Electronic Journal of Statistics
Vol. 17 (2023) 950–979
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2119

Estimating the proportion of signal variables under arbitrary covariance dependence

X. Jessie Jeng

Department of Statistics, North Carolina State University,
2311 Stinson Dr., Raleigh, NC 27695-8203, USA
e-mail: xjjeng@ncsu.edu

Abstract: Estimating the proportion of signals hidden in a large number of noise variables is a pervasive objective in scientific research. In this paper, we consider realistic, yet theoretically challenging scenarios with arbitrary covariance dependencies between variables. We quantify the overall level of covariance dependence using mean absolute correlation (MAC), and investigate the performance of a family of estimators across the full range of MAC values. We explore the joint effect of MAC dependence, signal sparsity, and signal intensity on estimator performance, and find that no single estimator in the family performs optimally across all MAC dependence levels. Based on this theoretical insight, we propose a new estimator that is better suited to arbitrary covariance dependencies. Our method compares favorably to several existing methods in a variety of finite-sample settings, including those with strong or weak covariance dependencies and real dependence structures from genetic association studies.

MSC2020 subject classifications: Primary 62F35; secondary 62F10.
Keywords and phrases: Dependence adaptivity, high-dimension data, lower bound estimator, sparse signal.

Received November 2021.

Contents

1. Introduction

We consider the problem of estimating the proportion of information-bearing signals that are sparsely located within a large number of noise variables. This problem is of interest in many scientific inquiries. For example, estimation of the signal proportion is required by multiple testing methods to calculate the local false discovery rate [10], to derive q-values [27], and to improve power [26, 12]. Furthermore, in many multi-stage studies, estimation of the signal proportion can assist in efficient pre-screening and sample size calculation [6]. A recent line of research, which focuses on retaining a high proportion of signals through efficient false negative control, also relies on the estimation of the signal proportion as a benchmark for signal inclusion [16, 18, 17].

Although the estimation of the signal proportion is in high demand, methodology development has encountered two major challenges. First, signals with different sparsity levels often require different estimation methods, while the sparsity levels of signals are unknown a priori. Secondly, the large set of variables under investigation may have complex dependence structures. There are a number of rigorously developed methods, however, most of them assume independence between variables [13, 24, 20] and sparsity levels within a certain range [7, 19]. An extensive review of the existing methods can be found in [8]. More recent developments extend the study to consider specific dependence structures. For example, [18] studies the problem assuming block-diagonal covariance structures for the variables; [15] considers the problem in linear regression and imposes certain dependence and sparsity conditions to facilitate accurate precision matrix estimation and bias mitigation. However, there is a lack of methods to consistently estimate the signal proportion under arbitrary covariance dependence when the signal sparsity is unknown and possibly falls in a wide range. Such an estimator could have far-reaching impact in real-world applications.

In this paper, we focus on the family of estimators introduced in [24], which covers several existing estimators in the literature. Members in the family have a desirable lower bound property that applies regardless of the characteristics of the signals being studied. In other words, estimators in the family provide conservative estimates for the signal proportion, regardless of how sparse or how weak the signals are. By analyzing the limiting distributions of empirical processes of the p-values, the most powerful estimator in the family has been discovered under independence.

Here, we consider variables that may be correlated in complex ways, and for which the limiting distribution of the empirical process of p-values is unknown and cannot be expressed analytically. The impact of dependence on the family members also depends on the level of signal sparsity, which is unknown. These complexities greatly complicate the problem and require new approaches to be developed.

In this paper, we present a novel strategy to quantify the degree of covariance dependence using mean absolute correlation (MAC), and investigate the performance of the family of estimators across the full range of MAC dependence levels and the full range of signal sparsiy. We find that the most powerful estimator under independence is no longer the optimal choice under arbitrary covariance dependence. Moreover, our study reveals that no single member in the family is most powerful under various MAC dependence levels.

Based on the theoretical insight, we propose a new estimator in an omnibus form that borrows power from different members in the family. In order to identify the most promising candidates among the infinitely many members of the family, we investigate two general scenarios classified by the joint effect of MAC dependence and signal sparsity, and study both the efficiency and power limits of the members. As a result, our new estimator has a surprisingly simple form that involves only two members of the family. This new estimator is proven to be more powerful than any other member of the family under almost arbitrary covariance dependence.

The new estimator is compared with several existing estimators in a variety of simulation settings, including those with strong or weak covariance dependencies and real dependence structures from genetic association studies. The existing estimators include members in the original family of [24], consistent estimators developed for independent variables and less sparse signals [13, 20], and lower bound estimators constructed based on recent developments in post hoc confidence bounds in multiple testing [2]. The results demonstrate that, while the winner among existing methods changes in different settings, the new estimator consistently outperforms or is comparable to the winner in most cases. Furthermore, the new estimator appears to be the most stable one over different levels of dependence and sparsity.

We apply the new estimator in two real-data analyses. The first dataset is from an expression quantitative trait loci (eQTL) study with 8637 candidate single-nucleotide polymorphisms (SNPs) that possess weak overall dependence in terms of the MAC level. The second dataset is from a classical association study, in which the microarray data of 4088 candidate genes possess strong overall dependence, as indicated by a high MAC level. These two real dependence structures have been investigated in simulation studies, and results in the real data analyses seem to be consistent with the patterns observed in the simulation, which supports the effectiveness of the new estimator in real-world applications.

2. Method and theory

Denote I_0 and I_1 as the sets of indices for noise and signal variables, respectively. We consider the marginal distribution of m variables as

$$X_j \sim F_0 \cdot 1\{j \in I_0\} + F_1 \cdot 1\{j \in I_1\}, \qquad j = 1, \ldots, m, \tag{1}$$

where F_0 and F_1 are the marginal null and signal distribution, respectively. We assume that F_0 is continuous and that the joint null distribution of $(X_1, \ldots, X_m)$ is known a priori, i.e., when $|I_1| = \emptyset$, $(X_1, \ldots, X_m)$ follows a known joint distribution. Define the signal proportion

$$\pi = |I_1|/m. \tag{2}$$

Our goal is to estimate the signal proportion π without needing to specify F_1. Note that the variables may be arbitrarily dependent on each other.

In this section, we first present a family of estimators originally introduced in [24] for independent variables. The original family was built upon the empirical processes of p-values and analyzed under independence. When variables are arbitrarily dependent, much of the original analysis cannot be applied. As the family of estimators is indexed by the choice of a bounding function, we first derive a general result on the effect of bounding functions under arbitrary dependence and sparsity levels. Then, we quantify the overall covariance dependence of the variables by the MAC level and derive new concentration inequalities to characterize the empirical processes of the variables across the full range of the MAC level. As a result, we are able to explicate the joint effect of MAC level and signal sparsity, and obtain valuable insight on the efficiency of different members in the family. These analyses eventually motivate us to come up with a new more powerful estimator under arbitrary covariance dependence.

2.1. A family of estimators

For presentation purposes, we perform inverse normal transformation as $Z_j = \Phi^{-1}(F_0(X_j))$, where Φ^{-1} is the inverse of the cumulative distribution function of $N(0, 1)$. After the transformation, we have

$$Z_j \sim \Phi \cdot 1\{j \in I_0\} + G \cdot 1\{j \in I_1\}, \qquad j = 1, \ldots, m, \tag{3}$$

where G denotes the signal distribution after the transformation, which remains unknown.

The family of estimators introduced in [24] is indexed by a strictly positive function, which is called a bounding function and denoted as $\delta(t)$. Given a bounding function $\delta(t)$, define

$$\hat{\pi}(\delta) = \sup_{t>0} \frac{\bar{F}_m(t) - 2\bar{\Phi}(t) - c_m(\delta; \alpha)\delta(t)}{1 - 2\bar{\Phi}(t)}, \tag{4}$$

where

$$\bar{F}_m(t) = m^{-1} \sum_{j=1}^{m} 1\{|Z_j| > t\},$$

$\bar{\Phi}(t) = 1 - \Phi(t)$, and $c_m(\delta;\alpha)$ is the so-called bounding sequence, whose value is determined as follows. Let

$$\bar{W}_m(t) = m^{-1} \sum_{j=1}^{m} 1\{|W_j| > t\},$$

where $(W_1, \ldots, W_m)$ are generated from the joint null distribution of $(Z_1, \ldots, Z_m)$, which is known a priori. Define

$$V_m(\delta) = \sup_{t>0} \frac{|\bar{W}_m(t) - 2\bar{\Phi}(t)|}{\delta\,(t)}. \tag{5}$$

For a given $\alpha > 0$, define the bounding sequence $c_m(\delta;\alpha)$ as a quantity that satisfies the following properties:

(a) $mc_m(\delta;\alpha) > m_0 c_{m_0}(\delta;\alpha)$, where $m_0 = |I_0|$, and

(b) $P(V_m(\delta) > c_m(\delta;\alpha)) < \alpha$ for all m.

It can be seen that $c_m(\delta;\alpha)$ is an upper bound of $V_m(\delta)$ at the level of $1-\alpha$, and $V_m(\delta)$ relies on the joint null distribution of $(Z_1, \ldots, Z_m)$ and the choice of the bounding function $\delta(t)$. We modified the original construction in [24] by adding the absolute sign in the numerator of (5) to stabilize $V_m(\delta)$ under dependence. To implement the $\hat{\pi}(\delta)$ estimators in real practice, $c_m(\delta;\alpha)$ can be numerically simulated as the $(1-\alpha)$-quantile of $V_m(\delta)$. More details of the implementation are provided at the end of Section 2.3. This version of $\hat{\pi}(\delta)$ is for signals of two-sided effects. Minor changes to accommodate one-sided signal effect is straightforward.

Intuitively speaking, for a given $\delta(t)$, $c_m(\delta;\alpha)$ indicates a normal range of $[\bar{F}_m(t) - 2\bar{\Phi}(t)]/\delta(t)$ when all variables are noise. Therefore, value of the observed $\bar{F}_m(t)$ that exceeds the normal range presents evidence for the existence of signals. Further, the property in (b) implies that $\bar{F}_m(t) - 2\bar{\Phi}(t) - c_m(\delta;\alpha)\delta(t)$ is a lower bound estimate for π for any $t > 0$. Therefore, a more efficient estimator can be constructed by taking the supremum of $\bar{F}_m(t) - 2\bar{\Phi}(t) - c_m(\delta;\alpha)\delta(t)$ over $t > 0$. The denominator $1 - 2\bar{\Phi}(t)$ is a term appeared in refined analysis, which can further improve the efficiency of the estimator.

The $\hat{\pi}(\delta)$ family covers a wide range of candidates depending on what the bounding function is. These candidates can perform very differently and excel in different settings. It has been shown that for independent variables, performances of the candidates would depend on the sparsity level of signals. Specifically, the signal proportion can be re-parameterized as $\pi = m^{-\gamma}$, $\gamma \in (0,1)$, with $\gamma \in (0,1/2)$ representing the relatively dense case and $\gamma \in [1/2,1)$ representing the more sparse case. Consistency of different members in the original family has been studied for $\gamma \in (0,1/2)$ and $\gamma \in [1/2,1)$, separately [24].

Here, we consider the estimation problem in the more general and realistic setting with arbitrary dependence and signal sparsity. When the joint null distribution of $(Z_1, \ldots, Z_m)$ has an arbitrary dependence structure, the limiting distribution of $V_m(\delta)$ in (5) is generally unknown and may not even have an analytic expression. This imposes the major challenge in analyzing the $\hat{\pi}(\delta)$ family under dependence. Our strategy to tackle the problem is to first derive a general result showing the effects of the bounding function and bounding sequence in providing lower bound and upper bound estimates of the signal proportion in Section 2.1. Then, we explicate the joint effects of dependence and signal sparsity under different choices of bounding function in Section 2.2.

Theorem 2.1 *Consider model (3). For a given bounding function $\delta(t)$, if a bounding sequence $c_m(\delta;\alpha)$ satisfying the properties in (a) and (b) is implemented in (4), then*

$$P(\hat{\pi}(\delta) < \pi) \geq 1 - \alpha. \tag{6}$$

On the other hand, if $\delta(t)$ is non-increasing with respect to t, and $G = G_m$ such that $G_m(\tau_m) \to 0$ or $G_m(-\tau_m) \to 1$ for some τ_m such that $\tau_m \gg 1$ and $\delta(\tau_m) \ll \pi/c_m(\delta;\alpha)$, then

$$P(\hat{\pi}(\delta) > (1-\epsilon)\pi) \to 1 \tag{7}$$

for any constant $\epsilon > 0$.

The expression $a \ll b$ (or $a \gg b$) means that $a/b = o(1)$ (or $b/a = o(1)$). The above theorem says that $\hat{\pi}(\delta)$ is a lower bound of π at $1-\alpha$ level, as shown in (6), if a bounding sequence $c_m(\delta;\alpha)$ satisfying (a) and (b) is implemented. On the other hand, as shown in (7), $\hat{\pi}(\delta)$ is an upper bound of $(1-\epsilon)\pi$ for any $\epsilon > 0$ under certain conditions on the signal distribution G, which essentially says that the signal effect, either positive ($G < \Phi$) or negative ($G > \Phi$), is strong enough, and the required magnitude depends on the bounding function $\delta(t)$, the bounding sequence $c_m(\delta;\alpha)$, and the true signal sparsity π.

If both the lower bound (with $\alpha = \alpha_m \to 0$) and the upper bound results hold for a given $\delta(t)$, then $\hat{\pi}(\delta)$ consistently estimates the true signal proportion, i.e., for any constant $\epsilon > 0$,

$$P((1-\epsilon)\pi \leq \hat{\pi}(\delta) < \pi) \to 1. \tag{8}$$

This general result holds for signals of arbitrary sparsity levels under general dependence.

2.2. Joint effect of dependence and sparsity on the $\hat{\pi}(\delta)$ family

As individual estimators in the $\hat{\pi}(\delta)$ family hinge on the choice of the bounding function, we further study their consistency with specific $\delta(t)$ functions. We focus on $\delta(t)$ of the form $\delta(t) = [\bar{\Phi}(t)]^{\theta}$, $\theta \in [0,1]$, as their effects on $\hat{\pi}(\delta)$ have been studied under independence [24]. Valuable insight can be obtained by comparing the estimators' performances under independence and dependence, which eventually helps us construct a new and more powerful estimator for dependent variables.

In order to explicate the effect of dependence on $\hat{\pi}(\delta)$, we assume that

$$(Z_1, \ldots, Z_m) \sim N_m(\mu, \Sigma), \tag{9}$$

where μ is a m-dimensional sparse vector with $\mu_j = A_j \cdot 1\{j \in I_1\}$, $A_j > 0$, and Σ is an arbitrary correlation matrix, known a priori. The condition that all $A_j > 0$ is for presentation simplicity and can be easily relaxed to $|A_j| > 0$.

Different from the existing analyses on $\hat{\pi}(\delta)$ under independence, where the limiting distributions of $V_m(\delta)$ with different $\delta(t)$ functions have readily applicable forms, theoretical study under arbitrary covariance dependence is more challenging as the limiting distributions of $V_m(\delta)$ are generally unknown. To proceed with the analysis, we calibrate covariance dependence through the measure of Mean Absolute Correlation (MAC) and derive new concentration inequalities to reflect the effects MAC dependence, signal sparsity, and signal intensity levels on the limiting behaviors of $V_m(\delta)$. The MAC measure is defines as

$$\bar{\rho}_\Sigma = \sum_{i=1}^{m} \sum_{j=1}^{m} |\Sigma_{ij}|/m^2. \tag{10}$$

A larger value of $\bar{\rho}_\Sigma$ indicates stronger overall dependence. Specifically, $\bar{\rho}_\Sigma = 1/m$ corresponds to the independent case, and $\bar{\rho}_\Sigma = O(1)$ occurs when, e.g., each variable is correlated with all the other variables at non-degenerating levels. The $\bar{\rho}_\Sigma$ measure has been discussed in the literature on multiple testing, as shown in, e.g., [25] and [9].

We also employ a discretization technique from [1] and [15] as follows. Define $\mathbb{T} = [\sqrt{\log\log m}, \sqrt{5\log m}] \cap \mathbb{N}$ and the discretized version of $V_m(\delta)$ as

$$V_m^*(\delta) = \max_{t \in \mathbb{T}} \frac{|\bar{W}_m(t) - 2\bar{\Phi}(t)|}{\delta(t)}. \tag{11}$$

Denote $c_m^*(\delta;\alpha)$ as the bounding sequence for $V_m^*(\delta)$, and define the corresponding proportion estimator as

$$\hat{\pi}^*(\delta) = \max_{t \in \mathbb{T}} \frac{\bar{F}_m(t) - 2\bar{\Phi}(t) - c_m^*(\delta;\alpha)\delta(t)}{1 - 2\bar{\Phi}(t)}. \tag{12}$$

Next, we explicates how the MAC dependence interacts with signal sparsity and signal intensity to influence the consistency of $\hat{\pi}^*(\delta)$ with $\delta(t) = [\bar{\Phi}(t)]^\theta$ and reveals very different results for $\theta \in [0, 1/2]$ and $\theta \in (1/2, 1]$.

Theorem 2.2 *Consider model (9). Let $\delta(t) = [\bar{\Phi}(t)]^\theta$ with $\theta \in [0, 1/2]$. Then, there exists a sequence $c_m^*(\delta;\alpha) = C_\alpha \sqrt{\bar{\rho}_\Sigma (\log m)^{\theta+1/2}}$, where $C_\alpha > 0$ is a large enough constant depending on α, that satisfies properties (a) and (b); and the corresponding estimator $\hat{\pi}^*(\delta)$ satisfies $P(\hat{\pi}^*(\delta) < \pi) \geq 1 - \alpha$.*

On the other hand, if $A_m (= \min_{j \in I_1} A_j)$ satisfies $A_m \gg 1$ and

$$A_m - \bar{\Phi}^{-1}\left(\frac{\pi^{1/\theta}}{\bar{\rho}_\Sigma^{-1/(2\theta)} (\log m)^{(\theta+1/2)/(2\theta)}}\right) \gg 1, \tag{13}$$

then $P(\hat{\pi}^*(\delta) > (1-\epsilon)\pi) \to 1$ for any constant $\epsilon > 0$.

The above theorem says that for $\delta(t) = [\bar{\Phi}(t)]^\theta$ with $\theta \in [0, 1/2]$, one can find a bounding sequence $c_m^*(\delta; \alpha) = C_\alpha \sqrt{\bar{\rho}_\Sigma (\log m)^{\theta+1/2}}$ to construct a lower bound estimator of π ($P(\hat{\pi}^*(\delta) < \pi) \geq 1 - \alpha$). On the other hand, for the upper bound result ($P(\hat{\pi}^*(\delta) > (1-\epsilon)\pi) \to 1$), effect of the constant C_α is asymptotically negligible and only the order of $c_m^*(\delta; \alpha)$ is used to derive condition (11) on signal intensity. Note that such a $c_m^*(\delta; \alpha)$ sequence may not be unique depending on the value of the constant C_α, and a smaller value of $c_m^*(\delta; \alpha)$ is preferred as indicated in Theorem 2.1. In the numerical implementation, we choose the bounding sequence as the $(1-\alpha)$-quantile of the empirical distribution of $V_m(\delta)$ as shown at the end of Section 2.3.

It can be seen that condition (13) becomes more stringent with sparser signals (smaller π) or stronger MAC dependence (larger $\bar{\rho}_\Sigma$). For the special case with independent variables, $\bar{\rho}_\Sigma = 1/m$ and (13) degenerates to

$$A_m - \bar{\Phi}^{-1}\left(\pi^{1/\theta} m^{1/2\theta} / (\log m)^{(\theta+1/2)/(2\theta)}\right) \to \infty,$$

which agrees with the sufficient and necessary condition for the consistency of $\hat{\pi}(\delta)$ under independence for relatively sparse signals. The comparison can be made by adopting the same parameterization as in Theorem 3 of [24] with $\pi = m^{-\gamma}$, $\gamma \in [1/2, 1)$, $\nu = \theta$, and $\kappa = 2$. Note that $\bar{\Phi}^{-1}(\pi^{1/\theta} / (\bar{\rho}_\Sigma^{1/(2\theta)} (\log m)^{(\theta+1/2)/(2\theta)}))$ is well-defined only for $\pi < \sqrt{\bar{\rho}_\Sigma (\log m)^{\theta+1/2}}$. In the case $\pi \geq \sqrt{\bar{\rho}_\Sigma (\log m)^{\theta+1/2}}$, condition (13) is simply $A_m \gg 1$.

Results in Theorem 2.2 indicate how the performance of $\hat{\pi}^*(\delta)$ with $\delta(t) = [\bar{\Phi}(t)]^\theta, \theta \in [0, 1/2]$, deteriorates as the MAC dependence gets stronger. These results, however, cannot be extended to $\hat{\pi}^*(\delta)$ with $\theta \in (1/2, 1]$. For the latter, we present the following theorem that applies to the full range of $\theta \in [0, 1]$.

Theorem 2.3 *Consider model (9). Let $\delta(t) = [\bar{\Phi}(t)]^\theta$ with $\theta \in [0, 1]$. Then, there exists a bounding sequence $c_m^*(\delta; \alpha) = C'_\alpha \sqrt{\log m}$, where $C'_\alpha > 0$ is a large enough constant depending on α, that satisfies properties (a) and (b); and the corresponding estimator $\hat{\pi}^*(\delta)$ satisfies $P(\hat{\pi}^*(\delta) < \pi) \geq 1 - \alpha$.*

On the other hand, if $A_m (= \min_{j \in I_1} A_j)$ satisfies

$$A_m - \bar{\Phi}^{-1}\left(\frac{\pi^{1/\theta}}{(\log m)^{1/(2\theta)}}\right) \gg 1, \tag{14}$$

then $P(\hat{\pi}^*(\delta) > (1-\epsilon)\pi) \to 1$ for any constant $\epsilon > 0$.

This theorem shows that for bounding function $\delta(t) = [\bar{\Phi}(t)]^\theta$ with $\theta \in [0, 1]$, one can find a bounding sequence $c_m^*(\delta; \alpha) = C'_\alpha \sqrt{\log m}$, whose order does not involve the MAC dependence level $\bar{\rho}_\Sigma$. Consequently, the condition in (14) for signal intensity only depends on the signal sparsity level (π). Results in Theorem 2.3 implies that $\hat{\pi}^*(\delta)$ with $\theta \in [0, 1]$ is consistent under (14), no matter how strong the MAC dependence is.

2.3. A new estimator for dependent variables

For independent variables, [24] shows that when $\delta(t) = [\bar{\Phi}(t)]^{\theta}$, the most powerful estimator in the $\hat{\pi}(\delta)$ class has $\theta = 1/2$, meaning that this single estimator is consistent under the most relaxed conditions compared to the other members in the class. This powerful estimator under independence, however, can be strongly affected by the MAC dependence as shown in Theorem 2.2. On the other hand, Theorem 2.3 indicates that candidates with $\theta \in (1/2, 1]$, which are not as powerful under independence, may outperform candidates with $\theta \in [0, 1/2]$ under strong MAC dependence. However it seems that no single candidate from the class is most powerful under arbitrary covariance dependence. Motivated by the theoretical insight, we propose to construct a new estimator of the form $\hat{\pi}_{adap} = \max\{\hat{\pi}(\delta), \delta \in \Delta\}$, where Δ is the set of $\delta(t)$ functions that render the most powerful estimators in different dependence scenarios.

The power of $\hat{\pi}_{adap}$ depends on the $\delta(t)$ functions in Δ. However, even with the specific form $\delta(t) = [\bar{\Phi}(t)]^{\theta}, \theta \in [0, 1]$, there are infinitely many candidates to choose from. Denote $\hat{\pi}_{\theta}$ as $\hat{\pi}(\delta)$ with $\delta = [\bar{\Phi}(t)]^{\theta}$. In order to find the most promising candidates in this class, we investigate efficiency as well as power limits of all the $\hat{\pi}_{\theta}$ members in two general scenarios classified by the joint effect of MAC dependence and signal sparsity. As a result, we identify two most promising candidates, $\hat{\pi}_{0.5}$ and $\hat{\pi}_1$, and construct the new estimator as

$$\hat{\pi}_{adap} = \max\{\hat{\pi}_{0.5}, \hat{\pi}_1\}. \tag{15}$$

The consistency of $\hat{\pi}_{adap}$ is directly implied by the consistency of $\hat{\pi}_{0.5}$ and $\hat{\pi}_1$ with a degenerating $\alpha = \alpha_m \to 0$. Despite the surprisingly simple form, the new estimator has significant power gain over any member in the $\hat{\pi}_{\theta}$ class under arbitrary MAC dependence. To provide more insight on this, we explicitly study the discretized version $\hat{\pi}^*_{adap} = \max\{\hat{\pi}^*_{0.5}, \hat{\pi}^*_1\}$. The following theorem is based on the monotonicity of $\hat{\pi}^*_{\theta}$ with respect to θ, the sufficient conditions in Theorem 2.2 and Theorem 2.3, and an almost necessary condition in Lemma A.3 for the consistency of $\hat{\pi}^*_{\theta}$ members.

Theorem 2.4 *Consider model (9). Let $\delta(t) = [\bar{\Phi}(t)]^{\theta}$ with $\theta \in [0, 1]$. If*

$$\bar{\Phi}^{-1}\left(\frac{\pi^2}{\bar{\rho}_{\Sigma} \log m}\right) \ll \bar{\Phi}^{-1}\left(\frac{\pi}{\sqrt{\log m}}\right), \tag{16}$$

then $\hat{\pi}^*_{0.5}$ is the most powerful estimator in $\{\hat{\pi}^*_{\theta} : \theta \in [0, 1]\}$, and we have $P(\hat{\pi}^*_{adap} = \hat{\pi}^*_{0.5}) \to 1$. On the other hand, if

$$\bar{\Phi}^{-1}\left(\frac{\pi^2}{\bar{\rho}_{\Sigma} \log m}\right) \gg \bar{\Phi}^{-1}\left(\frac{\pi}{\sqrt{\log m}}\right), \tag{17}$$

then $\hat{\pi}^*_1$ is the most powerful estimator in $\{\hat{\pi}^*_{\theta} : \theta \in [0, 1]\}$, and we have $P(\hat{\pi}^*_{adap} = \hat{\pi}^*_1) \to 1$.

Conditions (16) and (17) are almost complementary to each other and correspond to relatively weak and strong covariance dependence, respectively. For example, when variables are generated from an AR(1) model, $\bar{\rho}_\Sigma = O(m^{-1})$ and condition (16) holds for signals of almost all the sparsity levels. Consider another example where variables are equally correlated, we have $\bar{\rho}_\Sigma = O(1)$ and (17) is easily satisfied. In less extreme cases, the levels of $\bar{\rho}_\Sigma$ and π jointly affect the difficulty of estimation, and the proposed $\hat{\pi}^*_{adap}$ matches the winner in the $\hat{\pi}^*_\theta$ class with high probability. The theoretical analyses in Section 2.2 and Section 2.3, although derived for the discretized version of $\hat{\pi}(\delta)$ and $\hat{\pi}_{adap}$, provide important insights on their properties under dependence, which are supported by extensive simulation studies in Section 3.2.

Numerical Implementation. We provide additional descriptions on the numerical implementation of $\hat{\pi}_{adap}$. Specifically, we need to generate $(W_1, \dots, W_m)$ following the joint null distribution of Z_j. There are application scenarios where the joint null distribution is available. For example, in genome-wide association studies (GWAS), summary statistics for all the genetic variants are obtained based on marginal linear regression. The joint null distribution of the summary statistics follows $N(0, \Sigma)$, where Σ is the sample correlation matrix of the genetic variants that is often observable [3, 11].

In some application scenarios where the joint null distribution is unknown, $(W_1, \dots, W_m)$ may be simulated non-parametrically. For example, when the test statistics are derived from a method of association that does not render an observable null distribution, one can simulate $(W_1, \dots, W_m)$ by randomly shuffling the observations of the response variable to break the relationship between the response and the predictors. More details for such permutation approaches can be found in [28].

Based on the simulated $(W_1, \dots, W_m)$, we follow (5) to simulate $V_{m,0.5}$ and $V_{m,1}$ corresponding to $\theta = 1/2$ and 1. The maximums are taken over $t = w_1, \dots, w_m$. Consequently, we set $\alpha = 0.1$ and generate $c_{m,0.5}$ and $c_{m,1}$ as the $(1-\alpha)$-quantile of N replicates of $V_{m,0.5}$ and $V_{m,1}$, respectively. In our numerical studies, N is set to be 1000. The simulated $c_{m,0.5}$ and $c_{m,1}$ are implemented to calculate $\hat{\pi}_{0.5}$ and $\hat{\pi}_1$ as in (4) by taking the maximums over $t = z_1, \dots z_m$. Finally, the new estimator $\hat{\pi}_{adap}$ is calculated by (15).

2.4. Other existing lower bound estimators

Theorem 2.2–Theorem 2.4 imply that $P(\hat{\pi}^*_{adap} < \pi) \geq 1 - \alpha + o(1)$ under condition (16) or (17), and the two conditions are almost complementary to each other. This result indicates that the new estimator is likely to be an asymptotic lower bound of π at $1 - \alpha$ level. We note that other lower bound estimators can be readily constructed utilizing recent developments in post hoc confidence bounds in multiple testing (see e.g. [2], [22], and [14]). Specifically, for the candidate set that include all the variables, a confidence bound on false positives at $1 - \alpha$ level provides a lower bound of true positives in the full set at $1 - \alpha$

level. Such a lower bound, divided by m, results in a lower bound of the signal proportion at $1-\alpha$ level. Investigating the *consistency* of these lower bound estimators under different dependence structures would be an interesting topic for future research.

In Section 3.3, we compare our proposed estimator with two lower bound estimators, $\hat{\pi}_{SM}$ and $\hat{\pi}_{BL}$, in simulation. $\hat{\pi}_{SM}$ and $\hat{\pi}_{BL}$ are derived from the post hoc confidence bounds in [2] as follows. Given the p-values $p_1,\dots,p_m$ of $X_1,\dots,X_m$, let $\hat{\pi}_{SM}=(m-V_{SM})/m$, where V_{SM} is called the Simes post hoc bound and is defined as

$$V_{SM}=\min_{k\in\{1,\dots,m\}}\left\{\sum_{i=1}^{m}1\{p_i\geq \alpha k/m\}+k-1\right\}.$$

V_{SM} is derived based on the classical Simes inequality. It has been proved to be a level $1-\alpha$ confidence bound for the number of false positives in the full set of variables. Consequently, $\hat{\pi}_{SM}$ is a level $1-\alpha$ lower bound of π, i.e., $P(\hat{\pi}_{SM}<\pi)\geq 1-\alpha$.

The other lower bound estimator $\hat{\pi}_{BL}$ can be derived in a similar way as $\hat{\pi}_{BL}=(m-V_{BL})/m$, where V_{BL} is an improved confidence bound and is defined as

$$V_{BL}=\min_{k\in\{1,\dots,m\}}\left\{\sum_{i=1}^{m}1\{p_i\geq F_k^{-1}(\lambda(\alpha))\}+k-1\right\}.$$

$F_k^{-1}(\lambda(\alpha))$ is call a balanced template and is defined as the $\lambda(\alpha)$-quantile of F_k, where F_k is the cumulative distribution function of $p^0_{(k)}$, and $p^0_{(k)}$ is the kth ordered p-values under the joint null distribution of $p_1,\dots,p_m$. Moreover, $\lambda(\alpha)$ is the α-quantile of $\min_{k\in\{1,\dots,m\}}F_k(p^0_{(k)})$. In our numerical examples, $F_k^{-1}(\lambda(\alpha))$, $k=1,\dots,m$, are simulated based on the empirical distributions of F_k and $\min_{k\in\{1,\dots,m\}}F_k(p^0_{(k)})$, which are obtained from 1000 sets of p-values generated under the joint null distribution.

3. Simulation study

In the following simulation examples, we consider six dependence structures: cases (a)-(d) are commonly observed correlation structures in literature and cases (e)-(f) are real correlation structures from genetic association studies. In all the examples, $\Sigma_{ii}=1, i=1,\dots,m$.

(a) **Autocorrelation**. $\Sigma_{ij}=r^{|i-j|}$ and $r=0.9$.

(b) **Equal correlation**. $\Sigma_{ij}=0.5$ for $i\neq j$.

(c) **Block dependence**. Σ has square diagonal blocks. The off-diagonal elements in the blocks are 0.5, and the elements outside the blocks are zero.

(d) **Sparse dependence**. Σ has nonzero elements randomly located. The data generation process is similar to Model 3 in [5]. Let $\Sigma^*=(\sigma_{ij})$, where $\sigma_{ii}=1$, $\sigma_{ij}=0.9*$ Bernoulli$(1,0.1)$ for $i<j$ and $\sigma_{ji}=\sigma_{ij}$. Then $\Sigma=I^{1/2}(\Sigma^*+\delta I)/(1+\delta)I^{1/2}$, where $\delta=|\lambda_{\min}(\Sigma^*)|+0.05$.

(e) **SNP dependence**. Σ is the sample correlation matrix of the real SNP data on Chromosome 21 from 90 individuals in the International HapMap project.
(f) **Gene dependence**. Σ is the sample correlation matrix of the real gene expression data from 71 individuals in a riboflavin production study.

We generate test statistics $Z_1, \ldots, Z_m \sim N((\mu_1, \ldots, \mu_m), \Sigma)$ and set $m = 2000$ for cases (a)-(d) above. Case (e) has $m = 8657$, which is the number of SNPs in the dataset, and case (f) has $m = 4088$, which is the number of genes in the dataset. Additional details of the datasets can be found in Section 4. The block size in case (c) is set as 400×400. $(\mu_1, \ldots, \mu_m)$ is a sparse vector with randomly located non-zero elements. We consider both relatively sparse signals with $\pi = 0.02$ and more dense signals with $\pi = 0.1$.

3.1. MAC dependence effect on bounding sequences

We first calculate the MAC levels as defined in (10) for the dependence structures in (a)-(f) above and report the values of $c_{m,0.5}$ and $c_{m,1}$, which are generated by the procedure described in Numerical Implementation in Section 2.3 under the joint null distribution $N(0, \Sigma)$. Recall that a larger value of $\bar{\rho}_\Sigma$ indicates stronger overall covariance dependence. It can be seen in Table 1 that $\bar{\rho}_\Sigma$ is fairly small for cases (a) and (d), moderately small for cases (c) and (e), and fairly large for cases (b) and (f). Moreover, $c_{m,0.5}$ seems to vary positively with $\bar{\rho}_\Sigma$, whereas $c_{m,1}$ does not show such tendency. These patterns are demonstrated more clearly in Figure 1, which seem to agree with the theoretical results in Theorem 2.2 and Theorem 2.3 about the MAC dependence effect on $c_{m,0.5}$ and $c_{m,1}$.

TABLE 1
MAC levels and realized values of bounding sequences for different dependence structures.

	(a) Auto	(b) Equal	(c) Block	(d) Sparse	(e) SNP	(f) Gene
$\bar{\rho}_\Sigma$	0.0095	0.5003	0.1003	0.0042	0.0869	0.3353
$c_{m,0.5}$	0.178	0.87	0.397	0.099	0.222	0.706
$c_{m,1}$	8.46	4.39	5.58	6.79	13.6	6.42

3.2. Comparison with members in the $\hat{\pi}(\delta)$ family

We first compare the new estimator $\hat{\pi}_{adap}$ with members in the $\hat{\pi}(\delta)$ family. Since $\hat{\pi}_{0.5}$ and $\hat{\pi}_1$ are likely to outperform other $\hat{\pi}_\theta$ members with $\theta \in [0, 1/2)$ and $\theta \in (1/2, 1)$, respectively, as indicated in Theorem 2.4, we specifically compare $\hat{\pi}_{adap}$ with $\hat{\pi}_{0.5}$ and $\hat{\pi}_1$. Note that $\hat{\pi}_{0.5}$ has been shown in [24] as the most powerful estimator in the $\hat{\pi}_\theta$ class for independent variables. Besides various dependence structures in (a)-(f), we consider sparse and relatively dense signals with $\pi = 0.02$ and 0.1, respectively, and varying signal intensity with non-zero $\mu_j = 3, 4, 5, 6$.

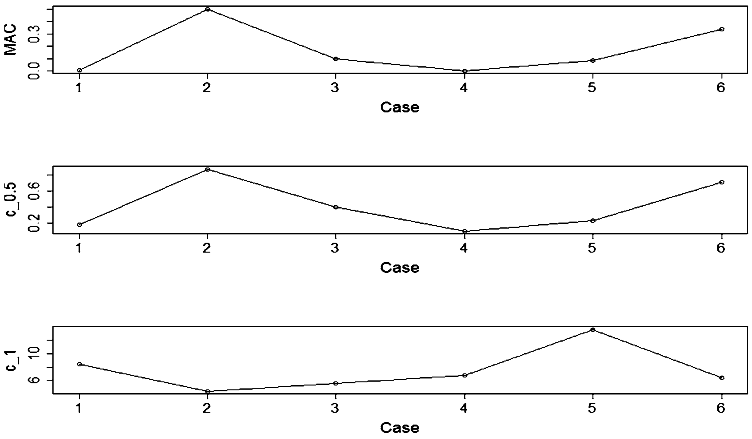

FIG 1. *Trends of MAC levels and bounding sequences for different dependence structures.*

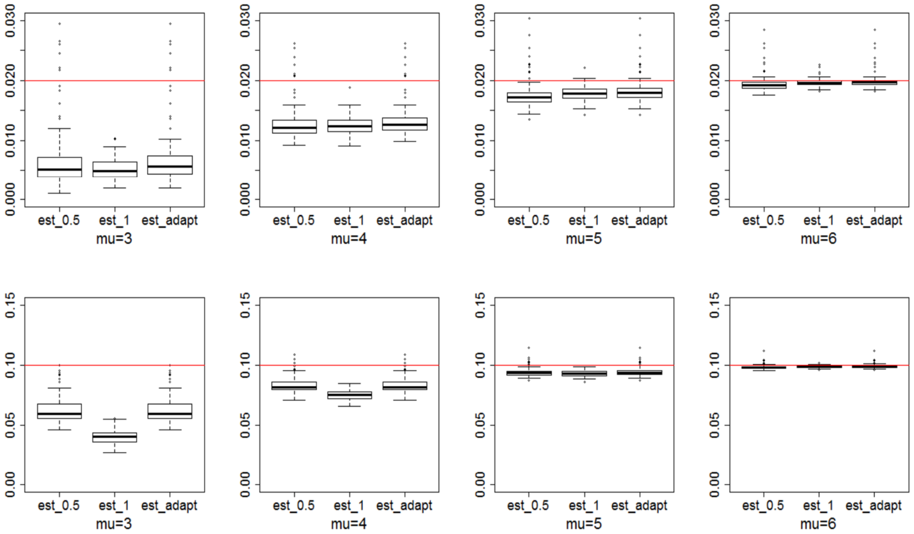

FIG 2. *Comparison under (a) autocorrelation. "est_0.05", "est_1", and "est_adapt" represent $\hat{\pi}_{0.5}$, $\hat{\pi}_1$, and $\hat{\pi}_{adap}$, respectively. The top row has $\pi = 0.02$, and the bottom row has $\pi = 0.1$. The true π values are highlighted by the red horizontal lines.*

Recall the theoretical results in Section 2.3 that $\hat{\pi}_{0.5}$ may outperform $\hat{\pi}_1$ when dependence is weak enough or signals are less sparse, and $\hat{\pi}_1$ may perform better in the other scenarios. We observe such tendencies in Figure 2-Figure 7. Specifically, the autocorrelation case (Figure 2) has small $\bar{\rho}_\Sigma = 0.0095$. It shows that $\hat{\pi}_{0.5}$ has comparable results as those of $\hat{\pi}_1$ for small $\pi = 0.02$, and outperforms $\hat{\pi}_1$ for larger $\pi = 0.1$. The equal correlation case (Figure 3) has the largest $\bar{\rho}_\Sigma = 0.5$. It shows that $\hat{\pi}_1$ outperforms $\hat{\pi}_{0.5}$ for both $\pi = 0.02$ and 0.1.

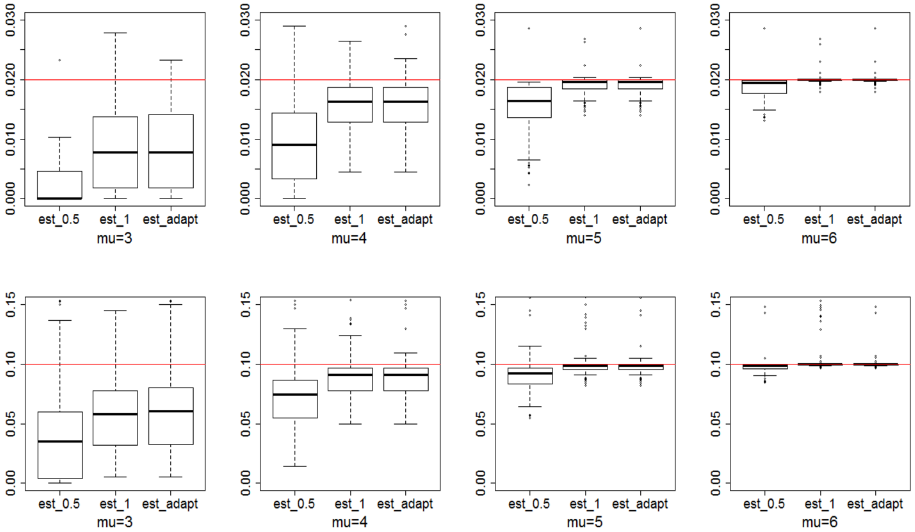

Fig 3. *Comparison under (b) equal correlation. The top row has $\pi = 0.02$, and the bottom row has $\pi = 0.1$. Notations and symbols are the same as in Figure 2.*

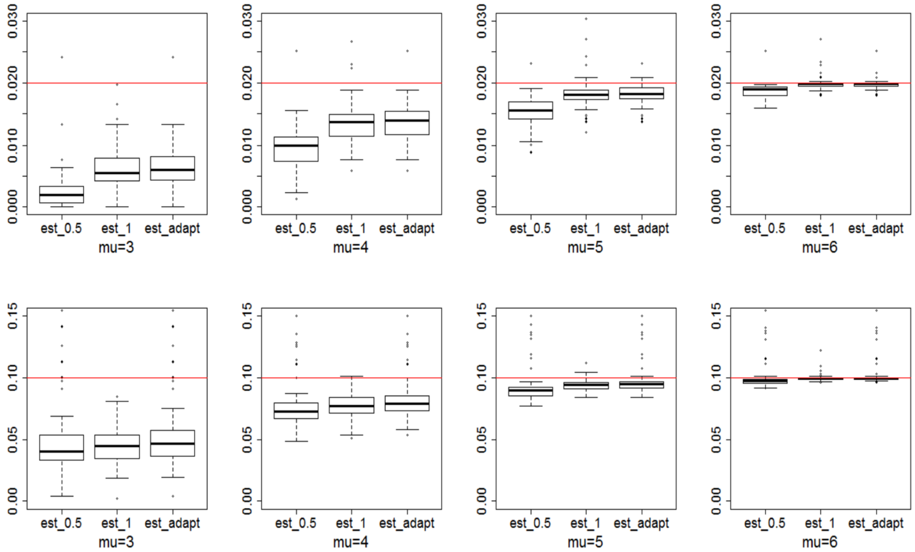

Fig 4. *Comparison under (c) block dependence. The top row has $\pi = 0.02$, and the bottom row has $\pi = 0.1$. Notations and symbols are the same as in Figure 2.*

The block diagonal case (Figure 4) has moderate $\bar{\rho}_\Sigma = 0.1$. It shows that $\hat{\pi}_1$ outperforms $\hat{\pi}_{0.5}$ for small π, and is comparable to $\hat{\pi}_{0.5}$ for larger π. The sparse correlation case (Figure 5) has the smallest $\bar{\rho}_\Sigma = 0.0042$, we see that $\hat{\pi}_{0.5}$ out-

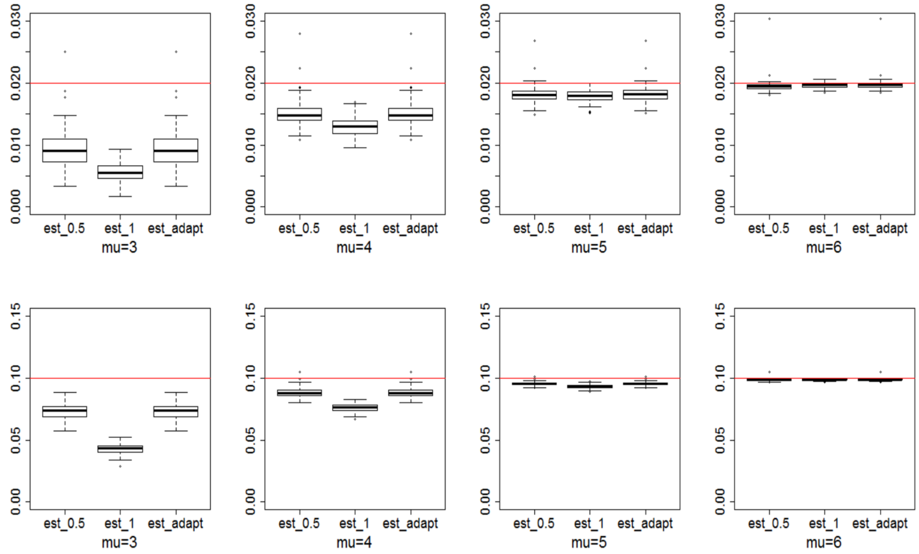

Fig 5. *Comparison under (d) sparse dependence. The top row has* $\pi = 0.02$*, and the bottom row has* $\pi = 0.1$*. Notations and symbols are the same as in Figure 2.*

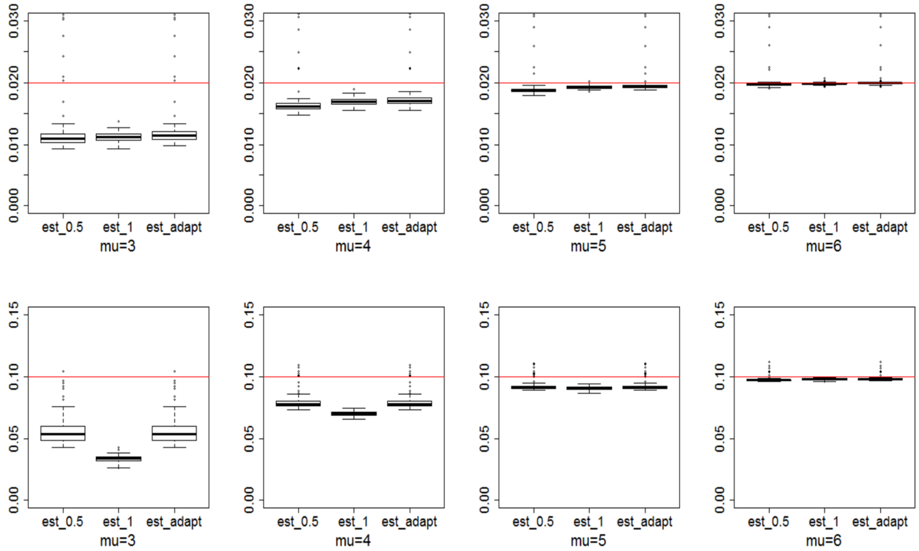

Fig 6. *Comparison under (e) SNP dependence. The top row has* $\pi = 0.02$*, and the bottom row has* $\pi = 0.1$*. Notations and symbols are the same as in Figure 2.*

performs $\hat{\pi}_1$ for both $\pi = 0.02$ and 0.1. The SNP correlation case (Figure 6) has $\bar{\rho}_\Sigma = 0.0869$, which is moderately small. We see that $\hat{\pi}_1$ is slightly better for small π, and $\hat{\pi}_{0.5}$ is better for larger π. The gene correlation case (Figure 7)

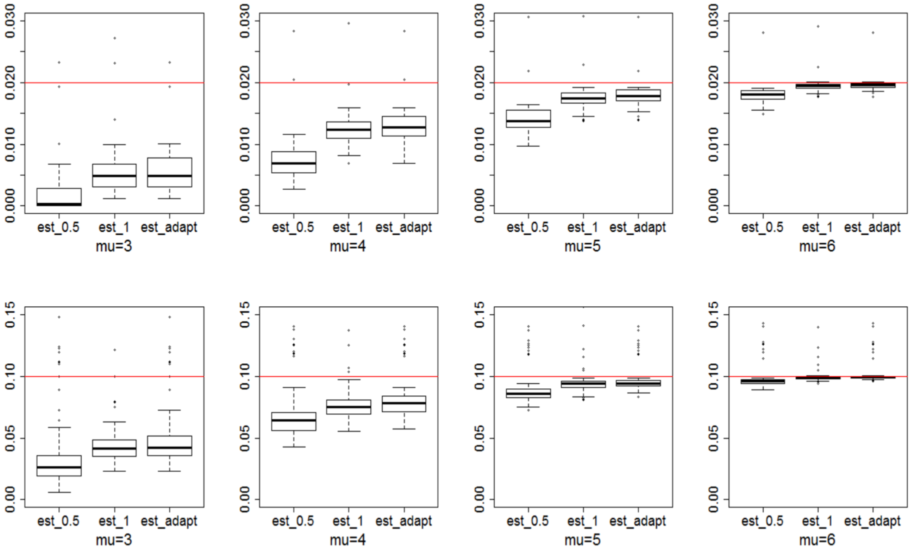

Fig 7. *Comparison under (f) gene dependence. The top row has* $\pi = 0.02$*, and the bottom row has* $\pi = 0.1$*. Notations and symbols are the same as in Figure 2.*

has $\bar{\rho}_\Sigma = 0.3353$, which is fairly large. It shows that $\hat{\pi}_1$ outperforms $\hat{\pi}_{0.5}$ for both small and larger π. In all these examples, the new estimator $\hat{\pi}_{adap}$ always matches the winner of $\hat{\pi}_{0.5}$ and $\hat{\pi}_1$ and exhibits better adaptivity to different dependence structures and signal sparsity levels.

3.3. Comparison with other existing estimators

In this section, we compare the proposed $\hat{\pi}_{adap}$ with two existing consistent estimators, $\hat{\pi}_{GW}$ and $\hat{\pi}_{JC}$, which are developed for relatively dense signals under independence [13, 21], and two lower bound estimators, $\hat{\pi}_{SM}$ and $\hat{\pi}_{BL}$, constructed based on the confidence bounds in multiple testing as described in Section 2.4. We set $\alpha = 0.1$ for $\hat{\pi}_{adap}$, $\hat{\pi}_{SM}$ and $\hat{\pi}_{BL}$. Table 2 and Table 3 report the summarized results of the five estimators from 100 replications.

Table 2 is for the settings with relatively sparse signals ($\pi = 0.02$). It can be seen that for the cases with relatively weak MAC dependence ((a), (d), and (e)), $\hat{\pi}_{adap}$ generally outperforms the other estimators, and the runner-up seems to be $\hat{\pi}_{BL}$. In the cases with moderately strong MAC dependence ((c) and (f)), $\hat{\pi}_{adap}$ outperforms the others when signals are relatively weak ($\mu = 3$ and 4) but not as well as $\hat{\pi}_{SM}$ and $\hat{\pi}_{BL}$ when signals are very strong ($\mu = 6$). In the case with extremely strong MAC dependence ((b)), $\hat{\pi}_{BL}$ performs the best. In this set of examples, $\hat{\pi}_{GW}$ and $\hat{\pi}_{JC}$ are not competitive, which is not surprising because they are not developed for very sparse signals.

Table 3 shows the comparison results for settings with relatively dense signals ($\pi = 0.1$). In the cases with relatively weak MAC dependence ((a), (d), and (e)),

TABLE 2

Mean values and standard deviations (in brackets) of different methods when signals are relatively sparse ($\pi = 0.02$). For each scenario with a fixed dependence structure and a fixed μ value, the mean value(s) that is closest to the true π is highlighted in **bold**.

Dependence	Method	$\mu = 3$	$\mu = 4$	$\mu = 5$	$\mu = 6$
(a) Auto	$\hat{\pi}_{adap}$	0.008 (0.006)	**0.014** (0.004)	**0.019** (0.003)	**0.020** (0.002)
	$\hat{\pi}_{SM}$	0.004 (0.001)	0.011 (0.002)	0.017 (0.001)	**0.020** (0.001)
	$\hat{\pi}_{BL}$	0.005 (0.004)	0.012 (0.003)	0.016 (0.002)	0.019 (0.001)
	$\hat{\pi}_{GW}$	**0.014** (0.024)	**0.014** (0.024)	0.015 (0.024)	0.014 (0.024)
	$\hat{\pi}_{JC}$	0.048 (0.048)	0.049 (0.046)	0.050 (0.045)	0.051 (0.045)
(b) Equal	$\hat{\pi}_{adap}$	0.031 (0.060)	0.039 (0.058)	0.042 (0.057)	0.043 (0.056)
	$\hat{\pi}_{SM}$	0.005 (0.006)	0.012 (0.007)	0.017 (0.004)	**0.020** (0.002)
	$\hat{\pi}_{BL}$	**0.010** (0.023)	**0.017** (0.022)	**0.022** (0.020)	0.024 (0.020)
	$\hat{\pi}_{GW}$	0.193 (0.273)	0.194 (0.272)	0.193 (0.273)	0.194 (0.274)
	$\hat{\pi}_{JC}$	0.295 (0.403)	0.296 (0.402)	0.295 (0.401)	0.297 (0.402)
(c) Block	$\hat{\pi}_{adap}$	**0.012** (0.019)	**0.018** (0.017)	**0.022** (0.016)	0.024 (0.016)
	$\hat{\pi}_{SM}$	0.005 (0.005)	0.012 (0.005)	**0.018** (0.004)	**0.020** (0.004)
	$\hat{\pi}_{BL}$	0.006 (0.012)	0.013 (0.011)	**0.018** (0.010)	0.021 (0.010)
	$\hat{\pi}_{GW}$	0.060 (0.091)	0.060 (0.091)	0.059 (0.091)	0.060 (0.091)
	$\hat{\pi}_{JC}$	0.115 (0.143)	0.115 (0.142)	0.116 (0.141)	0.118 (0.140)
(d) Sparse	$\hat{\pi}_{adap}$	0.009 (0.003)	**0.015** (0.022)	**0.018** (0.001)	**0.020** (0.001)
	$\hat{\pi}_{SM}$	0.004 (0.001)	0.011 (0.002)	0.017 (0.001)	0.019 (0.001)
	$\hat{\pi}_{BL}$	**0.010** (0.003)	**0.015** (0.002)	**0.018** (0.001)	0.019 (0.001)
	$\hat{\pi}_{GW}$	0.003 (0.006)	0.003 (0.007)	0.003 (0.006)	0.003 (0.007)
	$\hat{\pi}_{JC}$	0.030 (0.021)	0.031 (0.020)	0.032 (0.020)	0.033 (0.020)
(e) SNP	$\hat{\pi}_{adap}$	**0.006** (0.006)	**0.013** (0.005)	**0.018** (0.003)	**0.020** (0.003)
	$\hat{\pi}_{SM}$	0.003 (0.001)	0.011 (0.001)	0.017 (0.001)	0.019 (0.001)
	$\hat{\pi}_{BL}$	0.004 (0.002)	0.012 (0.001)	0.017 (0.001)	0.019 (0.001)
	$\hat{\pi}_{GW}$	0.034 (0.040)	0.034 (0.039)	0.035 (0.039)	0.035 (0.039)
	$\hat{\pi}_{JC}$	0.061 (0.069)	0.063 (0.069)	0.064 (0.068)	0.064 (0.068)
(f) Gene	$\hat{\pi}_{adap}$	**0.012** (0.031)	**0.020** (0.030)	0.025 (0.029)	0.026 (0.029)
	$\hat{\pi}_{SM}$	0.003 (0.002)	0.011 (0.002)	**0.017** (0.001)	**0.019** (0.001)
	$\hat{\pi}_{BL}$	0.004 (0.004)	0.012 (0.003)	**0.017** (0.002)	**0.019** (0.002)
	$\hat{\pi}_{GW}$	0.088 (0.145)	0.089 (0.145)	0.089 (0.146)	0.089 (0.145)
	$\hat{\pi}_{JC}$	0.151 (0.232)	0.151 (0.230)	0.152 (0.230)	0.154 (0.230)

all the methods perform quite well. In the cases with moderately strong MAC dependence ((c) and (f)), $\hat{\pi}_{adap}$, $\hat{\pi}_{SM}$ and $\hat{\pi}_{BL}$ outperform $\hat{\pi}_{GW}$ and $\hat{\pi}_{JC}$. In the case with extremely strong MAC dependence ((b)), $\hat{\pi}_{adap}$ outperforms the others when signals are relatively weak ($\mu = 3$ and 4) and performs comparably to $\hat{\pi}_{SM}$ and $\hat{\pi}_{BL}$ when signals are strong ($\mu = 5$ and 6). Overall, the performance of $\hat{\pi}_{adap}$ appears to be competitive and most stable across various dependence structures and levels of sparsity.

4. Real application

We apply the proposed method to two real datasets. The first dataset is from an eQTL study with the goal to identify SNPs that potentially govern the expres-

Table 3
Mean values and standard deviations (in brackets) of different methods when signals are relatively dense ($\pi = 0.1$). The mean value(s) that is closest to the true π for each scenario is highlighted in **bold**.

Dependence	Method	$\mu = 3$	$\mu = 4$	$\mu = 5$	$\mu = 6$
(a) Auto	$\hat{\pi}_{adap}$	0.06 (0.01)	0.08 (0.01)	**0.09** (0.00)	**0.10** (0.00)
	$\hat{\pi}_{SM}$	0.03 (0.01)	0.07 (0.00)	**0.09** (0.00)	**0.10** (0.00)
	$\hat{\pi}_{BL}$	0.06 (0.01)	0.08 (0.01)	**0.09** (0.00)	**0.10** (0.00)
	$\hat{\pi}_{GW}$	**0.08** (0.03)	**0.09** (0.03)	0.09 (0.02)	0.09 (0.02)
	$\hat{\pi}_{JC}$	0.12 (0.05)	0.12 (0.04)	0.13 (0.04)	0.13 (0.04)
(b) Equal	$\hat{\pi}_{adap}$	**0.07** (0.05)	**0.10** (0.04)	0.11 (0.04)	0.11 (0.04)
	$\hat{\pi}_{SM}$	0.03 (0.03)	0.07 (0.02)	0.09 (0.01)	**0.10** (0.00)
	$\hat{\pi}_{BL}$	0.04 (0.03)	0.08 (0.02)	**0.10** (0.01)	**0.10** (0.01)
	$\hat{\pi}_{GW}$	0.23 (0.26)	0.25 (0.26)	0.25 (0.25)	0.25 (0.25)
	$\hat{\pi}_{JC}$	0.33 (0.39)	0.33 (0.37)	0.34 (0.37)	0.34 (0.37)
(c) Block	$\hat{\pi}_{adap}$	0.05 (0.03)	**0.08** (0.02)	**0.10** (0.01)	**0.10** (0.01)
	$\hat{\pi}_{SM}$	0.03 (0.01)	0.07 (0.01)	0.09 (0.01)	0.10 (0.00)
	$\hat{\pi}_{BL}$	0.04 (0.02)	0.07 (0.01)	0.09 (0.01)	0.10 (0.01)
	$\hat{\pi}_{GW}$	**0.11** (0.09)	0.12 (0.08)	0.12 (0.08)	0.12 (0.08)
	$\hat{\pi}_{JC}$	0.17 (0.14)	0.17 (0.13)	0.18 (0.13)	0.18 (0.13)
(d) Sparse	$\hat{\pi}_{adap}$	0.07 (0.01)	**0.09** (0.00)	**0.10** (0.00)	**0.10** (0.00)
	$\hat{\pi}_{SM}$	0.03 (0.01)	0.07 (0.00)	0.09 (0.00)	**0.10** (0.00)
	$\hat{\pi}_{BL}$	0.07 (0.01)	**0.09** (0.00)	0.09 (0.00)	**0.10** (0.00)
	$\hat{\pi}_{GW}$	0.07 (0.01)	0.07 (0.01)	0.08 (0.01)	0.08 (0.01)
	$\hat{\pi}_{JC}$	**0.10** (0.02)	0.11 (0.01)	0.11 (0.02)	0.11 (0.02)
(e) SNP	$\hat{\pi}_{adap}$	0.06 (0.01)	0.08 (0.01)	**0.09** (0.00)	**0.10** (0.00)
	$\hat{\pi}_{SM}$	0.03 (0.01)	0.07 (0.00)	**0.09** (0.00)	**0.10** (0.00)
	$\hat{\pi}_{BL}$	0.05 (0.01)	0.07 (0.00)	**0.09** (0.00)	**0.10** (0.00)
	$\hat{\pi}_{GW}$	**0.10** (0.04)	**0.11** (0.04)	0.11 (0.03)	0.11 (0.03)
	$\hat{\pi}_{JC}$	0.13 (0.07)	0.13 (0.07)	0.14 (0.06)	0.14 (0.06)
(f) Gene	$\hat{\pi}_{adap}$	**0.05** (0.03)	**0.08** (0.03)	**0.10** (0.02)	**0.10** (0.02)
	$\hat{\pi}_{SM}$	0.03 (0.01)	0.07 (0.01)	0.09 (0.00)	**0.10** (0.00)
	$\hat{\pi}_{BL}$	0.04 (0.01)	0.07 (0.01)	0.09 (0.01)	**0.10** (0.00)
	$\hat{\pi}_{GW}$	0.15 (0.14)	0.15 (0.13)	0.16 (0.13)	0.16 (0.13)
	$\hat{\pi}_{JC}$	0.20 (0.23)	0.20 (0.22)	0.21 (0.22)	0.22 (0.21)

sion of gene CCT8 on chromosome 21. This gene has been found to be relevant to Down Syndrome [3, 11]. We obtain the SNP data and the gene expression data of 90 unaffected subjects from the Asian population in the International HapMap project (45 Japanese in Tokyo, Japan, and 45 Han Chinese in Beijing; http://zzz.bwh.harvard.edu/plink/res.shtml#hapmap). After removing SNPs with missing values, we have 8657 candidate SNPs left.

Test statistics for the associations between each SNP and the expression level of CCT8 are derived by fitting marginal linear regressions as in [3] and [11]. Histogram of the test statistics is presented in the left panel of Figure 8, where the long and thin right tail indicates possibly a small number of signals with positive signal effects. The correlation matrix of the test statistics, which is the same as the correlation matrix of the SNPs, has the MAC level of $\bar{\rho}_{\Sigma} = 0.087$.

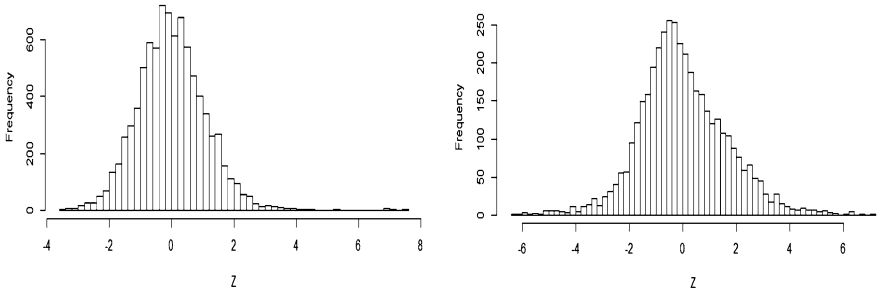

Fig 8. *Histograms in two application examples. The left panel is for the test statistics in an eQTL analysis. The right panel is for the test statistics in a gene expression association analysis.*

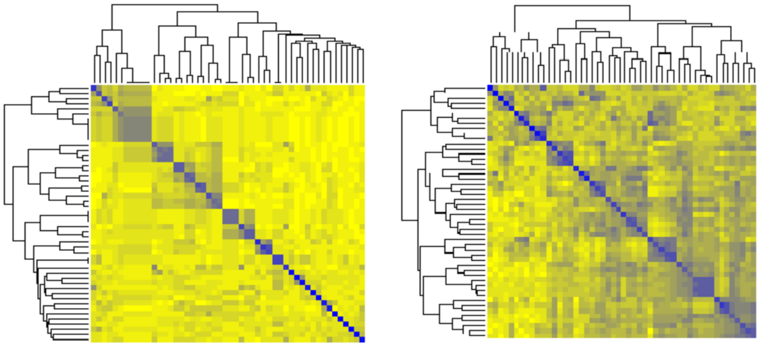

Fig 9. *Heatmaps in two application examples. The left panel shows the absolute value of correlations for 50 SNPs in the Hapmap data. The right panel shows the absolute value of correlations for 50 gene expressions in the riboflavin production study.*

Heatmap of the correlation matrix of the first 50 SNPs is illustrated in the left panel of Figure 9.

We apply the proposed estimator $\hat{\pi}_{adap}$, two lower bound estimators ($\hat{\pi}_{SM}$ and $\hat{\pi}_{BL}$) and two existing consistent estimators ($\hat{\pi}_{GW}$ and $\hat{\pi}_{JC}$) to the dataset. The estimated proportion values for the eQTL study are presented in Table 4, where it shows that $\hat{\pi}_{adap}$ is larger than $\hat{\pi}_{SM}$ and $\hat{\pi}_{BL}$, but smaller than $\hat{\pi}_{GW}$ and $\hat{\pi}_{JC}$. Note that this SNP correlation structure has been investigated in simulation studies. Results of this real application seem to be consistent with the findings in simulation, see, e.g. case (e) of Table 2, where it shows that $\hat{\pi}_{adap}$ could be closer to the true π under the SNP dependence structure when signals are very sparse.

The second real application example has microarray data from a study on riboflavin (vitamin B_2) production in bacillus subtilis. This dataset is available at https://www.annualreviews.org/doi/suppl/10.1146/annurev-statistics-022513-115545 and has been studied in [4]. The dataset includes the expression levels of 4088 genes and the logarithm of riboflavin production rate of 71

Table 4
Results of the real applications in eQTL study and gene expression association analysis.

	$\bar{\rho}_\Sigma$	$\hat{\pi}_{adap}$	$\hat{\pi}_{SM}$	$\hat{\pi}_{BL}$	$\hat{\pi}_{GW}$	$\hat{\pi}_{JC}$
eQTL	0.087	0.0016	0.0009	0.0009	0.0068	0.0186
Microarray	0.335	0.0638	0.0355	0.0443	0.2575	0.3742

individuals. Marginal regression coefficients are used as test statistics for associations between genes and riboflavin production. Histogram of the test statistics is presented in the right panel of Figure 8, which suggests possibly a larger signal proportion than in the eQTL data example. The right panel of Figure 9 shows the heatmap of the correlation matrix of the first 50 genes, which indicates a more complicated dependence structure. The MAC level of the genes is $\bar{\rho}_\Sigma = 0.335$, which is fairly large.

For this example, the estimated proportion values are reported in the bottom row of Table 4, where it shows that $\hat{\pi}_{adap}$ is larger than $\hat{\pi}_{SM}$ and $\hat{\pi}_{BL}$, yet smaller than $\hat{\pi}_{GW}$ and $\hat{\pi}_{JC}$. Recall that this gene dependence structure has been investigated in simulation studies. Results of this application example seem to be consistent with the simulation results reported in case (f) of Table 3, where it shows that the new estimator $\hat{\pi}_{adap}$ could be more accurate under the gene dependence structure when signals are relatively dense.

5. Conclusion and discussion

Estimating the proportion of information-bearing signals is notoriously difficult when dealing with large-scale datasets with complex dependence structures. In this paper, we quantify arbitrary covariance dependence by the MAC level and study the MAC dependence effect on a family of estimators that was originally developed under independence.

Different from the analysis under independence, key components of the estimators do not have readily applicable limiting distributions under arbitrary dependence. We develop new concentration inequalities to explicate the joint effects of MAC dependence, signal sparsity and signal intensity. Different from the previous conclusion that there exists a single member in the family that is most powerful for signals of different sparsity levels under independence, we find that no single estimator in the family is most powerful under different MAC dependence levels. We identify candidate estimators that are most powerful in different MAC dependence scenarios and develop a new estimator $\hat{\pi}_{adap}$ that better adapts to arbitrary covariance dependence.

The new estimator inherits the lower bound property of the family and provides a conservative estimate under very general conditions. This property is valuable in real applications as it requires no conditions on the unknown signals. Moreover, the new estimator is more powerful than any member in the family and compares favorably to other existing methods in extensive numerical examples considering weak to strong covariance dependence and real dependence structures from genetic associations studies. As the estimation problem is frequently met in real applications with complex data structures and the challenge to address arbitrary covariance dependence is at the frontier of high-dimensional sparse inference, we expect the impact of the proposed research to be far-reaching.

Appendix

This section presents proofs of the theoretical results appeared in Section 2. The symbol C denotes a genetic, finite constant whose values can be different at different occurrences.

A.1. Proof of Theorem 2.1

We first show $P(\hat{\pi}(\delta) < \pi) \geq 1 - \alpha$. Let $Z_j^0 = Z_j$ for $j \in I_0$ and $Z_j^1 = Z_j$ for $j \in I_1$. Denote $m_0 = |I_0|$ and $s = |I_1|$. Then

$$\begin{aligned} \bar{F}_m(t) &= m^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} + m^{-1} \sum_{j \in I_1} 1_{\{|Z_j^1| > t\}} \leq m^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} + m^{-1} s \\ &= (1 - \pi) m_0^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} + \pi. \end{aligned}$$

Consequently,

$$\begin{aligned} & P(\hat{\pi}_\delta > \pi) \\ \leq\, & P\left(\sup_{t>0} \left\{ \frac{(1-\pi)\left(m_0^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0|>t\}} - 2\bar{\Phi}(t)\right) + \pi\left(1 - 2\bar{\Phi}(t)\right) - c_m(\delta;\alpha)\delta(t)}{1 - 2\bar{\Phi}(t)} \right\} > \pi \right) \\ \leq\, & P\left(\sup_{t>0} \left\{ (1-\pi)\left(m_0^{-1} \textstyle\sum_{j \in I_0} 1_{\{|Z_j^0|>t\}} - 2\bar{\Phi}(t)\right) - c_m(\delta;\alpha)\delta(t) \right\} > 0 \right) \\ \leq\, & P\left(\sup_{t>0} \left\{ m_0^{-1} \textstyle\sum_{j \in I_0} 1_{\{|Z_j^0|>t\}} - 2\bar{\Phi}(t) - c_{m_0}(\delta;\alpha)\delta(t) \right\} > 0 \right) \\ \leq\, & P\left(V_{m_0} > c_{m_0,\delta}\right) = \alpha, \end{aligned}$$

where the second and last inequalities are by properties (a) and (b) of $c_m(\delta;\alpha)$, respectively. The claim $P(\hat{\pi}(\delta) < \pi) \geq 1 - \alpha$ follows.

Next, we show $P(\hat{\pi}(\delta) > (1 - \epsilon)\pi) \to 1$. Because $\hat{\pi}_\delta > \bar{F}_m(t) - 2\bar{\Phi}(t) - c_m(\delta;\alpha)\delta\,(t)$ for any $t > 0$ and

$$\bar{F}_m(t) = \frac{1 - \pi}{m_0} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} + \frac{\pi}{s} \sum_{j \in I_1} 1_{\{|Z_j^1| > t\}},$$

then

$$\frac{\hat{\pi}_\delta}{\pi} - 1 \quad > \quad -\frac{1}{\pi} c_m(\delta;\alpha)\delta(t) + \frac{1 - \pi}{\pi} \left(m_0^{-1} \textstyle\sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} - 2\bar{\Phi}(t) \right)$$

$$+ \quad \left(s^{-1}\sum\nolimits_{j\in I_1} 1_{\{|Z_j^1|>t\}} - 1\right) - 2\bar{\Phi}(t) \tag{18}$$

for any $t > 0$. Now, set t in (18) at τ_m such that $\tau_m \gg 1$ and $\delta(\tau_m) \ll \pi/c_m(\delta;\alpha)$. We will show that each term on the right hand side of (18) at $t = \tau_m$ is of $o_p(1)$.

First, by the condition $\delta(\tau_m) \ll \pi/c_m(\delta;\alpha)$, we have the first term $A_1 = -c_m(\delta;\alpha)\delta(\tau_m)/\pi = o(1)$.

Consider the second term $A_2 = \pi^{-1}(1-\pi)\left(m_0^{-1}\sum_{j\in I_0} 1_{\{|Z_j^0|>\tau_m\}} - 2\bar{\Phi}(\tau_m)\right)$ in (18). The following lemma is proved in Section A.2. Therefore $A_2 = o_p(1)$.

Lemma A.1 *For any τ_m such that $\tau_m \gg 1$ and $\delta(\tau_m) \ll \pi/c_m(\delta;\alpha)$, we have*

$$\pi^{-1}\left(m_0^{-1}\sum\nolimits_{j\in I_0} 1_{\{|Z_j^0|>\tau_m\}} - 2\bar{\Phi}(\tau_m)\right) = o_p(1).$$

For the third term $A_3 = s^{-1}\sum_{j\in I_1} 1_{\{|Z_j^1|>\tau_m\}} - 1$ in (18), we have

$$\begin{aligned} P(|A_3| > a) &= P(1 - s^{-1}\sum\nolimits_{j\in I_1} 1_{\{|Z_j^1|>\tau_m\}} > a) \\ &\le a^{-1}(1 - P(|Z_j^1| > \tau_m)) \\ &= a^{-1}\left(G(\tau_m) - G(-\tau_m)\right) = o(1) \end{aligned}$$

for any fixed $a > 0$, where the third step is by $Z_j^1 \sim G$ for $j \in I_1$, and the last the step is by the condition $G(\tau_m) \to 0$ or $G(-\tau_m) \to 1$.

Last but not least, the forth term in (18): $A_4 = -2\bar{\Phi}(\tau_m) = o(1)$ given $\tau_m \gg 1$.

Summarizing the above gives the desired result $P(\hat{\pi}(\delta)/\pi > 1 - \epsilon) \to 1$.

A.2. Proof of Lemma A.1

Recall the definitions of $V_m(\delta)$ in (5) and

$$P\left(\sup_{t>0} \frac{|m^{-1}\sum_{j=1}^m 1_{\{|W_j|>t\}} - 2\bar{\Phi}(t)|}{\delta\,(t)} > c_m(\delta;\alpha)\right) < \alpha.$$

where $W_1, \ldots, W_m \sim N_m(0, \Sigma)$. Then, for $t = \tau_m$,

$$P\left(\frac{|m^{-1}\sum_{j=1}^m 1_{\{|W_j|>\tau_m\}} - 2\bar{\Phi}(\tau_m)|}{\pi} > \frac{c_m(\delta;\alpha)\delta(\tau_m)}{\pi}\right) < \alpha.$$

Given $c_{m,\delta}\delta(\tau_m)/\pi = o(1)$, we have

$$\frac{|m^{-1}\sum_{j=1}^m 1_{\{|W_j|>\tau_m\}} - 2\bar{\Phi}(\tau_m)|}{\pi} = o_p(1). \tag{19}$$

Decompose the left hand side above as

$$\pi^{-1}\left|m^{-1}\sum\nolimits_{j=1}^m 1_{\{|W_j|>\tau_m\}} - 2\bar{\Phi}(\tau_m)\right|$$

$$
\begin{aligned}
\geq\ & \pi^{-1}\left|m_0^{-1}\sum\nolimits_{j\in I_0} 1_{\{|W_j|>\tau_m\}} - 2\bar{\Phi}(\tau_m)\right| \\
& -\pi^{-1}\left|m^{-1}\sum_{j=1}^{m} 1_{\{|W_j|>\tau_m\}} - m_0^{-1}\sum\nolimits_{j\in I_0} 1_{\{|W_j|>\tau_m\}}\right|
\end{aligned}
$$

For the second term on the right hand side,

$$
\begin{aligned}
& \pi^{-1}\left|m^{-1}\sum\nolimits_{j=1}^{m} 1_{\{|W_j|>\tau_m\}} - m_0^{-1}\sum\nolimits_{j\in I_0} 1_{\{|W_j|>\tau_m\}}\right| \\
=\ & \pi^{-1}\left|m^{-1}\sum\nolimits_{j\in I_1} 1_{\{|W_j|>\tau_m\}} - \pi m_0^{-1}\sum\nolimits_{j\in I_0} 1_{\{|W_j|>\tau_m\}}\right| \\
\leq\ & s^{-1}\sum\nolimits_{j\in I_1} 1_{\{|W_j|>\tau_m\}} + m_0^{-1}\sum\nolimits_{j\in I_0} 1_{\{|W_j|>\tau_m\}}. \qquad (20)
\end{aligned}
$$

Since $\tau_m \gg 1$, both $s^{-1}\sum_{j\in I_1} 1_{\{|W_j|>\tau_m\}} = o_p(1)$ and $m_0^{-1}\sum_{j\in I_0} 1_{\{|W_j|>\tau_m\}} = o_p(1)$ by Markov's inequality. Combining this with (19) and (20) implies that the first term on the right hand side of (20) satisfies

$$
\pi^{-1}\left|m_0^{-1}\sum\nolimits_{j\in I_0} 1_{\{|W_j|>\tau_m\}} - 2\bar{\Phi}(\tau_m)\right| = o_p(1).
$$

Now, because the joint distribution of $Z_j^0, j \in I_0$, is the same as the joint distribution of $W_j, j \in I_0$, claim in Lemma A.1 follows.

A.3. Proof of Theorem 2.2

First, we show that the sequence $c_m^*(\delta;\alpha) = C_\alpha\sqrt{\bar{\rho}_\Sigma(\log m)^{\theta+1/2}}$, with a large enough constant C_α, satisfies properties (a) $mc_m^*(\delta;\alpha) > m_0 c_{m_0,\delta}^*$ and (b) $P(V_m^*(\delta) > c_m^*(\delta;\alpha)) < \alpha$ for all m.

Consider property (a). Define Σ_0 as the covariance matrix of $W_j, j \in I_0$ and

$$
\bar{\rho}_{\Sigma_0} = \sum_{i\in I_0}\sum_{j\in I_0} |\Sigma_{ij}|/m_0^2.
$$

It can be shown that

$$
\bar{\rho}_\Sigma > \frac{1}{m^2}\sum_{i\in I_0}\sum_{j\in I_0}|\Sigma_{ij}| = \frac{(1-\pi)^2}{m_0^2}\sum_{i\in I_0}\sum_{j\in I_0}|\Sigma_{ij}| = (1-\pi)^2\bar{\rho}_{\Sigma_0}.
$$

Then it follows that

$$
c_m^*(\delta;\alpha) > C(1-\pi)\sqrt{\bar{\rho}_{\Sigma_0}(\log m)^{\theta+1/2}} > (1-\pi)c_{m_0,\delta}^* = (m_0/m)c_{m_0,\delta}^*,
$$

and property (a) is verified.

Next consider property (b). By Chebyshev's inequality and direct calculation,

$$
P(V_m(\delta)^* > c_m^*(\delta;\alpha)) \quad \leq \quad (c_m^*(\delta;\alpha))^{-2}\,\mathsf{Var}(V_m(\delta)^*)
$$

$$
\begin{aligned}
&\le (c_m^*(\delta;\alpha))^{-2}\, E([V_m(\delta)^*]^2)\\
&= (c_m^*(\delta;\alpha))^{-2}\, E\left[\max_{t\in\mathbb{T}}\left(\frac{|\bar{W}_m(t)-2\bar{\Phi}(t)|}{[\bar{\Phi}(t)]^\theta}\right)^2\right]
\end{aligned}
$$

Let $A(t) = [\bar{\Phi}(t)]^{-2\theta}(\bar{W}_m(t) - 2\bar{\Phi}(t))^2$. It can be shown that

$$
\begin{aligned}
E\left(\max_{t\in\mathbb{T}} A(t)\right) &= \int_0^\infty P(\max_{t\in\mathbb{T}} A(t) > c)dc \le \int_0^\infty \sum_{t\in\mathbb{T}} P(A(t) > c)dc\\
&= \sum_{t\in\mathbb{T}} E[A(t)] \le C\sqrt{\log m}\cdot \max_{t\in\mathbb{T}} E[A(t)]\\
&= C\sqrt{\log m}\cdot \max_{t\in\mathbb{T}}\left\{[\bar{\Phi}(t)]^{-2\theta}\mathsf{Var}\left(\bar{W}_m(t)\right)\right\}
\end{aligned}
$$

The following lemma provides the order of $\mathsf{Var}(\bar{W}_m(t))$.

Lemma A.2 *For $W_1, \dots, W_m \sim N_m(0,\Sigma)$ and $\bar{\rho}_\Sigma$ in (10),*

$$
\mathsf{Var}\left(\bar{W}_m(t)\right) = O\left(\bar{\rho}_\Sigma \cdot e^{-t^2/2}\right). \tag{21}
$$

Therefore,

$$
\begin{aligned}
[\bar{\Phi}(t)]^{-2\theta}\cdot \mathsf{Var}\left(\bar{W}_m(t)\right) &\le C[\bar{\Phi}(t)]^{-2\theta}\cdot \bar{\rho}_\Sigma \cdot e^{-t^2/2}\\
&\le C\left(\frac{t}{e^{-t^2/2}}\right)^{2\theta}\cdot \bar{\rho}_\Sigma\cdot e^{-t^2/2}\\
&\le C(\log m)^\theta \cdot \bar{\rho}_\Sigma \cdot e^{(\theta-1/2)t^2}\\
&\le C\bar{\rho}_\Sigma\cdot (\log m)^\theta
\end{aligned}
$$

where the first step above is by Lemma A.2, the second step is by Mill's ratio, the third step is by $t \in \mathbb{T}$, and the last step is by $\theta \in [0, 1/2]$. Combining the above, we have

$$
P(V_m(\delta)^* > c_m^*(\delta;\alpha)) \le C\,(c_m^*(\delta;\alpha))^{-2}\cdot \bar{\rho}_\Sigma \cdot (\log m)^{\theta+1/2},
$$

and the above is bounded by α if $c_m^*(\delta;\alpha) = C_\alpha\sqrt{\bar{\rho}_\Sigma(\log m)^{\theta+1/2}}$ for some large enough constant C_α.

Next, we demonstrate the upper bound property of $\hat{\pi}(\delta)^*$. Denote

$$
B_m = \bar{\Phi}^{-1}\left(\frac{\pi^{1/\theta}}{\bar{\rho}_\Sigma^{1/(2\theta)}(\log m)^{(\theta+1/2)/(2\theta)}}\right).
$$

Let $\tau_m = (A_m + B_m)/2$. Then, $A_m \gg 1$ and condition (13) imply that $\tau_m \gg 1$ and $\tau_m - B_m \to \infty$, which further imply

$$
[\bar{\Phi}(\tau_m)]^\theta \ll \pi/c_m^*(\delta;\alpha).
$$

On the other hand,

$$
G_m(\tau_m) = \Phi(\tau_m - A_m) = \Phi(-(A_m - B_m)/2) = o(1),
$$

where the last step is by condition (13). The rest is straightforward by applying Theorem 2.1.

A.4. Proof of Lemma A.2

$$\mathsf{Var}\left(\bar{W}_m(t)\right) = m^{-2}\sum_{j=1}^{m} Var(1_{\{|W_j|>t\}}) + m^{-2}\sum_{i\neq j} Cov(1_{\{|W_i|>t\}}, 1_{\{|W_j|>t\}}).$$

By Mill's ratio,

$$m^{-2}\sum_{j=1}^{m} Var(1_{\{|W_j|>t\}}) \leq m^{-1}2\bar{\Phi}(t)(1-2\bar{\Phi}(t)) \leq Cm^{-1}e^{-t^2/2}.$$

For $m^{-2}\sum_{i\neq j} Cov(1_{\{|W_i|>t\}}, 1_{\{|W_j|>t\}})$, we have

$$\begin{aligned} Cov(1_{\{|W_i|>t\}}, 1_{\{|W_j|>t\}}) &= 4\int_{-\infty}^{t}\int_{-\infty}^{t} f(x,y)dxdy - 4\int_{-\infty}^{t}\phi(x)dx\int_{-\infty}^{t}\phi(y)dy \\ &\leq C|\Sigma_{ij}|e^{-t^2/2}, \end{aligned}$$

where the last step follows from Corollary 2.1 in [23]. Combining the above with the definition of $\bar{\rho}_\Sigma$ results in (21).

A.5. Proof of Theorem 2.3

First, it is easy to see that $c_m^*(\delta;\alpha) = C'_\alpha\sqrt{\log m}$ satisfies property (a) $mc_m(\delta;\alpha) > m_0c_{m_0}(\delta;\alpha)$.

For property (b), by Markov's inequality,

$$P(V_m(\delta)^* > c_m^*(\delta;\alpha)) \leq (c_m^*(\delta;\alpha))^{-1}\,\mathsf{E}\left(\max_{t\in\mathbb{T}}\frac{|\bar{W}_m(t) - 2\bar{\Phi}(t)|}{[\bar{\Phi}(t)]^\theta}\right).$$

Let $B(t) = [\bar{\Phi}(t)]^{-\theta}|\bar{W}_m(t) - 2\bar{\Phi}(t)|$, and by the similar arguments as in Section A.3, we have

$$\mathsf{E}[\max_{t\in\mathbb{T}} B(t)] \leq C\sqrt{\log m}\cdot\max_{t\in\mathbb{T}}\mathsf{E}[B(t)].$$

Further, $\mathsf{E}[B(t)] \leq [\bar{\Phi}(t)]^{-\theta}(\mathsf{E}[\bar{W}_m(t)] + 2\bar{\Phi}(t)) = 4[\bar{\Phi}(t)]^{1-\theta} \leq 4$ for $\theta \in (1/2, 1]$.

Summing up the above, we have

$$P(V_m(\delta)^* > c_m^*(\delta;\alpha)) \leq C\left(c_m^*(\delta;\alpha)\right)^{-1}\sqrt{\log m} < \alpha,$$

where the last step is by $c_m^*(\delta;\alpha) = C'_\alpha\sqrt{\log m}$ with a large enough constant C'_α.

Next, we demonstrate the upper bound property of $\hat{\pi}(\delta)^*$ with $\delta(t) = [\bar{\Phi}(t)]^\theta$, $\theta \in (1/2, 1]$. Similar arguments as in the proof of Theorem 2.2 for the upper bound can be applied with condition (13) replaced by condition (14). We omit the details to save space.

A.6. Proof of Theorem 2.4

Recall the definition of $\hat{\pi}^*_{adap}$. It is enough to show the following four claims.
Claim 1: $\hat{\pi}^*_{0.5}$ is more powerful than any $\hat{\pi}^*_\theta$ with $\theta \in [0, 1/2)$.
Claim 2: $\hat{\pi}^*_1$ is more powerful than any $\hat{\pi}^*_\theta$ with $\theta \in (1/2, 1)$.
Claim 3: Under condition (16), $\hat{\pi}^*_{0.5}$ is more powerful than $\hat{\pi}^*_1$.
Claim 4: Under condition (17), $\hat{\pi}^*_1$ is more powerful than $\hat{\pi}^*_{0.5}$.

Denote c^*_θ as the bounding sequence $c^*_m(\delta; \alpha)$ with $\delta(t) = [\bar{\Phi}(t)]^\theta$. It can seen that $c^*_\theta[\bar{\Phi}(t)]^\theta$ is strictly decreasing with respect to $\theta \in [0, 1/2]$ for $t \in \mathbb{T}$ as shown in Theorem 2.2. Then $\hat{\pi}^*_\theta$ is strictly increasing with respect to $\theta \in [0, 1/2]$ and is most powerful when $\theta = 1/2$. This implies Claim 1. Similar arguments can be applied to prove Claim 2.

Claim 3 and 4 rely on the almost necessary condition on the consistency of $\hat{\pi}^*_\theta$ in the following lemma.

Lemma A.3 *Consider model (9). Let $\delta(t) = [\bar{\Phi}(t)]^\theta$ with $\theta \in [0, 1]$ and construct $\hat{\pi}^*_\theta$ with a degenerating $\alpha = \alpha_m \to 0$. If A_m satisfies $A_m \gg 1$ and*

$$\bar{\Phi}^{-1}\left((\pi/c^*_\theta)^{1/\theta}\right) - A_m \gg 1,$$

then $\hat{\pi}^*_\theta/\pi \to 0$ in probability.

For $\theta = 0.5$, we have $(\pi/c^*_\theta)^{1/\theta} = (\pi/c^*_{0.5})^2 = \pi^2/(C\bar{\rho}_\Sigma \log m)$. By Theorem 2.2 and Lemma A.3, the sufficient and almost necessary condition for the consistency of $\hat{\pi}^*_{0.5}$ is

$$A_m - \bar{\Phi}^{-1}\left(\frac{\pi^2}{C\bar{\rho}_\Sigma \log m}\right) \gg 1.$$

Similarly, for $\theta = 1$, we have $(\pi/c^*_\theta)^{1/\theta} = \pi/c^*_1 = \pi/(C\sqrt{\log m})$. By Theorem 2.3 and Lemma A.3, the sufficient and almost necessary condition for the consistency of $\hat{\pi}^*_1$ is

$$A_m - \bar{\Phi}^{-1}\left(\frac{\pi}{C\sqrt{\log m}}\right) \gg 1.$$

The above implies Claim 3 and 4. This concludes the proof of Theorem 2.4.

A.7. Proof of Lemma A.3

First, pick a $\tau_m = [\bar{\Phi}^{-1}\left((\pi/c^*_\theta)^{1/\theta}\right) + A_m]/2$. Then, we have

$$\bar{\Phi}(A_m) \gg \bar{\Phi}(\tau_m) \gg (\pi/c^*_\theta)^{1/\theta} \qquad \text{and} \qquad \tau_m - A_m \gg 1. \tag{22}$$

Define

$$D_\theta(t) = \bar{F}_m(t) - 2\bar{\Phi}(t) - c^*_\theta[\bar{\Phi}(t)]^\theta.$$

By definition of $\hat{\pi}^*_\theta$, for any $\epsilon > 0$,

$$P\left(\frac{\hat{\pi}^*_\theta}{\pi} > \epsilon\right) \quad = \quad P\left(\max_{t \in \mathbb{T}} \frac{D_\theta(t)}{1 - 2\bar{\Phi}(t)} > \epsilon\pi\right)$$

$$\begin{aligned}
&\leq P\left(\max_{t\in[1,\tau_m]\cap\mathbb{N}} \frac{D_\theta(t)}{1-2\bar{\Phi}(t)} > \epsilon\pi\right)\\
&\quad +P\left(\max_{t\in[\tau_m,\sqrt{5\log m}]\cap\mathbb{N}} \frac{D_\theta(t)}{1-2\bar{\Phi}(t)} > \epsilon\pi\right)\\
&\leq P\left(\max_{t\in[1,\tau_m]\cap\mathbb{N}} D_\theta(t) > 0\right)\\
&\quad +P\left(\max_{t\in[\tau_m,\sqrt{5\log m}]\cap\mathbb{N}} D_\theta(t) > \frac{1}{2}\epsilon\pi\right). \qquad (23)
\end{aligned}$$

Now, consider the first term in (23).

$$\begin{aligned}
&P\left(\max_{t\in[1,\tau_m]\cap\mathbb{N}} D_\theta(t) > 0\right)\\
&= P\Bigg(\max_{t\in[1,\tau_m]\cap\mathbb{N}} \Bigg\{(1-\pi)[m_0^{-1}\sum_{j\in I_0} 1_{\{|Z_j^0|>t\}} - 2\bar{\Phi}(t)]\\
&+\pi\ [s^{-1}\sum_{j\in I_1} 1_{\{|Z_j^1|>t\}} - 2\bar{\Phi}(t)] - c_\theta^*[\bar{\Phi}(t)]^\theta\Bigg\} > 0\Bigg)\\
&\leq P\left(\max_{t\in[1,\tau_m]\cap\mathbb{N}} \left\{(1-\pi)[m_0^{-1}\sum_{j\in I_0} 1_{\{|Z_j^0|>t\}} - 2\bar{\Phi}(t)] + \pi - c_\theta^*[\bar{\Phi}(t)]^\theta\right\} > 0\right)\\
&\leq P\left(\max_{t\in[1,\tau_m]\cap\mathbb{N}} \left\{(1-\pi)[m_0^{-1}\sum_{j\in I_0} 1_{\{|Z_j^0|>t\}} - 2\bar{\Phi}(t)] - \frac{1}{2}c_\theta^*[\bar{\Phi}(t)]^\theta\right\} > 0\right) \qquad (24)\\
&+P\left(\max_{t\in[1,\tau_m]\cap\mathbb{N}} \left\{\pi - \frac{1}{2}c_\theta^*[\bar{\Phi}(t)]^\theta\right\} > 0\right), \qquad (25)
\end{aligned}$$

where (24) goes to 0 by similar arguments as in the proof of Theorem 2.1 that rely on the bounding sequence property of c_θ^*. On the other hand, because $\bar{\Phi}(\tau_m) \gg (\pi/c_\theta^*)^{1/\theta}$ as in (22), (25) also goes to 0.

Next, consider the second term in (23).

$$\begin{aligned}
&P\left(\max_{t\in[\tau_m,\sqrt{5\log m}]\cap\mathbb{N}} D_\theta(t) > \frac{1}{2}\epsilon\pi\right)\\
&\leq P\Bigg(\max_{t\in[\tau_m,\sqrt{5\log m}]\cap\mathbb{N}} \Bigg\{(1-\pi)[m_0^{-1}\sum_{j\in I_0} 1_{\{|Z_j^0|>t\}}\\
&\quad - 2\bar{\Phi}(t)] - c_\theta^*[\bar{\Phi}(t)]^\theta\Bigg\} > \frac{1}{4}\epsilon\pi\Bigg) \qquad (26)\\
&+P\left(\max_{t\in[\tau_m,\sqrt{5\log m}]\cap\mathbb{N}} \left\{\pi[s^{-1}\sum_{j\in I_1} 1_{\{|Z_j^1|>t\}} - 2\bar{\Phi}(t)]\right\} > \frac{1}{4}\epsilon\pi\right), \qquad (27)
\end{aligned}$$

where (26) goes to zero by the bounding sequence property of c_θ^*. To investigate (27), we have

$$\begin{aligned}
& P\left(\max_{t\in[\tau_m,\sqrt{5\log m}]\cap\mathbb{N}}\left\{\pi[s^{-1}\sum_{j\in I_1}1_{\{|Z_j^1|>t\}}-2\bar{\Phi}(t)]\right\}>\frac{1}{4}\epsilon\pi\right)\\
\leq\ & P\left(\max_{t\in[\tau_m,\sqrt{5\log m}]\cap\mathbb{N}}\left\{s^{-1}\sum_{j\in I_1}1_{\{|Z_j^1|>t\}}\right\}>\frac{1}{4}\epsilon\right)\\
\leq\ & P\left(s^{-1}\sum_{j\in I_1}1_{\{|Z_j^1|>\tau_m\}}>\frac{1}{4}\epsilon\right)\leq C\epsilon^{-1}E(s^{-1}\sum_{j\in I_1}1_{\{|Z_j^1|>\tau_m\}})\\
\leq\ & CP(|Z_j^1|>t)\leq C\bar{\Phi}(\tau_m-A_m)\to 0
\end{aligned}$$

where the last step is because $\tau_m - A_m \gg 1$ as in (22). This concludes the proof of Lemma A.3.

References

[1] Arias-Castro, E., Candès, E. J. and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics* **39** 2533–2556. MR2906877

[2] Blanchard, G., Neuvial, P. and Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics* **48** 1281–1303. MR4124323

[3] Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B* **73** 325–349. MR2815779

[4] Buhlmann, P., Kalisch, K. and Meier, L. (2014). High-Dimensional Statistics with a View Toward Applications in Biology. *Annu Rev Stat Appl.* **1** 255–278.

[5] Cai, T., Liu, W. and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association* **108** 265–277. MR3174618

[6] Cai, T. and Sun, W. (2017). Optimal screening and discovery of sparse signals with applications to multistage high-throughput studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 197. MR3597970

[7] Cai, T. T., Jin, J., Low, M. G. et al. (2007). Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics* **35** 2421–2449. MR2382653

[8] Chen, X. (2019). Uniformly consistently estimating the proportion of false null hypotheses via Lebesgue–Stieltjes integral equations. *Journal of Multivariate Analysis* **173** 724–744. MR3980489

[9] Delattre, S. and Roquain, E. (2016). On empirical distribution function of high-dimensional Gaussian vector components with an application to multiple testing. *Bernoulli* **22** 302–324. MR3449784

[10] Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics* **35** 1351–1377. MR2351089

[11] Fan, J., Han, X. and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association* **107** 1019–1035. MR3010887

[12] Finner, H. and Gontscharuk, V. (2009). Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 1031–1048. MR2750256

[13] Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics* **32** 1035–1061. MR2065197

[14] Hemerik, J., Solari, A. and Goeman, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika* **106** 635–649. MR3992394

[15] Jeng, X. J. and Chen, X. (2019). Variable selection via adaptive false negative control in linear regression. *Electronic Journal of Statistics* **13** 5306–5333. MR4043074

[16] Jeng, X. J., Daye, Z. J., Lu, W. and Tzeng, J.-Y. (2016). Rare variants association analysis in large-scale sequencing studies at the single locus level. *PLoS computational biology* **12** e1004993.

[17] Jeng, X. J., Hu, Y., Sun, Q. and Li, Y. (2022). Weak Signal Inclusion Under Dependence and Applications in Genome-wide Association Study. *arXiv preprint arXiv:2212.13574*.

[18] Jeng, X. J., Zhang, T. and Tzeng, J.-Y. (2019). Efficient signal inclusion with genomic applications. *Journal of the American Statistical Association* **114** 1787–1799. MR4047300

[19] Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society: Series B* **70** 461–493. MR2420411

[20] Jin, J. and Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association* **102** 495–506. MR2325113

[21] Jin, J. and Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association* **102** 495–506. MR2325113

[22] Katsevich, E. and Ramdas, A. (2020). Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *The Annals of Statistics* **48** 3465–3487. MR4185816

[23] Li, W. V. and Shao, Q.-M. (2002). A normal comparison inequality and its applications. *Probability Theory and Related Fields* **122** 494–508. MR1902188

[24] Meinshausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses.

The Annals of Statistics **34** 373–393. MR2275246

[25] Schwartzman, A. and Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika* **98** 199–214. MR2804220

[26] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 479–498. MR1924302

[27] Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* **31** 2013–2035. MR2036398

[28] Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* **279**. John Wiley & Sons.