Published in final edited form as:

J Am Stat Assoc. 2019 ; 114(528): 1787–1799. doi:10.1080/01621459.2018.1518236.

Efficient Signal Inclusion With Genomic Applications

X. Jessie Jeng¹, Teng Zhang¹, Jung-Ying Tzeng^{1,2,3}

¹Department of Statistics, North Carolina State University

²Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan

³Department of Statistics, National Cheng-Kung University, Tainan, Taiwan

Abstract

This paper addresses the challenge of efficiently capturing a high proportion of true signals for subsequent data analyses when sample sizes are relatively limited with respect to data dimension. We propose the signal missing rate as a new measure for false negative control to account for the variability of false negative proportion. Novel data-adaptive procedures are developed to control signal missing rate without incurring many unnecessary false positives under dependence. We justify the efficiency and adaptivity of the proposed methods via theory and simulation. The proposed methods are applied to GWAS on human height to effectively remove irrelevant SNPs while retaining a high proportion of relevant SNPs for subsequent polygenic analysis.

Keywords

Dimension reduction; False negative control; False positive control; Ultrahigh dimension; Variable screening

1 Introduction

High-throughput technology in biology stimulates new challenges in high-dimensional data analysis. For example, recent genomic studies have suggested complex, polygenic bases for complex traits — hundreds of genetic variants are involved in conferring disease risk; individual variants have low effect but variants in aggregate modify disease susceptibility at the gene, pathway or network level. One major goal of the high-throughput biological research, such as genome-wide association studies (GWAS), epigenome-wide association studies (EWAS), and expression quantitative trait locus (eQTL), is to elucidate the joint mechanisms of a genome-wide set of genomic variables on the trait of interests. The exploration of poly-genicity requires sophisticated approaches based on simultaneously analyzing genome-wide variables, e.g., pathway based analysis (Kao et al. (2017) and reference therein), polygenic modeling for assessing SNP main or interaction effects (Waldmann et al. (2013) and reference therein, Wu et al. (2010), Hung et al. (2016)) or for genetic prediction (Abraham et al. (2013) and reference therein). However, due to the high

dimension of genome-wide variables and limited sample size, these simultaneous analyses have to be coupled with a pre-screening step to reduce the data dimension.

Pre-screening can greatly impact subsequent analyses. When individual variants have small effect, pre-screening that is too stringent may fail to capture them for follow-up studies. On the other hand, a pre-screening that is too liberal can hurt the performance of the subsequent simultaneous analyses by including too many noise variables. In current practice, pre-screening on genome-wide variables is often performed by selecting the SNPs with p-values less than an arbitrary threshold (e.g., p-value < 0.001 in Zhou et al. (2011), and < 0.0001 in Wu et al. (2009)). In this work, we aim to develop a data-driven method that is adaptive to the underlying data features so that a high proportion of signals can be selected without incurring many unnecessary noise variables.

One major challenge in developing a data-adaptive method for signal inclusion is how to effectively accommodate the unknown signal information such as signal sparsity and intensity. When signals are much rarer than noise, inference based on signal information is more challenging than inference based on noise distribution. Consequently, retaining signals through false negative control requires different techniques from those used for false positive control. Another challenge for data-adaptive signal inclusion is how to accommodate dependence among variables in real applications. For example, in genomic data analysis, a Single Nucleotide Polymorphism (SNP) is usually strongly correlated with the SNPs nearby due to linkage disequilibrium. The dependence among SNPs can dramatically effect the test statistics and confound inference.

In this paper, we propose a new analytic framework for efficient signal inclusion under dependence. We first discuss sensible criteria for false negative control and propose a new measure called signal missing rate (SMR). Compared to existing measures, SMR assesses the exceedance probability of false negative proportion and incorporates the variability of false negative proportion into inferences.

Next, we develop data-adaptive procedures to control SMR under dependence. The first procedure, conservative SMR (cvSMR), utilizes existing techniques in multiple testing to control false discovery proportion at a stochastic level involving signal information and, consequently, control the measure of SMR at a low level. cvSMR is quite intuitive and easy to implement. However, it tends to be overly conservative for false negative control and can include too many noise variables.

In order to improve the efficiency of signal inclusion, we propose the second method, Adaptive SMR (AdSMR). The main difference between AdSMR and cvSMR is that AdSMR implements a much relaxed critical sequence in its selection rule, which results in a smaller subset of selected variables. The new critical sequence is established by novel theoretical analysis on the variability of false negative proportion through concentration properties of order statistics under dependence. The improved AdSMR procedure and the new analytic techniques guarantee the control of SMR at a degenerating level and the control of unnecessary false positives. Although the implementation of AdSMR does not need signal information, the cut-off position of AdSMR automatically vary with signal sparsity and

intensity, and, as a result, when signal intensity become stronger, both false negative and false positive can be better controlled by AdSMR.

A by-product in the study is a consistent estimator for the number of signals under block dependence that is widely observed in genomic data. Existing studies on signal proportion estimation mainly assume independence (Meinshausen and Rice, 2006; Jin and Cai, 2007). Consistent estimation under dependence is not only useful for signal inclusion as described in this paper, but also valuable in other areas, such as to improve the performance of FDR-based methods in multiple testing.

We compare the finite-sample performances of the proposed AdSMR method and existing methods in simulation. The simulation settings include different sparsity and intensity levels of signals, block dependence with various block sizes, sparse dependence without block structures, and dependence structure from a multi-factor model. While all the methods seem to be effective in false negative control, AdSMR generally outperforms other methods in incurring less false positives.

We apply AdSMR to a GWA analysis on human height using the CoLaus data, in which all 340, 359 autosomal SNPs explain 53.7% of the phenotypic variability of human height. Multiple testing based on the full set of SNPs identifies zero significant candidates because individual variants have small effects. In order to significantly reduce data dimension and carry as many relevant SNPs to subsequent polygenic analyses, we apply AdSMR and select only a small proportion (0.021) of the total SNPs, which explains nearly all the 53.7% of the height variation attained using the full set of SNPs. We further apply penalized regression on the SNPs selected by AdSMR and narrow down the number of selected SNPs to 1, 563. The estimated heritability is still close to 53.7% based on only the 1, 563 SNPs. The selected subset would include a high proportion of truly relevant SNPs for further downstream analyses such as gene annotation, pathway mapping, polygenic risk score, etc.

The rest of the paper is organized as follows. Section 2 introduces the signal missing rate and develop procedures for SMR control under dependence. Consistent estimation of signal proportion is also discussed. Section 3 demonstrates the finite-sample performance of the proposed methods in simulation. Comparisons with other methods are provided. Section 4 presents an application of our method to genomic data analysis. Concluding remarks are provided in Section 5.

2 Efficient Signal Inclusion Under Dependence

2.1 Signal Missing Rate

We first discuss sensible measures for false negatives. Table 1 summarizes notations in classification of variables where TP, FN, FP, and TN are numbers of true positives, false negatives, false positives, and true negatives, respectively; s is the total number of signal variables. Our goal of signal inclusion corresponds to seeking low FN/s .

In the multiple testing literature, False Nondiscovery Rate (FNR) has been proposed as an analogue of False Discovery Rate (FDR) to measure false negatives. It is defined as the

expectation of the proportion of false negatives among the unselected variables, $E(FN/(m - R))$ (Genovese and Wasserman, 2004; Sarkar, 2006). The notion of FNR, unfortunately, does not suit our need for signal inclusion, because FNR is mostly very close to zero when signals are sparse, and a large (or small) FNR does not correspond to a high (or low) FN/s . We provide a simulation example to illustrate this point in Supplementary Material.

Recently, Cai and Sun (2017) constructed the Missed Discovery Rate (MDR) for false negative control. MDR is the expected value of false negative proportion, $E(FN/s)$. To utilize MDR for false negative control, all the variables are ranked by their estimated local FDR values $(\widehat{Lfd_r})$. Because $E(FN)$ can be approximated by $\sum (1 - \widehat{Lfd_r})$, given an estimate for s , a cut-off position on the ranked local FDR values can be determined to control MDR at a pre-fixed level. MDR control is intuitive and easy to implement. However, the measure of MDR does not consider the variability of false negative proportion; and the control of MDR has not been studied under dependence.

In this paper, we propose a new measure for false negative control called Signal Missing Rate (SMR). SMR is defined as

$$SMR^\epsilon = P(FN/s > \epsilon), \quad (2.1)$$

where $\epsilon > 0$ is a constant between 0 and 1. Signal missing rate evaluates the probability of neglecting at least a certain proportion of signals. By controlling SMR at a low level with a small ϵ , a high proportion of signals can be captured. Compared to MDR, SMR measures the exceedance probability of false negative proportion and incorporates the variability of FN/s into inference.

2.2 Controlling SMR under dependence

To assure generality of our work, we do not assume any specific distribution for the test statistic. Specifically, define I_0 and I_1 as collections of indices of the noise and signal variables, respectively. Let P_j be the p -value of the j th variable. Assume

$$P_j \sim U \cdot 1\{j \in I_0\} + G \cdot 1\{j \in I_1\}, \quad j = 1, \dots, m, \quad (2.2)$$

where U represents the cumulative distribution function of the uniform distribution at $[0,1]$ and G is some unknown cumulative distribution function dominating U , i.e., $G(t) > U(t)$ for all $t \in (0,1)$. This mixture model on p -values in (2.2) provides a convenient framework for large-scale inference. It can be used in a wide range of applications as long as the baseline distribution of the noise can be reasonably estimated from either asymptotic or empirical approaches, such as by permutation or parametric bootstrap.

We define the signal missing rate of a procedure selecting the top k candidates along the ranked p -values as

$$SMR^\epsilon(k) = P(FN(k)/s > \epsilon),$$

where $\epsilon \in (0,1)$ is a constant and $FN(k)$ represents the number of false negatives for selecting the top k candidates, which equals to the number of true signals ranked after k .

We develop two procedures for SMR control. Both procedures are easy to implement in applications. The first procedure is more in line with existing techniques in multiple testing. The second procedure improves the efficiency of the first approach by developing new analytic techniques based on concentration inequalities of order statistics under dependence.

2.2.1 The conservative SMR procedure—Suppose that we know the number of signals s , then a procedure controlling $FDP(= FP/R)$ at the level of $(R - s)/R$ includes the number of signals as $TP = R - FP = R - FDP \times R - s$. Therefore, one can modify a method that controls FDP at the level of $(R - s)/R$ to include a high proportion of signals. Motivated by this idea, we develop the conservative SMR (cvSMR), a procedure that determines the cut-off position on the ranked p -values $p_{(1)}, \dots, p_{(m)}$ by

$$k_{cv}^* = \hat{s} + \min \left\{ j \geq 1 : p_{(\hat{s} + j)} \leq \frac{j}{m} \alpha \right\} 1 \{ \hat{s} > t_1 \}, \quad (2.3)$$

where $\hat{s}$ is an estimate for the number of signals, α is a prefixed small constant, and $t_1 = \max \{ j : p_{(j)} < \alpha_m/m \}$ with $\alpha_m = \alpha(1)$. The top $\{1, \dots, k_{cv}^*\}$ candidates are selected.

cvSMR is a step-down procedure with critical sequence $\alpha_j = (j/m)\alpha$ that is frequently used in methods controlling FDR or FDP. Compared to an existing step-down procedure studied in Lehmann and Romano (2005), cvSMR uses an opposite sign when comparing p -value with the critical sequence. This is because the procedure starts at the position $\hat{s}$, where FDP is not controlled in general, and stops as soon as FDP is controlled at a desirable level. On the other hand, the procedure in Lehmann and Romano (2005) starts from the first position where FDP is controlled and stops once FDP cannot be controlled. cvSMR also differs in the index of α_j , where j is not the index of the ordered p -values (k), but $k - \hat{s}$. It can be proved that the step-down procedure of cvSMR controls FDP at a stochastic level of $(k_{cv}^* - \hat{s})/k_{cv}^*$. Consequently, the number of true signals included in the top k_{cv}^* variables is greater than $\hat{s}$ with high probability.

The following theoretical results show that cvSMR asymptotically controls SMR at the level of α . $P_1^0, \dots, P_{m-s}^0$ denote the p -values corresponding to the $m - s$ noise variables and $P_1^1, \dots, P_s^1$ denote the p -values corresponding to the s signals. We consider the same dependence condition as in Lehmann and Romano (2005): for any $j = 1, \dots, m - s$,

$$P(P_j^0 \leq u | P_1^1, \dots, P_s^1) \leq u. \quad (2.4)$$

This condition says that the p -value of a noise variable is conditionally dominated by a uniform distribution. This condition allows arbitrary joint dependence within noise variables and within signal variables.

Proposition 2.1 Consider model (2.2) under condition (2.4). Given a consistent estimator $\hat{s}$ for the number of signals and a constant α for the SMR control level, cvSMR asymptotically controls SMR^ϵ at the level of α for any $\epsilon > 0$, i.e.,

$$SMR^\epsilon(k_{cv}^*) \leq \alpha + \Delta_m, \quad (2.5)$$

where $\Delta_m = \alpha(1)$ for any $\epsilon > 0$.

Note that $\Delta_m = \alpha(1)$ as long as ϵ is a constant. If $\epsilon = \epsilon_m \rightarrow 0$, the asymptotic control on SMR by cvSMR may not hold. This follows our intuition that the control of false negatives is harder when smaller number of false negatives is allowed.

Generally speaking, cvSMR would work well in situations where FDR/FDP methods work well. However, cvSMR inherits the same issue as FDR/FDP methods and tends to be conservative under high dimensionality. Here, the conservativeness is on false negative control, which means that too many variables could be selected. In fact, this disadvantage can be more severe for cvSMR as the event $\{p_{(\hat{s}+j)} \leq (j/m)\alpha\}$ is less likely to happen than the event $\{p_{(j)} \leq (j/m)\alpha\}$ in FDR/FDP methods. Table 2 presents the number of selected candidates in a simulation example with $m = 5000$ and $s = 250$. The test statistics are generated from multivariate normal distribution $N(\mathbf{A}, \mathbf{\Sigma})$, where $A_j = 0$ for noise and $A_j = \mu > 0$ for signal. The covariance matrix $\mathbf{\Sigma}$ is a block-diagonal matrix with equal block size $l = 50$ and within-block correlation $\rho = 0.7$. cvSMR with $\alpha = 0.1$ appears too conservative in this example by selecting almost all variables.

Because cvSMR controls SMR^ϵ for arbitrarily small constant $\epsilon > 0$, one way to mitigate the conservativeness of cvSMR is to weaken the control of false negatives and allow for a fixed proportion of false negatives. Consequently, the critical sequence $\alpha_j = (j/m)\alpha$ can be relaxed by involving the fixed proportion. Considering our motivation for retaining as many true signals as possible for subsequent data analysis, we would like to propose a different strategy to significantly reduce the number of false positives without weakening the theoretical control on false negatives.

2.2.2 The Adaptive SMR Procedure—In order to develop a method that incurs less false positives and is more applicable in Big Data applications, we propose the second procedure, the Adaptive SMR (AdSMR). AdSMR has the cut-off position on the ranked p -values as

$$k^* = \hat{s} + \min\left\{j \geq 1: p_{(\hat{s}+j)} \leq b_j\right\} 1\{\hat{s} > t_1\}. \quad (2.6)$$

Similar to (2.3), $\hat{s}$ is an estimate of the number of signals and $t_1 = \max\{j: p_{(j)} < \alpha_m/m\}$ with $\alpha_m = \alpha(1)$. The key difference between (2.3) and (2.6) is that AdSMR has the critical sequence b_j defined as the median of $Beta(j, m - \hat{s} - j + 1)$. The rationale to use beta distribution to determine b_j is because our proposed method is based on ranked p -values, and the j -th ranked p -value of $m - \hat{s}$ noise variables follows $Beta(j, m - \hat{s} - j + 1)$ under independence. Therefore, it is natural to utilize $Beta(j, m - \hat{s} - j + 1)$ to perform inference.

Although the median of a beta distribution does not have an explicit form, it is known that the median is bounded by the mode and mean of beta distribution. Therefore,

$$\frac{j-1}{m-\hat{s}-1} < b_j < \frac{j}{m-\hat{s}+1}, \quad (2.7)$$

and b_j is approximately α^{-1} times as large as α_j of cvSMR. Larger b_j results in less variables being selected as shown in Table 2.

To justify this new procedure in theory, existing techniques for FDP control cannot be used anymore. We develop novel techniques to analyze the procedure based on concentration properties of the ordered p -values under dependence. Figure 1 illustrates a sequence of ordered p -values, where T_1 denotes the location before the first noise variable and T_2 denotes the location of the last signal variable. Signals and noise are mixed indistinguishably between T_1 and T_2 . We show that the proposed AdSMR method is able to capture a high proportion of signals ranked before T_2 and, at the same time, avoid unnecessary false positives ranked after T_2 .

Let L be the total number of noise variables ranked before T_2 , then $L = T_2 - s$. Note that L is a random variable varying from sample to sample. Generally speaking, lower signal intensity results in larger L . The specific relationship depends on the model that generates the data. In this section, we assume that L is bounded almost surely by a number $\bar{l}$, and

$$1 \ll \bar{l} \log(\bar{l}) \ll \min(s^2, \sqrt{ms}). \quad (2.8)$$

Condition (2.8) says that the number of indistinguishable noise is not too large. For example, in an association study with 500,000 total variables and 100 truly associated signals, we request $\bar{l} \log(\bar{l}) \ll 7,071$. This condition is fairly general as it allows the existence of weak signals that may rank after many noise variables and “pseudo” signals which are indeed noise but show up as signals due to high dependence with true signals. We note that the number of such “pseudo” signals is often much smaller than the total number of variables in applications when true signals are sparse.

Our next condition is on dependence in the data. Let $P_{(1)}^0, \dots, P_{(m-s)}^0$ denote the ordered p -values corresponding to the $m-s$ noise variables. Assume that, for any $r = 1, \dots, \bar{l}$,

$$P(P_{(r)}^0 \leq u | P_1^1, \dots, P_s^1) \leq c_1 F_r(u), \quad (2.9)$$

where $F_r(\cdot)$ is the counterpart of the left side probability for independent p -values, and $c_1 - 1$ is some constant. Since order statistics of independent noise p -values follow beta distribution, $F_r(\cdot)$ is the cumulative distribution function of $Beta(\nu_1, \nu_2)$ with $\nu_1 = r$ and $\nu_2 = m-s-r+1$. This condition essentially says that the $\bar{l}$ smallest noise p -values (given the signal p -values) can be more or less extreme than their counterparts under independence as long as c_1 is a constant no less than 1. There are no constraints on the noise p -values ranked after $\bar{l}$ or the p -values of signal variables.

The following theorem shows that given the above conditions, AdSMR has a degenerating SMR, which is equivalent to say that the FN proportion/sensitivity of AdSMR converges to 0/1 in probability.

Theorem 2.2 Consider model (2.2) under conditions (2.8) and (2.9). Given a consistent estimator $\hat{s}$ for the number of signals, AdSMR has a degenerating SMR^ϵ for any $\epsilon > 0$, i.e.,

$$SMR^\epsilon(k^*) \rightarrow 0 \quad (2.10)$$

as $m \rightarrow \infty$ for any constant $\epsilon > 0$.

Comparing Theorem 2.2 with Proposition 2.1, it can be seen that AdSMR and cvSMR control SMR differently. cvSMR asymptotically controls SMR at a prefixed level α , whereas AdSMR controls at a degenerating level. The theoretical justification coupled with the more relaxed critical sequence make AdSMR a more efficient method for false negative control.

The asymptotic result in Theorem 2.2 holds for any constant $\epsilon > 0$ but may not hold for $\epsilon \rightarrow 0$. In other words, AdSMR may allow a number of false negatives as long as the number is not greater than a proportion of the total number of signals.

Next, we show that AdSMR can avoid selecting unnecessary false positives. Recall the locations of T_1 and T_2 in Figure 1. Although signal and noise variables mix indistinguishably between T_1 and T_2 , noise variables ranked after T_2 should be avoided. The next theorem shows that under suitable conditions on the dependence and the estimator $\hat{s}$, AdSMR controls the selection of noise variables ranked after T_2 .

Theorem 2.3 Consider model (2.2). Define T_2 and L as in Figure 1. Assume that $L \gg 1$ with probability tending to 1, and for some small constant $\delta > 0$,

$$P(P_{(L+k)}^0 > u_k, k = 1, \dots, \delta L | P_1^1, \dots, P_s^1) \leq c_2 H(u_k, k = 1, \dots, \delta L), \quad (2.11)$$

where $H(\cdot)$ is the counterpart of the left side probability for independent p -values, and $c_2 \geq 1$ is some constant. Then, AdSMR with $\hat{s}$ satisfying $P(\hat{s} < s) \rightarrow 1$ has

$$P(k^* > (1 + \delta)T_2) \rightarrow 0. \quad (2.12)$$

Condition (2.11) is quite general as c_2 is an arbitrary constant no less than 1. Theorem 2.2 and 2.3 imply that AdSMR achieves both SMR control and false positive control when (a) dependence conditions in (2.9) and (2.11) are satisfied, (b) signal intensity is strong enough so that condition (2.8) is satisfied, and (c) $P(1 - \delta) < \hat{s} < s \rightarrow 1$ for arbitrarily small constant $\delta > 0$. The $\hat{s}$ estimator studied in the next section has the property in (c).

2.3 AdSMR with MR estimator

An important component of the SMR control-based methods is a consistent estimator for the number of signals. Estimation of signal proportion among all variables ($\pi = s/m$) has inspired profound research in high-dimensional inference. For example, Storey (2002), Genovese and Wasserman (2004), and Jin and Cai (2007) have developed consistent

estimators for relatively dense signals with proportion $\gg m^{-1/2}$; Cai et al. (2007) has considered sparse signals with proportion $\ll m^{-1/2}$; and Meinshausen and Rice (2006) has considered both dense and sparse signals for proportion estimation. However, existing studies mainly focus on independent variables. Rigorous analysis under realistic dependence structure is scarce.

In genomic data analysis, the covariance matrix of p -values of SNPs often exhibits a block structure. In Figure 2, we present heatmap of the absolute values of sample correlations of the first 50 SNPs in Chromosome 1 from Cohorte Lausannoise (CoLaus) study samples. Details of the real data are described in Section 4. The heatmap shows blocks of high correlations along the diagonal regions with different block sizes. Such dependence structures are frequently observed in genomic data (Efron, 2007; Fan et al., 2012).

In this section, we study the consistency of the estimator developed in Meinshausen and Rice (2006) in the situation where variables have block dependence. We refer to this estimator as the MR estimator which is constructed as follows. Define the empirical distribution of p -values

$$F_m(t) = \frac{1}{m} \sum_{j=1}^m 1\{P_j \leq t\} = (1 - \pi)U_{m_0}(t) + \pi G_s(t), \quad (2.13)$$

where U_{m_0} and G_s denote the empirical distributions of the p -values corresponding to $m_0 (= m - s)$ noise variables and s signals, respectively. Similarly, denote U_m as the empirical distributions of the p -values when all m candidates are noise. Define

$$V_m = \sup_{t \in (0, 1)} \frac{U_m(t) - t}{\sqrt{t(1-t)}}, \quad (2.14)$$

and denote c_m as a bounding sequence of V_m such that mc_m is monotonically increasing with m and $P(V_m > c_m) = \alpha_m \rightarrow 0$. Then, the MR estimator for π is constructed as

$$\hat{\pi}_{MR} = \sup_{t \in (0, 1)} \frac{F_m(t) - t - c_m \sqrt{t(1-t)}}{1-t}. \quad (2.15)$$

Consistency of $\hat{\pi}_{MR}$ has been shown for independent p -values in Meinshausen and Rice (2006). In this section, we study the consistency of $\hat{\pi}_{MR}$ under block dependence and implement $\hat{\pi}_{MR}$ into AdSMR. Assume that $\{P_j\}_{j=1}^m$ can be divided into independent groups with arbitrary dependence in each group. Denote l as the upper bound of the group sizes and assume

$$l = O(m^\kappa) \quad \text{for some constant } \kappa \in [0, 1). \quad (2.16)$$

The next theorem summarizes the conditions for $\hat{\pi}_{MR}$ consistency and the SMR control of AdSMR with MR estimator.

Theorem 2.4 Consider model (2.2) under condition (2.9) and block dependence in (2.16). Let $\pi = m^{-\eta}$ for some $\eta \in [0, 1)$. Assume either one of the following conditions:

- (i) $\eta \in [0, (1 - \kappa)/2)$, $\inf_{t \in (0, 1)} G(t) = 0$, and $1 \ll \bar{l} \log(\bar{l}) \ll m^{1 - \eta/2}$.
- (ii) $\eta \in [(1 - \kappa)/2, 2/3)$, $G(m^{-\tau}) \rightarrow 1$ for some $\tau > 2\eta - (1 - \kappa)$, and $1 \ll \bar{l} \log(\bar{l}) \ll m^{1 - \eta/2}$.
- (iii) $\eta \in [2/3, 1)$, $G(m^{-\tau}) \rightarrow 1$ for some $\tau > 2\eta - (1 - \kappa)$, and $1 \ll \bar{l} \log(\bar{l}) \ll m^{2(1 - \eta)}$.

Then $P(1 - \delta < \hat{\pi}_{MR}/\pi < 1) \rightarrow 1$ as $m \rightarrow \infty$ for arbitrarily small constant $\delta > 0$ and AdSMR with $\hat{s} = m\hat{\pi}_{MR}$ has $SMR^\epsilon(k^*) \rightarrow 0$ as $m \rightarrow \infty$ for any constant $\epsilon > 0$.

Theorem 2.4 considers the scenario where the same dataset is used to derive $\hat{\pi}_{MR}$ and to retain signals by AdSMR. Therefore, the SMR control of AdSMR would be affected by the conditions for the consistency of $\hat{\pi}_{MR}$. Given the general block structure of dependence, condition (2.9) implies constraints on the block size and within-block dependence. Under the constraints of dependence, conditions in (i) - (iii) can be satisfied if signal intensity is strong enough in different ranges of sparsity level.

It can be seen that the more sparse the signals (larger η), the stronger the condition on signal intensity. When $\eta \in [0, (1 - \kappa)/2)$, the condition $\inf_{t \in (0, 1)} G(t) = 0$ is quite general and has been described as the “pure” case in Genovese and Wasserman (2004). When $\eta \in [(1 - \kappa)/2, 1)$, the stronger condition $G(m^{-\tau}) \rightarrow 1$ says that the distribution of signal p -values is highly concentrated around 0. One can also see the effect of block size demonstrated through κ . Generally speaking, the larger the κ , the stronger the conditions in (i) - (iii).

AdSMR with MR estimator also controls unnecessary false positives ranked after T_2 as presented in the following corollary. The proof duplicates parts of the proofs for Theorem 2.3 and 2.4, and thus are omitted.

Corollary 2.5 Assume conditions in Theorem 2.3 and block dependence in (2.16). AdSMR with $\hat{s} = m\hat{\pi}_{MR}$ has $P(k^* > (1 + \delta)T_2) \rightarrow 0$ for arbitrarily small constant $\delta > 0$.

The next corollary shows that AdSMR with MR estimator has asymptotically zero false positives in two special scenarios.

Corollary 2.6 Assume model (2.2) with block dependence in (2.16). Consider two scenarios: (i) there is no signals exists, i.e. $s = 0$; and (ii) signals are strong enough such that $s\bar{G}(m^{-\tau_0}) = o(1)$. In both scenarios, AdSMR with $\hat{s} = m\hat{\pi}_{MR}$ has $P(FP(k^*) > 0) = o(1)$.

We conclude this section by a complete algorithm for AdSMR with the MR estimator. For simplicity, the same α_m is used to simulate the bounding sequence c_m and to obtain t_1 in AdSMR. More specifically, one can simulation V_m from the empirical null distribution. Then, c_m can be determined as the $(1 - \alpha_m)$ th quantile of the empirical distribution of V_m .

from 1000 simulations. To save computation, we approximate b_j by $j/(m - \hat{s})$ as shown in (2.7) and set an upper limit for k^* at $\lfloor m/2 \rfloor$. The step-by-step algorithm is as follows. A toy example demonstrating the algorithm is provided in Appendix 5.4.

Algorithm 1:

AdSMR with MR estimator

-
1. Simulate the bounding sequence c_m from the empirical null distribution of V_m with $\alpha_m = 1/\sqrt{\log m}$.
 2. Obtain $\hat{\pi}_{MR}$ by (2.15) using the bounding sequence c_m .
 3. Sort the observed p -values as $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$.
 4. Calculate the cut-off position by k^* in (2.6) with $\hat{s} = m\hat{\pi}_{MR}$, $\alpha_m = 1/\sqrt{\log m}$ and $b_j = j/(m - \hat{s})$. Set an upper limit for k^* at $\lfloor m/2 \rfloor$.
 5. Select the top ranked candidates with p -values $p_{(1)}, \dots, p_{(k^*)}$.
-

3 Simulation Study

We compare the finite-sample performances of AdSMR and existing methods for false negative control. These methods include the MDR procedure (Cai and Sun, 2017) and BH-FDR with high nominal levels. For fair comparison, both AdSMR and MDR use $\hat{\pi}_{MR}$ for proportion estimation. MDR aims to control the expectation of FN proportion through estimating local FDR. We use the software “locfdr” to estimate local FDR and apply MDR at the recommended level $1/\log(m)$. BH-FDR with high nominal levels are ad-hoc procedures that apply the original FDR method in Benjamini and Hochberg (1995) with high nominal levels 0.5 and 0.7 to capture more signals.

We demonstrate FN control of these methods by reporting their FN proportions (FN/s). Then, we show their efficiency by reporting the false discovery proportion ($FDP = FP/R$) of these methods. Higher FDP can be viewed as higher price paid to achieve low FN proportion. In addition, we employ the F-measure as a summary metric for FN proportion and FDP (Powers, 2011). By definition, F-measure is the harmonic mean of precision ($1 - FP/R$) and recall ($1 - FN/s$) and calculated by $2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$.

We simulate test statistics from multivariate normal distribution $N(\mathbf{A}, \mathbf{\Sigma})$, where $A_j = 0$ for noise and $A_j = \mu > 0$ for signal. The locations of the signals are selected randomly. We consider settings with different signal sparsity and intensity levels and various dependence.

Example 1 has data dimension $m = 5000$. The covariance matrix $\mathbf{\Sigma}$ is a block diagonal matrix with equal block size $l = 50$ and within-block correlation $\rho = 0.7$. The diagonal elements of $\mathbf{\Sigma}$ are set to be 1. We demonstrate different signal sparsity levels: $\pi = 0.02$ and 0.1. Signal intensity μ increases from 3 to 5.5. Figure 3 presents the median values of FN proportion, FDP, and the F-measure for all the methods from 100 simulations. It shows that the FN proportion of AdSMR is lower for signals with $\pi = 0.1$ than for signals with $\pi = 0.02$. This agrees with the theoretical insights of Theorem 2.4 where stronger condition on signal intensity is required for more sparse signals. Compared to other methods, AdSMR has slightly larger FN proportion but smaller FDP. AdSMR generally outperforms other methods

in F-measure especially for sparse signals with $\pi = 0.02$. We follow the convention in high-dimensional screening literatures, e.g., Fan and Lv (2008), and report the median of these measures. Examples of the mean and standard deviation of the measures are provided in Supplementary Material.

Example 2 has block dependence with larger block size $l = 200$. Figure 4 compares all the methods under different signal sparsity and intensity levels. Compared to Example 1, the performance of AdSMR deteriorates a little for FN control, which agree with the theoretical insights in Theorem 2.4 on the effect of block size. Overall, AdSMR still mostly outperforms other methods in FDP control and F-measure.

Example 3 simulates data with $m = 1000$ and a sparse Σ whose nonzero elements are randomly located. The data generation process is similar to Model 3 in Cai et al. (2013). Let $\Sigma^* = (\sigma_{ij})$, where $\sigma_{ii} = 1$, $\sigma_{ij} = 0.7 * \text{Bernoulli}(1, 0.1)$ for $i < j$ and $\sigma_{ji} = \sigma_{ij}$. Then $\Sigma = \mathbf{I}^{1/2}(\Sigma^* + \delta \mathbf{I})(\mathbf{I} + \delta \mathbf{I})^{1/2}$, where $\delta = |\lambda_{\min}(\Sigma^*)| + 0.05$. Figure 5 shows that the performance of AdSMR is comparable to its performance in Example 1, although the covariance matrix in this example does not have a block structure. MDR performs relatively better in this example than in previous examples. AdSMR and MDR generally outperform high level BH-FDR procedures in this example.

Example 4 considers dependence structure from a two-factor model as in Fan et al. (2012). Let Σ be the correlation matrix of a random sample with size 100 of m -dimensional vector $\mathbf{X} = (X_1, \dots, X_m)$, where $X_j = \rho_j^{(1)} W^{(1)} + \rho_j^{(2)} W^{(2)} + H_j$, $W^{(1)}$ and $W^{(2)}$ are iid $N(0, 1)$, $\rho_j^{(1)}$ and $\rho_j^{(2)}$ are iid $U(-1, 1)$, and H_j are iid $N(0, 1)$. Figure 6 shows that the performance of AdSMR is comparable to its performance in Example 2, where Σ has large diagonal blocks. MDR and high level BH-FDR procedures perform relatively better in this example than in the examples with block dependence.

In all the examples, the cut-off position of AdSMR automatically vary with signal sparsity and intensity and, as a result, when signal intensity becomes stronger and the noise and signal p -values are better separated, AdSMR controls both false negative and false positive better. BH-FDR with high nominal levels do not have such property as their FDPs remain high with increasing signal intensity. MDR method performs better than high level BH-FDR in terms of adapting to signal intensity but worse than AdSMR when signals are more sparse. Additional simulation for the empirical SMR of AdSMR is presented in Appendix 5.5. Finite-sample performance of the MR estimator along is demonstrated in Supplementary Material.

4 Real Data Analysis

Recent heritability analyses of GWAS data have suggested that a large proportion of variation of human height can be explained by all autosomal SNPs although only a small proportion of associated variants have been successfully identified. For example, Yang et al. (2011) showed that about 45% of the variation can be accounted for by common SNPs from a sample of around 4000 Australians with ancestry in the British Isles. In the analysis of Kostem and Eskin (2013), 62% of the variation can be explained by all autosomal SNPs

from the Northern Finland Birth Cohort of 1966 from 5319 unrelated individuals. However, the discovered associated SNPs have only explained a small proportion of the total heritability. For example, in Allen et al. (2010), only 10.5% of the variance can be explained by the 180 SNPs with genome-wide significance. It indicates that there exist a large number of weak signals, for which current methods have limited power even with a sample size of $10^3 \sim 10^4$.

We obtained the GWAS height data from the Cohorte Lausannoise (CoLaus) study (Firmann et al., 2008). The dataset includes 1874 subjects with available information on age, sex, and 340359 autosomal SNPs with minor allele frequencies greater than 1%. We calculated the p -values corresponding to each SNP by fitting marginal linear models while adjusting 12 clinical covariates including sex, age, and the top 10 principal components. The FDR method in Benjamini and Hochberg (1995) cannot identify any associated SNPs based on the full set of SNPs due to small effects of individual SNPs. On the other hand, the Lasso procedure (Tibshirani, 1996) applied to the full dataset identifies only two candidate SNPs.

In order to reduce data dimension and carry as many true signals to subsequent analyses, we apply both AdSMR and MDR procedures. To implement the MR estimator to AdSMR, we first generate a set of p -value sequences under the global null using the permutation approach introduced in Westfall and Young (1993). In each permutation, the phenotype values are randomly shuffled and reassigned to individuals, the null p -values are calculated in each permutation with the correlation structure of the SNPs preserved. We replicate the permutation process 1000 times. Based on the simulated null p -values, the MR estimate and the cut-off position of AdSMR are calculated by the algorithm in Section 2.3. The estimated number of associated SNPs is equal to 743; AdSMR selects 7,204 SNPs, while MDR selects 110,778 SNPs.

To examine the contribution of the selected SNPs, we perform heritability analysis using GCTA package (Yang et al. (2011)). First, we use GCTA to estimate the genetic relationship matrix (GRM) of all the GWAS SNPs and then fit a random effects model to estimate the proportion of variance explained by all the autosomal SNPs. We found that 53.7% of the phenotypic variance in height can be explained by all the autosomal SNPs. We then repeat the analysis on the partitioned SNP sets (e.g., the 7,204 SNPs identified by AdSMR vs. the rest) to estimate the proportion of variance in height explained by the selected SNPs while adjusting for the remaining unselected SNPs. Note that the random effects model in GCTA is very different from the marginal linear regression model that we used to select SNPs. In addition, the heritability percentage is also different from the R^2 statistic of goodness-of-fit test for prediction.

Table 3 shows that AdSMR reduces the number of SNPs from 340,359 to 7,204, but the explained variability of the selected SNPs is still 53.7%, suggesting that almost all important SNPs to human height are retained by AdSMR for this dataset. MDR also explained 53.7% of the height variation, but appears less efficient than AdSMR by selecting a much larger set of SNPs.

With the reduced set of SNPs from AdSMR, we can apply joint modeling using Lasso and further narrow down the number of SNPs to 1,563. By repeating the above GCTA analysis on the 1,563 selected SNPs vs. the rest, the estimated heritability is still 53.7%. These selected SNPs provide promising candidates for further downstream analyses such as gene annotation, pathway mapping, polygenic risk score, etc.

5 Conclusion and Further Discussion

In this paper, we consider the problem of efficient signal inclusion under dependence. Our motivation comes from Big Data applications where sample sizes are relatively limited with respect to data dimension, and there are great needs to significantly reduce data dimension while retaining as many true signals as possible for subsequent analyses. However, in applications where signals are much rarer than noise, inference for false negative control based on signal information requires different techniques from those used for false positive control. Furthermore, the dependence among variables, especially between noise and signal variables, adds another layer of difficulty. To address these challenges, we develop data-adaptive methods whose implementations do not need the signal information. Nevertheless, the cut-off position of the proposed AdSMR procedure automatically vary with signal sparsity and intensity and, as a result, a high proportion of signals can be selected without incurring many unnecessary false positives. When signal intensity becomes stronger and the noise and signal p -values are better separated, the cut-off position of AdSMR controls both false negative and false positive better. These properties are presented in Theorem 2.2 – 2.4 and Corollary 2.5 – 2.6, and illustrated in simulation examples where signal intensity increases in each setting.

We have also proposed a new measure, SMR, for false negative control. The notion of SMR includes two parameters: the FN proportion ϵ and the SMR control level α . The methods developed in the paper, cvSMR and AdSMR, control SMR at different levels. The construction of cvSMR is more in line with the existing techniques based on marginal distribution of p -values, whereas AdSMR benefits by new explorations on the concentration properties of the ordered p -values under dependence. Both methods control SMR for arbitrarily small ϵ . However, AdSMR is shown to incur much less false positives.

We note that false positives can be further reduced by allowing a small fixed FN proportion. Such extension on AdSMR would involve more delicate analyses on the concentration properties of order statistics under dependence. The estimator $\hat{s}$ implemented in cvSMR and AdSMR is another subject to be re-investigated under the new request. Detailed studies are deferred to future research. Another interesting topic for future research is SMR control for specific data generating models. While the current paper considers a general p -value model without model assumptions on test statistics, it will be interesting to relate conditions in the paper to parameters of the specific models. We expect that the characterizations would be very different for models with sparse or dense signal component.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

The authors are grateful to Dr. Peter Vollenweider and Dr. Gerard Waeber, PIs of the CoLaus study, and Dr. Meg Ehm and Dr. Matthew Nelson, collaborators at GlaxoSmithKline, for providing the CoLaus phenotype and genetic data. The authors appreciate the very helpful comments and suggestions from an associate editor and three reviewers. Dr. Jeng was partially supported by National Human Genome Research Institute of the National Institute of Health under grant R03HG008642. Dr. Tzeng was partially supported by National Institutes of Health grant P01CA142538.

Appendix

Appendix 5.1 – 5.3 provide proofs of Proposition 2.1 and Theorem 2.2. A toy example demonstrating the AdSMR algorithm is shown in Appendix 5.4, and additional simulation results are presented in Appendix 5.5. Proofs of Theorem 2.3, Theorem 2.4, and Corollary 2.6 and more simulation results are provided in Supplementary Material.

5.1 Proof of Proposition 2.1

For notation simplicity, let $j^* = \min\{j \geq 1 : P(\hat{s}_{+j}) \leq \alpha_j\}$, where $\alpha_j = (j/m)\alpha$. Then $k_{cv}^* = \hat{s} + j^*$. Denote $TP(k)$ and $FP(k)$ as the numbers of true positives and false positives in the top $\{1, \dots, k\}$ candidates, then $k_{cv}^* = TP(k_{cv}^*) + FP(k_{cv}^*)$. We also have $s = TP(k_{cv}^*) + FN(k_{cv}^*)$. Now, for any $\epsilon > 0$,

$$\begin{aligned} SMR^\epsilon(k_{cv}^*) &= P(FN(k_{cv}^*)/s > \epsilon) = P(TP(k_{cv}^*) < s - \epsilon s) = P(FP(k_{cv}^*) > k_{cv}^* - (1 - \epsilon)s) \\ &\leq P(FP(k_{cv}^*) > \hat{s} + j^* - (1 - \epsilon)s, \hat{s} \geq (1 - \epsilon)s) + P(\hat{s} < (1 - \epsilon)s) \\ &\leq P(FP(k_{cv}^*) > j^*) + o(1), \end{aligned} \quad (5.1)$$

where the last step is by the consistency of $\hat{s}$.

In the case of $\hat{s} \leq t_1$, we have $k_{cv}^* = \hat{s} \leq t_1$ and $j^* = 0$, then

$$\begin{aligned} P(FP(k_{cv}^*) > j^*) &\leq P(FP(t_1) > 0) \\ &= P(\text{at least one of the } m_0 \text{ noise variables rank ahead of } t_1) \\ &\leq P(\text{at least one of the } m \text{ noise variables have p-value} < \alpha_m/m) \\ &\leq m \frac{\alpha_m}{m} = \alpha_m = o(1). \end{aligned} \quad (5.2)$$

Combining (5.1) and (5.2) implies (2.5).

In the case of $\hat{s} > t_1$, consider the conditional probability $P(FP(k_{cv}^*) > j^* | P_1^1, \dots, P_s^1)$. By Markov's inequality,

$$\begin{aligned}
 P(FP(k_{cv}^*) > j^* | P_1^1, \dots, P_s^1) &\leq \frac{1}{j^*} E(FP(k_{cv}^*) | P_1^1, \dots, P_s^1) = \frac{1}{j^*} E \left(\sum_{j=1}^{m_0} 1(P_j^0 \leq \alpha_{j^*} | P_1^1, \dots, P_s^1) \right) \\
 &= \frac{1}{j^*} \sum_{j=1}^{m_0} P(P_j^0 \leq \alpha_{j^*} | P_1^1, \dots, P_s^1) \leq \frac{m_0}{j^*} \alpha_{j^*} = \frac{m_0}{j^*} \frac{j^*}{m} \alpha \leq \alpha,
 \end{aligned}$$

where the fourth step is by condition (2.4). The above implies

$$P(FP(k_{cv}^*) > j^*) \leq \alpha \quad (5.3)$$

Combining (5.1) and (5.3) gives (2.5).

5.2 Proof of Theorem 2.2

We consider $\bar{l} \leq \epsilon s/2$ and $\bar{l} > \epsilon s/2$ separately.

When $\bar{l} \leq \epsilon s/2$, the proof is relatively straight-forward. First, by the definitions T_2 and $\bar{l}$,

$$s = TP(T_2) = T_2 - FP(T_2) > T_2 - \bar{l} \geq T_2 - \epsilon s/2,$$

which implies $T_2 < (1 + \epsilon/2)s$ with probability tending to 1. On the other hand,

$$k^* \geq \hat{s} > (1 - \epsilon/4)s > \frac{1 - \epsilon/4}{1 + \epsilon/2} T_2 > T_2 - (2\epsilon/3)T_2,$$

which implies that

$$TP(k^*) \geq TP(T_2) - (2\epsilon/3)T_2 \geq s - (2\epsilon/3)(1 + \epsilon/2)s > (1 - \epsilon)s$$

with probability tending to 1. This conclude the case with $\bar{l} \leq \epsilon s/2$.

The following proof is for $\bar{l} > \epsilon s/2$. By the definition of SMR^ϵ , it is enough to show that for any $\epsilon > 0$,

$$P(FN(k^*) > \epsilon s) \rightarrow 0. \quad (5.4)$$

The case $\hat{s} \leq t_1$ can be proved by similar arguments leading to (5.2).

Consider the case $\hat{s} > t_1$. Without loss of generality, assume ϵs is an integer. Denote $T_2^{(1-\epsilon)s}$ as the position for the $(1 - \epsilon)s$ -th signal. Then

$$\{FN(k^*) > \epsilon s\} \subseteq \{k^* < T_2^{(1-\epsilon)s}\} \subseteq \left\{ \exists j \in \{1, \dots, T_2^{(1-\epsilon)s} - \hat{s}\} \text{ s.t. } P(\hat{s} + j) \leq b_j \right\}$$

Note that $P_{(1)}, \dots, P_{(T_2)}$ are composed of $P_{(1)}^1, \dots, P_{(s)}^1$ and $P_{(1)}^0, \dots, P_{(L)}^0$. Denote b_{j_r} and b_{j_q} as the critical values corresponding to $P_{(r)}^0$ and $P_{(q)}^1$ in $\left\{P_{(\hat{s}+1)}, \dots, P_{(T_2^{(1-\epsilon)s})}\right\}$, respectively.

Then

$$\begin{aligned} P(FN(k^*) > \epsilon s) &\leq P\left(\exists j \in \{1, \dots, T_2^{(1-\epsilon)s} - \hat{s}\} \text{ s.t. } P_{(\hat{s}+j)} \leq b_j\right) \\ &\leq P\left(\exists r \text{ s.t. } P_{(r)}^0 \leq b_{j_r}\right) + P\left(\exists q \text{ s.t. } P_{(q)}^1 \leq b_{j_q}\right) \end{aligned} \quad (5.5)$$

Let $P_{(r_q)}^0$ be the largest $P_{(r)}^0$ before $P_{(q)}^1$. The following lemma shows the relationships between r and j_r and between r_q and j_q .

Lemma 5.1 Given a consistent estimator $\hat{s}$ for the number of signals and $\bar{l} > \epsilon s/2$, we have

$$j_r + \epsilon s/2 < r < L \leq \bar{l} \quad (5.6)$$

and

$$j_q + \epsilon s/2 < r_q < L \leq \bar{l} \quad (5.7)$$

with high probability.

Now consider the first term of (5.5). By condition (2.9),

$$P\left(P_{(r)}^0 \leq b_{j_r} \mid P_1^1, \dots, P_s^1\right) \leq P\left(B_r \leq b_{j_r}\right),$$

where B_r is a random variable following $\text{Beta}(\nu_1, \nu_2)$ with $\nu_1 = r$ and $\nu_2 = m - s - r + 1$. Further, by (5.6),

$$P\left(B_r \leq b_{j_r}\right) < P\left(B_r \leq b_{r - \epsilon s/2}\right) \leq P\left(B_{\bar{l}} \leq b_{\bar{l} - \epsilon s/2}\right),$$

where the second inequality is by the properties of Beta distribution that $B_{(r)}$ is less positively skewed as r increases and the change of skewness gets slower as r approaches to $(m - s)/2$. Then the above implies

$$P\left(P_{(r)}^0 \leq b_{j_r}\right) \leq P\left(B_{\bar{l}} \leq b_{\bar{l} - \epsilon s/2}\right). \quad (5.8)$$

Consider the second term of (5.5). Clearly $P\left(P_{(q)}^1 \leq b_{j_q}\right) < P\left(P_{(r_q)}^0 \leq b_{j_q}\right)$. Similar arguments as above combining condition (2.9) and (5.7) in Lemma 5.1 give

$$P\left(P_{(r_q)}^0 \leq b_{j_q} \middle| P_{(1)}^1, \dots, P_{(s)}^1\right) \leq P\left(B_{r_q} \leq b_{j_q}\right) \leq P\left(B_{r_q} \leq b_{r_q - \epsilon s/2}\right) \leq P\left(B_{\bar{l}} \leq b_{\bar{l} - \epsilon s/2}\right),$$

which implies

$$P\left(P_{(q)}^1 \leq b_{j_q}\right) < P\left(P_{(r_q)}^0 \leq b_{j_q}\right) \leq P\left(B_{\bar{l}} \leq b_{\bar{l} - \epsilon s/2}\right). \quad (5.9)$$

Combining (5.8) and (5.9) with (5.5) gives

$$P(FN(k^*) > \epsilon s) \leq (\bar{l} + s) \cdot P\left(B_{\bar{l}} \leq b_{\bar{l} - \epsilon s/2}\right). \quad (5.10)$$

Note that there is no explicit form for the cumulative distribution function of Beta distribution. To derive the probability in (5.10), let

$$F = \frac{\nu_2 B_{\bar{l}}}{\nu_1 (1 - B_{\bar{l}})}.$$

By the relationship between Beta and F distributions, F has an $F_{2\nu_1, 2\nu_2}$ distribution (Johnson and Kotz, 1970). In our case, $1 \ll \nu_1 \ll \nu_2$ by condition (2.8), then $F_{2\nu_1, 2\nu_2}$ is highly concentrated at 1 with mean ≈ 1 and variance $\approx 1/\nu_1$. On the other hand, we know that the median of Beta distribution is bounded by its mean for $1 < \nu_1 < \nu_2$, then

$$b_{\bar{l} - \epsilon s/2} < E(B_{\bar{l} - \epsilon s/2}) = \frac{\bar{l} - \epsilon s/2}{m - \hat{s} + 1}$$

and

$$\begin{aligned} P(B_{\bar{l}} \leq b_{\bar{l} - \epsilon s/2}) &= P\left(F \leq \frac{\nu_2}{\nu_1} \frac{b_{\bar{l} - \epsilon s/2}}{1 - b_{\bar{l} - \epsilon s/2}}\right) \leq P\left(F \leq \frac{\nu_2}{\nu_1} b_{\bar{l} - \epsilon s/2} + C \frac{\nu_2}{\nu_1} b_{\bar{l} - \epsilon s/2}^2\right) \\ &\leq P\left(F \leq \frac{\nu_2}{\nu_1} \frac{\bar{l} - \epsilon s/2}{m - \hat{s} + 1} + C \frac{\bar{l}}{m}\right) \leq P\left(F \leq 1 - C \frac{\epsilon s}{\bar{l}}\right), \end{aligned} \quad (5.11)$$

where the last step is by condition (2.8). Further, let

$$Z = \frac{F^{1/3} \left(1 - \frac{2}{9\nu_2}\right) + \frac{2}{9\nu_1} - 1}{\sqrt{\frac{2F^{2/3}}{9\nu_2} + \frac{2}{9\nu_1}}},$$

then by Wilson-Hilferty approximation to χ^2 and the fact that F is a ratio of χ^2 distribution, Z approximately follows $N(0,1)$ distribution (Johnson and Kotz, 1970). It is clear that

$$Z < \frac{F^{1/3} + \frac{2}{9\nu_1} - 1}{\sqrt{\frac{2}{9\nu_1}}},$$

then

$$\begin{aligned} P\left(F \leq 1 - C\frac{\epsilon s}{l}\right) &< P\left(\left(\sqrt{\frac{2}{9\nu_1}}Z + 1 - \frac{2}{9\nu_1}\right)^3 \leq 1 - C\frac{\epsilon s}{l}\right) \\ &= P\left(Z \leq \sqrt{\frac{9\nu_1}{2}}\left(\left(1 - C\frac{\epsilon s}{l}\right)^{1/3} - 1 + \frac{2}{9\nu_1}\right)\right) \end{aligned} \quad (5.12)$$

Apply Taylor expansion on $\left(1 - C\frac{\epsilon s}{l}\right)^{1/3}$,

$$\left(1 - C\frac{\epsilon s}{l}\right)^{1/3} - 1 + \frac{2}{9\nu_1} \leq -C\frac{\epsilon s}{l} + \frac{2}{9l} \leq -C\frac{\epsilon s}{l}$$

as $s \gg 1$ is implied by condition (2.8). Then

$$\begin{aligned} P\left(Z \leq \sqrt{\frac{9\nu_1}{2}}\left(\left(1 - C\frac{\epsilon s}{l}\right)^{1/3} - 1 + \frac{2}{9\nu_1}\right)\right) &\leq P\left(Z \leq \sqrt{\frac{9l}{2}}\left(-C\frac{\epsilon s}{l}\right)\right) \\ &\leq P\left(Z \leq -C\frac{s}{\sqrt{l}}\right) \end{aligned} \quad (5.13)$$

Finally, by the Normal approximation of Z and Mill's inequality,

$$(\bar{l} + s)P\left(Z \leq -C\frac{s}{\sqrt{l}}\right) \leq C\bar{l}^{3/2}\exp\left(-\frac{Cs^2}{l}\right) + Cs\bar{l}^{1/2}\exp\left(-\frac{Cs^2}{l}\right) = o(1), \quad (5.14)$$

where the last step is by condition (2.8). Summarizing (5.10) - (5.14) gives (5.4).

5.3 Proof of Lemma 5.1

Consider (5.6) first. At the position of $\hat{s} + j_r$,

$$\hat{s} + j_r = FP(\hat{s} + j_r) + TP(\hat{s} + j_r) \leq r + TP\left(T_2^{(1-\epsilon)s}\right) = r + (1-\epsilon)s.$$

On the other hand, the consistency of s implies $\hat{s} > (1 - \epsilon/2)s$ with high probability, then

$$\hat{s} + j_r > (1 - \epsilon/2)s + j_r$$

with high probability. Combining the above gives

$$r > j_r + \epsilon s/2$$

with high probability. Further, by the definitions of r , L and $\bar{L}$,

$$r \leq FP(T_2^{(1-\epsilon)^s}) \leq FP(T_2) = L \leq \bar{L}.$$

Then (5.6) follows.

(5.7) can be proved in the similar way as above given the fact that $FP(\hat{s} + j_q) = r_q$.

5.4 A Toy example for AdSMR algorithm

We provide a simple example demonstrating the AdSMR algorithm presented at the end of Section 2.3. Suppose the data yield $\hat{s} = 1$ and 10 ordered p -values

$$\{0.02, 0.11, 0.12, 0.21, 0.36, 0.49, 0.69, 0.77, 0.87, 0.99\}.$$

Recall the cut-off position in (2.6)

$$k_{cv}^* = \hat{s} + \min\left\{j \geq 1: p_{(\hat{s}+j)} \leq b_j\right\} 1\{\hat{s} > t_1\},$$

where $t_1 = \max\{j: p_{(j)} < \alpha_m/m\}$ with $\alpha_m = 1/\sqrt{\log m}$. For this example with $m = 10$, $t_1 = \max\{j: p_{(j)} < 0.066\} = 1$. Then, the indicator function $1\{\hat{s} > t_1\} = 0$; and $k^* = \hat{s} = 1$.

On the other hand, if the data yield $\hat{s} = 2$, then the indicator function $1\{\hat{s} > t_1\} = 1$, and the procedure continues to check $p_{(2+j)} \leq b_j$. In this example, $p_{(2+j)} = 0.12, 0.21, 0.36, \dots$ and $b_j = 1/8, 2/8, 3/8, \dots$ for $j = 1, 2, 3, \dots$. Then $k^* = \hat{s} + 1 = 3$.

5.5 Empirical SMR of AdSMR

We illustrate the empirical SMR of AdSMR with the MR estimator under the settings of Example 1 in simulation. Recall that Theorem 2.4 has $SMR^\epsilon(k^*) \rightarrow 0$ for any constant $\epsilon > 0$ under suitable conditions. Table 4 reports the empirical SMR^ϵ of AdSMR for $\epsilon = 0.1, 0.2$ and 0.3 from 100 replications. The results agree with Figure 3. For example, Figure 3 shows the median of FN proportion around 0.18 for AdSMR when $\pi = 0.02$ and $\mu = 4.5$. Table 4 shows that when $\mu = 4.5$, the FN proportion of AdSMR is greater than 0.1 in 65 (out of 100) replications, greater than 0.2 in 29 replications, and greater than 0.3 in 0 replications. The empirical SMR decreases as μ increases.

References

- Abraham G, Kowalczyk A, Zobel J, and Inouye M (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic Epidemiology* 37(2), 184–195. [PubMed: 23203348]

- Allen HL, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467(7317), 832–838. [PubMed: 20881960]
- Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 289–300.
- 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 (501), 265–277.
- Cai TT, Jin J, Low MG, et al. (2007). Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics* 35(6), 2421–2449.
- Cai TT 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* 79 (1), 197–223.
- Efron B (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* 102, 93–103.
- Fan J, Han X, and Gu W (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association* 107(499), 1019–1035. [PubMed: 24729644]
- Fan J and Lv J (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5), 849–911.
- Firmann M, Mayor V, Vidal PM, Bochud M, Pécoud A, Hayoz D, Paccaud F, Preisig M, Song KS, Yuan X, et al. (2008). The colaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC cardiovascular disorders* 8(1), 6. [PubMed: 18366642]
- Genovese C and Wasserman L (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, 1035–1061.
- Hung H, Lin Y, Chen P, Wang C, Huang S, and Tzeng J (2016). Detection of gene-gene interactions using multistage sparse and low-rank regression. *Biometrics* 72, 8594.
- Jin J and Cai TT (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association* 102 (478), 495–506.
- Johnson NL and Kotz S (1970). *Continuous Univariate Distributions-2, Volume 2* John Wiley & Sons.
- Kao P, Leung KH, C. C. LW, Yip S, and Yap M (2017). Pathway analysis of complex diseases for gwas, extending to consider rare variants, multi-omics and interactions. *Biochim Biophys Acta* 1861, 335353.
- Kostem E and Eskin E (2013). Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *The American Journal of Human Genetics* 92 (4), 558–564. [PubMed: 23561845]
- Lehmann EL and Romano JP (2005). Generalizations of the familywise error rate. *The Annals of Statistics* 33(3), 1138–54.
- 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 (1), 373–393.
- Powers D (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *J. of Machine Learning Technologies* 2(1), 37–63.
- Sarkar SK (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures. *The Annals of Statistics* 34, 394–415.
- Storey JD (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B* 64 (3), 479–498.
- Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- Waldmann P, Meszaros G, Gredler B, Fuerst C, and Solkner J (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics* 4, 270. [PubMed: 24363662]
- Westfall PH and Young SS (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Volume 279 John Wiley & Sons.

- Wu J, Devlin B, Ringquist S, Trucco M, and Roeder K (2010). Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic epidemiology* 34 (3), 275–285. [PubMed: 20088021]
- Wu TT, Chen YF, Hastie T, Sobel E, and Lange K (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25(6), 714–721. [PubMed: 19176549]
- Yang J, Lee SH, Goddard ME, and Visscher PM (2011). Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* 88(1), 76–82. [PubMed: 21167468]
- Zhou H, Alexander DH, Sehl ME, Sinsheimer JS, Sobel E, and Lange K (2011). Penalized regression for genome-wide association screening of sequence data In *Biocomputing 2011*, pp. 106–117. World Scientific.

Figure 1:

A sequence of ordered p -values. T_1 denotes the location right before the first noise variable and T_2 denotes the location of the last signal variable.

Figure 2:
Heatmap of the absolute value of correlations for 50 SNPs in CoLaus data.

Figure 3:

Comparison of methods in false negative proportion, false discovery proportion, and F-measure for block dependence with $l = 50$. The top row has $\pi = 0.1$, and the bottom row has $\pi = 0.02$.

Figure 4:

Comparison of methods in false negative proportion, false discovery proportion, and F-measure for block dependence with $l = 200$. The top row has $\pi = 0.1$, and the bottom row has $\pi = 0.02$.

Figure 5:

Comparison of methods in false negative proportion, false discovery proportion, and F-measure for sparse covariance matrix. The top row has $\pi = 0:1$, and the bottom row has $\pi = 0:02$.

Figure 6:

Comparison of methods in false negative proportion, false discovery proportion, and F-measure under dependence structure from a two-factor model. The top row has $\pi = 0:1$, and the bottom row has $\pi = 0:02$.

Table 1:

Classifications of variables.

	Selected	Not selected	Total
Signal	TP	FN	s
Noise	FP	TN	$m - s$
	R	$m - R$	m

Table 2:

The average cut-off positions from 100 replications with $m = 5000$ and $s = 250$.

	$\mu = 3.5$	$\mu = 4.5$	$\mu = 5.5$
cvSMR	5000	5000	4905
AdSMR	348	300	288

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3:

Numbers of selected SNPs and the estimated heritability.

	Number of SNPs	Estimated Heritability
Total	340, 359	53.7%
AdSMR	7, 204	53.7%
MDR	110, 778	53.7%

Table 4:

Empirical $SMR^{\epsilon}(I^*)$ of AdSMR in the setting of Figure 3 with $\pi = 0.02$.

	μ				
	4.5	5	5.5	6	6.5
$\epsilon = 0.1$	0.65	0.43	0.03	0	0
$\epsilon = 0.2$	0.29	0	0	0	0
$\epsilon = 0.3$	0	0	0	0	0