) \in \mathbb{R}^{p\times K}$ and $\boldsymbol{\Sigma}_u = \sum_{k=K+1}^p \lambda_k \boldsymbol{\gamma}_k \boldsymbol{\gamma}_k^\top$. Then, the vector of test statistics $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is stochastically decomposed by $\mathbf{Z} = \boldsymbol{\mu} + \mathbf{B}\mathbf{W} + \mathbf{u}$. Here $\mathbf{W} = (W_1, \ldots, W_K)^\top \sim N_K(\mathbf{0}, \mathbf{I}_K)$ are common factors and $\mathbf{u} = (u_1, \ldots, u_p)^\top \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_u)$ are the errors, independent of $\mathbf{W}$.

We note that the condition (C1) implies that the errors $\{u_1, \ldots, u_p\}$ are *weakly dependent*, that is,

$$\lim_{p\to\infty} p^{-2} \sum_{j_1, j_2=1}^p |\sigma_{u, j_1 j_2}| = 0, \tag{1.4}$$

where $\sigma_{u,j_1 j_2}$ denotes the (j_1, j_2)-th element of $\boldsymbol{\Sigma}_u$.

Define $a_j = (1 - \|\mathbf{b}_j\|_2^2)^{-1/2}$ and $\eta_j = \mathbf{b}_j^\top \mathbf{W}$, and $\mathbf{b}_j^\top$ is the j-th row of $\mathbf{B}$. Using a similar argument as one used in Fan et al. (2012), we can show that

$$V_{\text{orc}}(t) = \mathbb{E}\{V(t)|\mathbf{W}\} = \sum_{j\in\mathcal{H}_0} \Phi(a_j(\mu_j + z_t + \eta_j)) \tag{1.5}$$

approximates the number of false rejections

$$V(t) = \sum_{j\in\mathcal{H}_0} I(P_j \leq t) = \sum_{j\in\mathcal{H}_0} I(Z_j \geq -z_t) \tag{1.6}$$

where $z_t = \Phi^{-1}(t)$ is the t-quantile of the standard normal distribution.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

1.4 Discarding adaptively with bounding (DAB)

Since $V_{\text{orc}}(t)$ can not be used in practice, we may consider the following upper bound of $V_{\text{orc}}(t)$.

$$V_{\text{U,naive}}(t) = \sum_{j=1}^{p} \Phi(a_j(z_t + \eta_j)) \tag{1.7}$$

However, the difference between $V_{\text{orc}}(t)$ and $V_{\text{U,naive}}(t)$ is not small and results in a significant loss of power. There are two sources that make the difference large; the first source is conservative null p values, and the other source is a nonignorable proportion of nonnulls. Both are directly related to the way that the upper bound is built, where μ_j is replaced with 0 and the summation is expanded over the true null index set $\mathcal{H}_0$ to the entire index set $\{1, 2, \ldots, p\}$.

To have a precise upper bound of $V_{\text{orc}}(t)$, we propose discarding p values close to 0 or 1, following the idea by Tian and Ramdas (Tian and Ramdas). This can be simply done by introducing an indicator $I(\lambda < P_j \leq \tau)$ for $\lambda, \tau \in (0, 1)$ as done in Tian and Ramdas (Tian and Ramdas). However, to consistently approximate $V_{\text{orc}}(t)$ under the general dependency (among the statistics), we require much more delicate terms in the upper bound that is obtainable. For example, under the independence assumption, we may consider the naive truncated term $\sum_{j=1}^{p} \Phi(a_j(z_t + \eta_j))I(\lambda < P_j \leq \tau)/(\tau - \lambda)$. However, under the unknown dependency, the denominator of the naive truncated term should be replaced with a function of cumulative probabilities of the standard normal distribution with an estimated mean and variance.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

1.5 Structure of this paper

In the following, we summarize major theoretical developments of this paper and a road map to show how we build a precise upper bound of $V_{\text{orc}}(t)$ and summarize its theoretical properties. Note that there is no sparsity assumption of the nonnull, which is common in existing literature, including Fan et al. (2012) and Fan and Han (2017), to make the difference of the sum over $\mathcal{H}_0$ and over the entire index set $\{1, 2, \ldots, p\}$ small.

Proposition 1 The term we would like to estimate in (1.2) is

$$V(t) = \sum_{j \in \mathcal{H}_0} I(\Phi(-Z_j) \le t). \tag{1.8}$$

In Proposition 1, we will show that $V(t) \simeq V_{\text{orc}}(t)$ where

$$V_{\text{orc}}(t) := \sum_{j \in \mathcal{H}_0} \Phi(a_j(\mu_j + z_t + \eta_j)), \tag{1.9}$$

and compute the convergence rate between $\text{FDP}_{\text{orc}}(t) := V_{\text{orc}}(t)/\{R(t) \vee 1\}$ and $\text{FDP}(t)$, which is the same rate shown by Fan and Han (2017) for multiple testing of two-sided hypotheses. We notice that we can set μ_j to 0 under the two-sided null hypothesis $\mathcal{H}_{0j} : \mu_j = 0$ while we still need to estimate μ_j since it can be a negative value under the one-sided null hypothesis $\mathcal{H}_{0j} : \mu_j \le 0$.

Lemma 1 We note that the indicator $I(P_j \le t)$ is approximated by $\Phi(a_j(\mu_j + z_t + \eta_j))$ in Proposition 1. Similarly, one may expect that the indicator $I(\lambda < P_j \le \tau)$ can be approximated by $\Phi(a_j(\mu_j + z_\tau + \eta_j)) - \Phi(a_j(\mu_j + z_\lambda + \eta_j))$. Lemma 1 shows that $V_{\text{orc}}(t) \simeq V_{\text{orc}}^{\text{DA}}(t)$ where

$$V_{\text{orc}}^{\text{DA}}(t) := \sum_{j \in \mathcal{H}_0} \frac{\Phi(a_j(\mu_j + z_t + \eta_j)) I(\lambda < P_j \le \tau)}{\Phi(a_j(\mu_j + z_\tau + \eta_j)) - \Phi(a_j(\mu_j + z_\lambda + \eta_j))}. \tag{1.10}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Theorem 2 In practice, we find that the denominator of $V_{\text{orc}}^{\text{DA}}(t)$ often becomes very small and $V_{\text{orc}}^{\text{DA}}(t)$ has high variability. To address this issue, we round up the denominators in $V_{\text{orc}}^{\text{DA}}(t)$ by ϵ in Theorem 2 and approximate them as $V_{\text{orc}}^{\text{DA}}(t) \simeq V_{\text{orc}}^{\text{DAB}}(t)$ where

$$V_{\text{orc}}^{\text{DAB}}(t) := \sum_{j \in \mathcal{H}_0} \frac{\Phi(a_j(\mu_j + z_t + \eta_j))I(\lambda < P_j \leq \tau)}{\{\Phi(a_j(\mu_j + z_\tau + \eta_j)) - \Phi(a_j(\mu_j + z_\lambda + \eta_j))\} \vee \epsilon}. \tag{1.11}$$

In Theorem 2, we find the asymptotic decay rate of the difference between $V_{\text{orc}}^{\text{DA}}(t)$ and $V_{\text{orc}}^{\text{DAB}}(t)$ with ϵ. The results recommend choosing a small value of ϵ for $V_{\text{orc}}^{\text{DAB}}(t)$.

Lemma 2 $V_{\text{orc}}^{\text{DAB}}(t)$ still can not be used in practice without knowing the true null set $\mathcal{H}_0$ and the mean value μ_js. To address the true null set issue, a easy solution is to use an upper bound of $V_{\text{orc}}^{\text{DAB}}(t)$ by extending the summation over the true null set $\mathcal{H}_0$ to the entire index set. However, it is not straightforward to show that the inequality can still hold after removing the true mean values $\{\mu_j : j \in \mathcal{H}_0\}$ from $V_{\text{orc}}^{\text{DAB}}(t)$. Lemma 2 show that this is the case, and we have an upper bound of $V_{\text{orc}}^{\text{DAB}}(t)$ as $V_{\text{orc}}^{\text{DAB}}(t) \leq V_{\text{U}}^{\text{DAB}}(t)$ where

$$V_{\text{U}}^{\text{DAB}}(t) := \sum_{j=1}^{p} \frac{\Phi(a_j(z_t + \eta_j))I(\lambda < P_j \leq \tau)}{\{\Phi(a_j(z_\tau + \eta_j)) - \Phi(a_j(z_\lambda + \eta_j))\} \vee \epsilon}. \tag{1.12}$$

Theorem 3 To use $V_{\text{U}}^{\text{DAB}}(t)$ in practice, we need to estimate the unknown values a_j and η_j. Following Fan and Han (2017), we estimate these unknown values using the eigenvalue and eigenvector estimators $\{(\widehat{\lambda}_j, \widehat{\boldsymbol{\gamma}}_j) : j = 1, \ldots, p\}$ of Σ. In Theorem 3, we show that the plug-in estimator $\widehat{V}_{\text{U}}^{\text{DAB}}(t)$ converges to realized $V_{\text{U}}^{\text{DAB}}(t)$ and computes the convergence rate in terms of the estimation accuracy of the eigenvalues and eigenvectors. The rate obtained is similar to that for the two-sided multiple hypothesis in Fan and Han (2017). The only difference between our rate and the rate for the two-sided multiple hypothesis is that our rate is inversely proportional to ϵ, which encourages us not to take ϵ to too small a value, which is the opposite of Theorem 2. Thus, there

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

exists a trade-off in choosing a proper ϵ as $V_{\text{orc}}^{\text{DAB}}(t)$ may not approximate $V_{\text{orc}}^{\text{DA}}(t)$ well if it is too large, and if it is too small, a higher accuracy in estimating eigenvalues and eigenvectors to approximate $V_{\text{U}}^{\text{DAB}}(t)$ is required.

Theorem 4 From Weyl's inequality, the estimation accuracy of eigenvalues and eigenvectors can be represented by a correlation matrix. In Theorem 4, we rewrite the convergence rate of $\widehat{V}_{\text{U}}^{\text{DAB}}(t)$ to realized $V_{\text{U}}^{\text{DAB}}(t)$ in terms of the estimation accuracy of the correlation matrix of test statistics under the assumption that the leading eigenvalues are distinct.

In short, we derive an upper bound of the approximation of $V(t)$ and estimate the realized upper bound with $\widehat{V}_{\text{U}}^{\text{DAB}}(t)$.

$$V(t) \approx V_{\text{orc}}(t) \approx V_{\text{orc}}^{\text{DA}}(t) \approx V_{\text{orc}}^{\text{DAB}}(t) \leq V_{\text{U}}^{\text{DAB}}(t)$$

All aforementioned theoretical results for the approximation of $V(t)$ can be applied to the approximation of $\text{FDP}(t) = V(t)/(R(t) \vee 1)$ and we propose the following estimator to approximate realized $\text{FDP}(t)$ to control FDR.

$$\widehat{\text{FDP}}_{\text{U}}^{\text{DAB}}(t) = \widehat{V}_{\text{U}}^{\text{DAB}}(t)/\{R(t) \vee 1\}.$$

The remainder of the paper is organized as follows. In Section 2, we show the theoretical properties of the proposed method. In Section 3, we numerically investigate the theoretical results, provide performance comparisons of the proposed method with other methods, and present sensitivity analysis regarding the choice of thresholding parameters. In Section 4, we demonstrate the methodology via an application to the protein phosphorylation analysis of ovarian serous adenocarcinoma samples to identify protein modification levels that are uniquely elevated in each of the five molecular subtypes. In Section 5, we conclude with a discussion.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

2. Approximation of false discovery proportion

2.1 Principal factor approximation under known dependence

Suppose that we wish to test p one-sided hypotheses

$$\mathcal{H}_{0j} : \mu_j \leq 0 \quad \text{vs.} \quad \mathcal{H}_{1j} : \mu_j > 0 \tag{2.1}$$

for $j = 1, 2, \ldots, p$. We have a test statistic Z_j with mean μ_j for each hypothesis. We assume that the p-dimensional vector $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$ of test statistics follows a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ and $\mathrm{Cov}(\mathbf{Z}) = \boldsymbol{\Sigma}$. Since we assume that the test statistics are normalized, $\mathrm{Var}(Z_j) = 1$ for $j = 1, \ldots, n$ and $\boldsymbol{\Sigma}$ is the correlation matrix.

Let $\mathcal{H}_0 = \{j : \mu_j \leq 0\}$ be the set of indices of true nulls. For each $j \in \{1, 2, \ldots, p\}$, the j-th test statistic Z_j marginally follows the standard normal distribution $N(0, 1)$, so the p-value for the j-th hypothesis is calculated as $P_j = \Phi(-Z_j)$. We let $t \in (0, 1)$ be a common threshold value for the multiple tests. That is, we reject the j-th null hypothesis $\mathcal{H}_{0j}$ if the corresponding p-value does not exceed the threshold value t. Define $R(t) = \#\{j : P_j \leq t\}$ as the number of discoveries, and $V(t) = \#\{j \in \mathcal{H}_0 : P_j \leq t\}$ as the number of false discoveries. We are interested in controlling the false discovery rate (FDR) under a predetermined level $\alpha \in (0, 1)$. To do this, we need to approximate the realized false discovery proportion $\mathrm{FDP}(t) = V(t)/\{R(t) \vee 1\}$. We note that $R(t)$ is observable, while $V(t)$ is a realized but unobservable random variable. The number of falsely rejected hypotheses $V(t)$ can be expressed as

$$V(t) = V_{\mathrm{os}}(t) = \sum_{j \in \mathcal{H}_0} I(P_j \leq t) = \sum_{j \in \mathcal{H}_0} I(Z_j \geq -z_t). \tag{2.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Here, the subscript "os" stands for "one-sided". It is worth noting that this representation of $V(t)$ is different from that of multiple two-sided tests. When testing two-sided hypotheses, p-values are calculated by $P_j = 2\Phi(-|Z_j|)$ $(j = 1, \ldots, p)$, and thus the number of false rejections $V_{\text{ts}}(t)$ is represented as

$$V_{\text{ts}}(t) = \sum_{j \in \mathcal{H}_0} I(P_j \le t) = \sum_{j \in \mathcal{H}_0} \left\{ I(Z_j \le z_{t/2}) + I(Z_j \ge -z_{t/2}) \right\}, \tag{2.3}$$

in which $\mathcal{H}_0 = \{j : \mu_j = 0\}$ and the subscript "ts" stands for "two-sided".

Based on the principal factor approximation (PFA) method, Fan et al. (2012) proposed approximating $V_{\text{ts}}(t)$ by

$$V_{\text{ts,orc}}(t) = \sum_{j \in \mathcal{H}_0} \left\{ \Phi(a_j(z_{t/2} + \eta_j)) + \Phi(a_j(z_{t/2} - \eta_j)) \right\}. \tag{2.4}$$

where a_j and η_j are defined in (1.5).

The PFA method can be applied to the multiple one-sided tests scheme. However, there is a critical difference between one-sided tests and two-sided tests. The mean value μ_j is fixed at zero for the two-sided null hypothesis $\mathcal{H}_{0j} : \mu_j = 0$ (*i.e.* $\mathcal{H}_{0j}$ is simple), while the mean μ_j is not determined in the one-sided null hypothesis $\mathcal{H}_{0j} : \mu_j \le 0$ (*i.e.* $\mathcal{H}_{0j}$ is complex). To reflect this difference, we show that the approximation of $V_{\text{os}}(t)$ in (2.2) is given as

$$V_{\text{os,orc}}(t) = V_{\text{orc}}(t) = \sum_{j \in \mathcal{H}_0} \Phi(a_j(\mu_j + z_t + \eta_j)). \tag{2.5}$$

Compared to the approximation (2.4), the true mean values $\{\mu_j : j \in \mathcal{H}_0\}$ remain in the approximation (2.5). As in Proposition 1 of Fan and Han (2017), we obtain the same convergence rate result for $\text{FDP}_{\text{orc}}(t) = V_{\text{orc}}(t)/(R(t) \vee 1)$ under the weak dependence assumption on $\mathbf{\Sigma}_u$ in (1.3).

Proposition 1. *If condition (C1) is satisfied, we have*

$$|\text{FDP}_{\text{orc}}(t) - \text{FDP}(t)| = O_p(p^{-(\delta/2-\theta)}),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

on the event $\mathcal{E}_0 = \{p^{-1}R(t) > cp^{-\theta}\}$ *for some* $c > 0$ *and* $\theta \geq 0$.

Sketch of proof. Similar to the proof of Theorem 1 by Fan et al. (2012), we can show that

$$p_0^{-1}|V_{\text{orc}}(t) - V(t)| = O_p(p^{-\delta/2}),$$

where $p_0 = \#\{j : \mu_j \leq 0\}$ is the number of true nulls. Hence, the desired result holds for the event $\mathcal{E}_0 = \{p^{-1}R(t) > cp^{-\theta}\}$. □

2.2 Discarding adaptively with bounding

In the previous section, we showed that $\text{FDP}_{\text{orc}}(t)$ approximates $\text{FDP}(t)$. In practice, however, we cannot observe $\text{FDP}_{\text{orc}}(t)$ directly for the following three reasons: we have no information about (i) the true null set $\mathcal{H}_0$, (ii) the true mean values $\{\mu_j : j \in \mathcal{H}_0\}$, and (iii) the other unknown (or unobserved) values $\{(a_j, \eta_j) : j = 1, \ldots, p\}$, which are functions of the eigenvalues and eigenvectors of $\mathbf{\Sigma}$.

To address the first problem, Fan et al. (2012) suggested using

$$V_{\text{U}}(t) = \sum_{j=1}^{p} \{\Phi(a_j(z_{t/2} + \eta_j)) + \Phi(a_j(z_{t/2} - \eta_j))\} \tag{2.6}$$

as a conservative surrogate for $V_{\text{ts,orc}}(t)$ in (2.4). Since they assumed that the mean vector $\boldsymbol{\mu}$ is sparse, the extra terms in (2.6) are negligible. However, in a multiple one-sided test scheme without the negligible non-null assumption, the additional terms are nonignorable. We note that the original method dropped the indicator terms $I(j \in \mathcal{H}_0)$. Instead, we propose using alternative indicator terms $I(\lambda < P_j \leq \tau)$ for some fixed values $\lambda, \tau \in (0, 1)$. This indicator has two purposes. One purpose is to adaptively estimate the proportion of true nulls, which was originally proposed by Storey (2002) in an offline FDR control setting and later utilized by Ramdas et al. (2018) in an online FDR control setting. The other purpose is to discard

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

obvious conservative nulls, which was suggested by Tian and Ramdas (Tian and Ramdas) to enhance the power in an online FDR control setting under conservative nulls. We note that the candidate threshold λ adaptively chooses whether $\mathcal{H}_{0j}$ is a candidate for rejection, and the discarding parameter τ determines whether $\mathcal{H}_{0j}$ is selected for testing. In Tian and Ramdas (Tian and Ramdas), the indicator terms $I(\lambda < P_j \leq \tau)$ were divided by $(\tau - \lambda)$ to make the estimator unbiased. Likewise, we also need to multiply the indicator terms by a proper weight to make our estimator approximate the realized FDP. The modified version, which we call the discarding adaptively with principal factor approximation (DA-PFA), is given as follows:

$$V_{\text{orc}}^{\text{DA}}(t;\lambda,\tau) = \sum_{j\in\mathcal{H}_0} \frac{\Phi(a_j(\mu_j + z_t + \eta_j))I(\lambda < P_j \leq \tau)}{\Phi(a_j(\mu_j + z_\tau + \eta_j)) - \Phi(a_j(\mu_j + z_\lambda + \eta_j))}. \tag{2.7}$$

Practically, the denominator terms in (2.7) might be unstable, so we put a fixed positive number ϵ as a lower bound of the denominator. Finally, the discarding adaptively with bounding on the principal factor approximation (DAB-PFA) is given as

$$V_{\text{orc}}^{\text{DAB}}(t;\lambda,\tau,\epsilon) = \sum_{j\in\mathcal{H}_0} \frac{\Phi(a_j(\mu_j + z_t + \eta_j))I(\lambda < P_j \leq \tau)}{\{\Phi(a_j(\mu_j + z_\tau + \eta_j)) - \Phi(a_j(\mu_j + z_\lambda + \eta_j))\} \vee \epsilon}. \tag{2.8}$$

The following lemma shows the convergence rate of $V_{\text{orc}}^{\text{DAB}}(t)$ to $V_{\text{orc}}(t)$.

Lemma 1. *For fixed constants $\lambda, \tau \in (0,1)$ with $\lambda < \tau$, define*

$$G(t;\lambda,\tau) = \sum_{j\in\mathcal{H}_0} g_j(\mathbf{W})I(\lambda < P_j \leq \tau)$$

where $g_j : \mathbb{R}^K \to \mathbb{R}$ is a continuous function for all $j \in \mathcal{H}_0$. Assume that condition (C1) is satisfied. Then,

$$p_0^{-1}\left|G(t;\lambda,\tau) - \sum_{j\in\mathcal{H}_0} g_j(\mathbf{W})\left\{\Phi(a_j(\mu_j + z_\tau + \eta_j)) - \Phi(a_j(\mu_j + z_\lambda + \eta_j))\right\}\right| = O_p(p^{-\delta/2}).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Let

$$g_j(\mathbf{w}) = \frac{\Phi(a_j(\mu_j + z_t + \mathbf{b}_j^\top \mathbf{w}))}{\Phi(a_j(\mu_j + z_\tau + \mathbf{b}_j^\top \mathbf{w})) - \Phi(a_j(\mu_j + z_\lambda + \mathbf{b}_j^\top \mathbf{w}))}.$$

Then, the above lemma leads to the convergence rate of $V_{\mathrm{orc}}^{\mathrm{DA}}(t)$ to $V_{\mathrm{orc}}(t)$.

In the following theorem, we show the convergence rate of $\mathrm{FDP}_{\mathrm{orc}}^{\mathrm{DA}}(t;\lambda,\tau)$, which is a generalized version of Proposition 1.

Theorem 1. *Assume that condition (C1) holds. Then, for fixed values $\lambda, \tau \in (0,1)$ with $\lambda < \tau$, it holds that*

$$p_0^{-1}|V_{\mathrm{orc}}^{\mathrm{DA}}(t;\lambda,\tau) - V_{\mathrm{orc}}(t)| = O_p(p^{-\delta/2}),$$

and thus,

$$p_0^{-1}|V_{\mathrm{orc}}^{\mathrm{DA}}(t;\lambda,\tau) - V(t)| = O_p(p^{-\delta/2}).$$

Furthermore, on the event $\mathcal{E}_0 = \{p^{-1}R(t) > cp^{-\theta}\}$ for some constants $c > 0$ and $\theta \geq 0$, it holds that

$$|\mathrm{FDP}_{\mathrm{orc}}^{\mathrm{DA}}(t;\lambda,\tau) - \mathrm{FDP}(t)| = O_p(p^{-(\delta/2-\theta)}).$$

The next theorem shows that the DAB-PFA estimator $V_{\mathrm{orc}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon)$ is close to the DA-PFA estimator $V_{\mathrm{orc}}^{\mathrm{DA}}(t;\lambda,\tau)$.

Theorem 2. *Assume that condition (C1) holds,*

(C2) $a_j \leq C_a$ $\forall j = 1, 2, \ldots, p$ for some finite constant $C_a > 1$, and

(C3) $\epsilon = O(p^{-\alpha})$ for a positive constant α.

Then,

$$p_0^{-1}|V_{\mathrm{orc}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon) - V_{\mathrm{orc}}^{\mathrm{DA}}(t;\lambda,\tau)| = O_p(p^{-\alpha\beta}) \tag{2.9}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

for a positive constant $\beta < 1$*, and therefore, we have*

$$p_0^{-1}|V_{\text{orc}}^{\text{DAB}}(t;\lambda,\tau,\epsilon) - V(t)| = O_p(p^{-\delta/2} + p^{-\alpha\beta}).$$

Furthermore, on the event $\mathcal{E}_0 = \{p^{-1}R(t) > cp^{-\theta}\}$ *for some* $c > 0$ *and* $\theta \geq 0$*, it holds that*

$$|\text{FDP}_{\text{orc}}^{\text{DAB}}(t;\lambda,\tau,\epsilon) - \text{FDP}(t)| = O_p(p^{\theta}(p^{-\delta/2} + p^{-\alpha\beta})). \tag{2.10}$$

Remark 1. In the proof of Theorem 2, we show that the equation (2.9) holds for a positive constant $\beta = C_3/C_a^2 < 1$. Here, $C_3 \in (0,1)$ does not depend on any parameters. Hence, the magnitude of β is only affected by C_a, which is an upper bound of $\{a_j : j = 1,\ldots,p\}$. We note that a smaller value of C_a makes condition (C2) stronger, resulting in a guaranteed faster convergence rate. Recall the definition of a_js: $a_j = (1-\|\mathbf{b}_j\|^2)^{-1/2}$ for each $j = 1,\ldots,p$ where $\mathbf{b}_j^\top$ is the j-th row of $\mathbf{B} = (\sqrt{\lambda_1}\boldsymbol{\gamma}_1,\ldots,\sqrt{\lambda_K}\boldsymbol{\gamma}_K) \in \mathbb{R}^{p\times K}$. Since we assume that $\boldsymbol{\Sigma}$ is a correlation matrix, we have $\sum_{k=1}^p \lambda_k\gamma_{jk}^2 = 1$. By definition, $\|\mathbf{b}_j\|^2 = \sum_{k=1}^K \lambda_k\gamma_{jk}^2$. We can expect a small value of C_a when the ratio of $\sum_{k=1}^K \lambda_k\gamma_{jk}^2$ to $\sum_{k=1}^p \lambda_k\gamma_{jk}^2$ is small. It is related to the choice of K, which is the number of factors. If we take a larger value of K, we can obtain an advantage in the convergence rate through a larger value of δ. However, a larger value of K makes a_js larger, resulting in a slower convergence rate. Thus, there is a kind of trade-off between the first and second terms in the convergence rate (2.10).

There are a few issues to use $\text{FDP}_{\text{orc}}^{\text{DAB}}(t)$ in practice. First, we have no information for true null set $\mathcal{H}_0$. To address this issue, We replace the indicator $I(j \in \mathcal{H}_0)$ with $I(\lambda < P_j \leq \tau)$. We note that changing the summation over the true null set to the summation over the entire index set does not affect the overall summed value since the set of indices of non-nulls has a negligible intersection with $\{j : \lambda < P_j \leq \tau\}$. *i.e.*

$$I(j \in \mathcal{H}_0, \lambda < P_j \leq \tau) \simeq I(\lambda < P_j \leq \tau).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Another issue is due to the unknown true mean μ_js. One naive approach to solve this issue is to replace μ_j with any plug estimator $\widehat{\mu}_j$. However, this approach has a structural problem. Instead, we suggest using the discarding parameter τ. Using τ enforces the discarding of test statistics less than $-z_\tau$. For example, hypotheses with negative Z values are discarded when $\tau = 0.5$. Based on the positive value of the test statistic, a reasonable estimate of the true mean value μ_j under the null hypothesis $\mathcal{H}_{0j} : \mu_j \leq 0$ is zero. It is equivalent to ignoring μ_j terms in (2.8). Now, we consider the following upper bound.

$$V_{\mathrm{U}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon) = \sum_{j=1}^{p} \frac{\Phi(a_j(z_t+\eta_j))I(\lambda < P_j \leq \tau)}{\{\Phi(a_j(z_\tau+\eta_j)) - \Phi(a_j(z_\lambda+\eta_j))\} \vee \epsilon}. \tag{2.11}$$

Here, the subscript "U" stands for "upper". We note that it is not trivial that $V_{\mathrm{U}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon)$ is always larger than $V_{\mathrm{orc}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon)$. To show the increment by removing μ_j, we introduce the following lemma.

Lemma 2. *For any $x \leq 0$ and $A > B > C$, we have*

$$\frac{\Phi(x+C)}{\Phi(x+A) - \Phi(x+B)} \leq \frac{\Phi(C)}{\Phi(A) - \Phi(B)}.$$

Typically, we take a candidate threshold λ larger than threshold t. This implies that if we let $x = a_j\mu_j$, $A = a_j(z_\tau + \eta_j)$, $B = a_j(z_\lambda + \eta_j)$, and $C = a_j(z_t + \eta_j)$, then $x \leq 0$ and $A > B > C$. By Lemma 2,

$$\frac{\Phi(a_j(\mu_j + z_t + \eta_j))}{\Phi(a_j(\mu_j + z_\tau + \eta_j)) - \Phi(a_j(\mu_j + z_\lambda + \eta_j))} \leq \frac{\Phi(a_j(z_t+\eta_j))}{\Phi(a_j(z_\tau+\eta_j)) - \Phi(a_j(z_\lambda+\eta_j))}$$

for all $j \in \mathcal{H}_0$, and thus $V_{\mathrm{U}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon)$ is larger than $V_{\mathrm{orc}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon)$. We consider $\mathrm{FDP}_{\mathrm{U}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon) = V_{\mathrm{U}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon)/\{R(t) \vee 1\}$ as a conservative surrogate.

Finally, it should be noted that we still cannot directly use $\mathrm{FDP}_{\mathrm{U}}^{\mathrm{DAB}}(t)$ due to unknown quantities $\{(a_j, \eta_j) : j = 1, \ldots, p\}$, which are functions of the eigenvalues and eigenvectors

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

of the unknown covariance matrix $\mathbf{\Sigma}$. As in Fan and Han (2017), we plug-in the eigenvalue and eigenvector estimators. The next section discusses the accuracy of the eigenvalue and eigenvector estimation.

2.3 Estimation accuracy of unknown dependence structure

Practically, the dependence structure is unknown; hence, it is necessary to estimate $\mathbf{\Sigma}$ to obtain the proposed FDP approximation. To obtain an optimal convergence rate in the approximation of the realized FDP, we need some requirements on $\widehat{\mathbf{\Sigma}}$. In this section, we discuss the relationship between the accuracy of $\widehat{\mathbf{\Sigma}}$ and the convergence rate of the proposed FDP approximation.

Let $\widehat{\mathbf{\Sigma}}$ be an estimator of $\mathbf{\Sigma}$, and let $\{(\widehat{\lambda}_j, \widehat{\boldsymbol{\gamma}}_j) : j = 1, \ldots, p\}$ be the eigenvalues and the eigenvectors of $\widehat{\mathbf{\Sigma}}$, respectively. Analogous to the decomposition (1.3), $\widehat{\mathbf{\Sigma}}$ is decomposed by

$$\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top + \widehat{\mathbf{\Sigma}}_u$$

where $\widehat{\mathbf{B}} = (\widehat{\lambda}_1^{1/2}\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\lambda}_K^{1/2}\widehat{\boldsymbol{\gamma}}_K) \in \mathbb{R}^{p\times K}$ and $\widehat{\mathbf{\Sigma}}_u = \sum_{j=K+1}^{p} \widehat{\lambda}_j \widehat{\boldsymbol{\gamma}}_j \widehat{\boldsymbol{\gamma}}_j^\top$. Here, K denotes the number of factors satisfying condition (C1). Commonly, K is unknown and should be estimated. Following Fan and Han (2017), we apply the eigenvalue ratio (ER) estimator proposed by Ahn and Horenstein (2013). The ER estimator is defined as

$$\widehat{K} = \underset{1\le K\le K_{\max}}{\operatorname{argmax}} \frac{\tilde{\lambda}_K}{\tilde{\lambda}_{K+1}},$$

where $\tilde{\lambda}_j$ is the j-th largest eigenvalue of the sample correlation matrix and $K_{\max}$ is a predetermined maximum possible number of factors. Other methods for estimating the number of factors can also be used. See Dobriban (2020) and the references therein. In the rest of this work, we assume that the number of factors K is known.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Based on the correlation matrix estimate, we need to estimate the realized common factors $\mathbf{W}$. Practically, we use the least square method. That is, we use the least squares estimate $\widehat{\mathbf{W}} = (\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}})^{-1} \widehat{\mathbf{B}}^\top \mathbf{Z}$. This also simplifies the technical arguments.

Define the DAB-PFA estimator as

$$\widehat{\mathrm{FDP}}^{\mathrm{DAB}}(t; \lambda, \tau, \epsilon) := \left[\sum_{j=1}^{p} \frac{\Phi(\widehat{a}_j(z_t + \widehat{\eta}_j)) I(\lambda < P_j \le \tau)}{\{\Phi(\widehat{a}_j(z_\tau + \widehat{\eta}_j)) - \Phi(\widehat{a}_j(z_\lambda + \widehat{\eta}_j))\} \vee \epsilon} \right] \Big/ \{R(t) \vee 1\}, \quad (2.12)$$

where $\widehat{a}_j = (1 - \|\widehat{\mathbf{b}}_j\|_2^2)^{-1/2}$, $\widehat{\eta}_j = \widehat{\mathbf{b}}_j^\top \widehat{\mathbf{W}}$, and $\widehat{\mathbf{b}}_j^\top$ is the j-th row of $\widehat{\mathbf{B}}$. The following theorem shows how the accuracy of the eigenvalue and eigenvector estimators affects the proposed FDP estimator.

Theorem 3. *On the event $\mathcal{E}$ that*

(C2)′ $a_j \le C_a$ and $\widehat{a}_j \le C_a$ $\forall j = 1, 2, \ldots, p$ for a finite constant $C_a > 1$,

(C3) $\epsilon = O(p^{-\alpha})$ for a positive constant α,

(C4) $\max\{|\eta_j|, |\widehat{\eta}_j|, |\tilde{\eta}_j| : \lambda < P_j \le \tau\} \le C_\eta$ for a finite constant $C_\eta > 0$ where $\tilde{\eta}_j = \mathbf{b}_j^\top \widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{W}} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{Z}$,

(C5) $\max_{1 \le k \le K} \|\widehat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k\| = O_p(p^{-\nu_1})$ for $\nu_1 > 0$,

(C6) $\sum_{k=1}^{K} |\widehat{\lambda}_k - \lambda_k| = O_p(p^{1-\nu_2})$ for $\nu_2 > 0$,

(C7) $p^{-1} R(t) > c p^{-\theta}$,

we have

$$|\widehat{\mathrm{FDP}}^{\mathrm{DAB}}(t) - \mathrm{FDP}_{\mathrm{U}}^{\mathrm{DAB}}(t)| = O_p\left\{ p^{\alpha+\theta} (K p^{-\nu_1} + p^{-\nu_2} + p^{-1/2} \|\boldsymbol{\mu}\|) \right\} \quad (2.13)$$

where $\mathrm{FDP}_{\mathrm{U}}^{\mathrm{DAB}}(t; \lambda, \tau, \epsilon) = V_{\mathrm{U}}^{\mathrm{DAB}}(t; \lambda, \tau, \epsilon) / \{R(t) \vee 1\}$ with $V_{\mathrm{U}}^{\mathrm{DAB}}(t; \lambda, \tau, \epsilon)$ defined in (2.11).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Remark 2. As opposed to the effect of ϵ on the convergence rate results in Theorem 2, using a small ϵ is not beneficial to the overall convergence rate in the above theorem. According to the convergence rate in (2.13), if the eigenvalues and eigenvectors of the correlation matrix are not precisely estimated, which is equivalent to small values of ν_1 and ν_2, then we have to use $\epsilon = O(p^{-\alpha})$ with small α. In other words, poor estimation of eigenvalues and eigenvectors may result in large values of $\widehat{a}_j$s, which make the denominator in (2.12) very small. Therefore, we need a bounding parameter $\epsilon = O(p^{-\alpha})$ with a small α to prevent the denominator from being too large.

Remark 3. In practice, the normality assumption is often violated for various reasons. Fan and Han (2017) addressed this issue by providing both theoretical and numerical analyses of cases where the normality assumption is violated. They derived theoretical results for test statistics that are heavy-tailed due to the estimation of marginal variances and reported numerical simulation results with data drawn from a t-distribution. A similar theoretical extension can be applied to our problem. Assume that the marginal variances σ_j^2 of the statistics Z_j are unknown, but we have estimates $\widehat{\sigma}_j^2$, where each estimate $\widehat{\sigma}_j^2$ is independent of Z_j and follows a χ^2 distribution with degrees of freedom d. In this case, the standardized test statistics $T_j = Z_j/\widehat{\sigma}_j$ follow a t distribution with degrees of freedom d. The p-values are calculated as $P_{\mathrm{T},j} = F_d(-T_j)$, where F_d denotes the cumulative distribution function of a t_d random variable, and the subscript 'T' indicates the use of t_d test statistics. In this setting, we define the DAB-PFA estimator as:

$$\widehat{\mathrm{FDP}}_{\mathrm{T}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon) := \left[\sum_{j=1}^{p} \frac{\Phi(\widehat{a}_j(z_t+\widehat{\eta}_{\mathrm{T},j}))I(\lambda < P_{\mathrm{T},j} \le \tau)}{\{\Phi(\widehat{a}_j(z_\tau+\widehat{\eta}_{\mathrm{T},j})) - \Phi(\widehat{a}_j(z_\lambda+\widehat{\eta}_{\mathrm{T},j}))\} \vee \epsilon}\right] \Big/ \{R(t) \vee 1\}, \quad (2.14)$$

where $\widehat{\eta}_{\mathrm{T},j} = \widehat{\mathbf{b}}_j^\top \widehat{\mathbf{W}}_{\mathrm{T}}$ and $\widehat{\mathbf{W}}_{\mathrm{T}} = (\widehat{\mathbf{B}}^\top\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^\top\mathbf{T}$. One might follow the theoretical approach of Fan and Han (2017) to analyze the asymptotic properties of $\widehat{\mathrm{FDP}}_{G}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon)$. However,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

we leave this problem for future research. Additionally, we evaluated the performance of the proposed method when the normality condition is violated. For more details, see Model 3 of the simulation studies in Section 3.

Remark 4. Compared to previous works (Fan et al., 2012; Fan and Han, 2017), our theoretical results do not explicitly require sparsity of the non-nulls. However, the term $\|\boldsymbol{\mu}\|$ in equation (2.13) can be interpreted as a regulation on the signal strength of both the non-nulls and true nulls. For simplicity, assume that each non-null mean value μ_j is equal to $\mu_{\mathrm{A}} > 0$ and each true null mean value μ_j is equal to $\mu_{\mathrm{N}} \leq 0$. Then, $\|\boldsymbol{\mu}\| = O(p_1^{1/2}|\mu_{\mathrm{A}}| + p_0^{1/2}|\mu_{\mathrm{N}}|)$. Theorem 3 implies that a sufficient condition for the right-hand side of equation (2.13) to converge to zero is $\|\boldsymbol{\mu}\| = O(p^{1/2-\alpha-\theta})$. Now, we consider two scenarios: strong and sparse signals, and weak and dense signals. In the first scenario, we assume $p_1 = O(p^r)$ for some $0 < r < 1$, implying $p_0 = O(p)$. If $|\mu_{\mathrm{A}}| = O(p^{(1-r)/2-\alpha-\theta})$ and $|\mu_{\mathrm{N}}| = O(p^{-\alpha-\theta})$, we have $\|\boldsymbol{\mu}\| = O(p^{1/2-\alpha-\theta})$. In the second scenario, we assume $p_1 = O(p)$ and $p_0 = O(p)$. If $|\mu_{\mathrm{A}}| = O(p^{-\alpha-\theta})$ and $|\mu_{\mathrm{N}}| = O(p^{-\alpha-\theta})$, we again obtain $\|\boldsymbol{\mu}\| = O(p^{1/2-\alpha-\theta})$. From this observation, we note that the proposed method can accommodate a broader range of scenarios compared to previous works, which only cover the strong and sparse case.

Following Fan and Han (2017), we now study sufficient conditions under which (C5) and (C6) are satisfied. The following lemma is a restatement of Lemma 1 of Fan and Han (2017).

Lemma 3. *For any matrices* $\boldsymbol{\Sigma}$ *and* $\widehat{\boldsymbol{\Sigma}}$*, we have* $|\widehat{\lambda}_j - \lambda_j| \leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$ *and*

$$\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\| \leq \frac{\sqrt{2}\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|}{\min(|\widehat{\lambda}_{j-1} - \lambda_j|, |\lambda_j - \widehat{\lambda}_{j+1}|)},$$

where $\|\mathbf{M}\| = \max_{1\leq j\leq p} \lambda_j(\mathbf{M})$ *is the operator norm for a* $p \times p$ *positive-definite matrix* $\mathbf{M}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Table 1: Some literature on covariance matrix estimation and their convergence rate results, where α is the rate of decay of a banded matrix, q is a constant between 0 and 1 that determines the class of sparse covariance matrices, $c_0(p)$ and $s_0(p)$ are sparsity parameters of a sparse covariance matrix.

Paper	Estimator	Convergence rate
Bickel and Levina (2008b)	$\widehat{\mathbf{\Sigma}}^{\mathrm{BLa}}$	$\|\widehat{\mathbf{\Sigma}}^{\mathrm{BLa}} - \mathbf{\Sigma}\| = O_p\left(\left(\frac{\log p}{n}\right)^{\alpha/(2\alpha+2)}\right)$
Bickel and Levina (2008a)	$\widehat{\mathbf{\Sigma}}^{\mathrm{BLb}}$	$\|\widehat{\mathbf{\Sigma}}^{\mathrm{BLb}} - \mathbf{\Sigma}\| = O_p\left(c_0(p)\left(\frac{\log p}{n}\right)^{(1-q)/2}\right)$
Cai and Liu (2011)	$\widehat{\mathbf{\Sigma}}^{\mathrm{CL}}$	$\|\widehat{\mathbf{\Sigma}}^{\mathrm{CL}} - \mathbf{\Sigma}\| = O_p\left(s_0(p)\left(\frac{\log p}{n}\right)^{(1-q)/2}\right)$

The above lemma implies that the errors in terms of eigenvalues and eigenvectors are directly bounded by the operator norm error of the correlation estimator.

Theorem 4. *If* $\min_{1\le k\le K}(\lambda_k - \lambda_{k+1}) \ge d_p$ *for a sequence* $\{d_p\}$ *of positive numbers, then on the event* $\mathcal{E} \cap \{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O_p(d_p p^{-\nu})\}$ *with a constant* $\nu > 0$*, we have*

$$\begin{aligned} &|\widehat{\mathrm{FDP}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon) - \mathrm{FDP}_{\mathrm{U}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon)| \\ &\quad = O_p\left\{p^{\alpha+\theta}\left(K(d_p/p+1)p^{-\nu} + p^{-1/2}\|\boldsymbol{\mu}\|\right)\right\} \end{aligned} \tag{2.15}$$

where the event $\mathcal{E}$ *is defined in Theorem 3.*

Many studies on estimating the structured covariance matrix have been proposed over the last two decades. We list some representative works in Table 1. As noted in Fan and Han (2017), if we combine the convergence rates proved by these papers with some assumptions on the relation between the sample size n and the number of variables p, we can obtain the condition $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O_p(d_p p^{-\nu})$ in Theorem 4.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

3. Simulation studies

In this section, we numerically investigate the performance of the proposed FDP approximation method. We mainly perform two simulation studies. One simulation study checks the validity of the DAB-PFA by comparing the finite sample behavior of the true FDP and the proposed FDP estimator for a fixed threshold level $t = 0.01$. This threshold level is chosen because it yields an average FDR close to 0.1 in our simulation setting. The other simulation study checks the performance of the proposed FDR control method with comparisons to other multiple testing procedures. For each method, we choose a threshold level $\widehat{t} \in (0, 1)$ so that the corresponding FDP estimate is controlled under the predetermined level $\alpha = 0.1$. In addition, we conduct a sensitivity analysis of the choice of the threshold parameters $(\lambda, \tau, \epsilon)$.

In the simulation studies, we consider the following scenarios: sample size $n = 100$, dimensionality $p = 1,000$, the number of false null hypotheses $p_1 = p - p_0 \in \{100, 300, 500\}$, the threshold parameters $(\lambda, \tau, \epsilon) = (0.1, 0.5, 0.01)$, and the number of simulation rounds $R = 200$. We note that Tian and Ramdas (Tian and Ramdas) used $(\lambda, \tau) = (0.25, 0.5)$ as their default choice. In our studies, we adopt the same $\tau = 0.5$ but choose a smaller $\lambda = 0.1$, enabling the use of more samples when estimating the upper bound $\widehat{\text{FDP}}^{\text{DAB}}$. We randomly set p_1 elements of mean vector $\boldsymbol{\mu}$ as $\mu_{\text{A}} = 3$ and the other p_0 elements as $\mu_{\text{N}} \in \{0, -0.1, -0.2\}$. Unlike Tian and Ramdas (Tian and Ramdas), who considered $\mu_{\text{A}} = 3$ and $\mu_{\text{N}} \in \{0, -0.5, -1, -1.5\}$, we use smaller μ_{N} values, as larger negative values make the testing problem easier, making it harder to compare the performance of the considered methods. We consider six data generation models, referring to the covariance structures used in Fan and Han (2017). We generate the simulation data as follows. First, we generate a covariance matrix $\boldsymbol{\Sigma}_0$ according to each model. Since we assume that $\boldsymbol{\Sigma}$ is a correlation

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

matrix, we consider $\mathbf{\Sigma} = \mathbf{D}^{-1}\mathbf{\Sigma}_0\mathbf{D}^{-1}$, where $\mathbf{D} = \mathrm{diag}(\sigma_{0,jj}^{1/2} : j = 1, \ldots, p)$. We decompose the correlation matrix $\mathbf{\Sigma}$ as $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\top + \mathbf{M}\mathbf{M}^\top$, where $\mathbf{B} = (\lambda_1^{1/2}\boldsymbol{\gamma}_1, \ldots, \lambda_K^{1/2}\boldsymbol{\gamma}_K) \in \mathbb{R}^{p\times K}$ and $\mathbf{M} = (\lambda_{K+1}^{1/2}\boldsymbol{\gamma}_{K+1}, \ldots, \lambda_p^{1/2}\boldsymbol{\gamma}_p) \in \mathbb{R}^{p\times(p-K)}$. Then, we independently generate K-dimensional random vectors $\{\mathbf{w}_i\}_{i=1}^n$ from $N_K(\mathbf{0}, \mathbf{I}_K)$ and $(p-K)$-dimensional random vectors $\{\mathbf{v}_i\}_{i=1}^n$ from $N_{p-K}(\mathbf{0}, \mathbf{I}_{p-K})$, except in Model 3. In Model 3, we generate each element of $\mathbf{w}_i$ and $\mathbf{v}_i$ independently from $\sqrt{2/3}\cdot t_6$, where t_6 is the t distribution with degrees of freedom 6. We note that the covariance matrices of $\mathbf{w}_i$ and $\mathbf{v}_i$ are $\mathbf{I}_K$ and $\mathbf{I}_{p-K}$, respectively, as in the other models. Finally, we obtain the p-dimensional random vector $\mathbf{X}_i = \tilde{\boldsymbol{\mu}} + \mathbf{D}(\mathbf{B}\mathbf{w}_i + \mathbf{M}\mathbf{v}_i)$ where $\tilde{\boldsymbol{\mu}} = n^{-1/2}\boldsymbol{\mu}$. Based on this data, we calculate the test statistics $Z_j = n^{-1/2}\sum_{i=1}^n X_{ij}$ $(j = 1, \ldots, p)$. It is easy to see that $\mathbb{E}(\mathbf{Z}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{Z}) = \mathbf{\Sigma}_0$. In the following section, we describe each of the six dependent structures in detail.

3.1 Dependence structures

- [**Model 1: Strict factor model**] We consider a factor model with three factors,

$$\mathbf{X}_i = \tilde{\boldsymbol{\mu}} + \mathbf{L}\mathbf{f}_i + \boldsymbol{\epsilon}_i$$

  where $\mathbf{f}_i \sim N_3(\mathbf{0}, \mathbf{I}_3)$ and $\boldsymbol{\epsilon}_i \sim N_p(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$ are independent. Entries of the factor loading matrix $\mathbf{L}$ are independently generated from the uniform distribution $U(-1, 1)$. The covariance matrix $\mathbf{\Sigma}_\epsilon$ of error vectors is set as $\mathbf{I}_p$. Note that the covariance matrix of $\mathbf{X}_i$ is $\mathrm{Cov}(\mathbf{X}_i) = \mathbf{L}\mathbf{L}^\top + \mathbf{I}_p$. We use $\mathbf{\Sigma}_0 = \mathbf{L}\mathbf{L}^\top + \mathbf{I}_p$ with $K = 3$.

- [**Model 2: Approximate factor model**] We consider a factor model with three factors, as in Model 1. The difference between Model 1 and Model 2 is the covariance matrix $\mathbf{\Sigma}_\epsilon$ of error vectors. Unlike Model 1, the covariance matrix used in Model 2 is the same as the one used in the numerical study by Fan and Han (2017). In other

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

words, in Model 2, we set $\mathbf{\Sigma}_\epsilon$ as the nearest positive definite matrix of $0.5(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)$, where $\mathbf{\Sigma}_1$ is a symmetric sparse matrix and $\mathbf{\Sigma}_2$ is a symmetric banded matrix.

- [**Model 3: Non-normal model**] We consider a covariance matrix $\mathbf{\Sigma}_0 = \mathbf{L}\mathbf{L}^\top + \mathbf{I}_p$ with $K = 5$, where entries of $p \times 5$ matrix $\mathbf{L}$ are independently generated from the uniform distribution $U(-1, 1)$. As previously explained, we decompose $\mathbf{\Sigma}_0$ as $\mathbf{\Sigma}_0 = \mathbf{D}(\mathbf{B}\mathbf{B}^\top + \mathbf{M}\mathbf{M}^\top)\mathbf{D}$, generate each element of $\mathbf{w}_i$ and $\mathbf{v}_i$ independently from $\sqrt{2/3} \cdot t_6$, and obtain an p-dimensional random vector $\mathbf{X}_i = \tilde{\boldsymbol{\mu}} + \mathbf{D}(\mathbf{B}\mathbf{w}_i + \mathbf{M}\mathbf{v}_i)$. The normality assumption is violated in this model; thus, we can check how important the normality assumption is for the performance of the proposed method.

- [**Model 4: Cluster model**] In this model, the covariance matrix $\mathbf{\Sigma}_0$ is constructed as follows. First, we generate a p-dimensional vector $\mathbf{\Lambda} = (l_1, \ldots, l_p)^\top$, where

$$l_j \sim \begin{cases} U(150, 170) & \text{for } j = 1, 2, 3, 4, \\ U(3, 6) & \text{for } j = 5, 6, \ldots, 14, \\ U(0.1, 0.3) & \text{for } j = 15, 16, \ldots, p, \end{cases}$$

and generate a $p \times p$ matrix $\mathbf{Q}$ in which each element is generated independently from $N(0, 1)$. Let $\mathbf{\Gamma}$ be the orthonormal matrix that consists of eigenvectors of $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$. Finally, we let $\mathbf{\Sigma}_0 = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^\top$ with $K = 4$. We set the number of factors as $K = 4$ since the eigengap between $\{l_1, \ldots, l_4\}$ and $\{l_5, \ldots, l_{14}\}$ is larger than the gap between $\{l_5, \ldots, l_{14}\}$ and $\{l_{15}, \ldots, l_p\}$.

- [**Model 5: Sparse precision matrix model I**] We consider a precision matrix $\mathbf{\Omega} = \text{diag}(\mathbf{A}_1, \mathbf{A}_2)$ with $\mathbf{A}_1 = \mathbf{B} + c \cdot \mathbf{I}_{p/2 \times p/2}$ and $\mathbf{A}_2 = 4 \cdot \mathbf{I}_{p/2 \times p/2}$. Here, $\mathbf{B}$ is a sparse symmetric matrix in which each element takes a value of 0.5 with a probability of 0.1

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

and takes a value of 0 with a probability of 0.9, and $c = \max\{-\lambda_{\min}(\mathbf{B}), 0\} + 0.01$ is a constant that makes $\mathbf{A}_1$ positive definite. Finally, $\mathbf{\Sigma}_0 = \mathbf{\Omega}^{-1}$.

- [**Model 6: Sparse precision matrix model II**] Similar to Model 5, we consider a precision matrix $\mathbf{\Omega} = \text{diag}(\mathbf{A}_1, \mathbf{A}_2)$. However, in this model, $\mathbf{B}$ is a sparse symmetric matrix of which each element takes a value uniformly in $[0.3, 0.8]$ with probability 0.2 and takes value 0 with probability 0.8, and $\mathbf{\Sigma}_0 = \mathbf{\Omega}^{-1}$.

3.2 Comparison with other multiple testing procedures

In this section, we introduce two existing methods that address multiple testing problems. The first method is the BH procedure proposed by Benjamini and Hochberg (1995), and the second method is the empirical Bayes method introduced by Efron (2004). Unlike our procedure, the empirical Bayes method controls the local false discovery rate.

First, we briefly review the BH procedure. Based on the uniformity of p-values under null hypotheses, it is proved that the BH procedure controls the FDR at a prespecified level $\alpha \in (0, 1)$. Let $\{P_{(j)}\}_{j=1}^{p}$ be sorted p-values obtained from p tests in ascending order. Assuming the independence of p-values, the threshold of the BH procedure is defined as

$$t_{\text{BH}} = \max_{1 \le j \le p} \left\{ P_{(j)} : P_{(j)} \le \frac{j}{p}\alpha \right\}. \tag{3.1}$$

The j-th hypothesis $\mathcal{H}_{0j}$ is rejected if $P_j \le t_{\text{BH}}$. As shown in Benjamini and Hochberg (1995), the BH procedure controls FDR at α. The BH procedure still works well if the test statistics have positive regression dependency (Benjamini and Yekutieli, 2001).

Next, we introduce Efron's empirical Bayes method. Let $\{Z_j\}_{j=1}^{p}$ be the test statistics from p tests, and let π_0 and $\pi_1 = 1 - \pi_0$ be the prior probability of null and non-null

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

hypotheses, respectively. In addition, let $f_0(z)$ and $f_1(z)$ be the density from null and non-null hypotheses, respectively. Then, the marginal density f can be written as a mixture density $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$. The local FDR is defined as

$$\mathrm{fdr}(z) = \frac{\pi_0 f_0(z)}{f(z)}. \tag{3.2}$$

Therefore, the local FDR can be interpreted as a posterior probability of being null, given test statistic z. We need to estimate π_0, $f_0(z)$ and $f(z)$ to control the local FDR. Efron (2004) used the zero-assumption technique to estimate the null density $f_0(z)$ and a nonparametric estimator for estimating the marginal density $f(z)$. See Efron (2004) for details. We note that the local FDR controlling method for a given level is typically more conservative than the FDR controlling method with the same level. Hence, controlling the local FDR at the predetermined level α guarantees the FDR control at level α.

3.3 Fixed threshold level t setting

In this simulation study, we numerically compare the performance of the original PFA method and the proposed DAB-PFA method. For the comparison, we fix the threshold level $t = 0.01$, and consider the relative error of the approximation methods and the relative gap of the conservative surrogates. Here, the relative errors of the original PFA method and the proposed DAB-PFA method are defined by

$$\mathrm{RE}_{\mathrm{Org}} = \frac{\mathrm{FDP}_{\mathrm{orc}}(t) - \mathrm{FDP}(t)}{\mathrm{FDP}(t)},$$
$$\mathrm{RE}_{\mathrm{DAB}} = \frac{\mathrm{FDP}^{\mathrm{DAB}}_{\mathrm{orc}}(t; \lambda, \tau, \epsilon) - \mathrm{FDP}(t)}{\mathrm{FDP}(t)},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

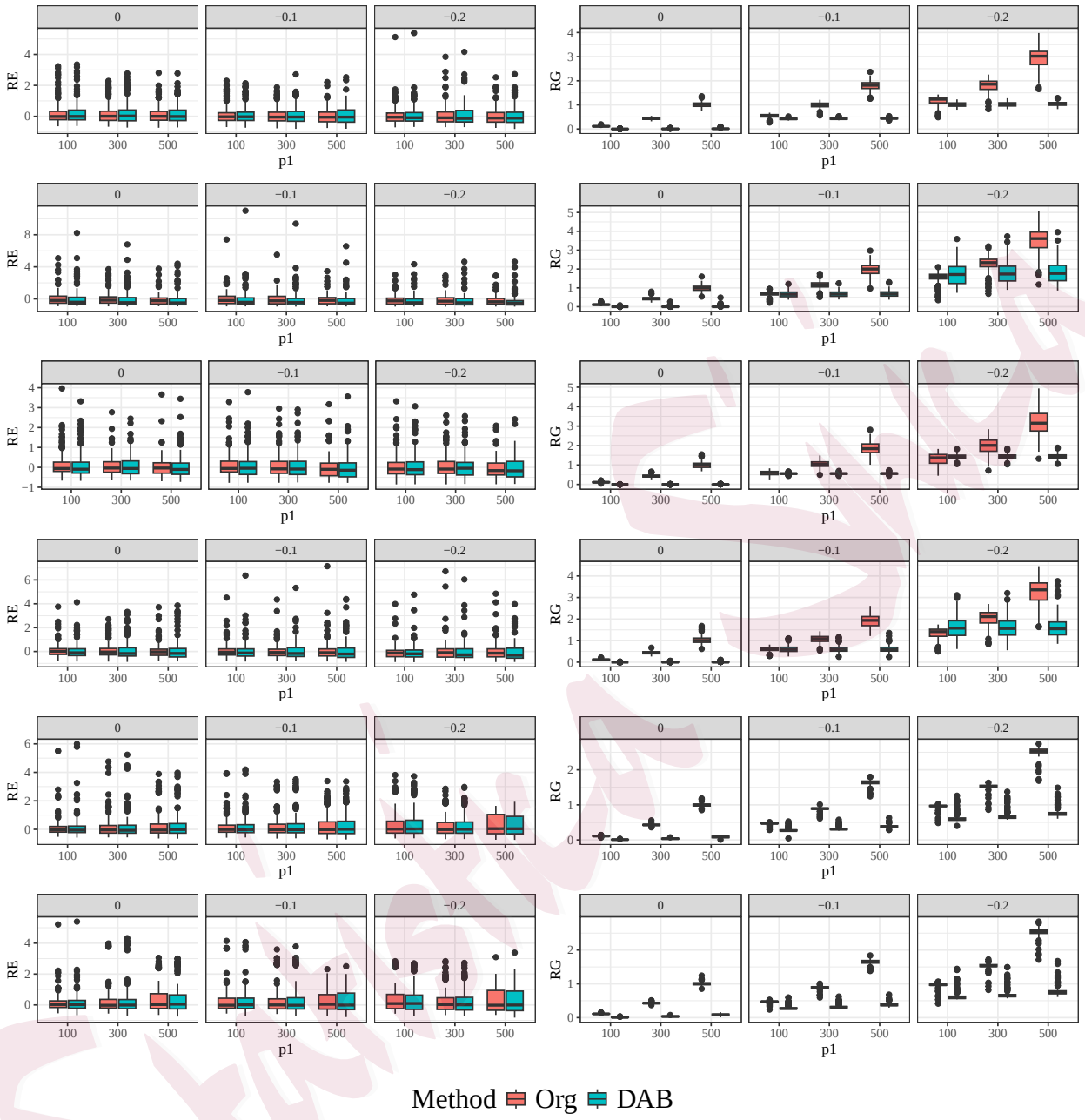

Figure 2: Box plots of relative errors (left column) and relative gaps (right column). Each row shows the result from each dependence structure (Model 1∼6). For each row, there are six panels: first three panels contains box plots of the relative errors and the next three panels contains box plots of the relative gaps for each $\mu_{\mathrm{N}} \in \{0, -0.1, -0.2\}$. Each panel shows the box plots for $p_1 \in \{100, 300, 500\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

respectively. The above relative error measures how much accurately each method approximates the FDP. Similarly, the relative gaps are defined by

$$\mathrm{RG}_{\mathrm{Org}} = \frac{\mathrm{FDP}_{\mathrm{U}}(t) - \mathrm{FDP}_{\mathrm{orc}}(t)}{\mathrm{FDP}_{\mathrm{orc}}(t)},$$
$$\mathrm{RG}_{\mathrm{DAB}} = \frac{\mathrm{FDP}_{\mathrm{U}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon) - \mathrm{FDP}_{\mathrm{orc}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon)}{\mathrm{FDP}_{\mathrm{orc}}^{\mathrm{DAB}}(t;\lambda,\tau,\epsilon)},$$

respectively. A smaller relative gap implies a closer conservative surrogate to the approximation. Figure 2 shows the box plots of the relative errors and the relative gaps. The left column of Figure 2 shows that there is not much difference between the original PFA method and the proposed method in terms of the relative error. However, the right column of Figure 2 shows that the relative gap of the proposed method is much smaller than that of the original method. In addition, as μ_{N} decreases, the relative gap of the original method increases rapidly while that of the proposed method is barely increased, implying that the proposed method is robust against the null mean value.

3.4 Practical setting

We compare the DAB-PFA method with the original PFA method, the BH procedure, and the empirical Bayes (EB) method in a practical setting. When applying the DAB-PFA method or the original PFA method, a covariance matrix estimator is required. We consider two covariance matrix estimators: POET (Fan et al., 2013) and S-POET (Wang and Fan, 2017). We denote two covariance matrix estimators by "P" and "S", respectively. We set the level $\alpha = 0.1$. For each FDP estimation method, we find the largest threshold $\widehat{t} \in (0, 1)$ that makes the estimated FDP value less than the level α. The threshold of the BH procedure is defined as (3.1), and the threshold of the EB method is set as the corresponding p-value of the smallest z-value that controls the estimated local FDR under level α. Then, we compute

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Table 2: The averages of the FDP values over 200 repetition are calculated with their standard error in parentheses for Model 1.

			Method					
Model	μ_N	π_1	P-PFA	DAB-P-PFA	S-PFA	DAB-S-PFA	EB	BH
M1	0	0.1	0.095 (0.038)	0.107 (0.046)	0.092 (0.037)	0.104 (0.045)	0.022 (0.052)	0.073 (0.091)
		0.3	0.072 (0.019)	0.104 (0.030)	0.071 (0.019)	0.103 (0.030)	0.019 (0.027)	0.062 (0.053)
		0.5	0.052 (0.014)	0.103 (0.023)	0.052 (0.013)	0.102 (0.023)	0.016 (0.020)	0.046 (0.032)
	-0.1	0.1	0.068 (0.031)	0.077 (0.034)	0.066 (0.029)	0.074 (0.033)	0.015 (0.041)	0.054 (0.076)
		0.3	0.056 (0.018)	0.080 (0.023)	0.054 (0.018)	0.078 (0.023)	0.013 (0.023)	0.048 (0.046)
		0.5	0.041 (0.012)	0.085 (0.020)	0.040 (0.012)	0.084 (0.019)	0.012 (0.015)	0.037 (0.028)
	-0.2	0.1	0.050 (0.026)	0.056 (0.028)	0.049 (0.026)	0.053 (0.028)	0.011 (0.036)	0.041 (0.064)
		0.3	0.042 (0.015)	0.061 (0.022)	0.041 (0.015)	0.060 (0.022)	0.009 (0.018)	0.037 (0.041)
		0.5	0.032 (0.010)	0.070 (0.018)	0.031 (0.010)	0.069 (0.018)	0.008 (0.013)	0.029 (0.025)

the false discovery proportion

$$\mathrm{FDP} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}$$

and the true positive proportion

$$\mathrm{TPP} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

where TP, FP, and FN are the numbers of true positives, false positives, and false negatives, respectively.

This scenario simulates a situation we encounter in a real-world research problem. Practically, we cannot obtain information about the FDP or the TPP of a multiple testing procedure since we do not know which hypotheses are true nulls. Through the simulation study in this practical setting for various data-generating models, we expect the performance comparison of the presented methods to represent the results in a real-world analysis well. The results of two measurements for Model 1 are summarized in Tables 2 and 3. The results for the other models are summarized in Tables S1, S2, S3, and S4, which can be found in

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Table 3: The averages of the TPP values over 200 repetition are calculated with their standard error in parentheses for Model 1.

			Method					
Model	μ_N	π_1	P-PFA	DAB-P-PFA	S-PFA	DAB-S-PFA	EB	BH
M1	0	0.1	0.786 (0.134)	0.797 (0.129)	0.781 (0.134)	0.793 (0.130)	0.519 (0.088)	0.711 (0.064)
		0.3	0.874 (0.107)	0.903 (0.091)	0.872 (0.107)	0.901 (0.092)	0.698 (0.064)	0.860 (0.051)
		0.5	0.909 (0.089)	0.953 (0.056)	0.908 (0.089)	0.952 (0.057)	0.791 (0.060)	0.911 (0.043)
	-0.1	0.1	0.782 (0.134)	0.791 (0.139)	0.779 (0.135)	0.786 (0.140)	0.437 (0.134)	0.707 (0.066)
		0.3	0.872 (0.107)	0.897 (0.100)	0.870 (0.108)	0.896 (0.101)	0.679 (0.053)	0.858 (0.052)
		0.5	0.909 (0.089)	0.949 (0.065)	0.908 (0.089)	0.948 (0.065)	0.777 (0.053)	0.909 (0.044)
	-0.2	0.1	0.781 (0.135)	0.787 (0.146)	0.777 (0.136)	0.783 (0.147)	0.423 (0.120)	0.705 (0.066)
		0.3	0.871 (0.108)	0.892 (0.108)	0.869 (0.109)	0.891 (0.109)	0.671 (0.056)	0.856 (0.053)
		0.5	0.908 (0.089)	0.945 (0.072)	0.907 (0.090)	0.945 (0.073)	0.771 (0.053)	0.909 (0.044)

the supplementary material.

We note that all compared methods control the FDR well under the prespecified level $\alpha = 0.1$ in most cases. However, the proposed method exhibits inflated FDR values when $\mu_N = 0$ and $\pi_1 = 100$ for Models 2, 3, and 4. This phenomenon can be explained by Figure 2. In Figure 2, when $\mu_N = 0$ and $\pi_1 = 100$, the relative errors of the proposed DAB-PFA method show a negative bias for Models 2, 3, and 4, while the relative gaps are nearly zero. This illustrates that although the discarding adaptively with bounding (DAB) approach significantly reduces the relative gap, it also introduces instability in the approximation of the FDP. When the underlying dependence structure satisfies the required conditions–weak dependence (condition (C1)) and the existence of a reliable covariance matrix estimator (conditions (C5) and (C6))–the proposed method outperforms the original PFA method. However, in Models 2, 3, and 4, where these conditions are not satisfied, a large relative gap can act as a buffer against the instability of the PFA method's approximation, whereas a

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

small relative gap may lead to an underestimation of the upper bound, resulting in slightly inflated FDR. Importantly, this inflation is not observed when μ_{N} takes negative values. Since the true parameter of interest is rarely exactly at the boundary of the null set in practical scenarios, we conclude that the DAB-PFA method demonstrates robustly better performance compared to other methods in terms of true positive rate (TPR).

3.5 Sensitivity analysis

In the previous simulation studies and the following case study, we use fixed threshold parameters $(\lambda, \tau, \epsilon) = (0.1, 0.5, 0.01)$. We numerically investigate how sensitive the performance of the DAB-PFA method is when we change the threshold parameters. We consider $\zeta = \lambda/\tau \in \{0.1, 0.2, 0.3\}$, $\tau \in \{0.4, 0.5, 0.6\}$, and $\epsilon \in \{0.001, 0.01, 0.1\}$. Since the number of all possible combinations of threshold parameters is considerably large, we only take into consideration the combinations that are different from the standard combination of threshold parameters $(\zeta, \tau, \epsilon) = (0.2, 0.5, 0.01)$ for only one parameter. The averages of the FDP and TPP values from 200 Monte Carlo simulations are summarized in Tables S5, S6, and S7 in the supplementary material.

We first focus on the effect of the selection of $\zeta = \lambda/\tau$. This parameter decides which proportion of hypotheses we regard as true nulls. From Table S5, we can find that using a smaller ζ yields a larger controlled FDR and a larger TPR when the null distribution is conservative. We note that as the null distribution becomes more conservative, the number of true null hypotheses used in approximating the FDP decreases. Then, using a small ζ value can increase the inclusion of true null hypotheses in approximating the FDP, which makes the approximation more accurate. However, using too small ζ may include some

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

non-null hypotheses in approximating the number of false rejections, so we suggest to use $\zeta = 0.2$.

We then focus on the sensitivity of the proposed method with respect to the value of τ. We note that τ is the essential parameter that makes our method robust when the true null distribution is conservative. From Table S6, we notice that using a smaller τ is advantageous because it prevents too conservative true nulls to be included in approximating the FDP. However, if we use too small τ, we have to approximate the FDP only with a small number of hypotheses. For these reasons, we propose to use $\tau = 0.5$ in general.

Finally, we examine the sensitivity analysis of ϵ. As mentioned in the remark following Theorem 3, The inclusion of ϵ prevents the denominators in the DAB-PFA estimator from becoming excessively small, a situation that may arise when estimates of $\{(a_j, \eta_j)\}_{j=1}^{p}$ pop out. In this sensitivity analysis, however, the estimations of $\widehat{\mathbf{\Sigma}}$ and its eigenvalues and eigenvectors are precisely conducted because the simulation data are generated from a well structured factor model. Hence, Table S7 shows that the selection of ϵ is not sensitive to the performance of the DAB-PFA method. Here, we provide $\epsilon = 0.01$ as a standard, but depending on the accuracy of the covariance matrix estimation, we need to choose a proper value of ϵ in practice.

4. Case study: Protein phosphorylation analysis

In this section, we demonstrate the proposed method for the protein phosphorylation analysis of ovarian serous adenocarcinoma samples compared with the other aforementioned procedures. Here, we aim to identify uniquely elevated protein modification levels in each of the five molecular subtypes, as defined by clustering analysis of mRNA-based gene expression

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

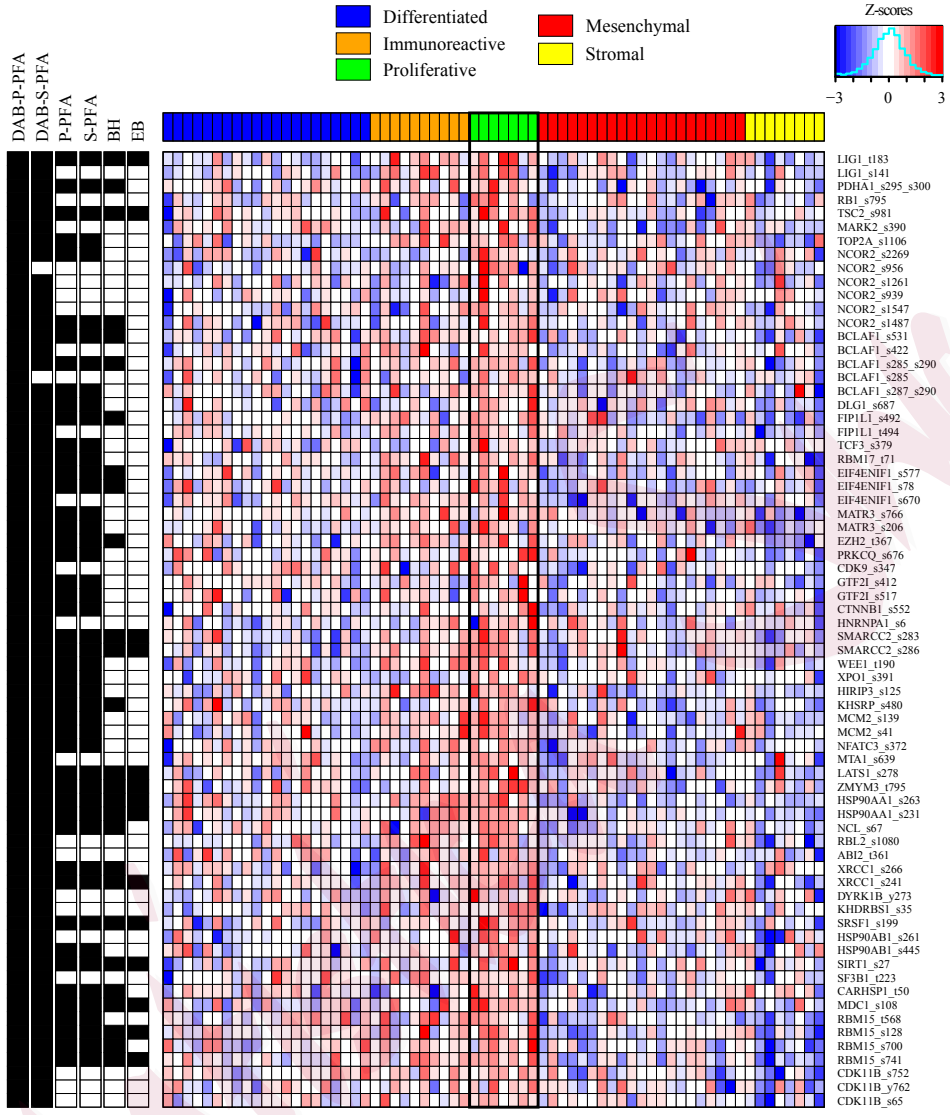

Figure 3: Heatmap of relative phosphorylation levels at high confidence phosphorylation sites with validated evidence of signaling by three or more kinases. Phosphorylation is uniquely elevated in these sites for the samples of the proliferative subtype. The black and white color bar to the far right side of the figure shows whether each phosphorylation site was above the threshold by each method (black = null hypothesis rejected, and white = not rejected).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Table 4: The number and the rate (%) of rejected hypotheses are shown for five molecular subtypes: differentiated subtype(A), immunoreactive subtype(B), proliferative subtype(C), mesenchymal subtype(D), and stromal subtype(E).

	Case				
Method	A	B	C	D	E
DAB-P-PFA	25 (0.44)	257 (4.47)	741 (12.90)	485 (8.44)	512 (8.91)
P-PFA	16 (0.28)	176 (3.06)	537 (9.35)	437 (7.61)	314 (5.46)
DAB-S-PFA	18 (0.31)	220 (3.83)	730 (12.70)	479 (8.34)	501 (8.72)
S-PFA	14 (0.24)	174 (3.03)	533 (9.28)	437 (7.61)	300 (5.22)
EB	1 (0.02)	253 (4.40)	202 (3.52)	200 (3.48)	73 (1.27)
BH	3 (0.05)	272 (4.73)	354 (6.16)	269 (4.68)	147 (2.56)

data (Zhang et al., 2016). Comparison of protein phosphorylation levels is a representative example of multiple one-sided tests since there is an implicit dependence structure among multiple phosphorylation sites on an identical substrate protein and the substrates phosphorylated by the same kinases. Moreover, mass spectrometry-based proteomics experiments do not always detect the same phosphopeptides because each experiment has an uncontrollable degree of variability. As a result, the dependence structure across the phosphorylation site is unique in each dataset, and multiple testing procedures that are robust against an arbitrary degree of dependence are of the utmost importance.

There are five molecular subtypes: differentiated (A), immunoreactive (B), proliferative (C), mesenchymal (D), and stromal (E). Each molecular subtype has $(n_A, n_B, n_C, n_D, n_E) = (21, 10, 7, 21, 8)$ samples. The protein phosphorylation level at the j-th site in the i-th sample of subtype g is denoted by $\{X_{g,ij}\}$, where $g \in G = \{A, B, C, D, E\}$, $i = 1, ..., n_g$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

$j = 1, ..., 5746$. In this section, we investigate whether the proliferative (C) molecular subtype samples have larger phosphorylation levels than others. Comparisons between other groups are given in the supplementary material. For each phosphorylation site $j = 1, ..., 5746$, we consider the hypotheses

$$\mathcal{H}_{0j} : \mu_{C,j} \leq \mu_{-C,j} \quad \text{vs.} \quad \mathcal{H}_{1j} : \mu_{C,j} > \mu_{-C,j}.$$

where $\mu_{C,j}$ is the mean phosphorylation level at site j of proliferative samples and $\mu_{-C,j}$ is the mean phosphorylation level at site j of other subtype samples. Then, the test statistic corresponding to phosphorylation site j is

$$T_j = \frac{\bar{X}_{C,j} - \bar{X}_{-C,j}}{s_j\sqrt{(1/n_C + 1/n_{-C})}}$$

where $\bar{X}_{C,j} = \frac{1}{n_C}\sum_{i=1}^{n_C} X_{C,ij}$ is the sample mean of phosphorylation levels at site j of proliferative samples, $\bar{X}_{-C,j} = \frac{1}{n-n_C}\sum_{g\in G\setminus\{C\}}\sum_{i=1}^{n_g} X_{g,ij}$ is the sample mean of phosphorylation levels at site j of other subtype samples, and $n_{-C} = n - n_C$ and $s_j^2 = \frac{1}{n-5}\sum_{g\in G}\sum_{i=1}^{n_g}(X_{g,ij} - \bar{X}_{g,j})^2$. Under the null hypothesis $\mathcal{H}_{0j}$, T_j follows the t distribution with degrees of freedom $n - 5$. Therefore, the p-value of the j-th test is $P_j = F_{n-5}^{-1}(-T_j)$, where F_{n-5} is the cumulative distribution function of t_{n-5}. The BH procedure and the empirical Bayes method are applied to these p-values $\{P_j\}_{j=1}^p$. To apply the DAB-PFA, we need to estimate the covariance structure of $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$ in which

$$Z_j = \frac{\bar{X}_{C,j} - \bar{X}_{-C,j}}{\sqrt{(1/n_C + 1/n_{-C})}}.$$

Under the homogeneous Gaussian assumption, $\mathbf{X}_{g,i} = (X_{g,i1}, \ldots, X_{g,ip})^\top \sim N_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_g = \boldsymbol{\mu}_C$ if $g = C$, and $\boldsymbol{\mu}_g = \boldsymbol{\mu}_{-C}$ otherwise. As a result, we have $\mathrm{Cov}(\mathbf{Z}) = \boldsymbol{\Sigma}$.

When estimating $\boldsymbol{\Sigma}$, we propose using POET or S-POET. In one sample test case, we apply POET and S-POET to the sample covariance matrix $S = (n-1)^{-1}\sum_i(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i -$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

$\bar{\mathbf{X}})^\top$. However, in this case, we estimate $\mathbf{\Sigma}$ by applying POET and S-POET to the pooled sample covariance matrix $S_{\text{pooled}} = (n-5)^{-1}\sum_{g\in G}\sum_{i=1}^{n_g}(\mathbf{X}_{g,i} - \bar{\mathbf{X}}_g)(\mathbf{X}_{g,i} - \bar{\mathbf{X}}_g)^\top$ instead. Using this covariance estimator, we implement our method for the FDP estimation and obtain the threshold.

In the following section, we compare the following six methods: DAB-P-PFA, P-PFA, DAB-S-PFA, S-PFA, EB, and BH. We set the FDR level $\alpha = 0.1$ and threshold parameters $(\lambda, \tau, \epsilon) = (0.1, 0.5, 0.01)$. Table 4 shows that the FDP estimation methods (DAB-P-PFA, P-PFA, DAB-S-PFA, and S-PFA) reject more hypotheses than the other methods (EB and BH). In particular, across the results for the five subtypes, the DAB methods consistently identify the largest number of phosphorylation sites as uniquely elevated in each subtype. To examine whether the additionally rejected hypotheses lead to biologically meaningful phosphorylation, we prioritize the substrate proteins known to be phosphorylated by at least three kinases in each comparison, selected by at least one of the six methods. These sites are most likely to be *true positives* in the sense that the activity of corresponding kinases on the exact sites has been validated in independent experiments of human cells (Hornbeck et al., 2015; Hu et al., 2014; Corwin et al., 2017). As expected, excluding the differentiated subtype with a low number of rejected hypotheses, the PFA procedures select a comparable number of phosphosites (see the supplementary material for the other heatmaps), often surpassing the number of rejections in both the Benjamini-Hochberg procedure and empirical Bayes method.

However, for the proliferative subtype, we observe that the DAB-PFA methods substantially improve the detection of additional true phosphosites, especially in the same substrates with other phosphosites detected by the original PFA methods. Figure 3 shows the trans-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

formed phosphorylation levels (Z-scores) for the sites, with consistently increased phosphorylation levels in the proliferative subtype ($n = 7$). For the heatmaps of other subtypes, see Figures S1, S2 and S3. It is evident that the DAB methods considerably improve the power to detect these sites, with adjacent phosphosites already detected by other methods.

5. Discussion

We have proposed a new multiple one-sided hypotheses testing procedure called the DAB-PFA to control the FDR under the general dependency of test statistics. We use a principal factor model to approximate the test statistics with a general dependency structure. As a result, we can express an approximation of the FDP as a function of the eigenvalues and eigenvectors of the covariance matrix of the test statistics. To account for the conservative null p-values from one-sided hypotheses, we suggest discarding p-values close to 0 or 1 by introducing an indicator $I(\lambda < P_j \leq \tau)$. We then use the upper bound of the FDP estimator to avoid estimating μ_j. In practice, we plug in the estimates of the eigenvalues and eigenvectors to obtain the approximation of the FDP. In our simulation studies, the proposed approximation shows an excellent approximation to the true FDP. We also show the FDR control of the proposed procedure and compare its power with other multiple testing procedures.

Instead of using the upper bound by removing μ_js, we can consider a plug-in estimator of μ_js to construct the upper bound. If we estimate μ_js consistently or accurately, we may obtain an upper bound with a smaller gap and a better approximation of the FDP. Here, we have a large number of parameters to be estimated, and no doubt need some structural assumptions to estimate them all consistently. It would be interesting to find the conditions

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

and propose an improved FDP approximation procedure.

Supplementary Material

The online Supplementary Material contains proofs of the main theorems, details of simulation results and the heatmaps from the case study.

Acknowledgments

The authors thank the co-editor, an associate editor, and three reviewers for their constructive suggestions and comments, which led to substantial improvements in the paper. Jang's research is supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 0769-20240034). Cho's research is supported by an INHA University Research Grant. Lim's research is supported by the National Research Foundation of Korea (No. NRF-2021R1A2C1010786) and the Brain Pool Program, which is also funded by the National Research Foundation of Korea and the Ministry of Education (NRF-2022H1D3A2A01063793)

References

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*(4), 1165–1188.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Bickel, P. J. and E. Levina (2008a). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577–2604.

Bickel, P. J. and E. Levina (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics 36*(1), 199–227.

Cai, T. T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association 106*(494), 672–684.

Cohen, A. and H. B. Sackrowitz (2005). Decision theory results for one-sided multiple comparison procedures. *Annals of Statistics 33*(1), 126–144.

Corwin, T., J. Woodsmith, F. Apelt, J.-F. Fontaine, D. Meierhofer, J. Helmuth, A. Grossmann, M. A. Andrade-Navarro, B. A. Ballif, and U. Stelzl (2017). Defining human tyrosine kinase phosphorylation networks using yeast as an in vivo model substrate. *Cell Systems 5*(2), 128–139.e4.

Dobriban, E. (2020). Permutation methods for factor analysis and PCA. *The Annals of Statistics 48*(5), 2824–2847.

Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association 99*(465), 96–104.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association 102*(477), 93–103.

Fan, J. and X. Han (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(4), 1143–1164.

Fan, J., X. Han, and W. Gu (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association 107*(499), 1019–1035.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(4), 603–680.

Finner, H. and M. Roters (2002). Multiple hypotheses testing and expected number of type i. errors. *The Annals of*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Statistics 30(1), 220–238.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 499–517.

Hornbeck, P. V., B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, and E. Skrzypek (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research 43*(D1), D512–D520.

Hu, J., H.-S. Rho, R. H. Newman, J. Zhang, H. Zhu, and J. Qian (2014). PhosphoNetworks: a database for human phosphorylation networks. *Bioinformatics 30*(1), 141–142.

Liu, J., C. Zhang, and D. Page (2016). Multiple testing under dependence via graphical models. *The Annals of Applied Statistics 10*(3), 1699–1724.

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(3), 411–426.

Ramdas, A., T. Zrnic, M. J. Wainwright, and M. Jordan (2018). SAFFRON: an adaptive algorithm for online control of the false discovery rate. In J. Dy and A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Volume 80 of *Proceedings of Machine Learning Research*, pp. 4286–4294. PMLR.

Romano, J. P., A. M. Shaikh, and M. Wolf (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST 17*(3), 417–442.

Sarkar, S. K. (2004). FDR-controlling stepwise procedures and their false negatives rates. *Journal of Statistical Planning and Inference 125*(1), 119–137.

Sarkar, S. K. (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures. *The Annals of Statistics 34*(1), 394–415.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 479–498.

Sun, W. and T. T. Cai (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society:*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Series B (Statistical Methodology) 71(2), 393–424.

Tian, J. and A. Ramdas (2019). ADDIS: An adaptive discarding algorithm for online FDR control with conservative nulls. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, pp. 9383–9391. NeurIPS.

Wang, W. and J. Fan (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics 45*(3), 1342–1374.

Wei, Z., W. Sun, K. Wang, and H. Hakonarson (2009). Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics 25*(21), 2802–2808.

Wu, W. B. (2008). On false discovery control under dependence. *The Annals of Statistics 36*(1), 364–380.

Xiao, J., W. Zhu, and J. Guo (2013). Large-scale multiple testing in genome-wide association studies via region-specific hidden markov models. *BMC Bioinformatics 14*(1), 282.

Zhang, H., T. Liu, Z. Zhang, S. H. Payne, B. Zhang, J. E. McDermott, J.-Y. Zhou, V. A. Petyuk, L. Chen, D. Ray, S. Sun, F. Yang, L. Chen, J. Wang, P. Shah, S. W. Cha, P. Aiyetan, S. Woo, Y. Tian, M. A. Gritsenko, T. R. Clauss, C. Choi, M. E. Monroe, S. Thomas, S. Nie, C. Wu, R. J. Moore, K.-H. Yu, D. L. Tabb, D. Fenyö, V. Bafna, Y. Wang, H. Rodriguez, E. S. Boja, T. Hiltke, R. C. Rivers, L. Sokoll, H. Zhu, I.-M. Shih, L. Cope, A. Pandey, B. Zhang, M. P. Snyder, D. A. Levine, R. D. Smith, D. W. Chan, K. D. Rodland, S. A. Carr, M. A. Gillette, K. R. Klauser, E. Kuhn, D. Mani, P. Mertins, K. A. Ketchum, R. Thangudu, S. Cai, M. Oberti, A. G. Paulovich, J. R. Whiteaker, N. J. Edwards, P. B. McGarvey, S. Madhavan, P. Wang, D. W. Chan, A. Pandey, I.-M. Shih, H. Zhang, Z. Zhang, H. Zhu, L. Cope, G. A. Whiteley, S. J. Skates, F. M. White, D. A. Levine, E. S. Boja, C. R. Kinsinger, T. Hiltke, M. Mesri, R. C. Rivers, H. Rodriguez, K. M. Shaw, S. E. Stein, D. Fenyo, T. Liu, J. E. McDermott, S. H. Payne, K. D. Rodland, R. D. Smith, P. Rudnick, M. Snyder, Y. Zhao, X. Chen, D. F. Ransohoff, A. N. Hoofnagle, D. C. Liebler, M. E. Sanders, Z. Shi, R. J. Slebos, D. L. Tabb, B. Zhang, L. J. Zimmerman, Y. Wang, S. R. Davies, L. Ding, M. J. Ellis, and R. R. Townsend (2016). Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell 166*(3), 755–765.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0022

Zhao, Q., D. S. Small, and W. Su (2019). Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *Journal of the American Statistical Association 114* (527), 1291–1304.

Seonghun Cho (Inha University)

E-mail: seonghun.cho@inha.ac.kr

Youngrae Kim (Seoul National University)

Johan Lim (Seoul National University)

E-mail: johanlim@snu.ac.kr

Hyungwon Choi (National University of Singapore)

DoHwan Park (University of Maryland at Baltimore County)

Woncheol Jang (Seoul National University)

E-mail: wcjang@snu.ac.kr