Estimating The Proportion of Signal Variables Under Arbitrary Covariance Dependence

X. Jessie Jeng*

Department of Statistics, North Carolina State University

August 27, 2020

Abstract

Accurately estimating the proportion of signals hidden in a large amount of noise variables is of interest in many scientific inquires. In this paper, we consider realistic but theoretically challenging settings with arbitrary covariance dependence between variables. We define mean absolute correlation (MAC) to measure the overall dependence strength and investigate a family of estimators for their performances in the full range of MAC. We explicit the joint effect of MAC and signal sparsity on the performances of the family of estimators and discover that the most powerful estimator under independence is no longer most effective when the MAC dependence is strong enough. Motivated by the theoretical insight, we propose a new estimator to better adapt to arbitrary covariance dependence. The proposed method compares favorably to several existing methods in extensive finite-sample settings with strong to weak covariance dependence and real dependence structures from genetic association studies.

Keywords: Dependence adaptivity; High-dimension data; Lower bound estimator; Sparse signal

*Address for correspondence: Department of Statistics, North Carolina State University, 2311 Stinson Dr., Raleigh, NC 27695-8203, USA. Email: xjjeng@ncsu.edu.
1 Introduction

We consider the problem of estimating the proportion of information bearing signals that are sparsely located among a large amount of noise variables. This problem is of great interest in many scientific inquiries. For example, many multiple testing methods need to implement estimates of signal proportion to calculate the local false discovery rate (Efron, 2007), to derive the q-value (Storey, 2003), and to further improve the power of existing multiple comparison approaches (Storey, 2002; Finner and Gontscharuk, 2009). Moreover, in many multi-stage signal discovery studies, a good estimation of signal proportion can assist efficient pre-screening, signal identification, and sample size calculation (Cai and Sun, 2017). A recent line of research, which focuses on retaining a high proportion of signals through efficient false negative control, also replies on the estimation of signal proportion as a benchmark for signal inclusion (Jeng et al., 2016, 2019; Jeng and Hu, 2020).

Although estimation of signal proportion is widely requested, methodology development has met two major challenges. First, signals of different sparsity levels often call for different estimation methods, and the sparsity levels are unknown a priori. Secondly, the set of variables under investigation may have complex dependence structures. There exist a number of rigorously developed methods. Most of them, however, assume independence between variables (Genovese and Wasserman, 2004; Meinshausen and Rice, 2006; Jin and Cai, 2007a) and consider sparsity levels within a certain range (Cai et al., 2007; Jin, 2008). A nice review of the existing methods can be found in Chen (2019). More recent developments extend the study to consider specific dependence structures of the variables such as block-diagonal covariance matrices (Jeng et al., 2019) and certain dependence conditions to facilitate accurate precision matrix estimation and bias mitigation in linear regression (Jeng and Chen, 2019). There lacks a method to consistently estimate the signal proportion under arbitrary covariance dependence when the sparsity of signals is unknown and possibly falls in a wide range. Such an estimator can have far-reaching impact in real practices.

In this paper, we define mean absolute correlation (MAC) to measure the overall dependence strength and investigate a family of estimators for their performances in the full extent of MAC. We explicit the joint effect of MAC and signal sparsity on the performances of the
family of estimators and discover that the most powerful estimator under independence is no longer most effective when the MAC dependence is strong enough given a signal sparsity level, or when signals are sparse enough given a MAC level. Motivated by the theoretical insight, we propose a new estimator to better adapt to arbitrary covariance dependence.

The new method is compared with several popular methods in extensive simulation examples including strong to weak covariance dependence between variables and real dependence structures from genetic association studies. It shows that although the winner of the several existing methods changes over different settings, performance of the new method is either comparable to or better than the performance of that winner in each setting. We apply the new methods to analyze two real datasets. The first dataset is from an expression quantitative trait loci (eQTL) study with 8637 candidate single-nucleotide polymorphisms (SNPs), for which the overall dependence in terms of the MAC level is rather weak. The second dataset is from a classical association study with microarray data, where 4088 candidate genes possess much stronger overall dependence with a high MAC level. We compare the estimate of the new method with those of the existing methods and relate the results to the simulation study with similar settings. The comparison results indicate better adaptivity of the new method to complex real dependence structures.

The rest of the paper is organized as follows. Section 2 first introduces a family of lower bound estimators and develops a general result on their estimation consistency under dependence. Then, for specific estimators in the family, the joint effects of covariance dependence and signal sparsity are explicated and compared in terms of power under dependence. Consequently, a new, more powerful estimator is developed for applications with arbitrary covariance dependence. Section 3 compares the proposed method with existing methods in simulation examples. Section 4 applies the new and existing methods to real genetic association studies and relates the results with the simulation examples that adopt similar settings. Section 5 concludes the work with further discussions. Technical proofs and supporting figures and tables are provided in Appendices.
2 Method and Theory

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

$$X_j \sim F_0 \cdot 1\{j \in I_0\} + F_1 \cdot 1\{j \in I_1\}, \quad j = 1, \ldots, p,$$

where $F_0$ and $F_1$ are the null and signal distribution, respectively. Define the signal proportion

$$\pi = |I_1|/p.$$

We assume that $F_0$ is continuous and known a priori. All the other components are unknown. Our goal is to estimate the signal proportion $\pi$ without the need to specify $F_1$ or to identify which variables are signals.

2.1 A family of estimators

Meinshausen and Rice (2006) introduced a family of proportion estimators that are built upon the empirical process of $p$-values under independence. This family of estimators have been proved to provide lower bound estimates for the true proportion $\pi$. Members in the family are indexed by the choice of a bounding function, and it has been shown that the estimator with bounding function $\delta(u) = \sqrt{u(1-u)}$ has the best overall performance for signals of different sparsity levels. This conclusion, unfortunately, does not hold anymore under arbitrary dependence. When variables are arbitrarily dependent, the limiting distribution of the empirical process of $p$-values is generally unknown and may not even have an analytic expression. Moreover, the dependence effect mingles with signal sparsity to influence the performances of different estimators in the family. These difficulties substantially complicate the estimation problem and motivate us to develop new techniques to study the family of estimators.

For presentation simplicity, we perform inverse normal transformation as $Z_j = \Phi^{-1}(F_0(X_j))$, where $\Phi^{-1}$ is the inverse of the cumulative distribution function of a standard Normal random
variable. Then, we have

\[ Z_j \sim \Phi \cdot 1\{j \in I_0\} + G \cdot 1\{j \in I_1\}, \quad j = 1, \ldots, p, \quad (1) \]

where \( G \) denotes the signal distribution after inverse normal transformation, which remains unknown.

Next, we construct a modified family of estimators to accommodate dependence among \( Z_j \). Let

\[ \bar{W}_p(t) = p^{-1} \sum_{j=1}^{p} 1\{|W_j| > t\}, \]

where \((W_1, \ldots, W_p)\) follow the joint null distribution of \((Z_1, \ldots, Z_p)\). Denote \( \bar{\Phi}(t) = 1 - \Phi(t) \).

For a given function \( \delta(t) \) that is strictly positive on \((0, \infty)\), define

\[ V_{p,\delta} = \sup_{t>0} \frac{|\bar{W}_p(t) - 2\bar{\Phi}(t)|}{\delta(t)}. \quad (2) \]

Apparently, \( V_{p,\delta} \) varies with the choice of \( \delta(t) \), and \( \delta(t) \) is called a bounding function. For a given \( \delta(t) \) and a control level \( \alpha \), we define the corresponding bounding sequence \( c_{p,\delta} \) as a function of \( p, \delta \) and \( \alpha \) that satisfies the following properties:

(a) \( pc_{p,\delta} > p_0c_{p_0,\delta} \), where \( p_0 = |I_0| \), and

(b) \( P(V_{p,\delta} > c_{p,\delta}) < \alpha \) for all \( p \).

From the above properties, it can be seen that \( c_{p,\delta} \) is an upper bound of \( V_{p,\delta} \), while \( V_{p,\delta} \) relies on only noise variables. Both \( V_{p,\delta} \) and \( c_{p,\delta} \) carry the information of dependence among the variables. Also, the absolute sign in the numerator of (2) stabilizes \( V_{p,\delta} \) as \( \bar{W}_p(t) - 2\bar{\Phi}(t) \) may not be asymptotically symmetric anymore under arbitrary dependence.

Then, given the observed \( Z_j \), a family of estimators indexed by \( \delta \) are constructed as

\[ \hat{\pi}_{\delta} = \sup_{t>0} \frac{\bar{F}_p(t) - 2\bar{\Phi}(t) - c_{p,\delta}\delta(t)}{1 - 2\Phi(t)}, \quad (3) \]

where

\[ \bar{F}_p(t) = p^{-1} \sum_{j=1}^{p} 1\{|Z_j| > t\}. \]
It can be seen that the $\hat{\pi}_\delta$ family of estimators hinge on the choices of the bounding function $\delta(t)$ and the corresponding bounding sequence $c_{p,\delta}$. As the bounding sequence $c_{p,\delta}$ can imply a normal range of $\bar{F}_p(u)$ if all variables are noise, the information carried by the observed $\bar{F}_p(u)$ that exceeds the normal range presents evidence for the existence of signals. This version of lower bound family is for signals of two-sided effects. Minor changes to accommodate one-sided signal effect is straightforward. More details for the numerical implementation of $\hat{\pi}_\delta$ can be found at the end of Section 2.3.

As shown in Meinshausen and Rice (2006), most estimators in the original family are not uniformly effective for signals of different sparsity levels. More specifically, the signal proportion can be re-parameterized as $\pi = p^{-\gamma}$ with $\gamma \in (0, 1/2)$ representing the relatively dense case and $\gamma \in [1/2, 1)$ representing the sparse case. Consistency of different estimators in the original family was proved under independence for $\gamma$ in the dense and sparse cases separately. Here, we consider the estimation problem in the more challenging setting with arbitrary dependence and modify the family of estimators to accommodate the dependence. In order to provide theoretical insight for the modified family, we impose a monotonicity constraint on the bounding function $\delta(t)$, which allows us to present the following general result on all the estimators in the family for signals in the full range of sparsity levels with $\gamma \in (0, 1)$.

**Lemma 2.1** Consider model (1). For a given bounding function $\delta(t)$, if there exist a bounding sequence $c_{p,\delta}$ satisfying the properties in (a) and (b), then

$$P(\hat{\pi}_\delta < \pi) \geq 1 - \alpha. \quad (4)$$

On the other hand, for $\pi$ satisfying $0 < \pi \ll 1$, if $\delta(t)$ is non-increasing with respect to $t$, and $G = G_p$ such that $G_p(\tau) \to 0$ or $G_p(-\tau) \to 1$ for some $\tau \gg \max\{\delta^{-1}(\pi/c_{p,\delta}), 1\}$, then

$$P(\hat{\pi}_\delta > (1 - \epsilon)\pi) \to 1 \quad (5)$$

for any constant $\epsilon > 0$. 

6
The above lemma says that the lower bound property of \( \hat{\pi}_\delta \) in (4) holds as long as a bounding sequence \( c_{p,\delta} \) satisfying (a) and (b) can be found. On the other hand, the upper bound property of \( \hat{\pi}_\delta \) in (5) holds 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. When both lower bound and upper bound conditions are satisfied for a given \( \delta(t) \) and a degenerating \( \alpha \) such that \( \alpha = \alpha_p \to 0 \), \( \hat{\pi}_\delta \) consistently estimate the true signal proportion, i.e., for any constant \( \epsilon > 0 \),

\[
P((1 - \epsilon)\pi \leq \hat{\pi}_\delta < \pi) \to 1.\]

The above results hold for arbitrarily dependent variables and for sparse signals that are not of a fixed proportion of all the variables.

### 2.2 Explicit joint effect of dependence and sparsity

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 \( \delta(t) = [\Phi(t)]^\theta, \theta \in [0, 1] \), as the corresponding \( \hat{\pi}_\delta \) are closely related to the existing estimators, whose consistency has been studied under independence in literature (Meinshausen and Rice (2006) and the references therein). Valuable insights can be obtained by comparing the estimators’ performances under independence and dependence, which later help us construct a more powerful estimator under dependence.

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

\[
(Z_1, \ldots, Z_p) \sim N_p(\mu, \Sigma),
\]

where \( \mu \) is a \( p \)-dimensional sparse vector with \( \mu_j = A_j \cdot 1\{j \in I_1\}, A_j \neq 0 \), and \( \Sigma \) is an arbitrary correlation matrix, i.e., \( \Sigma_{ij} = Corr(Z_i, Z_j) \). We define the Mean Absolute
Correlation (MAC) to calibrate the covariance dependence as

$$\bar{\rho}_\Sigma = \sum_{i=1}^{p} \sum_{j=1}^{p} |\Sigma_{ij}| / p^2. \quad (7)$$

A larger value of $\bar{\rho}_\Sigma$ indicates stronger overall dependence.

Moreover, we employ a discretization technique from Arias-Castro et al. (2011) and Jeng and Chen (2019) as follows. Define $T = [1, \sqrt{5 \log p}] \cap \mathbb{N}$ and the discretized version of $V_{p,\delta}$ as

$$V^*_{p,\delta} = \max_{t \in T} \left| \frac{\bar{W}_p(t) - 2\bar{\Phi}(t)}{\delta(t)} \right|. \quad (8)$$

Denote $c^*_{p,\delta}$ as the bounding sequence based on $V^*_{p,\delta}$, and define the corresponding proportion estimator as

$$\hat{\pi}_\delta^* = \max_{t \in T} \frac{\bar{P}_p(t) - 2\bar{\Phi}(t) - c^*_{p,\delta} \delta(t)}{1 - 2\bar{\Phi}(t)}. \quad (9)$$

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

**Theorem 2.1** Consider model (6). Let $\delta(t) = [\bar{\Phi}(t)]^\theta$ with $\theta \in [0, 1/2]$. Then, there exists a bounding sequence

$$c^*_{p,\delta} = O \left( \sqrt{\bar{\rho}_\Sigma (\log p)^{\theta + 1/2}} \right),$$

that satisfies properties (a) and (b), and the corresponding estimator $\hat{\pi}_\delta^*$ satisfies $P(\hat{\pi}_\delta^* < \pi) \geq 1 - \alpha$. Moreover, for $\pi$ satisfying $0 < \pi \ll 1$, if $G = G_p$ such that $G_p(\tau) \to 0$ or $G_p(-\tau) \to 1$ for some

$$\tau \gg \max \left\{ 1, \bar{\Phi}^{-1} \left( \frac{\pi^{1/\theta}}{\bar{\rho}_\Sigma^{1/(2\theta)} (\log p)^{(\theta + 1/2)/(2\theta)}} \right) \right\}, \quad (10)$$

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 $\theta \in [0, 1/2]$, we can find a bounding sequence $c^*_{p,\delta}$, whose order increases with respect to $\bar{\rho}_\Sigma$, $p$ and $\theta$, respectively. The second part
of the theorem provides a condition on $G$ in (10) for the consistency of $\hat{\pi}_\delta$. This condition is easier to be satisfied for less sparse signals (larger $\pi$) or less dependent variables (smaller $\bar{\rho}_\Sigma$). Note that $\Phi^{-1}(\pi^{1/\theta}/(\bar{\rho}_\Sigma^{1/(2\theta)}(\log p)^{(\theta+1)/2}))$ is well-defined only for $\pi < \sqrt{\bar{\rho}_\Sigma(\log p)^{\theta+1/2}}$. In the case $\pi \geq \sqrt{\bar{\rho}_\Sigma(\log p)^{\theta+1/2}}$, condition (10) is simply $\tau \gg 1$. For the special case with independent variables, $\bar{\rho}_\Sigma = 1/p$ and condition (10) degenerates to

$$\tau \gg \max\{1, \Phi^{-1}(\pi^{1/\theta}/(\log p)^{(\theta+1)/2})\},$$

which agrees with the sufficient and necessary condition in Meinshausen and Rice (2006) for the consistency of $\hat{\pi}_\delta$ under independence. The comparison can be made by adopting the same parameterization as in Theorem 3 of their paper with $\pi = p^{-\gamma}$, $\gamma \in [1/2, 1)$, $\nu = \theta$, and $\kappa = 2$.

Results in Theorem 2.1 explicate how the performance of $\hat{\pi}_\delta^*$ with $\delta(t) = [\Phi(t)]^\theta, \theta \in [0, 1/2]$ 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.

**Theorem 2.2** Consider model (6). Let $\delta(t) = [\Phi(t)]^\theta$ with $\theta \in (1/2, 1]$. Then, there exists a bounding sequence $c_{p,\delta}^* = O(\sqrt{\log p})$, that satisfies the properties (a) and (b), and the corresponding estimator $\hat{\pi}_\delta^*$ satisfies $P(\hat{\pi}_\delta^* < \pi) \geq 1 - \alpha$. Moreover, for $\pi$ satisfying $0 < \pi \ll 1$, if $G = G_p$ such that $G_p(\tau) \to 0$ or $G_p(-\tau) \to 1$ for some

$$\tau \gg \Phi^{-1}\left(\pi^{1/\theta}/(\log p)^{(\theta+1)/2}\right), \quad (11)$$

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

This theorem shows that for $\delta(t) = [\Phi(t)]^\theta$ with $\theta \in (1/2, 1]$, we can find a bounding sequence $c_{p,\delta}^*$, whose order does not involve the MAC dependence level $\bar{\rho}_\Sigma$. Moreover, the intensity condition in (11) does not involve $\bar{\rho}_\Sigma$, which means that $\hat{\pi}_\delta^*$ with $\theta \in (1/2, 1]$ is consistent under (11), no matter how strong the MAC dependence is.

Although the above analyses are for the discretized version of $\hat{\pi}_\delta$, it provides important insight on the performance of the $\hat{\pi}_\delta$ family under dependence, which is also supported by
extensive simulation studies in Section 3.

2.3 A new estimator for dependent variables

The study in Section 2.2 reveals that covariance dependence has different effects on different estimators in the $\hat{\pi}_\delta$ family. As condition (10) gets more relaxed as $\bar{\rho}_\Sigma$ decreases, estimators with $\theta \in [0, 1/2]$ are more powerful under weaker dependence in terms of the MAC level. On the other hand, when MAC dependence is strong enough, estimators with $\theta \in (1/2, 1]$ could be more powerful as condition (11) does not change with $\bar{\rho}_\Sigma$. Motivated by these findings, we propose to construct a new estimator of the form

$$\hat{\pi}_{\text{adap}} = \max\{\hat{\pi}_\delta, \delta \in \Delta\},$$

where $\Delta$ is the set of $\delta(t)$ functions that render the most powerful estimators in different dependence scenarios. Note that all the $\delta(t)$ functions in $\Delta$ result in conservative estimators with the lower bound property $P(\hat{\pi}_\delta < \pi) \geq 1 - \alpha$ as stated in Lemma 2.1. Therefore, the new estimator $\hat{\pi}_{\text{adap}}$ naturally inherit the lower bound property.

The power of $\hat{\pi}_{\text{adap}}$ depends on the candidate $\delta(t)$ functions in $\Delta$. We consider $\delta(t) = [\bar{\Phi}(t)]^\theta$ with specific $\theta$ values. Based on the results in Theorem 2.1, we find that among the estimators with $\theta \in [0, 1/2]$, the most powerful one has $\theta = 1/2$ as condition (10) is less stringent with larger $\theta$. On the other hand, based on the results in Theorem 2.2, the most powerful estimator among the estimators with $\theta \in (1/2, 1]$ has $\theta = 1$. The above analysis help us narrow down to two candidates that are most powerful in their own $\theta$ groups. Comparing these two candidates, we can see that $\theta = 1/2$ can result in a more (or less) powerful estimator than $\theta = 1$ when covariance dependence is relatively weak (or strong) with $\bar{\rho}_\Sigma \leq \pi/\sqrt{\log p}$ (or $\bar{\rho}_\Sigma > \pi/\sqrt{\log p}$). Since none of the candidates dominates the other under arbitrary dependence, our new estimator is constructed as

$$\hat{\pi}_{\text{adpt}} = \max\{\hat{\pi}_{0.5}, \hat{\pi}_1\}. \quad (12)$$

where $\hat{\pi}_{0.5}$ denotes $\hat{\pi}_\delta$ with $\delta = [\bar{\Phi}(t)]^{1/2}$, and $\hat{\pi}_1$ denotes $\hat{\pi}_\delta$ with $\delta = \bar{\Phi}(t)$. $\hat{\pi}_{\text{adpt}}$ is likely to be
comparable to $\hat{\pi}_{0.5}$ and outperform $\hat{\pi}_1$ under relatively weak dependence, and be comparable to $\hat{\pi}_1$ and outperform $\hat{\pi}_{0.5}$ when covariance dependence is strong.

**Numerical Implementation.** We conclude this section with additional notes on the numerical implementation of $\hat{\pi}_{adap}$. Specifically, we simulate $(w_1, \ldots, w_p)$ following the joint null distribution of $Z_j$. When the joint null distribution is unknown in real applications, $(w_1, \ldots, w_p)$ can often be simulated non-parametrically. For example, when $(Z_1, \ldots, Z_p)$ represent the associations between a set of explanatory variables and a response variable, a common practice to simulate $(w_1, \ldots, w_p)$ is by randomly shuffling only the sample of the response variable to remove the potential associations. More details for such permutation approaches can be found in Westfall and Young (1993). Then, following (2), we simulate $V_{p,0.5}$ and $V_{p,1}$ corresponding to $\theta = 1/2$ and 1 by taking the maximums over $t = w_1, \ldots, w_p$. Next, we set $\alpha = 0.1$ and generate $c_{p,0.5}$ and $c_{p,1}$ as the $(1 - \alpha)$th quantile of 1000 replicates of $V_{p,0.5}$ and $V_{p,1}$, respectively. The simulated $c_{p,0.5}$ and $c_{p,1}$ are implemented to calculate $\hat{\pi}_{0.5}$ and $\hat{\pi}_1$ as in (3) by taking the maximums over $t = z_1, \ldots, z_p$, where $z_1, \ldots, z_p$ are the observed variables after inverse normal transformation. Finally, the new estimator $\hat{\pi}_{adap}$ is calculated by (12).

### 3 Simulation Study

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

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

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

(c) **Block correlation.** $\Sigma$ has square diagonal blocks. The off-diagonal elements in the blocks are 0.5, and the elements outside the blocks are zero.

(d) **Sparse correlation.** $\Sigma$ has nonzero elements randomly located. The data generation process is similar to Model 3 in Cai et al. (2013). Let $\Sigma^* = (\sigma_{ij})$, where $\sigma_{ii} = 1,$
\[ \sigma_{ij} = 0.9 \text{Bernoulli}(1, 0.1) \text{ for } i < j \text{ and } \sigma_{ji} = \sigma_{ij}. \text{ Then } \Sigma = I^{1/2}(\Sigma^* + \delta I)/(1 + \delta)I^{1/2}, \]
where \( \delta = |\lambda_{\text{min}}(\Sigma^*)| + 0.05. \)

(e) **SNP correlation.** \( \Sigma \) is the sample correlation matrix of the real SNP data on Chromosome 21 from 90 individuals in the International HapMap project.

(f) **Gene correlation.** \( \Sigma \) is the sample correlation matrix of the real gene expression data from 71 individuals in a riboflavin production study.

We generate test statistics \( Z_1, \ldots, Z_p \sim N((\mu_1, \ldots, \mu_p), \Sigma) \) and set \( p = 2000 \) for cases (a)-(d) above. Case (e) has \( p = 8657 \), which is the number of SNPs in the dataset, and case (f) has \( p = 4088 \), 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 \( 400 \times 400 \). \( (\mu_1, \ldots, \mu_p) \) 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 Dependence effect on bounding sequences

We first demonstrate the MAC levels as defined in (7) for the dependence structures in (a)-(f) above and report the realized values for \( c_{p,0.5} \) and \( c_{p,1} \). 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, \( c_{p,0.5} \) seems to vary positively with \( \bar{\rho}_{\Sigma} \), whereas \( c_{p,1} \) does not show such tendency. The numerical results are demonstrated more clearly in Figure 1 and are consistent with Theorem 2.1 and 2.2 about the dependence effect on the two different bounding sequences.

|                                | Autocorr | Equal corr | Block corr | Sparse corr | SNP corr | Gene corr |
|--------------------------------|----------|------------|------------|-------------|----------|-----------|
| \( \bar{\rho}_{\Sigma} \)     | 0.0095   | 0.5003     | 0.1003     | 0.0042      | 0.0869   | 0.3353    |
| \( c_{p,0.5} \)               | 0.178    | 0.87       | 0.397      | 0.099       | 0.222    | 0.706     |
| \( c_{p,1} \)                 | 8.46     | 4.39       | 5.58       | 6.79        | 13.6     | 6.42      |
3.2 Comparison with existing methods

We compare the new estimator $\hat{\pi}_{adpt}$ with $\hat{\pi}_{0.5}$ and $\hat{\pi}_1$ from the estimator family as well as two other existing methods. 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 $\mu_j = 3, 4, 5, 6$. The comparisons are organized into two sets of examples.

The first set of examples demonstrate the performances of $\hat{\pi}_{adpt}$, $\hat{\pi}_{0.5}$, and $\hat{\pi}_1$ as they all possess the lower bound property. Recall the theoretical results in Section 2.2 and 2.3 that $\hat{\pi}_{0.5}$ may outperform (or underperform) $\hat{\pi}_1$ when dependence is weak (or strong) enough or signals are less (or more) sparse. We observe these tendencies in Figure 2-7. 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. 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 $\pi$, and is comparable to $\hat{\pi}_{0.5}$ for larger $\pi$. The sparse correlation case (Figure 5) has the smallest $\bar{\rho}_\Sigma = 0.0042$, we see that $\hat{\pi}_{0.5}$ outperforms $\hat{\pi}_1$
for both $\pi = 0.02$ and $0.1$. The SNP correlation case (Figure 6) has $\bar{\rho}_Z = 0.0869$, which is moderately small. We see that $\hat{\pi}_1$ is slightly better for small $\pi$, and $\hat{\pi}_{0.5}$ is better for larger $\pi$. The gene correlation case (Figure 7) has $\bar{\rho}_Z = 0.3353$, which is fairly large. It shows that $\hat{\pi}_1$ outperforms $\hat{\pi}_{0.5}$ for both small and larger $\pi$. In all these examples, the new estimator $\hat{\pi}_{adpt}$ always matches the winner of $\hat{\pi}_{0.5}$ and $\hat{\pi}_1$ and exhibits generally better adaptivity to different dependence structures and signal sparsity levels.

Figure 2: Comparison under autocorrelation. “est\_0.05”, “est\_1”, and “est\_adpt” represent $\hat{\pi}_{0.5}$, $\hat{\pi}_1$, and $\hat{\pi}_{adpt}$, respectively. The top row has $\pi = 0.02$, and the bottom row has $\pi = 0.1$. The true $\pi$ values are highlighted by the red horizontal lines.
Figure 3: Comparison under equal correlation. Notations and symbols are the same as in Figure 2.

Figure 4: Comparison under block dependence. Notations and symbols are the same as in Figure 2.
Figure 5: Comparison under sparse dependence. Notations and symbols are the same as in Figure 2.

Figure 6: Comparison under SNP dependence. Notations and symbols are the same as in Figure 2.
The second set of examples compares the performance of $\hat{\pi}_{\text{adpt}}$ with the estimator $\hat{\pi}_{GW}$ developed in Genovese and Wasserman (2004) and $\hat{\pi}_{JC}$ developed in Jin and Cai (2007b). These two existing methods have been studied for relatively dense signals under independence. Table 2 and 3 show that for the cases (a)-(f), $\hat{\pi}_{GW}$ and $\hat{\pi}_{JC}$ tend to over-estimate the true $\pi$ when dependence is strong (such as in (b) Equal correlation and (f) Gene correlation) or signals are sparse (such as for $\pi = 0.02$). On the other hand, the performance of $\hat{\pi}_{adpt}$ seems to be generally more accurate and stable over different dependence structures and $\pi$ values.
Table 2: Mean values and standard deviations (in brackets) of $\hat{\pi}_{\text{adpt}}$, $\hat{\pi}_{\text{GW}}$, and $\hat{\pi}_{\text{JC}}$ when signals are relatively spare with $\pi = 0.02$.

| Dependence   | Method   | $\mu = 3$ | $\mu = 4$ | $\mu = 5$ | $\mu = 6$ |
|--------------|----------|-----------|-----------|-----------|-----------|
| Autocorr     | $\hat{\pi}_{\text{adpt}}$ | 0.020(0.003) | 0.020(0.003) | 0.020(0.002) | 0.021(0.002) |
|              | $\hat{\pi}_{\text{GW}}$    | 0.014(0.024) | 0.015(0.024) | 0.014(0.024) | 0.014(0.024) |
|              | $\hat{\pi}_{\text{JC}}$    | 0.050(0.045) | 0.050(0.045) | 0.050(0.045) | 0.051(0.044) |
| Equal corr   | $\hat{\pi}_{\text{adpt}}$ | 0.034(0.051) | 0.036(0.050) | 0.038(0.049) | 0.038(0.049) |
|              | $\hat{\pi}_{\text{GW}}$    | 0.194(0.274) | 0.194(0.273) | 0.194(0.273) | 0.194(0.274) |
|              | $\hat{\pi}_{\text{JC}}$    | 0.295(0.402) | 0.295(0.402) | 0.295(0.401) | 0.296(0.402) |
| Block corr   | $\hat{\pi}_{\text{adpt}}$ | 0.012(0.019) | 0.018(0.017) | 0.023(0.016) | 0.024(0.016) |
|              | $\hat{\pi}_{\text{GW}}$    | 0.060(0.091) | 0.060(0.091) | 0.059(0.091) | 0.060(0.091) |
|              | $\hat{\pi}_{\text{JC}}$    | 0.115(0.143) | 0.115(0.142) | 0.116(0.141) | 0.118(0.140) |
| Sparse corr  | $\hat{\pi}_{\text{adpt}}$ | 0.009(0.003) | 0.015(0.022) | 0.018(0.001) | 0.020(0.001) |
|              | $\hat{\pi}_{\text{GW}}$    | 0.003(0.006) | 0.003(0.007) | 0.003(0.006) | 0.003(0.007) |
|              | $\hat{\pi}_{\text{JC}}$    | 0.030(0.021) | 0.031(0.020) | 0.032(0.020) | 0.033(0.020) |
| SNP corr     | $\hat{\pi}_{\text{adpt}}$ | 0.034(0.051) | 0.036(0.050) | 0.038(0.049) | 0.038(0.049) |
|              | $\hat{\pi}_{\text{GW}}$    | 0.194(0.274) | 0.194(0.273) | 0.194(0.273) | 0.194(0.274) |
|              | $\hat{\pi}_{\text{JC}}$    | 0.295(0.402) | 0.295(0.402) | 0.295(0.401) | 0.296(0.402) |
| Gene corr    | $\hat{\pi}_{\text{adpt}}$ | 0.019(0.038) | 0.026(0.037) | 0.030(0.035) | 0.032(0.035) |
|              | $\hat{\pi}_{\text{GW}}$    | 0.151(0.187) | 0.151(0.188) | 0.151(0.188) | 0.151(0.187) |
|              | $\hat{\pi}_{\text{JC}}$    | 0.257(0.308) | 0.256(0.307) | 0.257(0.307) | 0.258(0.306) |

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 expression of gene CCT8 on chromosome 21. This gene has been found to be relevant to Down Syndrome (Bridic et al., 2011; Fan et al., 2012). We obtain the SNP data of unaffected subjects from the International HapMap project (http://zzz.bwh.harvard.edu/plink/res.shtml#hapmap) and the gene expression data from ftp://ftp.sanger.ac.uk/pub/genevar/. Our data includes 90 samples from Asian population (45 Japanese in Tokyo, Japan (JPT), and 45 Han Chinese in Beijing). We consider SNPs without missing values, which results in 8657 candidate SNPs.

As in Bradic et al. (2011) and Fan et al. (2012), we use marginal linear regression to derive test statistics for the associations between each SNP and the expression level of CCT8. Histogram of the test statistics is presented in Figure 8, where the long and thin right tail
Table 3: Mean values and standard deviations (in brackets) of $\hat{\pi}_{adpt}$, $\hat{\pi}_{GW}$, and $\hat{\pi}_{JC}$ when signals are relatively dense with $\pi = 0.1$.

| Dependence Method | $\mu = 3$ | $\mu = 4$ | $\mu = 5$ | $\mu = 6$ |
|-------------------|-----------|-----------|-----------|-----------|
| Autocorr          | $\hat{\pi}_{adpt}$ | 0.063(0.011) | 0.083(0.007) | 0.094(0.004) | 0.100(0.002) |
|                   | $\hat{\pi}_{GW}$   | 0.078(0.029) | 0.086(0.026) | 0.088(0.023) | 0.088(0.024) |
|                   | $\hat{\pi}_{JC}$   | 0.119(0.046) | 0.124(0.043) | 0.127(0.043) | 0.125(0.041) |
| Equal corr        | $\hat{\pi}_{adpt}$ | 0.067(0.053) | 0.098(0.042) | 0.110(0.038) | 0.113(0.036) |
|                   | $\hat{\pi}_{GW}$   | 0.233(0.263) | 0.246(0.256) | 0.250(0.254) | 0.250(0.253) |
|                   | $\hat{\pi}_{JC}$   | 0.333(0.394) | 0.336(0.377) | 0.344(0.374) | 0.349(0.373) |
| Block diag        | $\hat{\pi}_{adpt}$ | 0.052(0.026) | 0.082(0.018) | 0.097(0.013) | 0.102(0.012) |
|                   | $\hat{\pi}_{GW}$   | 0.114(0.094) | 0.126(0.089) | 0.128(0.085) | 0.129(0.086) |
|                   | $\hat{\pi}_{JC}$   | 0.173(0.144) | 0.179(0.135) | 0.183(0.131) | 0.184(0.131) |
| Sparse corr       | $\hat{\pi}_{adpt}$ | 0.073(0.006) | 0.088(0.004) | 0.096(0.001) | 0.099(0.001) |
|                   | $\hat{\pi}_{GW}$   | 0.069(0.012) | 0.075(0.010) | 0.077(0.009) | 0.077(0.009) |
|                   | $\hat{\pi}_{JC}$   | 0.106(0.019) | 0.108(0.019) | 0.110(0.019) | 0.110(0.018) |
| SNP corr          | $\hat{\pi}_{adpt}$ | 0.057(0.013) | 0.080(0.007) | 0.093(0.004) | 0.098(0.002) |
|                   | $\hat{\pi}_{GW}$   | 0.100(0.043) | 0.107(0.038) | 0.110(0.036) | 0.111(0.035) |
|                   | $\hat{\pi}_{JC}$   | 0.129(0.070) | 0.133(0.066) | 0.137(0.064) | 0.138(0.063) |
| Gene corr         | $\hat{\pi}_{adpt}$ | 0.055(0.038) | 0.087(0.033) | 0.102(0.028) | 0.107(0.026) |
|                   | $\hat{\pi}_{GW}$   | 0.202(0.181) | 0.213(0.177) | 0.216(0.173) | 0.217(0.172) |
|                   | $\hat{\pi}_{JC}$   | 0.301(0.302) | 0.304(0.292) | 0.312(0.287) | 0.314(0.284) |

indicates possibly a small proportion 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 $\bar{\rho}_\Sigma = 0.087$. The heatmap of the correlation matrix of the first 50 SNPs is illustrated in Figure 9. For this real correlation structures, the MAC level and the values of the bounding sequences $c_{p,0.5}$ and $c_{p,1}$ have been demonstrated in Table 1, case (e).

We apply the proposed estimator $\hat{\pi}_{adpt}$ and the existing estimators $\hat{\pi}_{GW}$ and $\hat{\pi}_{JC}$ to the dataset. The results are $\hat{\pi}_{adpt} = 0.0016$, $\hat{\pi}_{GW} = 0.0068$, and $\hat{\pi}_{JC} = 0.0186$. Consequently, the estimated numbers of relevant SNPs are 14, 60, and 161 by the three methods, respectively. This real application has been simulated in Section 3. The simulated results in the top row of Figure 6 and the sixth row of Table 2 are most relevant. The real data results seem to be consistent with the findings in simulation, which show that $\hat{\pi}_{adpt}$ is smaller than $\hat{\pi}_{GW}$ and $\hat{\pi}_{JC}$ in this setting and is expected to be closer to the true $\pi$.

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.
The dataset includes the expression levels of 4088 genes and the logarithm of riboflavin production rate of 71 individuals. Marginal regression coefficients are used as test statistics for associations between genes and riboflavin production. The histogram of the test statistics is presented in Figure 10, which suggests a larger signal proportion than that in the first real data example. Figure 11 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 $\hat{\rho}_\Sigma = 0.335$, which is fairly large. The realized
value of the bounding sequences \( c_{p,0.5} \) and \( c_{p,1} \) have been demonstrated in Table 1, case (f).

Figure 10: Histogram of test statistics for gene expression association

![Histogram of test statistics for gene expression association](image)

Figure 11: Heatmap of the absolute value of correlations for 50 gene expressions in riboflavin production study.

![Heatmap of the absolute value of correlations for 50 gene expressions](image)

In this example, we have \( \hat{\pi}_{adpt} = 0.064, \hat{\pi}_{GW} = 0.258, \) and \( \hat{\pi}_{JC} = 0.374, \) which corresponds to 261, 1053, and 1530 relevant genes. This real application has been simulated in Section 3, and the most relevant simulation results are in the bottom row of Figure 7 and the last row of Table 3. We can see that the real application results seem to be consistent with the findings in simulation, which show that \( \hat{\pi}_{adpt} \) is smaller than \( \hat{\pi}_{GW} \) and \( \hat{\pi}_{JC} \) in this setting and is expected to be closer to the true \( \pi. \)
5 Conclusion and Discussion

Estimating the proportion of sparse signals under dependence is notoriously difficult due to the involved joint effects of dependence, signal sparsity, and signal intensity. In this paper, we define the MAC level to measure arbitrary covariance dependence and explicate the effect of MAC dependence on a family of estimators. Different from the result developed under independence that there exists a most powerful estimator in the family, we find that no single estimator in the family is most powerful in different dependence scenarios. We identify candidate estimators that are most powerful in different scenarios and develop a new estimator that better adapts to the arbitrary covariance dependence. The new estimator inherits the lower bound property of the family, which is to provide a conservative estimate with high probability. This property is valuable in real applications as it requires no conditions on the unknown signals. Moreover, the new estimator is generally more powerful than any members in the lower bound family under arbitrary covariance dependence, and compares favorably to other popular methods in extensive numerical examples including weak to strong covariance dependence and real dependence structures from genetic associations studies.

Appendix

This section presents the proofs of Lemma 2.1, Theorem 2.1, and Theorem 2.2. The symbol $C$ denotes a genetic, finite constant whose values can be different at different occurrences.

5.1 Proof of Lemma 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 $p_0 = |I_0|$ and $s = |I_1|$. Then

$$\bar{F}_p(t) = p^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} + p^{-1} \sum_{j \in I_1} 1_{\{|Z_j^1| > t\}} \leq p^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} + p^{-1} s$$

$$= (1 - \pi)p_0^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} + \pi.$$
Consequently,

\[
P(\hat{\pi}_\delta > \pi) \leq P\left( \sup_{t > 0} \left\{ (1 - \pi) \left( p_0^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} - 2\Phi(t) \right) - c_{p,\delta}(t) \right\} > 0 \right)
\]

\[
\leq P\left( \sup_{t > 0} \left\{ p_0^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} - 2\Phi(t) - c_{p_0,\delta}(t) \right\} > 0 \right)
\]

\[
\leq P(\mathcal{V}_{p_0} > c_{p_0,\delta}) = \alpha,
\]

where the second and last inequalities are by properties (a) and (b) of \(c_{p,\delta}\), 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 \(\hat{\pi}_\delta > \bar{F}_p(t) - 2\Phi(t) - c_{p,\delta}(t)\) for any \(t > 0\) and

\[
\bar{F}_p(t) = \frac{1 - \pi}{p_0} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} + \frac{\pi}{s} \sum_{j \in I_1} 1_{\{|Z_j^1| > t\}},
\]

then

\[
\frac{\hat{\pi}_\delta}{\pi} - 1 > -\frac{1}{\pi} c_{p,\delta}(t) + \frac{1 - \pi}{\pi} \left( p_0^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| > t\}} - 2\Phi(t) \right)
\]

\[
+ \left( s^{-1} \sum_{j \in I_1} 1_{\{|Z_j^1| > t\}} - 1 \right) - 2\Phi(t)
\]

(13)

for any \(t > 0\). Now, set \(t\) in (13) at \(\tau\) such that \(\tau \gg \max\{\delta^{-1}(\pi/c_{p,\delta}), 1\}\). We will show that each term on the right hand side of (13) at \(t = \tau\) is of \(o_p(1)\).

First, by the monotonicity of \(\delta(t)\) and the condition \(\tau \gg \delta^{-1}(\pi/c_{p,\delta})\), we have \(\delta(\tau) = o(\pi/c_{p,\delta})\) and the first term \(A_1 = -c_{p,\delta}(\tau)/\pi = o(1)\).

Consider the second term \(A_2 = \pi^{-1}(1 - \pi) \left( p_0^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| \leq \tau\}} - 2\Phi(\tau) \right)\) in (13). The following lemma is proved in Section 5.2.

**Lemma 5.1** For any \(\tau \gg \max\{\delta^{-1}(\pi/c_{p,\delta}), 1\}\), \(\pi^{-1} \left( p_0^{-1} \sum_{j \in I_0} 1_{\{|Z_j^0| \leq \tau\}} - 2\Phi(\tau) \right) = o_p(1)\).

**Lemma 5.1** and the condition \(\pi = o(1)\) are enough to show \(A_2 = o_p(1)\).
For the third term $A_3 = s^{-1} \sum_{j \in I_1} 1_{\{|Z_j| > \tau\}} - 1$ in (13), we have

$$P(|A_3| > a) = P(1 - s^{-1} \sum_{j \in I_1} 1_{\{|Z_j| > \tau\}} > a) \leq a^{-1} (1 - P(|Z_1^1| > \tau)) = a^{-1} (G(\tau) - G(-\tau)) = o(1)$$

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

Last but not least, the forth term in (13): $A_4 = -2\bar{\Phi}(\tau) = o(1)$ given $\tau \gg 1$.

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

5.2 Proof of Lemma 5.1

Recall the definitions of $V_{p,\delta}$ in (2) and

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

where $W_1, \ldots, W_p \sim N_p(0, \Sigma)$. Then, for $t = \tau$,

$$P\left(\frac{|p^{-1} \sum_{j=1}^p 1_{\{|W_j| > \tau\}} - 2\bar{\Phi}(\tau)|}{\pi} > c_{p,\delta}\frac{\delta(\tau)}{\pi}\right) < \alpha.$$  

Given $\tau \gg \delta^{-1}(\pi/c_{p,\delta})$, we have $c_{p,\delta}\delta(\tau)/\pi = o(1)$ and, consequently,

$$\frac{|p^{-1} \sum_{j=1}^p 1_{\{|W_j| > \tau\}} - 2\bar{\Phi}(\tau)|}{\pi} = o_p(1).$$  

(14)

Decompose the left hand side above as

$$\pi^{-1} \left|p^{-1} \sum_{j=1}^p 1_{\{|W_j| > \tau\}} - 2\bar{\Phi}(\tau)\right| \geq \pi^{-1} \left|p_0^{-1} \sum_{j \in I_0} 1_{\{|W_j| > \tau\}} - 2\bar{\Phi}(\tau)\right|$$

$$- \pi^{-1} \left|p^{-1} \sum_{j=1}^p 1_{\{|W_j| > \tau\}} - p_0^{-1} \sum_{j \in I_0} 1_{\{|W_j| > \tau\}}\right|$$

24
For the second term on the right hand side,

\[\pi^{-1}\left|p^{-1}\sum_{j=1}^{p} 1_{\{|W_j|>\tau\}} - p_0^{-1}\sum_{j \in I_0} 1_{\{|W_j|>\tau\}}\right| = \pi^{-1}\left|p^{-1}\sum_{j \in I_1} 1_{\{|W_j|>\tau\}} - \pi p_0^{-1}\sum_{j \in I_0} 1_{\{|W_j|>\tau\}}\right| \leq s^{-1}\sum_{j \in I_1} 1_{\{|W_j|>\tau\}} + p_0^{-1}\sum_{j \in I_0} 1_{\{|W_j|>\tau\}}. \quad (15)\]

Since \(\tau \gg 1\), both \(s^{-1}\sum_{j \in I_1} 1_{\{|W_j|>\tau\}} = o_p(1)\) and \(p_0^{-1}\sum_{j \in I_0} 1_{\{|W_j|>\tau\}} = o_p(1)\) by Markov’s inequality. Combining this with (14) and (15) gives

\[\pi^{-1}\left|p_0^{-1}\sum_{j \in I_0} 1_{\{|W_j|>\tau\}} - 2\Phi(\tau)\right| = o_p(1).\]

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 5.1 follows.

### 5.3 Proof of Theorem 2.1

First, we show that given \(\delta(t) = [\Phi(t)]^{\theta}, \theta \in [0, 1/2]\), \(c^*_{p,\delta} = C\sqrt{\bar{\rho}\Sigma_0 (\log p)^{\theta+1/2}}\), with a large enough constant \(C\), satisfies properties (a) \(pc^*_{p,\delta} > p_0c^*_{p_0,\delta}\), and (b) \(P(V^*_{p,\delta} > c^*_{p,\delta}) < \alpha\) for all \(p\).

Consider property (a). Define \(\Sigma_0\) as the covariance matrix of \(W_j, j \in I_0\) and

\[\bar{\rho}_{\Sigma_0} = \sum_{i \in I_0} \sum_{j \in I_0} |\Sigma_{ij}| / p_0^2.\]

It can be shown that

\[\bar{\rho}_\Sigma > \frac{1}{p^2} \sum_{i \in I_0} \sum_{j \in I_0} |\Sigma_{ij}| = \frac{(1-\pi)^2}{p_0^2} \sum_{i \in I_0} \sum_{j \in I_0} |\Sigma_{ij}| = (1-\pi)^2 \bar{\rho}_{\Sigma_0}.\]

Then it follows that

\[c^*_{p,\delta} > C(1-\pi)\sqrt{\bar{\rho}_{\Sigma_0} (\log p)^{\theta+1/2}} > (1-\pi)c^*_{p_0,\delta} = (p_0/p)c^*_{p_0,\delta}.\]

and property (a) is verified.
Next consider property (b). By Chebyshev’s inequality and direct calculation,

\[ P(V_{p,\delta}^* > c_{p,\delta}^*) \leq (c_{p,\delta}^*)^{-2} \text{Var}(V_{p,\delta}^*) \leq (c_{p,\delta}^*)^{-2} E([V_{p,\delta}^*]^2) \]

\[ = (c_{p,\delta}^*)^{-2} E \left[ \max_{t \in T} \left( \frac{|\bar{W}_p(t) - 2\Phi(t)|}{\Phi(t)^\theta} \right)^2 \right] \]

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

\[ E \left( \max_{t \in T} A(t) \right) = \int_0^\infty P(\max_{t \in T} A(t) > c) dc \leq \sum_{t \in T} P(A(t) > c) dc \]

\[ = \sum_{t \in T} E[A(t)] \leq C \sqrt{\log p} \cdot \max_{t \in T} E[A(t)] \]

\[ = C \sqrt{\log p} \cdot \max_{t \in T} \{ [\Phi(t)]^{-2\theta} \text{Var}(\bar{W}_p(t)) \} \]

The following lemma provides the order of \( \text{Var}(\bar{W}_p(t)) \).

**Lemma 5.2** For \( W_1, \ldots, W_p \sim N_p(0, \Sigma) \) and \( \bar{\rho}_\Sigma \) in (7),

\[ \text{Var}(\bar{W}_p(t)) = O \left( \bar{\rho}_\Sigma \cdot e^{-t^2/2} \right). \] (16)

Therefore,

\[ [\Phi(t)]^{-2\theta} \cdot \text{Var}(\bar{W}_p(t)) \leq C[\Phi(t)]^{-2\theta} \cdot \bar{\rho}_\Sigma \cdot e^{-t^2/2} \]

\[ \leq C \left( \frac{t}{e^{-t^2/2}} \right)^{2\theta} \cdot \bar{\rho}_\Sigma \cdot e^{-t^2/2} \leq C(\log p)^\theta \cdot \bar{\rho}_\Sigma \cdot e^{(\theta - 1/2)t^2} \leq C\bar{\rho}_\Sigma \cdot (\log p)^\theta \]

where the first step above is by Lemma 5.2, 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

\[ P(V_{p,\delta}^* > c_{p,\delta}^*) \leq C \left( c_{p,\delta}^* \right)^{-2} \cdot \bar{\rho}_\Sigma \cdot (\log p)^{\theta + 1/2}, \]

and \( c_{p,\delta}^* = O(\sqrt{\bar{\rho}_\Sigma (\log p)^{\theta + 1/2}}) \) follows.

Next, we demonstrate the upper bound property of \( \hat{\pi}_\delta^* \). Based on the general results in
Lemma 2.1, it is sufficient to show
\[ \delta^{-1} \left( \frac{\pi}{c^*_p,\delta} \right) \leq C \Phi^{-1} \left( \frac{\pi^{1/\theta}}{\bar{\rho}_\Sigma^{1/2} (\log p)^{(\theta+1)/2}} \right), \]
which is straightforward given \( \delta(t) = [\Phi(t)]^\theta \).

### 5.4 Proof of Lemma 5.2

\[ \text{Var} \left( \bar{W}_p(t) \right) = p^{-2} \sum_{j=1}^{p} \text{Var}(1_{\{|W_j| > t\}}) + p^{-2} \sum_{i \neq j} \text{Cov}(1_{\{|W_i| > t\}}, 1_{\{|W_j| > t\}}). \]

By Mill’s ratio,
\[ p^{-2} \sum_{j=1}^{p} \text{Var}(1_{\{|W_j| > t\}}) \leq p^{-1} 2\Phi(t)(1 - 2\Phi(t)) \leq Cp^{-1}e^{-t^2/2}. \]

For \( p^{-2} \sum_{i \neq j} \text{Cov}(1_{\{|W_i| > t\}}, 1_{\{|W_j| > t\}}) \), we have
\[ \text{Cov}(1_{\{|W_i| > t\}}, 1_{\{|W_j| > t\}}) = 4 \int_{-\infty}^{t} \int_{-\infty}^{t} f(x,y) dx dy - 4 \int_{-\infty}^{t} \phi(x) dx \int_{-\infty}^{t} \phi(y) dy \leq C|\Sigma_{ij}|e^{-t^2/2}, \]
where the last step follows from Corollary 2.1 in Li and Shao (2002). Combining the above with the definition of \( \bar{\rho}_\Sigma \) results in (16).

### 5.5 Proof of Theorem 2.2

First, it is easy to see that \( c^*_p,\delta = C \sqrt{\log p} \) satisfies property (a) \( pc_p,\delta > p_0 c_{p_0,\delta} \).

For property (b), by Markov’s inequality,
\[ P(V_{p,\delta}^* > c^*_p,\delta) \leq \left( c^*_p,\delta \right)^{-1} \mathbb{E} \left( \max_{t \in T} \left| \bar{W}_p(t) - 2\Phi(t) \right| \frac{\Phi(t)}{[\Phi(t)]^\theta} \right). \]

Let \( B(t) = [\Phi(t)]^{-\theta} |\bar{W}_p(t) - 2\Phi(t)| \), and by the similar arguments as in Section 5.3, we have
\[ \mathbb{E} \left[ \max_{t \in T} B(t) \right] \leq C \sqrt{\log p} \cdot \max_{t \in T} \mathbb{E}[B(t)]. \]
Further, \( E[B(t)] \leq [\Phi(t)]^{-\theta}(E[\tilde{W}_p(t)] + 2\Phi(t)) = 4[\Phi(t)]^{-\theta} \leq 4 \) for \( \theta \in (1/2, 1] \).

Summing up the above, we have

\[
P(V_{p,1}^* > c_{p,1}^*) \leq C(c_{p,1}^*)^{-1} \sqrt{\log p} < \alpha,
\]

where the last step is by \( c_{p,1}^* = C_0 \sqrt{\log p} \) with a large enough constant \( C_0 \).

Next, we demonstrate the upper bound property of \( \hat{\pi}_\delta^* \) with \( \delta(t) = [\Phi(t)]^\theta, \theta \in (1/2, 1] \).

According to Lemma 2.1, it is sufficient to show

\[
\delta^{-1} \left( \frac{\pi}{c_{p,\delta}^*} \right) \leq C\Phi^{-1} \left( \frac{\pi^{1/\theta}}{(\log p)^{1/2\theta}} \right),
\]

which is straightforward given \( \delta(t) = [\Phi(t)]^\theta \).

References

Arias-Castro, E., E. J. Candès, and Y. Plan (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics 39*(5), 2533–2556.

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

Buhlmann, P., K. Kalisch, and L. Meier (2014). High-dimensional statistics with a view toward applications in biology. *Annu Rev Stat Appl. 1*, 255–278.

Cai, T., W. Liu, and Y. Xia (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association 108*(501), 265–277.

Cai, T. and W. Sun (2017). Optimal screening and discovery of sparse signals with appli-
cations to multistage high-throughput studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79(1), 197.

Cai, T. T., J. Jin, M. G. Low, et al. (2007). Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics* 35(6), 2421–2449.

Chen, X. (2019). Uniformly consistently estimating the proportion of false null hypotheses via lebesgue–stieltjes integral equations. *Journal of Multivariate Analysis* 173, 724–744.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics* 35(4), 1351–1377.

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

Finner, H. and V. Gontscharuk (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(5), 1031–1048.

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

Jeng, X. J. and X. Chen (2019). Variable selection via adaptive false negative control in linear regression. *Electronic Journal of Statistics* 13(2), 5306–5333.

Jeng, X. J., Z. J. Daye, W. Lu, and J.-Y. Tzeng (2016). Rare variants association analysis in large-scale sequencing studies at the single locus level. *PLoS computational biology* 12(6), e1004993.

Jeng, X. J. and Y. Hu (2020). Dual control of testing errors in high-dimensional data analysis. *arXiv preprint arXiv:2006.15667*.

Jeng, X. J., T. Zhang, and J.-Y. Tzeng (2019). Efficient signal inclusion with genomic applications. *Journal of the American Statistical Association* 114(528), 1787–1799.
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(3), 461–493.

Jin, J. and T. T. Cai (2007a). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association* 102(478), 495–506.

Jin, J. and T. T. Cai (2007b). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association* 102(478), 495–506.

Li, W. V. and Q.-M. Shao (2002). A normal comparison inequality and its applications. *Probability Theory and Related Fields* 122(4), 494–508.

Meinshausen, N. and J. Rice (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics* 34(1), 373–393.

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

Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics* 31(6), 2013–2035.

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