ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 4, *pp.* 1143–1164

Estimation of the false discovery proportion with unknown dependence

Jianqing Fan

Princeton University, USA, and Fudan University, Shanghai, People's Republic of China

and Xu Han

Temple University, Philadelphia, USA

[Received May 2013. Final revision July 2016]

Summary. Large-scale multiple testing with correlated test statistics arises frequently in much scientific research. Incorporating correlation information in approximating the false discovery proportion (FDP) has attracted increasing attention in recent years. When the covariance matrix of test statistics is known, Fan and his colleagues provided an accurate approximation of the FDP under arbitrary dependence structure and some sparsity assumption. However, the covariance matrix is often unknown in many applications and such dependence information must be estimated before approximating the FDP. The estimation accuracy can greatly affect the FDP approximation. In the current paper, we study theoretically the effect of unknown dependence on the testing procedure and establish a general framework such that the FDP can be well approximated. The effects of unknown dependence on approximating the FDP are in the following two major aspects: through estimating eigenvalues or eigenvectors and through estimating marginal variances. To address the challenges in these two aspects, we firstly develop general requirements on estimates of eigenvalues and eigenvectors for a good approximation of the FDP. We then give conditions on the structures of covariance matrices that satisfy such requirements. Such dependence structures include banded or sparse covariance matrices and (conditional) sparse precision matrices. Within this framework, we also consider a special example to illustrate our method where data are sampled from an approximate factor model, which encompasses most practical situations. We provide a good approximation of the FDP via exploiting this specific dependence structure. The results are further generalized to the situation where the multivariate normality assumption is relaxed. Our results are demonstrated by simulation studies and some real data applications.

Keywords: Approximate factor model; Dependent test statistics; False discovery proportion; Large-scale multiple testing; Unknown covariance matrix

1. Introduction

The correlation effect of dependent test statistics in large-scale multiple testing has attracted considerable attention in recent years. In microarray experiments, thousands of gene expressions are usually correlated when cells are treated. Applying standard Benjamini and Hochberg (1995) or Storey's (2002) procedures for independent test statistics can lead to inaccurate false discovery control. Statisticians have now reached the conclusion that it is important and necessary to incorporate the dependence information in the multiple-testing procedure. See Efron (2007, 2010), Leek and Storey (2008), Schwartzman and Lin (2011) and Fan, Han and Gu (2012).

Address for correspondence: Jianqing Fan, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.
E-mail: jqfan@princeton.edu

 1369–7412/17/791143

Consideration of multiple-testing procedures for dependent test statistics dates back to the early 2000s. Benjamini and Yekutieli (2001) proved that the false discovery rate can be controlled by the Benjamini–Hochberg procedure when the test statistics satisfy positive regression dependence on subsets. Extension to a generalized stepwise procedure under positive regression dependence on subsets has been proved by Sarkar (2002). Later Storey *et al*. (2004) also showed that Storey's procedure can control the false discovery rate (FDR) under weak dependence. Sun and Cai (2009) developed a procedure where parameters underlying test statistics follow a hidden Markov model. Insightful results of validation for standard multiple-testing procedures under more general dependence structures have been shown in Clarke and Hall (2009). However, even if these procedures are valid under these special dependence structures, they still suffer from a loss of efficiency without considering the actual dependence information. In other words, there are universal upper bounds for a given class of covariance matrices.

A challenging question is how to incorporate the correlation effect in the testing procedure. Efron (2007, 2010) in his seminal work obtained repeated test statistics based on the bootstrap sample from the original raw data, took out the first eigenvector of the covariance matrix of the test statistics such that the correlation effect could be explained by a dispersion variate A and estimated A from the data to construct an estimate for the realized false discovery proportion (FDP). Friguet *et al*. (2009) and Desai and Storey (2012) assumed that the data come directly from a strict factor model with independent idiosyncratic errors and used the EM algorithm to estimate the number of factors, the factor loadings and the realized factors in the model and obtained an estimator for the FDP by subtracting out realized common factors. The drawbacks of the aforementioned procedures are, however, restricted model assumptions and the lack of formal justification.

Fan, Han and Gu (2012) considered a general set-up for approximating the FDP. They assumed that the test statistics are from a multivariate normal distribution with a known but arbitrary covariance matrix. Their idea is to apply spectral decomposition to the covariance matrix of test statistics and to use principal factors to account for dependence. This method is called the principal factor approximation (PFA). Under some sparsity assumption, they provided an accurate approximation of the FDP based on the eigenvalues and eigenvectors of the known covariance matrix.

A major restriction of the set-up in Fan, Han and Gu (2012) is that the covariance matrix of test statistics is known. Although they provided an interesting application with known covariance matrix, in many other cases, this matrix is usually unknown. For example, in microarray experiments, scientists are interested in testing whether genes are differently expressed under different experimental conditions (e.g. treatments or groups of patients). The dependence of test statistics is unknown in such applications. The problem of unknown dependence has at least two fundamental differences from the setting with known dependence.

(a) Impact through estimating marginal variances: when the population marginal variances of the observable random variables are unknown, they must be estimated first for standardization. In such a case, the popular choice of the test statistics will have a t-distribution with dependence rather than the multivariate normal distribution that was considered in Fan, Han and Gu (2012).
(b) Impact through estimating eigenvalues or eigenvectors: even if the population marginal variances of the observable random variables are known, estimation of eigenvalues or eigenvectors can still significantly affect the FDP approximation. In various situations, FDP approximation can have inferior performance even if a researcher chooses the 'best'

estimator for the unknown matrix. Therefore, more theoretical and methodological modifications are needed before directly applying the PFA to unknown dependence settings.

The current paper aims to study theoretically the effect of unknown dependence on the testing procedure and to establish a general framework for FDP approximation. For the independence case, this quantity depends asymptotically only on the number of true null hypotheses. For the general case, as to be elucidated in Section 2.2 (around equation (6)), it is far more complicated, depending on the whole set of the unknown true nulls. Therefore, consistently estimating the FDP is a hopeless task unless the signals are sparse. Under some sparsity assumption, the FDP can be conservatively estimated by taking the null proportion to be 1. But this will cause other technical problems. Instead, we shall focus on a statistical quantity $\mathrm{FDP}_{\mathrm{A}}$ (see equation (6)) and estimate it directly. $\mathrm{FDP}_{\mathrm{A}}$ can be viewed as an asymptotic upper bound of the FDP, and correspondingly the expectation of $\mathrm{FDP}_{\mathrm{A}}$ is the asymptotic upper bound of the conventional FDR. For the challenges from the unknown dependence, since the effect of aspect (b) is even more important than that of aspect (a), we shall first develop requirements for estimated eigenvalues and eigenvectors. Surprisingly, for a good estimate of this upper bound, we do not need these estimates of eigenvalues and eigenvectors to be consistent themselves. This finding relaxes the consistency restriction of covariance matrix estimation under operator norm. Our framework of FDP approximation encompasses both weak dependence and strong dependence, including banded matrices, (conditional) sparse matrices, (conditional) sparse precision matrices, etc.

As a specific example, we shall consider the covariance matrices with an approximate factor structure. This factor model encompasses a majority of statistical applications and is a generalization to the model in Friguet *et al*. (2009) and Desai and Storey (2012). After applying the principal orthogonal complement thresholding estimators POET (Fan *et al*., 2013) to estimate the unknown covariance matrix, we can then assess the FDP. This combination of POET to estimate the covariance matrix and the PFA to approximate the FDP should be applicable to most practical situations and is the method that we recommend for practice.

We shall also examine the effect of unknown marginal variances and generalize our results to the situation when the test statistics have a t-distribution with dependence, which is beyond the multivariate normal assumption. This dependent t-distribution is not the conventional multivariate t-distribution. We shall show that our proposed method is still applicable to this more general situation. The performance of our procedure is further evaluated by simulation studies and real data analysis.

The organization of the rest of the paper is as follows: Section 2 provides background information on large-scale multiple testing under dependence and the PFA, Section 3 includes the theoretical study on FDP approximations, Section 4 contains simulation studies and Section 5 illustrates the methodology via an application to a microarray data set. Throughout this paper, we use $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ to denote the minimum and maximum eigenvalues of a symmetric matrix $\mathbf{A}$. We also denote the Frobenius norm $\|\mathbf{A}\|_{\mathrm{F}} = \mathrm{tr}^{1/2}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$, the operator norm $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ and the induced norms $\|\mathbf{A}\|_1 = \max_{1 \leqslant j \leqslant p} \Sigma_{i=1}^{p} |a_{ij}|$ and $\|\mathbf{A}\|_\infty = \max_{1 \leqslant i \leqslant p} \Sigma_{j=1}^{p} |a_{ij}|$.

The proposed method, POET–PFA, can be easily implemented by the R package `pfa` (version 1.1) on `https://cran.r-project.org`. The simulation code and the data set are available from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Approximation of the false discovery proportion

Suppose that the observed data $\{\mathbf{X}_i\}_{i=1}^n$ are p-dimensional independent random vectors with

$\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ is a high dimensional sparse vector, but we do not know which are the non-vanishing signals. Let $p_0 = \#\{j : \mu_j = 0\}$ and $p_1 = \#\{j : \mu_j \neq 0\}$ so that $p_0 + p_1 = p$. We wish to test which co-ordinates of $\boldsymbol{\mu}$ are signals based on the realizations $\{\mathbf{x}_i\}_{i=1}^n$.

Consider the test statistics $\mathbf{Z} = \sqrt{n}\bar{\mathbf{X}}$ in which $\bar{\mathbf{X}}$ is the sample mean of $\{\mathbf{X}_i\}_{i=1}^n$. Then, $\mathbf{Z} \sim N_p(\sqrt{n}\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Standardizing the test statistics $\mathbf{Z}$, we assume for simplicity that $\boldsymbol{\Sigma}$ is a correlation matrix. Let $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_p^*)^{\mathrm{T}} = \sqrt{n}\boldsymbol{\mu}$. Then, multiple testing $H_{0j} : \mu_j = 0$ *versus* $H_{1j} : \mu_j \neq 0$ is equivalent to testing $H_{0j} : \mu_j^* = 0$ *versus* $H_{1j} : \mu_j^* \neq 0$ based on the test statistics $\mathbf{Z} = (Z_1, \ldots, Z_p)^{\mathrm{T}}$. The p-value for the jth hypothesis is $2\Phi(-|Z_j|)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. We use a threshold value t to reject the hypotheses which have p-values that are smaller than t. Define $R(t) = \#\{P_j : P_j \leqslant t\}$ as the number of discoveries and $V(t) = \#\{\text{true null} : P_j \leqslant t\}$ the number of false discoveries $V(t)$, where P_j is the p-value for testing the jth hypothesis. Our interest focuses on approximating the FDP $\mathrm{FDP}(t) = V(t)/R(t)$; here and hereafter the convention $0/0 = 0$ is always used. Note that $R(t)$ is observable, and $\mathrm{FDP}(t)$ is a realized but unobservable random variable. In comparison with $\mathrm{FDR}(t) = E[\mathrm{FDP}(t)]$, an average of FDPs for hypothetical replications of experiments, the FDP is concerned with the number of false discoveries given the experiment.

The normality assumption is an idealization. In the current paper, we shall show both theoretically and numerically that, even if the normality assumption is violated, our results are still applicable for a more general setting.

2.1. Effect of dependence on the false discoveries

The number of false discoveries $V(t)$ is an important quantity in multiple testing. It is a realized but unobservable value for a given experiment. To gain insight into how the dependence of test statistics impacts on the number of false discoveries, we first illustrate this by a simple example: the test statistic depends on a common unobservable factor W in the model

$$Z_i = \mu_i^* + b_i W + (1 - b_i^2)^{1/2}\varepsilon_i \sim N(\mu_i^*, 1), \tag{1}$$

where W and $\{\varepsilon_i\}_{i=1}^n$ are independent, having the standard normal distribution. Let z_α be the α-quantile of the standard normal distribution and $\mathcal{N} = \{i : \mu_i^* = 0\}$ is the true null set. Then,

$$V(t) = \sum_{i \in \mathcal{N}} I(|Z_i| > -z_{t/2}) = \sum_{i \in \mathcal{N}} [I\{\varepsilon_i > a_i(-z_{t/2} - b_i W)\} + I\{\varepsilon_i < a_i(z_{t/2} - b_i W)\}],$$

where $a_i = (1 - b_i^2)^{-1/2}$. By using the law of large numbers, conditioning on W, under some mild conditions, we have

$$p_0^{-1} V(t) = p_0^{-1} \sum_{i \in \mathcal{N}} [\Phi\{a_i(z_{t/2} + b_i W)\} + \Phi\{a_i(z_{t/2} - b_i W)\}] + o_p(1). \tag{2}$$

The dependence of $V(t)$ on the realization W is evidenced in equation (2). For example, if $b_i = \rho$,

$$p_0^{-1} V(t) = \Phi\left\{\frac{z_{t/2} + \rho W}{\sqrt{(1 - \rho^2)}}\right\} + \Phi\left\{\frac{z_{t/2} - \rho W}{\sqrt{(1 - \rho^2)}}\right\} + o_p(1). \tag{3}$$

When $\rho = 0$, $p_0^{-1} V(t) \approx t$ as expected. To quantify the dependence on the realization of W, let $p_0 = 1000$ and $t = 0.01$ and $\rho = 0.8$ so that

$$p_0^{-1} V(t) \approx [\Phi\{(-2.236 + 0.8W)/0.6\} + \Phi\{(-2.236 - 0.8W)/0.6\}].$$

When $W = -3, -2, -1, 0$, the values of $p_0^{-1}V(t)$ are approximately 0.608, 0.145, 0.008 and 0 respectively, which depend heavily on the realization of W. This is in contrast with the independence case in which $p_0^{-1}V(t)$ is always approximately 0.01.

Despite the dependence of $V(t)$ on the realized random variable W, the common factor can be inferred from the observed test statistics. For example, ignoring sparse μ_i^* in equation (1), we can estimate W via simple least squares: minimizing $\Sigma_{i=1}^p (Z_i - b_i W)^2$ with respect to W. Substituting the estimate into equation (3) and replacing p_0 by p, or more generally substituting the estimate into equation (2) and replacing $\mathcal{N}$ by the entire set, we obtain an estimate of $V(t)$ under dependence. A robust implementation is to use L_1-regression which finds the W to minimize $\Sigma_{i=1}^p |Z_i - b_i W|$ or to use penalized least squares such as $\Sigma_{i=1}^p (Z_i - \mu_i - b_i W)^2 + \lambda \Sigma_{i=1}^p |\mu_i|$ to explore the sparsity of μ. This is the basic idea behind Fan, Han and Gu (2012).

2.2. Principal factor approximation

The PFA, which was introduced by Fan, Han and Gu (2012), is a generalization of the idea in Section 2.1. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of correlation matrix $\mathbf{\Sigma}$ in non-increasing order, and $\gamma_1, \ldots, \gamma_p$ be their corresponding eigenvectors. For a given integer k, decompose $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^{\mathrm{T}} + \mathbf{A},$$

where $\mathbf{B} = (\sqrt{\lambda_1}\gamma_1, \ldots, \sqrt{\lambda_k}\gamma_k)$ are unnormalized first k principal components and

$$\mathbf{A} = \sum_{i=k+1}^{p} \lambda_i \gamma_i \gamma_i^{\mathrm{T}}.$$

Correspondingly, decompose the test statistics $\mathbf{Z} \sim N(\boldsymbol{\mu}^*, \mathbf{\Sigma})$ stochastically as

$$\mathbf{Z} = \boldsymbol{\mu}^* + \mathbf{B}\mathbf{W} + \mathbf{K}, \tag{4}$$

where $\mathbf{W} \sim N_k(0, \mathbf{I}_k)$ are k common factors and $\mathbf{K} \sim N(0, \mathbf{A})$ are the errors, independent of $\mathbf{W}$. Define the oracle FDP(t) as

$$\mathrm{FDP}_{\mathrm{oracle}}(t) = \sum_{i \in \{\text{true nulls}\}} \frac{\Phi\{a_i(z_{t/2} + \eta_i)\} + \Phi\{a_i(z_{t/2} - \eta_i)\}}{R(t)} \tag{5}$$

where $a_i = (1 - |\mathbf{b}_i|^2)^{-1/2}$, $\eta_i = \mathbf{b}_i^{\mathrm{T}}\mathbf{W}$ and $\mathbf{b}_i^{\mathrm{T}}$ is the ith row of $\mathbf{B}$. This is clearly a generalization of equation (2). Then, an examination of the proof of Fan, Han and Gu (2012) yields the following result.

Proposition 1. If $p^{-1}\sqrt{(\lambda_{k+1}^2 + \ldots + \lambda_p^2)} = O(p^{-\delta})$ for some $\delta > 0$ (condition 1), then, on the event $\{p^{-1}R(t) > cp^{-\theta}\}$ for some $c > 0$ and $\theta \geqslant 0$, we have

$$|\mathrm{FDP}_{\mathrm{oracle}}(t) - \mathrm{FDP}(t)| = O_p(p^{-(\delta/2-\theta)}).$$

Proposition 1 was established in the proof of theorem 1 of Fan, Han and Gu (2012) under condition 1 and the assumption that $\theta = 0$. Here we allow $\theta > 0$ and $R(t)$ can stochastically grow slower than p. Suppose that we choose $k' > k$. Then by condition 1 it is easy to see that the associated convergence rate is no slower than $p^{-(\delta/2-\theta)}$. This explains why, with more common factors in model (4), $|\mathrm{FDP}_{\mathrm{oracle}}(t) - \mathrm{FDP}(t)|$ converges to 0 faster as $p \to \infty$. This result will be useful for the discussion about determining the number of factors in Section 3.1. Condition 1 in proposition 1 implies that, if $|\mathbf{\Sigma}| = o(p^{1/2})$, we can take $k = 0$. In other words, $|\mathbf{\Sigma}| = o(p^{1/2})$ can be regarded as the condition for weak dependence of the multiple-testing problem. For the mean-square convergence of $V(t)$, see Azriel and Schwartzman (2015).

Since we do not know which co-ordinates of $\boldsymbol{\mu}$ vanish, $\mathrm{FDP}_{\mathrm{oracle}}(t)$ can be approximated by

$$\mathrm{FDP}_{\mathrm{A}}(t) = \sum_{i=1}^{p} \frac{\Phi\{a_i(z_{t/2} + \eta_i)\} + \Phi\{a_i(z_{t/2} - \eta_i)\}}{R(t)}. \tag{6}$$

This provides a useful upper bound for estimating $\mathrm{FDP}(t)$. For the independence case, in which $a_i = 1$ and $\|\mathbf{b}_i\| = 0$, $\mathrm{FDP}_{\mathrm{oracle}}(t) = p_0 t / R(t)$. It can be consistently estimated by estimating one parameter p_0. For the dependence case, however, we need to know the whole set of 'true null hypotheses' and this is an impossible task. Therefore the upper bound becomes an estimable statistical quantity that is frequently used in practice.

The PFA method of Fan, Han and Gu (2012) is to define

$$\widehat{\mathrm{FDP}}_{\mathrm{A}}(t) = \sum_{i=1}^{p} \frac{\Phi\{a_i(z_{t/2} + \tilde{\eta}_i)\} + \Phi\{a_i(z_{t/2} - \tilde{\eta}_i)\}}{R(t)}, \tag{7}$$

where $\tilde{\eta}_i = \mathbf{b}_i^{\mathrm{T}} \hat{\mathbf{W}}$ for an estimator $\hat{\mathbf{W}}$ of $\mathbf{W}$. Then, under mild conditions, Fan, Han and Gu (2012) showed that $|\widehat{\mathrm{FDP}}_{\mathrm{A}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = O_p(\|\hat{\mathbf{W}} - \mathbf{W}\|)$.

For the estimation of $\mathbf{W}$, since $\boldsymbol{\mu}^*$ is sparse, we can consider the following penalized least squares estimator based on model (4). Namely, $\hat{\mathbf{W}}$ is obtained by minimizing

$$\sum_{i=1}^{p} (z_i - \mu_i^* - \mathbf{b}_i^{\mathrm{T}} \mathbf{W})^2 + \sum_{i=1}^{p} p_\lambda(|\mu_i^*|) \tag{8}$$

with respect to $\boldsymbol{\mu}^*$ and $\mathbf{W}$, where p_λ can be the L_1 or smoothly clipped absolute deviation penalty function (Fan and Li, 2001). When $p_\lambda(|\mu_i^*|) = \lambda|\mu_i^*|$, the optimization problem in equation (8) is equivalent to

$$\min_{\mathbf{W}} \sum_{i=1}^{p} \psi(z_i - \mathbf{b}_i^{\mathrm{T}} \mathbf{W}) \tag{9}$$

where $\psi(\cdot)$ is the Huber loss function (Fan, Tang and Shi, 2012). Fan, Han and Gu (2012) also considered an alternative loss function for problem (9): the least absolute deviation loss

$$\min_{\mathbf{W}} \sum_{i=1}^{p} |z_i - \mathbf{b}_i^{\mathrm{T}} \mathbf{W}|. \tag{10}$$

Fan, Tang and Shi (2012) studied problem (8) rigorously. They showed that the penalized estimator of $\mathbf{W}$ is consistent and that its asymptotic distributions are Gaussian.

2.3. Principal factor approximation with unknown covariance

$\widehat{\mathrm{FDP}}_{\mathrm{A}}(t)$ in equation (7) is based on eigenvalues $\{\lambda_i\}_{i=1}^k$ and eigenvectors $\{\boldsymbol{\gamma}_i\}_{i=1}^k$ of the true covariance matrix $\boldsymbol{\Sigma}$. When $\boldsymbol{\Sigma}$ is unknown, we need an estimate $\hat{\boldsymbol{\Sigma}}$. Let $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ be eigenvalues of $\hat{\boldsymbol{\Sigma}}$ in a non-increasing order and $\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_p \in \mathbb{R}^p$ be their corresponding eigenvectors. We can obtain an approximation of the FDP by substituting unknown eigenvalues and eigenvectors in equation (7) by their corresponding estimates. Two questions arise naturally.

(a) What are the requirements for the estimates of $\{\lambda_i\}_{i=1}^k$ and $\{\boldsymbol{\gamma}_i\}_{i=1}^k$ such that $|\widehat{\mathrm{FDP}}_{\mathrm{A}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = o_p(1)$?
(b) Under what dependence structures of $\boldsymbol{\Sigma}$ can such estimates of $\{\lambda_i\}_{i=1}^k$ and $\{\boldsymbol{\gamma}_i\}_{i=1}^k$ be constructed?

The current paper will address these two questions.

3. Main result

We first present the results for a generic estimator $\hat{\mathbf{\Sigma}}$, and then we consider a special example in this general framework, the approximate factor model, to illustrate the effect of unknown dependence on the testing procedure.

3.1. Accuracy required

Suppose that condition 1 is satisfied for $\mathbf{\Sigma}$. Let $\hat{\mathbf{\Sigma}}$ be an estimator of $\mathbf{\Sigma}$, and correspondingly we have $\{\hat{\lambda}_i\}_{i=1}^k$ and $\{\hat{\boldsymbol{\gamma}}_i\}_{i=1}^k$ to estimate $\{\lambda_i\}_{i=1}^k$ and $\{\boldsymbol{\gamma}_i\}_{i=1}^k$. Analogously, we define $\hat{\mathbf{B}}$ and $\hat{\mathbf{b}}_i$. Note that we need to estimate only the first k eigenvalues and eigenvectors but not all of them.

The realized common factors $\mathbf{W}$ can be estimated robustly by using expressions (8) and (9) with $\mathbf{b}_i$ replaced by $\hat{\mathbf{b}}_i$. To simplify the technical arguments, we simply use the least squares estimate

$$\hat{\mathbf{W}} = (\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{Z}, \tag{11}$$

which ignores the $\boldsymbol{\mu}^*$ in equation (4) and replaces $\mathbf{B}$ by $\hat{\mathbf{B}}$. Define

$$\widehat{\mathrm{FDP}}_{\mathrm{U}}(t) = \sum_{i=1}^{p} \frac{\Phi\{\hat{a}_i(z_{t/2} + \hat{\eta}_i)\} + \Phi\{\hat{a}_i(z_{t/2} - \hat{\eta}_i)\}}{R(t)} \tag{12}$$

where $\hat{a}_i = (1 - \|\hat{\mathbf{b}}_i\|^2)^{-1/2}$ and $\hat{\eta}_i = \hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{W}}$. Then we have the following result.

Theorem 1. On the event $\mathcal{E}$ that

(a) $R(t)^{-1} = O(p^{-(1-\theta)})$ for some $\theta \geqslant 0$ (condition 2),
(b) $\max_{i \leqslant k} \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\| = O(p^{-\kappa})$ for $\kappa > 0$ (condition 3),
(c) $\Sigma_{i=1}^k |\hat{\lambda}_i - \lambda_i| = O(p^{1-\nu})$ for $\nu > 0$ (condition 4) and
(d) $\hat{a}_i \leqslant \tau_1$ and $a_i \leqslant \tau_2$ $\forall\, i = 1, \ldots, p$ for some finite constants τ_1 and τ_2 (condition 5),

we have

$$|\widehat{\mathrm{FDP}}_{\mathrm{U}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = O_p\{p^{\theta}(p^{-\nu} + kp^{-\kappa} + \|\boldsymbol{\mu}^*\|p^{-1/2})\}.$$

Note that $\mathrm{FDP}(t) = V(t)/R(t)$ in which $R(t)$ is observable and known. Approximating $\mathrm{FDP}(t)$ amounts to approximating $V(t)$, which does not rely on condition 2. In high dimensional applications, t can be chosen to decrease slowly with p, as in Donoho and Jin (2004, 2006). Our result on the approximation of $V(t)$ continues to hold for t that depends on p, i.e. t_p. If condition 2 holds for t_p, then theorem 1 follows for t_p.

Using $\Sigma_{i=1}^k \lambda_i \leqslant \mathrm{tr}(\mathbf{\Sigma}) = p$, we have $\Sigma_{i=1}^k |\hat{\lambda}_i - \lambda_i| \leqslant p \max_{i \leqslant k} |\hat{\lambda}_i/\lambda_i - 1|$. Thus, condition 4 holds with high probability when $\max_{i \leqslant k} |\hat{\lambda}_i/\lambda_i - 1| = O_p(p^{-\nu})$, which is particularly relevant when eigenvalues are spiked. The third term in the convergence result comes really from the least squares estimate. If a more sophisticated method such as expression (8) or (9) is used, the bias will be smaller (Fan, Tang and Shi, 2012). We do not plan to pursue this line to facilitate the presentation.

In theorem 1, we assume that the number of factors k is known. When k must be estimated, we shall apply the eigenvalue ratio (ER) estimator in Ahn and Horenstein (2013). The ER estimator is defined as $\hat{k}_{\mathrm{ER}} = \arg\max_{1 \leqslant k \leqslant k_{\max}} (\tilde{\lambda}_k/\tilde{\lambda}_{k+1})$, where $\tilde{\lambda}_i$ is the ith largest eigenvalue of the sample covariance matrix and $k_{\max}$ is the maximum possible number of factors. Under mild regularity conditions, this estimator has been shown to be consistent. A similar idea was also adopted by Lam and Yao (2012). Therefore, to simplify the presentation, we shall use a known k for the theoretical development in the current paper, but for the numerical studies in Section 4 and 5 we shall apply the ER estimator for estimating k. An overestimate of k does not

do as much harm to approximating the FDP, as long as the unobserved factors are estimated with reasonable accuracy. This is because condition 1 is also satisfied for a larger k and will be verified via simulation. In contrast, an underestimate of k can result in the approximate FDP with inferior performance, due to missing important factors to capture dependence.

3.2. Effect of estimating marginal variances

In the previous sections, we assume that $\mathbf{\Sigma}$ is a correlation matrix. In practice, the marginal variances $\{\sigma_j^2\}$ are unknown and need to be estimated. These estimates are used to normalize the testing problem. Suppose that $\{\hat{\sigma}_j^2\}_{j=1}^p$ are the diagonal elements of $\hat{\mathbf{\Sigma}}$, an estimate of $\mathbf{\Sigma}$. Conditioning on $\{\hat{\sigma}_j\}_{j=1}^p$, assume that $\hat{\mathbf{D}}^{-1}\sqrt{n}\bar{\mathbf{X}} \sim N(\sqrt{n}\hat{\mathbf{D}}^{-1}\boldsymbol{\mu}, \tilde{\mathbf{\Sigma}})$, $\tilde{\mathbf{\Sigma}} = \hat{\mathbf{D}}^{-1}\mathbf{\Sigma}\hat{\mathbf{D}}^{-1}$, where $\hat{\mathbf{D}} = \mathrm{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_p)$. When $\hat{\mathbf{\Sigma}}$ is the sample covariance matrix, it is well known that $\hat{\mathbf{\Sigma}}$ and $\bar{\mathbf{X}}$ are independent and the aforementioned assumption holds. Then $\tilde{\mathbf{\Sigma}}$ is approximately the same as the correlation matrix as long as $\{\hat{\sigma}_j\}_{j=1}^p$ converges uniformly to $\{\sigma_j\}_{j=1}^p$. Thanks to the Gaussian tails, this indeed holds for the sequence of the marginal sample covariances (Bickel and Levina, 2008a). Our simulations show the small effect of estimating the marginal variances.

The unconditional distribution of $\hat{\mathbf{D}}^{-1}\sqrt{n}\bar{\mathbf{X}}$ is not a multivariate normal distribution. To address this issue, let $\bar{X}_{(j)} = n^{-1}\Sigma_{i=1}^n X_{ij}$ and $\hat{\sigma}_j^2 = (n-1)^{-1}\Sigma_{i=1}^n (X_{ij} - \overline{X}_{(j)})^2$ and consider the standardized test statistics $T_j = \sqrt{n}\bar{X}_{(j)}/\hat{\sigma}_j$. Then, for the true null hypotheses, each T_j follows the t_{n-1}-distribution, and (T_j, T_l) have a bivariate t-distribution. See Siddiqui (1967). However, $\{T_j\}_{j=1}^p$ do not follow the multivariate t-distribution that was introduced in Kotz and Nadarajah (2004), because $\{\hat{\sigma}_j\}_{j=1}^p$ are also dependent on each other through $\mathbf{\Sigma}$. Therefore, in the following presentation, we shall call the joint distribution of $\{T_j\}_{j=1}^p$ a dependent t-distribution rather than a multivariate t-distribution to avoid any confusion. Let $F_{n-1}(\cdot)$ denote the cumulative distribution function of a t_{n-1} random variable, and let $q_{t/2}$ denote the $t/2$-quantile of F_{n-1}. The p-values are calculated as $P_j = 2F_{n-1}(-|T_j|)$. We use threshold t and reject the jth hypothesis if $P_j \leqslant t$.

Similarly to the definition of $\widehat{\mathrm{FDP}}_{\mathrm{U}}(t)$ in Section 3.1, we use the least squares estimate

$$\hat{\mathbf{W}}_{\mathrm{G}} = (\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{T},$$

where $\mathbf{T} = (T_1, \ldots, T_p)^{\mathrm{T}}$. Define $\widehat{\mathrm{FDP}}_{\mathrm{U,G}}(t) = \Sigma_{i=1}^p [\Phi\{\hat{a}_i(z_{t/2} + \hat{\eta}_{i,\mathrm{G}})\} + \Phi\{\hat{a}_i(z_{t/2} - \hat{\eta}_{i,\mathrm{G}})\}]/R(t)$, where $\hat{\eta}_{i,\mathrm{G}} = \hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{W}}_{\mathrm{G}}$. In these expressions, $\hat{\mathbf{B}}$, $\hat{\mathbf{b}}_i$ and $\hat{a}_i$ are calculated on the basis of the estimated correlation matrix of $\mathbf{X}$, and the subscript 'G' represents general covariance matrix $\mathbf{\Sigma}$.

Theorem 2. On the basis of the test statistics $\{T_j\}_{j=1}^p$, suppose that the correlation matrix of $\mathbf{X}$ satisfies condition 1. Then, on the event $\mathcal{E}$ in theorem 1, we have

$$|\mathrm{FDP}_{\mathrm{oracle}}(t) - \mathrm{FDP}(t)| = O_p\{p^{\theta}(p^{-\delta/2} + n^{-1/2})\},$$

where $\mathrm{FDP}_{\mathrm{oracle}}(t)$ is defined in equation (5) and

$$|\widehat{\mathrm{FDP}}_{\mathrm{U,G}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = O_p\{p^{\theta}(p^{-\nu} + kp^{-\kappa} + \|\boldsymbol{\mu}^*\| p^{-1/2} + n^{-1/2})\},$$

where $\mathrm{FDP}_{\mathrm{A}}(t)$ is defined in equation (6) corresponding to the correlation matrix of $\mathbf{X}$.

The first result in theorem 2 is similar to proposition 1, except for a term from the effect of the sample size n. This result suggests that, under some mild conditions, we can still apply the PFA method even if the effect of the marginal variance is considered. In the second result of theorem 2, $\{\hat{\lambda}_i\}$ and $\{\hat{\gamma}_i\}$ correspond to the estimated correlation matrix of $\mathbf{X}$, and $\{\lambda_i\}$ and $\{\gamma_i\}$ correspond to the population correlation matrix of $\mathbf{X}$. This result is very similar to that established in theorem 1. Therefore, to simplify the discussion and to highlight the effect of the estimator $\hat{\mathbf{\Sigma}}$ on the testing procedure, we shall assume in the following Sections 3.3–3.5 that the

diagonal elements of $\mathbf{\Sigma}$ are known and equal to 1. The simulation studies in Section 4 are still based on the set-up that $\mathbf{\Sigma}$ has general and unknown diagonal elements.

A direct derivation of the density function for the bivariate t random variables is complicated and not useful for our proof. The proof of theorem 2 is based on a Bayesian interpretation of bivariate t-distributions. The method is general and can be of independent interest for extending results under normality to dependent t distributions.

3.3. Results on eigenvectors and eigenvalues

In theorem 1, the rate of convergence of $\widehat{\mathrm{FDP}}_{\mathrm{U}}(t)$ critically depends on the estimated eigenvalues and eigenvectors. In the current section, we shall study under what situations conditions 3 and 4 can be satisfied.

Lemma 1. For any matrix $\hat{\mathbf{\Sigma}}$, we have

$$|\hat{\lambda}_i - \lambda_i| \leqslant \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|,$$

$$\|\hat{\gamma}_i - \gamma_i\| \leqslant \frac{\sqrt{2}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|}{\min(|\hat{\lambda}_{i-1} - \lambda_i|, |\lambda_i - \hat{\lambda}_{i+1}|)}.$$

The first result is referred to as Weyl's theorem (Horn and Johnson, 1990) and the second result is called the $\sin(\theta)$ theorem (Davis and Kahan, 1970). They have been applied in sparse covariance matrix estimation (El Karoui, 2008; Ma, 2013). By lemma 1, the consistency of eigenvectors and eigenvalues is directly associated with operator norm consistency. Several references have shown that, under various conditions on $\mathbf{\Sigma}$, $\hat{\mathbf{\Sigma}}$ can be constructed such that $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| \to 0$, which will be discussed in more detail after the following theorem 3.

Theorem 3. If $\lambda_i - \lambda_{i+1} \geqslant d_p$ for a sequence $d_p > 0$ for $i = 1, \ldots, k$, then on the event $\mathcal{E} \cap \{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O(d_p p^{-\tau})\}$ for some $\tau > 0$, for sufficiently large p, we have

$$|\widehat{\mathrm{FDP}}_{\mathrm{U}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = O_p[p^{\theta}\{kp^{-\tau}d_p/p + (k+1)p^{-\tau} + \|\boldsymbol{\mu}^*\|p^{-1/2}\}].$$

Note that the first k eigenvalues should be distinguishable by a certain amount of gap d_p. Theorem 3 is so written that it is applicable to both spike or non-spike cases. For the non-spike case, typically $d_p = d > 0$. In this case, the covariance is estimated consistently and the first term in theorem 3 now becomes $O_p(kp^{-\tau-1})$. For the spiked case such as the k-factor model (4), the first k eigenvalues are of order p and the $(k+1)$th eigenvalue is of order 1 (Fan *et al.*, 2013). Therefore, $d_p \asymp p$. In this case, the covariance matrix cannot be consistently estimated, and the first term is of order $O(kp^{-\tau})$. See Section 3.4 for additional details.

Depending on the structures of $\mathbf{\Sigma}$ and different choices of $\hat{\mathbf{\Sigma}}$, we shall have different requirements such that the event $\{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O(d_p p^{-\tau})\}$ occurs with high probability. It is impossible for us to list all the references in the area of large covariance matrix estimation, but we shall focus on several representative classes of $\mathbf{\Sigma}$-structures and present relevant results.

(a) *Banded matrix*: Bickel and Levina (2008a) considered a class of banded matrices with decaying rate α. After banding the sample covariance matrix, they constructed an estimator $\hat{\mathbf{\Sigma}}_1$, which has operator norm convergence rate $\|\hat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}\| = O_p[\{\log(p)/n\}^{\alpha/(2\alpha+2)}]$.

(b) *Sparse matrix*: Bickel and Levina (2008b) considered a class of sparse covariance matrices with sparsity parameters $c_0(p)$ and q where $0 \leqslant q \leqslant 1$. With a thresholding technique, they constructed an estimator $\hat{\mathbf{\Sigma}}_2$ which satisfies $\|\hat{\mathbf{\Sigma}}_2 - \mathbf{\Sigma}\| = O_p[c_0(p)\{\log(p)/n\}^{(1-q)/2}]$. In the special case when $q = 0$ and $c_0(p)$ is bounded, this convergence rate is $\{\log(p)/n\}^{1/2}$.

(c) *Sparse precision matrix*: Cai *et al.* (2011) considered a class of sparse precision matrices $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ with sparsity parameters $s_0(p)$ and q. By a constrained l_1-minimization

approach, they constructed an estimator $\hat{\boldsymbol{\Omega}}_3$ such that $\|\hat{\boldsymbol{\Omega}}_3 - \boldsymbol{\Omega}\| = O_p[s_0(p) \times \{\log(p)/n\}^{(1-q)/2}]$. Furthermore, for $\hat{\boldsymbol{\Sigma}}_3 = \hat{\boldsymbol{\Omega}}_3^{-1}$, under some mild conditions, it is easy to show that $\|\hat{\boldsymbol{\Sigma}}_3 - \boldsymbol{\Sigma}\| = O_p[s_0(p)\{\log(p)/n\}^{(1-q)/2}]$.

It is worth mentioning that the convergence rate $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$ leads to some requirement on the sample size n. For example, in the special case of the sparse matrix when $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| = O_p[\{\log(p)/n\}^{1/2}]$, if it also satisfies the condition in theorem 3 that $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| = O_p(p^{-\tau})$, then the sample size n must be greater than $p^{2\tau}\log(p)$. This requirement of n is of major importance in practice.

3.4. Approximate factor model

We shall study the multiple-testing problem where the test statistics have some strong dependence structure as a special example of theorem 3. Assume that the dependence of a high dimensional variable vector of interest can be captured by a few latent factors. This factor structure model has a long history in financial econometrics (Engle and Watson, 1981; Bai, 2003). It has also received considerable attention in genomic research (Friguet *et al.*, 2009; Desai and Storey, 2012). Major restrictions in these models are that the idiosyncratic errors are independent. A more practicable extension is the approximate factor model (Chamberlain and Rothschild, 1983; Fan *et al.*, 2011, 2013).

The approximate factor model takes the form

$$\mathbf{X}_i = \boldsymbol{\mu} + \mathbf{B}\mathbf{f}_i + \mathbf{u}_i, \qquad i = 1, \ldots, n, \tag{13}$$

for each observation, where $\boldsymbol{\mu}$ is a p-dimensional unknown sparse vector, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)^{\mathrm{T}}$ is the factor loading matrix and $\mathbf{f}_i$ is a vector of common factors to the ith observations, independent of the noise $\mathbf{u}_i \sim N_p(0, \boldsymbol{\Sigma}_u)$ where $\boldsymbol{\Sigma}_u$ is sparse. The unobserved common factors $\mathbf{f}_i$ drive the dependence of the measurements (e.g. gene expressions) within the ith sample. Under model (13), the covariance matrix of $\mathbf{X}_i$ is given by $\boldsymbol{\Sigma} = \mathbf{B}\,\mathrm{cov}(\mathbf{f})\mathbf{B}^{\mathrm{T}} + \boldsymbol{\Sigma}_u$. We can also assume without loss of generality the identifiability condition $\mathrm{cov}(\mathbf{f}) = \mathbf{I}_K$ and that the columns of $\mathbf{B}$ are orthogonal. See Fan *et al.* (2013).

For the random errors $\mathbf{u}$, let $\sigma_{u,ij}$ be the (i, j)th element of covariance matrix $\boldsymbol{\Sigma}_u$ of $\mathbf{u}$. Then we impose a sparsity condition on $\boldsymbol{\Sigma}_u$:

$$m_p = \max_{i \leqslant p} \sum_{j \leqslant p} |\sigma_{u,ij}|^q, \qquad m_p = o(p), \text{ for some } q \in [0, 1). \tag{14}$$

Under model (13), the test statistics $\mathbf{X}^* = \sqrt{n}\bar{\mathbf{X}}$ follow the approximate factor model

$$\mathbf{X}^* = \boldsymbol{\mu}^* + \mathbf{B}\mathbf{f}^* + \mathbf{u}^* \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}), \tag{15}$$

where $\boldsymbol{\mu}^* = \sqrt{n}\boldsymbol{\mu}$, $\mathbf{f}^* = \sqrt{n}\bar{\mathbf{f}}$ and $\mathbf{u}^* = \sqrt{n}\bar{\mathbf{u}}$ with $\bar{\mathbf{f}}$ and $\bar{\mathbf{u}}$ being the corresponding mean vector.

Fan *et al.* (2013) developed a method called POET to estimate the unknown $\boldsymbol{\Sigma}$ on the basis of samples $\{\mathbf{X}_i\}_{i=1}^n$ in equation (13). The basic idea is to take advantage of the factor model structure and the sparsity of the covariance matrix of idiosyncratic noise. Their idea combined with the PFA in Fan, Han and Gu (2012) yields the following *POET–PFA method*.

(a) Compute sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ and decompose $\hat{\boldsymbol{\Sigma}} = \Sigma_{i=1}^p \tilde{\lambda}_i \tilde{\boldsymbol{\gamma}}_i \tilde{\boldsymbol{\gamma}}_i^{\mathrm{T}}$, where $\{\tilde{\lambda}_i\}$ and $\{\tilde{\boldsymbol{\gamma}}_i\}$ are the eigenvalues and eigenvectors of $\hat{\boldsymbol{\Sigma}}$. Apply a thresholding method to $\Sigma_{i=k+1}^p \tilde{\lambda}_i \tilde{\boldsymbol{\gamma}}_i \tilde{\boldsymbol{\gamma}}_i^{\mathrm{T}}$ to obtain $\hat{\boldsymbol{\Sigma}}_u^{\mathcal{T}}$ (e.g. the adaptive thresholding method in the on-line supplementary materials with a general thresholding function in Antoniadis and Fan (2001)). Set $\hat{\boldsymbol{\Sigma}}_{\mathrm{POET}} = \Sigma_{i=1}^k \tilde{\lambda}_i \tilde{\boldsymbol{\gamma}}_i \tilde{\boldsymbol{\gamma}}_i^{\mathrm{T}} + \hat{\boldsymbol{\Sigma}}_u^{\mathcal{T}}$.

(b) Apply singular value decomposition to $\hat{\mathbf{\Sigma}}_{\mathrm{POET}}$. Obtain its eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$ in non-increasing order and the associated eigenvectors $\hat{\gamma}_1, \ldots, \hat{\gamma}_K$.
(c) Construct $\hat{\mathbf{B}} = (\hat{\lambda}_1^{1/2}\hat{\gamma}_1, \ldots, \hat{\lambda}_K^{1/2}\hat{\gamma}_K)$ and compute the least squares $\hat{\mathbf{f}}^* = (\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\sqrt{n}\bar{\mathbf{X}}$, which is the least squares estimate from model (15) with $\boldsymbol{\mu}^*$ ignored.
(d) With $\hat{\mathbf{b}}_i^{\mathrm{T}}$ denoting the ith row of $\hat{\mathbf{B}}$, compute

$$\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t) = \sum_{i=1}^{p} \frac{\Phi\{\hat{a}_i(z_{t/2} + \hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{f}}^*)\} + \Phi\{\hat{a}_i(z_{t/2} - \hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{f}}^*)\}}{R(t)} \tag{16}$$

for some threshold value t, where $\hat{a}_i = (1 - \|\hat{\mathbf{b}}_i\|^2)^{-1/2}$.

The convergence rate of $\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t)$ is as follows. Under assumptions 1–4 in the on-line supplementary materials, lemma 2 there holds with high probability. Call this event $\mathcal{E}^*$. Let $\mathcal{E}_1$ be the event that conditions 2 and 5 are satisfied.

Theorem 4. For the POET–PFA method, we have

$$|\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = O_p[p^{\theta}\{k(\omega_p + m_p\omega_p^{1-q}p^{-1}) + \|\boldsymbol{\mu}^*\|p^{-1/2}\}],$$

on the event $\mathcal{E}_1 \cap \mathcal{E}^*$, where $\omega_p = p^{-1/2} + \sqrt{\{\log(p)/n\}}$.

Theorem 4 can be considered as a corollary of theorems 1 and 3. However, since POET–PFA is the method that we recommend, we would like to state it as a theorem to emphasize its importance. It is worth noting that here $|\mathrm{FDP}_{\mathrm{oracle}}(t) - \mathrm{FDP}(t)| = O_p(p^{\theta}m_p^{1/2}p^{-1/2})$ by the examination of the proof of proposition 2 in Fan, Han and Gu (2012).

3.5. Dependence-adjusted procedure

The p-value of each test is determined completely by individual Z_i, which ignores the correlation structure. This method can be inefficient, as Fan, Han and Gu (2012) pointed out. This section shows how to use the dependent structure to improve the power of the test and how to provide an alternative ranking of statistical significance from ranking $\{|Z_i|\}_{i=1}^p$ under dependence.

Under model (4), $a_i(Z_i - \mathbf{b}_i^{\mathrm{T}}\mathbf{W}) \sim N(a_i\mu_i, 1)$. Since $a_i > 1$, this increases the strength of signals and provides an alternative ranking of the significance of each hypothesis. Indeed, the p-value based on this adjusted test statistic is now $2\Phi\{-|a_i(Z_i - \mathbf{b}_i^{\mathrm{T}}\mathbf{W})|\}$ and the null hypothesis H_{i0} is rejected when it is no larger than t. In other words, the critical region is $|a_i(Z_i - \mathbf{b}_i^{\mathrm{T}}\mathbf{W})| \leqslant |z_{t/2}|$. When the covariance matrix $\mathbf{\Sigma}$ is unknown, we calculate the p-values as

$$P_i = 2\Phi\{-|\hat{a}_i(Z_i - \hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{W}})|\},$$

where $\hat{a}_i$, $\hat{\mathbf{b}}_i$ and $\hat{\mathbf{W}}$ have been defined in equations (11) and (12). The theoretical investigation of this procedure is beyond the scope of the current paper. We shall show in simulation studies that this dependence-adjusted procedure is still more powerful than the fixed threshold procedure.

4. Simulation studies

In the simulation studies, we consider dimensionality $p = 1000$, sample sizes $n = 50, 100, 200$, the number of false null hypotheses $p_1 = 50$, threshold value $t = 0.01$ and the number of simulation rounds 500, unless stated otherwise. The data are generated from $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ except in the following model 3. The signal strength μ_i equals 1 for $i = 1, \ldots, 50$ and 0 otherwise. To investigate the effect of signal strength, we also consider non-zero μ_i as 0.8 and 1.2. For brevity, these results are shown in the on-line supplementary materials. We estimate the unknown number of factors k for POET–PFA by the data-driven eigenvalue ratio method that was described in Section 3.1

Table 1. Empirical mean absolute error between the true FDP(t) and $\widehat{\text{FDP}}(t)$†

n	*Results (%) for the following methods:*						
	POET–PFA	*Efron*	*FAMT*	*FAMT–PFA*	*HF–PFA*	*SS–PFA*	*LW–PFA*
Model 1							
50	4.39	19.72	11.48	5.90	5.40	6.95	5.94
100	3.66	19.53	10.26	4.91	4.83	4.90	4.56
200	3.34	19.58	11.86	5.33	3.60	3.85	3.71
Model 2							
50	5.09	17.49	10.15	5.56	5.69	7.49	6.93
100	3.93	17.80	11.42	5.61	5.53	5.28	5.14
200	3.81	18.37	10.49	5.17	5.11	4.24	4.20
Model 3							
50	5.61	15.05	12.23	6.67	6.29	7.57	6.50
100	4.24	14.37	12.69	6.29	5.22	5.87	5.35
200	3.84	14.63	12.27	5.54	4.55	4.77	4.60
Model 4							
50	4.62	19.49	11.40	6.62	5.50	7.26	7.10
100	4.07	19.01	11.25	6.75	5.41	4.97	5.09
200	3.48	18.71	10.14	6.05	3.80	3.94	3.98
Model 5							
50	5.44	10.46	10.09	5.18	6.95	7.38	5.66
100	5.65	10.57	10.64	5.33	6.81	6.46	5.86
200	5.29	10.64	10.76	4.65	7.03	5.78	5.47
Model 6							
50	4.60	10.12	9.84	4.83	4.73	6.08	4.60
100	4.03	9.44	8.59	3.89	3.67	4.82	4.03
200	4.13	9.36	10.20	4.40	4.83	4.45	4.13
Model 7							
50	4.50	10.18	5.88	4.68	4.98	6.24	4.63
100	4.30	10.33	6.19	4.77	4.66	5.29	4.43
200	4.13	9.99	6.17	4.58	5.21	4.66	4.21
Model 8							
50	4.53	11.66	6.35	4.77	5.76	6.72	5.02
100	4.25	11.13	6.30	4.81	5.16	5.26	4.41
200	4.02	10.62	6.01	4.42	6.07	4.59	4.14

†The non-zero $\mu_i = 1$.

with $k_{\max} = \lfloor 0.2n \rfloor$. To demonstrate the wide applicability of POET–PFA compared with other methods, we consider eight different model settings for dependence structures in Table 1 as well as Tables 3 and 4 in the supplementary materials.

4.1. Model 1: strict factor model

Consider a three-factor model

$$\mathbf{x}_i = \boldsymbol{\mu} + \mathbf{B}\mathbf{f}_i + \mathbf{u}_i, \qquad \mathbf{f}_i \sim N_3(0, \mathbf{I}_3) \text{ independent of } \mathbf{u}_i \sim N_p(0, \boldsymbol{\Sigma}_u).$$

Each entry of the factor loading matrix $\mathbf{B}_{ij}$ is an independent realization from the uniform distribution $U(-1, 1)$. In addition, $\boldsymbol{\Sigma}_u = \mathbf{I}_p$.

4.2. Model 2: approximate factor model

The model 2 set-up is the same as for model 1, except that we construct $\mathbf{\Sigma}_u$ as follows. First apply the method in Fan *et al*. (2013) to create a covariance matrix $\mathbf{\Sigma}_1$, which was calibrated to the returns of Standard & Poors 500 index constituent stocks. We omit the details. Then we construct a symmetric banded matrix $\mathbf{\Sigma}_2$. For the (i, j)th element, if $i \neq j$ and $|i-j| \leqslant 25$, set the element as 0.4 and 0 otherwise. Next we construct a symmetric matrix $\mathbf{\Sigma}_3$ as the nearest positive definite matrix of $\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$ by the algorithm of Higham (1988). Finally the covariance matrix $\mathbf{\Sigma}_u$ is set as $0.5\mathbf{\Sigma}_3$.

4.3. Model 3: non-normal model

Consider a five-factor model $\mathbf{x}_i = \boldsymbol{\mu} + \mathbf{B}\mathbf{f}_i + \mathbf{u}_i$. $\mathbf{B}$ is generated similarly to model 1, but each element of $\mathbf{f}_i$ and each element of $\mathbf{u}_i$ are independent realizations from $\sqrt{(2/3)}t_6$ where t_6 is a t-distribution with 6 degrees of freedom. Model 3 is constructed to show the performance of POET–PFA even when the normality assumption for the data-generating process is violated.

4.4. Model 4: cluster model

We first generate a p-dimensional vector Λ, where the first four elements are independent realizations from the uniform distribution $U(160, 190)$, the next 10 elements are independently distributed from $U(8, 12)$ and the rest are independently distributed from $U(0.1, 0.3)$. Next we generate a $p \times p$ matrix $\mathbf{Q}$ in which each element is an independent realization from $N(0, 1)$. Let $\mathbf{\Gamma}$ be the matrix, consisting of eigenvectors of $\mathbf{Q}\Lambda\mathbf{Q}^{\mathrm{T}}$. Finally, let $\mathbf{\Sigma} = \mathbf{\Gamma}\Lambda\mathbf{\Gamma}^{\mathrm{T}}$. Model 4 is designed against the eigengap condition in theorem 3 and also tests the robustness of determining the number of factors.

4.5. Model 5: long memory autocovariance model

Consider $\mathbf{\Sigma}$ where each element is defined as $\Sigma_{ij} = 0.5(||i-j|+1|^{2H} - 2|i-j|^{2H} + ||i-j| - 1|^{2H})$, $1 \leqslant i, j \leqslant p$, with $H = 0.9$. Model 5 is from Bickel and Levina (2008a) and has also recently been considered by Huang and Fryzlewicz (2015) for strong long memory dependence.

4.6. Model 6: normal perturbation model

Consider a symmetric matrix $\mathbf{Q}$ with diagonal elements 1 and each off-diagonal element independent realizations from $N(0.5, 0.1)$. Let $\mathbf{\Sigma}$ be the nearest positive definite matrix of $\mathbf{Q}$ based on the algorithm in Higham (1988). Model 6 is constructed lacking an apparent factor model pattern.

4.7. Model 7: sparse precision matrix model I

Consider the precision matrix $\mathbf{\Omega} = \mathrm{diag}(\mathbf{A}_1, \mathbf{A}_2)$, where $\mathbf{A}_2 = 4\mathbf{I}_{p/2 \times p/2}$ and $\mathbf{A}_1 = \mathbf{B} + \epsilon\mathbf{I}_{p/2 \times p/2}$. $\mathbf{B}$ is a symmetric matrix where each element b_{ij} takes value 0.5 with probability 0.1 and takes value 0 with probability 0.9. $\epsilon = \max\{-\lambda_{\min}(\mathbf{B}), 0\} + 0.01$ to ensure that $\mathbf{A}_1$ is positive definite. Finally, let $\mathbf{\Sigma} = (\mathbf{\Omega})^{-1}$. Construction of $\mathbf{A}_1$ is from Rothman *et al*. (2008) for a sparse precision matrix structure.

4.8. Model 8: sparse precision matrix model II

Consider the precision matrix $\mathbf{\Omega} = \mathrm{diag}(\mathbf{A}_1, \mathbf{A}_2)$ similarly to model 7 except that each b_{ij} takes a value uniformly in $[0.3, 0.8]$ with probability 0.2 and takes value 0 with probability 0.8. Finally, let $\mathbf{\Sigma} = (\mathbf{\Omega})^{-1}$. The sparsity structure in model 8 is from Cai and Liu (2011) but we consider this sparsity structure for the precision matrix. The final $\mathbf{\Sigma}$ is quite different from model 7.

4.9. Comparison with other methods for estimating the false discovery proportion

We compare our POET–PFA method with the methods in Efron (2007) and Fan *et al.* (2013). The latter assumed a strict factor model and used the EM algorithm to estimate the factor loadings $\mathbf{B}$ and the common factors $\{\mathbf{f}_i\}_{i=1}^n$. Correspondingly, Fan *et al.* (2013) constructed an estimator for FDP(t) based on their factor model and multiple-testing method FAMT. To see how well the EM algorithm estimates factor loadings $\hat{\mathbf{B}}$, we include the FAMT–PFA method, which replaces $\hat{\mathbf{B}}$ in the fourth step of our POET–PFA method with that computed by the EM algorithm, for comparison. In the above simulations, we used the R package FAMT from Friguet *et al.* (2009) to obtain the EM-based estimators $\hat{\mathbf{B}}$ and $\{\hat{\mathbf{f}}\}_{i=1}^n$. We further consider other methods for estimating the unknown $\mathbf{\Sigma}$ rather than POET and compare the performance of the corresponding $\widehat{\mathrm{FDP}}$. Exploration in this direction could be endless, and we consider only three representative types of shrinkage estimators here: Huang and Fryzlewicz (2015), HF, Schafer and Strimmer (2005), SS, and Ledoit and Wolf (2003), LW. Although these three methods do not involve estimating the number of factors k for the covariance matrix step, they still need to estimate k for the PFA step. Therefore, we apply the eigenvalue ratio method with their methods for a fair comparison with our POET–PFA method. The results in HF–PFA are based on 50 simulation rounds by its cross-validation-based algorithm NOVELIST. Other results are still based on 500 simulation rounds.

In Table 1, we calculate the empirical mean absolute error (the absolute difference between the true FDP and $\widehat{\mathrm{FDP}}$) for the seven methods. We recall that the FDP is a quantity that is measured in percentages and therefore the measurement unit for the mean absolute error that is reported in Table 1 is per cent. Generally, when the sample size increases, the mean absolute error of POET–PFA tends to be smaller. The results in model 6 seem to be a violation of this statement. However, considering that $\mathrm{FDP}_{\mathrm{A}}(t)$ tends to be an upper bound of FDP, the results here are still reasonable. Overall, our POET–PFA method performs the best compared with the other six methods, in terms of producing smaller mean absolute error. In model 5, FAMT–PFA outperforms POET–PFA; however, further investigation shows that the average of $\widehat{\mathrm{FDP}}$ by FAMT–PFA is an underestimate of the true FDR, whereas our POET–PFA method provides an overestimate, which is better for practical FDR control. Results for signal strengths 0.8 and 1.2 are shown in Tables 3 and 4 in the on-line supplementary materials and are consistent with the findings in Table 1 here.

Fig. 1 further demonstrates the performance of our POET–PFA method involving least squares estimation compared with Efron's method, FAMT and FAMT–PFA under models 1 and 2. The sample size $n = 50$. Our POET–PFA method approximates the true FDP(t) well. Efron's method captures the general trend of FDP(t) when the true values are relatively small and deviates from the true values in the opposite direction when FDP(t) becomes large. FAMT–PFA performs much better than FAMT, but it still could not capture the true value when FDP(t) is large. Comparisons under models 3–8 are shown in the supplementary materials.

4.10. Dependence-adjusted testing procedure

We compare the dependence-adjusted procedure that was described in Section 3.5 with the fixed threshold procedure, i.e. we compare the $|Z_i|$ with a universal threshold without using the correlation information. Define the false negative rate $\mathrm{FNR} = E[T/(p-R)]$ where T is the number of falsely accepted null hypotheses. With the same FDR level, a procedure with a smaller false negative rate is more powerful. Since the advantage of the dependence-adjusted procedure can be better demonstrated by an apparent factor model structure, Table 2 considers only models 1 and 2. In Table 2, we fix the threshold value $t = 0.001$ and reject the hypotheses when the dependence-adjusted p-value is smaller than 0.001. Then we find the corresponding threshold

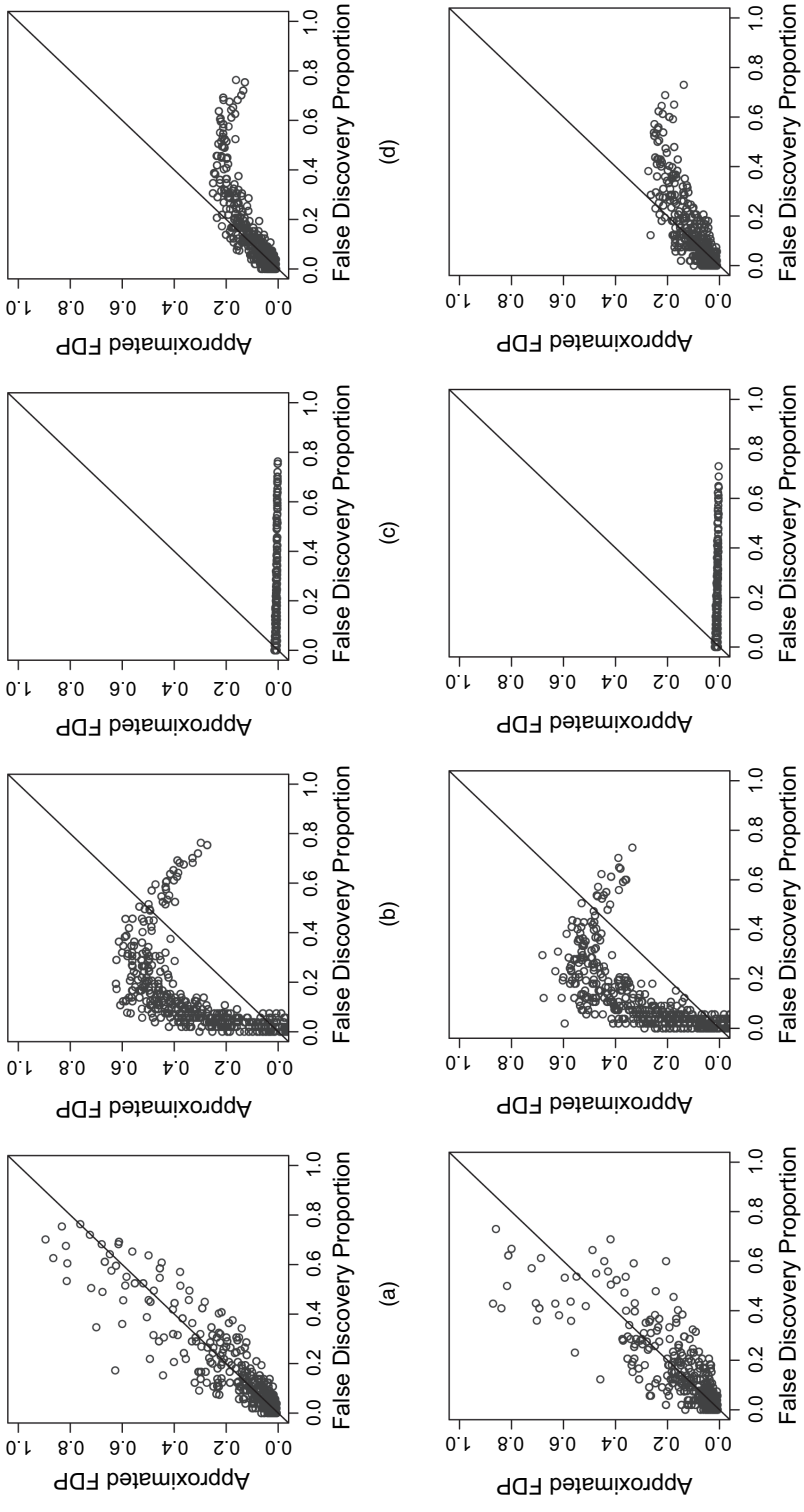

Fig. 1. Comparison of realized values of the FDP with $\widehat{\mathrm{FDP}}(t)$ ($n = 50$): (a) POET–PFA, model 1; (b) Efron, model 1; (c) FAMT, model 1; (d) FAMT–PFA, model 1; (e) POET–PFA, model 2; (f) Efron, model 2; (g) FAMT, model 2; (h) FAMT–PFA, model 2

Table 2. Comparison of the dependence-adjusted procedure with the fixed threshold procedure under the approximate factor model and strict factor model†

n	*Results for fixed threshold procedure*			*Results for dependence-adjusted procedure*		
	FDR (%)	*FNR (%)*	*Threshold*	*FDR (%)*	*FNR (%)*	*Threshold*
Model 1						
50	3.21	14.54	0.0026	3.24	1.96	0.001
100	2.48	9.53	0.0048	2.46	0.54	0.001
200	2.85	4.65	0.0074	2.89	0.08	0.001
Model 2						
50	2.64	15.03	0.0028	2.66	2.40	0.001
100	1.86	10.56	0.0034	1.85	0.70	0.001
200	1.86	5.65	0.0044	1.86	0.09	0.001

†The non-zero μ_i are simulated from $U(0.1, 0.5)$ and $p_1 = 200$.

value for the fixed threshold procedure such that the FDRs in the two testing procedures are approximately the same. To highlight the advantage of the dependence-adjusted procedure, we reset $\mathbf{\Sigma}_u$ as $0.1\mathbf{\Sigma}_3$. FNR for the dependence-adjusted procedure is smaller than that of the fixed threshold procedure, which suggests that the dependence-adjusted procedure is more powerful. Fan, Han and Gu (2012) showed numerically that, if the covariance is known, the advantage of the dependence-adjusted procedure is even more substantial. In Table 2, $p_1 = 200$ compared with $p = 1000$, implying that the better performance of the dependence-adjusted procedure is not limited to the sparse situation. This is expected since subtracting common factors out makes the problem have a higher signal-to-noise ratio.

Additional simulation results regarding comparisons with the known covariance matrix case can be found in the on-line supplementary materials. The basic findings are that under the apparent factor model structure the estimation errors of the covariance matrix have limited effect (see Figs 1 and 2 in the supplementary materials) and methods (the least absolute deviation (11), the least squares estimate (12) and the smoothly clipped absolute deviation (8)) for extracting unobservable realized latent factors are all effective.

5. Data analysis

In a well-known breast cancer study (Hedenfalk *et al.*, 2001; Efron, 2007), scientists compared gene expression levels in 15 patients. These observed gene expression levels have one of the two different genetic mutations BRCA1 and BRCA2 which are known to increase the lifetime risk of hereditary breast cancer. The study included seven women with the BRCA1 gene and eight women with the BRCA2 gene. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n, n = 7$, denote the microarray of expression levels on the $p = 3226$ genes for the first group, and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m, m = 8$, for that of the second group, so each $\mathbf{X}_i$ and $\mathbf{Y}_i$ are p-dimensional column vectors. Understanding the groups of genes that are expressed significantly differently in breast cancers can help scientists to identify cases of hereditary breast cancer on the basis of gene expression profiles.

Assume that the gene expressions of the two groups on each microarray are from two multivariate normal distributions with (potentially) different mean vector but the same covariance matrix, namely $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}^X, \mathbf{\Sigma})$ for $i = 1, \ldots, n$ and $\mathbf{Y}_i \sim N_p(\boldsymbol{\mu}^Y, \mathbf{\Sigma})$ for $i = 1, \ldots, m$. Then identifying differentially expressed genes is essentially a multiple-hypothesis test on $H_{0j} : \mu_j^X = \mu_j^Y$ *versus*

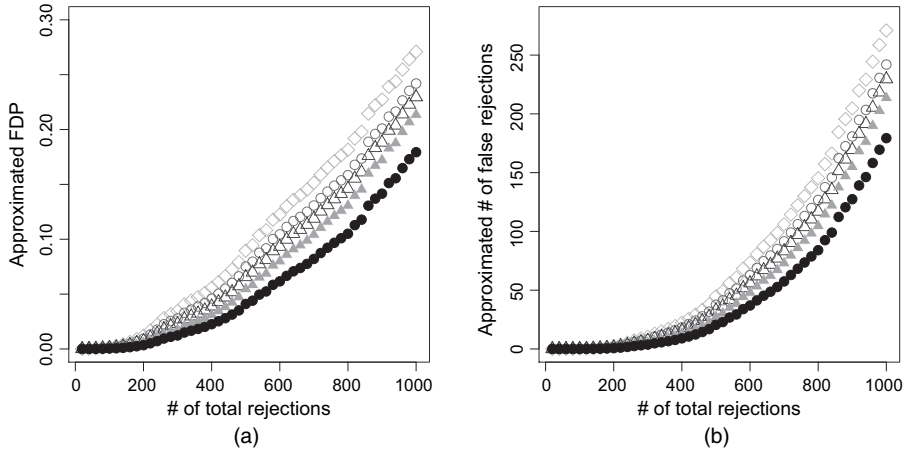

Fig. 2. (a) Approximated FDP and (b) approximated number of false discoveries as functions of the number of total discoveries for $p = 3226$ genes, where the estimated number of factors is 1 (◇) compared with other choices $k = 2$ (○), 3 (△), 4 (▲), 5 (●)

$H_{1j} : \mu_j^X \neq \mu_j^Y$, $j = 1, \ldots, p$. Consider the test statistics $\mathbf{Z} = \sqrt{\{nm/(n+m)\}}(\bar{\mathbf{X}} - \bar{\mathbf{Y}})$ where $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are the sample averages. Then we have $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \sqrt{\{nm/(n+m)\}}(\boldsymbol{\mu}^X - \boldsymbol{\mu}^Y)$, and the above two-sample comparison problem is equivalent to simultaneously testing $H_{0j} : \mu_j = 0$ *versus* $H_{1j} : \mu_j \neq 0$, $j = 1, \ldots, p$, based on $\mathbf{Z}$ and the unknown covariance matrix $\boldsymbol{\Sigma}$. It is also reasonable to assume that a large proportion of the genes are not differentially expressed, so that $\boldsymbol{\mu}$ is sparse.

Factor model structure has gained increasing popularity among biologists in the past decade, since it has been widely acknowledged that gene activities are usually driven by a small number of latent variables. See, for example, Friguet *et al.* (2009) and Desai and Storey (2012) for more details. We therefore apply the POET–PFA procedure (see Section 3.4) to the data set to obtain $\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t)$ for a given threshold value t. We apply the ER method as in Section 3.1 to estimate the unknown number of factors. The estimated k is 1 on the basis of the sample data. Because of the small sample size, this estimate could deviate from the true value. Therefore, we also report the results for $k = 2, 3, 4, 5$. The results of our analysis are depicted in Fig. 2. As can be seen, both $\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t)$ and $\hat{V}(t)$ increase with larger $R(t)$, and $\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t)$ is fairly close to 0 when $R(t)$ is below 200, suggesting that the rejected hypotheses in this range have high accuracy of being the true discoveries. Secondly, even when as many as 1000 hypotheses, corresponding to almost a third of the total number, have been rejected, the estimated FDPs are around 25%. Finally it is worth noting that, although our procedure seems robust under different choices of the number of factors, the estimated FDP tends to be relatively small with a larger number of factors. We also apply the dependence-adjusted procedure to the data. The relationship of $\widehat{\mathrm{FDP}}$ and the number of total rejections are summarized in Fig. 5 in the on-line supplementary materials. Compared with Fig. 2, $\widehat{\mathrm{FDP}}$ tends to be smaller with the same amount of total rejections. The same also happens to the estimated number of false rejections. This is consistent with the fact that the factor-adjusted test is more powerful. We conclude our analysis by presenting the list of the 40 most significantly differentially expressed genes in Table 4 and Table 5 of the supplementary materials with the POET–PFA method and the dependence-adjusted procedure respectively. Table 5 provides an alternative ranking of statistically significantly expressed genes for biologists, which have a lower FDP than the conventional method presented in Table 4.

Acknowledgements

This research was partly supported by National Institutes of Health grants R01-GM072611-11 and R01GM100474-04 and National Science Foundation grant DMS-1206464. We thank Dr Weijie Gu for early assistance on this project. We also thank the Joint Editor, a past Editor, the Associate Editors and the referees for many constructive comments which significantly improve the presentation of the paper.

Appendix A

A.1. Proof of theorem 1

First, note that, by equation (11), we have

$$\hat{\mathbf{B}}\hat{\mathbf{W}} = \hat{\mathbf{B}}(\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{Z} = \left(\sum_{i=1}^{k}\hat{\gamma}_i\hat{\gamma}_i^{\mathrm{T}}\right)\mathbf{Z}. \tag{17}$$

Similarly, let $\mathbf{B} = (\sqrt{\lambda_1}\gamma_1, \ldots, \sqrt{\lambda_k}\gamma_k)$ and $\tilde{\mathbf{W}} = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{Z}$. Then,

$$\mathbf{B}\tilde{\mathbf{W}} = \left(\sum_{i=1}^{k}\gamma_i\gamma_i^{\mathrm{T}}\right)\mathbf{Z}. \tag{18}$$

Denote by $\mathrm{FDP}_1(t)$ the estimator in equation (7) by using the infeasible estimator $\tilde{\mathbf{W}}$. Then,

$$\widehat{\mathrm{FDP}}_{\mathrm{U}}(t) - \mathrm{FDP}_{\mathrm{A}}(t) = \{\widehat{\mathrm{FDP}}_{\mathrm{U}}(t) - \mathrm{FDP}_1(t)\} + \{\mathrm{FDP}_1(t) - \mathrm{FDP}_{\mathrm{A}}(t)\}.$$

We shall bound these two terms separately.

Let us deal with the first term. Define

$$\Delta_1 = \sum_{i=1}^{p}[\Phi\{\hat{a_i}(z_{t/2} + \hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{W}})\} - \Phi\{a_i(z_{t/2} + \mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}})\}],$$

$$\Delta_2 = \sum_{i=1}^{p}[\Phi\{\hat{a_i}(z_{t/2} - \hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{W}})\} - \Phi\{a_i(z_{t/2} - \mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}})\}].$$

Then, we have

$$\widehat{\mathrm{FDP}}_{\mathrm{U}}(t) - \mathrm{FDP}_1(t) = (\Delta_1 + \Delta_2)/R(t). \tag{19}$$

We now deal with the term $\Delta_1 = \Sigma_{i=1}^{p}\Delta_{1i}$, in which

$$\begin{aligned}\Delta_{1i} =& \Phi\{\hat{a_i}(z_{t/2} + \hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{W}})\} - \Phi\{\hat{a_i}(z_{t/2} + \mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}})\} + \Phi\{\hat{a_i}(z_{t/2} + \mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}})\} - \Phi\{a_i(z_{t/2} + \mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}})\}\\ \equiv& \Delta_{11i} + \Delta_{12i}.\end{aligned}$$

Δ_2 can be dealt with analogously and hence has been omitted. For Δ_{12i}, by the mean value theorem, there exists $a_i^* \in (a_i, \hat{a_i})$ such that $\Delta_{12i} = \phi\{a_i^*(z_{t/2} + \mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}})\}(\hat{a_i} - a_i)(z_{t/2} + \mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}})$. Since $a_i > 1$ and $\hat{a_i} > 1$, we have $a_i^* > 1$ and hence $\phi\{a_i^*(z_{t/2} + \mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}})\}|z_{t/2} + \mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}}|$ is bounded. In other words, $|\Sigma_{i=1}^{p}\Delta_{12i}| \leqslant C\Sigma_{i=1}^{p}|\hat{a_i} - a_i|$, for a generic constant C. Using the definition of $\hat{a_i}$ and a_i, we have

$$|\hat{a_i} - a_i| = |(1 - |\hat{\mathbf{b}_i}\|^2)^{-1/2} - (1 - \|\mathbf{b}_i\|^2)^{-1/2}|.$$

Using the mean value theorem again, together with assumption 5, we have

$$|(1 - \|\hat{\mathbf{b}_i}\|^2)^{-1/2} - (1 - \|\mathbf{b}_i\|^2)^{-1/2}| \leqslant C(\|\hat{\mathbf{b}_i}\|^2 - \|\mathbf{b}_i\|^2).$$

Let $\gamma_h = (\gamma_{1h}, \ldots, \gamma_{ph})^{\mathrm{T}}$ and $\hat{\gamma}_h = (\hat{\gamma}_{1h}, \ldots, \hat{\gamma}_{ph})^{\mathrm{T}}$. Then

$$\begin{aligned}\sum_{i=1}^{p}|\|\hat{\mathbf{b}_i}\|^2 - \|\mathbf{b}_i\|^2| &= \sum_{i=1}^{p}\left|\sum_{h=1}^{k}(\hat{\lambda}_h - \lambda_h)\hat{\gamma}_{ih}^2 + \sum_{h=1}^{k}\lambda_h(\hat{\gamma}_{ih}^2 - \gamma_{ih}^2)\right|\\ &\leqslant \sum_{h=1}^{k}|\hat{\lambda}_h - \lambda_h| + \sum_{h=1}^{k}\lambda_h\sum_{i=1}^{p}|\hat{\gamma}_{ih}^2 - \gamma_{ih}^2|,\end{aligned}$$

where we used $\Sigma_{i=1}^{p}\hat{\gamma}_{ih}^2 = 1$. The second term of the last expression can be bounded as

$$\begin{aligned}\sum_{i=1}^{p}|\hat{\gamma}_{ih}^2-\gamma_{ih}^2| &\leqslant \left(\sum_{i=1}^{p}|\hat{\gamma}_{ih}-\gamma_{ih}|^2\sum_{i=1}^{p}|\hat{\gamma}_{ih}+\gamma_{ih}|^2\right)^{1/2}\\ &\leqslant \|\hat{\gamma}_h-\gamma_h\|\left\{2\sum_{i=1}^{p}(\hat{\gamma}_{ih}^2+\gamma_{ih}^2)\right\}^{1/2}\\ &= 2\|\hat{\gamma}_h-\gamma_h\|.\end{aligned}$$

Combining all the results that we have obtained, we conclude that

$$\left|\sum_{i=1}^{p}\Delta_{12i}\right| \leqslant C\left(\sum_{h=1}^{k}|\hat{\lambda}_h-\lambda_h|+\lambda_h\|\hat{\gamma}_h-\gamma_h\|\right). \tag{20}$$

Therefore, by using $\Sigma_{h=1}^{k}\lambda_h < p$ and assumptions 3 and 4, on the event $\mathcal{E}$, we conclude that $|\Sigma_{i=1}^{p}\Delta_{12i}| = O(p^{1-\min(\nu,\kappa)})$.

We now deal with the term Δ_{11i}. By the mean value theorem, there exists ξ_i between $\hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{W}}$ and $\mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}}$ such that $\Delta_{11i} = \phi\{\hat{a_i}(z_{t/2}+\xi_i)\}\hat{a_i}(\hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{W}}-\mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}})$. By condition 5 $\hat{a_i}$ is bounded and so is $\phi\{\hat{a_i}(z_{t/2}+\xi_i)\}\hat{a}_i$. Let $\mathbf{1}$ be a p-dimensional vector with each element 1. Then, by equations (17) and (18), we have

$$\sum_{i=1}^{p}|\hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{W}}-\mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}}| \leqslant \mathbf{1}^{\mathrm{T}}|\hat{\mathbf{B}}\hat{\mathbf{W}}-\mathbf{B}\tilde{\mathbf{W}}| = \mathbf{1}^{\mathrm{T}}\left|\sum_{h=1}^{k}(\hat{\gamma}_h\hat{\gamma}_h^{\mathrm{T}}-\gamma_h\gamma_h^{\mathrm{T}})\mathbf{Z}\right| \leqslant \sqrt{p}\left\|\sum_{h=1}^{k}(\hat{\gamma}_h\hat{\gamma}_h^{\mathrm{T}}-\gamma_h\gamma_h^{\mathrm{T}})\right\|\|\mathbf{Z}\| \tag{21}$$

where $|\mathbf{a}| = (|a_1|,\ldots,|a_p|)^{\mathrm{T}}$ for any vector $\mathbf{a}$ and the last inequality is obtained by the Cauchy–Schwartz inequality.

We now deal with the two factors in equation (21). The first factor is easily bounded by

$$\sum_{h=1}^{k}\|\hat{\gamma}_h(\hat{\gamma}_h-\gamma_h)^{\mathrm{T}}+(\hat{\gamma}_h-\gamma_h)\gamma_h^{\mathrm{T}}\| \leqslant 2\sum_{h=1}^{k}\|\hat{\gamma}_h-\gamma_h\|.$$

Let $\{\varepsilon_i\}_{i=1}^{p}$ be a sequence of independently and identically distributed $N(0,1)$ random variables. Then, stochastically, we have

$$E\|\mathbf{Z}\|^2 \leqslant 2\|\boldsymbol{\mu}^*\|^2+2\sum_{i=1}^{p}\lambda_i E[\varepsilon_i^2].$$

Therefore, $\|\mathbf{Z}\| = O_p(\|\mu^*\|+p^{1/2})$.

Substituting these two terms into expression (21), we have

$$\sum_{i=1}^{p}|\hat{\mathbf{b}}_i^{\mathrm{T}}\hat{\mathbf{W}}-\mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{W}}| = O_p\{kp^{1/2-\kappa}(\|\boldsymbol{\mu}^*\|+p^{1/2})\}.$$

Therefore, we can conclude that

$$\left|\sum_{i=1}^{p}\Delta_{11i}\right| = O_p\{kp^{1/2-\kappa}(\|\boldsymbol{\mu}^*\|+p^{1/2})\}. \tag{22}$$

Combination of the results in equations (20) and (22) leads to

$$\Delta_1 = O_p(p^{1-\min(\nu,\kappa)})+O_p(kp^{1-\kappa})+O_p(k\|\boldsymbol{\mu}^*\|p^{1/2-\kappa}).$$

In $\mathrm{FDP}_1(t)$, the least squares estimator is

$$\tilde{\mathbf{W}} = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\boldsymbol{\mu}^*+\mathbf{W}+(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{K} = \mathbf{W}+(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\boldsymbol{\mu}^* \tag{23}$$

in which we utilize the orthogonality between $\mathbf{B}$ and $\mathrm{var}(\mathbf{K})$. With a similar argument to that above, we can show that

$$|\mathrm{FDP}_1(t)-\mathrm{FDP}_{\mathrm{A}}(t)| = O\{|\mathbf{1}^{\mathrm{T}}\mathbf{B}(\tilde{\mathbf{W}}-\mathbf{W})|/R(t)\},$$

and we have

$$|(1,\ldots,1)\mathbf{B}(\tilde{\mathbf{W}}-\mathbf{W})| = \left|\mathbf{1}^{\mathrm{T}}\left(\sum_{h=1}^{k}\gamma_h\gamma_h^{\mathrm{T}}\right)\boldsymbol{\mu}^*\right| \leqslant p^{1/2}\|\boldsymbol{\mu}^*\|\left\|\sum_{h=1}^{k}\gamma_h\gamma_h^{\mathrm{T}}\right\| = p^{1/2}\|\boldsymbol{\mu}^*\|.$$

The proof is now complete.

A.2. Proof of theorem 2

For brevity, the proof of theorem 2 is relegated to the on-line supplementary material.

A.3. Proof of theorem 3

By the triangular inequality,

$$|\lambda_i - \hat{\lambda}_{i+1}| \geqslant ||\lambda_i - \lambda_{i+1}| - |\lambda_{i+1} - \hat{\lambda}_{i+1}||.$$

By Weyl's theorem in lemma 1, $|\lambda_{i+1} - \hat{\lambda}_{i+1}| \leqslant \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$. Therefore, on the event $\{\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| = O(d_p p^{-\tau})\}$

$$|\lambda_i - \hat{\lambda}_{i+1}| \geqslant d_p - \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \geqslant d_p/2$$

for sufficiently large p. Similarly, we have $|\hat{\lambda}_{i-1} - \lambda_i| \geqslant d_p/2$. By the $\sin(\theta)$ theorem in lemma 1, $\|\gamma_i - \hat{\gamma}_i\| = O(p^{-\tau})$. Hence, condition 3 holds with $\kappa = \tau$. Using Weyl's theorem again, we have

$$\sum_{i=1}^{k} |\lambda_{i+1} - \hat{\lambda}_{i+1}| \leqslant k\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| = O(k d_p p^{-\tau}).$$

Hence, condition 4 holds with $p^{-\delta} = k p^{-\tau} d_p / p$. The result now follows from theorem 1.

A.4. Proof of theorem 4

Let $\tilde{\mathbf{B}} = (\tilde{\lambda}_1^{1/2}\tilde{\gamma}_1, \ldots, \tilde{\lambda}_k^{1/2}\tilde{\gamma}_k)$. Note that

$$\|\hat{\boldsymbol{\Sigma}}_{\mathrm{POET}} - \boldsymbol{\Sigma}\| \leqslant \|\tilde{\mathbf{B}}\tilde{\mathbf{B}}^{\mathrm{T}} - \mathbf{B}\mathbf{B}^{\mathrm{T}}\| + \|\hat{\boldsymbol{\Sigma}}_u^{\mathcal{T}} - \boldsymbol{\Sigma}_u\|. \tag{24}$$

The bound for the second term is given by lemma 1 in the on-line supplementary materials. We now consider the first term in equation (24). By the triangular inequality, it follows that

$$\begin{aligned}\|\tilde{\mathbf{B}}\tilde{\mathbf{B}}^{\mathrm{T}} - \mathbf{B}\mathbf{B}^{\mathrm{T}}\| &\leqslant \|\mathbf{B}(\mathbf{H}^{\mathrm{T}}\mathbf{H} - \mathbf{I}_k)\mathbf{B}^{\mathrm{T}}\| + \|\mathbf{B}\mathbf{H}^{\mathrm{T}}(\tilde{\mathbf{B}} - \mathbf{B}\mathbf{H}^{\mathrm{T}})^{\mathrm{T}}\| + \|(\tilde{\mathbf{B}} - \mathbf{B}\mathbf{H}^{\mathrm{T}})\mathbf{H}\mathbf{B}^{\mathrm{T}}\| \\ &\quad + \|(\tilde{\mathbf{B}} - \mathbf{B}\mathbf{H}^{\mathrm{T}})(\tilde{\mathbf{B}} - \mathbf{B}\mathbf{H}^{\mathrm{T}})^{\mathrm{T}}\| \\ &\leqslant \|\mathbf{H}^{\mathrm{T}}\mathbf{H} - \mathbf{I}_k\|\|\mathbf{B}\|^2 + 2\|\mathbf{B}\|\|\mathbf{H}\|\|\tilde{\mathbf{B}} - \mathbf{B}\mathbf{H}^{\mathrm{T}}\| + \|(\tilde{\mathbf{B}} - \mathbf{B}\mathbf{H}^{\mathrm{T}})\|^2. \end{aligned} \tag{25}$$

Recall that $\{\tilde{\mathbf{b}}_j\}_{j=1}^k$ are columns of $\mathbf{B}$. Without loss of generality, assume that $\{\|\tilde{\mathbf{b}}_j\|\}$ are in non-increasing order. Since $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ is diagonal, $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ has non-vanishing eigenvalues $\{\|\tilde{\mathbf{b}}_j\|^2\}_{j=1}^K$ and $\|\mathbf{B}\| = \|\tilde{\mathbf{b}}_1\|$. Furthermore, by Weyl's theorem in lemma 1, $|\lambda_i - \|\tilde{\mathbf{b}}_i\|^2| \leqslant \|\boldsymbol{\Sigma} - \mathbf{B}\mathbf{B}^{\mathrm{T}}\| = \|\boldsymbol{\Sigma}_u\|$. Since the operator norm is bounded by the L_1-norm, we have

$$\|\boldsymbol{\Sigma}_u\| \leqslant \max_{i \leqslant p} \sum_{j=1}^{p} |\sigma_{u,ij}|^q |\sigma_{u,ii}\sigma_{u,jj}|^{(1-q)/2} \leqslant m_p. \tag{26}$$

Hence, $\|\tilde{\mathbf{b}}_i\|^2 \leqslant \lambda_1 + m_p = O(p)$.

We are now bounding each term in equation (25). Since the operator norm is bounded by the Frobenius norm, by lemma 2 in the supplementary materials, the first term in equation (25) is bounded by $O_p(p\omega_p)$, the second term in equation (25) is of order $O_p(\omega_p\sqrt{p})$ and the third term in equation (25) is $O_p(\omega_p^2)$. Combining these results leads to $\|\tilde{\mathbf{B}}\tilde{\mathbf{B}}^{\mathrm{T}} - \mathbf{B}\mathbf{B}\| = O_p(p\omega_p)$. Substituting this into inequality (24), we have

$$\|\hat{\boldsymbol{\Sigma}}_{\mathrm{POET}} - \boldsymbol{\Sigma}\| = O_p(p\omega_p + m_p\omega_p^{1-q}).$$

By Weyl's theorem in lemma 1, the conclusion for $|\hat{\lambda}_i - \lambda_i|$ follows.

Assumption 1 in the supplementary materials and Weyl's theorem imply that $\lambda_i = c_i p + o(p)$ for $i = 1, \ldots, k$ and that the c_is are distinct. By the triangular inequality, $|\lambda_i - \hat{\lambda}_{i+1}| \geqslant ||\lambda_i - \lambda_{i+1}| - |\lambda_{i+1} - \hat{\lambda}_{i+1}||$. By Weyl's theorem, $|\lambda_{i+1} - \hat{\lambda}_{i+1}| = o_p(p)$. Therefore, for sufficiently large n, $|\lambda_i - \hat{\lambda}_{i+1}| \geqslant \tilde{c}_i p$ for some constant $\tilde{c}_i > 0$ with probability tending to 1. By the $\sin(\theta)$ theorem, $\|\hat{\gamma}_i - \gamma_i\| = O_p(\omega_p + m_p\omega_p^{1-q}p^{-1})$. With direct application of theorems 1 and 3, we have

$$|\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = O_p[p^{\theta}\{k(\omega_p + m_p\omega_p^{1-q}p^{-1}) + \|\boldsymbol{\mu}^*\|p^{-1/2}\}].$$

The proof is now complete.

References

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

Antoniadis, A. and Fan, J. (2001) Regularized wavelet approximations (with discussion). *J. Am. Statist. Ass.*, **96**, 939–967.

Azriel, D. and Schwartzman, A. (2015) The empirical distribution of a large number of correlated normal variables. *J. Am. Statist. Ass.*, **110**, 1217–1228.

Bai, J. (2003) Inferential theory for factor models of large dimensions. *Econometrica*, **71**, 135–171.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

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

Bickel, P. and Levina, L. (2008a) Regularized estimation of large covariance matrices. *Ann. Statist.*, **36**, 199–227.

Bickel, P. and Levina, L. (2008b) Covariance regularization by thresholding. *Ann. Statist.*, **36**, 2577–2604.

Cai, T. and Liu, W. (2011) Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Ass.*, **106**, 672–684.

Cai, T., Liu, W. and Luo, X. (2011) A constrained l_1 minimization approach to sparse precision matrix estimation. *J. Am. Statist. Ass.*, **106**, 594–607.

Chamberlain, G. and Rothschild, M. (1983) Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica*, **51**, 1305–1324.

Clarke, S. and Hall, P. (2009) Robustness of multiple testing procedure against dependence. *Ann. Statist.*, **37**, 332–358.

Davis, C. and Kahan, W. (1970) The rotation of eigenvectors by a perturbation III. *SIAM J. Numer. Anal.*, **7**, 1–46.

Desai, K. H. and Storey, J. D. (2012) Cross-dimensional inference of dependent high-dimensional data. *J. Am. Statist. Ass.*, **107**, 135–151.

Donoho, D. L. and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 962–994.

Donoho, D. L. and Jin, J. (2006) Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist.*, **34**, 2980–3018.

Efron, B. (2007) Correlation and large-scale simultaneous significance testing. *J. Am. Statist. Ass.*, **102**, 93–103.

Efron, B. (2010) Correlated Z-values and the accuracy of large-scale statistical estimates (with discussion). *J. Am. Statist. Ass.*, **105**, 1042–1055.

El Karoui, N. (2008) Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.*, **36**, 2717–2756.

Engle, R. and Watson, M. (1981) A one-factor multivariate time series model of metropolitan wage rates. *J. Am. Statist. Ass.*, **76**, 774–781.

Fan, J., Han, X. and Gu, W. (2012) Estimating false discovery proportion under arbitrary covariance dependence (with discussion). *J. Am. Statist. Ass.*, **107**, 1019–1035.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, J., Liao, Y. and Mincheva, M. (2011) High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.*, **39**, 3320–3356.

Fan, J., Liao, Y. and Mincheva, M. (2013) Large covariance estimation by thresholding principal orthogonal complements (with discussion). *J. R. Statist. Soc.* B, **75**, 603–680.

Fan, J., Tang, R. and Shi, X. (2012) Partial consistency with sparse incidental parameters. *Preprint arXiv: 1210.6950*.

Friguet, C., Kloareg, M. and Causeur, D. (2009) A factor model approach to multiple testing under dependence. *J. Am. Statist. Ass.*, **104**, 1406–1415.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilford, B., Borg, A., Trent, J., Raffield, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrie, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H. and Sauter, G. (2001) Gene-expression profiles in hereditary breast cancer. *New Engl. J. Med.*, **344**, 539–548.

Higham, N. (1988) Computing a nearest symmetric positive semidefinite matrix. *Lin. Alg. Appl.*, **103**, 103–118.

Horn, R. and Johnson, C. (1990) *Matrix Analysis*. Cambridge: Cambridge Univesity Press.

Huang, N. and Fryzlewicz, P. (2015) NOVELIST estimator of large correlation and covariance matrices and their inverses. To be published.

Kotz, S. and Nadarajah, S. (2004) *Multivariate t Distributions and Their Applications*. Cambridge: Cambridge University Press.

Lam, C. and Yao, Q. (2012) Factor modeling for high-dimensional time series: inference for the number of factors. *Ann. Statist.*, **40**, 694–726.

Ledoit, O. and Wolf, M. (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Emp. Finan.*, **10**, 603–621.

Leek, J. and Storey, J. (2008) A general framework for multiple testing dependence. *Proc. Natn. Acad. Sci. USA*, **105**, 18718–18723.
Ma, Z. (2013) Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, **41**, 772–801.
Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008) Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, **2**, 494–515.
Sarkar, S. (2002) Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, **30**, 239–257.
Schafer, J. and Strimmer, K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Molec. Biol.*, **4**, article 32.
Schwartzman, A. and Lin, X. (2011) The effect of correlation in false discovery rate estimation. *Biometrika*, **98**, 199–214.
Siddiqui, M. (1967) A bivariate *t* distribution. *Ann. Math. Statist.*, **38**, 162–166.
Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.
Storey, J. D., Taylor, J. E. and Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc.* B, **66**, 187–205.
Sun, W. and Cai, T. T. (2009) Large-scale multiple testing under dependence. *J. R. Statist. Soc.* B, **71**, 393–424.

Supporting information

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

'Supplementary materials to "Estimation of false discovery proportion withs unknown dependence"'.