The Annals of Statistics
2017, Vol. 45, No. 2, 557–590
DOI: 10.1214/16-AOS1459
© Institute of Mathematical Statistics, 2017

A RATE OPTIMAL PROCEDURE FOR RECOVERING SPARSE DIFFERENCES BETWEEN HIGH-DIMENSIONAL MEANS UNDER DEPENDENCE

BY JUN LI AND PING-SHOU ZHONG

Kent State University and Michigan State University

The paper considers the problem of recovering the sparse different components between two high-dimensional means of column-wise dependent random vectors. We show that dependence can be utilized to lower the identification boundary for signal recovery. Moreover, an optimal convergence rate for the marginal false nondiscovery rate (mFNR) is established under dependence. The convergence rate is faster than the optimal rate without dependence. To recover the sparse signal bearing dimensions, we propose a Dependence-Assisted Thresholding and Excising (DATE) procedure, which is shown to be rate optimal for the mFNR with the marginal false discovery rate (mFDR) controlled at a pre-specified level. Extensions of the DATE to recover the differences in contrasts among multiple population means and differences between two covariance matrices are also provided. Simulation studies and case study are given to demonstrate the performance of the proposed signal identification procedure.

1. Introduction. In genetic studies, one important task is selecting the differentially expressed genes, which can be crucial in identifying novel biomarkers for cancers. Motivated by the problem of identifying differentially expressed genes, we consider the high-dimensional model

$$X_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma_i) \text{ for } i = 1, 2 \text{ and } 1 \le j \le n_i, \tag{1.1}$$

where μ_i is a p-dimensional population mean vector and Σ_i is a $p \times p$ covariance matrix. If we let $\delta = \mu_1 - \mu_2 = (\delta_1, \ldots, \delta_p)^T$, our interest is to determine which components of δ are nonzero.

Due to high dimensionality and relatively small sample sizes in modern statistical data such as microarray data, we consider $p \gg n_i$. Despite the large number of components, we assume that there are only a small number of signal bearing dimensions, which is thought to be reasonable in many applications. For instance, it is commonly believed that there are only a small number of genes that are significantly differentially expressed between two treatments in a study. Therefore, δ is

Received November 2015; revised February 2016.
MSC2010 subject classifications. Primary 62H15; secondary 62G20.
Key words and phrases. False discovery rate, high dimensional data, multiple testing, sparse signals, thresholding.

sparse in the sense that most of its components are zero but only a small portion of them are nonzero.

The magnitude of δ can be estimated by the statistic $J_n = \sqrt{n}(\bar{X}_1 - \bar{X}_2)$ with $n = (n_1 n_2)/(n_1 + n_2)$ based on sufficient statistics $\bar{X}_1 = {n_1}^{-1} \sum_{j=1}^{n_1} X_{1j}$ and $\bar{X}_2 = {n_2}^{-1} \sum_{j=1}^{n_2} X_{2j}$. From (1.1), it immediately follows that

$$J_n \sim \mathrm{N}\big(\sqrt{n}\delta, \Omega^{-1}\big), \qquad \text{where} \tag{1.2}$$

$$\begin{aligned} \Omega &= (\omega_{kl}) \\ &= \Sigma^{-1} = \big\{(1-\varsigma)\Sigma_1 + \varsigma\Sigma_2\big\}^{-1} \qquad \text{with } \varsigma = \lim_{n_1, n_2 \to \infty} \frac{n_1}{n_1 + n_2}. \end{aligned} \tag{1.3}$$

The model (1.2) is closely related to the Stein's normal means model, which has been carefully studied in Hall and Jin (2010) in the context of global testing. Specifically, the authors showed that with nonidentity covariance matrix Ω^{-1}, detecting whether δ is nonzero can be improved by incorporating data dependence. However, different from Hall and Jin (2010), the current work focuses on recovering sparse nonzero components of δ.

The Stein's normal means model has also been studied in the context of variable selection. To this end, Ji and Jin (2012) considered the following high-dimensional regression model:

$$Y = X\varrho + z, \tag{1.4}$$

where the rows of X are i.i.d. random vectors satisfying $\mathrm{N}(0, \tilde{\Omega}/n)$ for some sparse covariance $\tilde{\Omega}$ and $z = (z_1, \ldots, z_n)^T$ with z_i being i.i.d. $\mathrm{N}(0, 1)$. They showed that the model (1.4) can be reduced to the model (1.2) in the sense that $J_n^* = X^T Y \sim \mathrm{N}(\tilde{\Omega}\varrho, \tilde{\Omega})$. To recover the nonzero coefficients of ϱ in (1.4), they proposed a Univariate Penalization Screening (UPS) procedure for variable selection, which was shown to achieve the optimal rate of convergence with the risk measured by the Hamming distance. Different from the Hamming distance, a risk defined as a weighted sum of false negatives and false positives is more relevant for the model (1.2) in the context of multiple testing. As argued by Sun and Cai (2007) and Sun and Cai (2009), a procedure minimizing the weighted sum is also an optimal multiple testing procedure minimizing the false nondiscovery rate (FNR) with the false discovery rate (FDR) controlled at a pre-selected level. The connection between the optimal variable selection and optimal multiple testing was further elaborated in Jin (2012). Motivated by Jin (2012), Ji and Zhao (2014) recently extended the UPS procedure for variable selection to a Univariate Penalization Testing (UPT) procedure that was shown to be rate optimal in recovering the coefficients ϱ from (1.4).

Despite the connection between (1.4) and (1.1) through the Stein's normal means model (1.2), the parameter of interest δ considered in current work is different from ϱ in Ji and Jin (2012) and Ji and Zhao (2014). Furthermore, in Ji and

Jin (2012) and Ji and Zhao (2014), $\tilde{\Omega}$ is approximated by the known matrix $X^T X$ from (1.4). However, Ω defined by (1.3) is the inverse of a linear combination of two unknown covariance matrices. Therefore, the effect of estimating Ω on the signal recovering needs to be addressed in current work. Most importantly, similar to Hall and Jin (2010) who demonstrated the advantageous effect of data dependence on signal detection, a major contribution of the current work is to unveil a similar advantageous effect of dependence on the recovery of nonzero components of δ in (1.1).

A commonly used approach to recover nonzero components of δ is the multiple testing procedure. Each dimension $k \in \{1, \ldots, p\}$ is tested by a t-statistic which is expected to have significant value if $\delta_k \neq 0$ and, conversely, to be insignificant if $\delta_k = 0$. After all the p-values associated with the t-statistics are ranked, the dimensions with p-values smaller than a critical p-value threshold are selected and treated as signal bearing dimensions. In the multiple testing procedure, the threshold is chosen to control the FDR, which is defined as the fraction of false positives among all the rejected hypotheses. For this purpose, Benjamini and Hochberg (1995) introduced a novel procedure (BH procedure) which is shown to be more desirable than other procedures controlling the familywise error rate (FWER) such as the Bonferroni correction, since the former is less conservative than the latter. However, the BH procedure relies on the assumption that the test statistics corresponding to the true null hypotheses ($\delta_k = 0$) are independent. It has been shown that the presence of dependence among test statistics can substantially affect the number of reported nonnull hypotheses, since the empirical null distribution of dependent p-values can be significantly different from the theoretical null distribution under the assumption of independence [Efron (2007)]. As a result, the outcome of genetic studies by simply ignoring the intergene correlation is implausible, and a clear strategy to control the false positives in the multiple testing for dependent data is needed [Qiu, Klebanov and Yakovlev (2005)].

Some efforts have been made to address the effect of dependence on the multiple testing by assuming some special dependence structures. For example, Benjamini and Yekutieli (2001) showed that when the test statistics corresponding to the true null hypotheses ($\delta_k = 0$) have the positive regression dependence, the BH procedure is still able to be modified to control the FDR. Based on a hidden Markov model for the dependence structure, Sun and Cai (2009) proposed an oracle and an asymptotically optimal data-driven procedures which were shown to be able to minimize the FNR while controlling the FDR at a pre-specified level. Xie, Cai and Li (2011) established a Bayes oracle rule along with the corresponding data adaptive rule based on independent data, which were shown to be optimal in that it minimizes the sum of false negatives and false positives. They also argued that the proposed methods are still valid and remain optimal under short-range dependence.

The advantageous effect of dependence on signal detection boundary has been well established by Hall and Jin (2010), who showed that the detection boundary

can be lowered by incorporating the data dependence. Different from the signal detection boundary that separates the plane of signal sparsity and signal strength into the detectable region and the undetectable region, the identification boundary separates the same plane into the other two different regions. In the region above the boundary, signals can be recovered individually. But below the boundary, a successful identification is impossible [Donoho and Jin (2004); Hall and Jin (2010); Ji and Jin (2012)]. In this paper, we investigate the effect of dependence on the signal identification boundary for the model (1.1). Note that the benefit of dependence on signal identification has been addressed for the sparse regression model (1.4) by Genovese et al. (2012), Jin, Zhang and Zhang (2014) and Ke, Jin and Fan (2014). However, the setting addressed here is different from those because the parameter δ for the model (1.1) is different from ϱ for the model (1.4). Moreover, instead of considering variable selection, we focus on the multiple testing for the nonzero components of δ. Specifically, we show that the signal identification boundary for dependent data is lower than that for independent data. An explicit expression for the identification boundary is also established when dependence is present.

To recover the sparse nonzero components of δ, we are interested in the optimal procedure that minimizes the FNR while the FDR is controlled at a pre-specified level. To this purpose, we propose a dependence-assisted thresholding and excising (DATE) procedure. The proposed procedure is implemented by first transforming the original X_{ij} through Ω in (1.3) into $Z_{ij} = \Omega X_{ij}$. It will be shown in Section 3 that under certain sparse settings of signals and Ω, the standardized magnitude of the transformed signal is greater than that of the original data or the de-correlated data obtained by transforming the original data via $\Omega^{1/2}$, which potentially increases the probability of identifying signals. After the transformation, the null components of the transformed data are removed by conducting a marginal thresholding, which is followed by another step to excise the fake signals induced by the transformation. As we will show in Section 4, the proposed procedure attains not only the signal identification boundary under dependence but also the optimal convergence rate for the marginal false nondiscovery rate (mFNR) with the marginal false discovery rate (mFDR) controlled at a pre-selected level, and thus is superior compared with other methods without taking data dependence into account.

The rest of the paper is organized as follows. In Section 2, we establish two lower bounds: one for the risk function (2.2) and another for the convergence rate of the mFNR. To show the optimality of these two bounds, we first demonstrate the benefit of transforming data by the matrix Ω in (1.3) in Section 3. Then a thresholding and excising procedure based on the transformed data is introduced in Section 4. The proposed procedure is shown to be able to achieve two lower bounds established in Section 2, and thus is rate optimal. Extensions of the proposed procedure to recover differences in contrasts among multiple population means and differences between two covariance matrices are provided in Section 5. Section 6 illustrates some numerical studies and Section 7 reports an empirical study to select differentially expressed genes for a human breast cancer data set.

Discussion is given in Section 8. Due to limited space, all the proofs are relegated to the Supplementary Material [Li and Zhong (2016)].

2. Lower bounds for signal identification under dependence. We start with some notation and definitions. Denote $S_\beta = \{k : \delta_k \neq 0\}$ to be a set including the locations of the nonzero δ_k. The number of non-zero elements in S_β is $p^{1-\beta}$ for $\beta \in (0, 1)$. Define L_p to be a slowly varying logarithmic function in the form of $(a \log p)^b$ for some constants a and b. Without loss of generality, we assume both Σ_1 and Σ_2 are standardized to have unit diagonal elements. With matrix $\Omega = (\omega_{ij})$ defined in (1.3), let

$$\underline{\omega} = \varliminf_{p\to\infty} \min_{1\le k\le p} \omega_{kk} \quad \text{and} \quad \bar{\omega} = \varlimsup_{p\to\infty} \max_{1\le k\le p} \omega_{kk}. \tag{2.1}$$

We model δ to satisfy the following condition [see Ji and Jin (2012)]:

(C1) The components of δ follow a mixture distribution

$$\delta_k \overset{\text{i.i.d.}}{\sim} (1 - p^{-\beta})h_0 + p^{-\beta}\kappa_p, \qquad k = 1, \ldots, p,$$

where h_0 is a point mass at 0 and κ_p is a distribution with the support $[-\sqrt{2r \log p/n}, 0) \cup (0, \sqrt{2r \log p/n}]$ for $r > 0$ and $n = (n_1 n_2)/(n_1 + n_2)$.

Given δ_k for $1 \le k \le p$, $\hat{\delta}_k$ is denoted as an estimate of δ_k. For any signal identification procedure, there are generally two types of error related with the signal estimate $\hat{\delta}_k$: the false negative meaning that $\delta_k \neq 0$ but $\hat{\delta}_k = 0$, and the false positive representing that $\delta_k = 0$ but $\hat{\delta}_k \neq 0$. When identifying the nonzero components of δ, people are often interested in the optimal procedure that minimizes the FNR while the FDR is controlled at a certain level. For this purpose, Sun and Cai (2007) and Sun and Cai (2009) introduced an expected weighted sum of false negatives and false positives:

$$H(\Lambda) = \mathrm{E}\Big\{ \sum_{k\in S_\beta} \mathrm{I}(\hat{\delta}_k = 0) + p^{-\Lambda} \sum_{l\in S_\beta^c} \mathrm{I}(\hat{\delta}_l \neq 0) \Big\}, \tag{2.2}$$

where the weight $p^{-\Lambda}$ with $\Lambda \in [0, \infty)$ is chosen to adjust the level of false positives. The effect of Λ on false positives can be demonstrated by Figure 1. Assume that the minimization of $H(0)$ is achieved at the intersection point *diamond* of the false positives line (FP) and the false negatives line (FN). By multiplying FP with $p^{-\Lambda}$ (dash line), the FP becomes less important in $H(\Lambda)$ and $H(\Lambda)$ is minimized at the intersection point *star* which is on the right-hand side of the intersection point *diamond*. As a result, the expected false positives corresponding to the minimized $H(\Lambda)$ is larger than that corresponding to the minimized $H(0)$. With a specific choice of Λ, Sun and Cai (2007) and Sun and Cai (2009) showed that minimizing the above risk function leads to the optimal multiple testing procedure for the normal means model (1.2) with the FDR controlled at a specific level.

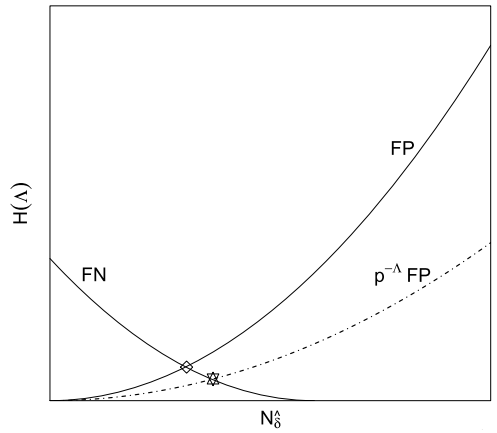

FIG. 1. *The horizontal axis $N_{\hat{\delta}}$ represents the number of $\hat{\delta}_k \neq 0$. The diamond is the intersection point of the false positives line (FP) and the false negatives line (FN) where $H(0)$ is minimized and the star is the intersection point where $H(\Lambda)$ is minimized.*

By choosing the weight $\Lambda = 0$ such that $H(0)$ becomes the classification error, Ji and Jin (2012) established the optimal convergence rate for the variable selection in the high-dimensional regression model (1.4). Moreover, Jin (2012) elaborated a connection between the optimal variable selection and the optimal multiple testing by showing that with a properly chosen Λ, an optimal variable selection procedure that minimizes the weighted risk function $H(\Lambda)$ is also an optimal procedure in the multiple testing. Recently, Ji and Zhao (2014) developed the idea of Jin (2012) and proposed the UPT procedure that was shown to attain the optimal rate of convergence in mFNR with the mFDR controlled at a pre-selected level. Interested readers may refer to Ji and Zhao (2014) for a comprehensive review of literature evolution about the connection between optimal procedure for variable selection and optimal procedure for multiple testing.

Inspired by Sun and Cai (2007, 2009), Ji and Jin (2012), Jin (2012) and Ji and Zhao (2014), the current article is to seek a rate optimal procedure for recovering sparse nonzero components of δ based on the connection between (1.1) and the normal means model (1.2). To begin with, we first establish a universal lower bound of the risk function $H(\Lambda)$. Let $\hat{\theta}_j = \mathrm{I}(\hat{\delta}_j \neq 0)$ for $j = 1, \ldots, p$. The Bayesian decision rule minimizing the risk function (2.2) is

$$\hat{\theta}_j = \mathrm{I}\left\{ \frac{(1-p^{-\beta}) f_{j,0}(X_{11}, \ldots, X_{1n_1}; X_{21}, \ldots, X_{2n_2})}{p^{-\beta} f_{j,1}(X_{11}, \ldots, X_{1n_1}; X_{21}, \ldots, X_{2n_2})} \leq p^{\Lambda} \right\}, \tag{2.3}$$

where $f_{j,0}$ and $f_{j,1}$, defined by (A.7) in the Supplementary Material [Li and Zhong (2016)], are the distributions of $(X_{11}, \ldots, X_{1n_1}; X_{21}, \ldots, X_{2n_2})$ conditional on $\delta_j = 0$ and $\delta_j \neq 0$, respectively. Based on the Bayesian rule, the universal lower bound of the risk function $H(\Lambda)$ at a fixed value Λ is established by the following theorem.

THEOREM 1. *Assume condition* (C1) *and the model* (1.1) *for* X_{ij}. *As* $p \to \infty$,

$$H(\Lambda) \geq \begin{cases} L_p p^{1-\beta-(\bar{\omega}r-\beta+\Lambda)^2/(4\bar{\omega}r)}, & -r < (\Lambda-\beta)/\underline{\omega} < r, \\ p^{1-\beta}, & r < (\beta-\Lambda)/\bar{\omega}, \\ p^{1-\Lambda}, & r < (\Lambda-\beta)/\bar{\omega}, \end{cases}$$

where $\underline{\omega}$ *and* $\bar{\omega}$ *are defined in* (2.1), *and* L_p *is a slowly varying logarithmic function*.

It is worth mentioning that the lower bounds do not depend on n since the signal strength has been normalized by $\sqrt{n}$ as shown in (C1). The universal lower bound varies with different values of r, β for each fixed value of Λ. If we choose $\Lambda = 0$, the classification error has the lower bound

$$H(0) \geq \begin{cases} L_p p^{1-\beta-(\bar{\omega}r-\beta)^2/(4\bar{\omega}r)}, & r > \beta/\underline{\omega}; \\ p^{1-\beta}, & r < \beta/\bar{\omega}. \end{cases}$$

Some key observations are as follows. First, if the signal strength $r < \beta/\bar{\omega}$, the classification error is no less than $p^{1-\beta}$, the number of nonzero δ_k, which implies that there exists no successful signal identification procedure. The area $r < \beta/\bar{\omega}$ in $r - \beta$ plane is thereafter called the region of no recovery. On the other hand, if the signal strength attains $r \geq (1+\sqrt{1-\beta})^2/\underline{\omega}$, the classification error asymptotically converges to zero and all the signals can be successfully recovered. The corresponding region is called the region of full recovery. The area sandwiched between the no recovery region and the full recovery region satisfies $\beta/\bar{\omega} < r < (1+\sqrt{1-\beta})^2/\underline{\omega}$, having the classification error less than the number of signals and greater than zero. This region is called region of partial recovery. Most importantly, since $\bar{\omega} \geq \underline{\omega} > 1$ under data dependency shown by Lemma 1 in the Supplementary Material [Li and Zhong (2016)], the partial recovery boundary $r = \beta/\bar{\omega}$ and full recovery boundary $r = (1+\sqrt{1-\beta})^2/\underline{\omega}$ used to separate three regions are lower than those without existence of data dependence.

To demonstrate the observations above, we consider $\Sigma_1 = \Sigma_2 = (\rho^{|i-j|})$ for $1 \leq i, j \leq p$ in model (1.1) such that the data dependence is exhibited by the value of ρ. If $\rho = 0$, $\bar{\omega} = \underline{\omega} = 1$ since there is no data dependence. On the other hand, if $\rho = 0.6$, we obtain $\underline{\omega} = 1.5625$ and $\bar{\omega} = 2.125$. The corresponding phase diagrams with and without data dependence are displayed in Figure 2 in which the partial signal identification boundary and the full recovery boundary with $\rho = 0.6$ are lower than those with $\rho = 0$ due to the fact that $\underline{\omega} > 1$ and $\bar{\omega} > 1$. As a result, even though the signals with $r < \beta$ are unable to be identified by any procedure if there exists no data dependence, some of them can be recovered as long as the signal strength $r > \beta/2.125$ with the existence of data dependence. The benefit to the full signal identification with the existence of dependence can be seen based on the similar derivation.

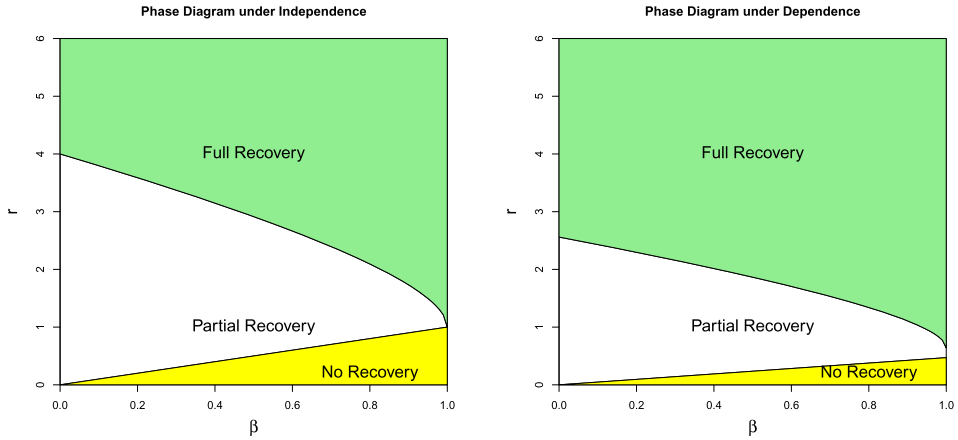

FIG. 2. *Left*: *phase diagram for signal recovery without data dependence. Right*: *phase diagram for signal recovery with* $\Sigma_1 = \Sigma_2 = (0.6^{|i-j|})$ *for* $1 \le i, j \le p$.

As pointed out by Sun and Cai (2007, 2009), Jin (2012) and Ji and Zhao (2014), with a properly chosen Λ, the decision rule that minimizes the weighted risk function $H(\Lambda)$ is also the optimal procedure that controls the mFDR at level α and minimizes the mFNR in the multiple testing. Let FP = false positives, TP = true positives, FN = false negatives and TN = true negatives. The mFDR and mFNR are defined as

$$\text{mFDR} = \left\{ \frac{\text{E(FP)}}{\text{E(FP)} + \text{E(TP)}} \right\} \quad \text{and} \quad \text{mFNR} = \left\{ \frac{\text{E(FN)}}{\text{E(FN)} + \text{E(TN)}} \right\}.$$

Genovese and Wasserman (2002) showed that mFDR and mFNR are asymptotically equivalent to FDR and FNR under weak conditions. In general, the connection between Λ and α is complicated. The following theorem provides a solution for choosing a proper $\Lambda(\alpha)$ such that the mFDR is controlled at the level of $\alpha < 1$. Moreover, it establishes a lower bound for the mFNR subject to the constraint that $\text{mFDR} \le \alpha$.

THEOREM 2. *Assume condition* (C1) *and* (1.1) *for* X_{ij}. *If we choose*

$$\Lambda(\alpha) = \underline{\omega} r + \beta - 2\sqrt{\underline{\omega} r \beta \{1 - g(\alpha, p)/\beta\}},$$

where $g(\alpha, p) = \log\{\frac{\alpha}{(1-\alpha)}\sqrt{4\pi\beta \log p}\}/\log p$, *then as* $p \to \infty$,

$$\text{mFNR} \ge L_p p^{-\beta - \{\sqrt{\bar{\omega} r} - \sqrt{\beta - g(\alpha,p)}\}^2} \quad \textit{and} \quad \text{mFDR} \le \alpha < 1.$$

Note that with $\bar{\omega} = \underline{\omega} = 1$, the lower bound of the mFNR above was also established in Ji and Zhao (2014) for the high-dimensional regression model (1.4). Since the model (1.1) considered in current work and the model (1.4) both are

related with the Stein's normal means model (1.2), it is not surprising to see the similar results established for two models. However, as discussed in Section 1, the current work is to investigate the advantageous effect of data dependence on signal recovering. More specifically, our Theorem 2 demonstrates that similar to the weighted risk function, the lower bound for the mFNR is accelerated by existence of dependence since $\bar{\omega} > 1$. To show that the lower bounds in Theorems 1 and 2 are tight, we need to search for a signal identification procedure that is able to attain the universal lower bounds. As we will see in next section, the key for this procedure is to take the data dependence into account, which can be done by transforming data via the matrix Ω defined in (1.3).

3. Data transformation. Data transformation by Ω to enhance signal strength was originally discovered by Hall and Jin (2010) for global testing of ϱ in the Stein's normal means model (1.2). To extend the same result to identification of the nonzero components of δ from the model (1.1), we need some additional assumptions.

(C2) Eigenvalues of Σ_i for $i = 1, 2$ satisfy $C_0^{-1} \le \lambda_{\min}(\Sigma_i) \le \lambda_{\max}(\Sigma_i) \le C_0$ for some constant $C_0 > 0$.

(C3) The matrix Ω in (1.3) is presumably sparse and belongs to the class

$$\mathcal{V}(c_p, M_p) = \left\{ \Omega : \|\Omega\|_{L_1} \le M_p, \max_{1 \le j \le p} \sum_{i=1}^{p} |\omega_{ij}|^q \le c_p \text{ for } 0 < q < 1 \right\},$$

where $M_p = O(\log^{b_1} p)$ and $c_p = O(\log^{b_2} p)$ for some constants $b_1 \ge 0$ and $b_2 \ge 0$.

(C4) As $n \to \infty$, $p \to \infty$ and $\log p = o(n^\theta)$ where $\theta = (1-q)/\{(2b_1+1) \times (1-q) + 2b_2\}$, and q, b_1 and b_2 are defined in (C3).

Conditions (C2) and (C3) define a class of matrices with the sparse structure similar to Cai, Liu and Luo (2011), where both M_p and c_p are allowed to grow with p logarithmically. Condition (C4) specifies the exponential growth of dimension p with n.

For signal identification, we need to construct a statistic to estimate the magnitude of the signal. Generally, if we let Q be a $p \times p$ invertible matrix and $\bar{Z}_{Q,i}^{(k)} = (Q\bar{X}_i)^{(k)}$ for $i = 1, 2$ where $\bar{Z}_{Q,i}^{(k)}$ is the kth component of $\bar{Z}_{Q,i}$, then a measure of the signal at kth dimension is

$$T_Q^k = \frac{n\{\bar{Z}_{Q,1}^{(k)} - \bar{Z}_{Q,2}^{(k)}\}^2}{a_{kk}}, \qquad k = 1, \ldots, p,$$

where $A = (a_{ij})$ is the covariance matrix of $\sqrt{n}(\bar{Z}_{Q,1} - \bar{Z}_{Q,2})$. In the above statistics, Q needs to be specified. The most common choice of Q is the identity matrix I and the corresponding statistic T_I^k depends on the standardized signal strength

$\sqrt{n}\delta_k/\sqrt{\sigma_{kk}}$ where σ_{kk}, the kth diagonal element of Σ defined in (1.3), becomes 1 if both Σ_1 and Σ_2 are standardized to have unit diagonal elements. Another choice of Q is $\Omega = (\omega_{kl})$ for $1 \le k, l \le p$ defined in (1.3), which has been considered in Hall and Jin (2010) for their innovated higher criticism test, and Cai, Liu and Xia (2014) for testing the equality of two sample mean vectors. The corresponding statistic T_Ω^k depends on the standardized signal strength $\sqrt{n}\delta_{\Omega,k}/\sqrt{\omega_{kk}}$, where δ_Ω is defined as the difference in two population mean vectors after the transformation.

Similar to a very important result established in Hall and Jin (2010), Lemmas 1 and 2 in the Supplementary Material [Li and Zhong (2016)] show that for sparse signals $\beta \in (1/2, 1)$, sparse Ω assumed in (C3) and $k \in S_\beta$,

$$\frac{\sqrt{n}\delta_{\Omega,k}}{\sqrt{\omega_{kk}}} \ge \frac{\sqrt{n}\delta_k}{\sqrt{\sigma_{kk}}}. \tag{3.1}$$

This implies that the standardized signal strength can be boosted by the transformation of Ω. The signal gain in practice is also explored by the simulation studies in Section 6 to confirm the above theoretical finding.

In addition to the transformation induced by I and Ω, another natural choice of Q is $\Omega^{1/2}$ that was considered by Allen and Tibshirani (2012) to de-correlate the original data so that they are independent. The corresponding statistic $T_{\Omega^{1/2}}^k$ depends on the standardized signal strength $\sqrt{n}\delta_{\Omega^{1/2},k}$ which can be approximated by $\sqrt{n}\varpi_{kk}\delta_k$, where $\delta_{\Omega^{1/2}}$ is the difference in two population mean vectors after the transformation induced by $\Omega^{1/2}$, and ϖ_{kk} is the kth diagonal element of $\Omega^{1/2}$. Due to the fact that $\omega_{kk} = \sum_l \varpi_{kl}^2$, $\sqrt{n}\sqrt{\omega_{kk}}\delta_k \ge \sqrt{n}\varpi_{kk}\delta_k$, or equivalently,

$$\frac{\sqrt{n}\delta_{\Omega,k}}{\sqrt{\omega_{kk}}} \ge \sqrt{n}\delta_{\Omega^{1/2},k}. \tag{3.2}$$

Both (3.1) and (3.2) suggest that the statistics T_Ω^k based on the data transformed by Ω obtain more gain in standardized signal strength, and thus are selected for the signal identification. For notation simplicity, we suppress the subscript Ω in the transformed data $Z_{\Omega,ij} = \Omega X_{ij}$ and corresponding statistics T_Ω^k.

In real applications, Ω is unknown and needs to be estimated by $\hat{\Omega}$. Observing that $\Omega = (1-\varsigma)^{-1}\Sigma_w^{-1}$ with $\Sigma_w \equiv \Sigma_1 + \{\varsigma/(1-\varsigma)\}\Sigma_2$, we only need to estimate Σ_w^{-1}. There are many methods available in the literature for estimating the precision matrix. When the precision matrix is bandable, it can be estimated through the Cholesky decomposition proposed by Bickel and Levina (2008). When the precision matrix is sparse, Cai, Liu and Luo (2011) introduced the CLIME estimator based on the constrained L_1 minimization approach for precision matrix estimate. More can be found in Friedman, Hastie and Tibshirani (2008). Although all the above methods are designed for estimating the precision matrix for one sample case, the estimator $\hat{\Sigma}_w^{-1}$ can be obtained from those methods [Friedman, Hastie and Tibshirani (2008); Cai, Liu and Luo (2011)] by replacing

the regular sample covariance with the following estimator based on two-sample U-statistics:

$$S_n^* = \frac{1}{n_1 n_2} \sum_{k=1}^{n_1} \sum_{l=1}^{n_2} Y_{n,kl} Y_{n,kl}^T, \tag{3.3}$$

where $Y_{n,kl} = X_{1k} - \bar{X}_1 - \sqrt{n_1/n_2}(X_{2l} - \bar{X}_2)$ for $k = 1, \ldots, n_1$ and $l = 1, \ldots, n_2$. Then $\hat{\Omega}$ can be obtained by $\hat{\Omega} = (1 + n_1/n_2)\hat{\Sigma}_w^{-1}$. The consistency of the estimator $\hat{\Omega}$ can be established under conditions (C2)–(C4) by changing the exponential inequality for the one-sample covariance to the exponential inequality for the above two-sample U-statistics [see Cai, Liu and Luo (2011)].

With estimated $\hat{\Omega}$ obtained from one of the above methods, the transformed signal for $k \in S_\beta$ is $\hat{\delta}_{\Omega,k} = \sum_{l \in S_\beta} \hat{\omega}_{kl}\delta_l$. Similar to (3.1) by assuming that both Σ_1 and Σ_2 have diagonal elements equal 1, Lemmas 1 and 2 show that under some mild conditions, with probability approaching 1,

$$\frac{\sqrt{n}\hat{\delta}_{\Omega,k}}{\sqrt{\hat{\omega}_{kk}}} \geq \sqrt{n}\delta_k. \tag{3.4}$$

Therefore, we consider the following test statistics based on the transformed data $\hat{Z}_{ij} = \hat{\Omega}X_{ij}$ as the starting point of the proposed signal identification procedure:

$$\hat{T}_k = \frac{n\{\bar{\hat{Z}}_1^{(k)} - \bar{\hat{Z}}_2^{(k)}\}^2}{\hat{\omega}_{kk}}, \qquad k = 1, \ldots, p. \tag{3.5}$$

The advantage of the statistics in (3.5) is that the standardized signal strength has been enhanced by incorporating dependence, which potentially increases the probability of weak signals being identified by the signal recovery procedure. However, since $\delta_{\Omega,k} = \sum_{l \in S_\beta} \omega_{kl}\delta_l$, a side effect of the transformation is that it generates some fake signals, that is, $\delta_k = 0$ but $\delta_{\Omega,k} \neq 0$ if $\omega_{kl} \neq 0$ for some $l \in S_\beta$. Therefore, a successful signal recovery procedure benefited by data transformation requires to remove these fake signals. As we will discuss in next section, fake signals can be successfully excised by a penalized method with L_0 penalty. As revealed by Ji and Jin (2012), this approach is very effective in cleaning fake signals but suffers the computational intensity if dimension p is large. To reduce the complexity of the original signal selection problem, we first need a dimension reduction procedure, which is fulfilled by a thresholding step as we will discuss in the next section.

4. DATE procedure to recover signals. To introduce our signal identification procedure, we first focus on the most interesting case where $\underline{\omega}r < (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$. According to Theorem 1, this case indicates that the weighted risk $H(\Lambda)$ does not converge to zero but is less than $p^{1-\beta}$. The corresponding region on $r - \beta$ plane is the partial recovery under a fixed value Λ. The case

$\underline{\omega} r \geq (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$ corresponding to the full recovery region is an easier problem due to the relatively larger signal strength. We will discuss it at the end of this section.

As we have discussed in the previous section, after data transformation, p coordinates consist of the signals, fake signals and noise. As the first step of the proposed method for signal recovery, a thresholding is conducted to remove the noise. After all the p dimensions are checked by a threshold function $2s \log p$, we set $\hat{\delta}_k = 0$ for $k \in \{1, \ldots, p\}$ if and only if

$$\hat{T}_k < 2s \log p, \tag{4.1}$$

where $s > 0$ is chosen to control the level of the threshold, and the decision on other coordinates with $\hat{T}_k \geq 2s \log p$ will be made in another step following the thresholding step. Although imposing the threshold is to prevent noise, it can potentially screen out signals, and thus produce the false negatives. Similar to Ji and Jin (2012) and Ji and Zhao (2014), the following lemma establishes the upper bound of the expected false negatives generated in the thresholding step (4.1).

LEMMA 3. *Assume* (C1), (C3) *and* (C4). *Let* $s \in (0, (\underline{\omega} r + \beta - \Lambda)^2/(4\underline{\omega} r))$, $\beta \in (1/2, 1)$ *and* $\beta - \Lambda < \underline{\omega} r < (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$. *As* $p \to \infty$,

$$\mathrm{E}\left\{\sum_{k=1}^{p} \mathrm{I}(\hat{\delta}_k = 0, \delta_k \neq 0)\right\} \leq L_p p^{1-\beta-(\underline{\omega} r - \beta + \Lambda)^2/(4\underline{\omega} r)}.$$

Since the error above is no more than the error rate established in Theorem 1 provided that $\underline{\omega} = \bar{\omega}$, it does not affect the rate optimality of the whole identification procedure as long as the error made in the following excising step is under control.

The fake signals generated by the transformation are able to survive from the thresholding if

$$\hat{T}_k \geq 2s \log p, \qquad k \notin S_\beta.$$

To excise fake signals, we implement the L_0 penalization approach. For the purpose of variable selection, this approach directly penalizes the number of nonzero parameters but is hampered by high dimensionality since it requires an exclusive search of all 2^p sub-models. However, as pointed out by Ji and Jin (2012), this NP hard problem can be circumvented thanks to an important consequence of conducting the thresholding. To see it, we let $\mathcal{U}(s)$ be a set including all components survived from the thresholding

$$\mathcal{U}(s) = \{k : \hat{T}_k \geq 2s \log p, 1 \leq k \leq p\}. \tag{4.2}$$

We define $V_0 = \{1, \ldots, p\}$ to be a set of notes and

$$\Omega^*(i, j) = \hat{\Omega}(i, j)\mathrm{I}_{\{|\hat{\Omega}(i,j)| \geq \log p/n\}} \tag{4.3}$$

to be regularized $\hat{\Omega}$. The reason for regularizing $\hat{\Omega}$ is that although it is in general a sparse estimate of Ω, it could contain some noisy elements. Therefore, $\log p/n$ is simply chosen to further remove those noisy elements if there exists any. According to the Gaussian graph theory, given the precision matrix Ω^*, any $i \neq j \in V_0$ are connected if and only if $\Omega^*(i, j) \neq 0$. Similar to Ji and Jin (2012), Lemma 4 summarizes the consequence of conducting the thresholding.

LEMMA 4. *Assume conditions* (C1)–(C4). *Let* $s \in (0, (\underline{\omega}r + \beta - \Lambda)^2/(4\underline{\omega}r))$, $\beta \in (1/2, 1)$ *and* $\beta - \Lambda < \underline{\omega}r < (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$. *With probability* $1 - L_p p^{-\beta-(\underline{\omega}r-\beta+\Lambda)^2/(4\underline{\omega}r)}$, $\mathcal{U}(s)$ *are split into disconnected clusters of size no more than a positive integer* K *with respect to* (V_0, Ω^*).

According to Lemma 4, the L_0 penalization approach can be effectively applied to each self-connected subset with relatively small size. Let $I_0 = \{i_1, \ldots, i_m\}$ be one of the self-connected subsets with size $m \leq K$, and $\hat{A} = \hat{\Omega}^{I_0, I_0}$ be an $m \times m$ matrix with $\hat{\Omega}^{I_0, I_0}(k, l) = \hat{\Omega}(i_k, i_l)$. To excise the fake signals in I_0, we find an m-dimensional vector $\hat{\delta}(I_0)$, each component of which is equal to either 0 or δ^{date} or $-\delta^{\text{date}}$, to minimize the following function:

$$n\{(\bar{\hat{Z}}_1 - \bar{\hat{Z}}_2)^{I_0} - \hat{A}\delta\}^T \hat{A}^{-1}\{(\bar{\hat{Z}}_1 - \bar{\hat{Z}}_2)^{I_0} - \hat{A}\delta\} + (\lambda^{\text{date}})^2\|\delta\|_0, \tag{4.4}$$

where λ^{date} and δ^{date} are two tuning parameters.

After we apply the L_0 penalization approach to all the self-connected subsets, each δ_k for $k = 1, \ldots, p$ is eventually determined by the proposed DATE procedure which can be summarized by the following algorithm:

(1) transform data X_{ij} to obtain $\hat{Z}_{ij} = \hat{\Omega}X_{ij}$ where $\hat{\Omega}$ is estimated Ω;

(2) conduct the thresholding described by (4.1) such that the coordinates $k = 1, \ldots, p$ are assigned to either $\mathcal{U}(s)$ or its complement $\mathcal{U}^c(s)$ where $\mathcal{U}(s)$ is defined in (4.2). For all $k \in \mathcal{U}^c(s)$, we set $\hat{\delta}_k = 0$;

(3) allocate $l \in \mathcal{U}(s)$ into $h \geq 1$ self-connected subsets $\{I_0^{(1)}, \ldots, I_0^{(h)}\}$ with respect to (V_0, Ω^*). For $I_0^{(1)}$, $\delta(I_0^{(1)})$ is equal to $\hat{\delta}(I_0^{(1)})$ each component of which is chosen to be either 0 or δ^{date} or $-\delta^{\text{date}}$ in order to minimize the penalized function (4.4). Repeat the same procedure to other $I_0^{(j)}$ where $j \in \{2, \ldots, h\}$ to determine δ_l for $l \in \mathcal{U}(s)$.

To easily measure the performance of the proposed DATE procedure, we further assume the following condition which is analogous to (C1) but requires a slightly stronger signal strength than (C1). A similar strategy was also taken by Ji and Jin (2012) for variable selection and Ji and Zhao (2014) for multiple testing in the high-dimensional regression problem.

(C1)′ Similar to (C1), the components of δ follow the mixture distribution with κ_p being a distribution on the support $[-(1+\eta)\sqrt{2r\log p/n}, -\sqrt{2r\log p/n}] \cup$

$[\sqrt{2r\log p/n}, (1+\eta)\sqrt{2r\log p/n}]$ where $\eta \le \frac{\beta-\Lambda}{\sqrt{C_0 r}} \frac{\sqrt{\beta r}}{\sqrt{(\underline{\omega} r-\beta+\Lambda)^2+4\underline{\omega} r\beta}}$, $\beta \in (1/2, 1)$ and the constant C_0 is defined in (C2).

Note that although the signal strength in (C1)$'$ can be stronger than that in (C1), the support in (C1) overlaps the support in (C1)$'$ at $-\sqrt{2r\log p/n}$ and $\sqrt{2r\log p/n}$. As we will discuss later, the overlapping of two supports is crucial to show the tightness of the lower bound established in Theorem 1. The following theorem establishes the upper bound of the risk (2.2) for the proposed DATE procedure.

THEOREM 3. *Assume* (C2)–(C4) *and* (C1)$'$. *Let* $s \in (0, (\underline{\omega} r+\beta-\Lambda)^2/(4\underline{\omega} r))$ *and* $\beta-\Lambda < \underline{\omega} r < (\sqrt{1-\Lambda}+\sqrt{1-\beta})^2$, *and set the tuning parameters in* (4.4) *to be*

$$\lambda^{\text{date}} = \sqrt{2(\beta-\Lambda)\log p}, \qquad \delta^{\text{date}} = \sqrt{2r\log p/n}.$$

As $p \to \infty$, *the weighted risk* (2.2) *for the DATE satisfies*

$$H(\Lambda) \le L_p p^{1-\beta-(\underline{\omega} r-\beta+\Lambda)^2/(4\underline{\omega} r)}.$$

Since $(\underline{\omega} r-\beta+\Lambda)^2/(4\underline{\omega} r) \le (\bar{\omega} r-\beta+\Lambda)^2/(4\bar{\omega} r)$, the lower bound in Theorem 1 is no greater than the upper bound in Theorem 3. Especially, these two bounds match each other if $\bar{\omega} = \underline{\omega}$. However, as noted by a reviewer, the two bounds are established under different supports for signals. To show the rate optimality of the proposed procedure, we let Π represent any mixture distribution of δ_k satisfying condition (C1), and let $\hat{\psi}$ be any decision rule. Theorem 1 shows that

$$\min_{\hat{\psi}} \min_{\Pi} H(\Lambda) \ge L_p p^{1-\beta-(\bar{\omega} r-\beta+\Lambda)^2/(4\bar{\omega} r)} \quad \text{for } -r < (\Lambda-\beta)/\underline{\omega} < r. \tag{4.5}$$

Note that the minimum in (4.5) is taken over with respect to all the decision rules $\hat{\psi}$ and signal distributions specified in condition (C1). The universal lowest rate in (4.5) can be attained [the equality in (4.5) holds] by setting $\hat{\psi}$ as the Bayesian rule $\hat{\theta}_j$ in (2.3) and the mixture distribution to be Π^*, a special distribution in (C1) where κ_p has support only on $\sqrt{2r\log p/n}$ or $-\sqrt{2r\log p/n}$. According to Theorem 3, the proposed DATE procedure is able to achieve the lowest rate in Theorem 1 when the signal distribution is Π^*, which shows that (4.5) is tight and the proposed procedure is rate optimal. A similar argument can be given for the rate optimality of the proposed procedure in the mFNR, which will be discussed as follows.

Our ultimate goal is to apply the DATE procedure to signal identification. So we need to ensure that it can successfully control the FDR at any desired level $\alpha < 1$. By carefully reviewing the whole procedure, we see that the thresholding

step (4.1) is designated to control the false negatives and the success of the FDR control is determined only by the excising step (4.4) where the role is played by the tuning parameter λ^{date}. Due to the adoption of L_0 penalty, smaller value of λ^{date} allows more toleration for the false positives, and thus leads to greater FDR. It turns out that if we subtract an additional term from the λ^{date} in Theorem 3, the mFDR can be successfully controlled at $\alpha < 1$ and the rate of the mFNR is accordingly established by Theorem 4.

THEOREM 4. *Assume conditions* (C2)–(C4) *and* (C1)′. *Choose* $s \in (0, \beta)$, $\beta - \Lambda < \underline{\omega} r < (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$ *and* $\Lambda = (\sqrt{\underline{\omega} r} - \sqrt{\beta})^2$. *As* $p \to \infty$, *by setting the tuning parameters of the DATE as*

$$\lambda^{\text{date}} = \sqrt{2(\beta - \Lambda)\log p - \Upsilon}, \qquad \delta^{\text{date}} = \sqrt{2r \log p / n},$$

where

$$\Upsilon = \frac{4\underline{\omega} r}{\underline{\omega} r + \beta - \Lambda}\left(\frac{1}{2}\log\log p + \log\left\{\frac{\alpha\sqrt{\pi}(\underline{\omega} r + \beta - \Lambda)}{2\sqrt{\underline{\omega} r}(1-\alpha)}\right\}\right).$$

Then

$$\text{mFDR} \le \alpha \quad \textit{and} \quad \text{mFNR} \le L_p p^{-\beta - (\sqrt{\underline{\omega} r} - \sqrt{\beta})^2}.$$

Although the upper bound of the mFNR above with $\bar{\omega} = \underline{\omega} = 1$, was also established in Ji and Zhao (2014), they were derived under different models and conditions. A more detailed discussion on the connection and difference between the current sample means problem and the linear regression problem will be provided in Section 8. Since $\bar{\omega} r \ge \underline{\omega} r > \beta$, the optimal rate of the mFNR in Theorem 2 is not faster than the rate in Theorem 4 and two rates are equal to each other asymptotically if $\bar{\omega} = \underline{\omega}$. This, combining with the fact that $\text{mFDR} \le \alpha < 1$, shows that the proposed DATE procedure is optimal in that it minimizes the mFNR subject to the constraint that mFDR is controlled at the desired level $\alpha < 1$.

There are three tuning parameters needed to estimated in the proposed signal identification procedure: the level of threshold s in (4.1), two tuning parameters δ^{date} and λ^{date} in (4.4). To select tuning parameters λ^{date} and δ^{date}, we estimate the sparsity β, the signal magnitude r and $\underline{\omega}$ by the following estimators:

$$\begin{aligned} \hat{\beta} &= -\log\left\{\frac{1}{p}\sum_{k=1}^{p} \mathrm{I}(\hat{T}_k > 2q\log p)\right\} \Big/ \log p, \\ \hat{r} &= \frac{1}{2p^{1-\hat{\beta}}\log p}\sum_{k=1}^{p}\frac{\hat{T}_k - 1}{\hat{\omega}_{kk}}\mathrm{I}(\hat{T}_k > 2q\log p), \qquad \hat{\underline{\omega}} = \min_{1\le k\le p}\hat{\omega}_{kk}, \end{aligned} \tag{4.6}$$

where q is another threshold level controlling the accuracy of estimate in β and r. The question of how to properly choose both s and q is addressed in Theorem 5.

With two tuning parameters λ^{date} and δ^{date} estimated by plugging the $\hat{\beta}, \hat{r}, \hat{\underline{\omega}}$ into the expressions defined in Theorem 4, the following theorem shows that the performance of the DATE procedure with estimated parameters (4.6) is asymptotically equivalent to the DATE in Theorem 4.

THEOREM 5. *Assume conditions* (C2)–(C4) *and* (C1)′. *As* $p \to \infty$, *by setting* $s \in (0, \beta)$ *in* (4.1), $q \in (\beta, \underline{\omega} r)$ *in* (4.6) *and estimating the tuning parameters as*

$$\hat{\lambda} = 2\hat{s}\log p, \qquad \hat{\lambda}^{\text{date}} = \sqrt{2(\hat{\beta} - \hat{\Lambda})\log p - \hat{\Upsilon}}, \qquad \hat{\delta}^{\text{date}} = \sqrt{2\hat{r}\log p/n},$$

where

$$\hat{\Lambda} = (\sqrt{\hat{\underline{\omega}}\hat{r}} - \sqrt{\hat{\beta}})^2,$$

$$\hat{\Upsilon} = \frac{4\hat{\underline{\omega}}\hat{r}}{\hat{\underline{\omega}}\hat{r} + \hat{\beta} - \hat{\Lambda}}\left(\frac{1}{2}\log\log p + \log\left\{\frac{\alpha\sqrt{\pi}(\hat{\underline{\omega}}\hat{r} + \hat{\beta} - \hat{\Lambda})}{2\sqrt{\hat{\underline{\omega}}\hat{r}}(1-\alpha)}\right\}\right),$$

and $\hat{\beta}$, $\hat{r}$ *and* $\hat{\underline{\omega}}$ *are given by* (4.6), *then*

$$\text{mFDR} \le \alpha \quad \textit{and} \quad \text{mFNR} \le L_p p^{-\beta - (\sqrt{\underline{\omega} r} - \sqrt{\beta})^2}.$$

Although two threshold levels s and q are not explicitly specified, simulation studies in Table 1 demonstrate that the proposed procedure is insensitive to (s, q) as long as $s \in (0, \beta)$ and $q \in (\beta, \underline{\omega} r)$, where β is assumed to be known in order to separate the two intervals. In practice, β is unknown and can be estimated by $\hat{\beta}$ in (4.6), which, however, relies on the properly chosen q. If $\underline{\omega} r > 1$, $q = 1$ will fall into $(\beta, \underline{\omega} r)$, the proposed procedure can be implemented by choosing $q = 1$ and $s \in (0, \hat{\beta})$ where $\hat{\beta}$ is obtained by (4.6) with $q = 1$. If $\underline{\omega} r \le 1$, choosing $q = 1$ in (4.6) can screen out too many signals and thus leads to an overestimated β. This obstacle can be overcome by utilizing other existing methods to estimate the sparsity β or equivalently, the number of (false) null hypotheses. For instance, Schweder and Spjøtvoll (1982) propose a method to estimate the number of true null hypotheses based on a linear fit of the empirical distribution of p-values. Storey (2002) considers estimating the number of true null hypotheses by the number of p-values greater than some threshold λ and then scaled by $1 - \lambda$. And Meinshausen and Bühlmann (2005) provide an estimator that is a lower bound for the number of false null hypotheses under general dependence structures between test statistics.

The optimality of the proposed DATE is established for the signal in the partial recovery region with $\underline{\omega} r < (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$. If $\underline{\omega} r \ge (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$, the region is the full recovery region. The lower bounds of the weighted risk $H(\Lambda)$ and the mFNR corresponding to this region converge to zero as r tends to infinity at each fixed large value of p as shown in Theorems 1 and 2. However, even when $\underline{\omega} r \ge (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$, the upper bounds for these two rates corresponding

to the full recovery region will not vanish, since the proposed DATE procedure involves data transformation, precision matrix and tuning parameters estimation each of which contributes non-negligible error at the order of $o(p^{-1})$. Although this error is very small, it becomes prominent and dominant as r is big enough to make two upper bounds established in Theorems 3, 4 and 5 smaller order of $o(p^{-1})$, and consequently the upper bounds of the weighted risk $H(\Lambda)$ and the mFNR will be at the rate of $o(p^{-1})$.

5. Some extensions. The proposed procedure can be extended to other signal recovery problems. One natural extension is to recover differences in the contrasts among multiple population mean vectors. Consider that $i = 1, \ldots, g$ in (1.1) where the number of populations $g > 2$, and suppose that the total g mean vectors are partitioned into two sub-groups: one has g_1 mean vectors $\mu_1, \ldots, \mu_{g_1}$ and the other consists of the remainder of $g_2 = g - g_1$ mean vectors $\mu_{g_1+1}, \ldots, \mu_g$. Letting

$$\iota = \frac{\mu_1 + \cdots + \mu_{g_1}}{g_1} - \frac{\mu_{g_1+1} + \cdots + \mu_g}{g_2} = (\iota_1, \ldots, \iota_p)^T, \tag{5.1}$$

we want to determine all nonzero components of ι, which, in genetic studies, corresponds to identify all the differentially expressed genes subject to the contrasts among different treatments.

Let

$$\begin{aligned} n_g = &\left(g_1^2 g_2^2 \prod_{i=1}^{g} n_i \right) \\ &\Big/ \left(g_1^2 \left(\prod_{i=1}^{g_1} n_i \right) \left(\sum_{i=g_1+1}^{g} \prod_{\substack{j \in \{g_1+1,\ldots,g\} \\ j \neq i}} n_j \right) \right. \\ &\left. + g_2^2 \left(\prod_{i=g_1+1}^{g} n_i \right) \left(\sum_{i=1}^{g_1} \prod_{\substack{j \in \{1,\ldots,g_1\} \\ j \neq i}} n_j \right) \right), \end{aligned}$$

and define

$$\Omega_g = \left\{ \frac{n_g}{g_1^2} \sum_{i=1}^{g_1} \frac{\Sigma_i}{n_i} + \frac{n_g}{g_2^2} \sum_{i=g_1+1}^{g} \frac{\Sigma_i}{n_i} \right\}^{-1}.$$

By assuming that all the conditions (C1)–(C4) are imposed to the recovery of sparse components of ι subject to some proper notation replacement, the proposed DATE procedure can be extended to the problem of multiple population means contrasts.

Specially, the proposed DATE procedure is implemented for the contrasts by first transforming the original X_{ij} into $\hat{\mathcal{Z}}_{ij} = \hat{\Omega}_g X_{ij}$ where $\hat{\Omega}_g$ is the estimated Ω_g

based on the method proposed in Section 3. The reason for the data transformation is to enhance the magnitude of the components of ι, which is similar to the problem of two-sample sparse differences recovery. With the transformed data, a similar thresholding step is conducted to remove nonsignal bearing dimensions by replacing $\hat{T}_k$ in (4.1) with

$$\hat{\mathscr{T}}_k = \frac{n_g(\frac{1}{g_1}\sum_{i=1}^{g_1}\bar{\hat{\mathcal{Z}}}_i^{(k)} - \frac{1}{g_2}\sum_{i=g_1+1}^{g}\bar{\hat{\mathcal{Z}}}_i^{(k)})^2}{\hat{\omega}_{g,kk}},$$

where $\bar{\hat{\mathcal{Z}}}_i^{(k)} = \sum_{j=1}^{n_i}\hat{\mathcal{Z}}_{ij}^{(k)}/n_i$ for $k = 1, \ldots, p$. The survival are then cleaned by choosing an m-dimensional vector $\hat{\iota}(I_0)$ each component of which is equal to either 0 or ι^{date} or $-\iota^{\text{date}}$ to minimize the following L_0 penalization similar to (4.4):

$$n_g\left\{\left(\frac{1}{g_1}\sum_{i=1}^{g_1}\bar{\hat{\mathcal{Z}}}_i - \frac{1}{g_2}\sum_{i=g_1+1}^{g}\bar{\hat{\mathcal{Z}}}_i\right)^{I_0} - \hat{A}_g\iota\right\}^T \hat{A}_g^{-1}$$
$$\times\left\{\left(\frac{1}{g_1}\sum_{i=1}^{g_1}\bar{\hat{\mathcal{Z}}}_i - \frac{1}{g_2}\sum_{i=g_1+1}^{g}\bar{\hat{\mathcal{Z}}}_i\right)^{I_0} - \hat{A}_g\iota\right\} + (\lambda^{\text{date}})^2\|\iota\|_0,$$

where $I_0 = \{i_1, \ldots, i_m\}$ is one of the self-connected subsets with size $m \le K$, and $\hat{A}_g = \hat{\Omega}_g^{I_0,I_0}$ is an $m \times m$ matrix with $\hat{\Omega}_g^{I_0,I_0}(k,l) = \hat{\Omega}_g(i_k, i_l)$. When all the tuning parameters are given in Theorem 5 with n replaced by n_g, it can be shown by similar derivations that the procedure for contrasts is rate optimal for the mFNR with the mFDR controlled at a pre-selected level, and the optimal rate of mFNR is specified in Theorem 5.

Another extension of the DATE procedure is to recover the sparse differences between two covariance matrices $(\delta_{kl})_{p\times p} = \Sigma_1 - \Sigma_2$, where $\delta_{kl} = \sigma_{kl}^{(1)} - \sigma_{kl}^{(2)}$ and $\Sigma_1 = (\sigma_{kl}^{(1)})_{p\times p}$ and $\Sigma_2 = (\sigma_{kl}^{(2)})_{p\times p}$. Testing the equality of two covariance matrices has been an important research topic [see Schott (2007), Srivastava and Yanagihara (2010), Li and Chen (2012), Cai, Liu and Xia (2013)], which has the practical application of comparing the difference in dependence among the measurements of the genes subject to different treatments. In addition to testing the equality of two covariance matrices, it is very often interesting to recover the differences between two covariance matrices. To generalize our procedure for recovering sparse nonzero components δ_{kl} for $1 \le k, l \le p$, we first define a $\tilde{p} = p(p+1)/2$ dimensional vector $\delta_{\Sigma_1-\Sigma_2} = (\delta_{11}, \delta_{12}, \ldots, \delta_{1p}, \delta_{22}, \ldots, \delta_{2p}, \ldots, \delta_{pp})^T$, which consists of the diagonal and upper triangular elements of $\Sigma_1 - \Sigma_2$. By letting $\tilde{X}_{ij} = (\tilde{X}_{ij}^{(11)}, \ldots, \tilde{X}_{ij}^{(1p)}, \ldots, \tilde{X}_{ij}^{(pp)})^T$ where $\tilde{X}_{ij}^{(kl)} = X_{ij}^{(k)}X_{ij}^{(l)} - \bar{X}_i^{(k)}\bar{X}_i^{(l)}$ for $i = 1, 2$, $\hat{\delta}_{\Sigma_1-\Sigma_2}$ is an unbiased estimate of $\delta_{\Sigma_1-\Sigma_2}$ with elements

$$\hat{\delta}_{kl} = \frac{1}{n_1}\sum_{j=1}^{n_1}\tilde{X}_{1j}^{(kl)} - \frac{1}{n_2}\sum_{j=1}^{n_2}\tilde{X}_{2j}^{(kl)}.$$

It can be shown that the leading order covariance matrix of $\sqrt{n}\hat{\delta}_{\Sigma_1-\Sigma_2}$ is

$$\begin{aligned}\tilde{V} &= (v_{kl,k'l'})_{\tilde{p}\times\tilde{p}} \\ &= \big((1-\varsigma)\{\sigma^{(1)}_{kk'}\sigma^{(1)}_{ll'} + \sigma^{(1)}_{kl'}\sigma^{(1)}_{lk'}\} + \varsigma\{\sigma^{(2)}_{kk'}\sigma^{(2)}_{ll'} + \sigma^{(2)}_{kl'}\sigma^{(2)}_{lk'}\}\big)_{\tilde{p}\times\tilde{p}}.\end{aligned}$$

By letting $\tilde{\Omega} = \tilde{V}^{-1}$ and $\hat{\tilde{\Omega}}$ be a consistent estimator of the sparse $\tilde{\Omega}$ specified in (C3), the transformed random vectors $\hat{\tilde{Z}}_{ij} = \hat{\tilde{\Omega}}\tilde{X}_{ij}$. With the transformed data, a thresholding step similar to (4.1) is conducted based on the statistics

$$\hat{\tilde{T}}_k = \frac{n\{\bar{\hat{\tilde{Z}}}^{(k)}_1 - \bar{\hat{\tilde{Z}}}^{(k)}_2\}^2}{\hat{\tilde{\omega}}_{kk}}, \qquad k = 1, \ldots, \tilde{p},$$

where $\hat{\tilde{\omega}}_{kk}$ is the kth diagonal element of $\hat{\tilde{\Omega}}$. All the survivals are then cleaned by applying the L_0-penalty method similar to (4.4). Specifically, let $I_0 = \{i_1, \ldots, i_m\}$ be one of the self-connected subsets with size $m \le K$, and $\hat{\tilde{A}} = \hat{\tilde{\Omega}}^{I_0,I_0}$ be an $m \times m$ matrix with $\hat{\tilde{\Omega}}^{I_0,I_0}(k,l) = \hat{\tilde{\Omega}}(i_k, i_l)$. Then we minimize the following function with respect to $\hat{\delta}_{\Sigma_1-\Sigma_2}(I_0)$ by setting each component of $\delta_{\Sigma_1-\Sigma_2}(I_0)$ to be either 0 or $\delta^{\text{date}}_{\Sigma_1-\Sigma_2}$ or $-\delta^{\text{date}}_{\Sigma_1-\Sigma_2}$:

$$n\big\{(\bar{\hat{\tilde{Z}}}_1 - \bar{\hat{\tilde{Z}}}_2)^{I_0} - \hat{\tilde{A}}\delta_{\Sigma_1-\Sigma_2}\big\}^T \hat{\tilde{A}}^{-1}\big\{(\bar{\hat{\tilde{Z}}}_1 - \bar{\hat{\tilde{Z}}}_2)^{I_0} - \hat{\tilde{A}}\delta_{\Sigma_1-\Sigma_2}\big\} + \big(\tilde{\lambda}^{\text{date}}\big)^2\|\delta_{\Sigma_1-\Sigma_2}\|_0,$$

where two tuning parameters $\tilde{\lambda}^{\text{date}}$ and $\delta^{\text{date}}_{\Sigma_1-\Sigma_2}$ can be chosen in Theorem 5 with p replaced by $\tilde{p}$. The asymptotic properties of the above procedure for recovering sparse differences between two covariance matrices are expected to be similar to those established in Theorems 1–5. Due to the limited space, we will not pursue them in this paper and leave explorations to future study.

6. Simulation study. Simulation studies were conducted to demonstrate the performance of the proposed procedure for signals recovery under different combinations of signal sparsity controlled by β, signal strength r and data dependence. The proposed procedure is denoted by DATE_Ω if Ω is known and $\text{DATE}_{\hat{\Omega}}$ if Ω is unknown. For comparison, other three signal recovery procedures were also considered. The first competitor is the BH procedure that was implemented as follows: each of p coordinates is tested by the two-sample t-test to obtain the ordered p-values $P_{(1)} < \cdots < P_{(p)}$. Based on the cutoff value $m = \max\{1 \le k \le p : P_{(k)} \le k\alpha/p\}$, the coordinates with $P_i \le P_{(m)}$ are treated as signal bearing dimensions.

The second competitor is the PFA procedure based on principle factor approximation proposed by Fan, Han and Gu (2012). Suppose that $(J_1, \ldots, J_p)^T \sim \text{N}((\sqrt{n}\delta_1, \ldots, \sqrt{n}\delta_1)^T, (1-\varsigma)\Sigma_1 + \varsigma\Sigma_2)$ where $J_i = \sqrt{n}\{\bar{X}^{(i)}_1 - \bar{X}^{(i)}_2\}$. The eigenvalues of $(1-\varsigma)\Sigma_1 + \varsigma\Sigma_2$ are $\lambda_1 > \cdots > \lambda_p$ whose corresponding orthonormal eigenvectors are $\gamma_1, \ldots, \gamma_p$. Then J_i can be written as

$$J_i = \sqrt{n}\delta_i + b_i^T W + V_i,$$

where $b_i = (b_{i1}, \ldots, b_{ik})^T$, $(b_{1j}, \ldots, b_{pj})^T = \sqrt{\lambda_j}\gamma_j$, the factors $W = (W_1, \ldots, W_k)^T \sim \mathrm{N}(0, \mathrm{I}_k)$ and the random errors $(V_1, \ldots, V_p)$ are weakly dependent by properly choosing k such that $\sqrt{\lambda_{k+1}^2 + \cdots + \lambda_p^2}/(\lambda_1 + \cdots + \lambda_p) < \varepsilon$ with a small ε. According to the authors, a dependence-adjusted procedure for signal recovery is conducted based on the test statistics $a_i(J_i - b_i^T \hat{W})$ for $i = 1, \ldots, p$ where $a_i = (1 - \sum_{h=1}^k b_{ih}^2)^{-1/2}$ and $\hat{W}$ is obtained by first choosing an integer m that corresponds to the smallest 95% of $|J_i|$'s, and then applying the L_1-regression to the equation

$$J_l = b_l^T W + V_l, \qquad l = 1, \ldots, m.$$

The BH procedure was then implemented to the dependence-adjusted p-values given by $2\Phi(-|a_i(J_i - b_i^T \hat{W})|)$ for $i = 1, \ldots, p$.

The third competitor is the Sphering procedure by Allen and Tibshirani (2012). The original procedure is proposed to utilize row and column covariances to decorrelate the noise in the transposable data matrix meaning that neither the row nor the column variables are considered to be independent. We modify it for our column dependent data as follows. Two population mean vectors μ_1 and μ_2 are estimated by the sample mean vectors $\bar{X}_1$ and $\bar{X}_2$, respectively. Define the $p \times 1$ noise vector $\hat{N}_{ij} = X_{ij} - \bar{X}_i$ for $i = 1, 2$ and $j = 1, \ldots, n_i$. After the noise vector is sphered by $\hat{\Omega}^{1/2}\hat{N}_{ij}$, the sphered data $\hat{X}_{ij} = \bar{X}_i + \hat{\Omega}^{1/2}\hat{N}_{ij}$. Then the BH procedure was implemented to the p-values obtained from the two-sample t-test based on the test statistics $\sqrt{n}\{\bar{\hat{X}}_1^{(k)} - \bar{\hat{X}}_2^{(k)}\}$ for $k = 1, \ldots, p$.

The random samples $\{X_{ij}\}$ were generated from $\mathrm{N}(\mu_i, \Sigma)$ for $i = 1, 2$. Without loss of generality, $\mu_1 = 0$ and μ_2 had $[p^{1-\beta}]$ nonzero coordinates which were uniformly and randomly drawn from $\{1, \ldots, p\}$. The magnitude of each nonzero entry of μ_2 was randomly drawn from the interval $[\sqrt{r \log p/n}, \sqrt{3r \log p/n}]$ and then multiplied by a random sign. Four models were considered for the covariance matrix $\Sigma = (\sigma_{ij})$:

(a) AR(1) model: $\sigma_{ij} = \rho^{|i-j|}$ for $1 \le i, j \le p$.

(b) Block diagonal model: $\sigma_{ii} = 1$ for $i = 1, \ldots, p$, and $\sigma_{ij} = 0.6$ for $2(k-1) + 1 \le i \ne j \le 2k$ where $k = 1, \ldots, [p/2]$.

(c) Penta-diagonal model: $\sigma_{ii} = 1$ for $i = 1, \ldots, p$, $\sigma_{ij} = 0.5$ for $|i - j| = 1$ and $\sigma_{ij} = 0.2$ for $|i - j| = 2$.

(d) Random sparse matrix model: first generate a $p \times p$ matrix Γ each row of which has only one nonzero element that is randomly chosen from $\{1, \ldots, p\}$ with magnitude generated from Unif(1, 2) multiplied by a random sign. Σ is then obtained by standardizing $\Gamma\Gamma^T + \mathbf{I}$ to have unit diagonal elements.

To apply the $\mathrm{DATE}_{\hat{\Omega}}$, we need to estimate Ω. For models (a)–(c), the Cholesky decomposition approach [Bickel and Levina (2008)] was implemented. Recall that the precision matrix Ω can be decomposed as $\Omega = (I - A)^T D^{-1}(I - A)$ where A is a lower triangular matrix with zero diagonals and D is a diagonal matrix. The

elements below the diagonal element on the kth row of A can be thought as the regression coefficients of the kth component on its predecessors, and the kth diagonal element of D is the corresponding residual variance. Let A_τ be the τ-banded lower triangular matrix of A and D_τ be the corresponding residual variances on the diagonals. The τ-banded precision matrix $\Omega_\tau = (I - A_\tau)^T D_\tau^{-1}(I - A_\tau)$. Given a sample, A_τ and D_τ can be estimated by the least square estimation, which leads to

$$\hat{\Omega}_\tau = (I - \hat{A}_\tau)^T \hat{D}_\tau^{-1}(I - \hat{A}_\tau),$$

where the banding width parameter τ in the estimation of Ω was chosen according to the data-driven procedure proposed by Bickel and Levina (2008). For a given data set, we divided it into two subsamples by repeated ($N = 50$ times) random data split. For the lth split, $l \in \{1, \ldots, N\}$, we let $\hat{\Sigma}_\tau^{(l)} = \{(I - \hat{A}_\tau^{(l)})^T\}^{-1} \hat{D}_\tau^{(l)} \times (I - \hat{A}_\tau^{(l)})^{-1}$ be the Cholesky decomposition of Σ obtained from the first subsample by taking the same approach described in previous section for $\hat{A}_\tau^{(l)}$ and $\hat{D}_\tau^{(l)}$. Also we let $S_n^{(l)}$ be the sample covariance matrix obtained from the second sub-sample. Then the banding parameter τ is selected as

$$\hat{\tau} = \min_\tau \frac{1}{N} \sum_{l=1}^{N} \| \hat{\Sigma}_\tau^{(l)} - S_n^{(l)} \|_F, \tag{6.1}$$

where $\| \cdot \|_F$ denotes the Frobenius norm.

The sparse Ω in model (d) can be estimated by some available packages such as *glasso*, *Covpath* and *CLIME* that are coded based on different estimation approaches discussed in Section 3. To implement a fast algorithm, we adopted the *glasso* which chooses the nonnegative definite matrix $\hat{\Omega}_{Glasso}$ to maximize the L_1-regularized log-likelihood:

$$\log\det(\Sigma^{-1}) - \operatorname{tr}(S_n^* \Sigma^{-1}) - \rho_g \| \Sigma^{-1} \|_1, \tag{6.2}$$

where S_n^* is given by (3.3) and ρ_g is a tuning parameter controlling the L_1 shrinkage. To select the regularization parameter ρ_g, we considered the package *huge* developed by Zhao et al. (2012) where three methods are provided: the stability approach for regularization selection, rotation information criterion and a likelihood-based extended Bayesian information criterion.

The theoretical signal enhancement demonstrated by (3.1) and (3.4) can be explored in practice based on N simulations via AR(1) model (a) by choosing $p = 500$, $n_1 = n_2 = 30$ and $\beta = 0.6$. The gain in lth simulation, denoted by $B^{(l)}$, is defined to be

$$B^{(l)} = \frac{1}{p^{1-\beta}} \sum_{k \in S_\beta^{(l)}} \frac{\delta_{\Omega,k}^{(l)}}{\sqrt{\omega_{kk}}\delta_k^{(l)}} \qquad \text{for } l = 1, \ldots, N. \tag{6.3}$$

The signal gain with known Ω was evaluated by averaging $B^{(1)}, \ldots, B^{(N)}$. The signal gain subject to estimated $\hat{\Omega}$ was conducted similarly by replacing Ω with

$\hat{\Omega}$ in (6.3). With $\rho = 0.6$ in AR(1) model, $\underline{\omega} = 1.56$ and $\bar{\omega} = 2.13$. The average signal gain based on $N = 100$ simulations with known Ω was 1.46. With estimated $\hat{\Omega}$, the average gain was 1.28, which was close to $\sqrt{\bar{\omega}} = 1.46$. So our simulation results demonstrate that the standardized signal strength can be boosted by the transformation Ω or $\hat{\Omega}$, confirming the theoretical findings in (3.1) and (3.4).

The performance of each signal recovery procedure was evaluated by mFDR, mFNR and the average number of true positives mean(TP) based on 100 replications. The sparsity parameter β was chosen to be 0.6. Therefore, the true positives that need to be recovered were $[500^{0.4}] = 12$ when $p = 500$, and $[1000^{0.4}] = 16$ when $p = 1000$. The nominal FDR level was set at $\alpha = 0.05$. Figure 3 displays the performance of two proposed procedures DATE_{Ω} and $\text{DATE}_{\hat{\Omega}}$, the BH procedure integrated with two-sample t-test, the PFA procedure and the Sphering procedure with different values of signal strength r and data dependence ρ under the AR(1) model (a) when $p = 500$. In the first column of the figure, data were weakly dependent and all five procedures had the mFDR controlled around the nominal level 0.05 except $r = 0.4$. The distortion of the mFDR at $r = 0.4$ is due to the fact that the signals fall into the region of no recovery since $r < \beta/\bar{\omega}$ with $\bar{\omega} = 1.08$ when $\rho = 0.2$. With the dependence increased from $\rho = 0.2$ to 0.6, the inflation of mFDR for DATE_{Ω} and $\text{DATE}_{\hat{\Omega}}$ was mitigated since $r > \beta/\underline{\omega}$ with $\underline{\omega} = 1.56$ when $\rho = 0.6$. Although all five procedures performed similarly in terms of the mFNR and mean(TP) under weak dependence $\rho = 0.2$, both DATE_{Ω} and $\text{DATE}_{\hat{\Omega}}$ identified more mean(TP) close to the number of true signals $[500^{0.4}] = 12$ for stronger signal strength r, and suffered less mFNR than the other three procedures with stronger dependence $\rho = 0.6$, which confirms that the data dependence can be utilized by the proposed procedures for signal identification. When dimension p was increased from 500 to 1000, Figure 4 demonstrates the results similar to Figure 3. Especially with stronger signal strength r, the recovery of signals by both DATE_{Ω} and $\text{DATE}_{\hat{\Omega}}$ was closer to the number of true signals $[1000^{0.4}] = 16$.

The performance of five procedures subject to various dependent structures defined in models (b)–(d) were displayed in Figures 5–7. Both DATE_{Ω} and $\text{DATE}_{\hat{\Omega}}$ performed better than the other three in terms of mFNR and mean(TP) with the mFDR reasonably controlled at the nominal level 0.05. Figure 8 demonstrates the effect of small sample sizes on the five procedures based on the AR(1) model (a). Again DATE_{Ω} and $\text{DATE}_{\hat{\Omega}}$ that employed data dependence were the top two procedures in terms of the mFNR and mean(TP) compared with the other three procedures. As the sample sizes were increased, DATE_{Ω} and $\text{DATE}_{\hat{\Omega}}$ were separated from the other three and the performance of the data-driven $\text{DATE}_{\hat{\Omega}}$ was closer to that of the best performer DATE_{Ω}.

DATE_{Ω} depends on the level of threshold s and $\text{DATE}_{\hat{\Omega}}$ depends on both s and q, which are required to be chosen from intervals $(0, \beta)$ and $(\beta, \underline{\omega} r)$, respectively. Table 1 displays the performance of both DATE_{Ω} and $\text{DATE}_{\hat{\Omega}}$ in terms of mFDR and mFNR with $\beta = 0.6$ subject to different values of s and q under model (a)

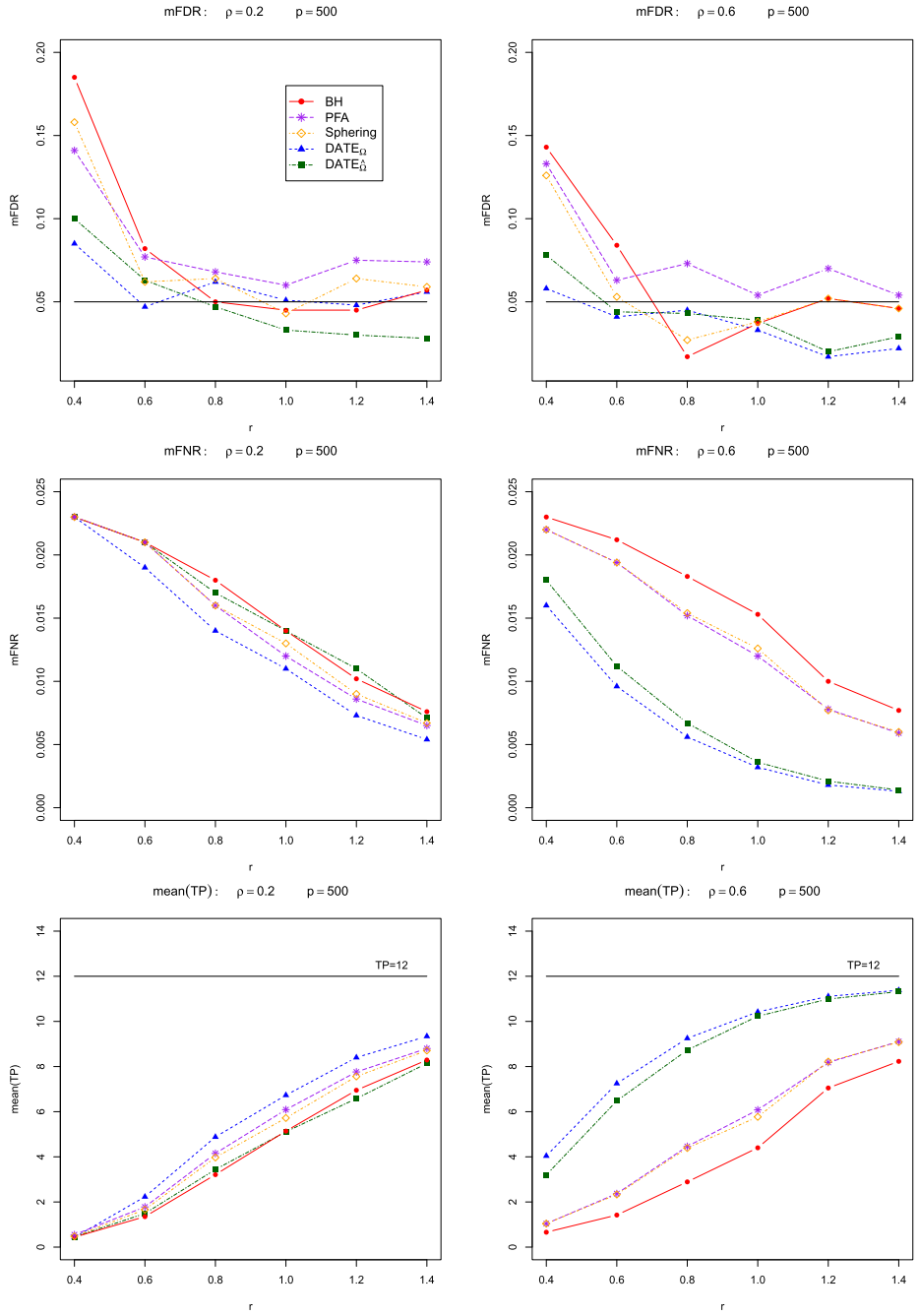

FIG. 3. *The mFDR, mFNR and mean(TP) yielded by* DATE_{Ω}, $\mathrm{DATE}_{\hat{\Omega}}$ *and the BH procedure integrated with t-test, PFA and Sphering subject to different dependence under the* AR(1) *model* (a). *The dimension* $p = 500$, *sample sizes* $n_1 = 60$ *and* $n_2 = 60$, *the sparsity parameter* $\beta = 0.6$ *and the nominal FDR level* $\alpha = 0.05$.

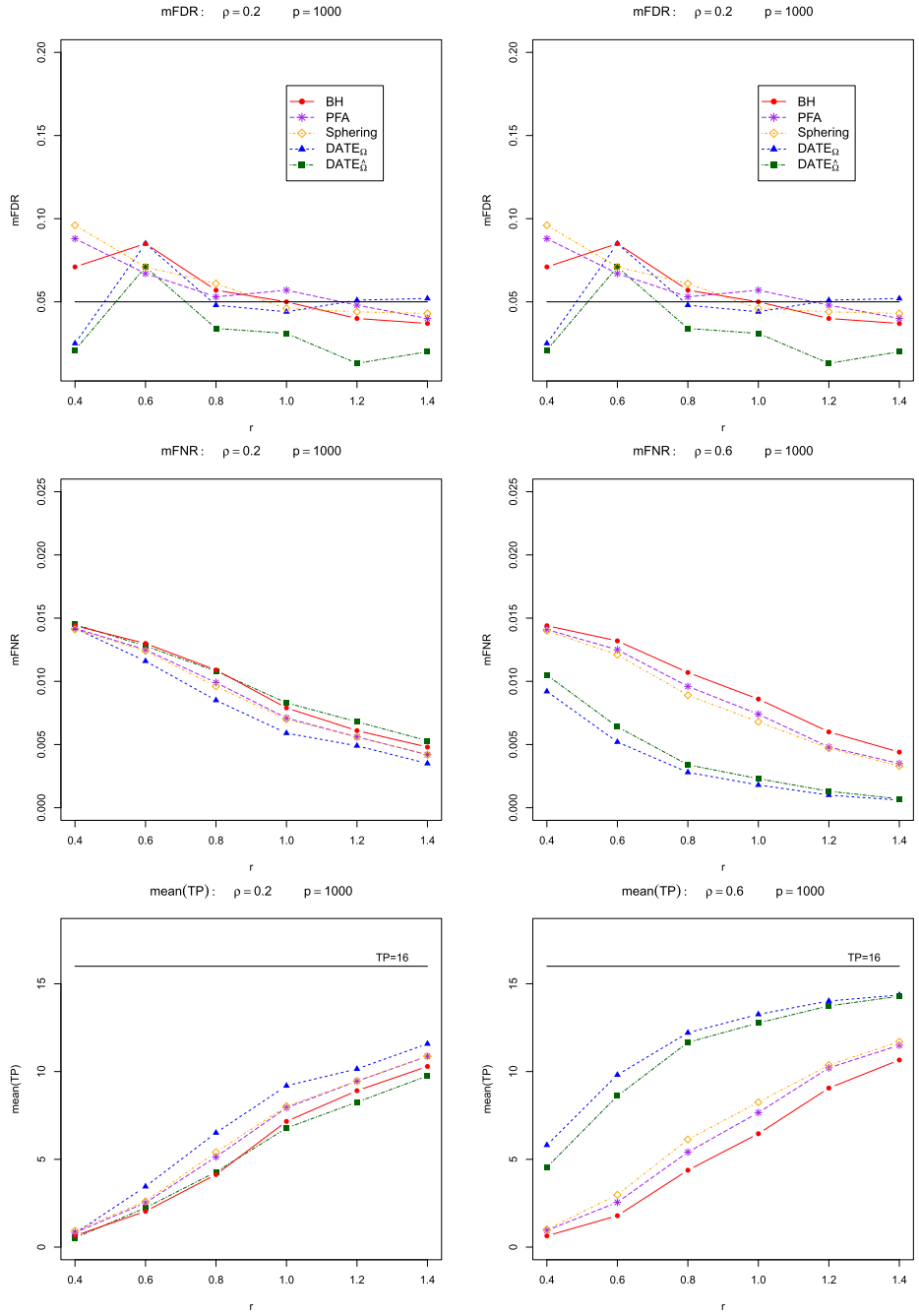

FIG. 4. *The mFDR, mFNR and mean(TP) yielded by* DATE_{Ω}, $\mathrm{DATE}_{\hat{\Omega}}$ *and the BH procedure integrated with t-test, PFA and Sphering subject to different dependence under the* AR(1) *model* (a). *The dimension* $p = 1000$, *sample sizes* $n_1 = 60$ *and* $n_2 = 60$, *the sparsity parameter* $\beta = 0.6$ *and the nominal FDR level* $\alpha = 0.05$.

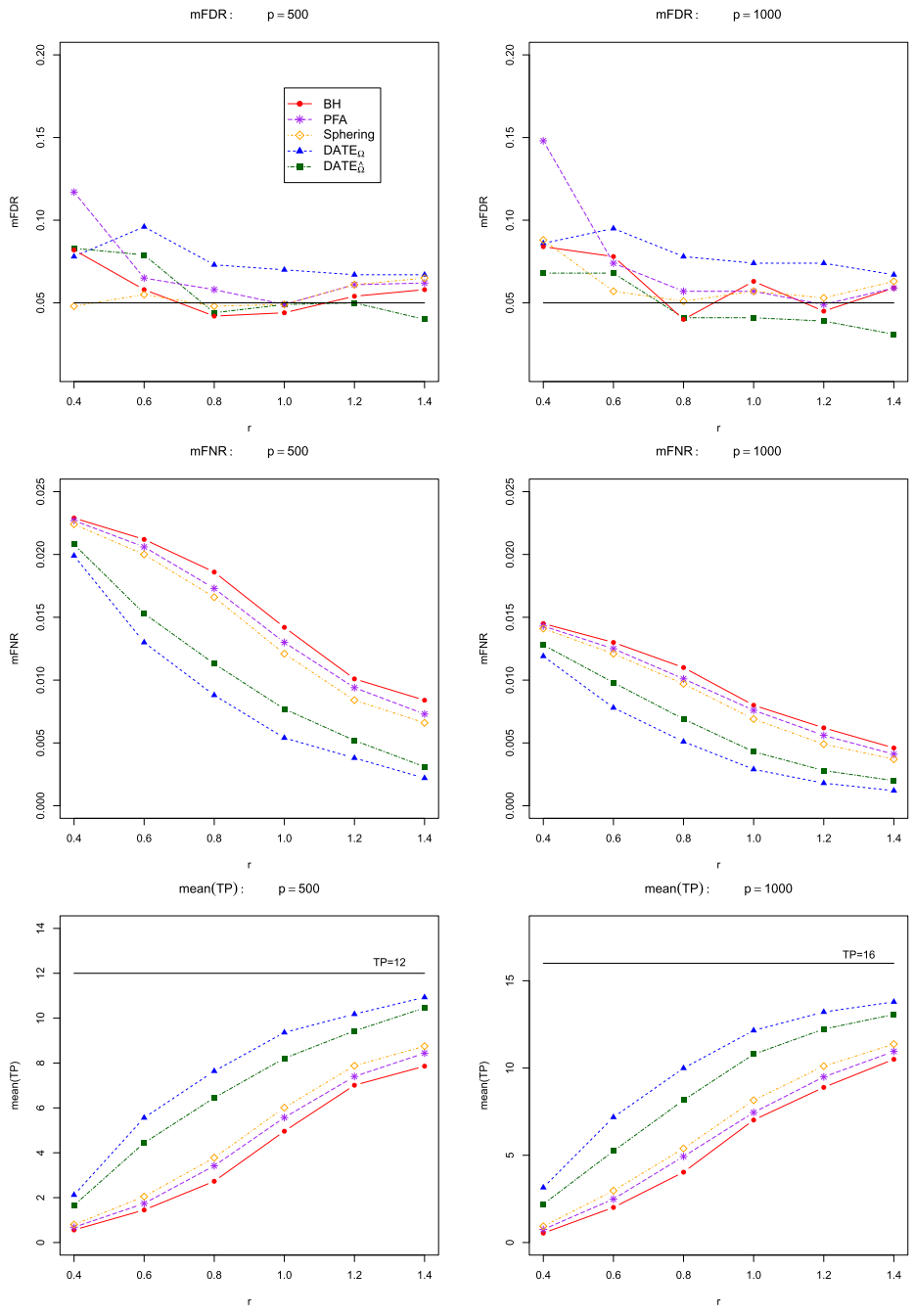

FIG. 5. *The mFDR, mFNR and mean(TP) yielded by* DATE_Ω, $\mathrm{DATE}_{\hat{\Omega}}$ *and the BH procedure integrated with t-test, PFA and Sphering under the block diagonal model* (b). *The sample sizes* $n_1 = 60$ *and* $n_2 = 60$, *the sparsity parameter* $\beta = 0.6$ *and the nominal FDR level* $\alpha = 0.05$.

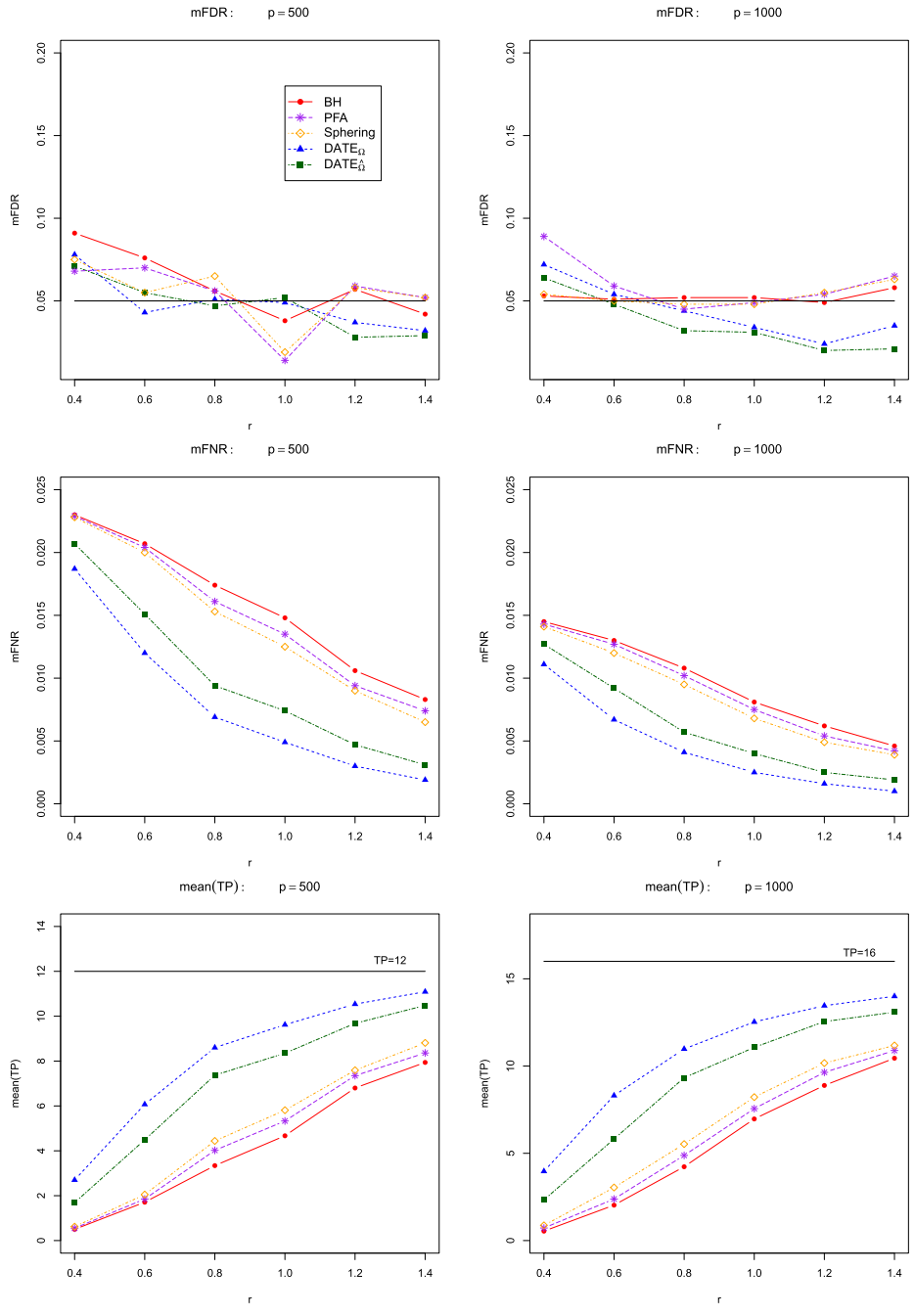

FIG. 6. *The mFDR, mFNR and mean(TP) yielded by* DATE_{Ω}, $\mathrm{DATE}_{\hat{\Omega}}$ *and the BH procedure integrated with t-test, PFA and Sphering under the penta-diagonal model* (c). *The sample sizes* $n_1 = 60$ *and* $n_2 = 60$, *the sparsity parameter* $\beta = 0.6$ *and the nominal FDR level* $\alpha = 0.05$.

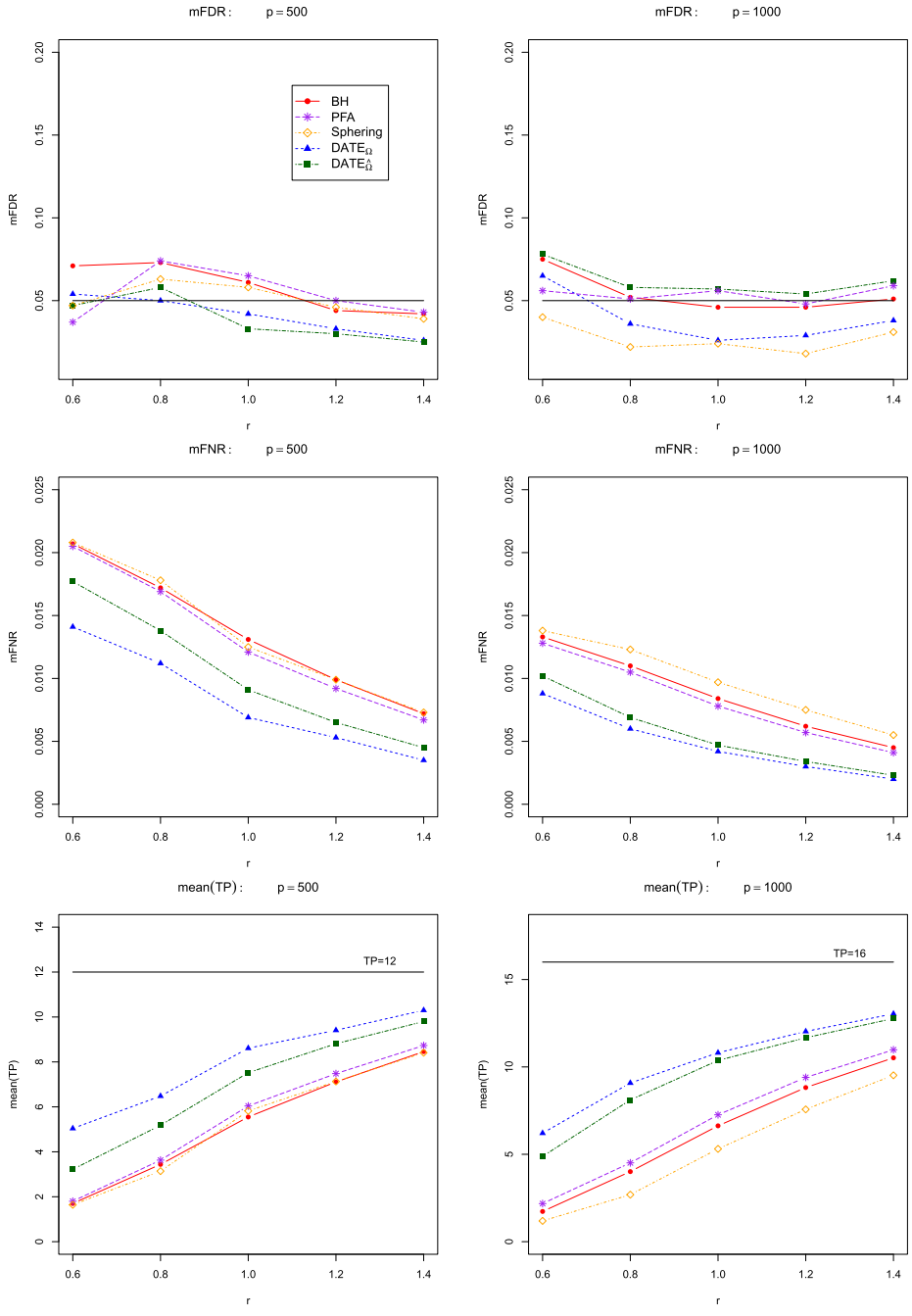

FIG. 7. *The mFDR, mFNR and mean(TP) yielded by* DATE_Ω, $\mathrm{DATE}_{\hat{\Omega}}$ *and the BH procedure integrated with t-test, PFA and Sphering under the random sparse model* (d). *The sample sizes* $n_1 = 60$ *and* $n_2 = 60$, *the sparsity parameter* $\beta = 0.6$ *and the nominal FDR level* $\alpha = 0.05$.

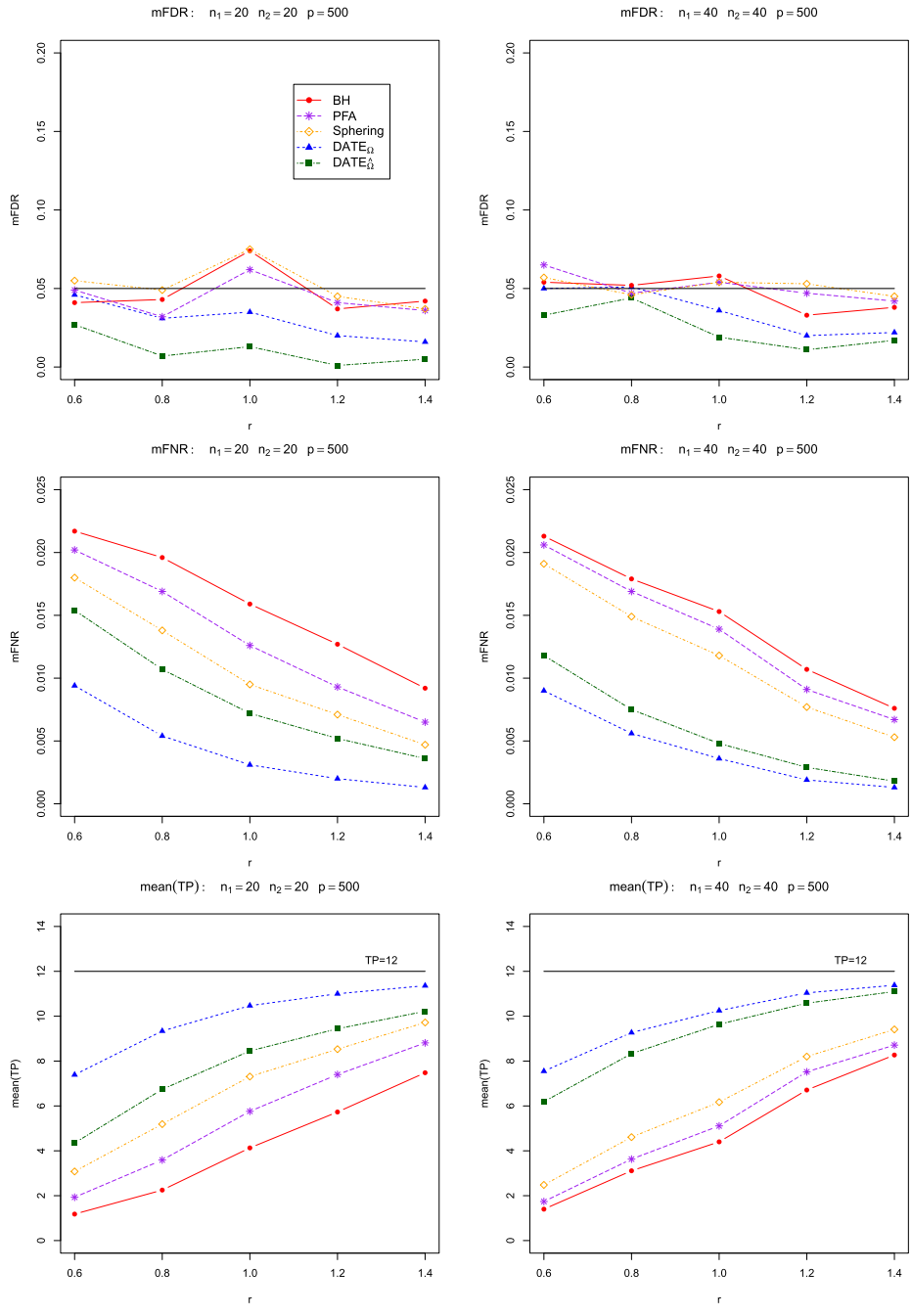

FIG. 8. *The mFDR, mFNR and mean(TP) yielded by* DATE_{Ω}, $\mathrm{DATE}_{\hat{\Omega}}$ *and the BH procedure integrated with t-test, PFA and Sphering subject to small sample sizes under the* AR(1) *model* (a). *The sparsity parameter* $\beta = 0.6$ *and the nominal FDR level* $\alpha = 0.05$.

TABLE 1

The performance of DATE_{Ω} *and* $\text{DATE}_{\hat{\Omega}}$ *in mFDR and mFNR under model* (a), *where* $\rho = 0.6$, $r = 0.8$, $p = 500$ *and* $n_1 = n_2 = 60$. *Different values of tuning parameters* s *and* q *are chosen from two intervals separated by* $\beta = 0.6$. *The computation is based on* 100 *simulations. The standard deviations of the false discovery proportion and the false nondiscovery proportion are listed in parentheses*

	q							
	0.75		**0.80**		**0.85**		**0.90**	
s	DATE_{Ω}	$\text{DATE}_{\hat{\Omega}}$	DATE_{Ω}	$\text{DATE}_{\hat{\Omega}}$	DATE_{Ω}	$\text{DATE}_{\hat{\Omega}}$	DATE_{Ω}	$\text{DATE}_{\hat{\Omega}}$
				mFDR				
0.35	0.039 (0.062)	0.035 (0.060)	0.042 (0.067)	0.036 (0.062)	0.045 (0.064)	0.035 (0.057)	0.039 (0.062)	0.032 (0.061)
0.40	0.039 (0.060)	0.046 (0.075)	0.040 (0.070)	0.037 (0.067)	0.039 (0.058)	0.031 (0.049)	0.047 (0.067)	0.039 (0.064)
0.45	0.043 (0.070)	0.035 (0.064)	0.039 (0.054)	0.033 (0.050)	0.035 (0.048)	0.025 (0.045)	0.043 (0.055)	0.027 (0.051)
0.50	0.044 (0.065)	0.041 (0.057)	0.046 (0.067)	0.037 (0.057)	0.031 (0.055)	0.025 (0.051)	0.038 (0.057)	0.031 (0.053)
				mFNR				
0.35	0.005 (0.003)	0.007 (0.004)	0.005 (0.003)	0.006 (0.004)	0.005 (0.003)	0.007 (0.004)	0.006 (0.003)	0.007 (0.004)
0.40	0.006 (0.003)	0.007 (0.004)	0.005 (0.003)	0.006 (0.003)	0.005 (0.003)	0.006 (0.004)	0.005 (0.003)	0.007 (0.003)
0.45	0.006 (0.003)	0.007 (0.003)	0.005 (0.003)	0.006 (0.004)	0.005 (0.003)	0.007 (0.003)	0.006 (0.003)	0.007 (0.004)
0.50	0.006 (0.003)	0.007 (0.003)	0.006 (0.003)	0.007 (0.004)	0.006 (0.003)	0.007 (0.003)	0.006 (0.003)	0.007 (0.004)

where $\rho = 0.6$, $r = 0.8$, $p = 500$ and $n_1 = n_2 = 60$. As we can see, the proposed procedure is insensitive to the choice of s and q as long as they are chosen properly from the intervals.

7. Empirical study. We applied the proposed procedure to a human breast cancer dataset (GDS2250) available at http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2250. The data were analyzed by Richardson et al. (2006) to provide insight into the molecular pathogenesis of Sporadic basal-like cancers (BLC) that is a distinct class of human breast cancers. As discussed by Richardson et al. (2006), BLC specimens display X chromosome abnormalities in the sense that most of the BLC cases lack markers of a normal inactive X chromosome, which are rare in non-BLC specimens. Our interest on this data set is to display these X chromosome abnormalities by identifying the differentially ex-

TABLE 2
The number of differentially expressed genes identified by the BH, the DATE and both on chromosome X with the FDR controlled at the level of $\alpha = 0.01, 0.005$ *and* 0.001

FDR-controlled level	BH	DATE	Both
0.01	52	56	38
0.005	43	50	33
0.001	27	39	22

pressed genes between the BLC and non-BLC. For this purpose, we formed two samples by taking 18 sporadic BLC specimens and 20 non-BLC specimens from the original data, and each sample contains 1438 genes obtained from chromosome X.

To apply the DATE procedure, we first estimated Ω in (1.3) where $\Sigma_1 \neq \Sigma_2$ in general. We applied the method proposed in Section 3 to estimate Ω where the regular sample covariance matrix is replaced by the two-sample version of sample covariance matrix S_n^* defined by (3.3). By replacing the regular sample covariance matrix S with S_n^* in (6.2), $\Sigma_w^{-1} = (\Sigma_1 + n_1/n_2\Sigma_2)^{-1}$ was first estimated by the *glasso* method described in Section 6. Then Ω was estimated accordingly based on the relationship $\Omega = (1 + n_1/n_2)\Sigma_w^{-1}$. Except the DATE procedure, we also considered the classical BH procedure integrated with two-sample t-test as a comparison.

In order to identify the differentially expressed genes, the FDR was chosen to be controlled at $\alpha = 0.001, 0.005$ and 0.01. Table 2 summarizes the number of differentially expressed genes identified by the BH only and the DATE only, and both procedures. By carefully investigating the genes identified by both procedures, we found that the XIST (X inactive specific transcript) gene was discovered. This gene is in charge of an early developmental process in females and provides dosage equivalence between males and females. The XIST difference is thought as one of the characteristics for the BLC according to Richardson et al. (2006). Moreover, the authors argue that there exists the over-expression of a small subset of genes on chromosome X for BLC. In Table 3, we list additional 17 genes that are identified by the DATE but missed by the BH with the FDR controlled at $\alpha = 0.001$. The association of these genes with the BLC may deserve some further biological investigation.

8. Discussion. Signal identification is different from its closely related problem of signal detection. Whereas the detection focuses purely on the presence of signals, the signal identification is designated for locating the signals. The advantage of dependence for signal detection was explored by Hall and Jin (2010) who showed that the detection boundary can be lowered by incorporating the data correlation. Moreover, the benefit of dependence on signal identification in the context

TABLE 3
The differentially expressed genes identified by the DATE not by the BH on chromosome X with the FDR controlled at level 0.001

Gene symbol	Location	Description
PTCHD1	Xp22.11	Patched domain containing 1
DMD	Xp21.2	Dystrophin
SLC9A6	Xq26.3	Solute carrier family 9 (sodium/hydrogen exchanger), member 6
KAL1	Xp22.32	Kallmann syndrome 1 sequence
TMSB15B	Xq22.2	Thymosin-like 8
GPR64	Xp22.13	G Protein-coupled receptor 64
ATP6AP1	Xq28	Atpase, H+ transporting, lysosomal accessory protein 1
NXT2	Xq23	Nuclear transport factor 2-like export factor 2
CLCN4	Xp22.3	Chloride channel 4
VGLL1	Xq26.3	Vestigial like 1 (Drosophila)
BEX1	Xq22	Brain expressed, X-linked 1
SLC6A14	Xq23	Solute carrier family 6 (amino acid transporter), member 14
BCOR	Xp21.2-p11.4	Bcl6 corepressor
BCORL1	Xq25-q26.1	Bcl6 corepressor-like 1
MUM1L1	Xq22.3	Melanoma associated antigen (mutated) 1-like 1
SYTL5	Xp21.1	Synaptotagmin-like 5
RLIM	Xq13-q21	Ring finger protein, LIM domain interacting

of variable selection has been addressed for the sparse regression model (1.4) by Genovese et al. (2012), Jin, Zhang and Zhang (2014) and Ke, Jin and Fan (2014). The current paper attempts to address the advantageous effect of dependence on recovering δ for the model (1.1) based on the multiple testing procedure. Our analysis shows that both full and partial signal identification boundaries for dependent data are lower than those without dependence. Our result, combined with the findings in Hall and Jin (2010), shows that data dependence is advantageous in both signal detection and signal identification. Furthermore, when both signals and precision matrix are sparse, the proposed DATE procedure takes advantage of dependence through the transformation to enhance the signal strength and is shown to have the faster convergence rate in mFNR than other procedures without taking data dependence into account.

At last, we would like to point out the connection and difference between the current work and that of Ji and Zhao (2014) in more detail. With $\bar{\omega} = \underline{\omega} = 1$, the lower and upper bounds of the mFNR in our Theorems 2 and 4, respectively, were also established in Ji and Zhao (2014) for the high-dimensional regression model (1.4). It is not surprising to see this because, as we have pointed out in Section 1, the model (1.1) considered in current work and the model (1.4) for Ji and Zhao (2014) are both related to the Stein's normal means model (1.2). However, the results established in both papers are developed for different parameters along different directions: the current article extends Hall and Jin (2010) from signal detection to signal identification, and Ji and Zhao (2014) extends Ji and Jin (2012)

from variable selection to multiple testing. Due to the different interests, the settings considered in both papers are different. In addition, some of our technical conditions are weaker than those in Ji and Zhao (2014). For example, Conditions (C2) and (C3) in our paper define a class of matrices with the sparse structure, where both M_p and c_p are allowed to grow with p logarithmically. However, in Ji and Zhao (2014), the corresponding parameters are fixed. Furthermore, Condition (C4) in our paper specifies the exponential growth of dimension p with n, but Ji and Zhao (2014) assumes the polynomial growth of dimension p with n.

Another significant difference is that in our two-sample testing problem (1.1), both Σ_1 and Σ_2 are assumed to be unknown. However, in Ji and Zhao (2014), the covariance of $X^T Y$ is $X^T X$, which is known. The effect of estimating the precision matrix Ω on the identification boundary is thereafter addressed in the current paper. An important contribution of the current work is utilizing dependence in recovering differences between two high-dimensional mean vectors. We have demonstrated the advantageous effect of dependence on recovering δ in (1.1), and unveiled the key roles of $\bar{\omega}$ and $\underline{\omega}$ by explicitly incorporating them in the expressions of the signal identification boundary and the rate of the mFNR. However, the UPT procedure in Ji and Zhao (2014) does not consider the advantageous effect of data dependence on signal enhancement.

Acknowledgments. We would like to thank Pengsheng Ji and Zhigen Zhao for helpful discussion, and insightful comments and suggestions. We also thank the Editor, an Associate Editor and two referees for their constructive comments which have significantly improved the quality and presentation of the paper.

SUPPLEMENTARY MATERIAL

Supplementary material for "A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence" (DOI: 10.1214/16-AOS1459SUPP; .pdf). The supplementary material provides the proofs of Lemmas 1–4 and Theorems 1–5.

REFERENCES

ALLEN, G. I. and TIBSHIRANI, R. (2012). Inference with transposable data: Modelling the effects of row and column correlations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 721–743. MR2965957

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CAI, T., LIU, W. and XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277. MR3174618

CAI, T. T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372. MR3164870

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302

FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 499–517. MR1924303

GENOVESE, C. R., JIN, J., WASSERMAN, L. and YAO, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13** 2107–2143. MR2956354

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357

JI, P. and JIN, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40** 73–103. MR3013180

JI, P. and ZHAO, Z. (2014). Rate optimal multiple testing procedure in high-dimensional regression. Technical report.

JIN, J. (2012). Comment: "Estimating false discovery proportion under arbitrary covariance dependence". *J. Amer. Statist. Assoc.* **107** 1042–1045. MR3010891

JIN, J., ZHANG, C.-H. and ZHANG, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15** 2723–2772. MR3270749

KE, Z. T., JIN, J. and FAN, J. (2014). Covariate assisted screening and estimation. *Ann. Statist.* **42** 2202–2242. MR3269978

LI, J. and CHEN, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* **40** 908–940. MR2985938

LI, J. and ZHONG, P. (2016). Supplement to "A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence." DOI:10.1214/16-AOS1459SUPP.

MEINSHAUSEN, N. and BÜHLMANN, P. (2005). Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika* **92** 893–907. MR2234193

QIU, X., KLEBANOV, L. and YAKOVLEV, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 34. MR2183944

RICHARDSON, A., WANG, Z., NICOLO, A., LU, X., BROWN, M., MIRON, A., LIAO, X., IGLEHART, J., LIVINGSTON, D. and GANESAN, S. (2006). X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell* **9** 121–132.

SCHOTT, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Comput. Statist. Data Anal.* **51** 6535–6542. MR2408613

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

SRIVASTAVA, M. S. and YANAGIHARA, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *J. Multivariate Anal.* **101** 1319–1329. MR2609494

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498. MR1924302

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657

SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 393–424. MR2649603

XIE, J., CAI, T. T. and LI, H. (2011). Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika* **98** 273–290. MR2806428

ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The `huge` package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062. MR2930633

DEPARTMENT OF MATHEMATICAL SCIENCES
KENT STATE UNIVERSITY
KENT, OHIO 44242
USA
E-MAIL: junli@math.kent.edu

DEPARTMENT OF STATISTICS AND PROBABILITY
MICHIGAN STATE UNIVERSITY
EAST LANSING, MICHIGAN 48823
USA
E-MAIL: pszhong@stt.msu.edu