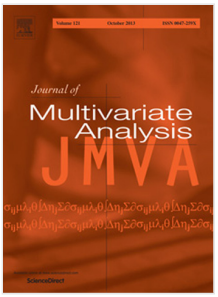

Detecting weak signals in high dimensions

X. Jessie Jeng

PII: S0047-259X(16)00025-7
DOI: http://dx.doi.org/10.1016/j.jmva.2016.02.004
Reference: YJMVA 4082

To appear in: *Journal of Multivariate Analysis*

Received date: 28 June 2014

Please cite this article as: X. Jessie Jeng, Detecting weak signals in high dimensions, *Journal of Multivariate Analysis* (2016), http://dx.doi.org/10.1016/j.jmva.2016.02.004

Detecting Weak Signals in High Dimensions

X. Jessie Jeng

Department of Statistics, North Carolina State University
SAS Hall, 2311 Stinson Dr., Raleigh, NC 27695, United States

E-mail address: xjjeng@ncsu.edu

Abstract

Fast emerging high-throughput technology advances scientific applications into a new era by enabling detection of information-bearing signals with unprecedented sizes. Despite its potential, the analysis of ultrahigh-dimensional data involves fundamental challenges, wherein the deluge of a large amount of irrelevant data can easily obscure the true signals. Classical statistical methods for low to moderate-dimensional data focus on identifying strong true signals using false positive control criteria. These methods, however, have limited power for identifying weak true signals embedded in an extremely large amount of noise. This paper seeks to facilitate the detection of weak signals by introducing a new approach based on false negative instead of false positive control. As a result, a high proportion of weak signals can be retained for follow-up study. The new procedure is completely data-driven and fast in computation. We show in theory its efficiency and adaptivity to the unknown features of the data including signal intensity and sparsity. Simulation studies further evaluate the method under various model settings. We apply the new method in a real-data analysis on detecting genomic variants with varying signal intensities.

Keywords: False negative control, Multiple testing, Variable screening, Variable selection, Trichotomous analysis

1 Introduction

1.1 Background and Motivation

Signal detection is a central topic in modern Statistics that considers the problem of discerning between information-bearing signals and random noise. Signal detection arises in a wide spectrum of applications such as the detection of astrophysical sources, surveillance for disease outbreaks, identification of genetic associations, etc. Fast emerging high-throughput technology advances scientific applications into a new era by enabling detection of information-bearing signals with unprecedented sizes. However, we pay the price for analyzing high-dimensional data, not only by extensive computational cost, but also by the capacity to identify the true signals, as they are more easily obscured by the large amount of noise. Examples can be found in Zaykin and Zhivotovsky [34], Han et al. [16], etc.

Contemporary statistical methods often use false positive control as the criterion to select true signals. Signals that are strong enough to stand outside the range of noise can be identified with high confidence. Popular false positive control criteria include family-wise error (FWER) control [9] and false discovery rate (FDR) control [2]. However, when data dimension is extremely high, such as in genome-wide association study, majority of the true signals may not be able to stand out from the noise, as the range of noise increases with the dimensionality of data. As a result, these true signals cannot be efficiently identified by methods based on false positive control.

Having recognized this limitation, researchers working with ultrahigh dimensional data often use subjective criterion to select an $a\%$ top-ranked candidates for follow-up study, where $a\%$ is a pre-fixed percentage determined by either prior knowledge in application or convention [29]. In high-dimensional regression, the problem of retaining a high proportion of signals has been studied from a different perspective. In the seminal paper of Fan and Lv [12], sure independent screening is proposed based on the ranking of the marginal sample correlations between the response and each predictor. It has been shown that when regression coefficients are large enough, the top d variables, where d is smaller than the sample size, include all the relevant vari-

ables with high probability. However, in real applications, the condition on regression coefficients is usually unverifiable. Even if we know *a priori* that the condition on regression coefficients is reasonable, always selecting a fixed number of d variables is not the most efficient approach as the selection does not adapt to the underlying sparsity and effect sizes of the relevant variables.

The limitation of the current methodology calls for statistical studies on efficient variable screening based on adaptive control on false negatives. Such development is highly relevant when the control of false negatives is of primary interest. For example, in early exploratory stage of ultrahigh dimensional data analysis, computationally efficient dimension reduction is often needed due to the extremely large number of candidates. However there has not been an agreement on how to select a subset of variables with a guarantee of retaining most of the ture signals. The proposed method could provide a data-driven dimension reduction for analyzing ultrahigh dimensional data. On the other hand, the results of the method can be combined with the results of multiple testing procedures to provide important insights of the data. For example, if the results of false negative control and false positive control are close, then it is likely that the signals in the dataset are strong and sparse; otherwise, signals are relatively weak and dense. By further analyzing these results, follow-up studies can be designed more efficiently.

1.2 A Trichotomous Framework

We propose a trichotomous framework to account for the fact that the high-dimensional data composed of a mixture of signal and noise would naturally fall into three subsets. The signal subset includes only strong signals that stand outside the range of noise. The noise subset includes only noise. The third subset is most interesting, where relatively weak signals mix indistinguishably with noise. We call this the mixed subset.

In this paper, we consider the trichotomous framework on p-values. Assume that the p-value

$$P_i \sim U1_{\{i\in S_0\}} + G1_{\{i\in S_1\}}, \qquad i \in \{1,\ldots,n\}, \tag{1}$$

where S_0 is the collection of the noise, S_1 is the collection of signals, U the uniform distribution on $[0,1]$, and G some unknown continuous distribution with $G(t) > U(t)$ for all $t \in (0,1)$. Rank the observed p-values increasingly, so that the p-values from S_1 would rank relatively higher than those from S_0. Define J_1 as the positions of the first noise minus one and J_2 the last signal, i.e., $J_1 = \min\{i : p_{(i)} \text{ from a noise}\} - 1$ and $J_2 = \max\{i : p_{(i)} \text{ from a signal}\}$. J_1 and J_2 are random, varying from sample to sample, and not directly observable from the data. We define the signal, mixed, and noise subsets as $\mathbb{S} = \{\text{candidates ranked } 1, \ldots, J_1\}$, $\mathbb{M} = \{\text{candidates ranked } J_1 + 1 \ldots, J_2\}$, and $\mathbb{N} = \{\text{candidates ranked } J_2 + 1, \ldots, n\}$, respectively.

The following example illustrate the natural formation of the three subsets. We generate n observations, among which 200 observations are signals generated from $\mathcal{N}(\mu, 1)$ and the rest are noise generated from $\mathcal{N}(0, 1)$. We calculate the p-values of the observations and rank them increasingly. Figure 1 shows the scatter plots of the ordered p-values. In Figure 1 (a), signal intensity $\mu = 7$ and data dimension $n = 1,000$. Signals are so strong that all of them rank ahead of the noise. Therefore, $J_1 = J_2 = 200$ and the mixed subset $\mathbb{M}$ does not exist. In Figure 1 (b), signal intensity reduces to $\mu = 3$, only 98 out of the 200 signals rank ahead of the noise, whereas the rest enter subset $\mathbb{M}$ between $J_1 = 98$ and $J_2 = 446$. In Figure 1 (c), signal intensity remains the same, but the data dimension increases to $n = 5,000$. In this case, fewer signals ($J_1 = 65$) stand outside the range of the noise and more signals enter $\mathbb{M}$ between $J_1 = 65$ and $J_2 = 1639$. This example shows that the mixed subset is more likely to exist when the intensity of signals is relatively low or the data dimension is sufficiently high.

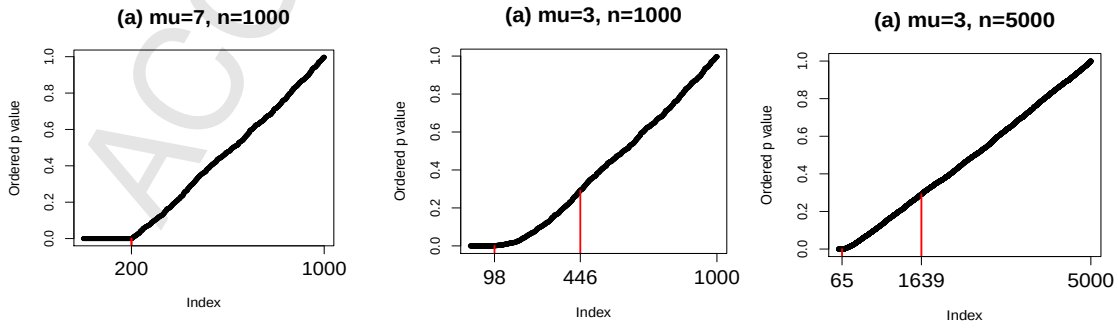

Figure 1: Formation of the signal, noise, and mixed subsets.

1.3 Our Contributions

We focus on the scientific challenge of detecting the true signals in $\mathbb{M}$ that are mixed indistinguishably with noise and, therefore, cannot be addressed by multiple testing. Although one can always relax the control level of false positives in multiple testing to include more signals, it is not clear by how much should the control level be relaxed. A pre-fixed high level would certainly not work for different datasets.

In this paper, we propose a data-driven screening procedure to identify the mixed and noise subsets. Theoretical analysis shows that our screening method can retain a large proportion of true signals with high probability. Our method is also readily adaptive to the underlying sparsity and intensity level of signals, so that the size of the identified mixed subset varies with these unknown features. The adaptivity of our method separates it from the existing screening procedures that often select a prefixed number or percentage of top ranked candidates. We further provide some insights on the characteristics of the three subsets by showing the sufficient and almost necessary conditions for the existence of the three subsets for a Gaussian model. The results can be connected to studies in mixture model detection and signal recovery, and provide additional insights to the phase diagram in high-dimensional inference.

The algorithm of our proposed method is easy to implement and efficient in computation with complexity at the order of $O(n \ln n)$. No prior information of the signal distribution is needed; neither are tuning parameters involved in the algorithm. These practical properties of the method make it easy to be adopted in a wide spectrum of applications. The method can be particularly valuable when data has relatively small signal-to-noise ratio under high-dimensionality. Application examples include association analysis for rare genomic variants, structural variation analysis in germline constitutional genome, etc. In this paper, we apply the trichotomous analysis to a real-data example to detect copy number variations based on genotyping data generated from Illumina HumanHap550 array.

From extensive literature search, we find that the idea of separating data into three categories has been pursued in various applications. For example, Drton and Perlman [8] introduced an "indeterminate set" for Gaussian graphical model selection. This set

is determined by looking at the scatter plot of the p-values and subjectively selecting the middle range where p-values are not too large or too small. This procedure is simple to implement but obviously ad hoc. Another related study can be found in Jeske et al. [18], where a Bayes classifier is proposed for microbial community profiling, and a "neutral zone" is defined as the set of data where the weight of evidence is too weak for the Bayes classifier to make decisions. This approach needs to pre-specify the density function of signals, which is not required for the implementation of our method. A fundamental difference between the existing studies and the study here is that the existing studies define the "middle" sets based on the results of certain methods, whereas we first define and characterize the three subsets from the data, and then develop data-driven methods to identify them with statistical accuracy.

2 Identification of the Three Subsets

Recall the definitions of signal, noise, and mixed subsets in Section 1.2 based on ranked p-values. Since no noise rank ahead of J_1, it is natural to identify $\mathbb{S}$ through a procedure controlling family-wise Type I error. Such procedures have been widely studied in literature [9]. Note that the main purpose of identifying $\mathbb{S}$ is to reveal the three-subset structure of the data. This is different from the purpose of multiple testing, which is to make dichotomous decisions about whether to reject each of the null hypotheses.

In this paper, we focus on the more challenging problem of identifying $\mathbb{M}$ and $\mathbb{N}$. We propose a new selection rule based on p-value ranking. We refer to the new selection rule as an Adaptive False Negative Control (AFNC) procedure. For illustrative purposes, we first study an ideal setting where the number of signals (s) is known *a priori*. Then, the AFNC procedure is developed for unknown s.

2.1 AFNC with the number of signals known a priori

In order to identify $\mathbb{M}$ and $\mathbb{N}$, we need to find the separation point J_2. Even when s is known, one still do not know the location of J_2 as signals may be mixed in-

distinguishably with noise. We propose the AFNC procedure, which traverses the ordered p-values until all signals are likely to be included. Given the number of signals, $s(= |S_1|)$, the AFNC procedure is constructed as follows.

(1) Order the p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$.

(2) Find the point

$$T_{fn} = s + \min\{j \geq 1 : p_{(s+j)} \leq d_{s,j}(\gamma)\}1_{\{s>t_1\}}, \tag{2}$$

where $d_{s,j}(\gamma)$ is a pre-specified critical sequence and $t_1 = \max\{j : p_{(j)} < d_{s,1}(\gamma)\}$.

(3) Retain the top $\{1, \ldots, T_{fn}\}$ candidates.

This simple procedure seeks to include a high proportion of signals while incurring as few false positives as possible. The critical sequence $d_{s,j}(\gamma)$ is generated from p-values of $n - s$ noise and defined as the 100γ-th percentile of $U(j)$, where $U(j)$ is the cdf of $P_{(j)}$ from $n - s$ noise and γ a prefixed small number. The occurrence of the event $p(s + j) \leq d_{s,j}(\gamma)$ implies that all signals have been included in the top $s + j$ candidates with probability of at least $1 - \gamma$. We suggest to stop as early as possible when this event occurs. On the other hand, the indicator $1_{\{s>t_1\}}$ separates the case where all signals are strong enough to rank ahead of noise from the case where some signals are relatively weak and mixed indistinguishably with noise.

Define $FN(i)$ as the number of false negatives for selecting the top $\{1, \ldots i\}$ candidates ranked by their p-values. We have the following theoretical results on the efficiency of AFNC.

Theorem 1 *Assume model (1) with $s = |S_1|$ known a priori. Then the AFNC procedure optimally controls the probability of having at least one false negatives at level γ based on the ranked p-values, i.e.,*

$$T_{fn} = \min\{i \geq 1 : \Pr\{FN(i) \geq 1\} \leq \gamma\}. \tag{3}$$

Theorem 1 says that the top $\{1, \ldots T_{fn}\}$ candidates include all the signals with high probability $(1-\gamma)$. Furthermore, this selected set of AFNC is the smallest set achieving this false negative control property. Given the efficiency of AFNC, T_{fn} can be used as a surrogate of J_2 to separate the mixed and noise subsets.

2.2 AFNC with unknown number of signals

In real-data applications, the number of signals is usually unknown. Various estimators have been developed in the literature to estimate the signal proportion $\pi(= s/n)$. For example, Genovese and Wasserman [15] and Meinshausen and Rice [26] proposed two proportion estimators for the p-value model as in (1). Cai et al. [5], Jin and Cai [22], and Jin [21] developed proportion estimators for normally distributed observations. Consistency of the aforementioned estimators has been studied under independence assumption on the variables. Proportion estimators can be used to improve power of multiple testing methods [15, 30]. Here, we incorporate an existing proportion estimator in AFNC when s is unknown. For algorithm simplicity, we choose the estimator introduced in Meinshausen and Rice [26], which is also constructed based on p-values.

$$\hat{\pi} = \max_{1<i<n/2} \frac{i/n - p_{(i)} - \sqrt{2\ln\ln n/n}\sqrt{p_{(i)}(1-p_{(i)})}}{1-p_{(i)}}. \tag{4}$$

The consistency of $\hat{\pi}$ for estimating the proportion of both dense and sparse signals has be proved for independent p-values [26]. Let $\pi = n^{-C}$ for some $C \in [0,1)$. Assume Model (1) with independent p-values and either

$$C \in [0, 1/2) \quad \text{and} \quad \inf_{t\in(0,1)} G'(t) = 0, \tag{5}$$

or

$$C \in [1/2, 1) \quad \text{and} \quad \forall q \in (0,1), \quad \lim_{n\to\infty} \{\ln G^{-1}(q)\}/(\ln n) = -r, \quad r > 2(C - 1/2). \tag{6}$$

Then, it has been shown that $\Pr\{(1-\epsilon)\pi \le \hat{\pi} \le \pi\} \to 1$ as $n \to \infty$ for any $\epsilon > 0$. Condition (5) considers relatively dense signals with $\pi n \gg \sqrt{n}$; and all we need is the "pure" condition $\inf_{t\in(0,1)} G'(t) = 0$ [15, 26]. Condition (6) considers sparse signals with $\pi n \le \sqrt{n}$. In this case, stronger condition is needed for signal intensity, which is implied by (6).

Utilizing a consistent estimator $\hat{\pi}$ for signal proportion, we develop the AFNC with

$$\hat{T}_{fn} = \hat{\pi} n + \min\{j \ge 1 : p_{(\hat{\pi}n+j)} \le d_{\hat{\pi}n,j}(\gamma)\} 1_{\{\hat{\pi}n > t_{\hat{\pi},1}\}}, \tag{7}$$

where the critical sequence $d_{\hat{\pi}n,j}(\gamma)$ is defined as the 100γ-th percentile of the distribution of the j-th ordered p-value from $(1-\hat{\pi})n$ noise, γ a prefixed small number, and $t_{\hat{\pi},1} = \max\{j : p_{(j)} < d_{\hat{\pi}n,1}(\gamma)\}$.

It can be shown that the AFNC can control false negatives by controlling the probability of missing a small proportion of signals. Recall the definition of $FN(i)$ as the number of false negatives for selecting the top $\{1, \ldots i\}$ candidates ranked by their p-values. We have the following theoretical results.

Theorem 2 *Assume model (1) with independent p-values and condition (5) or (6). Then the AFNC procedure asymptotically controls the probability of having at least a small proportion of false negatives at level γ, i.e.,*

$$\Pr\{FN(\hat{T}_{fn})/s > \epsilon\} \leq \gamma + \Delta, \tag{8}$$

where $\Delta = o(1)$ for arbitrarily small constant $\epsilon > 0$.

The result in (8) is weaker than that in (3). This is the price to pay for not knowing the number of signals *a priori*. To show a similar result of optimality as in Theorem 1, delicate analysis is necessary to explore the effect of the convergence speed of $\hat{\pi}$ to π on $\hat{T}_{fn}$. More specific assumptions on the model may be needed. This analysis is deferred to future work.

In addition to the false negative control property, AFNC is adaptive to the unknown signal proportion and intensity. When all signals are strong enough to rank ahead of noise. $\hat{T}_{fn}$ converges to s.

Theorem 3 *Assume model (1) with independent p-values. If signals are strong enough, such that $s\bar{G}(n^{-r}) \to 0$ for some $r > 1$. Then, with high probability, the mixed subset does not exist, and for any $\gamma = \gamma_n$ satisfying $\gamma_n = o(1)$ and $\gamma_n > n^{-(r-1)}$, the AFNC procedure consistently selects the top s candidates, i.e., $\Pr(\hat{T}_{fn} = J_1 = J_2 = s) \to 1$ as $n \to \infty$.*

An intuitive understanding for the condition $s\bar{G}(n^{-r}) \to 0$, $r > 1$, is that $\bar{G}(n^{-r}) \ll 1/s = 1/n^{1-C} = o(1)$, which means that the total mass of G is asymptotically between 0 and n^{-r}. Note that the expectation of the smallest p-value from n independent

noise is at the order of n^{-1}. Therefore, with $r > 1$, all the p-values of signals are well-separated from all the p-values of noise.

Theorem 2 and 3 demonstrate the validity and efficiency of the AFNC with $\hat{\pi}$ defined as in (4). If other proportion estimators are implemented in (7), conditions in Theorem 2 and 3 need to be changed accordingly. For example, the proportion estimator in Cai et al. [5] is designed for normally distributed noise and signals. Utilizing the additional assumptions for the model, this estimator is consistent under a weaker condition on the signal intensity in the sparse scenario compared to (6) [5]. The validity and efficiency of the AFNC can be proved in a similar way.

2.3 Prior Knowledge-Based AFNC

In applications, data may not satisfy the conditions for the existence of a consistent proportion estimator. However, prior knowledge can often allow practitioners to provide a possible range for the number of signals. Suppose s is bounded by

$$s^- \leq s \leq s^+. \tag{9}$$

Utilizing this information, the prior knowledge-based AFNC can be developed with

$$\tilde{T}_{fn} = s^+ + \min\{j \geq 1 : p_{(s^+ + j)} \leq d_j(\gamma)\} 1_{\{s^+ > t_1\}}.$$

Result on the validity of the prior knowledge-based AFNC can be proved similarly as in Theorem 2.

Corollary 2.1 *Assume model (1) and condition (9). Then we have*

$$\Pr\{FN(\tilde{T}_{fn}) > 1\} \leq \gamma.$$

Compared to $\hat{T}_{fn}$, $\tilde{T}_{fn}$ may include more noise. However, the prior knowledge-based AFNC can be useful in applications where conditions for the consistency of proportion estimation are hard to be satisfied and some informative knowledge of the number of signals is available.

3 Insights on the Formations of the Three Subsets

We provide some additional insights on the natural formations of $\mathbb{S}$, $\mathbb{M}$, and $\mathbb{N}$. The following theorem shows the sufficient and almost necessary conditions for the existence of the three subsets under a Gaussian model. This model has been widely studied in high-dimensional inference. Assume the data are generated by

$$X_i \sim \mathcal{N}(0,1)1_{\{i\in S_0\}} + \mathcal{N}(\mu,\sigma^2)1_{\{i\in S_1\}}, \quad i = 1,\ldots,n, \tag{10}$$

where $\mu > 0$ and $\sigma > 0$. Define $n_0 = |S_0|$.

Theorem 4 *Assume model (10). The observed p-values are ranked increasingly. Then, we have*

$$\Pr(\mathbb{S} \neq \emptyset) \to 1 \qquad \textit{if} \qquad \mu \geq \sqrt{2(1+\epsilon)\ln n_0} - \sigma\sqrt{2\ln s}, \tag{11}$$

$$\Pr(\mathbb{S} = \emptyset) \to 1 \qquad \textit{if} \qquad \mu \leq \sqrt{2(1-\epsilon)\ln n_0} - \sigma\sqrt{2\ln s}, \tag{12}$$

$$\Pr(\mathbb{M} \neq \emptyset) \to 1 \qquad \textit{if} \qquad \mu \leq \sqrt{2(1-\epsilon)\ln n_0} + \sigma\sqrt{2\ln s} \tag{13}$$

$$\Pr(\mathbb{M} = \emptyset) \to 1 \qquad \textit{if} \qquad \mu \geq \sqrt{2(1+\epsilon)\ln n_0} + \sigma\sqrt{2\ln s} \tag{14}$$

$$\Pr(\mathbb{N} \neq \emptyset) \to 1 \qquad \textit{if} \qquad \ln s \leq (1-\epsilon)\ln n_0, \tag{15}$$

$$\Pr(\mathbb{N} = \emptyset) \to 1 \qquad \textit{if} \qquad \ln s \geq (1+\epsilon)\ln n_0, \tag{16}$$

for arbitrarily small constant $\epsilon > 0$.

Theorem 4 says that all three subsets asymptotically exist when signals are sparse as in (15) and the signal intensity is between the two bounds in (11) and (13). Moreover, (13) shows that the higher the dimensionality is, the more likely that the mixed subset exists. Note that since ϵ is an arbitrarily small constant, (12), (14), and (16) imply that (11), (13), and (15) are sufficient and almost necessary conditions for the existence of the three subsets respectively.

We find some interesting connections between the formations of the three subsets to the problem of mixture model detection in Donoho and Jin [7] and Cai et al. [6], and to the problem of exact signal recovery in Xie et al. [33] and Ji and Jin [19]. Adopting the calibration used in the aforementioned papers, let

$$\pi = n^{-\gamma}, \quad 0 < \gamma < 1, \qquad \text{and} \qquad \mu = \mu_n = \sqrt{2r\ln n}, \quad r > 0. \tag{17}$$

We have the following result. The proof is straightforward, and thus, omitted.

Corollary 3.1 *Assume model (10) with calibration (17). Then, asymptotically, the noise subset always exists, and the sufficient and necessary condition for* $\Pr(\mathbb{S} \neq \emptyset) \to 1$ *is*

$$r > (1 - \sigma\sqrt{1-\gamma})^2, \tag{18}$$

and for $\Pr(\mathbb{M} \neq \emptyset) \to 1$ *is*

$$r < (1 + \sigma\sqrt{1-\gamma})^2. \tag{19}$$

We find that condition (18) coincides with the upper half of the detection boundary introduced in Cai et al. [6] for mixture model detection. The problem of mixture model detection is to test the global null, $X_i \sim \mathcal{N}(0,1)$ for all $i = 1, \ldots, n$, versus the global alternative, $X_i \sim (1-\pi)\mathcal{N}(0,1) + \pi\mathcal{N}(\mu, \sigma^2)$ for all $i = 1, \ldots, n$. Corollary 3.1 implies the following result.

Corollary 3.2 *For the problem of mixture model detection with calibration (17) and* $\gamma \in (1/2, 1)$*, only when the signal subset exists is it possible to successfully separate the global null from the global alternative.*

We also find that condition (19) delineates the complementary set of the exact recovery region in Xie et al. [33] and Ji and Jin [19]. The problem of exact signal recovery is to show under what condition, Pr(all signals can be separated from noise) $\to$ 1. Corollary 3.1 implies the following result.

Corollary 3.3 *For the problem of exact signal recovery with calibration (17), only when the mixed subset does not exist, is it possible to fully recover all the signals.*

The above study on the formations of the three subsets provides additional insights to the phase diagram in high-dimensional inference.

4 Simulation Studies

In this section, we demonstrate the finite sample performance of AFNC. In each example, $10,000$ observations are generated, in which the noise data points are sampled

independently from $\mathcal{N}(0, \sigma)$ and signals from $\mathcal{N}(\mu, 1)$. The results of AFNC with $\hat{T}_{fn}$ as in (7) and $\gamma = 0.05$ are compared to those of the Bonferroni with $\alpha = 0.05$, the BH-FDR with $\alpha = 0.05$ [2], and the adaptive FDR [3]. Note that we can use Bonferroni to separate the signal and mixed subsets ($\mathbb{S}$-$\mathbb{M}$ Separation) and $\hat{T}_{fn}$ to separate the mixed and noise subsets ($\mathbb{M}$-$\mathbb{N}$ Separation).

We report the number of top-ranked variables selected by AFNC as $\hat{T}_{fn}$, by Bonferroni as t_{bonf}, by BH-FDR as t_{FDR}, and by adaptive FDR as t_{aFDR}. Corresponding numbers of false positives (FP) and false negatives (FN) for each procedure are also computed. We repeatedly generate the observations and compute performance measures for 100 times in each simulation example. The median and mean absolute deviation (MAD) of these measures are reported for more robust comparison results against the outliers in the 100 replications.

Example 1 demonstrates the effect of signal proportion. Set $\sigma = 1$ and $\mu = 3$. The signal proportion π changes from 1% to 20%. As signal proportion increases, the adaptive FDR significantly outperforms BH-FDR. To save space, the results of the latter are omitted in this example. As shown in Table 1, the number of selected candidates of all the methods increases with π, and t_{aFDR} is always in between t_{bonf} and $\hat{T}_{fn}$. The FN of $\hat{T}_{fn}$ is much smaller than those of the other methods and remains fairly robust over different signal proportions. This shows that $\hat{T}_{fn}$ is a valid false negative control procedure which is adaptive to the unspecified signal proportion. The FP of t_{bonf} remains around 0 showing that this procedure only selects true signals and therefore can be used as a surrogate of J_1. The adaptive FDR has both FP and FN increasing with signal proportion. When the proportion of true signals is relatively large, large numbers of FN are observed for adaptive FDR. This example shows that the trichotomous analysis is a valuable addition to the dichotomous decision rules.

Example 2 demonstrates the effect of signal intensity. Set $\sigma = 1$ and $\pi = 0.01$. Signal mean μ varies from 2.5 to 5.5. Since the signal proportion is very small, the results of BH-FDR and the adaptive FDR are very close. To save space, the results of the latter are omitted in this example. Figure 2 presents the histograms of $\hat{T}_{fn}$ from the 100 replications for $\mu = 2.5$ and 5.5. It shows that as signal intensity increases, the distribution of $\hat{T}_{fn}$ becomes more concentrated. Table 5 shows that the cut-off

Table 1: Effect of signal proportion. Median and MAD (in parentheses) of t_{bonf}, t_{FDR}, $\hat{T}_{fn}$, and their corresponding FP and FN over 100 replications. μ is fixed at 3.

	$\mathbb{S}$-$\mathbb{M}$ Separation			adapFDR			$\mathbb{M}$-$\mathbb{N}$ Separation		
π	t_{bonf}	FP	FN	t_{aFDR}	FP	FN	$\hat{T}_{fn}$	FP	FN
1%	8(3)	0(0)	92(3)	27(7)	1(1)	74(6)	227(140)	147(135)	21(12)
5%	40(6)	0(0)	460(6)	259(13)	12(4)	253(13)	1119(494)	645(462)	31(28)
10%	81(8.9)	0(0)	919(9)	647(18)	31(4)	386(18)	1960(589)	996(548)	38(31)
20%	172(11)	0(0)	1828(10)	1543(32)	72(11)	529(24)	3224(660)	1268(614)	46(32)

locations of all three procedures converge to the number of true signals as signal intensity increases. This demonstrates the adaptivity of $\hat{T}_{fn}$ with finite sample.

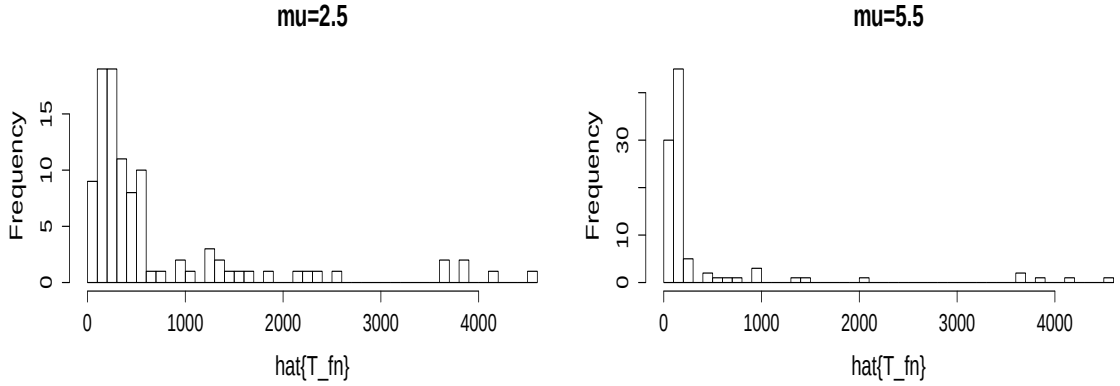

Figure 2: Histograms of $\hat{T}_{fn}$ for $\mu = 2.5$ and 5.5 from 100 replications.

Table 2: Effect of signal intensity. π is fixed at 1%.

	$\mathbb{S}$-$\mathbb{M}$ Separation			BH-FDR			$\mathbb{M}$-$\mathbb{N}$ Separation		
	t_{bonf}	FP	FN	t_{FDR}	FP	FN	$\hat{T}_{fn}$	FP	FN
$\mu = 2.5$	3(1)	0(0)	97(1)	8(4)	0(0)	92(4)	325(269)	261(253)	28(19)
$\mu = 3.5$	17(3)	0(0)	83(3)	54(7)	2(1)	48(6)	194(113)	103(97)	11(9)
$\mu = 4.5$	54(4)	0(0)	46(5)	92(4)	4(3)	12(3)	126(44)	29(34)	3(3)
$\mu = 5.5$	86(3)	0(0)	14(3)	103(1)	4(1)	1(1)	104(9)	4(3)	1(1)

Example 3 has heterogeneous noise generated for 10% of the observations. With signal intensity and proportion fixed at $\mu = 3.5$ and $\pi = 1\%$, the proportion of het-

erogeneous noise is 10 times the proportion of signals. This example demonstrates a common scenario in real-data applications where unjustified artifacts causes heterogeneity in the background noise. The heterogeneous noise in this example are randomly generated from $\mathcal{N}(0, \sigma)$ with $\sigma \sim \text{Gamma}(2, \theta)$. Let the scale parameter θ vary from 0.5 to 2, which results in increasing variability for the noise. Due to the small signal proportion, the results of the adaptive FDR are very close to those of the BH-FDR and omitted in this example. Table 3 shows that FPs of all procedures increase with θ. FNs, on the other hand, are very stable. Corollary 2.1 provides some explanation for the robustness of $\hat{T}_{fn}$ in controlling false negatives in this example. Since heterogeneous noise can result in large jumps, the estimated proportion $\hat{\pi}$ is larger than the true π. Constructed using this $\hat{\pi}$, $\hat{T}_{fn}$ is close to $\tilde{T}_{fn}$ in (2.3). The false negative control of $\tilde{T}_{fn}$ is presented in Corollary 2.1.

Table 3: Robustness for heterogeneous noise. Set $\mu = 3.5$ and $\pi = 1\%$.

	$\mathbb{S}$-$\mathbb{M}$ Separation			BH-FDR			$\mathbb{M}$-$\mathbb{N}$ Separation		
	t_{bonf}	FP	FN	t_{FDR}	FP	FN	$\hat{T}_{fn}$	FP	FN
$\theta = 0.5$	22(4)	5(3)	82(3)	69(9)	15(4)	45(6)	196(67)	107(60)	12(7)
$\theta = 1$	53(7)	35(6)	81(4)	132(12)	71(9)	38(4)	443(180)	347(174)	7(4)
$\theta = 1.5$	94(10)	75(9)	80(4)	195(15)	130(13)	35(4)	556(230)	459(223)	7(3)
$\theta = 2$	134(12)	113(10)	80(4)	249(12)	182(10)	33(4)	556(179)	466(175)	9(4)

Example 4 generates autocorrelated observations with $\rho_{ij} = a^{|i-j|}$ for $a = 0, 0.5, 0.7$ and 0.9. The number of observations are reduced to 1,000 to save computation time. Set $\sigma = 1$, $\pi = 0.05$, and $\mu = 3$. The results summarized in Table 4 are quite stable over different values of the autocorrelation parameter a with $\hat{T}_{fn}$ having slightly better control on false negatives for large a.

Example 5 illustrates the false negative control property presented in Theorem 2 under finite sample. Particularly, Theorem 2 shows that under certain conditions on the signal proportion and intensity, AFNC asymptotically controls the probability of having a small proportion of false negatives. Since the result in Theorem 2 with arbitrarily small constant ϵ is implied by $\Pr\{FN(\hat{T}_{fn})/s > \epsilon_n\} \leq \gamma + \Delta$ for $\epsilon_n \to 0$ as $n \to \infty$, we set $\epsilon_n = 1/\ln(n)$ and record the empirical probability of $FN(\hat{T}_{fn})/s > \epsilon_n$

Table 4: Robustness under autocorrelation. Set $\pi = 0.05$ and $\mu = 3$.

	$\mathbb{S}$-$\mathbb{M}$ Separation			BH-FDR			$\mathbb{M}$-$\mathbb{N}$ Separation		
	t_{bonf}	FP	FN	t_{FDR}	FP	FN	$\hat{T}_{fn}$	FP	FN
$a = 0$	14(3)	0(0)	36(3)	27(4)	1(1)	25(4)	74(45)	30(33)	7(7)
$a = 0.5$	13(4)	0(0)	37(4)	24(7)	1(1)	27(7)	69(35)	27(28)	7(7)
$a = 0.7$	13(7)	0(0)	37(6)	28(9)	1(1)	24(7)	67(44)	25(33)	5(8)
$a = 0.9$	14(13)	0(0)	36(13)	29(17)	0(0)	20(15)	72(51)	27(41)	3(4)

from 100 replications in Table 5. It is clear that except the very challenging case with sparse signal (small π) and low intensity level (small μ), the empirical probability is generally small, which shows the control of the probability for having a small proportion of false negatives by AFNC.

Table 5: Empirical probability of $FN(\hat{T}_{fn})/s > \epsilon_n$ with $\epsilon_n = 1/\ln(n)$.

	$\pi = 5\%$	$\pi = 10\%$	$\pi = 20\%$
$\mu = 2.5$	0.29	0.05	0
$\mu = 3.5$	0.02	0	0
$\mu = 4.5$	0	0	0

5 Application to DNA Structural Variation Analysis

We apply the trichotomous framework to analyze DNA structural variation based on high-throughput SNP array data. Here, we focus on an important type of structural variants called copy number variants (CNVs), which play important roles in population diversity and disease association [25]. Our dataset is generated from Illumina HumanHap 550 array, where genotypes are measured at $\sim 500,000$ SNP locations along the human genome. It has been reported that most CNVs from the germline constitutional genome are very sparse and short, ranging less than 20 SNPs [35]. Many of these subtle signals cannot reach significance levels of multiple testing in

genome-wide search.

In this paper, instead of presenting a list of candidates that pass certain significance level, we aim to provide more insight of the data by identifying the signal, noise, and mixed subsets. Our data includes three individuals from the Autism Genetics Resource Exchange (AGRE) collection [4]. We specifically consider the data on Chromosome 19, which include measurements at 9501 SNP locations for each individual. At each SNP location, the Log R ratio (LRR) is measured to represent the total intensity of both major and minor alleles. Due to the fact that LRR deviates from the baseline in CNV segments, LRR data are widely used for detecting CNVs [28].

For a given individual, LRR observations are first normalized, and then the likelihood ratio is calculated for each interval with length $\leq L$. The likelihood ratio of an interval is defined as the standardized sum of observations in that interval, and L is set at 20 as most of the CNVs cover less than 20 SNPs [35]. The efficiency and optimality of using likelihood ratios as the test statistics for CNV data have been studied in Jeng et al. [17]. There are $n = 9501 \times 20 = 190,020$ such likelihood ratio statistics for each individual. When the distributions of LRRs change in an interval, the corresponding likelihood ratio for that interval is expected to deviate from the baseline. Figure 3 demonstrates the empirical distribution of the $190,020$ likelihood ratios for an individual. The outliers at the left tail are likely to correspond to copy number deletions.

We calculate the p-values for these likelihood ratios assuming that the background noise follow $\mathcal{N}(0,1)$ after normalization. The likelihood ratios are locally dependent due to the fact that the intervals are short and overlapping. In this example we treat them as independent observations to illustrate the method. The robustness of our method under similar local dependence is shown in Example 4 of Simulation Studies. We find that estimating the signal proportion by (4) seems to result in a much larger proportion estimate than commonly expected for SNP array data, possibly due to artifacts involved in the data generation process [24]. Thus, we use a more reasonable bound of $0 \leq s \leq 50$ for this data set. Setting the upper bound at 50 means that the copy number deletions on Chromosome 19 are less than 50 [35]. Consequently,

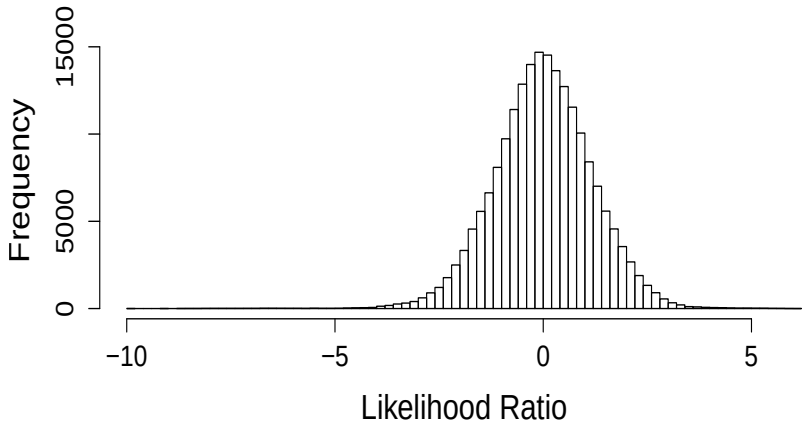

Figure 3: Histogram of the likelihood ratios of the intervals on Chromosome 19.

the AFNC procedure with $\tilde{T}_{fn}$ and $\gamma = 0.1$ is used to identify the mixed subset. To identify the signal subset, t_{bonf} with $\alpha = 0.1$ is used as it only select the true signals with high probability. Because the intervals are overlapping, we only keep intervals having minimum p-values among overlapping segments to indicate the locations of copy number deletions. All the other intervals overlapping with them are removed. t_{bonf} and $\tilde{T}_{fn}$ are then re-defined as the ranks among these non-overlapping intervals.

For the three individuals, $t_{bonf} = 1, 2, 1$ and $\tilde{T}_{fn} = 76, 18, 36$. These results show that only 1 or 2 candidates are strong enough to enter the signal subset. The candidates in between t_{bonf} and $\tilde{T}_{fn}$ include relatively weaker CNV signals, which are mixed indistinguishably with noise but desirable to be kept for follow-up study. Note that we start with $n = 190,020$ candidate intervals for each individual. The resulting $\tilde{T}_{fn}$ greatly reduce the number of tests in the follow-up study.

To exam whether true CNVs are included in the identified mixed subset between t_{bonf} and $\tilde{T}_{fn}$, we compare the candidates in each identified subset to the reported members in a CNV database maintained in The Centre for Applied Genomics (http://projects.tcag.ca/variation/project.html). A candidate region can overlap with zero, one, or more than one CNVs in the database. The mean value of the number of such CNVs in the database is presented for each subset in Table 6. Larger mean value represents stronger evidence for identifying true CNVs in a subset. In other words, let O_j = number of CNVs in the database that overlap with the j-

th candidate in the list of ranked intervals. Define ovlap-S = mean($O_j, 1 \leq j \leq t_{bonf}$, ovlap-M = mean($O_j, t_{bonf} < j \leq \tilde{T}_{fn}$), ovlap-N = mean($O_j, \tilde{T}_{fn} < j \leq$ total number of intervals). Table 6 shows that these mean values, in general, decrease from ovlap-S to ovlap-N. This agrees with our intuition. For example, the result of individual 3 shows that, on average, each candidate in the identified mixed subset overlaps with 6.8 CNVs in the database, while the number decreases to 2.0 for the identified noise subset. If we only select candidates based on false positive control, many of the true signals in the mixed subset will be missed for follow-up study. The sample correlation between the interval length and O_j are 0.17, 0.28, and 0.26 for the three individuals, respectively, indicating that the trend observed in Table 6 is not likely caused by the length factor.

Table 6: t_{bonf}, $\tilde{T}_{fn}$, and the mean value of O_j in each subset.

	ovlap-S	t_{bonf}	ovlap-M	$\tilde{T}_{fn}$	ovlap-N
Individual 1	4	1	4.7	76	2.1
Individual 2	10.5	2	3.4	18	2.5
Individual 3	0	1	6.8	36	2.0

6 Discussion

In this paper, we study the natural formation of three subsets when analyzing high dimensional data. An efficient screening procedure is proposed to retain the relatively weak signals in the mixed subset. This study provides important insights for the data and helps practitioners to quickly remove irrelevant information. Furthermore, the results of the trichotomous analysis can help practitioners to design more efficient follow-up studies. For example, unlike traditional sample-size calculations that need to pre-specify signal intensity level, we may infer the number and intensities of the signals based on the identified $\mathbb{M}$ and determine the sample size needed in a follow-up study objectively. Works along this line are deferred to future research. Additional insight for the quality of the data may also be achieved by examining the identified $\mathbb{M}$. For example, a large $\mathbb{M}$ suggests that there may exist many small non-null ob-

servations, which are either true signals or, very often, caused by artifacts involved during the data generation. Investigating the sources of possible artifacts in follow-up studies may significantly reduce the size of $\mathbb{M}$ and result in better separation between signals and noise.

The study in this paper is based on p-values. Other statistics carrying information about signal intensity, such as the local FDR values [10, 30] may be used in place of p-values. It will be interesting to investigate these possibilities in future research.

In this paper, we assumed independent p-values to allow a succinct theoretical study of the new method. Simulation examples in section 4 demonstrate the robustness of the proposed method for autocorrelated observations. We plan to study in depth the three-subset identification under dependence in future works. We find the works in multiple testing under dependence very helpful. Examples include Leek and Storey [23], Sun and Cai [31], Friguet et al. [14], Fan et al. [11], etc. According to these works, it is possible to adjust for the dependence structure of the test statistics to better separate signals from noise.

Last but not least, the identification of $\mathbb{S}$, $\mathbb{M}$, and $\mathbb{N}$ can be related to the problems of variable selection and variable screening in high-dimensional regression [32, 12, 13]. Existing studies show that under certain conditions, a pre-fixed number of top-ranked candidates include all the relevant predictors with high probability. In other words, those top-ranked candidates include the union of $\mathbb{S}$ and $\mathbb{M}$. However, how to efficiently choose the pre-fixed screening parameter remains an open question. We expect that the data-driven construction of $\hat{T}_{fn}$ may shed light on this open question. Furthermore, recent studies in high-dimensional regression investigate the challenging problem of controlling certain Type I error rate when performing variable selection. For example, Meinshausen and Bühlmann [27] consider the control of familywise error rate (FWER), Barber and Candès [1] consider the control of false discovery rate (FDR), and Ji and Zhao [20] consider the control of marginal false discovery rate (mFDR). Optimal multiple testing procedures for high-dimensional regression have also been studied in Ji and Zhao [20]. We find these works very interesting and would like to explore their connections to the problem of identifying the signal, noise, and mixed subsets in high-dimensional regression.

7 Proofs

7.1 Proof of Theorem 1

It is sufficient to show that

$$\Pr\{FN(T_{fn}) \geq 1\} \leq \gamma \tag{20}$$

and, for any $\tilde{k} < T_{fn}$,

$$\Pr\{FN(\tilde{k}) \geq 1\} > \gamma. \tag{21}$$

For (20), define $\hat{j} = \min\{j \geq 1 : p_{(s+j)} \leq d_{s,j}(\gamma)\}1_{\{s>t_1\}}$, then $T_{fn} = s + \hat{j}$. Define $TP(i)$ and $FP(i)$ as the numbers of true positives and false positives for selecting the top $\{1, \ldots i\}$ candidates. We have

$$\begin{aligned} \Pr\{FN(T_{fn}) \geq 1\} &= \Pr\{TP(T_{fn}) < s\} \\ &= \Pr\{FP(T_{fn}) > T_{fn} - s\} = \Pr\{FP(T_{fn}) > \hat{j}\}, \end{aligned} \tag{22}$$

where the second equality is by $T_{fn} = FP(T_{fn}) + TP(T_{fn})$.

In the case of $s \leq t_1$, we have $\hat{j} = 0$ and $T_{fn} = s$. Then

$$\Pr\{FP(T_{fn}) > \hat{j}\} = \Pr\{FP(s) > 0\} \leq \Pr\{FP(t_1) > 0\}. \tag{23}$$

Define $P^{n-s}_{(j)}$ as the j-th smallest p-value from $n-s$ noise and $P^{n}_{(j)}$ as the j-th smallest p-value from n noise. By the construction of t_1,

$$\Pr\{FP(t_1) > 0\} = \Pr\{P^{n-s}_{(1)} < p_{(t_1)}\} \leq \Pr\{P^{n-s}_{(1)} < d_{s,1}(\gamma)\} = \gamma. \tag{24}$$

Combining (22) - (24) gives (20).

In the case of $s > t_1$,

$$\Pr\{FP(T_{fn}) > \hat{j}\} = \Pr\{P^{n-s}_{(\hat{j})} < p_{(T_{fn})}\} \leq \Pr\{P^{n-s}_{(\hat{j})} < d_{s,j}(\gamma)\} = \gamma, \tag{25}$$

where the first step is because when the elements from S_0 are more than $\hat{j}$ in $\{1, \ldots, T_{fn}\}$, the $\hat{j}$-th smallest P_j from S_0 must rank before $p_{(T_{fn})}$, and the second step is by the construction of T_{fn}. Combining (22) and (25) gives (20).

Next, consider (21). Similar calculation as in (22) gives

$$\Pr\{FN(\tilde{k}) \geq 1\} = \Pr\{FP(\tilde{k}) > \tilde{j}\}, \tag{26}$$

where $\tilde{k} = s + \tilde{j}$. The construction of T_{fn} in (2) implies that $p_{(\tilde{k})} > d_{s,\tilde{j}}(\gamma)$, then

$$\Pr\{FP(\tilde{k}) > \tilde{j}\} \;=\; \Pr\{P^{n-s}_{(\tilde{j})} < p_{(\tilde{k})}\} > \Pr\{P^{n-s}_{(\tilde{j})} < d_{s,\tilde{j}}(\gamma)\} = \gamma. \tag{27}$$

(21) follows from (26) and (27).

7.2 Proof of Theorem 2

Let $\hat{j} = \min\{j \geq 1 :\ p_{(\hat{\pi}n+j)} \leq d_{\hat{\pi}n,j}(\gamma)\}1_{\{\hat{\pi}n > t_{\hat{\pi},1}\}}$. Then, it can be shown that

$$\begin{aligned}
& \Pr\{FN(\hat{T}_{fn})/s > \epsilon\} \\
= \;& \Pr\{TP(\hat{T}_{fn}) < (1-\epsilon)s\} = \Pr\{FP(\hat{T}_{fn}) > \hat{T}_{fn} - (1-\epsilon)s\} \\
= \;& \Pr\{FP(\hat{T}_{fn}) > \hat{\pi}n + \hat{j} - (1-\epsilon)\pi n\} \\
\leq \;& \Pr\{FP(\hat{T}_{fn}) > \hat{\pi}n + \hat{j} - (1-\epsilon)\pi n, \hat{\pi} \geq (1-\epsilon)\pi\} + \Pr\{\hat{\pi} < (1-\epsilon)\pi\}
\end{aligned} \tag{28}$$

The following lemma summarizes part of the results in Theorem 1 and 2 in [26]. The proof of the lemma is omitted.

Lemma 7.1 *Assume the same conditions as in Theorem 2. Let $\hat{\pi}$ be defined as in (4). Then, as $n \to \infty$,*

$$\Pr\{(1-\epsilon)\pi \leq \hat{\pi} \leq \pi\} \to 1$$

for any $\epsilon > 0$.

Then it is easy to see that

$$\begin{aligned}
& \Pr\{FP(\hat{T}_{fn}) > \hat{\pi}n + \hat{j} - (1-\epsilon)\pi n, \hat{\pi} \geq (1-\epsilon)\pi\} + \Pr\{\hat{\pi} < (1-\epsilon)\pi\} \\
\leq \;& \Pr\{FP(\hat{T}_{fn}) > \hat{j}\} + o(1).
\end{aligned} \tag{29}$$

When $\hat{\pi}n \leq t_{\hat{\pi},1}$, similar arguments as in (23) - (24) give

$$\Pr\{FP(\hat{T}_{fn}) > \hat{j}\} \leq \Pr\{P^{n-s}_{(1)} < d_{\hat{\pi}n,1}(\gamma)\}.$$

Since

$$\begin{aligned}\Pr\{P_{(1)}^{n-s} < d_{\hat{\pi}n,1}(\gamma)\} &\leq \Pr\{P_{(1)}^{n-s} < d_{\hat{\pi}n,1}(\gamma), \hat{\pi} < \pi\} + \Pr(\hat{\pi} > \pi) \\ &\leq \Pr\{P_{(1)}^{n-s} < d_{s,1}(\gamma)\} + o(1) = \gamma + o(1). \end{aligned} \tag{30}$$

Then (8) follows.

When $\hat{\pi}n > t_{\hat{\pi},1}$,

$$\begin{aligned}\Pr\{FP(\hat{T}_{fn}) > \hat{j}\} &= \Pr\{P_{(\hat{j})}^{n-s} < p_{(\hat{T}_{fn})}\} \leq \Pr\{P_{(\hat{j})}^{n-s} < d_{\hat{\pi}n,\hat{j}}(\gamma)\} \\ &\leq \Pr\{P_{(\hat{j})}^{n-s} < d_{\hat{\pi}n,\hat{j}}(\gamma), \hat{\pi} \leq \pi\} + \Pr(\hat{\pi} > \pi) \\ &\leq \Pr\{P_{(\hat{j})}^{n-s} < d_{s,\hat{j}}(\gamma)\} + o(1) = \gamma + o(1).\end{aligned}$$

Then (8) follows.

7.3 Proof of Theorem 3

Defines events $A = \{t_{\hat{\pi},1} \geq \hat{\pi}n\}$, $B = \{J_1 = \pi n\}$, $C = \{J_2 = \pi n\}$, and $D = \{\hat{\pi}n = \pi n\}$. It is enough to show that

$$\Pr(A \cap B \cap C \cap D) \to 1,$$

which is implied by

$$\Pr(A^c) + \Pr(B^c) + \Pr(C^c) + \Pr(D^c) \to 0. \tag{31}$$

Consider $\Pr(A^c)$ first. By the definition of $t_{\hat{\pi},1}$ and Lemma (7.1),

$$\begin{aligned}\Pr(A^c) &\leq \Pr(t_{\hat{\pi},1} < \pi n) + \Pr(\hat{\pi} > \pi) \\ &\leq \Pr\{\exists i \in S_1 : P_i > d_{\hat{\pi}n,1}(\gamma_n)\} + o(1) \\ &\leq s\bar{G}\{d_{\hat{\pi}n,1}(\gamma_n)\} + o(1) \\ &\leq s\bar{G}(n^{-r}) + o(1) = o(1),\end{aligned}$$

where the fourth inequality is by $d_{\hat{\pi}n,1}(\gamma_n) \geq C\gamma_n n^{-1}$ and $\gamma_n > n^{-(r-1)}$, and the last step is by the condition $s\bar{G}(n^{-r}) \to 0$.

For $\Pr(B^c)$, by the definition of J_1, J_1 is always less or equal to s, then

$$\begin{aligned}\Pr(B^c) &= \Pr(J_1 < \pi n) = \Pr(\exists i \in S_1 : P_i > P_{(1)}^{n-s}) \\ &\leq \Pr(\exists i \in S_1 : P_i > n^{-r}) + \Pr(P_{(1)}^{n-s} < n^{-r}) \\ &\leq s\bar{G}(n^{-r}) + o(1) = o(1).\end{aligned}$$

Similar argument can be applied to show $\Pr(C^c) = o(1)$

Now consider $\Pr(D^c)$. By lemma 7.1, it is enough to show that

$$\Pr(\hat{\pi}n \leq \pi n - 1) \to 0,$$

which is implied by

$$\Pr(\frac{\hat{\pi}}{\pi} - 1 < -\frac{1}{\pi n}) \to 0. \tag{32}$$

Define

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1(P_i \leq t), \qquad U_{n-s}(t) = \frac{1}{n-s}\sum_{i=1}^{n-s} 1(P_i^0 \leq t), \qquad G_s(t) = \frac{1}{s}\sum_{i=1}^{s} 1(P_i^1 \leq t).$$

Then, by the construction of $\hat{\pi}$ in (4), for any $t \in [0,1]$,

$$\begin{aligned}\frac{\hat{\pi}}{\pi} - 1 &\geq \frac{F_n(t) - t - \pi}{\pi} - \frac{\sqrt{2\ln\ln n}\sqrt{t(1-t)}}{\pi\sqrt{n}} \\ &= \frac{(1-\pi)U_{n-s}(t) + \pi G_s(t) - t - \pi}{\pi} - \frac{\sqrt{2\ln\ln n}\sqrt{t(1-t)}}{\pi\sqrt{n}} \\ &= \{G(t) - 1\} + \{G_s(t) - G(t)\} + \frac{1-\pi}{\pi}\{U_{n-s}(t) - t\} - t - \frac{\sqrt{2\ln\ln n}\sqrt{t(1-t)}}{\pi\sqrt{n}}.\end{aligned}$$

Let $t = n^{-r}$. Then by condition $s\bar{G}(n^{-r}) \to 0$ and $r > 1$,

$$|G(t) - 1)| = \bar{G}(n^{-r}) = o(\frac{1}{s}),$$

$$|G_s(t) - G(t)| = O_p(\sqrt{\frac{G(t)(1-G(t))}{s}}) = O_p(\sqrt{\frac{\bar{G}(n^{-r})}{s}}) = o_p(\frac{1}{s}),$$

$$\frac{1-\pi}{\pi}|U_{n-s}(t) - t| = O_p(\frac{1-\pi}{\pi}\sqrt{\frac{t(1-t)}{n-s}}) = O_p(\frac{\sqrt{1-\pi}}{\pi}\frac{1}{n^{(1+r)/2}}) = o_p(\frac{1}{s}),$$

$$\frac{\sqrt{2\ln\ln n}\sqrt{t(1-t)}}{\pi\sqrt{n}} = \frac{\sqrt{2\ln\ln n}}{\pi}\frac{1}{n^{(1+r)/2}} = o(\frac{1}{s}).$$

Therefore, (32) follows. Combining the above results for $\Pr(A^c)$, $\Pr(B^c)$, $\Pr(C^c)$, and $\Pr(D^c)$ gives (31).

7.4 Proof of Theorem 4

For notation simplicity, let $n_0 = n - s$.

Consider (11) first. It can be shown that

$$\begin{aligned}\Pr(\mathbb{S} = \emptyset) &= \Pr(\max\{X_i, i \in S_1\} \le \max\{X_i, i \in S_0\}) \\ &\le \Pr(\max\{X_i, i \in S_1\} \le \sqrt{2\ln n_0}) + \Pr(\max\{X_i, i \in S_0\} > \sqrt{2\ln n_0}) \quad (33)\end{aligned}$$

where

$$\Pr(\max\{X_i, i \in S_0\} > \sqrt{2\ln n_0}) \le n_0 \Pr\{\mathcal{N}(0,\sigma^2) > \sqrt{2\ln n_0}\} \le C/\sqrt{\ln n_0} = o(1) \quad (34)$$

and

$$\begin{aligned}&\Pr(\max\{X_i, i \in S_1\} \le \sqrt{2\ln n_0}) \\ =\ &\Pr(\max\{X_i, i \in S_1\} - \mu \le \sqrt{2\ln n_0} - \mu) \\ =\ &\Pr\{\max\{X_i, i \in S_1\} - \mu \le \sigma\sqrt{2\ln s} + (\sqrt{2\ln n_0} - \mu - \sigma\sqrt{2\ln s})\} \\ \le\ &\Pr[\max\{X_i, i \in S_1\} - \mu \le \sigma\sqrt{2\ln s} + \{\sqrt{2\ln n_0} - \sqrt{2(1+\epsilon)\ln n_0}\}] \\ \le\ &\Pr(\max\{X_i, i \in S_1\} - \mu \le \sigma\sqrt{2\ln s} - C\sqrt{\ln n_0}) = o(1), \quad (35)\end{aligned}$$

where the first inequality is by condition (11) and the last step is by the extreme value theory of normal random variables. Combining (33) - (35) gives (11).

Next consider (12). It can be shown that

$$\begin{aligned}&\Pr(\mathbb{S} \ne \emptyset) \\ =\ &\Pr(\max\{X_i, i \in S_1\} > \max\{X_i, i \in S_0\}) \\ \le\ &\Pr(\max\{X_i, i \in S_1\} > \sqrt{2\ln n_0} - \ln\ln n_0) + \Pr(\max\{X_i, i \in S_0\} < \sqrt{2\ln n_0} - \ln\ln n_0) \\ \le\ &\Pr(\max\{X_i, i \in S_1\} > \sqrt{2\ln n_0} - \ln\ln n_0) + o(1), \quad (36)\end{aligned}$$

where the second inequality is by the extreme value theory of standard normal random variables. Also,

$$\begin{aligned}&\Pr(\max\{X_i, i \in S_1\} > \sqrt{2\ln n_0} - \ln\ln n_0) \\ =\ &\Pr\{\max\{X_i, i \in S_1\} - \mu > \sigma\sqrt{2\ln s} + (\sqrt{2\ln n_0} - \ln\ln n_0 - \mu - \sigma\sqrt{2\ln s})\} \\ \le\ &\Pr[\max\{X_i, i \in S_1\} - \mu > \sigma\sqrt{2\ln s} + \{\sqrt{2\ln n_0} - \ln\ln n_0 - \sqrt{2(1-\epsilon)\ln n_0}\}] \\ \le\ &\Pr(\max\{X_i, i \in S_1\} - \mu > \sigma\sqrt{2\ln s} + C\sqrt{\ln n_0}) = o(1), \quad (37)\end{aligned}$$

where the first inequality is by condition (12). Combining (36) and (37) gives (12).

The claims in (13) - (16) can be proved similarly.

Acknowledgments

The author thanks Dr. Leonard Stefanski and Dr. John Daye for helpful discussions and comments.

References

[1] Barber, R. F. and Candès, E. J. (2015), "Controlling the false discovery rate via knockoffs," *Ann. Statist.*, 43, 2055–2085.

[2] Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. Royal Stat. Soc. B*, 57, 289–300.

[3] — (2000), "On the adaptive control of the false discovery rate in multiple testing with independent statistics," *J. of Educational and Behavioral Statistics*, 25, 60–83.

[4] Bucan, M., Abrahams, B., Wang, K., Glessner, J., Herman, E., Sonnenblick, L., Retuerto, A. A., Imielinski, M., Hadley, D., Bradfield, J., Kim, C., Gidaya, N., Lindquist, I., Hutman, T., , Sigman, M., Kustanovich, V., Lajonchere, C., Singleton, A., Kim, J., , Wassink, T., McMahon, W., Owley, T., Sweeney, J., Coon, H., Nurnberger, J., Li, M., Cantor, R., Minshew, N., Sutcliffe, J., Cook, E., Dawson, G., Buxbaum, J., Grant, S., Schellenberg, G., Geschwind, D., and Hakonarson, H. (2009), "Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes," *PLoS Genetics*, 5(6), e1000536.

[5] Cai, T., Jin, J., and Low, M. (2007), "Estimation and confidence sets For sparse normal mixtures," *Ann. Statist.*, 35, 2421–2449.

[6] Cai, T. T., Jeng, X. J., and Jin, J. (2011), "Optimal detection of heterogeneous and heteroscedastic mixtures," *J. Royal Stat. Soc. B*, 73, 629–662.

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

[8] Drton, M. and Perlman, M. D. (2008), "A SINful approach to Gaussian graphical model selection," *J. of Statistical Planning and Inference*, 138(4), 1179–1200.

[9] Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple hypothesis testing in microarray experiments," *Statist. Sci.*, 18(1), 71–103.

[10] Efron, B. (2007), "Size, power and false discovery rates," *Ann. Statist*, 35, 1351–1377.

[11] Fan, J., Han, X., and Gu, W. (2012), "Control of the false discovery rate under arbitrary covariance dependence," *J. Am. Statist. Ass.*, 107, 1019–1045.

[12] Fan, J. and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space," *J. Royal Stat. Soc. B*, 70, 849–911.

[13] Fan, Y. and Tang, C. (2013), "Tuning parameter selection in high dimensional penalized likelihood," *J. Royal Stat. Soc. B*, 75, 531–552.

[14] Friguet, C., Kloareg, M., and Causeur, D. (2009), "A factor model approach to multiple testing under dependence," *J. Am. Statist. Ass.*, 488, 1406–1415.

[15] Genovese, C. and Wasserman, L. (2004), "A stochastic process approach to false discovery control," *Ann. Statist*, 32, 1035–1061.

[16] Han, B., Kang, H. M., and Eskin, E. (2009), "Rapid and accurate multiple testing correction and power estimation for millions of correlated markers," *PLos Genet.*, 5, 1–13.

[17] Jeng, X. J., Cai, T. T., and Li, H. (2010), "Optimal sparse segment identification with application in copy number variation analysis," *J. Am. Statist. Ass.*, 105, 1156–1166.

[18] Jeske, D., Liu, Z., Bent, E., and Borneman, J. (2007), "Classification rules that include neutral zones and their application to microbial community profiling," *Communication in Statistics - Theory and Methods*, 36(10), 1965–1980.

[19] Ji, P. and Jin, J. (2012), "UPS delivers optimal phase diagram in high-dimensional variable selection," *Ann. Statist.*, 40, 73–103.

[20] Ji, P. and Zhao, Z. (2014), "Rate optimal multiple testing procedure in high-dimensional regression," arXiv:1404.2961.

[21] Jin, J. (2008), "Proportion of nonzero normal means: oracle equivalence and uniformly consistent estimators," *J. Royal Stat. Soc. B*, 70, 461–493.

[22] Jin, J. and Cai, T. (2007), "Estimating the null and the proportion of non-null effects in large-scale multiple comparisons," *J. Am. Statist. Ass.*, 102, 495–506.

[23] Leek, J. and Storey, J. (2008), "A general framework for multiple testing dependence," *Proc Natl Acad Sci*, 105, 18718–18723.

[24] Marioni, J. C., Thorne, N. P., Valsesia, A., Fitzgerald, T., Redon, R., Fiegler, H., Andrews, T. D., Stranger, B. E., Lynch, A. G., Dermitzakis, E. T., Carter, N. P., Tavar, S., and Hurles, M. E. (2007), "Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization," *Genome Biology*, 8(10), R228.

[25] McCarroll, S. S. and Altshuler, D. M. (2007), "Copy-number variation and association studies of human disease," *Nature Genetics*, 39, S37–S42.

[26] Meinshausen, M. and Rice, J. (2006), "Estimating the proportion of false null hypotheses among a large number of independent tested hypotheses," *Ann. Statist.*, 34, 373–393.

[27] Meinshausen, N. and Bühlmann, P. (2010), "Stability Selection," *J. Royal Stat. Soc. B*, 72, 417–473.

[28] Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C. A., Belmont, J., Cheung, S. W., Shen, R. M., Barker, D. L., and Gunderson, K. L. (2006), "High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping," *Genome Res*, 16, 1136–1148.

[29] Skol, A., Scott, L., Abecasis, G., and Boehnke, M. (2006), "Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies," *Nature Genetics*, 38, 209–213.

[30] Sun, W. and Cai, T. (2007), "Oracle and adaptive compound decision rules for false discovery rate control," *J. Am. Statist. Ass.*, 102, 901–912.

[31] — (2009), "Large-scale multiple testing under dependency," *J. Royal Stat. Soc. B*, 71, 393–424.

[32] Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc. B*, 58, 267–288.

[33] Xie, J., Cai, T., and Li, H. (2011), "Sample size and power analysis for sparse signal recovery in genome-wide association studies," *Biometrika*, 98, 273–290.

[34] Zaykin, D. V. and Zhivotovsky, L. A. (2005), "Ranks of genuine associations in whole-genome scans," *Genetics*, 171, 813–823.

[35] Zhang, F., Gu, W., Hurles, M., and Lupski, J. (2009), "Copy number variation in human health, disease and evolutions," *Annual Review of Genomics and Human Genetics*, 10, 451–481.