
Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

# Estimating False Discovery Proportion Under Arbitrary Covariance Dependence

Jianqing Fan [a b], Xu Han [c] & Weijie Gu [a]

[a] Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ, 08544, USA

[b] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

[c] Department of Statistics, Fox Business School, Temple University, Philadelphia, PA, 19122

Accepted author version posted online: 20 Aug 2012.Version of record first published: 08 Oct 2012.

To cite this article: Jianqing Fan, Xu Han & Weijie Gu (2012): Estimating False Discovery Proportion Under Arbitrary Covariance Dependence, Journal of the American Statistical Association, 107:499, 1019-1035

To link to this article: http://dx.doi.org/10.1080/01621459.2012.720478

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

# Estimating False Discovery Proportion Under Arbitrary Covariance Dependence

Jianqing FAN, Xu HAN, and Weijie GU

Multiple hypothesis testing is a fundamental problem in high-dimensional inference, with wide applications in many scientific fields. In genome-wide association studies, tens of thousands of tests are performed simultaneously to find if any single-nucleotide polymorphisms (SNPs) are associated with some traits and those tests are correlated. When test statistics are correlated, false discovery control becomes very challenging under arbitrary dependence. In this article, we propose a novel method—based on principal factor approximation—that successfully subtracts the common dependence and weakens significantly the correlation structure, to deal with an arbitrary dependence structure. We derive an approximate expression for false discovery proportion (FDP) in large-scale multiple testing when a common threshold is used and provide a consistent estimate of realized FDP. This result has important applications in controlling false discovery rate and FDP. Our estimate of realized FDP compares favorably with Efron's approach, as demonstrated in the simulated examples. Our approach is further illustrated by some real data applications. We also propose a dependence-adjusted procedure that is more powerful than the fixed-threshold procedure. Supplementary material for this article is available online.

KEY WORDS: Arbitrary dependence structure; False discovery rate; Genome-wide association studies; High-dimensional inference; Multiple hypothesis testing.



## 1. INTRODUCTION

Multiple hypothesis testing is a fundamental problem in modern research for high-dimensional inference, with wide applications in scientific fields, such as biology, medicine, genetics, neuroscience, economics, and finance. For example, in genome-wide association studies, massive amount of genomic data (e.g., single-nucleotide polymorphism (SNPs), expression quantitative trait loci (eQTLs)) are collected and tens of thousands of hypotheses are tested simultaneously to find if any of these genomic data are associated with some observable traits (e.g., blood pressure, weight, some disease); in finance, thousands of tests are performed to see which fund managers have winning ability (Barras, Scaillet, and Wermers 2010).

False discovery rate (FDR) was introduced in the celebrated article by Benjamini and Hochberg (1995) for large-scale multiple testing. By definition, FDR is the expected proportion of falsely rejected null hypotheses among all of the rejected null hypotheses. The classification of tested hypotheses can be summarized in Table 1.

Various testing procedures have been developed for controlling FDR, among which there are two major approaches. One is to compare the $p$-values with a data-driven threshold as in Benjamini and Hochberg (1995). Specifically, let $p_{(1)} \le p_{(2)} \le \cdots \le p_{(p)}$ be the ordered observed $p$-values of $p$ hypotheses. Define $k = \max\{i : p_{(i)} \le i\alpha/p\}$ and reject $H^0_{(1)}, \ldots, H^0_{(k)}$, where $\alpha$ is a specified control rate. If no such $i$ exists, reject no hypothesis. The other related approach is to fix a threshold value $t$, estimate the FDR, and choose $t$ so that the estimated FDR is no larger than $\alpha$ (Storey 2002). Finding such a common threshold is based on a conservative estimate of FDR. Specifically, let $\widehat{\mathrm{FDR}}(t) = \widehat{p}_0 t/(R(t) \vee 1)$, where $R(t)$ is the number of total discoveries with the threshold $t$ and $\widehat{p}_0$ is an estimate of $p_0$. Then, solve $t$ such that $\widehat{\mathrm{FDR}}(t) \le \alpha$. The equivalence between the two methods has been theoretically studied by Storey, Taylor, and Siegmund (2004) and Ferreira and Zwinderman (2006).

Both procedures have been shown to control the FDR for independent test statistics. However, in practice, test statistics are usually correlated. Although Benjamini and Yekutieli (2001) and Clarke and Hall (2009) argued that when the null distribution of test statistics satisfies some conditions, the dependence case in the multiple testing is asymptotically the same as the independence case, multiple testing under general dependence structures is still a very challenging and important open problem. Efron (2007) pioneered the work in the field and noted that correlation must be accounted for in deciding which null hypotheses are significant because the accuracy of FDR techniques will be compromised in high correlation situations. There are several articles that show that the Benjamini–Hochberg (B–H) procedure or Storey's procedure can control FDR under some special dependence structures, for example, positive regression dependence on subsets (Benjamini and Yekutieli 2001) and weak dependence (Storey, Taylor, and Siegmund 2004). Sarkar (2002) also showed that FDR can be controlled by a generalized stepwise multiple testing procedure under positive regression dependence on subsets. However, even if the procedures are valid under these special dependence structures, they will still suffer from efficiency loss without considering the actual dependence information. In other words, there are universal upper bounds for a given class of covariance matrices.

In this article, we develop a procedure for high-dimensional multiple testing, which can deal with any arbitrary dependence

Jianqing Fan is Frederick L. Moore'18 Professor, Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ 08544, USA and Honorary Professor, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China (E-mail: *jqfan@princeton.edu*). Xu Han is Assistant Professor, Department of Statistics, Fox Business School, Temple University, Philadelphia, PA 19122 (E-mail: *hanxu3@temple.edu*). Weijie Gu is graduate student, Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ 08544, USA (E-mail: *wgu@princeton.edu*). The article was completed while Xu Han was a postdoctoral fellow at Princeton University. This research was partly supported by NSF grants DMS-0704337 and DMS-0714554 and NIH grant R01-GM072611. The authors are grateful to the editor, associate editor, and referees for helpful comments.

Table 1. Classification of tested hypotheses

| Number | Number not rejected | Number rejected | Total |
|---|---|---|---|
| True null | $U$ | $V$ | $p_0$ |
| False null | $T$ | $S$ | $p_1$ |
| | $p-R$ | $R$ | $p$ |

structure and fully incorporate the covariance information. This is in contrast with previous articles that considers multiple testing under special dependence structures; for example, Sun and Cai (2009) developed a multiple testing procedure where parameters underlying test statistics follow a hidden Markov model, and Leek and Storey (2008) and Friguet, Kloareg, and Causeur (2009) studied multiple testing under the factor models. More specifically, consider the test statistics

$$(Z_1, \ldots, Z_p)^T \sim N((\mu_1, \ldots, \mu_p)^T, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is known and $p$ is large. We would like to simultaneously test $H_{0i} : \mu_i = 0$ versus $H_{1i} : \mu_i \neq 0$ for $i = 1, \ldots, p$. Note that $\mathbf{\Sigma}$ can be any nonnegative definite matrix. Our procedure is called principal factor approximation (PFA). The basic idea is to first take out the principal factors that derive the strong dependence among observed data $Z_1, \ldots, Z_p$ and to account for such dependence in calculation of false discovery proportion (FDP). This is accomplished by the spectral decomposition of $\mathbf{\Sigma}$ and taking out the largest common factors so that the remaining dependence is weak. We then derive the asymptotic expression of the FDP, defined as $V/R$, which accounts for the strong dependence. The realized but unobserved principal factors that derive the strong dependence are then consistently estimated. The estimate of the realized FDP is obtained by substituting the consistent estimate of the unobserved principal factors.

We are especially interested in estimating FDP under the high-dimensional sparse problem, that is, $p$ is very large, but the number of $\mu_i \neq 0$ is very small. In Section 2, we will explain the situation under which $\mathbf{\Sigma}$ is known. Sections 3 and 4 present the theoretical results and the proposed procedures. In Section 5, the performance of our procedures is critically evaluated by various simulation studies. Section 6 is about the real data analysis. All the proofs are relegated to the Appendix and the Supplementary material.

## 2. MOTIVATION OF THE STUDY

In genome-wide association studies, consider $p$ SNP genotype data for $n$ individual samples, and further suppose that a response of interest (i.e., gene expression level or a measure of phenotype such as blood pressure or weight) is recorded for each sample. The SNP data are conventionally stored in an $n \times p$ matrix $\mathbf{X} = (x_{ij})$, with rows corresponding to individual samples and columns corresponding to individual SNPs . The total number $n$ of samples is in the order of hundreds, and the number $p$ of SNPs is in the order of tens of thousands.

Let $X_j$ and $Y$ denote, respectively, the random variables that correspond to the $j$th SNP coding and the outcome. The biological question of the association between genotype and phenotype can be restated as a problem in multiple hypothesis testing, that is, the simultaneous tests for each SNP $j$ of the null hypothesis $H_j$ of no association between the SNP $X_j$ and $Y$. Let $\{X_j^i\}_{i=1}^n$ be the sample data of $X_j$, $\{Y^i\}_{i=1}^n$ be the independent sample random variables of $Y$. Consider the marginal linear regression between $\{Y^i\}_{i=1}^n$ and $\{X_j^i\}_{i=1}^n$:

$$(\alpha_j, \beta_j) = \operatorname{argmin}_{a_j, b_j} \frac{1}{n} \sum_{i=1}^{n} E\big(Y^i - a_j - b_j X_j^i\big)^2, \quad j = 1, \ldots, p. \tag{1}$$

where expectation is taken conditionally given $\{X_j^i\}_{i=1}^n$.

We wish to simultaneously test the hypotheses

$$H_{0j} : \beta_j = 0 \quad \text{versus} \quad H_{1j} : \beta_j \neq 0, \quad j = 1, \ldots, p \tag{2}$$

to see which SNPs are correlated with the phenotype.

Recently, statisticians have shown increasing interest in the high-dimensional sparse problem: although the number of hypotheses to be tested is large, the number of false nulls ($\beta_j \neq 0$) is very small. For example, among 2000 SNPs, there are maybe only 10 SNPs that contribute to the variation in phenotypes or certain gene expression level. Our purpose is to find these 10 SNPs by multiple testing with some statistical accuracy.

Because of the sample correlations among $\{X_j^i\}_{i=1,j=1}^{i=n,j=p}$, the least-square estimators $\{\widehat{\beta}_j\}_{j=1}^p$ for $\{\beta_j\}_{j=1}^p$ in (1) are also correlated. The following result describes the joint distribution of $\{\widehat{\beta}_j\}_{j=1}^p$. The proof is straightforward.

*Proposition 1.* Let $\widehat{\beta}_j$ be the least-square estimator for $\beta_j$ in (1) based on $n$ data points, $r_{kl}$ be the sample correlation between $X_k$ and $X_l$ and $s_{kk}$ be the sample standard deviation of $X_k$. Assume that the conditional distribution of $Y^i$ given $X_1^i, \ldots, X_p^i$ is $N(\mu(X_1^i, \ldots, X_p^i), \sigma^2)$. Then, conditioning on $\{X_j^i\}_{i=1,j=1}^{i=n,j=p}$, the joint distribution of $\{\widehat{\beta}_j\}_{j=1}^p$ is $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^T \sim N((\beta_1, \ldots, \beta_p)^T, \mathbf{\Sigma}^*)$, where the $(k, l)$th element in $\mathbf{\Sigma}^*$ is $\mathbf{\Sigma}^*_{kl} = \sigma^2 r_{kl}/(n s_{kk} s_{ll})$.

For ease of notation, let $Z_1, \ldots, Z_p$ be the standardized random variables of $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$, that is,

$$Z_i = \frac{\widehat{\beta}_i}{\mathrm{SD}(\widehat{\beta}_i)} = \frac{\widehat{\beta}_i}{\sigma/(\sqrt{n} s_{ii})}, \quad i = 1, \ldots, p. \tag{3}$$

In the previous equation, we implicitly assume that $\sigma$ is known and the above standardized random variables are $Z$-test statistics. The estimate of residual variance $\sigma^2$ will be discussed in Section 6 via refitted cross-validation (RCV; Fan, Guo, and Hao 2012). Then, conditioning on $\{X_j^i\}$,

$$(Z_1, \ldots, Z_p)^T \sim N((\mu_1, \ldots, \mu_p)^T, \mathbf{\Sigma}), \tag{4}$$

where $\mu_i = \sqrt{n}\beta_i s_{ii}/\sigma$ and $\mathbf{\Sigma}$ has the $(k, l)$th element as $r_{kl}$. Simultaneously testing (2) based on $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^T$ is thus equivalent to testing

$$H_{0j} : \mu_j = 0 \quad \text{versus} \quad H_{1j} : \mu_j \neq 0, \quad j = 1, \ldots, p \tag{5}$$

based on $(Z_1, \ldots, Z_p)^T$.

In (4), $\mathbf{\Sigma}$ is the population covariance matrix of $(Z_1, \ldots, Z_p)^T$, and is known based on the sample data $\{X_j^i\}$. The covariance matrix $\mathbf{\Sigma}$ can have arbitrary dependence structure. We would like to clarify that $\mathbf{\Sigma}$ is known and there is no estimation of the covariance matrix of $X_1, \ldots, X_p$ in this setup.

## 3. ESTIMATING FALSE DISCOVERY PROPORTION

From now on, assume that among all the $p$ null hypotheses, $p_0$ of them are true and $p_1$ hypotheses ($p_1 = p - p_0$) are false, and $p_1$ is supposed to be very small compared to $p$. For ease of presentation, we let $p$ be the sole asymptotic parameter, and assume $p_0 \to \infty$ when $p \to \infty$. For a fixed rejection threshold $t$, we will reject those $p$-values no greater than $t$ and regard them as statistically significant. Because of its powerful applicability, this procedure has been widely adopted; see, for example, Storey (2002), Efron (2007, 2010), among others. Our goal is to estimate the realized FDP for a given $t$ in a multiple testing problem (5) based on the observations (4) under an arbitrary dependence structure of $\mathbf{\Sigma}$. Our methods and results have direct implications on the situation in which $\mathbf{\Sigma}$ is unknown, the setting studied by Efron (2007, 2010) and Friguet, Kloareg, and Causeur (2009). In the latter case, $\mathbf{\Sigma}$ needs to be estimated with certain accuracy.

### 3.1 Approximation of FDP

Define the following empirical processes:

$$\begin{aligned} V(t) &= \#\{\text{true null } P_i : P_i \le t\}, \\ S(t) &= \#\{\text{false null } P_i : P_i \le t\}, \quad \text{and} \\ R(t) &= \#\{P_i : P_i \le t\}, \end{aligned}$$

where $t \in [0, 1]$. Then, $V(t)$, $S(t)$, and $R(t)$ are the number of false discoveries, the number of true discoveries, and the number of total discoveries, respectively. Obviously, $R(t) = V(t) + S(t)$, and $V(t)$, $S(t)$, and $R(t)$ are all random variables, depending on the test statistics $(Z_1, \ldots, Z_p)^T$. Moreover, $R(t)$ is observed in an experiment, but $V(t)$ and $S(t)$ are both unobserved.

By definition, $\mathrm{FDP}(t) = V(t)/R(t)$ and $\mathrm{FDR}(t) = E[V(t)/R(t)]$. One of the interests is to control $\mathrm{FDR}(t)$ at a predetermined rate $\alpha$, say 15%. There are also substantial research interests in the statistical behavior of the number of false discoveries $V(t)$ and the false discovery proportion $\mathrm{FDP}(t)$, which are unknown but realized given an experiment. One may even argue that controlling FDP is more relevant, since it is directly related to the current experiment.

We now approximate $V(t)/R(t)$ for the high-dimensional sparse case $p_1 \ll p$. Suppose $(Z_1, \ldots, Z_p)^T \sim N((\mu_1, \ldots, \mu_p)^T, \mathbf{\Sigma})$ in which $\mathbf{\Sigma}$ is a correlation matrix. We need the following definition for weakly dependent normal random variables; other definitions can be found in Farcomeni (2007).

*Definition 1.* Suppose $(K_1, \ldots, K_p)^T \sim N((\theta_1, \ldots, \theta_p)^T, \mathbf{A})$. Then, $K_1, \ldots, K_p$ are called weakly dependent normal variables, if

$$\lim_{p\to\infty} p^{-2} \sum_{i,j} |a_{ij}| = 0, \tag{6}$$

where $a_{ij}$ denote the $(i, j)$th element of the covariance matrix $\mathbf{A}$.

Our procedure is called PFA. The basic idea is to decompose any dependent normal random vector as a factor model with weakly dependent normal random errors. The details are shown as follows. First, apply the spectral decomposition to the covariance matrix $\mathbf{\Sigma}$. Suppose the eigenvalues of $\mathbf{\Sigma}$ are $\lambda_1, \ldots, \lambda_p$ that have been arranged in decreasing order. If the corresponding orthonormal eigenvectors are denoted as $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$, then

$$\mathbf{\Sigma} = \sum_{i=1}^{p} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T. \tag{7}$$

If we further denote $\mathbf{A} = \sum_{i=k+1}^{p} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T$ for an integer $k$, then

$$\|\mathbf{A}\|_F^2 = \lambda_{k+1}^2 + \cdots + \lambda_p^2, \tag{8}$$

where $\|\cdot\|_F$ is the Frobenius norm. Let $\mathbf{L} = (\sqrt{\lambda_1}\boldsymbol{\gamma}_1, \ldots, \sqrt{\lambda_k}\boldsymbol{\gamma}_k)$, which is a $p \times k$ matrix. Then, the covariance matrix $\mathbf{\Sigma}$ can be expressed as

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{A}, \tag{9}$$

and $Z_1, \ldots, Z_p$ can be written as

$$Z_i = \mu_i + \mathbf{b}_i^T \mathbf{W} + K_i, \qquad i = 1, \ldots, p, \tag{10}$$

where $\mathbf{b}_i = (b_{i1}, \ldots, b_{ik})^T$, $(b_{1j}, \ldots, b_{pj})^T = \sqrt{\lambda_j}\boldsymbol{\gamma}_j$, the factors are $\mathbf{W} = (W_1, \ldots, W_k)^T \sim N_k(0, \mathbf{I}_k)$, and the random errors are $(K_1, \ldots, K_p)^T \sim N(0, \mathbf{A})$. Furthermore, $W_1, \ldots, W_k$ are independent of each other and independent of $K_1, \ldots, K_p$. Changing a probability if necessary, we can assume that (10) is the data-generation process. In expression (10), $\{\mu_i = 0\}$ correspond to the true null hypotheses, while $\{\mu_i \ne 0\}$ correspond to the false ones. Note that although (10) is not exactly a classical multifactor model because of the existence of dependence among $K_1, \ldots, K_p$, we can nevertheless show that $(K_1, \ldots, K_p)^T$ is a weakly dependent vector if the number of factors $k$ is appropriately chosen.

We now discuss how to choose $k$ such that $(K_1, \ldots, K_p)^T$ is weakly dependent. Denote by $a_{ij}$ the $(i, j)$th element in the covariance matrix $\mathbf{A}$. If we have

$$p^{-1}\left(\lambda_{k+1}^2 + \cdots + \lambda_p^2\right)^{1/2} \longrightarrow 0 \text{ as } p \to \infty, \tag{11}$$

then by the Cauchy–Schwartz inequality

$$\begin{aligned} p^{-2} \sum_{i,j} |a_{ij}| &\le p^{-1}\|\mathbf{A}\|_F \\ &= p^{-1}\left(\lambda_{k+1}^2 + \cdots + \lambda_p^2\right)^{1/2} \longrightarrow 0 \text{ as } p \to \infty. \end{aligned}$$

Note that $\sum_{i=1}^{p} \lambda_i = \mathrm{tr}(\mathbf{\Sigma}) = p$, so that (11) is self-normalized. Note also that the left-hand side of (11) is bounded by $p^{-1/2}\lambda_{k+1}$, which tends to zero whenever $\lambda_{k+1} = o(p^{1/2})$. Therefore, we can assume that $\lambda_k > cp^{1/2}$ for some $c > 0$. In particular, if $\lambda_1 = o(p^{1/2})$, the matrix $\mathbf{\Sigma}$ is weakly dependent and $k = 0$. In practice, we always choose the smallest $k$ such that

$$\frac{\sqrt{\lambda_{k+1}^2 + \cdots + \lambda_p^2}}{\lambda_1 + \cdots + \lambda_p} < \varepsilon$$

holds for a predetermined small $\varepsilon$, say, 0.01.

*Theorem 1.* Suppose $(Z_1, \ldots, Z_p)^T \sim N((\mu_1, \ldots, \mu_p)^T, \mathbf{\Sigma})$. Choose an appropriate $k$ such that

$$\text{(C0)} \qquad \frac{\sqrt{\lambda_{k+1}^2 + \cdots + \lambda_p^2}}{\lambda_1 + \cdots + \lambda_p} = O(p^{-\delta}) \quad \text{for } \delta > 0.$$

Let $\sqrt{\lambda_j}\boldsymbol{\gamma}_j = (b_{1j}, \ldots, b_{pj})^T$ for $j = 1, \ldots, k$. Then,

$$\lim_{p\to\infty}\left\{\mathrm{FDP}(t) - \frac{\sum_{i\in\{\text{true null}\}}[\Phi(a_i(z_{t/2}+\eta_i))+\Phi(a_i(z_{t/2}-\eta_i))]}{\sum_{i=1}^{p}[\Phi(a_i(z_{t/2}+\eta_i+\mu_i))+\Phi(a_i(z_{t/2}-\eta_i-\mu_i))]}\right\} = 0 \text{ a.s.}, \tag{12}$$

where $a_i = (1-\sum_{h=1}^{k} b_{ih}^2)^{-1/2}$, $\eta_i = \mathbf{b}_i^T\mathbf{W}$ with $\mathbf{b}_i = (b_{i1}, \ldots, b_{ik})^T$ and $\mathbf{W} \sim N_k(0, \mathbf{I}_k)$ in (10), and $\Phi(\cdot)$ and $z_{t/2} = \Phi^{-1}(t/2)$ are the cumulative distribution function and the $t/2$ lower quantile of a standard normal distribution, respectively.

Note that condition (C0) implies that $K_1, \ldots, K_p$ are weakly dependent random variables, but (11) converges to zero at some polynomial rate of $p$.

Theorem 1 gives an asymptotic approximation for FDP($t$) under general dependence structure. To the best of our knowledge, it is the first result to explicitly spell out the impact of dependence. It is also closely connected with the existing results for independence case and weak dependence case. Let $b_{ih} = 0$ for $i = 1, \ldots, p$ and $h = 1, \ldots, k$ in (10) and $K_1, \ldots, K_p$ be weakly dependent or independent normal random variables, then it reduces to the weak dependence case or independence case, respectively. In the above two specific cases, the numerator of (12) is just $p_0 t$. Storey (2002) used an estimate for $p_0$, resulting in an estimator of FDP($t$) as $\widehat{p}_0 t/R(t)$. This estimator has been shown to control the FDR under independency and weak dependency. However, for general dependency, Storey's procedure will not work well because it ignores the correlation effect among the test statistics, as shown by (12). Further discussions for the relationship between our result and the other leading research for multiple testing under dependence are shown in Section 3.4.

The results in Theorem 1 can be better understood by some special dependence structures as follows. These specific cases are also considered by Roquain and Villers (2011), Finner, Dickhaus, and Roters (2007), and Friguet, Kloareg, and Causeur (2009) under somewhat different settings.

*Example 1.* **[Equal Correlation]** If $\boldsymbol{\Sigma}$ has $\rho_{ij} = \rho \in [0, 1)$ for $i \neq j$, then we can write

$$Z_i = \mu_i + \sqrt{\rho}W + \sqrt{1-\rho}K_i \quad i = 1, \ldots, p,$$

where $W \sim N(0, 1)$, $K_i \sim N(0, 1)$, and $W$ and all $K_i$'s are independent of each other. Thus, we have

$$\lim_{p\to\infty}\left[\mathrm{FDP}(t) - \frac{p_0[\Phi(d(z_{t/2}+\sqrt{\rho}W))+\Phi(d(z_{t/2}-\sqrt{\rho}W))]}{\sum_{i=1}^{p}[\Phi(d(z_{t/2}+\sqrt{\rho}W+\mu_i))+\Phi(d(z_{t/2}-\sqrt{\rho}W-\mu_i))]}\right] = 0 \text{ a.s.},$$

where $d = (1-\rho)^{-1/2}$.

Note that Delattre and Roquain (2011) studied the FDP in a particular case of equal correlation. They provided a slightly different decomposition of $\{Z_i\}_{i=1}^p$ in the proof of Lemma 3.3 where the errors $K_i$'s have a sum equal to 0. Finner, Dickhaus, and Roters (2007) in their Theorem 2.1 also showed a result similar to Theorem 1 for the equal correlation case.

*Example 2.* **[Multifactor Model]** Consider a multifactor model

$$Z_i = \mu_i + \eta_i + a_i^{-1}K_i, \qquad i = 1, \ldots, p, \tag{13}$$

where $\eta_i$ and $a_i$ are defined in Theorem 1 and $K_i \sim N(0, 1)$ for $i = 1, \ldots, p$. All the $W_h$'s and $K_i$'s are independent of each other. In this model, $W_1, \ldots, W_k$ are the $k$ common factors. By Theorem 1, expression (12) holds.

Note that the covariance matrix for model (13) is

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathrm{diag}\big(a_1^{-2}, \ldots, a_p^{-2}\big). \tag{14}$$

When $\{a_j\}$ is not a constant, columns of $L$ are not necessarily eigenvectors of $\boldsymbol{\Sigma}$. In other words, when the principal component analysis is used, the decomposition of (9) can yield a different $L$ and condition (11) can require a different value of $k$. In this sense, there is a subtle difference between our approach and that in Friguet, Kloareg, and Causeur (2009) when the principal component analysis is used. Theorem 1 should be understood as a result for any decomposition (9) that satisfies condition (C0). Because we use principal components as approximated factors, our procedure is called PFA. In practice, if one knows that the test statistics come from a factor model structure (13), a multiple testing procedure based on the factor analysis (14) is preferable, since formula (12) becomes exact. In this case, the matrices $\mathbf{L}$ from the principal analysis (9) and from the factor analysis (14) are approximately the same when $p$ is large, under some mild conditions. On the other hand, when such a factor structure is not granted, the principal component analysis is more advantageous, due in part, to computational expediency and theoretical support (Theorem 1).

In Theorem 1, since FDP is bounded by 1, taking expectation on both sides of Equation (12) and by the Portmanteau lemma, we have the convergence of FDR:

*Corollary 1.* Under the assumptions in Theorem 1,

$$\lim_{p\to\infty}\left\{\mathrm{FDR}(t) - E\left[\frac{\sum_{i\in\{\text{true null}\}}\{\Phi(a_i(z_{t/2}+\eta_i))+\Phi(a_i(z_{t/2}-\eta_i))\}}{\sum_{i=1}^{p}\{\Phi(a_i(z_{t/2}+\eta_i+\mu_i))+\Phi(a_i(z_{t/2}-\eta_i-\mu_i))\}}\right]\right\} = 0. \tag{15}$$

The expectation on the left-hand side of (15) is with respect to standard multivariate normal variables $(W_1, \ldots, W_k)^T \sim N_k(0, \mathbf{I}_k)$.

The proof of Theorem 1 is based on the following result.

*Proposition 2.* Under the assumptions in Theorem 1,

$$\lim_{p\to\infty}\left[p^{-1}R(t) - p^{-1}\sum_{i=1}^{p}[\Phi(a_i(z_{t/2}+\eta_i+\mu_i)) + \Phi(a_i(z_{t/2}-\eta_i-\mu_i))]\right] = 0 \text{ a.s.}, \tag{16}$$

$$\lim_{p\to\infty}\Bigg[p_0^{-1}V(t)-p_0^{-1}\sum_{i\in\{\text{true null}\}}[\Phi(a_i(z_{t/2}+\eta_i))+\Phi(a_i(z_{t/2}-\eta_i))]\Bigg]=0 \text{ a.s.} \tag{17}$$

The proofs of Theorem 1 and Proposition 2 are given in the Appendix.

### 3.2 Estimating Realized FDP

In Theorem 1 and Proposition 2, the summation over the set of true null hypotheses is unknown. However, due to the high dimensionality and sparsity, both $p$ and $p_0$ are large and $p_1$ is relatively small. Therefore, we can use

$$\sum_{i=1}^{p}[\Phi(a_i(z_{t/2}+\eta_i))+\Phi(a_i(z_{t/2}-\eta_i))] \tag{18}$$

as a conservative surrogate for

$$\sum_{i\in\{\text{true null}\}}[\Phi(a_i(z_{t/2}+\eta_i))+\Phi(a_i(z_{t/2}-\eta_i))]. \tag{19}$$

Since only $p_1$ extra terms are included in (18), the substitution is accurate enough for many applications.

Recall that $\mathrm{FDP}(t)=V(t)/R(t)$, in which $R(t)$ is observable and known. Thus, only the realization of $V(t)$ is unknown. The mean of $V(t)$ is $E[\sum_{i\in\{\text{true null}\}}I(P_i\le t)]=p_0t$, since the $p$-values corresponding to the true null hypotheses are uniformly distributed. However, the dependence structure affects the variance of $V(t)$, which can be much larger than the binomial formula $p_0t(1-t)$. Owen (2005) has theoretically studied the variance of the number of false discoveries. In our framework, expression (18) is a function of iid standard normal variables. Given $t$, the variance of (18) can be obtained by simulations and hence variance of $V(t)$ is approximated via (18). Relevant simulation studies will be presented in Section 5.

In recent years, there has been substantial interest in the realized random variable FDP itself in a given experiment, instead of controlling FDR, as we are usually concerned about the number of false discoveries given the observed sample of test statistics, rather than an average of FDP for hypothetical replications of the experiment. See Genovese and Wasserman (2004), Meinshausen (2006), Efron (2007), Friguet, Kloareg, and Causeur (2009), etc. In our problem, by Proposition 2, it is known that $V(t)$ is approximately

$$\sum_{i=1}^{p}[\Phi(a_i(z_{t/2}+\eta_i))+\Phi(a_i(z_{t/2}-\eta_i))]. \tag{20}$$

Let

$$\mathrm{FDP_A}(t)=\Bigg(\sum_{i=1}^{p}[\Phi(a_i(z_{t/2}+\eta_i))+\Phi(a_i(z_{t/2}-\eta_i))]\Bigg)/R(t),$$

if $R(t)\neq 0$ and $\mathrm{FDP_A}(t)=0$ when $R(t)=0$. Given observations $z_1,\dots,z_p$ of the test statistics $Z_1,\dots,Z_p$, if the unobserved but realized factors $W_1,\dots,W_k$ can be estimated by $\widehat{W}_1,\dots,\widehat{W}_k$, then we can obtain an estimator of $\mathrm{FDP_A}(t)$ by

$$\widehat{\mathrm{FDP}}(t)=\min\Bigg(\sum_{i=1}^{p}[\Phi(a_i(z_{t/2}+\widehat{\eta}_i))+\Phi(a_i(z_{t/2}-\widehat{\eta}_i))],\,R(t)\Bigg)/R(t), \tag{21}$$

when $R(t)\neq 0$ and $\widehat{\mathrm{FDP}}(t)=0$ when $R(t)=0$. Note that in (21), $\widehat{\eta}_i=\sum_{h=1}^{k}b_{ih}\widehat{W}_h$ is an estimator for $\eta_i=\mathbf{b}_i^T\mathbf{W}$.

The following procedure is one practical way to estimate $\mathbf{W}=(W_1,\dots,W_k)^T$ based on the data. For observed values $z_1,\dots,z_p$, we choose the smallest 90% of $|z_i|$'s, say. For ease of notation, assume the first $m$ $z_i$'s have the smallest absolute values. Then, approximately

$$Z_i=\mathbf{b}_i^T\mathbf{W}+K_i,\quad i=1,\dots,m. \tag{22}$$

The approximation from (10) to (22) stems from the intuition that large $|\mu_i|$'s tend to produce large $|z_i|$'s as $Z_i\sim N(\mu_i,1)$ $1\le i\le p$ and the sparsity makes approximation errors negligible. Finally, we apply the robust $L_1$-regression to the equation set (22) and obtain the least-absolute deviation estimates $\widehat{W}_1,\dots,\widehat{W}_k$. We use $L_1$-regression rather than $L_2$-regression because there might be nonzero $\mu_i$ involved in Equation (22) and $L_1$ is more robust to the outliers than $L_2$. Other possible methods include using penalized method such as least absolute shrinkage and selection operator (LASSO) or smoothly clipped absolute deviation (SCAD) to explore the sparsity. For example, one can minimize

$$\sum_{i=1}^{p}\left(Z_i-\mu_i-\mathbf{b}_i^T\mathbf{W}\right)^2+\sum_{i=1}^{p}p_\lambda(|\mu_i|)$$

with respect to $\{\mu_i\}_{i=1}^p$ and $\mathbf{W}$, where $p_\lambda(\cdot)$ is a folded-concave penalty function (Fan and Li 2001).

The estimator (21) performs significantly better than Efron's (2007) estimator in our simulation studies. One difference is that in our setting $\mathbf{\Sigma}$ is known. The other is that we give a better approximation as shown in Section 3.4.

Efron (2007) proposed the concept of conditional FDR. Consider $E(V(t))/R(t)$ as one type of FDR definitions (see Efron 2007, expression (46)). The numerator $E(V(t))$ represents over-replications of the experiment, and equals a constant $p_0t$. But if the actual correlation structure in a given experiment is taken into consideration, then Efron (2007) defines the conditional FDR as $E(V(t)|A)/R(t)$, where $A$ is a random variable that measures the dependency information of the test statistics. Estimating the realized value of $A$ in a given experiment by $\widehat{A}$, one can have the estimated conditional FDR as $E(V(t)|\widehat{A})/R(t)$. Following Efron's proposition, Friguet, Kloareg, and Causeur (2009) gave the estimated conditional FDR by $E(V(t)|\widehat{\mathbf{W}})/R(t)$, where $\widehat{\mathbf{W}}$ is an estimate of the realized random factors $\mathbf{W}$ in a given experiment.

Our estimator in (21) for the realized FDP in a given experiment can be understood as an estimate of conditional FDR. Note that (19) is actually $E(V(t)|\{\eta_i\}_{i=1}^p)$. By Proposition 2, we can approximate $V(t)$ by $E(V(t)|\{\eta_i\}_{i=1}^p)$. Thus, the estimate of conditional FDR $E(V(t)|\{\widehat{\eta}_i\}_{i=1}^p)/R(t)$ is directly an estimate of the realized FDP $V(t)/R(t)$ in a given experiment.

### 3.3 Asymptotic Justification

Let $\mathbf{w} = (w_1, \ldots, w_k)^T$ be the realized values of $\{W_h\}_{h=1}^k$, and $\widehat{\mathbf{w}}$ be an estimator for $\mathbf{w}$. We now show in Theorem 2 that $\widehat{\text{FDP}}(t)$ in (21) based on a consistent estimator $\widehat{\mathbf{w}}$ has the same convergence rate as $\widehat{\mathbf{w}}$ under some mild conditions.

*Theorem 2.* If the following conditions are satisfied:

(C1) $R(t)/p > H$ for $H > 0$ as $p \to \infty$,
(C2) $\min_{1\le i\le p} \min(|z_{t/2} + \mathbf{b}_i^T\mathbf{w}|, |z_{t/2} - \mathbf{b}_i^T\mathbf{w}|) \ge \tau > 0$,
(C3) $\|\widehat{\mathbf{w}} - \mathbf{w}\|_2 = O_p(p^{-r})$ for some $r > 0$,

then $|\widehat{\text{FDP}}(t) - \text{FDP}_\text{A}(t)| = O(\|\widehat{\mathbf{w}} - \mathbf{w}\|_2)$.

In Theorem 2, (C2) is a reasonable condition because $z_{t/2}$ is a large negative number when threshold $t$ is small and $\mathbf{b}_i^T\mathbf{w}$ is a realization from a normal distribution $N(0, \sum_{h=1}^k b_{ih}^2)$ with $\sum_{h=1}^k b_{ih}^2 < 1$. Thus, $z_{t/2} + \mathbf{b}_i^T\mathbf{w}$ or $z_{t/2} - \mathbf{b}_i^T\mathbf{w}$ is unlikely close to zero.

Theorem 3 shows the asymptotic consistency of $L_1$-regression estimators under model (22). Portnoy (1984b) proved the asymptotic consistency for robust regression estimation when the random errors are iid. However, that proof does not work here because of the weak dependence of random errors. Our result allows $k$ to grow with $m$, even at a faster rate of $o(m^{1/4})$ imposed by Portnoy (1984b).

*Theorem 3.* Suppose (22) is a correct model. Let $\widehat{\mathbf{w}}$ be the $L_1$-regression estimator:

$$\widehat{\mathbf{w}} \equiv \text{argmin}_{\boldsymbol{\beta}\in R^k} \sum_{i=1}^m \big|Z_i - \mathbf{b}_i^T\boldsymbol{\beta}\big|, \tag{23}$$

where $\mathbf{b}_i = (b_{i1}, \ldots, b_{ik})^T$. Let $\mathbf{w} = (w_1, \ldots, w_k)^T$ be the realized values of $\{W_h\}_{h=1}^k$. Suppose $k = O(m^\kappa)$ for $0 \le \kappa < 1 - \delta$. Under the assumptions

(C4) $\sum_{j=k+1}^p \lambda_j^2 \le \eta$ for $\eta = O(m^{2\kappa})$,
(C5)

$$\lim_{m\to\infty} \sup_{\|\mathbf{u}\|=1} m^{-1} \sum_{i=1}^m I\big(\big|\mathbf{b}_i^T\mathbf{u}\big| \le d\big) = 0$$

for a constant $d > 0$,
(C6) $a_{\max}/a_{\min} \le S$ for some constant $S$ when $m \to \infty$, where $1/a_i$ is the standard deviation of $K_i$,
(C7) $a_{\min} = O(m^{(1-\kappa)/2})$.

We have $\|\widehat{\mathbf{w}} - \mathbf{w}\|_2 = O_p(\sqrt{\frac{k}{m}})$.

(C4) is stronger than (C0) in Theorem 1 as (C0) only requires $\sum_{j=k+1}^p \lambda_j^2 = O(p^{2-2\delta})$. (C5) ensures the identifiability of $\boldsymbol{\beta}$, which is similar to Proposition 3.3 in Portnoy (1984a). (C6) and (C7) are imposed to facilitate the technical proof.

We now briefly discuss the role of the factor $k$. To make the approximation in Theorem 1 hold, we need $k$ to be large. On the other hand, to make the realized factors estimable with reasonable accuracy, we hope to choose a small $k$ as demonstrated in Theorem 3. Thus, the practical choice of $k$ should be done with care.

Since $m$ is chosen as a certain large proportion of $p$, combination of Theorem 2 and Theorem 3 thus shows the asymptotic consistency of $\widehat{\text{FDP}}(t)$ based on $L_1$-regression estimator of $\mathbf{w} = (w_1, \ldots, w_k)^T$ in model (22):

$$|\widehat{\text{FDP}}(t) - \text{FDP}_\text{A}(t)| = O_p\left(\sqrt{\frac{k}{m}}\right).$$

The result in Theorem 3 is based on the assumption that (22) is a correct model. In the following, we will show that even if (22) is not a correct model, the effects of misspecification are negligible when $p$ is sufficiently large. To facilitate the mathematical derivations, we instead consider the least-square estimator. Suppose we are estimating $\mathbf{W} = (W_1, \ldots, W_k)^T$ from (10). Let $\mathbf{X}$ be the design matrix of model (10). Then, the least-square estimator for $\mathbf{W}$ is $\widehat{\mathbf{W}}_\text{LS} = \mathbf{W} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{\mu} + \mathbf{K})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and $\mathbf{K} = (K_1, \ldots, K_p)^T$. Instead, we estimate $W_1, \ldots, W_k$ based on the simplified model (22), which ignores sparse $\{\mu_i\}$. Then, the least-square estimator for $\mathbf{W}$ is $\widehat{\mathbf{W}}_\text{LS}^* = \mathbf{W} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{K} = \mathbf{W}$, in which we use the orthogonality between $\mathbf{X}$ and var($\mathbf{K}$). The following result shows that the effect of misspecification in model (22) is negligible when $p \to \infty$ and when the least-square estimator is consistent.

*Theorem 4.* The bias due to ignoring nonnulls is controlled by

$$\|\widehat{\mathbf{W}}_\text{LS} - \mathbf{W}\|_2 = \|\widehat{\mathbf{W}}_\text{LS} - \widehat{\mathbf{W}}_\text{LS}^*\|_2 \le \|\boldsymbol{\mu}\|_2\left(\sum_{i=1}^k \lambda_i^{-1}\right)^{1/2}.$$

In Theorem 1, we can choose appropriate $k$ such that $\lambda_k > cp^{1/2}$ as noted in Theorem 1. Therefore, $\sum_{i=1}^k \lambda_i^{-1} \to 0$ as $p \to \infty$ is a reasonable condition. When $\{\mu_i\}_{i=1}^p$ are truly sparse, it is expected that $\|\boldsymbol{\mu}\|_2$ grows slowly or is even bounded so that the bound in Theorem 4 is small. For $L_1$-regression, it is expected to be even more robust to the outliers in the sparse vector $\{\mu_i\}_{i=1}^p$.

### 3.4 Dependence-Adjusted Procedure

A problem of the method used thus far is that the ranking of statistical significance is completely determined by the ranking of the test statistics $\{|Z_i|\}$. This is undesirable and can be inefficient for the dependent case: the correlation structure should also be taken into account. We now show how to use the correlation structure to improve the signal-to-noise ratio.

Note that by (10), $Z_i - \mathbf{b}_i^T\mathbf{W} \sim N(\mu_i, a_i^{-2})$, where $a_i$ is defined in Theorem 1. Since $a_i^{-1} \le 1$, the signal-to-noise ratio increases, which makes the resulting procedure more powerful. Thus, if we know the true values of the common factors $\mathbf{W} = (W_1, \ldots, W_k)^T$, we can use $a_i(Z_i - \mathbf{b}_i^T\mathbf{W})$ as the test statistics. The dependence-adjusted $p$-values $2\Phi(-|a_i(Z_i - \mathbf{b}_i^T\mathbf{W})|)$ can then be used. Note that this testing procedure has different thresholds for different hypotheses based on the magnitude of $Z_i$, and has incorporated the correlation information among hypotheses. In practice, given $Z_i$, the common factors $\{W_h\}_{h=1}^k$ are realized but are unobservable. As shown in Section 3.2, they can be estimated. The dependence-adjusted $p$-values are then given by

$$2\Phi\big(-\big|a_i\big(Z_i - \mathbf{b}_i^T\widehat{\mathbf{W}}\big)\big|\big) \tag{24}$$

for ranking the hypotheses, where $\widehat{\mathbf{W}} = (\widehat{W}_1, \ldots, \widehat{W}_k)^T$ is an estimate of the principal factors. We will show in Section 5

by simulation studies that this dependence-adjusted procedure is more powerful. The "factor-adjusted multiple testing procedure" in Friguet, Kloareg, and Causeur (2009) shares a similar idea.

### 3.5 Relation With Other Methods

Efron (2007) proposed a novel parametric model for $V(t)$,

$$V(t) = p_0 t\left[1 + 2A\frac{(-z_{t/2})\phi(z_{t/2})}{\sqrt{2}t}\right], \tag{25}$$

where $A \sim N(0, \alpha^2)$ for some real number $\alpha$ and $\phi(\cdot)$ stands for the probability density function of the standard normal distribution. The correlation effect is explained by the dispersion variate $A$. His procedure is to estimate $A$ from the data and use

$$p_0 t\left[1 + 2\widehat{A}\frac{(-z_{t/2})\phi(z_{t/2})}{\sqrt{2}t}\right]\Big/ R(t) \tag{26}$$

as an estimator for realized FDP($t$). Note that the above expressions are adaptations from his procedure for the one-sided test to our two-sided test setting. In his simulation, the above estimator captures the general trend of the FDP, but it is not accurate and deviates from the true FDP with large amount of variability. Consider our estimator $\widehat{\mathrm{FDP}}(t)$ in (21). Write $\widehat{\eta}_i = \sigma_i Q_i$, where $Q_i \sim N(0, 1)$. When $\sigma_i \to 0$ for $\forall i \in \{\text{true null}\}$, by the second-order Taylor expansion,

$$\widehat{\mathrm{FDP}}(t) \approx \frac{p_0 t}{R(t)}\left[1 + \sum_{i\in\{\text{true null}\}} \sigma_i^2\big(Q_i^2 - 1\big)\frac{(-z_{t/2})\phi(z_{t/2})}{p_0 t}\right]. \tag{27}$$

By comparison with Efron's estimator, we can see that

$$\widehat{A} = \frac{1}{\sqrt{2}p_0}\sum_{i\in\{\text{true null}\}}\big[\widehat{\eta}_i^2 - E\big(\widehat{\eta}_i^2\big)\big]. \tag{28}$$

Thus, our method is more general.

Leek and Storey (2008) considered a general framework for modeling the dependence in multiple testing. Their idea is to model the dependence via a factor model and reduce the multiple testing problem from dependence to independence case via accounting the effects of common factors. They also provided a method of estimating the common factors. In contrast, our problem is different from Leek and Storey's and we estimate common factors from PFA and other methods. In addition, we provide the approximated FDP formula and its consistent estimate.

Friguet, Kloareg, and Causeur (2009) followed closely the framework of Leek and Storey (2008). They assumed that the data come directly from a multifactor model with independent random errors, and then used the expectation–maximization (EM) algorithm to estimate all the parameters in the model and obtained an estimator for FDP($t$). In particular, they subtracted $\eta_i$ out of (13) based on their estimate from the EM algorithm to improve the efficiency. However, the ratio of estimated number of factors to the true number of factors in their studies varies according to the dependence structures by their EM algorithm, thus leading to inaccurate estimated FDP($t$). Moreover, it is hard to derive theoretical results based on the estimator from their EM algorithm. Compared with their results, our procedure does not assume any specific dependence structure of the test statistics. What we do is to decompose the test statistics into an approximate factor model with weakly dependent errors, derive the factor loadings, and estimate the unobserved but realized factors by $L_1$-regression. Since the theoretical distribution of $V(t)$ is known, estimator (21) performs well based on a good estimation for $W_1, \ldots, W_k$.

## 4. APPROXIMATE ESTIMATION OF FDR

In this section, we propose some ideas that can asymptotically control the FDR, not the FDP, under arbitrary dependency. Although their validity is yet to be established, promising results reveal in the simulation studies. Therefore, they are worth some discussion and serve as a direction of our future work.

Suppose that the number of false null hypotheses $p_1$ is known. If the signal $\mu_i$ for $i \in \{\text{false null}\}$ is strong enough such that

$$\Phi(a_i(z_{t/2} + \eta_i + \mu_i)) + \Phi(a_i(z_{t/2} - \eta_i - \mu_i)) \approx 1, \tag{29}$$

then asymptotically the FDR is approximately given by

$$\mathrm{FDR}(t) = E\left\{\frac{\sum_{i=1}^p[\Phi(a_i(z_{t/2} + \eta_i)) + \Phi(a_i(z_{t/2} - \eta_i))]}{\sum_{i=1}^p[\Phi(a_i(z_{t/2} + \eta_i)) + \Phi(a_i(z_{t/2} - \eta_i))] + p_1}\right\}, \tag{30}$$

which is the expectation of a function of $W_1, \ldots, W_k$. Note that FDR($t$) is a known function and can be computed by Monte Carlo simulation. For any predetermined error rate $\alpha$, we can use the bisection method to solve $t$ so that $\mathrm{FDR}(t) = \alpha$. Since $k$ is not large, the Monte Carlo computation is sufficiently fast for most applications.

The requirement (29) is not very strong. First of all, $\Phi(3) \approx 0.9987$, so (29) will hold if any number inside the $\Phi(\cdot)$ is greater than 3. Second, $1 - \sum_{h=1}^k b_{ih}^2$ is usually very small. For example, if it is 0.01, then $a_i = (1 - \sum_{h=1}^k b_{ih}^2)^{-1/2} \approx 10$, which means that if either $z_{t/2} + \eta_i + \mu_i$ or $z_{t/2} - \eta_i - \mu_i$ exceed 0.3, then (29) is approximately satisfied. Since the effect of sample size $n$ is involved in the problem in Section 2, (29) is not a very strong condition on the signal strength $\{\beta_i\}$.

Note that Finner, Dickhaus, and Roters (2007) considered a "Dirac uniform model," where the $p$-values corresponding to a false hypothesis are exactly equal to 0. This model might be potentially useful for FDR control. The calculation of (30) requires the knowledge of the proportion $p_1$ of signal in the data. Since $p_1$ is usually unknown in practice, there is also future research interest in estimating $p_1$ under arbitrary dependency.

## 5. SIMULATION STUDIES

In the simulation studies, we consider $p = 2000$, $n = 100$, $\sigma = 2$, the number of false null hypotheses $p_1 = 10$, and the nonzero $\beta_i = 1$, unless stated otherwise. We will present six different dependence structures for $\mathbf{\Sigma}$ of the test statistics $(Z_1, \ldots, Z_p)^T \sim N((\mu_1, \ldots, \mu_p)^T, \mathbf{\Sigma})$. Following the setting in Section 2, $\mathbf{\Sigma}$ is the correlation matrix of a random sample of size $n$ of $p$-dimensional vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$, and $\mu_j = \sqrt{n}\beta_j\widehat{\sigma}_j/\sigma$, $j = 1, \ldots, p$. The data-generating process vector $\mathbf{X}_i$'s are as follows.

- **[Equal correlation]** Let $\mathbf{X}^T = (X_1, \ldots, X_p)^T \sim N_p(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ has diagonal element 1 and off-diagonal element $1/2$.
- **[Fan & Song's model]** For $\mathbf{X} = (X_1, \ldots, X_p)$, let $\{X_k\}_{k=1}^{1900}$ be iid $N(0, 1)$ and

$$X_k = \sum_{l=1}^{10} X_l(-1)^{l+1}/5 + \sqrt{1 - \frac{10}{25}}\epsilon_k, \quad k = 1901, \ldots, 2000,$$

  where $\{\epsilon_k\}_{k=1901}^{2000}$ are standard normally distributed.
- **[Independent Cauchy]** For $\mathbf{X} = (X_1, \ldots, X_p)$, let $\{X_k\}_{k=1}^{2000}$ be iid. Cauchy random variables with location parameter 0 and scale parameter 1.
- **[Three factor model]** For $\mathbf{X} = (X_1, \ldots, X_p)$, let

$$X_j = \rho_j^{(1)} W^{(1)} + \rho_j^{(2)} W^{(2)} + \rho_j^{(3)} W^{(3)} + H_j,$$

  where $W^{(1)} \sim N(-2, 1)$, $W^{(2)} \sim N(1, 1)$, $W^{(3)} \sim N(4, 1)$, $\rho_j^{(1)}, \rho_j^{(2)}, \rho_j^{(3)}$ are iid $U(-1, 1)$, and $H_j$ are iid $N(0, 1)$.
- **[Two factor model]** For $\mathbf{X} = (X_1, \ldots, X_p)$, let

$$X_j = \rho_j^{(1)} W^{(1)} + \rho_j^{(2)} W^{(2)} + H_j,$$

  where $W^{(1)}$ and $W^{(2)}$ are iid $N(0, 1)$, $\rho_j^{(1)}$ and $\rho_j^{(2)}$ are iid $U(-1, 1)$, and $H_j$ are iid $N(0, 1)$.
- **[Nonlinear factor model]** For $\mathbf{X} = (X_1, \ldots, X_p)$, let

$$X_j = \sin\left(\rho_j^{(1)} W^{(1)}\right) + \operatorname{sgn}\left(\rho_j^{(2)}\right) \exp\left(\left|\rho_j^{(2)}\right| W^{(2)}\right) + H_j,$$

  where $W^{(1)}$ and $W^{(2)}$ are iid $N(0, 1)$, $\rho_j^{(1)}$ and $\rho_j^{(2)}$ are iid $U(-1, 1)$, and $H_j$ are iid $N(0, 1)$.

Fan and Song's model was considered by Fan and Song (2010) for high-dimensional variable selection. This model is close to the independent case but has some special dependence structure. Note that although we have used the term "factor model" above to describe the dependence structure, it is not the factor model for the test statistics $Z_1, \ldots, Z_p$ directly. The covariance matrix of these test statistics is the sample correlation matrix of $X_1, \ldots, X_p$.

We examine the effectiveness of our method in several aspects. We first examine the goodness of approximation in Theorem 1 by comparing the marginal distributions and variances. We then compare the accuracy of FDP estimates with other methods. Finally, we demonstrate the improvement of the power with dependence adjustment.

### 5.1 Distributions of FDP and Its Approximation

Without loss of generality, we consider a dependence structure based on the two-factor model above. Let $n = 100$, $p_1 = 10$, and $\sigma = 2$. Let $p$ vary from 100 to 1000 and $t$ be either 0.01 or 0.005. The distributions of FDP($t$) and its approximated expression in Theorem 1 are plotted in Figure 1. The convergences of the distributions are self-evidenced. Table 2 summarizes the total variation distance between the two distributions.

### 5.2 Variance of $V(t)$

Variance of false discoveries in the correlated test statistics is usually large compared with that of the independent case that is $p_0 t(1 - t)$, due to correlation structures. Thus, the ratio of variance of false discoveries in the dependent case to that in the independent test statistics can be considered as a measure of correlation effect (see Owen 2005). Estimating the variance of false discoveries is an interesting problem. With approximation (17), this can easily be computed. In Table 3, we compare the true

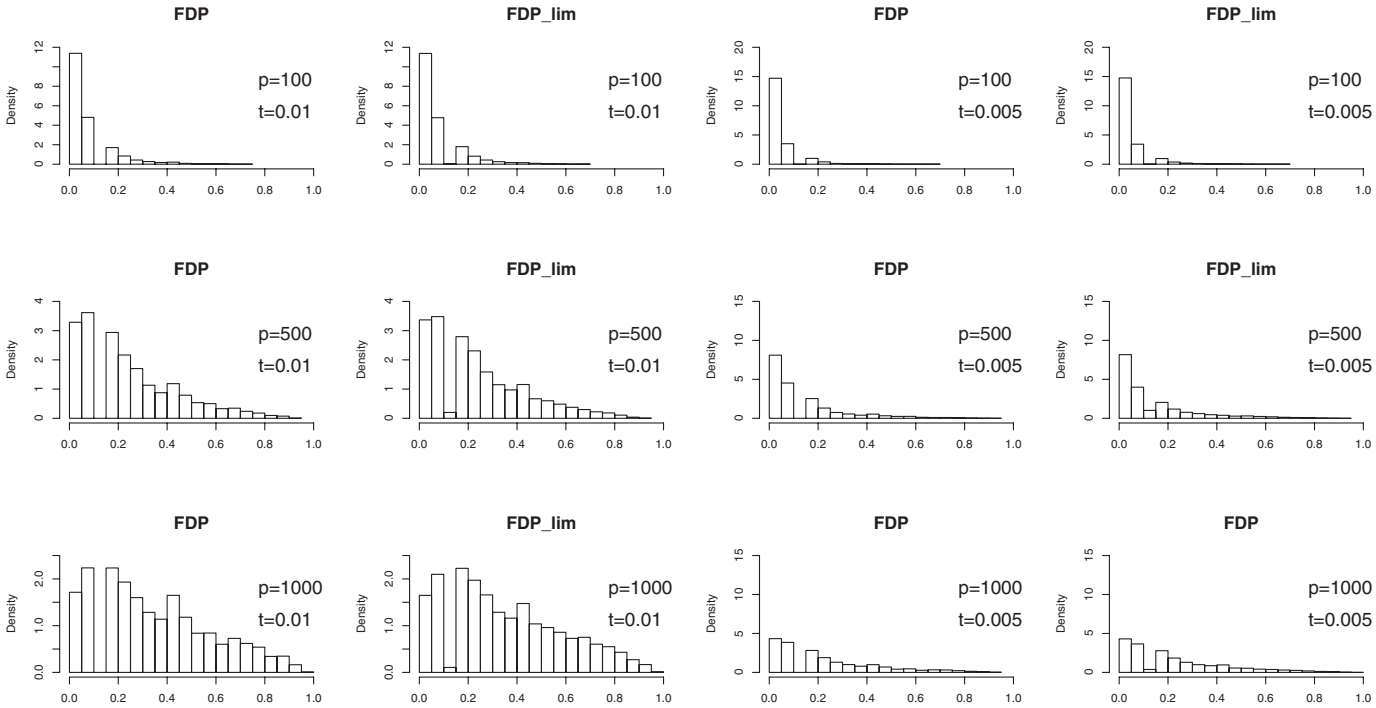


Figure 1. Comparison for the distribution of the FDP with that of its approximated expression, based on the two-factor model over 10,000 simulations. From the top row to the bottom, $p = 100, 500, 1000$, respectively. The first two columns correspond to $t = 0.01$ and the last two correspond to $t = 0.005$.

Table 2. Total variation distance between the distribution of FDP and the limiting distribution of FDP in Figure 1. The total variation distance is calculated based on "TotalVarDist" function with "smooth" option in R software

| | $p = 100$ | $p = 500$ | $p = 1000$ |
|---|---|---|---|
| $t = 0.01$ | 0.6668 | 0.1455 | 0.0679 |
| $t = 0.005$ | 0.6906 | 0.2792 | 0.1862 |

variance of the number of false discoveries, the variance of expression (19) (which is infeasible in practice), and the variance of expression (18) under six different dependence structures. It shows that the variance computed based on expression (18) approximately equals the variance of number of false discoveries. Therefore, we provide a fast and alternative method to estimate the variance of number of false discoveries in addition to the results in Owen (2005). Note that the variance for independent case is merely less than 2. The impact of dependence is very substantial.

### 5.3 Comparing Methods of Estimating FDP

Under different dependence structures, we compare FDP values using our procedure PFA in Equation (30) without taking expectation and with $p_1$ known, Storey's procedure with $p_1$ known ($(1 - p_1)t/R(t)$), and B–H procedure. Note that B–H procedure is an FDR control procedure rather than an FDP estimating procedure. The B–H FDP is obtained by using the mean of "True FDP" in Table 4 as the control rate in B–H procedure. Table 4 shows that our method performs much better than Storey's procedure and the B–H procedure, especially under strong dependence structures (rows 1, 4, 5, and 6), in terms of both mean and variance of the distribution of FDP. Recall that the expected value of FDP is the FDR. Table 3 also compares the FDR of three procedures by looking at the averages. Note that the actual FDR from B–H procedure under dependence is much smaller than the control rate, which suggests that B–H procedure can be quite conservative under dependence.

### 5.4 Comparison With Efron's Methods

We now compare the estimated values of our method PFA (21) and Efron's (2007) estimator with true values of FDP, under six different dependence structures. Efron's (2007) estimator is developed for estimating FDP under unknown $\mathbf{\Sigma}$. In our simulation study, we have used a known $\Sigma$ for Efron's estimator for fair comparisons. The results are depicted in Figures 2 and 3 and Table 5. Figure 2 shows that our estimated values correctly track the trends of FDP with smaller amount of noise. It also shows that both our estimator and Efron's estimator tend to overestimate the true FDP, since $\mathrm{FDP_A}(t)$ is an upper bound of the true FDP($t$). They are close only when the number of false nulls $p_1$ is very small. In the current simulation setting, we choose $p_1 = 50$ compared with $p = 1000$, therefore, it is not a very sparse case. However, even under this case, our estimator still performs very well for six different dependence structures. Efron's (2007) estimator is illustrated in Figure 2 with his suggestions for estimating parameters, which captures the general trend of true FDP but with large amount of noise. Figure 3 shows that the relative errors (REs) of PFA concentrate around 0, which suggests good accuracy of our method in estimating FDP. Table 5 summarizes the REs of the two methods.

Table 3. Comparison for variance of number of false discoveries (column 2), variance of expression (19) (column 3), and variance of expression (18) (column 4) with $t = 0.001$ based on 10,000 simulations

| Dependence structure | var($V(t)$) | var($V$) | var($V.up$) |
|---|---|---|---|
| Equal correlation | 180.9673 | 178.5939 | 180.6155 |
| Fan & Song's model | 5.2487 | 5.2032 | 5.2461 |
| Independent Cauchy | 9.0846 | 8.8182 | 8.9316 |
| Three-factor model | 81.1915 | 81.9373 | 83.0818 |
| Two-factor model | 53.9515 | 53.6883 | 54.0297 |
| Nonlinear factor model | 48.3414 | 48.7013 | 49.1645 |

Table 4. Comparison of FDP values for our method based on Equation (30) without taking expectation (PFA) with Storey's procedure and Benjamini–Hochberg's procedure under six different dependence structures, where $p = 2000$, $n = 200$, $t = 0.001$, and $\beta_i = 1$ for $i \in \{\text{falsenull}\}$. The computation is based on 10,000 simulations. The means of FDP are listed with the standard deviations in the brackets

| | True FDP | PFA | Storey | B–H |
|---|---|---|---|---|
| Equal correlation | 6.67% (15.87%) | 6.61% (15.88%) | 2.99% (10.53%) | 3.90% (14.58%) |
| Fan & Song's model | 14.85% (11.76%) | 14.85% (11.58%) | 13.27% (11.21%) | 14.46% (13.46%) |
| Independent Cauchy | 13.85% (13.60%) | 13.62% (13.15%) | 11.48% (12.39%) | 13.21% (15.40%) |
| Three-factor model | 8.08% (16.31%) | 8.29% (16.39%) | 4.00% (11.10%) | 5.46% (16.10%) |
| Two-factor model | 8.62% (16.44%) | 8.50% (16.27%) | 4.70% (11.97%) | 5.87% (16.55%) |
| Nonlinear factor model | 6.63% (15.56%) | 6.81% (15.94%) | 3.20% (10.91%) | 4.19% (15.31%) |

### 5.5 Dependence-Adjusted Procedure

We compare the dependence-adjusted procedure described in Section 3.4 with the testing procedure based only on the observed test statistics without using correlation information.

Table 5. Means and standard deviations of the relative error (RE) between true values of FDP and estimated FDP under the six dependence structures in Figure 2. $\mathrm{RE_P}$ and $\mathrm{RE_E}$ are the REs of our PFA estimator and Efron's (2007) estimator, respectively. RE is defined in Figure 3

| | $\mathrm{RE_P}$ | | $\mathrm{RE_E}$ | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Equal correlation | 0.0241 | 0.1262 | 1.4841 | 3.6736 |
| Fan & Song's model | 0.0689 | 0.1939 | 1.2521 | 1.9632 |
| Independent Cauchy | 0.0594 | 0.1736 | 1.3066 | 2.1864 |
| Three-factor model | 0.0421 | 0.1657 | 1.4504 | 2.6937 |
| Two-factor model | 0.0397 | 0.1323 | 1.1227 | 2.0912 |
| Nonlinear factor model | 0.0433 | 0.1648 | 1.3134 | 4.0254 |

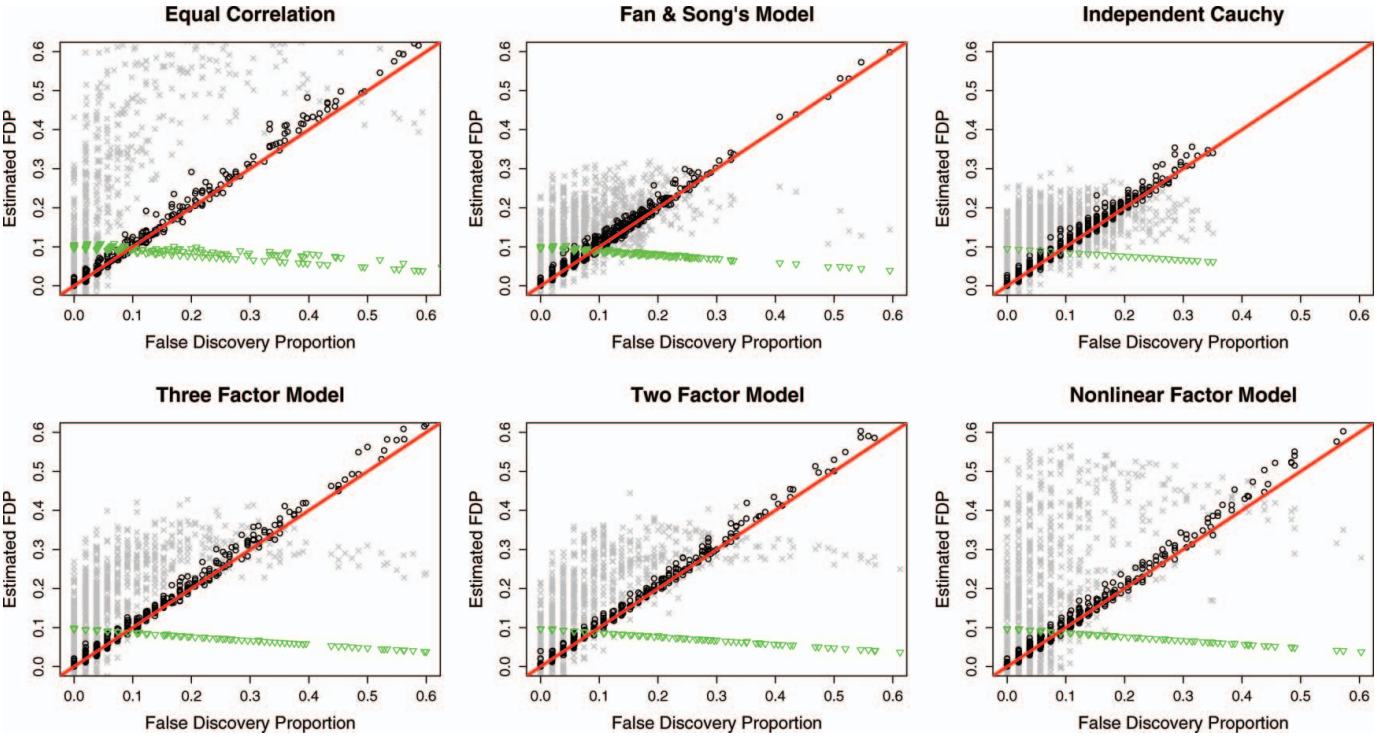


Figure 2. Comparison of true values of false discovery proportion (FDP) with estimated FDP by Efron's (2007) procedure (gray crosses) and our PFA method (black dots) under six different dependence structures, with $p = 1000$, $p_1 = 50$, $n = 100$, $\sigma = 2$, $t = 0.005$, and $\beta_i = 1$ for $i \in \{\text{falsenull}\}$ based on 1000 simulations. The $Z$-statistics with absolute value less than or equal to $x_0 = 1$ are used to estimate the dispersion variate $A$ in Efron's (2007) estimator. The unconditional estimate of FDR($t$) is $p_0 t / R(t)$ shown as green triangles. The online version of this figure is in color.

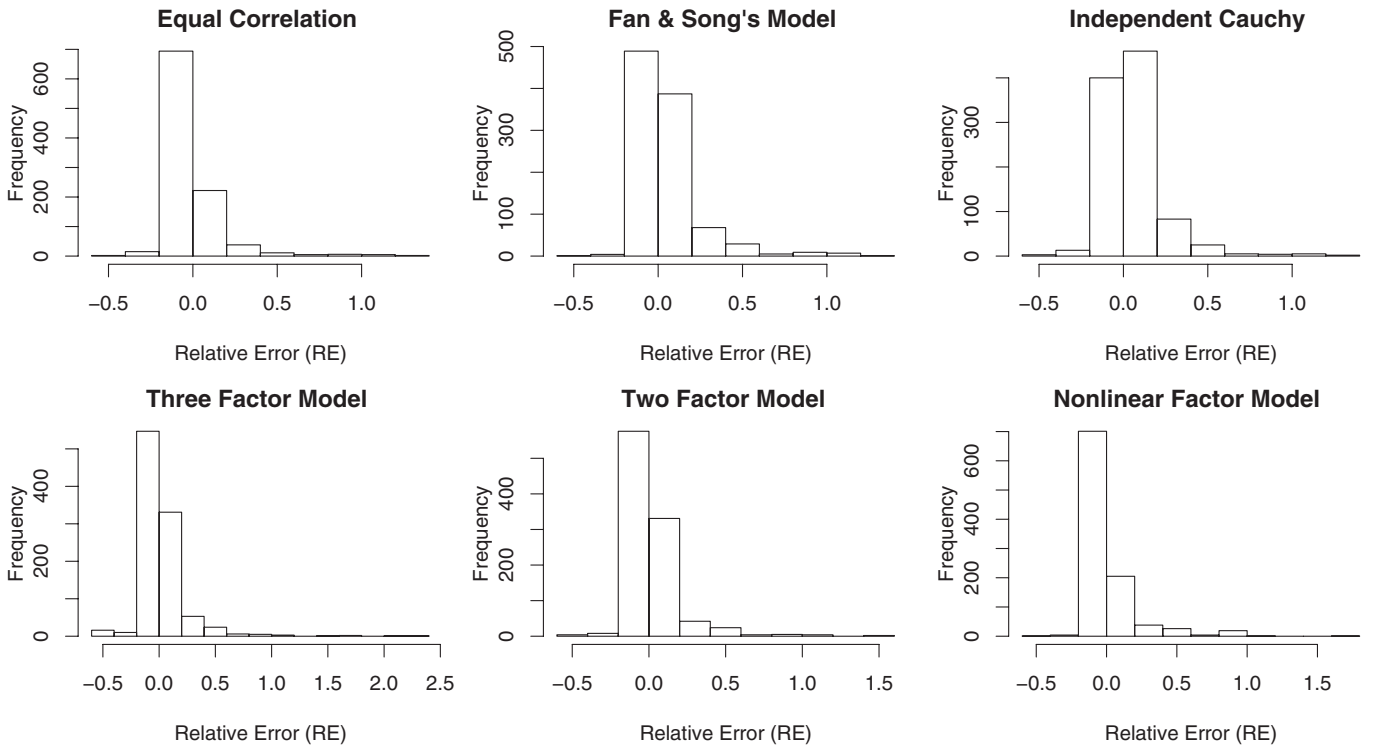


Figure 3. Histograms of the relative error (RE) between true values of FDP and estimated FDP by our PFA method under the six dependence structures in Figure 2. RE is defined as $(\widehat{\text{FDP}}(t) - \text{FDP}(t))/\text{FDP}(t)$ if $\text{FDP}(t) \neq 0$ and 0 otherwise.

Table 6. Comparison of dependence-adjusted procedure with fixed threshold procedure under six different dependence structures, where $p = 1000$, $n = 100$, $\sigma = 1$, $p_1 = 200$, nonzero $\beta_i$ simulated from $U(0, 1)$, and $k = n - 3$ over 1000 simulations

| | Fixed-threshold procedure | | | Dependence-adjusted procedure | | |
|---|---|---|---|---|---|---|
| | FDR | FNR | Threshold | FDR | FNR | Threshold |
| Equal correlation | 17.06% | 4.82% | 0.06 | 17.34% | 0.35% | 0.001 |
| Fan & Song's model | 6.69% | 6.32% | 0.0145 | 6.73% | 1.20% | 0.001 |
| Independent Cauchy | 7.12% | 0.45% | 0.019 | 7.12% | 0.13% | 0.001 |
| Three-factor model | 5.46% | 3.97% | 0.014 | 5.53% | 0.31% | 0.001 |
| Two-factor model | 5.00% | 4.60% | 0.012 | 5.05% | 0.39% | 0.001 |
| Nonlinear factor model | 6.42% | 3.73% | 0.019 | 6.38% | 0.68% | 0.001 |

The latter is to compare the original $Z$-statistics with a fixed threshold value and is labeled as "fixed threshold procedure" in Table 6. With the same FDR level, a procedure with smaller false nondiscovery rate (FNR) is more powerful, where FNR $= E[T/(p - R)]$ using the notation in Table 1.

In Table 6, without loss of generality, for each dependence structure we fix threshold value 0.001 and reject the hypotheses when the dependence-adjusted $p$-values (24) are smaller than 0.001. Then, we find the corresponding threshold value for the fixed threshold procedure such that the FDR in the two testing procedures are approximately the same. The FNR for the dependence-adjusted procedure is much smaller than that of the fixed threshold procedure, which suggests that the dependence-adjusted procedure is more powerful. Note that in Table 6, $p_1 = 200$ compared with $p = 1000$, implying that the better performance of the dependence-adjusted procedure is not limited to a sparse situation. This is expected since subtracting common factors out results in the problem having a higher signal-to-noise ratio.

## 6. REAL DATA ANALYSIS

Our proposed multiple testing procedures are now applied to the genome-wide association studies, in particular the eQTL mapping. It is known that the expression levels of gene CCT8 are highly related to Down Syndrome phenotypes. In our analysis, we use more than two million SNP genotype data and CCT8 gene expression data for 210 individuals from three different populations, testing which SNPs are associated with the variation in CCT8 expression levels. The SNP data are from the International HapMap project, which include 45 Japanese in Tokyo, Japan (JPT), 45 Han Chinese in Beijing, China (CHB), 60 Utah residents with ancestry from northern and western Europe (CEU), and 60 Yoruba in Ibadan, Nigeria (YRI). The Japanese and Chinese population are further grouped together to form the Asian population (JPTCHB). To save space, we omit the description of the data preprocessing procedures. Interested readers can find more details from the Web sites: *http://pngu.mgh.harvard.edu/purcell/plink/res.shtml* and *ftp://ftp.sanger.ac.uk/pub/genevar/*, and the article by Bradic, Fan, and Wang (2010).

We further introduce two sets of dummy variables $(\mathbf{d}_1, \mathbf{d}_2)$ to recode the SNP data, where $\mathbf{d}_1 = (d_{1,1}, \ldots, d_{1,p})$ and $\mathbf{d}_2 = (d_{2,1}, \ldots, d_{2,p})$, representing three categories of polymorphisms, namely, $(d_{1,j}, d_{2,j}) = (0, 0)$ for $\text{SNP}_j = 0$ (no polymorphism), $(d_{1,j}, d_{2,j}) = (1, 0)$ for $\text{SNP}_j = 1$ (one nucleotide has polymorphism), and $(d_{1,j}, d_{2,j}) = (0, 1)$ for $\text{SNP}_j = 2$ (both nucleotides have polymorphisms). Let $\{Y^i\}_{i=1}^n$ be the independent sample random variables of $Y$, $\{d_{1,j}^i\}_{i=1}^n$ and $\{d_{2,j}^i\}_{i=1}^n$ be the sample values of $d_{1,j}$ and $d_{2,j}$, respectively. Thus, instead of using model (1), we consider two marginal linear regression models between $\{Y^i\}_{i=1}^n$ and $\{d_{1,j}^i\}_{i=1}^n$:

$$\min_{\alpha_{1,j},\beta_{1,j}} \frac{1}{n} \sum_{i=1}^{n} E\big(Y^i - \alpha_{1,j} - \beta_{1,j} d_{1,j}^i\big)^2, \quad j = 1, \ldots, p \tag{31}$$

and between $\{Y^i\}_{i=1}^n$ and $\{d_{2,j}^i\}_{i=1}^n$:

$$\min_{\alpha_{2,j},\beta_{2,j}} \frac{1}{n} \sum_{i=1}^{n} E\big(Y^i - \alpha_{2,j} - \beta_{2,j} d_{2,j}^i\big)^2, \quad j = 1, \ldots, p. \tag{32}$$

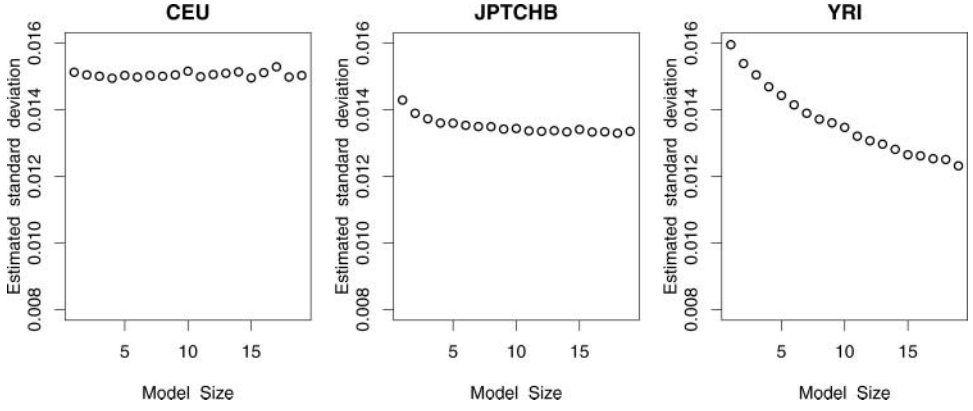


Figure 4. $\widehat{\sigma}$ of the three populations with respect to the selected model sizes, derived by using refitted cross-validation (RCV).

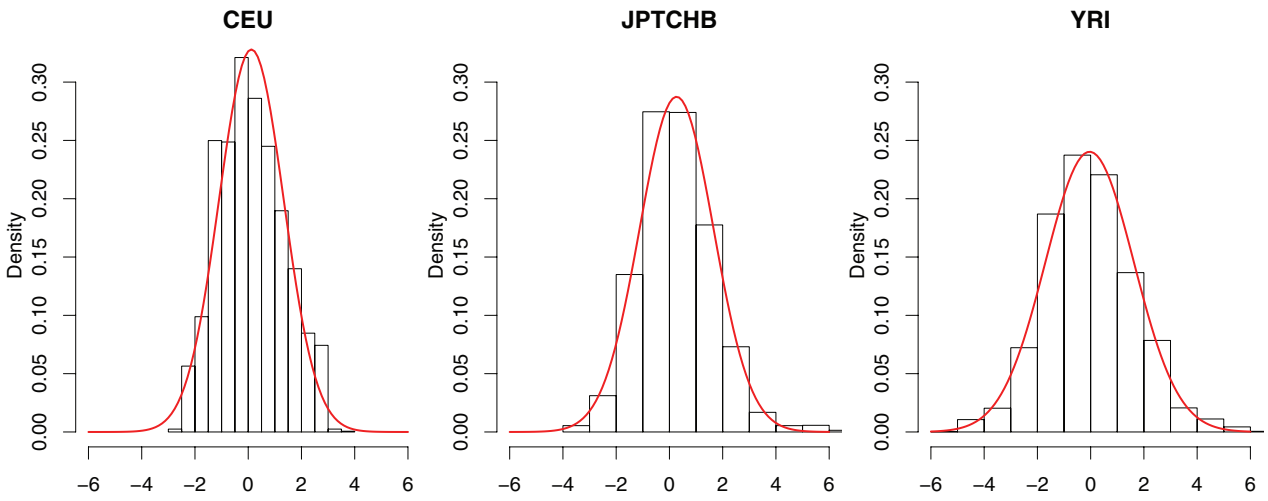


Figure 5. Empirical distributions and fitted normal density curves of the $Z$-values for each of the three populations. Because of dependency, the $Z$-values are no longer $N(0, 1)$ distributed. The empirical distributions, instead, are $N(0.12, 1.22^2)$ for CEU, $N(0.27, 1.39^2)$ for JPT and CHB, and $N(-0.04, 1.66^2)$ for YRI, respectively. The density curve for CEU is closest to $N(0, 1)$ and the least dispersed among the three. The online version of this figure is in color.



For ease of notation, we denote the recoded $n \times 2p$ dimensional design matrix as $\mathbf{X}$. The missing SNP measurement is imputed as 0 and the redundant SNP data are excluded. Finally, the logarithm-transform of the raw CCT8 gene expression data are used. The details of our testing procedures are summarized as follows.

- To begin with, consider the full model $Y = \alpha + \mathbf{X}\beta + \epsilon$, where $Y$ is the CCT8 gene expression data, $\mathbf{X}$ is the $n \times 2p$ dimensional design matrix of the SNP codings, and $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$ are the independent random errors. We adopt the RCV (Fan, Guo, and Hao 2012) technique to estimate $\sigma$ by $\widehat{\sigma}$, where LASSO is used in the first (variable selection) stage.
- Fit the marginal linear models (31) and (32) for each (recoded) SNP and obtain the least-square estimate $\widehat{\beta}_j$ for $j = 1, \ldots, 2p$. Compute the values of $Z$-statistics using formula (3), except that $\sigma$ is replaced by $\widehat{\sigma}$.
- Calculate the $p$-values based on the $Z$-statistics and compute $R(t) = \#\{P_j : P_j \le t\}$ for a fixed threshold $t$.
- Apply eigenvalue decomposition to the population covariance matrix $\mathbf{\Sigma}$ of the $Z$-statistics. By Proposition 1, $\mathbf{\Sigma}$ is the sample correlation matrix of $(d_{1,1}, d_{2,1}, \ldots, d_{1,p}, d_{2,p})^T$. Determine an appropriate number of factors $k$ and derive the corresponding factor loading coefficients $\{b_{ih}\}_{i=1,\ h=1}^{i=2p,\ h=k}$.
- Order the absolute-valued $Z$-statistics and choose the first $m = 95\% \times 2p$ of them. Apply $L_1$-regression to the equation set (22) and obtain its solution $\widehat{W}_1, \ldots, \widehat{W}_k$. Insert them into (21) and get the estimated FDP($t$).

For each intermediate step of the above procedure, the outcomes are summarized in the following figures. Figure 4 illustrates the trend of the RCV-estimated standard deviation $\widehat{\sigma}$ with respect to different model sizes. Our result is similar to that in Fan, Guo, and Hao (2012), in that although $\widehat{\sigma}$ is influenced by the selected model size, it is relatively stable and thus provides reasonable accuracy. The empirical distributions of the $Z$-values are presented in Figure 5, together with the fitted normal density curves. As pointed out in Efron (2007, 2010), due to the existence of dependency among the $Z$-values, their densities are either narrowed or widened and are not $N(0, 1)$ distributed. The histograms of the $p$-values are further provided in Figure 6,

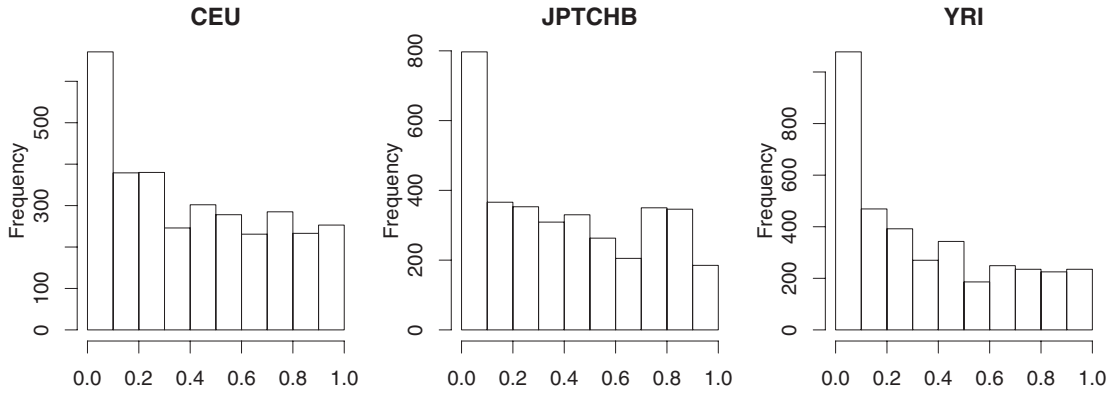


Figure 6. Histograms of the $p$-values for each of the three populations.

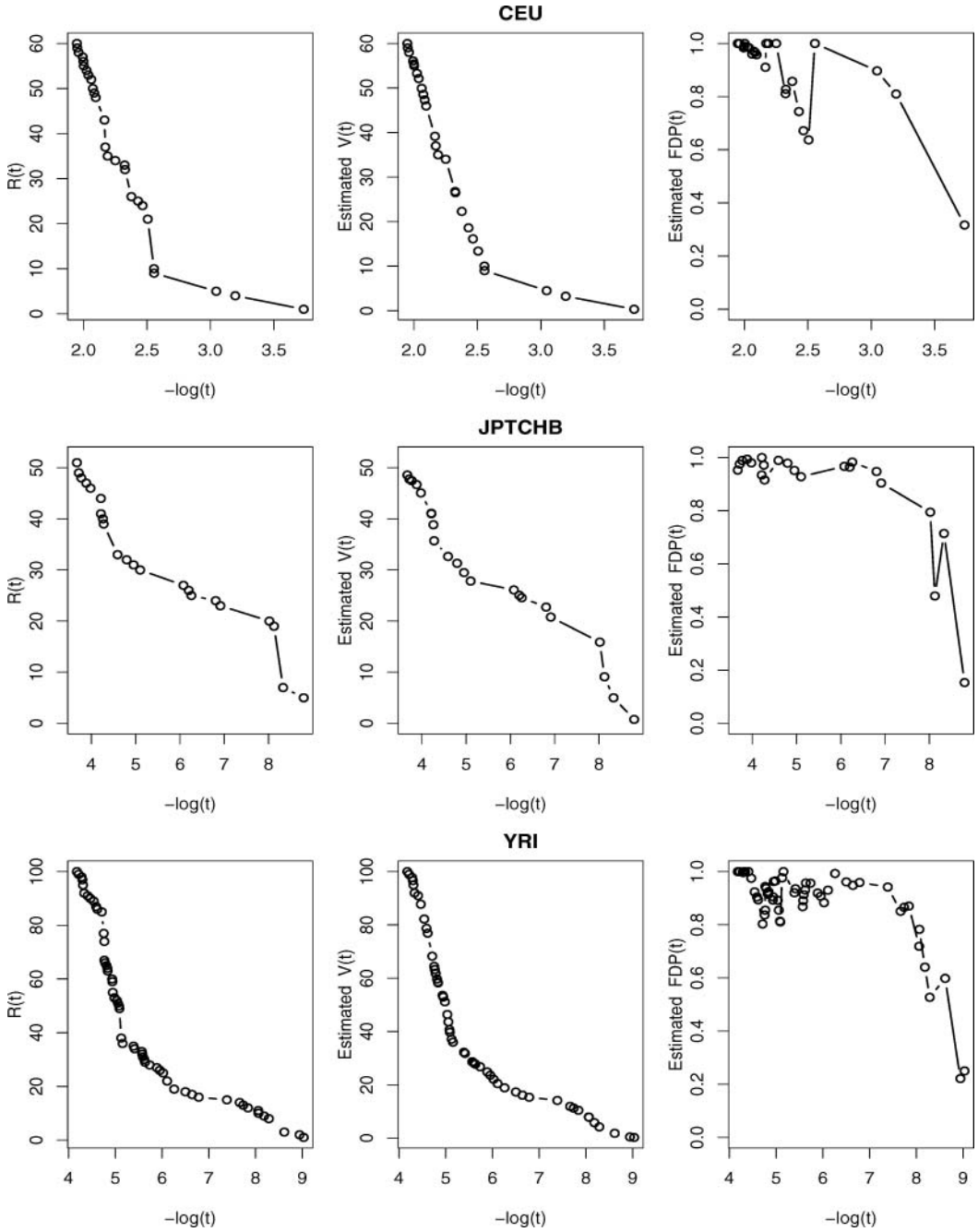


Figure 7. Number of total discoveries, estimated number of false discoveries, and estimated FDP as functions of thresholding $t$ for CEU population (row 1), JPT and CHB (row 2), and YRI (row 3). The $x$-coordinate is $-\log t$, the minus $\log_{10}$-transformed thresholding.

giving a crude estimate of the proportion of the false nulls for each of the three populations.

The main results of our analysis are presented in Figure 7 which depicts the number of total discoveries $R(t)$, the estimated number of false discoveries $\widehat{V}(t)$, and the estimated false discovery proportion $\widehat{\mathrm{FDP}}(t)$ as functions of (the minus $\log_{10}$-transformed) thresholding $t$ for the three populations. As can be seen, in each case both $R(t)$ and $\widehat{V}(t)$ are decreasing when $t$ decreases, but $\widehat{\mathrm{FDP}}(t)$ exhibits zigzag patterns and does not always decrease along with $t$, which results from the cluster effect of the $p$-values. A closer study of the outputs further shows that for all populations, the estimated FDP has a general trend of decreasing to the limit of around 0.1–0.2, which backs up the intuition that a large proportion of the smallest $p$-values should correspond to the false nulls (true discoveries) when Z-statistics is very large; however, in most other thresholding values, the estimated FDPs are at a high level. This is possibly due to small signal-to-noise ratios in eQTL studies.

The results of the selected SNPs, together with the estimated FDPs, are depicted in Table 7. It is worth mentioning that Deutsch et al. (2005) and Bradic, Fan, and Wang (2010) had also worked on the same CCT8 data to identify the significant SNPs in CEU population. Deutsch et al. (2005) performed association analysis for each SNP using analysis of variance (ANOVA), while Bradic, Fan, and Wang (2010) proposed the penalized composite quasi-likelihood variable selection method.

Table 7. Information of the selected SNPs and the associated FDP for a particular threshold. Note that the density curve of the $Z$-values for CEU population is close to $N(0, 1)$, so the approximate $\widehat{\text{FDP}}(t)$ equals $pt/R(t) \approx 0.631$. Therefore, our high estimated FDP is reasonable

| Population | Threshold | No. of discoveries | Estimated FDP | Selected SNPs |
|---|---|---|---|---|
| JPTCHB | $1.61 \times 10^{-9}$ | 5 | 0.1535 | rs965951 rs2070611 rs2832159 rs8133819 rs2832160 |
| YRI | $1.14 \times 10^{-9}$ | 2 | 0.2215 | rs9985076 rs965951 |
| CEU | $6.38 \times 10^{-4}$ | 4 | 0.8099 | rs965951 rs2832159 rs8133819 rs2832160 |

Table 8. Information of the selected SNPs for a particular threshold based on the dependence-adjusted procedure. The number of factors $k$ in Equation (24) equals 10. The estimated FDP is based on estimator (21) by applying PFA to the adjusted $Z$-values. $*$ is the indicator for SNP equal to 2 and otherwise is the indicator for 1

| Population | Threshold | No. of discoveries | Estimated FDP | Selected SNPs |
|---|---|---|---|---|
| JPTCHB | $2.89 \times 10^{-4}$ | 5 | 0.1205 | rs965951 rs2070611 rs2832159 rs8133819 rs2832160 |
| YRI | $8.03 \times 10^{-5}$ | 4 | 0.2080 | rs7283791 rs11910981 rs8128844 rs965951 |
| CEU | $5.16 \times 10^{-2}$ | 6 | 0.2501 | rs464144* rs4817271 rs2832195 rs2831528* rs1571671* rs6516819* |



Their findings were different as well, for the first group identified four SNPs (exactly the same as ours) that have the smallest $p$-values but the second group only discovered one SNP rs965951 among those four, arguing that the other three SNPs make little additional contributions conditioning on the presence of rs965951. Our results for CEU population coincide with that of the latter group, in the sense that the FDR is high in our findings and our association study is marginal rather than joint modeling among several SNPs.

Table 8 lists the SNP selection based on the dependence-adjusted procedure. For JPTCHB, with slightly smaller estimated FDP, the dependence-adjusted procedure selects the same SNPs with the group selected by the fixed-threshold procedure, which suggests that these five SNPs are significantly associated with the variation in gene CCT8 expression levels. For YRI, rs965951 is selected by both the procedures, but the dependence-adjusted procedure selects other three SNPs which do not appear in Table 7. For CEU, the selections based on the two procedures are quite different. However, since the estimated FDP for CEU is much smaller in Table 8 and the signal-to-noise ratio of the test statistics is higher from the dependence-adjusted procedure, the selection group in Table 8 seems more reliable.

## 7. DISCUSSION

We have proposed a new method (PFA) for high-dimensional multiple testing where the test statistics have an arbitrary dependence structure. For multivariate normal test statistics with a known covariance matrix, we can express the test statistics as an approximate factor model with weakly dependent random errors, by applying spectral decomposition to the covariance matrix. We then obtain an explicit expression for the FDP in large-scale simultaneous tests. This result has important applications in controlling FDP and FDR. We also provide a procedure to estimate the realized FDP, which, in our simulation studies, correctly tracks the trend of FDP with smaller amount of noise.

To take into account the dependence structure in the test statistics, we propose a dependence-adjusted procedure with different threshold values for magnitude of $Z_i$ in different hypotheses. This procedure has been shown in simulation studies to be more powerful than the fixed threshold procedure. An interesting research question is how to take advantage of the dependence structure such that the testing procedure is more powerful or even optimal under arbitrary dependence structures.

While our procedure is based on a known correlation matrix, we would expect that it can be adapted to the case with estimated covariance matrix. The question is then how accurate the covariance matrix should be so that a simple substitution procedure will give an accurate estimate of FDP.

We provide a simple method to estimate the realized principal factors. A more accurate method is probably the use of the penalized least-square method to explore the sparsity and to estimate the realized principal factor.

## APPENDIX

Lemma 1 is fundamental to our proof of Theorem 1 and Proposition 2. The result is known in probability, but has the formal statement and proof in Lyons (1988).

*Lemma 1* (*Strong Law of Large Numbers for Weakly Correlated Variables*). Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of real-valued random variables such that $E|X_n|^2 \le 1$. If $|X_n| \le 1$ a.s. and $\sum_{N\ge1} \frac{1}{N} E|\frac{1}{N}\sum_{n\le N} X_n|^2 < \infty$, then $\lim_{N\to\infty} \frac{1}{N}\sum_{n\le N} X_n = 0$ a.s.

*Proof of Proposition 2:* Note that $P_i = 2\Phi(-|Z_i|)$. Based on the expression of $(Z_1, \ldots, Z_p)^T$ in (10), $\{I(P_i \le t|W_1, \ldots, W_k)\}_{i=1}^p$ are dependent random variables. Nevertheless, we want to prove

$$p^{-1}\sum_{i=1}^{p}[I(P_i \le t|W_1, \ldots, W_k) - P(P_i \le t|W_1, \ldots, W_k)] \xrightarrow{p\to\infty} 0 \text{ a.s.} \tag{A.1}$$

Letting $X_i = I(P_i \le t|W_1, \ldots, W_k) - P(P_i \le t|W_1, \ldots, W_k)$, by Lemma 1 the conclusion (A.1) is correct if we can show

$$\operatorname{var}\left(p^{-1}\sum_{i=1}^{p} I(P_i \le t|W_1, \ldots, W_k)\right) = O_p(p^{-\delta}) \text{ for some } \delta > 0.$$

To begin with, note that

$$\begin{aligned}&\operatorname{var}\left(p^{-1}\sum_{i=1}^{p} I(P_i \le t|W_1, \ldots, W_k)\right)\\ &= p^{-2}\sum_{i=1}^{p}\operatorname{var}(I(P_i \le t|W_1, \ldots, W_k))\\ &+ 2p^{-2}\sum_{1\le i<j\le p}\operatorname{cov}(I(P_i \le t|W_1, \ldots, W_k), I(P_j \le t|W_1, \ldots, W_k)).\end{aligned}$$

Since $\operatorname{var}(I(P_i \le t|W_1, \ldots, W_k)) \le \frac{1}{4}$, the first term in the right-hand side of the last equation is $O_p(p^{-1})$. For the second term, the covariance is given by

$$\begin{aligned}&P(P_i \le t, P_j \le t|W_1, \ldots, W_k) - P(P_i \le t|W_1, \ldots, W_k)P\\ &\quad\times(P_j \le t|W_1, \ldots, W_k)\\ &= P(|Z_i| < -\Phi^{-1}(t/2), |Z_j| < -\Phi^{-1}(t/2)|W_1, \ldots, W_k)\\ &\quad - P(|Z_i| < -\Phi^{-1}(t/2)|W_1, \ldots, W_k)P(|Z_j|\\ &\quad < -\Phi^{-1}(t/2)|W_1, \ldots, W_k).\end{aligned}$$

To simplify the notation, let $\rho_{ij}^k$ be the correlation between $K_i$ and $K_j$. Without loss of generality, we assume $\rho_{ij}^k > 0$ (for $\rho_{ij}^k < 0$, the calculation is similar). Denote by

$$c_{1,i} = a_i(-z_{t/2} - \eta_i - \mu_i), \quad c_{2,i} = a_i(z_{t/2} - \eta_i - \mu_i).$$

Then, from the joint normality, it can be shown that

$$\begin{aligned}&P(|Z_i| < -\Phi^{-1}(t/2), |Z_j| < -\Phi^{-1}(t/2)|W_1, \ldots, W_k)\\ &= P(c_{2,i}/a_i < K_i < c_{1,i}/a_i, c_{2,j}/a_j < K_j < c_{1,j}/a_j)\\ &= \int_{-\infty}^{\infty}\left[\Phi\left(\frac{(\rho_{ij}^k)^{1/2}z + c_{1,i}}{(1-\rho_{ij}^k)^{1/2}}\right) - \Phi\left(\frac{(\rho_{ij}^k)^{1/2}z + c_{2,i}}{(1-\rho_{ij}^k)^{1/2}}\right)\right]\\ &\quad\times\left[\Phi\left(\frac{(\rho_{ij}^k)^{1/2}z + c_{1,j}}{(1-\rho_{ij}^k)^{1/2}}\right) - \Phi\left(\frac{(\rho_{ij}^k)^{1/2}z + c_{2,j}}{(1-\rho_{ij}^k)^{1/2}}\right)\right]\phi(z)dz.\end{aligned} \tag{A.2}$$

Next, we will use Taylor expansion to analyze the joint probability further. We have shown that $(K_1, \ldots, K_p)^T \sim N(0, \mathbf{A})$ are weakly dependent random variables. Let $\operatorname{cov}_{ij}^k$ denote the covariance of $K_i$ and $K_j$, which is the $(i, j)$th element of the covariance matrix $\mathbf{A}$. We also let $b_{ij}^k = (1 - \sum_{h=1}^k b_{ih}^2)^{1/2}(1 - \sum_{h=1}^k b_{jh}^2)^{1/2}$. By the Hölder inequality,

$$\begin{aligned}p^{-2}\sum_{i,j=1}^{p}|\operatorname{cov}_{ij}^k|^{1/2} &\le p^{-1/2}\left(\sum_{i,j=1}^{p}|\operatorname{cov}_{ij}^k|^2\right)^{1/4}\\ &= \left[p^{-2}\left(\sum_{i=k+1}^{p}\lambda_i^2\right)^{1/2}\right]^{1/4} \to 0\end{aligned}$$

as $p \to \infty$. For each $\Phi(\cdot)$, we apply Taylor expansion with respect to $(\operatorname{cov}_{ij}^k)^{1/2}$,

$$\begin{aligned}\Phi\left(\frac{(\rho_{ij}^k)^{1/2}z + c_{1,i}}{(1-\rho_{ij}^k)^{1/2}}\right) &= \Phi\left(\frac{(\operatorname{cov}_{ij}^k)^{1/2}z + (b_{ij}^k)^{1/2}c_{1,i}}{(b_{ij}^k - \operatorname{cov}_{ij}^k)^{1/2}}\right)\\ &= \Phi(c_{1,i}) + \phi(c_{1,i})(b_{ij}^k)^{-1/2}z(\operatorname{cov}_{ij}^k)^{1/2}\\ &\quad + \frac{1}{2}\phi(c_{1,i})c_{1,i}(b_{ij}^k)^{-1}(1-z^2)\operatorname{cov}_{ij}^k + R(\operatorname{cov}_{ij}^k),\end{aligned}$$

where $R(\operatorname{cov}_{ij}^k)$ is the Lagrange residual term in the Taylor's expansion, and $R(\operatorname{cov}_{ij}^k) = f(z)O(|\operatorname{cov}_{ij}^k|^{3/2})$ in which $f(z)$ is a polynomial function of $z$ with the highest order as 6.

Therefore, we have (A.2) equals

$$\begin{aligned}&[\Phi(c_{1,i}) - \Phi(c_{2,i})][\Phi(c_{1,j}) - \Phi(c_{2,j})]\\ &\quad + (\phi(c_{1,i}) - \phi(c_{2,i}))(\phi(c_{1,j}) - \phi(c_{2,j}))(b_{ij}^k)^{-1}\operatorname{cov}_{ij}^k + O(|\operatorname{cov}_{ij}^k|^{3/2}),\end{aligned}$$

where we have used the fact that $\int_{-\infty}^{\infty} z\phi(z)dz = 0$, $\int_{-\infty}^{\infty}(1 - z^2)\phi(z)dz = 0$, and the finite moments of standard normal distribution are finite. Now since $P(|Z_i| < -\Phi^{-1}(t/2)|W_1, \ldots, W_k) = \Phi(c_{1,i}) - \Phi(c_{2,i})$, we have

$$\begin{aligned}&\operatorname{cov}(I(P_i \le t|W_1, \ldots, W_k), I(P_j \le t|W_1, \ldots, W_k))\\ &= (\phi(c_{1,i}) - \phi(c_{2,i}))(\phi(c_{1,j}) - \phi(c_{2,j}))a_ia_j\operatorname{cov}_{ij}^k + O(|\operatorname{cov}_{ij}^k|^{3/2}).\end{aligned}$$

In the last line, $(\phi(c_{1,i}) - \phi(c_{2,i}))(\phi(c_{1,j}) - \phi(c_{2,j}))a_ia_j$ is bounded by some constant except on a countable collection of measure zero sets. Let $C_i$ be defined as the set $\{z_{t/2} + \eta_i + \mu_i = 0\} \cup \{z_{t/2} - \eta_i - \mu_i = 0\}$. On the set $C_i^c$, $(\phi(c_{1,i}) - \phi(c_{2,i}))a_i$ converges to zero as $a_i \to \infty$. Therefore, $(\phi(c_{1,i}) - \phi(c_{2,i}))(\phi(c_{1,j}) - \phi(c_{2,j}))a_ia_j$ is bounded by some constant on $(\bigcup_{i=1}^p C_i)^c$.

By the Cauchy–Schwartz inequality and (C0) in Theorem 1, $p^{-2}\sum_{i,j}|\operatorname{cov}_{i,j}^k| = O(p^{-\delta})$. Also, we have $|\operatorname{cov}_{ij}^k|^{3/2} < |\operatorname{cov}_{ij}^k|$. On the set $(\bigcup_{i=1}^p C_i)^c$, we conclude that

$$\operatorname{var}\left(p^{-1}\sum_{i=1}^{p} I(P_i \le t|W_1, \ldots, W_k)\right) = O_p(p^{-\delta}).$$

Hence, by Lemma 1, for fixed $(w_1, \ldots, w_k)^T$,

$$\begin{aligned}&p^{-1}\sum_{i=1}^{p}\{I(P_i \le t|W_1 = w_1, \ldots, W_k = w_k) - P(P_i \le t|W_1\\ &\quad = w_1, \ldots, W_k = w_k)\} \xrightarrow{p\to\infty} 0 \text{ a.s.}\end{aligned} \tag{A.3}$$

If we define the probability space on which $(W_1, \ldots, W_k)$ and $(K_1, \ldots, K_p)$ are constructed as in (10) to be $(\Omega, \mathcal{F}, \nu)$, with $\mathcal{F}$ and $\nu$ being the associated $\sigma-$algebra and (Lebesgue) measure, then in a more formal way, (A.3) is equivalent to

$$\begin{aligned}&p^{-1}\sum_{i=1}^{p}\{I(P_i(\omega) \le t|W_1 = w_1, \ldots, W_k = w_k) - P(P_i \le t|W_1\\ &\quad = w_1, \ldots, W_k = w_k)\} \xrightarrow{p\to\infty} 0\end{aligned}$$

for each fixed $(w_1, \ldots, w_k)^T$ and almost every $\omega \in \Omega$, leading further to

$$p^{-1}\sum_{i=1}^{p}\{I(P_i(\omega) \le t) - P(P_i \le t|W_1(\omega), \ldots, W_k(\omega))\} \xrightarrow{p\to\infty} 0$$

for almost every $\omega \in \Omega$, which is the definition for

$$p^{-1}\sum_{i=1}^{p}\{I(P_i \le t) - P(P_i \le t|W_1, \ldots, W_k)\} \xrightarrow{p\to\infty} 0 \text{ a.s.}$$

Therefore,

$$\begin{aligned}\lim_{p\to\infty} p^{-1}\sum_{i=1}^{p}\{I(P_i \le t) - [\Phi(a_i(z_{t/2} + \eta_i + \mu_i))\\ + \Phi(a_i(z_{t/2} - \eta_i - \mu_i))]\} = 0 \text{ a.s.}\end{aligned}$$

With the same argument, we can show

$$\lim_{p\to\infty} p_0^{-1}\left\{V(t) - \sum_{i\in\{\text{true null}\}} [\Phi(a_i(z_{t/2}+\eta_i)) + \Phi(a_i(z_{t/2}-\eta_i))]\right\} = 0 \text{ a.s.}$$

for the high-dimensional sparse case. The proof of Proposition 2 is now complete.

*Proof of Theorem 1:*

For ease of notation, denote $\sum_{i=1}^{p}[\Phi(a_i(z_{t/2}+\eta_i+\mu_i)) + \Phi(a_i(z_{t/2}-\eta_i-\mu_i))]$ as $\tilde{R}(t)$ and $\sum_{i\in\{\text{true null}\}}[\Phi(a_i(z_{t/2}+\eta_i)) + \Phi(a_i(z_{t/2}-\eta_i))]$ as $\tilde{V}(t)$, then

$$\begin{aligned}
&\lim_{p\to\infty}\left\{\text{FDP}(t) - \frac{\sum_{i\in\{\text{true null}\}}[\Phi(a_i(z_{t/2}+\eta_i)) + \Phi(a_i(z_{t/2}-\eta_i))]}{\sum_{i=1}^{p}[\Phi(a_i(z_{t/2}+\eta_i+\mu_i)) + \Phi(a_i(z_{t/2}-\eta_i-\mu_i))]}\right\}\\
&= \lim_{p\to\infty}\left\{\frac{V(t)}{R(t)} - \frac{\tilde{V}(t)}{\tilde{R}(t)}\right\}\\
&= \lim_{p\to\infty}\frac{(V(t)/p_0)[(\tilde{R}(t)-R(t))/p] + (R(t)/p)[(V(t)-\tilde{V}(t))/p_0]}{R(t)\tilde{R}(t)/(p_0 p)}\\
&= 0 \text{ a.s.}
\end{aligned}$$

by the results in Proposition 2 and the fact that both $p_0^{-1}V(t)$ and $p^{-1}R(t)$ are bounded random variables. The proof of Theorem 1 is complete. □

*Proof of Theorem 2:* Letting

$$\Delta_1 = \sum_{i=1}^{p}\left[\Phi(a_i(z_{t/2}+\mathbf{b}_i^T\widehat{\mathbf{w}})) - \Phi(a_i(z_{t/2}+\mathbf{b}_i^T\mathbf{w}))\right]$$

and

$$\Delta_2 = \sum_{i=1}^{p}\left[\Phi(a_i(z_{t/2}-\mathbf{b}_i^T\widehat{\mathbf{w}})) - \Phi(a_i(z_{t/2}-\mathbf{b}_i^T\mathbf{w}))\right],$$

we have

$$\widehat{\text{FDP}}(t) - \text{FDP}_{\text{A}}(t) = (\Delta_1+\Delta_2)/R(t).$$

Consider $\Delta_1 = \sum_{i=1}^{p}\Delta_{1i}$. By the mean value theorem, there exists $\xi_i$ in the interval of $(\mathbf{b}_i^T\widehat{\mathbf{w}}, \mathbf{b}_i^T\mathbf{w})$, such that $\Delta_{1i} = \phi(a_i(z_{t/2}+\xi_i))a_i\mathbf{b}_i^T(\widehat{\mathbf{w}}-\mathbf{w})$, where $\phi(\cdot)$ is the standard normal density function.

Next, we will show that $\phi(a_i(z_{t/2}+\xi_i))a_i$ is bounded by a constant. Without loss of generality, we discuss about the case in (C2) when $z_{t/2}+\mathbf{b}_i^T\mathbf{w} < -\tau$. By (C3), we can choose sufficiently large $p$ such that $z_{t/2}+\xi_i < -\tau/2$. For the function $g(a) = \exp(-a^2x^2/8)a$, $g(a)$ is maximized when $a = 2/x$. Therefore,

$$\sqrt{2\pi}\phi(a_i(z_{t/2}+\xi_i))a_i < a_i\exp(-a_i^2\tau^2/8) \le 2\exp(-1/2)/\tau.$$

For $z_{t/2}+\mathbf{b}_i^T\mathbf{w} > \tau$, we have the same result. In both cases, we can use a constant $D$ such that $\phi(a_i(z_{t/2}+\xi_i))a_i \le D$.

By the Cauchy–Schwartz inequality, we have $\sum_{i=1}^{p}|b_{ih}| \le (p\sum_{i=1}^{p}b_{ih}^2)^{1/2} = (p\lambda_h)^{1/2}$. Therefore, by the Cauchy–Schwartz inequality and the fact that $\sum_{h=1}^{k}\lambda_h < p$, we have

$$\begin{aligned}
|\Delta_1| &\le D\sum_{i=1}^{p}\left[\sum_{h=1}^{k}|b_{ih}||\widehat{w}_h - w_h|\right]\\
&\le D\sum_{h=1}^{k}(p\lambda_h)^{1/2}|\widehat{w}_h - w_h|\\
&\le D\sqrt{p}\left(\sum_{h=1}^{k}\lambda_h\sum_{h=1}^{k}(\widehat{w}_h - w_h)^2\right)^{1/2}\\
&< Dp\|\widehat{\mathbf{w}} - \mathbf{w}\|_2.
\end{aligned}$$

By (C1) in Theorem 2, $R(t)/p > H$ for $H > 0$ when $p\to\infty$. Therefore, $|\Delta_1/R(t)| = O(\|\widehat{\mathbf{w}} - \mathbf{w}\|_2)$. For $\Delta_2$, the result is the same. The proof of Theorem 2 is now complete. □

*Proof of Theorem 3:* The proof is technical. To save space, it is relegated to the supplementary material.

*Proof of Theorem 4:* Note that $\|\widehat{\mathbf{W}}_{\text{LS}} - \widehat{\mathbf{W}}^*_{\text{LS}}\|_2 = \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\mu}\|_2$. By the definition of $\mathbf{X}$, we have $\mathbf{X}^T\mathbf{X} = \Lambda$, where $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_k)$. Therefore, by the Cauchy–Schwartz inequality,

$$\|\widehat{\mathbf{W}}_{\text{LS}} - \widehat{\mathbf{W}}^*_{\text{LS}}\|_2 = \left[\sum_{i=1}^{k}\left(\frac{\sqrt{\lambda_i}\boldsymbol{\gamma}_i^T\boldsymbol{\mu}}{\lambda_i}\right)^2\right]^{1/2} \le \|\boldsymbol{\mu}\|_2\left(\sum_{i=1}^{k}\frac{1}{\lambda_i}\right)^{1/2}.$$

The proof is complete. □

## SUPPLEMENTARY MATERIAL

**Proof of Theorem 3:** This supplement consists of the proof of Theorem 3. (pdf)

*[Received November 2010. Revised October 2011.]*

## REFERENCES


Barras, L., Scaillet, O., and Wermers, R. (2010), "False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas," *Journal of Finance*, 65, 179–216. [1019]

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, 57, 289–300. [1019]

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. [1019]

Bradic, J., Fan, J., and Wang, W. (2010), "Penalized Composite Quasi-Likelihood For Ultrahigh-Dimensional Variable Selection," *Journal of the Royal Statistical Society*, Series B, 73, 325–349. [1029,1031]

Clarke, S., and Hall, P. (2009), "Robustness of Multiple Testing Procedures Against Dependence," *The Annals of Statistics*, 37, 332–358. [1019]

Delattre, S., and Roquain, E. (2011), "On the False Discovery Proportion Convergence Under Gaussian Equi-Correlation," *Statistics and Probability Letters*, 81, 111–115. [1022]

Deutsch, S., Lyle, R., Dermitzakis, E. T., Attar, H., Subrahmanyan, L., Gehrig, C., Parand, L., Gagnebin, M., Rougemont, J., Jongeneel, C. V., and Antonarakis, S. E. (2005), "Gene Expression Variation and Expression Quantitative Trait Mapping of Human Chromosome 21 Genes," *Human Molecular Genetics*, 14, 3741–3749. [1031]

Efron, B. (2007), "Correlation and Large-Scale Simultaneous Significance Testing," *Journal of the American Statistical Association*, 102, 93–103. [1019,1021,1023,1025,1027,1030]

——— (2010), "Correlated Z-Values and the Accuracy of Large-Scale Statistical Estimates," *Journal of the American Statistical Association*, 105, 1042–1055. [1021]

Fan, J., Guo, S., and Hao, N. (2012), "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the Royal Statistical Society*, Series B, 74, 37–65. [1020,1030]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of American Statistical Association*, 96, 1348–1360. [1023]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1026]

Farcomeni, A. (2007), "Some Results on the Control of the False Discovery Rate Under Dependence," *Scandinavian Journal of Statistics*, 34, 275–297. [1021]

Ferreira, J., and Zwinderman, A. (2006), "On the Benjamini-Hochberg Method," *The Annals of Statistics*, 34, 1827–1849. [1019]

Finner, H., Dickhaus, T., and Roters, M. (2007), "Dependency and False Discovery Rate: Asymptotics," *The Annals of Statistics*, 35, 1432–1455. [1022,1025]

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. [1020,1021,1022,1023,1025,1025]

Genovese, C., and Wasserman, L. (2004), "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32, 1035–1061. [1023]
Leek, J. T., and Storey, J. D. (2008), "A General Framework for Multiple Testing Dependence," *Proceedings of the National Academy of Sciences of the United States of America*, 105, 18718–18723. [1020,1025]
Lyons, R. (1988), "Strong Laws of Large Numbers for Weakly Correlated Random Variables," *The Michigan Mathematical Journal*, 35, 353–359. [1032]
Meinshausen, N. (2006), "False Discovery Control for Multiple Tests of Association Under General Dependence," *Scandinavian Journal of Statistics*, 33, 227–237. [1023]
Owen, A. B. (2005), "Variance of the Number of False Discoveries," *Journal of the Royal Statistical Society,* Series B, 67, 411–426. [1023,1026]
Portnoy, S. (1984a), "Tightness of the Sequence of C.D.F. Processes Defined From Regression Fractiles," in *Robust and Nonlinear Time Series Analysis*, eds. J. Franke, W. Hardle, and D. Martin, New York: Springer-Verlag, pp. 231–246. [1024]
——— (1984b), "Asymptotic Behavior of M-Estimators of $p$ Regression Parameters When $p^2/n$ is Large; I. Consistency," *The Annals of Statistics*, 12, 1298–1309. [1024]
Roquain, E., and Villers, F. (2011), "Exact Calculations For False Discovery Proportion With Application To Least Favorable Configurations," *The Annals of Statistics*, 39, 584–612. [1022]
Sarkar, S. (2002), "Some Results on False Discovery Rate in Stepwise Multiple Testing Procedures," *The Annals of Statistics*, 30, 239–257. [1019]
Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Series B, 64, 479–498. [1019,1021,1022]
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, 66, 187–205. [1019]
Sun, W., and Cai, T. (2009), "Large-Scale Multiple Testing Under Dependency," *Journal of the Royal Statistical Society*, Series B, 71, 393–424. [1020]



# Comment

Larry WASSERMAN

## 1. INTRODUCTION

Fan, Hu, and Gu (FHG) have derived some elegant methods and theory for controlling the false discovery proportion (FDP) in the case of dependent test statistics. The work is very interesting and the technical tools that are developed will no doubt be useful in other contexts. The article raises some general questions: should we be doing hypothesis testing in these problems? should we interpret coefficients in linear models? and should we use marginal regression?

## 2. TESTING MARGINAL REGRESSION COEFFICIENTS

When I was a student, I learned a few rules of thumb that I teach in my courses:

1. Do not use hypothesis testing unless it is absolutely necessary. Focus on estimation, confidence intervals, or prediction error instead.
2. The parameters in a linear model are meaningless unless (a) the model is exactly correct and (b) there are no unobserved confounding variables.
3. The coefficients of a marginal regression are difficult to interpret. Even if the coefficients in a linear model are well defined and interpretable (which is itself rare), the coefficients in the marginal regression are unrelated to them. You can have a huge regression effect and a zero marginal effect. Conversely, you can have a tiny regression effect and a huge marginal effect. In more traditional language, correlation is not causation.

Instead of asking "which parameter coefficients are significant," we can ask (and answer) the simpler question: what is a good sparse linear predictor? We can answer this with, say, the Lasso coupled with cross-validation estimates of prediction risk. No tests and no interpreting coefficients. This is consistent with the view espoused by Breiman (2001). I am not suggesting that hypothesis testing and interpreting parameters are always bad. But I do wonder if we, as a field, have put too much attention on interpreting and testing parameters lately. I would be very interested in hearing the authors' views on this point.

## 3. OTHER RANDOM COMMENTS

1. The example in Section 6 is quite interesting. But why use the Lasso to estimate $\sigma$? Is it even reasonable to assume that $\sigma$ is constant?
2. If one is going to use FDP control, then it is worth noting that, simultaneously, Genovese and Wasserman (2006) and van der Laan, Dudoit, and Pollard (2004) proposed a very simple method that works for arbitrary dependence. The method is as follows:
   (a) Find the test statistics $R$ rejected by some method that controls $\alpha$-familywise error (such as Bonferroni).
   (b) Add the next $k$ test statistics to the rejection set where $k$ is chosen so that $k/(k + |R|) = c$.
   (c) It follows that $\mathbb{P}(\text{FDP} > c) \leq \alpha$.
   How does this compare to the proposed methods?
3. Romano and Wolf (2007) proposed some methods for controlling FDP. Are there any connections with the method in FHG?

Larry Wasserman Department of Statistics and the Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213-3891 (E-mail: *larry@stat.cmu.edu*).


Journal of the American Statistical Association
September 2012, Vol. 107, No. 499, Theory and Methods
DOI: 10.1080/01621459.2012.711729